Expression of the Genes for

Arylamine N-acetyltransferases in Mice

Submitted for the degree of Doctor of Philosophy

Hilary Term 2002

Sotiria Boukouvala

St. Peter's College

Department of Pharmacology

University of Oxford

Supervisor: Professor Edith Sim

To Giannoulis and my family

Expression of the Genes for Arylamine N-acetyltransferases in Mice

Sotiria Boukouvala, St. Peter's College, University of Oxford

Submitted for the degree of Doctor of Philosophy, Hilary Term, 2002

ABSTRACT

Arylamine N-acetyltransferases (NATs) are polymorphic enzymes involved in the metabolism of arylamine and hydrazine xenobiotics. Murine NAT isoenzymes are encoded by three *Nat* genes. Only *Nat2* was previously known to be polymorphic, a single nucleotide substitution causing the slow acetylator phenotype in the A/J strain. The present study (Chapter 3) describes novel polymorphisms in all three *Nat* genes of the wild-derived inbred strains *Mus spretus* and *Mus musculus castaneus*. Functional analysis of hepatic and heterologously expressed NAT variants from the two strains demonstrated that *M.m. castaneus* is a fast and *M. spretus* a slow acetylator.

A previously isolated 14.3kb *Nat*-positive mouse genomic clone of 129/Ola strain origin (clone A) was sequenced (Chapter 4). A 8.6kb *Hind*III fragment, containing the entire *Nat2* coding region, was subcloned from clone A and used to generate a targeting construct for the production of *Nat2* knock-out mice. Probes and appropriate PCR methodologies were developed for screening of embryonic stem cells for targeted incorporation of the construct (Chapter 4). Computational analysis of clone A sequence revealed a number of important elements around the *Nat2* gene, including four microsatellite markers which were found to be polymorphic among different mouse strains. Combined with preliminary physical mapping work (Chapter 4), these markers will assist accurate localisation of the *Nat* genes on mouse chromosome 8.

A non-coding exon was mapped 6.4-6.1kb upstream of the intronless *Nat2* coding region. A functional polyadenylation signal was also identified 0.45kb downstream of the mouse *Nat2* coding region. The genomic structure of other genes for mammalian NAT was also analysed, by comparison of ESTs and genomic sequences deposited in electronic databases. Reverse transcription PCR confirmed that *Nat2* is expressed in many tissues, while *Nat1* and *Nat3* are expressed in the liver and spleen, respectively. The *Nat2* transcript was also detected in mouse embryonic stem cells, suggesting a possible involvement of murine NAT2 early in development (Chapter 5). The elements constituting the core promoter of *Nat2* were characterised, and a preliminary search for other transcriptional regulatory sequences was carried out, using reporter gene assays and electrophoretic mobility shift assays (Chapter 6).

Acknowledgements

I am grateful to my supervisor Professor Edith Sim for giving me the opportunity to study in Oxford. Her guidance and support during the three years of hard work in the lab, as well as while writing this thesis are greatly appreciated. I also wish to thank the Action Research/SPARKS for providing generous funding, as well as the Bodossaki Foundation (Greece) and St. Peter's College (Oxford) for the award of a graduate scholarship.

I would like to thank Dr. Kathryn Plant for her kind and generous help, as well as Professor Nicholas Proudfoot, Dr. Shona Murphy and Dr. Jane Mellor for their expert advice. I am also thankful to Dr. Rodney Minchin for the long and stimulating discussions we had during his stay in Oxford. His opinion about NAT, Opera and restaurants around the city is always appreciated!

Warmest thanks go to all past and present members of the "Sim group" for transforming the dull bench work into an agreeable experience. Nikki and Anna, I'll miss our chats about shopping, boyfriends and diamonds! Akane, I have enjoyed your company in the evenings! Adeel, thanks for tolerating with my ruthless teasing! Fred, thanks for sharing the Birmingham canal cruise experience with me! I look forward to another cruise along the river Seine! James, molecular biologists are people of expensive tastes, so many thanks for doing all my shopping! Ed, I hope you never give up travelling! Valerie, Rupika and John, thanks for your useful hints! Also many thanks to Katalin, Sanjib, Carolyn, Cedric and Georgina, as well as to all the students who worked in the Sim lab during the past three and a half years. The help of Mrs V. Moar, K. Butler and the administration staff of the Department is also appreciated.

I would like to acknowledge the contribution of Naomi Price and wish her all the best with her D.Phil. Also thanks to a number of dear friends, especially Katerina and Myrto, who have additionally been valuable advisors on scientific matters.

Coming to Oxford would not have been possible without the efforts of my parents, Rigas and Ephie, who have inspired me with the enthusiasm to pursue knowledge. Thanks for your trust, patience and love. *Σας ευχαριστώ!* I am also grateful to Costas and Tetty for their enormous help and support. Finally, heartfelt thanks to Giannoulis, my dearest colleague and life's companion, the person with whom I shared all the good and bad moments of this effort.

Publications arising from this Thesis

Journal Articles

1. Smelt VA, Upton A, Adjaye J, Payton MA, Boukouvala S, Johnson N, Mardon HJ, Sim E (2000). Expression of arylamine N-acetyltransferases in pre-term placentas and in human pre-implantation embryos. *Human Molecular Genetics* **9:** 1101-1107.

2. Fakis G, Boukouvala S, Buckle V, Payton M, Denning C, Sim E (2000). Chromosome mapping of the genes for murine arylamine N-acetyltransferases (NATs), enzymes involved in the metabolism of carcinogens: identification of a novel upstream non-coding exon for murine *Nat2*. *Cytogenetics and Cell Genetics* **90:** 134-138.

3. Butcher N, Boukouvala S, Sim E, Minchin R (2002). Pharmacogenetics of the arylamine N-acetyltransferases. *The Pharmacogenomics Journal* **2:** 30-43.

4. Boukouvala S, Price N, Sim E (2002). Identification and functional characterisation of novel polymorphisms associated with the genes for arylamine N-acetyltransferases in mice. *Pharmacogenetics* **(in press)**.

Conference abstracts

1. Boukouvala S, Fakis G, Johnson N, Payton M, Smelt VA, Upton A, Adjaye J, Brook F, Priddle H, Sim E (1998). Investigation of the role of arylamine N-acetyltransferase, an enzyme of folate metabolism, in development. Presentation abstract for the 9th Mammalian Genetics and Development Workshop of the Genetical Society (London, UK). Published in *Genetical Research Cambridge* 1999; **74:** 102.

2. Boukouvala S, Fakis G, Pinter K, Buckle V, Priddle H, Sim E (1999). Expression and localisation of the genes for arylamine N-acetyltransferases in mice. Presentation abstract for the 10th Mammalian Genetics and Development Workshop of the Genetical Society (London, UK). Published in *Genetical Research Cambridge* 2000; **76:** 212.

3. Payton M, Boukouvala S, Upton A, Smelt V, Stanley L, Sim E (2000). Expression of arylamine N-acetyltransferase isoenzymes during development. Poster abstract for the Joint Symposium of the American Society for Biochemistry and Molecular Biology and the American Society for Pharmacology and Experimental Therapeutics (Boston, USA).

4. Boukouvala S, Sim E (2000). Tissue-specific expression of arylamine N-acetyltransferases in mice. Poster abstract for the 18th International Congress of Biochemistry and Molecular Biology (Birmingham, UK).

5. Boukouvala S, Price N, Sim E (2001). Investigation of polymorphism in and around the genes for murine arylamine N-acetyltransferases, enzymes of phase II drug metabolism. Poster abstract for the Human Genome Meeting (HUGO) (Edinburgh, UK).

6. Boukouvala S, Price N, Plant K, Sim E (2001). Structure and regulation of the murine homologue of the gene for human arylamine N-acetyltransferase 1, an enzyme involved in folate metabolism. Poster abstract for the 10th International Congress of Human Genetics (Vienna, Austria). Published in the *European Journal of Human Genetics* 2001; **9(S1):** 242.

7. Cornish V, Pinter K, Smith A, Boukouvala S, Mo M, Payton M, Sim E (2001). Murine arylamine N-acetyltransferase type 2-a knockout mouse progress report. Abstract for the 2nd International NAT Workshop (Oxford, UK).

8. Boukouvala S, Price N, Fakis G, Sim E (2001). Localisation, expression and novel functional polymorphism of the genes for arylamine N-acetyltransferases in mice. Presentation abstract for the 2nd International NAT Workshop (Oxford, UK).

List of Contents

List of Figures

List of Tables

Abbreviations

2-AF	2-aminofluorene
5-AS	5-aminosalicylic acid
acetyl-CoA	acetyl coenzyme A
APS	ammonium persulphate
ATCC	American Type Culture Collection
BAC	bacterial artificial chromosome
bHLH	basic helix-loop-helix
BLAST	Basic Local Alignment Search Tool
BRE	TFIIB-recognition element
BSA	bovine serum albumin
bZIP	basic leucine zipper
C/EBP	CCAAT enhancer binding protein
CASP	CTD-associated SR-like proteins
CBP	CREB-binding protein
cCAP	The Cancer Chromosome Aberration Project
CDD	The Conserved Domain Database
cDNA	complementary DNA
CF	cleavage factor
CGAP	The Cancer Genome Anatomy Project
CHO cells	Chinese hamster ovary cells
CIAP	calf intestine alkaline phosphatase
CMV	cytomegalovirus
COGs	Clusters of Orthologous Groups
CPSF	cleavage/polyadenylation specificity factor
CRE	cAMP response element
CREB	cAMP response element binding protein
CstF	cleavage stimulatory factor
CTD	C-terminal domain
CTF	CCAAT-box transcription factor
DCE	downstream core element
DDBJ	DNA Database of Japan
ddH_2O	double deionised water
DEPC	diethyl pyrocarbonate
dH_2O	deionised water
DHMHD	Dysmorphic Human-Mouse Homology Database
DIG	digoxigenin
DMAB	dimethylaminobenzaldehyde
DMBA	7,12-dimethylbenz[α]anthracene
DMEM	Dulbecco's modified Eagle's medium
DMSO	dimethyl sulphoxide
DNaseI	deoxyribonuclease I
dNTPs	deoxyribonucleotide triphosphates
DPE	downstream promoter element
DSE	downstream element
DTT	dithiothreitol
EBI	European Bioinformatics Institute
EDTA	ethylene-diamine-tetra-acetate
ELSI programme	"Ethical Legal and Social Implications" programme

EMBL	European Molecular Biology Laboratory
EMSA	electrophoretic mobility shift assay
ENU	N-ethyl-N-nitrosourea
ES cells	embryonic stem cells
ESE	exonic splicing enhancer
EST	expressed sequence tag
FBS	foetal bovine serum
FISH	fluorescence *in situ* hybridisation
GCG	Genetics Computer Group
GCV	ganciclovir
GEO	Gene Expression Omibus
GSS database	Genome Sequencing Survey database
GST	glutathione S-transferase
GTF	general transcription factor
HAT	histone acetyltransferase
HBS	HEPES-buffered saline
HGMP-RC	Human Genome Mapping Project Resource Center
HGP	Human Genome Project
hnRNP	heterogeneous ribonucleoprotein particles
HTG database	High-Throughput Genomic database
IHGSC	The International Human Genome Sequencing Consortium
Inr	initiator
IPTG	isopropyl-β-D-thiogalactopyranoside
IRES	internal ribosomal entry site
IRS	interspersed repeat sequence
LA-PCR	long and accurate PCR
LIF	leukaemia inhibitory factor
LINE	long interspersed element
LOH	loss of heterozygosity
LPS	lipopolysaccharide
LTR	long terminal repeat
MCA	*Mus musculus castaneus*
mEH	microsomal epoxide hydrolase
MEM	minimum essential medium
MGD	Mouse Genome Database
MGP	Mouse Genome Project
MKMD	Mouse Knock-out and Mutation Database
MMDB	The Molecular Modelling Database
MRC	Medical Research Council
***Ms*NAT**	*Mycobacterium smegmatis* NAT
MSP	*Mus spretus*
MTHFR	methylenetetrahydrofolate reductase
N-apABGlu	N-acetyl p-aminobenzoyl glutamate
NAT	arylamine N-acetyltransferase
NCBI	National Centre for Biotechnology Information
NCE	non-coding exon
NQO1	NADPH:quinone oxidoreductase 1
NTD	neural tube defect
OD	optical density
OMIM	Online Mendelian Inheritance in Man
P/CAF	p300/CBP-associated factor

pABA	p-aminobenzoic acid
pABGlu	p-aminobenzoyl glutamate
PABP II	polyA-binding protein II
PAC	P1 artificial chromosome
pANIS	p-anisidine
PAP	polyA-polymerase
pAS	p-aminosalicylic acid
pBS vector	pBluescript vector
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PDB	Protein Database
PhIP	2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine
PIC	preinitiation complex
PMSF	phenylmethylsulphonyl fluoride
polyA signal	polyadenylation signal
RACE	rapid amplification of cDNA ends
RAPD	randomly amplified polymorphic DNA
RFLP	restriction fragment length polymorphism
RNase	ribonuclease
RT-PCR	reverse transcription-polymerase chain reaction
SAGE	serial analysis of gene expression
SBTI	soybean trypsin inhibitor
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
SF1	splicing factor 1
SINE	short interspersed element
SMZ	sulphamethazine
SNP	single nucleotide polymorphism
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein particle
SR-proteins	Serine/Arginine-proteins
SRS	Sequence Retrieval System
SSCP	single strand conformational polymorphism
SSLP	simple sequence length polymorphism
SSR	simple sequence repeat
***St*NAT**	*Salmonella typhimurium* NAT
STS	sequence-tagged site
TAF	TBP-associated factor
TBP	TATA-binding protein
TCA	trichloroacetic acid
TCR	T-cell receptor
TEMED	N,N,N′,N′-tetramethylethylenediamine
TIC	TAF-Inr co-dependent co-factor
Tm	melting temperature
TSG	tumour suppressor gene
U2AF	U2snRNP auxiliary factor
UTR	untranslated region
UV	ultraviolet
VA RNA genes	virus associated RNA genes
x-gal	5-bromo-4-chloro-3-indoyl-β-D-galactopyranoside
YAC	yeast artificial chromosome

CHAPTER 1

Introduction

Arylamine N-acetyltransferase (NAT; Enzyme Commission number EC 2.3.1.5) is the enzymatic activity responsible for the N-acetylation of arylamine and hydrazine xenobiotics, including several drugs and carcinogens. The NAT isoenzymes form a protein family distinct from other acetyltransferases, such as the serotonin and histone acetyltransferases (Wade *et al.*, 1997; Borjigin *et al.*, 1999). This introductory Chapter will review the current knowledge of the enzymology and molecular genetics of human NATs, as well as their postulated involvement in chemical carcinogenesis and other disease conditions. The value of the laboratory mouse as a model for studying the role of human NATs will be discussed, with special reference to the contribution of the Mouse Genome Project. Finally, an overview of eukaryotic gene expression will be provided, as an essential background to the work described in subsequent experimental Chapters.

1.1 Xenobiotic metabolism and the history of pharmacogenetics

Humans are constantly exposed to xenobiotics, i.e. foreign compounds that enter the body, but do not constitute nutrients or essential factors of endogenous metabolism. Drugs, food additives, environmental pollutants and industrial chemicals are commonly encountered xenobiotics with potentially beneficial or harmful effects (Caldwell *et al.*, 1995). Upon ingestion, xenobiotic compounds become subject to metabolic transformation by a group of enzymes, referred to as xenobiotic metabolising enzymes. Xenobiotic metabolism is a biphasic process, comprising phase I (functionalisation) and phase II (conjugation) reactions (Williams, 1959). Phase I reactions, such as oxidation, reduction and hydrolysis, introduce a functional group (e.g. -OH, $-NH_2$) to the parent molecule, increasing its polarity and reactivity.

Phase II reactions, such as glucuronidation, sulphation, acetylation and methylation, involve covalent binding of small molecules (e.g. glutathione, SO_4^{2-}, CH_3CO-, CH_3-) to phase I metabolites or chemically suitable parent compounds. Xenobiotic metabolism is regarded as a detoxification process, responsible for the elimination of foreign compounds from the body. However, both phase I and phase II reactions can also lead to the bioactivation of otherwise harmless compounds, revealing their toxic or carcinogenic potential (Idle *et al.*, 1992; Smith *et al.*, 1994; Caldwell *et al.*, 1995; Delaforge, 1998). The fate of xenobiotics in the human body, from absorption to excretion, is outlined in figure 1.1. A summary of the reactions and enzymes of phase I and II metabolism is provided in table 1.1.

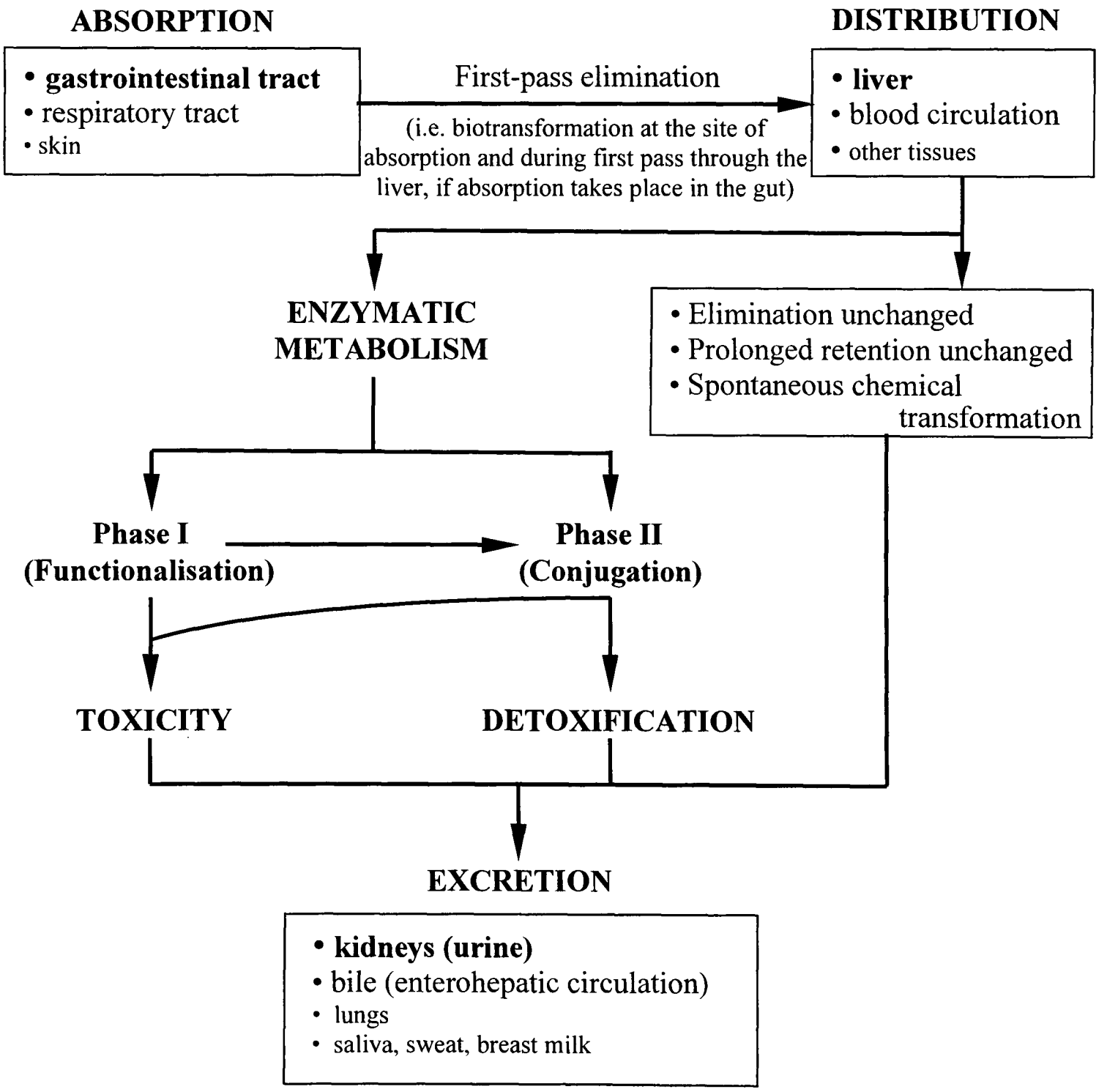


Figure 1.1: The fate of xenobiotics entering the human body. *(Based on Caldwell et al., 1995).*

Table 1.1: Common reactions and enzymes of xenobiotic metabolism. *(Based on Idle et al., 1992; Smith et al., 1994; Delaforge, 1998).*

Phase I		**Phase II**	
Reaction	**Enzyme**	**Reaction**	**Enzyme**
Oxidation	Cytochromes P450 Alcohol & aldehyde dehydrogenases Aromatases Mono-amine oxidases S-oxidase	Glucuronidation	UDP-glucuronosyl transferases
		Glucosidation	UDP-glucosyl transferases
		Sulphation	Sulphotransferases
Reduction	Azo-reductases Nitro-reductases N-oxide reductases	Methylation	Methyltransferases (N-, O, and S-)
		Acetylation	N-acetyltransferases
Hydrolysis	Epoxide hydrolases Carboxy-ester hydrolases Amidases Cholinesterases Arylesterases	Glutathione conjugation	Glutathione S-transferases
		Amino acid conjugation	N-acyl transferases
		Esterification	Fatty acid transferases

The concept that the commonly observed interindividual variation in the reaction to drugs must be under genetic control has long been appreciated. In 1931, A. Garrod suggested that such variation should be expected, because of the unique genetic constitution of each individual. In 1957, A. Motulsky attributed the side effects of drugs to genetically determined enzymatic deficiencies of certain subjects. In 1959, F. Vogel established the discipline of *Pharmacogenetics*, as the study of the genetic factors underlying differences in the response to drugs (La Du, 1992; Pfost *et al.*, 2000, for historical reviews). The term *Ecogenetics* was later introduced to describe similar variations in the response to environmental chemicals (Brewer, 1971), while *Xenogenetics* refers to the entirety of foreign chemicals, combining the above disciplines (Sim *et al.*, 1995). It is now established that many xenobiotic metabolising enzymes exhibit genetic polymorphism, the clinical consequences of which may be reduced drug efficacy, increased risk for adverse drug reactions or susceptibility to certain types of disease, especially cancer (May, 1994; Nebert, 1997).

In the past fifty years, numerous inherited traits affecting drug disposition have been characterised at the phenotypic, and more recently, the genotypic level. Carson *et al.* (1956) discovered that a deficiency in glucose-6-phosphate dehydrogenase was responsible for the susceptibility of certain individuals to

haemolytic anaemia, induced by the antimalarial drug primaquine. The prolonged muscle relaxation sometimes observed during treatment with suxamethonium was attributed to an inherited deficiency of plasma cholinesterase (Kalow and Genest, 1957). Polymorphism associated with the CYP2D6 enzyme of the P450 family has also been extensively studied as an important pharmacogenetic trait. Mahgoub *et al.* (1977) first observed that the exaggerated reaction of certain individuals to the standard dose of the antihypertensive agent debrisoquine is due to ineffective hydroxylation of the drug. The "poor metabolisers" of the drug were later identified to carry mutant *CYP2D6* alleles, encoding for defective variants of the enzyme (Skoda *et al.*, 1988). It is now established that *CYP2D6* polymorphisms are responsible for interindividual variation in the reaction to at least 30 common drugs, including β-blockers, antiarrhythmics, antidepressants, antipsychotics and opioids (Gonzalez and Nebert, 1990; May, 1994; Ingelman-Sundberg, 1998; Meyer, 2000). "Ultrarapid metabolisers", carrying multiple copies of the *CYP2D6* gene, have also been identified (Bertilsson *et al.*, 1993). Standard doses are generally inadequate to treat these individuals, who -on the other hand- are more prone to the side effects of drugs undergoing bioactivation by CYP2D6 in the body (e.g. codeine, which is converted into morphine) (Ingelman-Sundberg, 1998).

The N-acetylation polymorphism was originally identified as the genetic factor underlying variability in the response of tuberculosis patients to the drug isoniazid (Evans, 1992; Vatsis and Weber, 1994, for historical reviews). Early data suggested a bimodal distribution of the drug in the blood and urine. Individuals with high serum concentrations of the active drug excreted only small amounts of inactive metabolites in their urine and were more prone to neuropathy. On the other hand, individuals metabolising isoniazid efficiently were more resistant to the tuberculostatic effect of the drug and required higher doses for effective treatment. Mitchell and Bell (1957) suggested the classification of individuals as *fast* or *slow inactivators* of isoniazid. The *slow* phenotype was subsequently shown to be inherited as an autosomal co-dominant trait (Evans *et al.*, 1960). Following the work of Jenne (1965), NAT was identified as the enzymatic activity responsible for the inactivation of isoniazid and the term *acetylator* replaced *inactivator*. It was later identified that other drugs are also polymorphically acetylated by NAT, with the slow acetylator phenotype generally regarded as a risk factor for adverse reactions, including

hydralazine-induced systemic lupus erythematosus and dapsone-induced haemolysis (Weber and Hein, 1985; Evans, 1992). Since its discovery, the acetylation polymorphism has been recognised as a pharmacogenetic trait of major importance (Grant *et al.*, 2000; Sim *et al.*, 2000).

1.2 Arylamine N-acetyltransferases in humans

1.2.1 The human NAT isoenzymes

NAT is a phase II xenobiotic metabolising enzyme, catalysing the conjugation of an acetyl group to the amino group of primary arylamines and hydrazines or to the hydroxy group of hydroxyarylamines. The reaction is acetyl coenzyme A (acetyl-CoA)-dependent and proceeds in a two-step "ping-pong Bi-Bi" mechanism (Weber and Cohen, 1967; Riddle and Jencks, 1971). The first step involves binding of the acetyl group of acetyl-CoA to the sulphydryl group of an active site cysteine, to form a covalent acetyl-enzyme intermediate (a cysteinyl thioester). The second step involves transfer of the acetyl group from the enzyme to the substrate (Jencks *et al.*, 1972; Andres *et al.*, 1983a). A NAT-mediated acetyl-CoA-independent intramolecular N,O-acetyltransfer reaction has also been described with arylhydroxamic acid substrates (Grant, 1993; King *et al.*, 1997).

In addition to isoniazid, other clinically useful NAT substrates include the anti-arrhythmic drug procainamide, the chemotherapeutic agent aminoglutethimide, several anti-bacterial sulphonamides (e.g. sulphamethoxazole), the anti-inflammatory drug 5-aminosalicylate, the antihypertensive drug hydralazine, the monoamine oxidase inhibitor phenelzine, and dapsone, a drug used for the treatment of leprosy and malaria. The secondary arylamine or hydrazine metabolites of several drugs, such as sulphasalazine, acebutolol and caffeine, are also subject to N-acetylation by NAT. Of great importance is the role of NAT in the metabolism of many industrial or environmental carcinogens, including β-naphthylamine, benzidine, 2-aminofluorene and 4-aminobiphenyl (Weber and Hein, 1985; Evans, 1992; Hanna, 1996; Grant *et al.*, 2000). The chemical structure of important NAT substrates is shown in figure 1.2.

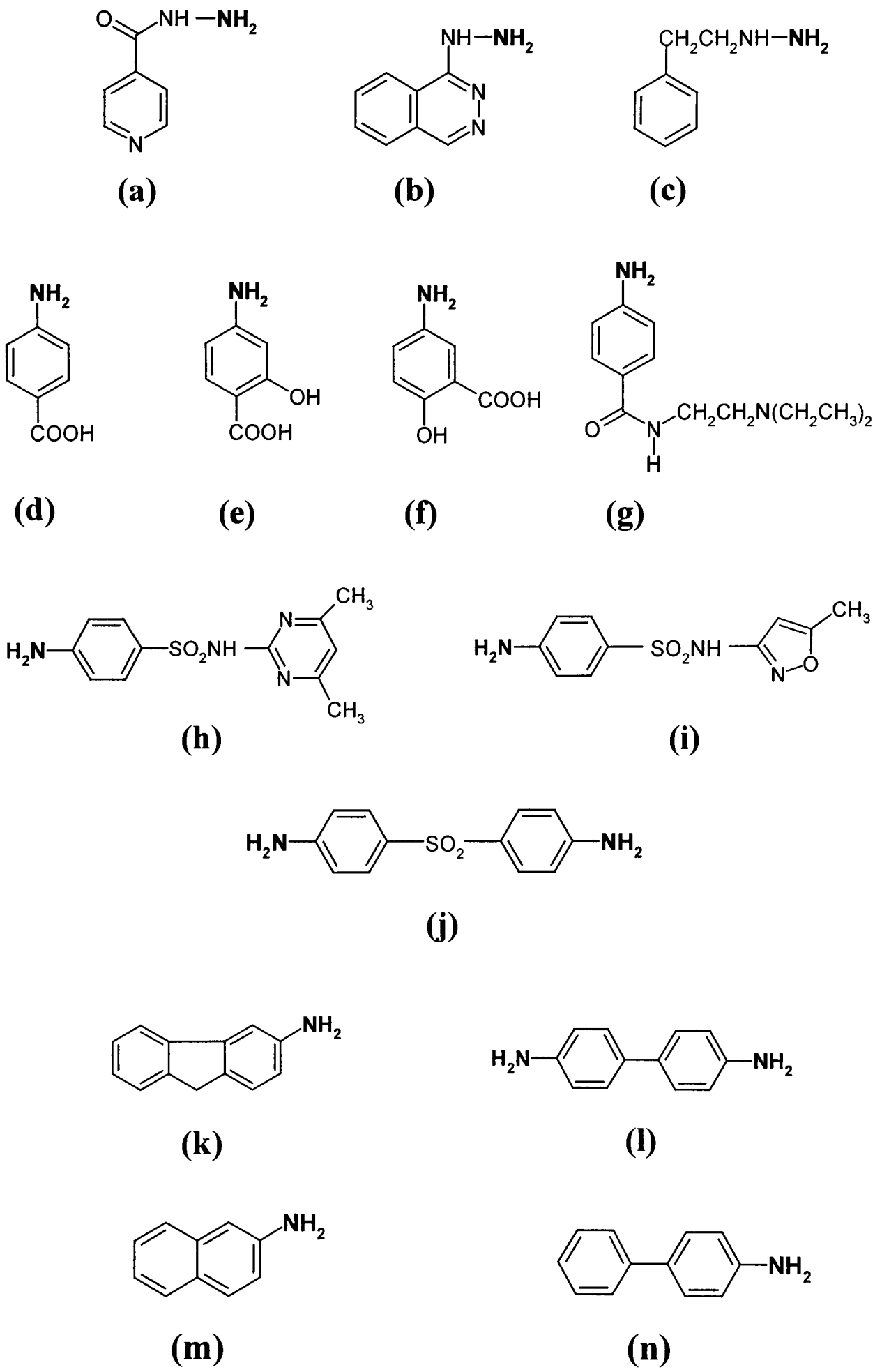


Figure 1.2: Chemical structure of common NAT substrates. Hydrazine drugs: a) isoniazid, b) hydralazine, c) phenelzine. Arylamine drugs: d) p-aminobenzoic acid, e) p-aminosalicylic acid, f) 5-aminosalicylic acid, g) procainamide, h) sulphamethazine, i) sulphamethoxazole, j) dapsone. Arylamine carcinogens: k) 2-aminofluorene, l) benzidine, m) β-naphthylamine, n) 4-aminobiphenyl. Amino groups in bold are acetylated by NAT. *(From Weber and Hein, 1985; Hanna, 1996).*

Early experiments measuring the metabolic inactivation rates of drugs in fast and slow acetylators indicated the presence of two acetyltransferase activities, responsible for the polymorphic and monomorphic acetylation of isoniazid and p-aminosalicylic acid (pAS), respectively (Jenne *et al.*, 1961; Jenne, 1965). Other studies demonstrated polymorphic acetylation of sulphamethazine (SMZ) and hydralazine, as opposed to the monomorphic acetylation of p-aminobenzoic acid (pABA) (Evans, 1964; Evans and White, 1964). The physicochemical and kinetic properties of human NAT were first described by Jenne (1965), who purified the enzyme 300-fold from post-mortem liver of fast and slow acetylators. Many years later, Grant *et al.* (1989b) separated two NAT enzymatic activities from the liver. However, even after 1000-fold purification, the enzyme preparations were less than 10% pure, indicating that NAT must represent a minor fraction of total hepatic protein. The partially purified enzymes had the same molecular weight of 31kDa, but differed in their capacity to acetylate SMZ. Following cloning and heterologous expression of the human *NAT* genes (Blum *et al.*, 1990a; Ohsako and Deguchi, 1990), it was realised that these two enzymes were variants of the same NAT isoenzyme (designated NAT2), responsible for the polymorphic acetylation of SMZ and procainamide (figure 1.2) (Grant *et al.*, 1991). The same investigators also isolated a 33kDa NAT isoenzyme (designated NAT1), responsible for the monomorphic acetylation of pAS and pABA (figure 1.2). With the introduction of molecular biology techniques, it soon became evident that NAT1 is also polymorphic (Vatsis and Weber, 1993). Consequently, the terms monomorphic and polymorphic NAT were replaced by the symbols NAT1 and NAT2, respectively (Vatsis *et al.*, 1995).

The use of cloning techniques, allowing the quick production of recombinant protein in bacterial expression systems, has opened the way towards detailed functional analysis of the human NAT isoenzymes. NAT1 and NAT2 proteins are 290 amino acids long and share 81% identity (Vatsis *et al.*, 1995). Using site-directed mutagenesis, it was established that a conserved cysteine at position 68 of mammalian NATs is essential for the formation of the acetyl-enzyme intermediate during catalysis (Dupret and Grant, 1992). Cloning and expression of the first 204 amino acids of human NAT1 further demonstrated that this part of the molecule is responsible for acetyl-CoA binding, but not acetyltransfer (Sinclair and Sim, 1997). Two arginines,

conserved at positions 9 and 64 of mammalian NATs, were also proposed to play a role in maintaining structural stability of the molecule (Deloménie *et al.*, 1997).

Expression of NAT1/NAT2 chimeric proteins identified regions determining the substrate selectivity of each NAT isoenzyme (Dupret *et al.*, 1994). More refined analysis, based on site-directed mutagenesis, further demonstrated that introduction of NAT2 residues at positions 125, 127 and 129 of recombinant NAT1 protein was enough to switch the selectivity of the molecule from NAT1 substrates (e.g. pAS) to NAT2 substrates (e.g. SMZ), and vice versa. It appears that the bulky Phe^{125}, Arg^{127} and Tyr^{129} residues of NAT1 may prevent binding of the relatively large substrates preferred by NAT2, which instead has three small serine residues at these sites (Goodfellow *et al.*, 2000).

A major advancement towards understanding of the NAT function at the molecular level has been the resolution of the three-dimensional structure of *Salmonella typhimurium* NAT (*St*NAT) (Sinclair *et al.*, 2000). The molecule consists of three domains (I-III) of roughly the same length, and a shorter interdomain between domains II and III. There are 9 α-helices (5 in domain I, 2 in the interdomain and 2 in domain III) and 14 β-sheets (2 in domain I, 8 in domain II and 4 in domain III) of variable length. The interface of domains II and III, together with the interdomain, form the active site cleft of the molecule. The Cys^{69} residue, which is equivalent to Cys^{68} of mammalian NAT (Watanabe *et al.* 1992), is located within the active site cleft and forms a catalytic triad with His^{107} and Asp^{122}. This triad is a characteristic feature of the cysteine protease superfamily and was an entirely unexpected finding. It appears that NATs share the same ancestral catalytic motif with cysteine proteases, evolved to perform acetyltransfer reactions instead of proteolytic cleavage (Sinclair *et al.*, 2000). The recently resolved structure of *Mycobacterium smegmatis* NAT (*Ms*NAT) is very similar to that of *St*NAT (Sandy *et al.*, 2002; Pompeo *et al.*, 2002). Prokaryotic NATs show a remarkable amino acid similarity to the mammalian NATs, which also possess the same Cys^{68}-His^{107}-Asp^{122} catalytic triad (Sinclair *et al.*, 2000). Other features of the active site cleft, such as a loop determining substrate selectivity of the molecule, are also conserved between prokaryotic (residues 122-133) and eukaryotic (residues 122-131) NATs, based on homology modelling (Rodrigues-Lima *et al.*, 2001).

1.2.2 Cloning and chromosomal localisation of the human *NAT* genes

Although the human acetylation polymorphism has been an area of pharmacological research for almost fifty years, its molecular basis has only become evident during the past ten years, following cloning of the genes encoding for the NAT1 and NAT2 isoenzymes. The first sequence of a human *NAT* gene (Grant *et al.*, 1989a) was determined by sequencing of a clone, obtained from a genomic library with a rabbit *Nat* cDNA clone (Blum *et al.*, 1989a) as probe. Screening of additional genomic clones provided the sequences of two more *NAT* genes (Blum *et al.*, 1990a). All three *NAT* genes are intronless and share more than 80% homology in their 870bp coding sequence. One of the genes was further identified to contain several frameshift and nonsense mutations, suggesting that it is a pseudogene (*NATP*) (Blum *et al.*, 1990a). The predicted open reading frame of the three *NAT* genes was transiently expressed in mammalian cells, to allow characterisation of the corresponding protein products. One of the two functional genes (designated *NAT1*) was identified to encode for a NAT isoenzyme with high affinity to pAS and pABA, while the other (designated *NAT2*) encoded for a distinct NAT isoenzyme with high affinity to SMZ and procainamide. The *NATP* pseudogene did not produce any NAT activity or immunoreactive NAT protein, as expected (Blum *et al.*, 1990a; Grant *et al.*, 1991).

In parallel with the search for genomic sequences encoding for human NAT, other researchers achieved the isolation and heterologous expression of *NAT* cDNA clones from human liver (Ohsako and Deguchi, 1990). Of the three positive clones identified, one corresponded to *NAT1*, while the other two were distinct alleles of the *NAT2* gene. Though different by only one amino acid, the products of these alleles had different acetylation capacities. The molecular aetiology of the acetylation polymorphism was, thus, resolved for the first time. In addition, comparison of the genomic and liver cDNA *NAT* sequences provided a first indication of the genomic structure of the *NAT* genes. The sequence of *NAT1* cDNA completely matched the available genomic sequence, indicating that the entire transcribed region of the *NAT1* gene is contained in a single exon. In contrast, the 5′ untranslated region (5′ UTR) of the *NAT2* gene forms a separate exon, located about 8kb upstream of the second exon, which contains the coding region and the 3′ UTR (Ebisawa and Deguchi, 1991).

The *NAT1* and *NAT2* genes were initially localised to human chromosome 8, by Southern analysis of a human-rodent somatic cell hybrid DNA panel (Blum *et al.*, 1990a). Subsequent cytogenetic mapping, using a series of *NAT*-positive yeast artificial chromosome (YAC) clones as probes for fluorescent *in situ* hybridisation (FISH), placed the *NAT1* and *NAT2* genes on the short arm of chromosome 8, within region 8p21.3-23.1 (Hickman *et al.*, 1994). A more refined assignment to cytogenetic band 8p22 was achieved using a smaller *NAT2*-positive cosmid clone as a FISH probe (Franke *et al.*, 1994).

A YAC contig spanning the 8p22 region (Bookstein *et al.*, 1994) was subsequently used to carry out physical mapping of the human *NAT* locus (Matas *et al.*, 1997). *NAT1* and *NAT2* were mapped within 170-360kb from each other, while *NATP* was placed between them. The cluster co-localises with microsatellite marker *D8S261* and restriction fragment length polymorphism (RFLP) marker *D8S21*. The latter corresponds to a polymorphic *Rsa*I restriction site, located within the coding region of the *NAT2* gene. The order of the loci on chromosome 8p (telomere to centromere) is *D8S261-NAT1-NATP-NAT2(D8S21)*. A number of CpG islands, i.e. sequences of high GC content often associated with gene promoters (Kundu and Rao, 1999), were also mapped on the same YAC clones as the *NAT* locus, suggesting the presence of other genes nearby (Matas *et al.*, 1997). Detailed study of the region surrounding the *NAT* locus has now become possible (figure 1.3), following the announcement of the sequence draft of the human genome [The International Human Genome Sequencing Consortium (IHGSC), 2001; Venter *et al.*, 2001].

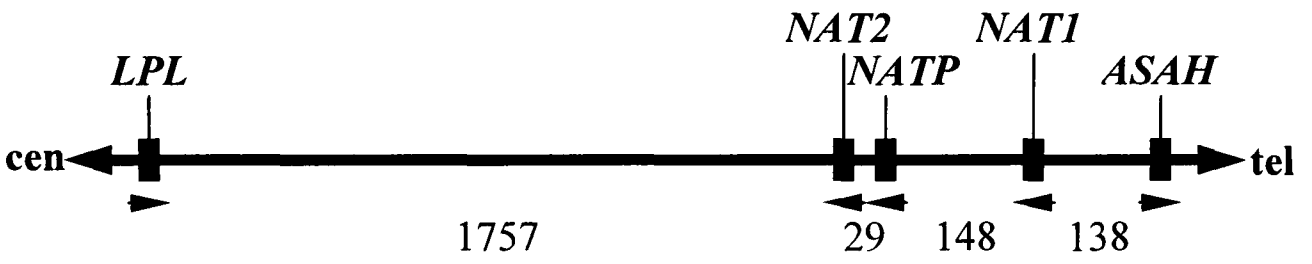


Figure 1.3: Arrangement of *NAT* and neighbouring loci on human chromosome 8p. Bacterial artificial chromosome (BAC) contig NT008271 contains all three *NAT* loci (overlapping BAC clones AB020868, AC025062 & AC025648). The closest known loci to *NAT* are the genes for lipoprotein lipase (*LPL*) (accession no. M15856), on the centromeric side, and N-acylsphingosine amidohydrolase (*ASAH*) (accession no. U70063), on the telomeric side. At least 13 hypothetical loci are also predicted between *LPL* and *NAT2* by sequence analysis. The exact cytogenetic position of the genes is at the border between bands 8p23.1 and 8p22. The arrowheads indicate the direction of the reading frame of each gene. The numbers represent kilobases of sequence between two genes. The results can be viewed using MapViewer, available from the National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/).

1.2.3 The molecular basis of the NAT polymorphism in humans

Cloning of the human *NAT* genes and the application of rapid genotyping methods has led to the identification of numerous polymorphisms in both *NAT1* and *NAT2*. Soon after the first few alleles were described, researchers in the field agreed on a consensus nomenclature (Vatsis *et al.*, 1995), to allow the systematic classification of novel alleles and facilitate the interpretation of results from different research groups. A NAT nomenclature committee was formed (Ilett *et al.*, 1999), with the authority to assign symbols to new alleles and maintain a regularly updated electronic database (www.louisville.edu/medschool/pharmacology/NAT.html) of all known *NAT* alleles in humans and other species (Hein *et al.*, 2000b).

Genotyping of the human *NAT* genes has been PCR-based. Methods involving PCR-RFLP or allele-specific PCR are most commonly used (e.g. Hickman and Sim, 1991; Bell *et al.*, 1993; Lin *et al.*, 1993; Cascorbi *et al.*, 1995; Doll *et al.*, 1995). Other PCR-based techniques have also been introduced, to allow quick genotyping of large

numbers of samples for high-throughput epidemiological research. These include single strand conformational polymorphism (SSCP) (Lin *et al.*, 1994; 1998; Lo-Guidice *et al.*, 2000), oligonucleotide ligation assay (Bigler *et al.*, 1997), allele-specific oligonucleotide hybridisation (Labuda *et al.*, 1999; Bunschoten *et al.*, 2000), heteroduplex analysis (Hubbard *et al.*, 1998; Lin *et al.*, 1998) and allele-specific amplification using a light-cycler (Wikman *et al.*, 2001).

1.2.3.1 *NAT1* polymorphism

Although early studies reported up to 6-fold interindividual variation in the acetylation rate of substrates such as pAS and pABA, a bimodal distribution of fast and slow phenotypes was not observed, leading to the view that the NAT1 isoenzyme is monomorphic (Weber and Vatsis, 1993, for historical review). Following cloning of the human *NAT* genes, however, it became evident that the *NAT1* locus has multiple alleles, similarly to *NAT2* (Vatsis and Weber, 1993). A total of 26 *NAT1* alleles have been identified to date, according to the NAT nomenclature web-site.

Of the first *NAT1* alleles to be described were the wild-type *NAT1*4* allele, responsible for the fast acetylator phenotype, and the mutant *NAT1*10* allele (Vatsis and Weber, 1993). The two alleles have identical coding regions, but *NAT1*10* contains two mutations in the 3′ UTR, one of which ($T^{1088}\rightarrow A$) abolishes the polyadenylation (polyA) signal of the gene. It was initially supported that *NAT1*10* is associated with increased acetylation activity (Bell *et al.*, 1995a), presumably via activation of an adjacent, stronger polyA signal, created by the mutation. However, subsequent studies failed to confirm these findings and suggested that the activity conferred by *NAT1*10* is the same as that of the wild type *NAT1*4* allele (Payton and Sim, 1998; Smelt *et al.*, 1998; Bruhn *et al.*, 1999; Yang *et al.*, 2000; Vaziri *et al.*, 2001). This view was further supported by functional analysis of recombinant NAT1 10 protein (De Leon *et al.*, 2000). In contrast, the insertion of 3 adenosines immediately after the polyA signal of the recently identified *NAT1*16* allele was demonstrated to compromise stability of the mRNA (De Leon *et al.*, 2000). Other alleles, such as *NAT1*18A* and **18B* (Yang *et al.*, 2000), *NAT1*26A* and **26B* (GeneBank accession nos. AF032677 and AF032678), and *NAT1*28* and **29* (Lo-Guidice *et al.*, 2000), also contain mutations, insertions or deletions in the region of the polyA signal, but their phenotypic impact has not been examined yet.

The high variation in acetylation activity observed between individuals of the same *NAT1* genotype (Ward *et al.*, 1992; Payton and Sim, 1998; Hughes *et al.*, 1998a; Smelt *et al.*, 1998) does not allow safe prediction of the acetylator phenotype from the genotype. A slight bimodality in phenotype distribution was observed by Butcher *et al.* (1998) and Bruhn *et al.* (1999), who reported that about 8% of the individuals they genotyped for *NAT1* were slow acetylators. These were carriers of mutant alleles, such as *NAT1*14A*, **14B* and **17* associated with decreased acetylation capacities (Payton and Sim, 1998; Butcher *et al.*, 1998; Hubbard *et al.*, 1998; Hughes *et al.*, 1998a; Lin *et al.*, 1998), or completely defective alleles, such as *NAT1*15* and *NAT1*19* that carry nonsense mutations in the *NAT1* coding region (Hubbard *et al.*, 1998; Hughes *et al.*, 1998a; Lin *et al.*, 1998). Interestingly, a NAT1-deficient, but otherwise normal, *NAT1*15/*15* homozygote was identified by Bruhn *et al.* (1999). The same investigators also suggested that the *NAT1*11* allele may confer an acetylator phenotype slower than normal, although other studies have not shown such an association (Doll *et al.*, 1997; Hughes *et al.*, 1998a; De Leon *et al.*, 2000). Allele *NAT1*22* also appears to be non-functional (Lin *et al.*, 1998), while other alleles have been associated with normal (*NAT1*3*, **20, *23* and **27*) (Blum *et al.*, 1990a; Lin *et al.*, 1998; Smelt *et al.*, 2000) or increased (*NAT1*21*, **24* and **25*) (Lin *et al.*, 1998) NAT1 activity.

Several studies have reported typical frequencies of 75% for *NAT1*4* and 20% for *NAT1*10* alleles in Caucasian populations. Alleles *NAT1*3*, **11* and **14* are less common, reaching frequencies of up 4% (Smelt *et al.*, 1998; Bruhn *et al.*, 1999; Labuda *et al.*, 1999; Lo-Guidice *et al.*, 2000; Smelt *et al.*, 2000). In Japanese, the frequencies of *NAT1*4* and *NAT1*10* are about 55% and 40%, respectively (Katoh *et al.*, 1998; Sekine *et al.*, 2001), while in Chinese, the *NAT1*4*, **10* and **3* alleles are present at roughly equal frequencies of 30-35% (Zhao *et al.*, 1998). Other *NAT1* alleles are extremely rare (<1% frequency) in all ethnic groups studied (Lin *et al.*, 1998).

1.2.3.2 *NAT2* polymorphism

NAT2 polymorphism has been studied extensively, due to its long known impact on the acetylator phenotype. To date, 29 alleles have been described for the human *NAT2* gene, according to the NAT nomenclature web-site. Of those, *NAT2*4*

is considered as the wild type allele, responsible for the fast acetylator phenotype. The *NAT2*11*, **12A*, **12B*, **12C* and **13* alleles are also considered as "fast", because they contain silent mutations and/or misense mutations which lead to conservative amino acid changes (e.g. the Lys^{268}→Arg substitution in the NAT2 12 protein) (Lin *et al.*, 1993; Hein *et al.*, 1994; Cascorbi *et al.*, 1996a). Functional analysis has not yet been carried out for the *NAT2*10*, **17*, **18* and **19* alleles, but they all contain misense mutations which cause non-conservative amino acid changes (Lin *et al.*, 1994; Doll *et al.*, 1995; Shishikura *et al.*, 2000). The remaining 19 *NAT2* alleles have been demonstrated to encode for enzymes with reduced activity and/or stability and, therefore, confer slow acetylator phenotypes (Hein *et al.*, 1994; Ferguson *et al.*, 1994b; Hickman *et al.*, 1995; Leff *et al.*, 1999b; Fretland *et al.*, 2001).

It is of interest that different combinations of only 13 single nucleotide polymorphisms (SNPs) at positions 111, 190, 191, 282, 341, 434, 481, 499, 590, 759, 803, 845 and 857 of the coding region form all known *NAT2* alleles. The SNP at position 282 is shared between nine *NAT2* alleles (*NAT2*6A*, **6C*, **6D*, **7B*, **12B*, **13*, **14B*, **14D* and **14G*), the SNP at position 341 between eight alleles (*NAT2*5A*, **5B*, **5C*, **5D*, **5E*, **5F*, **14C* and **14F*), the SNP at position 481 between six alleles (*NAT2*5A*, **5B*, **5F*, **11*, **12C* and **14C*), the SNP at position 590 between six alleles (*NAT2*5E*, **6A*, **6B*, **6C*, **6D* and **14D*), and the SNP at position 803 between eleven alleles (*NAT2*5B*, **5C*, **5F*, **6C*, **12A*, **12B*, **12C*, **14C*, **14E*, **14F* and **14G*). Some non-conservative SNPs have been used to designate specific allelic groups. For example, the SNP at position 191 (Arg^{64}→Gln) is characteristic of the *NAT2*14*, the SNP at position 341 (Ile^{114}→Thr) of the *NAT2*5*, and the SNP at position 590 (Arg^{197}→Gln) of the *NAT2*6* allelic group (data from the NAT nomenclature web-page).

The first identified and most commonly encountered *NAT2* alleles in Caucasian populations are the wild type *NAT2*4* and the mutant *NAT2*5* and *NAT2*6* alleles (Hickman and Sim, 1991; Blum *et al.*, 1991; Vatsis *et al.*, 1991; Abe *et al.*, 1993). In constrast, the "slow" *NAT2*5* allele is extremely rare in Oriental populations, a difference compensated for by an increased frequency of the "fast" *NAT2*4* allele (Deguchi *et al.*, 1990; Deguchi, 1992). This variance in the allelic frequency of *NAT2*5* accounts for the long known difference in the incidence of the

slow acetylator phenotype between Caucasians (52-68%) and the Japanese (10-15%) (Sim and Hickman, 1991). The frequencies of the most common alleles in different ethnic groups are shown in table 1.2. It has been observed that the *NAT1*10/NAT2*4* haplotype is up to 3.5-fold over-represented in Caucasian populations, implying a linkage disequilibrium between these particular *NAT* alleles (Smelt *et al.*, 1998; Henning *et al.*, 1999; Westphal *et al.*, 2000). However, the selective advantage conferred by this allelic combination remains unknown.

Table 1.2: Frequencies of most common *NAT2* alleles in different ethnic groups.

Ethnic group[1]	**Allelic frequency (%)**				**References**
	NAT2*4	***NAT2*5***[2]	***NAT2*6***[3]	***NAT2*7***[4]	
Caucasians					
British	26	45	27	1.6	Hickman & Sim, 1991
Scottish	20	49	27	3.7	Smith *et al.*, 1997
Swedish	19	51	28	2	
German	23	47	28	1.3	Cascorbi *et al.*, 1995
French	18	52	25	1	Deloménie *et al.*, 1996
Spanish[5]	24	42	25	1.3	Agundez *et al.*, 1994; 1996c
Portuguese	21	43	33	2.7	Lemos & Regateiro, 1998
USA White[5]	24	43	30	2.5	Bell *et al.*, 1993; Lin *et al.*, 1994; Gross *et al.*, 1999;
Arab	18	55	21	4	Woolhouse *et al.*, 1997
Indian	26	33	38	3	Lin *et al.*, 1994
Orientals					
Japanese[5]	61	1.8	25	10	Deguchi, *et al.*, 1990; Lin *et al.*, 1994; Sekine *et al.*, 2001
Chinese	61	7	27	5	Rothman *et al.*, 1993
Korean	68	1.8	18	11	Lin *et al.*, 1994
Hong Kong	47	5.7	31	16	
Taiwanese	51	2.5	31	15	
Filipino	40	6	36	18	
Africans					
US Black[5,6]	34	30	23	5	Bell *et al.*, 1993; Lin *et al.*, 1994
Gabonese[7]	6	41	22	2	Deloménie *et al.*, 1996
Dogons[8,9]	12	33	25	1.5	

*[1]Only studies genotyping healthy individuals are presented; [2]Combined frequencies of all NAT2*5 alleles. NAT2*5B is the most common; [3]Combined frequencies of all NAT2*6 alleles. NAT2*6A is the most common; [4]Combined frequencies of all NAT2*7 alleles. NAT2*7B is the most common; [5]Mean allelic frequencies from different studies, reporting similar results; [6]Also 8-9% of NAT2*14A allele; [7]Also 13% NAT2*12, 8% NAT2*13 and 8% NAT2*14; [8]Mean allelic frequencies of caste and non-caste Dogons; [9]Also 8-20% NAT2*12, 4-10% NAT2*13 and 5-7% NAT2*14.*

1.2.4 NAT in disease

It is well-recognised today that individual susceptibility to complex disorders, such as autoimmune, neurodegenerative and cardiovascular disease, diabetes, obesity, osteoporosis and cancer, is influenced by genetic and environmental factors (Sim *et al.*, 1995; Moore, 1999). The environmental influence is primarily exerted via exposure to pollutants, industrial chemicals or harmful components of the diet. Studying the genetic response of individuals to xenobiotics is, therefore, essential for understanding the genetic-environmental interactions leading to disease. Numerous epidemiological studies have investigated the links between specific polymorphisms of xenobiotic metabolising enzymes and predisposition to diseases with a suspected environmental component (Smith *et al.*, 1994; Nebert *et al.*, 1996; Hirvonen, 1999).

The NAT2 polymorphism has been studied in relation with various disorders, including Parkinson's (Agundez *et al.*, 1998a; Harhangi *et al.*, 1999) and Alzheimer's (Rocha *et al.*, 1999) disease, diabetes mellitus (Mrozikiewicz *et al.*, 1994; Agundez *et al.*, 1996a) and allergy (Gawronska-Szklarz *et al.*, 1999). However, the results of different studies are often controversial and do not suggest a role for NAT as a susceptibility factor. The progression of certain diseases, e.g. diabetes mellitus and AIDS, may alter the acetylator status of affected individuals, leading to a high incidence of discordance between the NAT2 phenotype and genotype (Agundez *et al.*, 1996a; O'Neil *et al.*, 1997).

The NAT polymorphism has been extensively studied in relation to cancer (Hein *et al.*, 2000a). Arylamines have been recognised as the compounds responsible for the high incidence of bladder cancer among workers in the dye industry, an observation first made more than a century ago (Weisburger, 1997, for historical review). The carcinogenicity of arylamines is exerted via conversion to N-hydroxyarylamines, a reaction catalysed by cytochrome P450. N-acetylation by NAT prevents N-hydroxylation and subsequent bioactivation of arylamines, protecting the cell from the genotoxic effect of these potent carcinogens (King *et al.*, 1997). Indeed, slow acetylators have been shown to carry more carcinogen-DNA adducts in their bladder cells, compared with fast acetylators. The influence of the acetylator phenotype is most evident when exposure is low (e.g. through tobacco smoke or urban pollution), but declines when exposure becomes very high (Vineis *et al.*, 1994).

Apart from detoxification, NAT can also mediate the bioactivation of arylamines via two pathways. The first involves N-hydroxylation by cytochrome P450, followed by NAT-catalysed O-acetylation. The resulting acetoxy ester is unstable and decomposes quickly to form an arylnitrenium ion. The latter is electrophilic and binds to DNA, causing mutations that may lead to cell transformation and cancer. The second pathway involves NAT-mediated N-acetylation of the arylamine, followed by N-hydroxylation of the resulting arylamide by P450, to form an arylhydroxamic acid. This is then converted into an unstable acetoxy ester, via NAT-catalysed N,O-acetyltransfer (Grant, 1993; King *et al.*, 1997). The rates of competing reactions in these metabolic pathways can determine the fate of arylamines in the body. Thus, polymorphisms affecting the amount and catalytic efficiency of NAT isoenzymes can influence susceptibility to cancer, as they may shift the balance towards carcinogen bioactivation, rather than detoxification.

1.2.4.1 NAT in bladder cancer

The involvement of the NAT2 polymorphism in arylamine-induced carcinogenesis of the bladder has been thoroughly studied. Early evidence suggested that the slow acetylator phenotype may be a risk factor for the disease (reviewed in Evans, 1992). The association was most evident among patients occupationally exposed to arylamines (Cartwright *et al.*, 1982). Recent studies, based on *NAT2* genotyping, have generally confirmed these initial findings. Risch *et al.* (1995) observed an over-representation of slow *NAT2* genotypes among bladder cancer patients exposed to arylamines through their occupation. In addition, they showed that 72% of the patients, who were also smokers, were slow acetylators. In the control group of healthy smokers, the proportion of slow acetylators was 39%. Other researchers have also implicated the slow *NAT2* genotype as a susceptibility factor for bladder cancer (Filiadis *et al.*, 1999; Inatomi *et al.*, 1999; Katoh *et al.*, 1999), especially among smokers (Brockmöller *et al.*, 1996; Okkels *et al.*, 1997; Inatomi *et al.*, 1999). However, there are also conflicting reports in literature (Schnakenberg *et al.*, 1998; Taylor *et al.*, 1998; Hsieh *et al.*, 1999). Hayes *et al.* (1993) reported lack of association between the acetylator phenotype/genotype and the risk for bladder cancer among Chinese workers in the benzidine industry. However, it is likely that at high

exposure levels to such a potent carcinogen the protection conferred by the fast acetylator phenotype becomes insignificant.

A number of meta-analyses have recently been published, re-evaluating the combined results of different studies (Green *et al.*, 2000; Johns and Houlston, 2000; Marcus *et al.*, 2000b). The data of up to 22 published studies that meet a series of eligibility criteria were analysed. It has been concluded that the slow NAT2 phenotype/genotype is a moderate risk factor for bladder cancer (combined odds ratio 1.37, 95% confidence interval 1.19-1.58). A meta-analysis of 16 studies that provide adequate information about the smoking history of patient and control subjects, further suggested a link between the slow acetylator status and increased risk for smoking-related bladder cancer (Marcus *et al.*, 2000a). The combined odds ratio for smokers is 1.5 (95% confidence interval 1.1-1.9), i.e. slightly higher than that reported for the general population.

The *NAT1* polymorphism does not appear to affect susceptibility to bladder cancer on its own (Okkels *et al.*, 1997; Katoh *et al.*, 1999; Hsieh *et al.*, 1999). However, certain studies have suggested that carriers of two "slow" *NAT2* alleles, who are also homozygous for *NAT1*10*, are more susceptible to the disease, especially if smokers (Taylor *et al.*, 1998; Katoh *et al.*, 1999; Hsieh *et al.*, 1999).

1.2.4.2 NAT in colon cancer

The NAT isoenzymes have long been suspected to play a role in chemical carcinogenesis of the colon, because of their ability to bioactivate pro-carcinogens present in cooked meat. At high temperatures, polycyclic aromatic hydrocarbons and heterocyclic amines are formed from creatinine, sugar and amino acid precursors. Subsequent hydroxylation by cytochrome P450 and O-acetylation by NAT leads to the production of highly mutagenic compounds that may trigger colon carcinogenesis (Nagao and Sugimura, 1993). Early studies have suggested a possible link between the fast NAT2 phenotype and increased predisposition to the disease (reviewed in Minchin *et al.*, 1993). However, studies conducted after the introduction of *NAT2* genotyping methods have generally failed to confirm these results (Rodriguez *et al.*, 1993; Shibuta *et al.*, 1994; Probst-Hensch *et al.*, 1995; Spurr *et al.*, 1995; Bell *et al.*, 1995b; Hubbard *et al.*, 1997; Lee *et al.*, 1998). It is likely that fast acetylators are

more prone to developing colon cancer, if they consume large amounts of well-done red meat on a regular basis (Lang *et al.*, 1994; Welfare *et al.*, 1997; Chen *et al.*, 1998; Kiss *et al.*, 2000). Individuals carrying "fast" alleles for both NAT2 and cytochrome P450 1A2 (CYP1A2) also appear to be at increased risk (Lang *et al.*, 1994).

The effect of *NAT1* polymorphism has also been investigated, but the results have been conflicting. Some researchers have reported lack of association between the *NAT1* genotype and susceptibility to colon cancer (Probst-Hensch *et al.*, 1996; Chen *et al.*, 1998), while others have suggested the *NAT1*10* allele as a putative risk factor (Bell *et al.*, 1995b). Two recent studies have also associated *NAT1*10* with higher predisposition to gastric adenocarcinoma (Boissy *et al.*, 2000; Katoh *et al.*, 2000).

1.2.4.3 NAT in other types of cancer

The involvement of NAT in chemical carcinogenesis has been most extensively studied in tissues sensitive to the effect of environmental carcinogens. In addition to the colon and bladder, the breast, lung, oral/pharyngeal cavity, liver and prostate are considered as target organs for carcinogens entering the body. Breast cancer is a multifactorial disease with red meat consumption and tobacco smoking as potential predisposing factors (Williams and Phillips, 2000). Susceptibility to breast carcinogenesis varies among individuals, suggesting the involvement of polymorphic enzymes of xenobiotic metabolism. Epidemiological evidence does not point to NAT2 as a strong risk factor for the disease (Agundez *et al.*, 1995; Hunter *et al.*, 1997; Ambrosone *et al.*, 1998; Gertig *et al.*, 1999; Delfino *et al.*, 2000a; 2000b). However, the fast acetylator phenotype has been implicated as a predisposing factor when other risk factors, such as smoking, excessive meat consumption or the menopause, are taken into account (Ambrosone *et al.*, 1996; Millikan *et al.*, 1998; Huang *et al.*, 1999a; Deitz *et al.*, 2000; Morabia *et al.*, 2000; Krajinovic *et al.*, 2001). A few studies have also implicated the *NAT1*10* and **11* alleles as putative risk factors (Millikan *et al.*, 1998; Zheng *et al.*, 1999; Krajinovic *et al.*, 2001), but the results are ambiguous (Millikan, 2000).

Recent studies have suggested that the "fast" *NAT2* genotype may be linked to higher predisposition to cancers of the oral/pharyngeal cavity (Henning *et al.*, 1999; Jourenkova-Mironova *et al.*, 1999) and the lung (Cascorbi *et al.*, 1996b; Nyberg *et al.*,

1998). In contrast, the "slow" *NAT2* genotypes may act as predisposing factors for hepatocellular carcinoma, especially in carriers of the hepatitis virus (Agundez *et al.*, 1996b). There appears to be no association between the *NAT2* genotype and prostate cancer (Agundez *et al.*, 1998b; Wadelius *et al.*, 1999). However, the above results require further confirmation, as conflicting evidence has been reported in literature (Gonzalez *et al.*, 1998; Fukutome *et al.*, 1999; Hou *et al.*, 2000; Wikman *et al.*, 2001).

1.3 Arylamine N-acetyltransferases in other organisms

1.3.1 NATs in prokaryotes

NATs are a very well conserved group of proteins, found in both eukaryotic and prokaryotic organisms. The enzymatic properties and structure of prokaryotic NATs have been extensively studied in *S. typhimurium* (Watanabe *et al.*, 1992; Sinclair *et al.*, 2000) and *M. smegmatis* (Payton *et al.*, 1999a; Sandy *et al.*, 2002). Moreover, analysis of 55 sequenced prokaryotic genomes (www.tigr.org/tdb/mdb/mdb.html) has provided *nat*-like gene sequences for many representatives of the *Proteobacteria* and *Firmicutes* phyla. It appears that, in certain mycobacteria (e.g. *M. tuberculosis, M. bovis* and *M. avium*), *nat* belongs to a cluster of genes encoding for enzymes involved in the metabolism of aromatic and biphenyl compounds (Payton *et al.*, 2001).

About 45 prokaryotic NAT enzymes have been biochemically characterised to date, including those from *M. tuberculosis*, *Bacillus subtilis* and many enterobacteria, such as *Escherichia coli* and species of *Klebsiella* and *Citrobacter* (Deloménie *et al.*, 2001; Payton *et al.*, 2001). The degree of conservation between prokaryotic and eukaryotic NAT enzymes is high, with *St*NAT showing 38.3% and *Ms*NAT 43.2% amino acid homology with human NAT1. Moreover, functional polymorphism has recently been detected in the *nat* genes of *B. subtilis*, *S. typhimurium* and *M. tuberculosis* strains (Payton *et al.*, 2001; Upton *et al.*, 2001). The ability of NATs from pathogenic bacteria to metabolise drugs, such as isoniazid and 5-aminosalicylic acid (5-AS), may affect the efficacy of therapies for common microbial diseases. A SNP recently found in *M. tuberculosis nat* could be pharmacologically important, as it may contribute to the isoniazid susceptibility of carrier strains (Upton *et al.*, 2001).

1.3.2 NATs in eukaryotes

NAT has been found in almost all eukaryotic taxa examined to date. NAT enzymatic activity has been detected in fungi (Fang *et al.*, 1997), many edible fruits and vegetables (Chung *et al.*, 1997), amphibians (Ho *et al.*, 1996), birds (Andres *et al.*, 1983b; Rougraff and Paxton, 1987) and domestic mammals (Güray and Güvenç, 1997a; 1997b). The use of molecular biology methods has provided the complete nucleotide sequence of a chicken *Nat* gene (Ohsako *et al.*, 1988), as well as partial sequences for a bovine (Upton and Sim, 1999) and a feline (Trepanier *et al.*, 1998) *Nat* gene. The latter is the only *Nat* gene found in the cat genome, and its product has substrate specificity similar to human NAT1 (Trepanier *et al.*, 1998). The genome of canids, however, does not contain any *Nat* genes (Trepanier *et al.*, 1997).

NATs have been extensively studied in laboratory mammals, commonly used for pharmacological research. Fast and slow acetylating rabbit strains have been widely used as models for the human acetylation polymorphism, since the 1960s (Frymoyer and Jacox, 1963). The properties of two NAT isoenzymes, partially purified from various rabbit tissues, were initially characterised by Hearse and Weber (1973). One isoenzyme, whose levels varied markedly between fast and slow acetylators, was found only in the liver and gut, and acetylated isoniazid, pABA and SMZ. The other isoenzyme, which appeared to be monomorphic, was found mostly in extrahepatic tissues and had high affinity for pABA. Sequencing of tryptic peptides from purified rabbit NAT protein (Andres *et al.*, 1987) allowed the production of oligonucleotide probes, used for screening of rabbit cDNA libraries (Blum *et al.*, 1989a; Sasaki *et al.*, 1991). The identified *Nat*-positive cDNA clones were subsequently employed to isolate the *Nat1* and *Nat2* genes of the rabbit (Blum *et al.*, 1990b; 1990c). It was further demonstrated by Southern blot analysis that deletion of the entire *Nat2* gene is responsible for the slow acetylator phenotype of certain rabbit strains (Blum *et al.*, 1989b; Sasaki *et al.*, 1991).

Genetically determined acetylator polymorphism has also been described for different laboratory strains of the Syrian hamster (Hein *et al.*, 1982) and rat (Hein *et al.*, 1991; Juberg *et al.*, 1991). In contrast to humans and rabbits, it is pABA and pAS, rather than isoniazid and SMZ, which are polymorphically acetylated in these species. Each species has two NAT isoenzymes (Ozawa *et al.* 1990; Hein *et al.*, 1991), the

genes for which have been cloned and sequenced (Abu-Zeid *et al.*, 1991; Ferguson *et al.*, 1994a; Land *et al.*, 1994; Nagata *et al.*, 1994; Doll and Hein, 1995; Ferguson *et al.*, 1996). Specific SNPs in the *Nat2* gene are responsible for the slow acetylator phenotype of some Syrian hamster and rat strains (table 1.3).

All genes for mammalian NATs, characterised to date, have 870bp intronless coding regions, encoding for 290 amino acid polypeptide chains of 29-33kDa molecular weight (Vatsis *et al.*, 1995). Mammalian NATs are remarkably conserved, both at the level of nucleotide and amino acid sequence (table 1.4). A phylogenetic analysis and amino acid sequence comparison of different NAT proteins is presented in figures 1.4 and 1.5, respectively.

Table 1.3: Known polymorphisms at the *Nat2* locus of rodent species.

Species	Name of allele[1]	SNPs	Amino acid changes	Acetylator phenotype
Syrian hamster[2]	*Nat2*15*	none	none	fast
	*Nat2*16A*	$T^{36}\rightarrow C$	none	no activity
		$A^{633}\rightarrow G$	none	
		$C^{727}\rightarrow T$	$Arg^{243}\rightarrow stop$	
	*Nat2*16B*	$T^{36}\rightarrow C$	none	no activity
		$C^{325}\rightarrow T$	none	
		$A^{633}\rightarrow G$	none	
		$C^{727}\rightarrow T$	$Arg^{243}\rightarrow stop$	
Rat[3]	*Nat2*20*	none	none	fast
	*Nat2*21A*	$G^{361}\rightarrow A$	$Val^{121}\rightarrow Ile$	slow
		$G^{399}\rightarrow A$	none	
		$G^{522}\rightarrow A$	none	
		$G^{796}\rightarrow A$	$Val^{266}\rightarrow Ile$	
	*Nat2*21B*	$G^{361}\rightarrow A$	$Val^{121}\rightarrow Ile$	slow
		$G^{399}\rightarrow A$	none	
		$C^{672}\rightarrow T$	none	
		$G^{796}\rightarrow A$	$Val^{266}\rightarrow Ile$	
Mouse[4]	*Nat2*8*	none	none	fast
	*Nat2*9*	$A^{296}\rightarrow T$	$Asn^{99}\rightarrow Ile$	slow

[1]Nomenclature of alleles according to Vatsis et al., 1995; [2]From Nagata et al. (1994) and Ferguson et al. (1996); [3]From Doll and Hein (1995); [4]From Martell et al. (1991).

		Rabbit		Mouse			Rat		Syrian hamster		Human			
		NAT1	**NAT2**	**NAT1**	**NAT2**	**NAT3**	**NAT1**	**NAT2**	**NAT1**	**NAT2**	**NAT1**	**NAT2**		
Rabbit	***Nat1***		95.2	65.2	72.4	62.8	68.6	72.0	69.7	73.4	74.8	72.8	**NAT1**	**Rabbit**
	Nat2	97.2		66.9	73.1	61.4	70.0	72.4	70.3	74.1	75.5	73.1	**NAT2**	
Mouse	***Nat1***	75.7	76.3		81.0	67.2	91.7	81.3	83.8	82.1	74.5	71.7	**NAT1**	**Mouse**
	Nat2	77.2	77.3	83.9		73.1	83.1	94.1	78.3	92.8	81.7	74.5	**NAT2**	
	Nat3	73.0	72.7	74.5	77.9		68.6	73.4	67.2	73.1	68.3	65.9	**NAT3**	
Rat	***Nat1***	76.2	76.3	93.0	83.8	74.8		83.1	85.8	84.8	81.0	74.8	**NAT1**	**Rat**
	Nat2	78.0	78.2	84.3	93.1	77.1	84.9		77.2	91.7	81.3	73.7	**NAT2**	
Syrian hamster	***Nat1***	75.9	76.1	86.5	81.7	74.3	87.6	81.5		79.0	71.4	68.6	**NAT1**	**Syrian hamster**
	Nat2	78.3	78.2	83.7	90.3	78.2	84.2	90.1	83.4		81.7	74.8	**NAT2**	
Human	***NAT1***	82.0	82.2	80.2	83.2	75.2	80.5	84.0	77.3	83.3		81.0	**NAT1**	**Human**
	NAT2	82.0	82.0	78.7	80.1	74.7	79.3	80.3	76.5	80.2	87.4		**NAT2**	
		Nat1	***Nat2***	***Nat1***	***Nat2***	***Nat3***	***Nat1***	***Nat2***	***Nat1***	***Nat2***	***NAT1***	***NAT2***		
		Rabbit		**Mouse**			**Rat**		**Syrian hamster**		**Human**			

Table 1.4: Comparison of known NAT isoenzymes and their encoding genes between mammalian species. The top half of the table shows the percentage identity of NAT amino acid sequences from rabbit, mouse, rat, Syrian hamster and human. The bottom half of the table shows the percentage identity of the nucleotide sequences of the corresponding genes. The gene sequences were obtained from the GeneBank database and used to deduce the corresponding amino acid sequences. Percentage identities were determined using the Genetics Computer Group (GCG) package, Version 10.2 for UNIX.

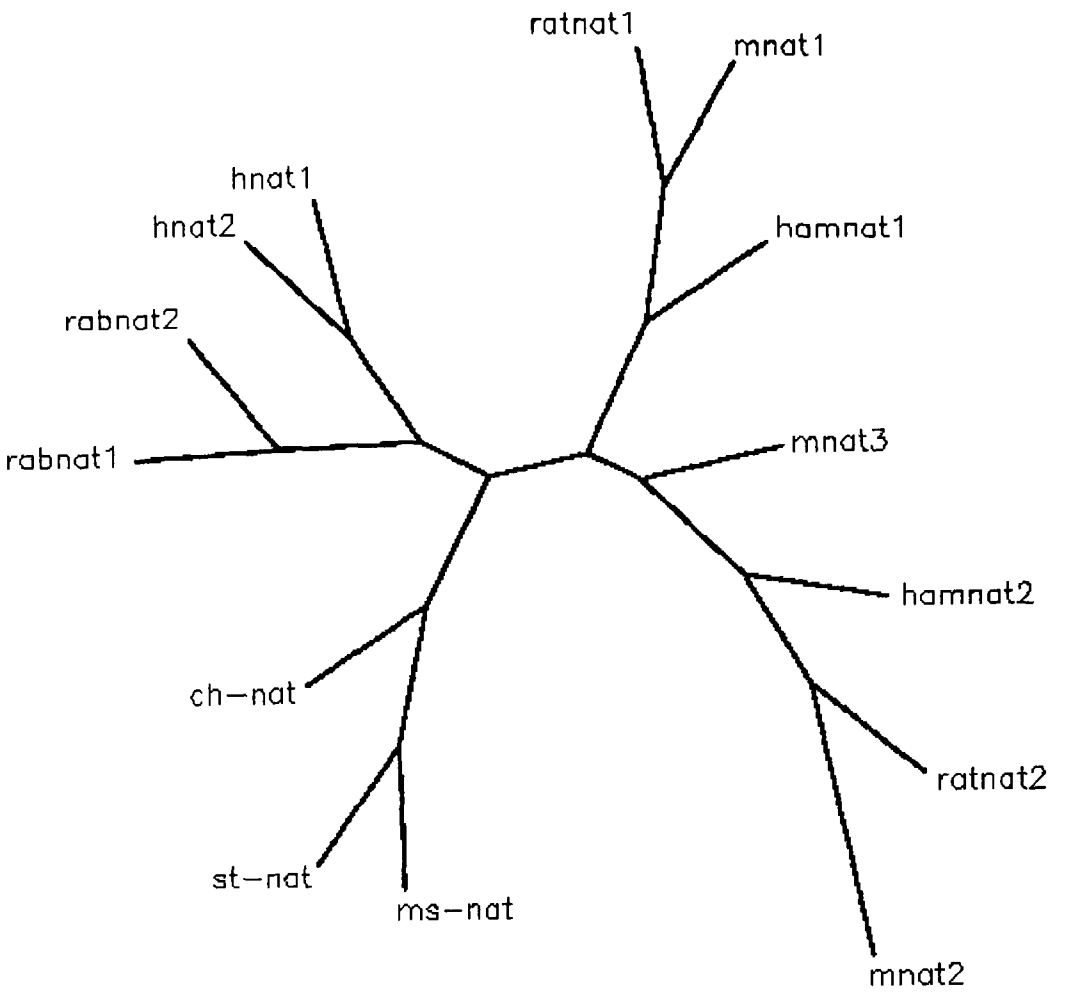


Figure 1.4: Phylogenetic tree of NAT proteins from different species. The amino acid sequences of NAT proteins were deduced from the corresponding gene sequences, deposited in the GeneBank database. The multiple alignment of proteins was performed using the PILEUP programme of the GCG package. The Phylip programme was used for phylogenetic analysis and the results were viewed through the Pie interphase, available from the Human Genome Mapping Project Resource Center (HGMP-RC) (http://menu.hgmp.mrc.ac.uk/menu-bin/PIE/pie.pl). The genetic distances were calculated using the protdist algorithm, executed with default options. The bootstrap option was applied to test the robustness of the tree. The codes represent rabbit NAT1 (rabnat1) and NAT2 (rabnat2), mouse NAT1 (mnat1), NAT2 (mnat2) and NAT3 (mnat3), rat NAT1 (ratnat1) and NAT2 (ratnat2), Syrian hamster NAT1 (hamnat1) and NAT2 (hamnat2), human NAT1 (hnat1) and NAT2 (hnat 2), chicken NAT (ch-nat), *M. smegmatis* NAT (ms-nat) and *S. typhimurium* NAT (st-nat).

Figure 1.5: Alignment of NAT amino acid sequences from different mammalian species. Amino acids are represented by the single-letter code. Numbers indicate the amino acid position. The codes representing NAT protein sequences are the same as in figure 1.4. Conserved amino acids are highlighted in yellow, while those partially conserved are highlighted in red.

```
                    .    .    .   10    .    .    .   20    .    .    .   30    .    .    .   40    .    .    .   50    .    .    .   60
rabnat1      1:MDIEAYYQRIGYKNPRNKLDLESLTDIFQHQIRTVPYENLSIHCGESMELDLEAIFDQIV: 60
rabnat2      1:MDIEAYYQRIGYKNPRNKLDLESLTDIFQHQIRTVPYENLSIHCGESMELDLEAIFDQIV: 60
mnat1        1:MDIEAYFERIGYKNSVNKLDLATLTEVLQHQMRAVPFENLNMHCGEAMHLDLQDIFDHIV: 60
ratnat1      1:MDIEAYFERIGYKNSVNKLDLATLTEVLQHQMRAVPFENLSMHCGEAMCLGLEATFDHIV: 60
hamnat1      1:MDIEAYFERIGYNNPVYTLDLATLTEVLQHQMRTIPFENLNMHCGEAMDLGLEATFDQIV: 60
mnat2        1:MDIEAYFERIGYQSTRSKLDLKTLTEILQHQIRAIPFENLNIHCGESMELSLEAIFDQIV: 60
ratnat2      1:MDIEAYFERIGYQSSRNKLDLEELTEILQHQIRAIPFENLNIHCGESMELNLEVIFDQVV: 60
hamnat2      1:MDIEAYFERIGYQNSRNKLDLQTLTEILQHQIRAIPFENLNIHCGESMELSLETIFDQIV: 60
hnat1        1:MDIEAYLERIGYKKSRNKLDLETLTDILQHQIRAVPFENLNIHCGDAMDLGLEAIFDQVV: 60
hnat2        1:MDIEAYFERIGYKNSRNKLDLETLTDILEHQIRAVPFENLNMHCGQAMELGLEAIFDHIV: 60
mnat3        1:MDIEAYFERIGYQKSSNKLDLQTLTEILQHQIRAIPFENLNIHCGKTMELSLEDTFHQIV: 60

                    .    .    .   70    .    .    .   80    .    .    .   90    .    .    .  100    .    .    .  110    .    .    .  120
rabnat1     61:RRNRGGWCLQVNYLLYWALTTTGFETTMLGGFVCGSHTDKYSTGMIHLIVQVTINGRNYI:120
rabnat2     61:RRNRGGWCLQVNYLLYWALTTTGFETTMLGGFVYGSNNDKYSTGMIHLIVQVTINGRNYI:120
mnat1       61:RKKRGGWCLQVNHLLYWALTKMGFETTMLGGYVYITPVSKYSSEMVHLLVQVTISDRKYI:120
ratnat1     61:RKKRGGWCLQVNHLLYWALTKMGFETTMLGGYVYITPVNKYSSEMVHLLVQVTISDRNYI:120
hamnat1     61:RKKRGGWCLQVNHLLYWALTQMGFETTMLGGYVYIVPVSKYSSEMIHLLVQVTISDRNYI:120
mnat2       61:RKKRGGWCLQVNHLLYWALTKLGFETTMLGGYVFNTPAIKYSSGMIHLLVQVTISGKDYI:120
ratnat2     61:RKKRGGWCLQVNHLLYWALTKMGFEATMLGGYVFNTPANKYSSGMIHLLVQVTLSGKDYI:120
hamnat2     61:RKKRGGWCLQVNHLLYWALTKMGFETTMLGGYVFNTPANKYSSGMIHLLVQVTISDRNYI:120
hnat1       61:RRNRGGWCLQVNHLLYWALTTIGFETTMLGGYVYSTPAKKYSTGMIHLLLQVTIDGRNYI:120
hnat2       61:RRNRGGWCLQVNQLLYWALTTIGFQTTMLGGYFYIPPVNKYSTGMVHLLLQVTIDGRNYI:120
mnat3       61:RKKRGGWCLQVNHLLYWALAMIGFETTMLGGCVYVPSACKYSNTMIHLLLQVTISGKTYI:120

                    .    .    .  130    .    .    .  140    .    .    .  150    .    .    .  160    .    .    .  170    .    .    .  180
rabnat1    121:VDAGFGRSYQMWQPVELISGKDQPQVPSIFRLREEGETWYLDQIRRQQHVPDQEFLNSEL:180
rabnat2    121:VDAGFGRSYQMWQPVELISGKDQPQVPSIFRLREEGETWYLDQIRRQQHVPDQEFLNSEL:180
mnat1      121:VDSAYGGSYQMWEPLELTSGKDQPQVPAIFLLTEENGTWYLDQIRREQYVPNEEFVNSDL:180
ratnat1    121:VDSAYGSSYQMWEPLELTSGKDQPQVPAIFRLTEENGTWYLDQIRREQDVPNQEFVNSDL:180
hamnat1    121:VDAAYGGSYQMWEPVELASGKDQPQVPAIFRLTEENETWYLDQIRREQHVPNQEFVNSDL:180
mnat2      121:VDAGFGRSYQMWEPLELTSGKDQPQVPAIFRLTEENGTWYLDQIRREQYVPNQEFINSDL:180
ratnat2    121:VDAGFGRSYQMWEPLELTSGKDQPQVPAIFRLTEENGTWYLDQIRREQYVPNQEFVNSDL:180
hamnat2    121:VDAGFGRSLQMWEPLELVSGKDHPQVPAIFRLTEENETWYLDQIRREQYVPNQAFVNSDL:180
hnat1      121:VDAGFGRSYQMWQPLELISGKDQPQVPCVFRLTEENGFWYLDQIRREQYIPNEEFLHSDL:180
hnat2      121:VDAGSGSSSQMWQPLELISGKDQPQVPCIFCLTEERGIWYLDQIRREQYITNKEFLNSHL:180
mnat3      121:VDSAFPFSCQLWEPLELTSGKDQPQVPAIFHLREENGTWYLEQTKRQEYVSNQEFIDSNF:180

                    .    .    .  190    .    .    .  200    .    .    .  210    .    .    .  220    .    .    .  230    .    .    .  240
rabnat1    181:LERKTHRKLYCFTLQPRTIEEFESANTYLQISPSSSPFLDKSICSLQTPEGVHCLVGLILT:240
rabnat2    181:LEKKIYQKLYCFTLQPRTIEEFESANTYLQESPSSSVFLDKSICSLQTPEGVHCLVGLTLT:240
mnat1      181:LEKNKYRKIYSFTLEPRVIEDFEYVNSYLQTSPASVFVSTSFCSLQTSEGVHCLVGSTFT:240
ratnat1    181:LEKSKYRKIYSFTLEPRTIEDFEYVNTYLQTSPASVFVSTSFCSLQTSEGVCCLIGSTLT:240
hamnat1    181:LEKNTYRKIYSFTLQPRTIEDFEYANTYLQISPVSVFVNTSFCSLQTSEGVCCLIGSTIA:240
mnat2      181:LEKNKYRKIYSFTLEPRTIEDFESMNTYLQTSPASVFTSKSFCSLQTPEGVHCLVGSTLT:240
ratnat2    181:LEKNKYRKIYSFTLEPRTIEDFESINTYLQTSPASLFTSKSFCSLQTLEGVHCLVGSTLT:240
hamnat2    181:LEKNKYRKIYSFTLEPRTIEDFESMNTYLQTSPASVFTSKSFCSLQTPEGVHCLVGCTLT:240
hnat1      181:LEDSKYRKIYSFTLKPRTIEDFESMNTYLQTSPSSVFTSKSFCSLQTPDGVHCLVGFTLT:240
hnat2      181:LPKKKHQKIYLFTLEPRTIEDFESMNTYLQTSPTSSFITTSFCSLQTPEGVYCLVGFILT:240
mnat3      181:LEKNTHRKIYSFTLEPRTIEDFWSISTYYQVSRTSVMTNTSLCSLHTKDGVHGLMGTILA:240

                    .    .    .  250    .    .    .  260    .    .    .  270    .    .    .  280    .    .    .  290
rabnat1    241:FRTYNYKENTDLVEFKVLTEEEVEEVLKTIFNISLGKKLVSKNGNLFFTI:290
rabnat2    241:SRTYNYKENTDLVEFKVLTEEEVEGVLKTIFNISLGKKLVSKNGHLSFTI:290
mnat1      241:SRRFSYKDDVDLVEFKYVNEEEIEDVLKTAFGISLERKFVPKHGELVFTI:290
ratnat1    241:SRRFSYKDNVDLVEFKSLTEEEIEDVLKTTFGISLEKKFVPKHGELVFTI:290
hamnat1    241:RRKFSYKENVDLVEFKNVSEEEIEDVLKTAFGVSLERKFVPKNGNLSFSI:290
mnat2      241:YRRFSYKDNVDLVEFKSLTEEEIEDVLRTIFGVSLERKLVPKHGDRFFTI:290
ratnat2    241:YRRFSYKDNIDLVEFKSLTEEEIEDVLKTIFGVSLERKLVPKHGDRFFTI:290
hamnat2    241:YRRFSYKDNVDLVEFKSLKEEEIEDVLKTIFGISLEKKLVPKHGDRFFTI:290
hnat1      241:HRRFNYKDNTDLIEFKTLSEEEIEKVLKNIFNISLQRKLVPKHGDRFFTI:290
hnat2      241:YRKFNYKDNTDLVEFKTLTEEEVEEVLKNIFKISLGRNLVPKPGDGSLTI:290
mnat3      241:YKKFNYKDNIDLVEFKTLKEEEIEEVLKSVFGIHLETKLVPKCGNVFFTI:290
```

1.3.3 Mouse NATs

The mouse has been used in genetic studies for almost a century. It has many of the advantages of less complex organisms, such as the yeast, fruit fly and nematode, in that it is relatively easy to maintain and breed, has a short generation time and produces sizeable litters. Most importantly, it is a mammal, hence, its physiology, development and genetics is very similar to the human. The mouse is an excellent model for studying complex disorders involving gene-gene and gene-environment interactions. For example, inbred strains can be used to measure the toxic, teratogenic or carcinogenic potency of foreign compounds, without the confounding effect of genetic heterogeneity. Furthermore, using congenic strains, the contribution of a spontaneous or induced mutation to a disease phenotype can be evaluated against a uniform genetic background and under controlled environmental conditions (Levy *et al.*, 1996; Avner, 1998; Moore, 1999; West *et al.*, 2000).

The mouse has been a popular model for studying the effect of NAT polymorphism on the hereditary response to xenobiotics (Levy *et al.*, 1992; 1996; Hein *et al.*, 1997). Variability in N-acetylation activity among mouse strains was first observed twenty years ago by Glowinski and Weber (1982b). Three of the twenty inbred strains they examined, namely the A/J, A/HeJ and X/Gf strains, were identified to be slow acetylators of common NAT substrates, such as benzidine and pABA. Up to 10-fold higher acetylation rates were measured in the cytosols of fast acetylating strains, including the commonly used C57Bl/6J, Balb/c, C3H/HeJ, CBA/J and DBA/J strains. Intercrosses between fast and slow acetylating strains produced litters with intermediate acetylation capacities, indicating inheritance of the acetylator phenotype in an autosomal co-dominant fashion, as in humans. Kinetic analysis of cytosolic NAT activity from fast acetylating C57Bl/6J and slow acetylating A/J mice showed the two variants to have similar apparent K_m, but 2-fold different apparent V_{max}, values for pABA (Mattano and Weber, 1987).

Early evidence supported the presence of at least two NAT isoenzymes, one monomorphic and one polymorphic, in the mouse liver (Hein *et al.*, 1988). However, the application of standard protein purification methods failed to separate the two NAT enzymatic activities, which persistently co-purified as a single protein of approximately 31kDa molecular weight (Glowinski and Weber, 1982a; Hein *et al.*,

1988; Mattano *et al.*, 1989; Watson *et al.*, 1990). Using a 95% pure protein preparation from C57Bl/6J mouse liver, Mattano *et al.* (1989) showed that, apart from arylamine N-acetylation, NAT also catalyses hydroxyarylamine O-acetylation and arylhydroxamic acid N,O-acetyltransfer. They also suggested involvement of an active site cysteine in the transacetylation from acetyl-CoA.

The enzymatic properties of the mouse NAT isoenzymes and the molecular basis of the acetylation polymorphism were better understood following cloning of the murine *Nat* genes (Martell *et al.*, 1991; Kelly and Sim, 1994). A mouse *Nat*-specific probe was generated by PCR with degenerate primers, designed to anneal to conserved regions of the rabbit and human genes for NAT. The probe was used to isolate *Nat*-positive clones from genomic libraries constructed with DNA of C57Bl/6J and A/J mice. Analysis of the clones revealed the presence of two *Nat* genes, named *Nat1* and *Nat2* (Martell *et al.*, 1991). A third gene, called *Nat3*, was later identified by Southern blot analysis and screening of a genomic library generated from the Balb/c strain (Kelly and Sim, 1994). All three *Nat* genes have 870bp intronless coding regions and encode for polypeptide chains of 290 amino acids, like all known genes for mammalian NAT. A comparison between the murine and other mammalian NAT isoenzymes and their genes is included in table 1.4 and figures 1.4 and 1.5.

Restriction mapping of plasmid clones carrying both murine *Nat1* and *Nat2*, demonstrated that the two genes are positioned within 9-9.5kb (Martell *et al.*, 1991; Fakis *et al.*, 2000). The *Nat3* gene has been found on the same 130kb P1 artificial chromosome (PAC) clone with *Nat1* and *Nat2*, indicating co-localisation of all three *Nat* genes in the same chromosomal region. The estimated distance between *Nat3* and the other *Nat* loci is 22-130kb, as judged from the size and preliminary restriction analysis of the PAC clone (Fakis *et al.*, 2000). Early genetic mapping demonstrated linkage between *Nat2* and the polymorphic *Es1* gene for esterase 1, previously localised at 43cM on mouse chromosome 8 (Mattano *et al.*, 1988). The estimated genetic distance between *Nat2* and *Es1* is 12cM, placing *Nat2* either at 31 or 55cM (i.e. on the centromeric or telomeric side of *Es1*, respectively) on chromosome 8. The weak linkage observed between *Nat2* and the polymorphic *Emv2* gene for endogenous ecotropic MuLV2 protein, previously mapped at 67cM on chromosome 8, led to assignment of the *Nat2* gene to the 31cM position (Mattano *et al.*, 1988). Localisation

of the *Nat* genes on mouse chromosome 8 was later confirmed by FISH analysis, using *Nat*-positive PAC clones as probes. The three genes lie on cytogenetic band B3.1-3.3, around the middle of the chromosome (Fakis *et al.*, 2000). This region is syntenic to human 8p22, where the homologous genes are located (section 1.2.2).

Sequencing of the *Nat1* and *Nat3* genes did not reveal any differences between fast (C57Bl/6J, Balb/c, C3H/HeJ) and slow (A/J, A/HeJ) acetylating strains (Martell *et al.*, 1991; Kelly and Sim, 1994; Fretland *et al.*, 1997). In contrast, the *Nat2* gene of slow strains was shown to contain an A→T nucleotide substitution at position 296 of the coding region, resulting in an Asn→Ile amino acid change at position 99 of the NAT2 protein (table 1.3) (Martell *et al.*, 1991; Fretland *et al.*, 1997). Recombinant NAT1 protein was further demonstrated to N-acetylate mouse "monomorphic" substrates, such as isoniazid, while recombinant NAT2 N-acetylates "polymorphic" substrates, such as pABA (Martell *et al.*, 1992). Other compounds, e.g. p-anisidine (pANIS) and 2-aminofluorene (2-AF), are good substrates for both NAT1 and NAT2, while SMZ is a poor substrate for either isoenzyme. In addition, both recombinant NAT1 and NAT2 catalyse O-acetylation and N,O-acetyltransfer reactions. The function of NAT3 remains obscure, since the acetyltransferase activity of recombinant protein, measured towards common NAT substrates, is extremely low (Martell *et al.*, 1992; Kelly and Sim, 1994; Fretland *et al.*, 1997; Estrada-Rodgers, 1998b).

The enzymatic properties of the NAT2 allozymes from fast and slow acetylating strains have been extensively studied. Martell *et al.* (1991) predicted an increase in the local hydrophobicity of slow NAT2, due to the Asn^{99}→Ile substitution. The mutation also affected stability of the molecule, providing a plausible explanation for the decreased NAT2 activity in slow acetylating mice (Martell *et al.*, 1992; De Leon *et al.*, 1995; Fretland *et al.*, 1997). De Leon *et al.* (1995) reported higher apparent K_m values for cytosolic NAT2 from slow strains, suggesting that the mutant enzyme is also catalytically compromised. However, other researchers measured similar K_m values for recombinant NAT2 from fast and slow strains (Martell *et al.*, 1992; Fretland *et al.*, 1997). Northern blot analysis showed that fast and slow acetylators produce similar amounts of *Nat2* mRNA in their liver cells (Martell *et al.*, 1992; De Leon *et al.*, 1995). However, the Asn^{99}→Ile substitution appears to affect the rate of *in vitro* NAT2 translation (Martell *et al.*, 1992).

NAT2 congenic strains have been generated, in order to assess the significance of the acetylation polymorphism, independent of other polymorphisms between the C57Bl/6J and A/J inbred strains (Mattano *et al.*, 1988). The congenic lines carry either the "fast" *Nat2*8* allele (table 1.3) on an A/J genetic background (*Nat2*8*-A/J), or the "slow" *Nat2*9* allele (table 1.3) on a C57Bl/6J genetic background (*Nat2*9*-C57Bl/6J). Together with their parental inbred strains, these congenic strains have been useful tools for studying how the NAT2 phenotype affects response to carcinogens. Slow acetylating *Nat2*9*-C57Bl/6J congenic mice accumulated more 2-AF-DNA adducts in their bladder cells than fast acetylating C57Bl/6J inbred mice (Levy and Weber, 1992). This is similar to the situation observed in humans, where slow acetylators are more susceptible to the effects of bladder carcinogens (section 1.2.4.1). The reverse pattern was observed in the liver and mononuclear leukocytes, where 2-AF-DNA adduct levels were higher in the fast *Nat2*8*-A/J congenic mice, compared with the slow A/J inbred mice (Levy and Weber, 1989; Levy *et al.*, 1994). These studies also demonstrated that more factors, such as the genetic background, sex and age, may influence the extent of carcinogen-induced DNA damage.

1.4 Recent developments in genome research

1.4.1 The genome projects

1.4.1.1 The Human Genome Project

With the advancement of methodologies allowing cloning, sequencing and mapping of genomic DNA, by the mid-1980s, the idea of sequencing the entire human genome began materialising. In 1990, the Department of Energy and the National Institute of Health in the United States launched the Human Genome Project (HGP), aiming to map and sequence the genome of humans and several model organisms (Watson, 1990; Cantor, 1990). Other countries soon joined forces; genomic research was funded by the Medical Research Council (MRC) and the Wellcome Trust in the UK, the Centre d' Etude du Polymorphisme Humain in France and the Science and Technology Agency in Japan. The final phase of the HGP, involving sequencing of the entire human genome, has been the product of a collaborative effort

of 20 groups from the United States, UK, Japan, France, Germany and China [The International Human Genome Sequencing Consortium (IHGSC), 2001].

The first eight years of the HGP were dedicated to the development of information, material and technology resources, sufficient to support the final sequencing effort (Guyer and Collins, 1995; Collins *et al.*, 1998). Most important has been the generation of detailed maps of the human genome. These include various genetic maps (Gyapay *et al.*, 1994; Dib *et al.*, 1996), the most recent one consisting of 5,264 simple sequence length polymorphic (SSLP) markers with an average spacing of 1.6cM. A number of radiation hybrid maps have also been constructed (Gyapay *et al.*, 1996; Olivier *et al.*, 2001), as well as YAC-based (Chumakov *et al.*, 1995) and sequence-tagged site (STS)-based (Hudson *et al.*, 1995) physical maps. The latter consists of 15,086 STSs with an average spacing of 199kb. Physical mapping has been facilitated by high-resolution FISH techniques, allowing rapid validation of clone ordering during the construction of contigs (Windle *et al.*, 1995; Palotie *et al.*, 1996). Gene-based maps are also available (Schuler *et al.*, 1996; Deloukas *et al.*, 1998), providing the position of about 30,000 genes in the human genome.

In addition to YACs, widely used in positional cloning projects, a number of recently developed cloning systems have been employed by the HGP, such as PACs (Ioannou *et al.*, 1994) and BACs (Shizuya *et al.*, 1992). These *E. coli*-based systems are more efficient than plasmids and cosmids, as they can incorporate larger inserts, up to 300kb in size. They also overcome many of the disadvantages of YACs, such as instability and chimerism (Monaco and Larin, 1994). BACs have been the basis of the "clone-by-clone" sequencing strategy adopted by the HGP. This involves ordering of BAC clones, isolated from large genomic libraries, to form contigs that cover the entire genome (The International Human Genome Mapping Consortium, 2001). Each of the mapped clones is then fragmented and subjected to shotgun sequencing. By establishing overlap between the fragments, the sequence of each BAC clone is assembled and linked to the sequence of the flanking clones. This process eventually leads to sequencing of entire chromosomes (McPherson, 1997; Olson, 2001). The alternative "whole-genome shotgun" sequencing method involves fragmentation and cloning of genomic DNA into plasmid or cosmid vectors, suitable for direct sequencing. Assembly of the sequence is based on the overlap between the sequenced

clones and limited mapping data, required to anchor these clones to the genome (Venter *et al.*, 2001). This method is quick, but more prone to assembly errors, due to the presence of repeat elements in the human genome.

Following a period of pilot projects, the full-scale HGP sequencing effort was launched in 1999, aiming to produce a "working draft" sequence by 2000 and the finished sequence no later than 2003 (Collins *et al.*, 1998). At around the same time, a biotechnology company in the United States, Celera Genomics, also announced plans for sequencing the human genome, using the "shotgun" sequencing approach described above (Marshall, 1999). The completion of two "working draft" sequences of the human genome was jointly announced by the directors of the two projects in June 2000 and the results were published in February 2001 (IHGSC, 2001; Venter *et al.*, 2001). The current progress of the public HGP is summarised in table 1.5. Efforts have now focused on the production of the "finished" sequence, which will provide almost complete coverage of the euchromatic portion of the genome, with an error rate of less than 1 per 10,000 bases (Bork and Copley, 2001). In addition, the large-scale production of expressed sequence tags (ESTs) (Adams *et al.*, 1991), combined with Serial Analysis of Gene Expression (SAGE) techniques (Velculescu *et al.*, 1995), is anticipated to provide extensive information about the composition of the human transcriptome (Strachan *et al.*, 1997; Caron *et al.*, 2001; Riggins and Strausberg, 2001).

Table 1.5: Current progress of the HGP and overview of the human genome.

Finished sequence[1]:	63%
Draft sequence[1]:	34.8%
Total sequence read[1]:	97.8%
Size of euchromatic genome[1]:	3.2Gb
Estimated no. of genes[2]:	30,000-40,000
Finished chromosomes[3]:	20, 21, 22
Nearly finished chromosomes[1]:	6, 7, 9, 10, 13, 14, 19, Y
Estimated no. of CpG islands[2]:	28,890
Percent of genome classified as SINEs[2, 4]:	13.14%
Percent of genome classified as LINEs[2, 4]:	20.42%
Percent of genome classified as LTRs[2, 4]:	8.29%
Percent of genome classified as SSRs[2, 4]:	3%
Percent of genome classified as repeats[2]:	47.83%

[1]From the NCBI web-site (www.ncbi.nlm.nih.gov/genome/seq/HsProgress.shtml), as on 31/12/2001; [2]From IHGSC, 2001; [3]From Deloukas et al. (2001), Hattori et al. (2000) and Dunham et al. (1999); [4]SINE: short interspersed element, LINE: long interspersed element, LTR:long terminal repeat, SSR: simple sequence repeat.

1.4.1.2 The Mouse Genome Project

Mapping and sequencing the genomes of model organisms has been a major objective of the HGP (Watson, 1990; Collins *et al.*, 1998). Genome comparisons can provide valuable information about the localisation, structure and evolution of genes and crucial regulatory elements (Rubin, 2001). Moreover, many of the analysed genomes belong to organisms of great medical or agronomic interest. To date, the genome of more than 55 bacteria, most of which are pathogenic (e.g. *Haemophilus influenzae*, *Neisseria meningitidis*), has been sequenced (www.tigr.org/tdb/mdb/mdb.html). The complete genome sequence is also available for several eukaryotic organisms, including the yeast *Saccharomyces cerevisiae* (Mewes *et al.*, 1997), the nematode *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998), the insect *Drosophila melanogaster* (Adams *et al.*, 2000; Myers *et al.*, 2000) and the plant *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative, 2000).

Sequencing of the mouse genome is important for the interpretation of the data generated for the human genome. Comparison of the two sequences will accelerate gene discovery in humans, since genomic regions conserved in the two species would be anticipated to contain functional elements (Collins *et al.*, 1998; West *et al.*, 2000). The Mouse Genome Project (MGP) has evolved in parallel with the HGP and has already provided the scientific community with valuable tools for genetic research. Detailed linkage maps of the mouse genome have been constructed, incorporating markers, such as gene polymorphisms, SSLPs, RFLPs, SSCPs, interspersed repeat sequence (IRS) polymorphisms and randomly amplified polymorphic DNAs (RAPDs) (Dietrich *et al.*, 1995; McCarthy *et al.*, 1995; Brady *et al.*, 1997; Rhodes *et al.*, 1998). The most recent genetic map (Rhodes *et al.*, 1998) has an average resolution of 0.61cM and consists of 3,368 SSLP markers. It has been the basis for the construction of a YAC-based physical map, anchored to the genetic map via 9,787 STSs, and providing 92% coverage of the mouse genome (Nusbaum *et al.*, 1999). Additional physical mapping data have been generated via cytogenetic (Matsuda and Chapman, 1995) and radiation hybrid (Van Etten *et al.*, 1999; Hudson *et al.*, 2001) mapping. The production, sequencing and annotation of more than 350,000 ESTs has provided valuable information about the composition of the murine transcriptome (Marra *et al.*, 1999), while recently developed high-throughput systems have further allowed the production and accurate sequencing of more than 21,000 full-length cDNAs from a

variety of mouse tissues (Shibata *et al.*, 2000; The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, 2001).

An initial plan for sequencing the mouse genome was put forward in 1999 (Battey *et al.*, 1999), followed by establishment of a public-private consortium, committed to fund the effort (Smaglik and Abbott, 2000). Celera Genomics launched a separate mouse sequencing project and announced the completion of a draft sequence in Spring 2001. However, these data are available only to subscribers (Marshall, 2001). At around the same time, the public laboratories released a set of raw data, generated by shotgun sequencing of unordered clones. The production of a good-quality "working draft", expected by 2003, will be based on a "clone-by-clone" sequencing strategy, using a set of ordered BAC clones, as for the human genome. The finished sequence is anticipated to become available no later than 2005 (Marshall, 2001). Table 1.6 summarises the recent progress towards sequencing of the mouse genome by the public laboratories and Celera.

Table 1.6: Progress of the public and private mouse genome sequencing projects.

	Public consortium[1]	**Celera Genomics**[2]
Mouse strain used as DNA source	C57Bl/6J	129X1/SvJ, DBA/2J, A/J
Raw sequence from whole genome shotgun	30 million reads of about 500bp each	15.9Gb
Genome coverage (shotgun)	more than 90%	99%
DNA redundancy (shotgun)	at least 3-fold	6-fold
Assembled genome (percent of size of the euchromatic genome[3]**)**	finished: 75.5Mb (2.5%), draft: 514.2Mb (17.1%)	2.6Gb (85%)

[1]From Marshall, 2001 and the NCBI web site (www.ncbi.nlm.nih.gov/genome/seq/MmProgress.shtml), as on 6/12/2001; [2]From Marshall, 2001; [3]According to the NCBI, the euchromatic portion of the murine genome has an estimated size of 3.1Gb.

1.4.2 Pharmacogenomics

The HGP is anticipated to benefit medical research more than any other scientific discovery to date. Knowledge of the human genome sequence will allow identification of novel disease genes, as well as genes whose products can be important drug targets. Moreover, sequencing of the DNA of many individuals will uncover a plethora of functional polymorphisms, affecting susceptibility to various

diseases or the adverse reactions of drugs (Lu, 1998; Evans and Relling, 1999; Peltonen and McKusick, 2001). By the time the draft sequence of the human genome was published, more than 1.42 million SNPs had already been mapped, providing an average density of one SNP every 1900 bases. At least 4% of these SNPs were predicted to fall within exons (The International SNP Map Working Group, 2001). Comparison of the draft sequences produced by the public HGP and Celera has revealed the location of more than 2.1 million putative SNPs (Venter *et al.*, 2001).

With the tremendous advancement of genomic research, *Pharmacogenetics* has evolved into the new discipline of *Pharmacogenomics*, which employs genomic resources to identify novel drug targets and study variation in the entirety of genes that determine individual response to drugs. This revolutionary approach is anticipated to change the way that drugs are developed and prescribed, allowing individualised treatments (Lu, 1998; Evans and Relling, 1999; Omenn, 2001). To achieve this aim, it is necessary to identify the bulk of SNPs that affect the metabolism and efficacy of drugs, as well as to develop tools for high-throughput screening of genes and management of the generated data (Pfost *et al.*, 2000; McLeod and Evans, 2001). The introduction of microarrays has opened the way towards large-scale genotyping of thousands of SNPs (Lander, 1999; McLeod and Evans, 2001). Fluorescence-labelled PCR products, spanning the regions of interest, are hybridised to oligonucleotide probes, immobilised onto a silicone-coated glass slide. The genotype is determined by analysis of the hybridisation pattern, as each probe is designed to contain a specific mutation. Microarrays have also been used to study gene expression patterns, by hybridising fluorescence-labelled total cDNA to arrays of gene-specific probes (Brown and Botstein, 1999; Duggan *et al.*, 1999; Southern *et al.*, 1999).

While the promise of Pharmacogenomics is enormous, scientists and medical workers will have to deal with the many ethical concerns of the public. Fears have been expressed that genetic testing will be used as a means to deny employment or health insurance to "high risk" individuals. The use of genetic testing is also questioned when medical intervention cannot be of benefit. The production of drugs targeted to a limited number of people may increase the cost of some therapies, while practitioners will have to cope with the extra responsibility of choosing the right drug

for the right person (Pfost *et al.*, 2000; Nebert and Bingham, 2001; Rothstein and Epps, 2001). The "ethical, legal and social implications" (ELSI) programme, launched together with the HGP (Watson, 1990), investigates the impact of genomic research on society, monitors the public opinion and provides training to clinical staff, social workers and educators (Guyer and Collins, 1995; Collins *et al.*, 1998).

1.4.3 Genomic databases

Large-scale genome projects generate vast amounts of data, which need to be quickly processed, annotated and stored, then readily distributed to the international scientific community. The discipline of bioinformatics, occupied with the application of computer and statistical techniques to the management of biological information, has evolved as an essential part of genomic research in the past ten years. Its progress has been accelerated by the enormous advancement of computer technologies and the development of the World Wide Web. Today, researchers can employ user-friendly web-based bioinformatics tools to easily access, analyse and compare the results of the genome projects on their personal computer (Eppig *et al.*, 1998; Gelbart, 1998).

Databases play a key role in communicating the results of biological research to scientists across the world. Nucleic acid sequences from individual laboratories and large-scale sequencing projects are deposited in the GenBank, EMBL (European Molecular Biology Laboratory) and DDBJ (DNA Database of Japan) databases, maintained by the National Centre for Biotechnology Information (NCBI) in the USA (www.ncbi.nlm.nih.gov), the European Bioinformatics Institute (EBI) in the UK (www.ebi.ac.uk) and the Center for Info-Biology in Japan (www.ddbj.nig.ac.jp), respectively. Data are regularly exchanged between these databases, to ensure world-wide coverage (Ouellette and Boguski, 1997; Eppig *et al.*, 1998; Benson *et al.*, 2000). Other databases, e.g. SWISS-PROT and PDB (Protein Database), are specialised for the storage of information on protein sequence and structure (Eppig *et al.*, 1998; Mallon and Strivens, 1998). Tools, such as NCBI's Entrez and EBI's SRS (Sequence Retrieval System), allow easy data retrieval, while BLAST (Basic Local Alignment Search Tool) is used for sequence-similarity searches (Mallon and Strivens, 1998).

For mouse geneticists, the most comprehensive source of information is the Mouse Genome Database (MGD) (www.informatics.jax.org), maintained by the

Jackson Laboratory in the USA. Cytogenetic, genetic, physical and radiation hybrid mapping data are displayed in a user-friendly manner and the details of probes, STSs, polymorphic markers, YAC clones and contigs can be readily accessed. The database provides information on mammalian homologies, descriptions of backcross and somatic cell hybrid experiments, as well as advice on nomenclature issues. It also incorporates data on mouse gene expression and has links to all major databases (Blake *et al.*, 1997).

Bioinformatics centers, such as the NCBI and EBI, have also developed a number of web-based resources, allowing retrieval and handling of specialised information. Examples are the dbSNP, dbEST and dbSTS divisions of GenBank and EMBL databases, the Human Genome MapViewer, allowing visualisation of HGP data, and the Online Mendelian Inheritance in Man (OMIM) database, providing information about human disease genes (Ouelette and Boguski, 1997; Wheeler *et al.*, 2001). The Human Genome Mapping Project Resource Centre (HGMP-RC) in the UK (www.hgmp.mrc.ac.uk) provides a variety of bioinformatics tools through the web, including programmes for sequence alignment, exon-prediction, phylogenetic analysis and protein structure modelling (Claverie, 1997; Mallon and Strivens, 1998). It also has links to many databases. A selection of databases is described in table 1.7. Web-based resources accessed via the NCBI web-site are listed in table 1.8. Other databases and bioinformatics tools are described in the experimental chapters.

Table 1.7: Specialised web-based genetic databases. Some of these databases are divisions of all three GenBank, EMBL and DDBJ databases, but only their NCBI address is provided for convenience (marked with an asterisk). *(From Ouellette and Boguski, 1997; Eppig et al., 1998; Mallon and Strivens, 1998; Lijnzaad et al, 1998).*

Database	Description	Electronic address
dbEST	Expressed Sequence Tag database	http://www.ncbi.nlm.nih.gov/dbEST*
dbSTS	Sequence Tagged Site database	http://www.ncbi.nlm.nih.gov/dbSTS*
GSS	Genome Sequencing Survey (GSS) database.	http://www.ncbi.nlm.nih.gov/dbGSS*
HTG	High-Throughput Genomic (HTG) database	http://www.ncbi.nlm.nih.gov/HTGS*
RHdb	Radiation Hybrid Database	http://www.ebi.ac.uk/RHdb
MKMD	Mouse Knock-out and Mutation Database	http://research.bmn.com/mkmd/
DHMHD	The Dysmorphic Human-Mouse Homology Database	http://www.hgmp.mrc.ac.uk/DHMHD/dysmorph.html

Table 1.8: Selected web-based data resources available from the NCBI. These are accessed via the NCBI web-site (www.ncbi.nlm.nih.gov) and allow search of specialised data sets from the GenBank. *(From Maglott et al., 2000; Smigielski et al., 2000; Wheeler et al., 2000; 2001).*

Resource	Description
The Taxonomy Browser	Searches the NCBI taxonomy database, which lists over 79,000 organisms represented in sequence databases.
LocusLink	Provides official gene names and other identifiers, plus descriptive information about genes and their products.
UniGene	Partitions GenBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters.
HomoloGene	Lists sets of homologous and orthologous UniGene clusters for human and various model organisms.
RefSeq	Database of reference sequence standards for mRNAs and proteins deposited in general databases.
dbSNP	Single Nucleotide Polymorphism database.
ORF Finder	Tool for location of open reading frames.
Electronic PCR	Tool for locating STSs within a nucleotide sequence query, by comparison with information deposited in the dbSTS database.
Human Genome MapViewer	Genome browser, showing an integrated view of the human genome physical and genetic maps.
Human Genome Sequencing	Tracks progress and provides access to human genome sequencing data.
GeneMap'99	Provides mapping information for over 30,000 genes.
The Human-Mouse Homology Maps	Provides information about genetic loci in homologous segments of DNA from mouse and human.
The Cancer Chromosome Aberration Project (cCAP)	A comprehensive list of recurrent neoplasia-associated chromosomal aberrations.
Entrez Genomes	Organises and provides access to genomic mapping and sequencing data for more than 900 species.
Clusters of Orthologous Groups (COGs)	Clusters of orthologous groups of proteins from completely sequenced bacteria, archea and eukaryotes.
The Cancer Genome Anatomy Project (CGAP)	Provides access to genetic data on normal, precancerous and malignant cells.
Gene Expression Omnibus (GEO)	A database of gene expression data obtained by various methodologies.
SAGEmap	Tool for the analysis of SAGE-generated data.
Online Mendelian Inheritance in Man (OMIM)	Comprehensive catalogue of human genes and disorders.
The Conserved Domain Database (CDD)	A library of conserved protein domains.
The Molecular Modelling Database (MMDB)	Structural database derived from the Protein Databank (PDB).

1.4.4 Mouse transgenesis and mutagenesis technologies

Although comparative sequencing and EST analysis are anticipated to reveal the location and genomic organisation of all human genes in the foreseeable future, it will take many years for scientists to unravel the function of every single gene in the human genome. *In vivo* studies of newly identified genes will be essential, in order to elucidate their functional significance and potential involvement in genetic disease, as well as to assess their value in pharmacological and toxicological investigations. Genetic manipulation is a very efficient way of dissecting gene function, but can be performed only on suitable animal models. The mouse is almost exclusively used for this type of experiment, due to its genetic similarity to humans and the extensive understanding of its genetics (Rosenberg, 1997; You *et al.*, 1997; Sauer, 1998; Williams and Wagner, 2000).

There are two basic strategies for engineering the mouse genome. The first involves microinjection of the gene of interest into a fertilised oocyte, followed by random integration into the genome. The transgene is usually cloned adjacent to a strong promoter or elements driving inducible or tissue-specific expression. This approach is very useful for studying the consequences of excessive or ectopic gene expression, as well as the effects of heterologous expression of human genes in the mouse. On the other hand, transgene expression levels may vary between transgenic lines, due to differences in the number of integrating copies or the presence of endogenous control elements at the sites of incorporation. Random integration of the transgene may also interfere with the normal function of endogenous genes, leading to false interpretation of the observed phenotype (Moreadith and Radford, 1997; Hardouin and Nagy, 2000; Williams and Wagner, 2000).

The second strategy allows targeted replacement of an endogenous gene by a heterologous or mutated copy, incorporated to the mouse genome via homologous recombination. The technique is based on the ability of mouse embryonic stem (ES) cells to promote recombination between native and exogenous DNA sequences, provided that they share sufficient homology. In the majority of cases, the aim is to specifically inactivate a gene of interest, producing a knock-out mouse. *Replacement* or *insertion* targeting constructs are used for this purpose. These contain the gene of interest, flanked by two homology arms of sufficient length to allow homologous

recombination. Gene disruption is achieved by insertion of a cassette, carrying a positive selection marker (e.g. a gene providing resistance to neomycin, hygromycin or puromycin). In replacement constructs, a negative selection marker [e.g. the *thymidine kinase* (*tk*) gene, conferring sensitivity to ganciclovir] is also cloned outside the homology arms. A replacement construct is linearised outside the homology arms and is used to completely replace the endogenous locus. In contrast, the insertion construct is linearised within the region of homology and its targeted integration results in disruption of the exon-intron structure of the endogenous gene (Capecchi, 1989; Hasty and Bradley, 1992; Galli-Taliadoros *et al.*, 1995; Moreadith and Radford, 1997; Müller, 1999). Following selection of targeted ES cells, the production of knock-out mice is carried out as outlined in figure 1.6.

New technologies allow gene inactivation in a tissue- or developmental-specific manner (conditional or inducible gene targeting). These exploit the properties of *Cre*, a P1 bacteriophage site-specific DNA recombinase, catalysing reciprocal conservative DNA recombination between two 34bp recognition sites, called *loxP* (Gu *et al.*, 1994; Kühn *et al.*, 1995; Sauer, 1998). The *Cre/loxP* system has also been used to delete extended regions (up to 3-4cM) of the mouse genome (Justice *et al.*, 1997; Sauer, 1998; Hardouin and Nagy, 2000). Mice carrying deletions and random point mutations, e.g. induced by a potent chemical mutagen such as N-ethyl-N-nitrosourea (ENU), are being used to study the effect of recessive mutations in previously unknown genes (Meisler, 1996; Justice *et al.*, 1997). Gene trapping methods are also popular, as they allow discovery and simultaneous disruption of novel genes in the mouse (Jackson, 2001). Large-scale mutagenesis projects have already generated about 4000 engineered, 1500 ENU-induced and 1500 gene-trap mutants (The International Mouse Mutagenesis Consortium, 2001).

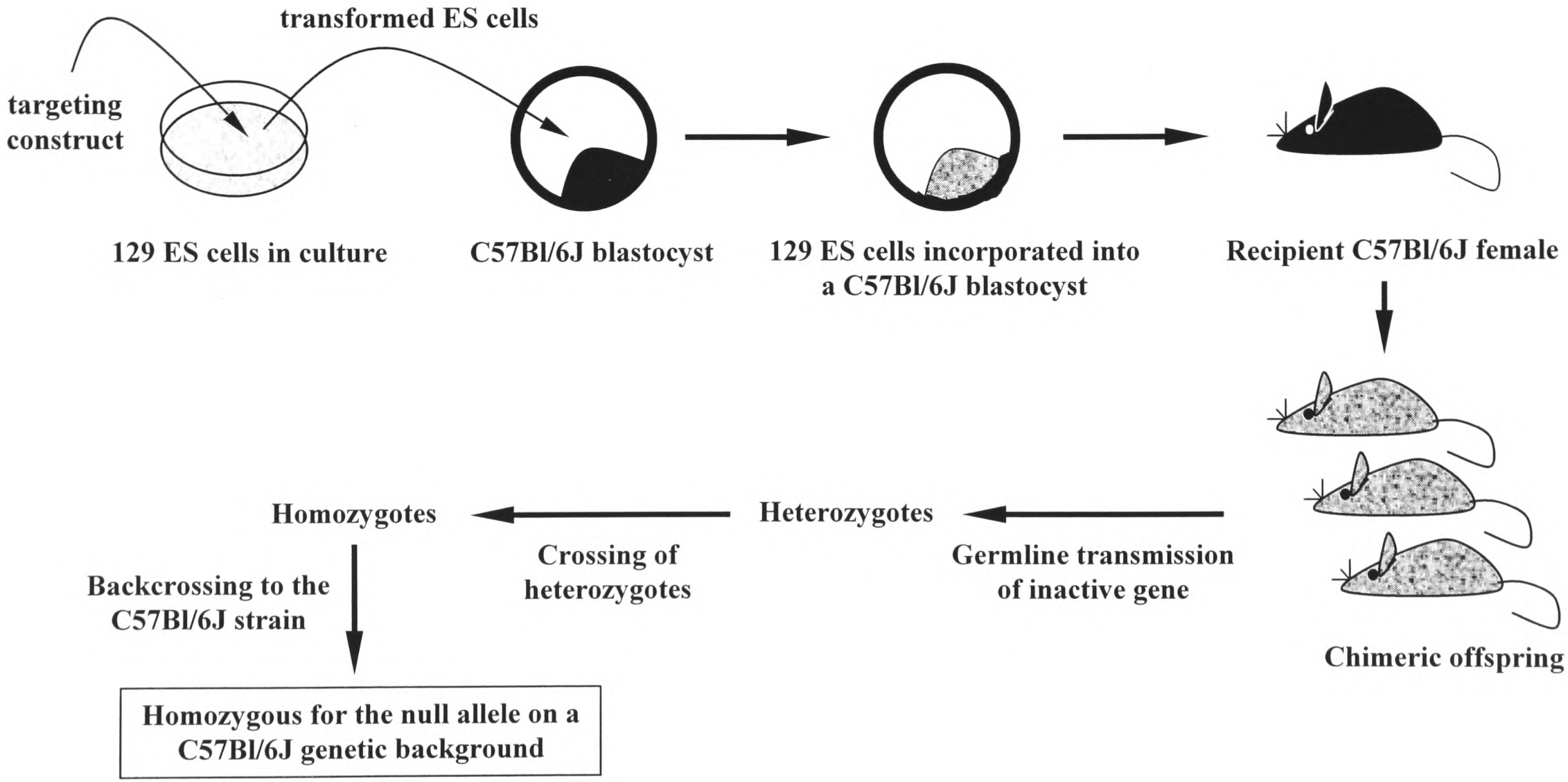


Figure 1.6: Overview of the strategy used for generating knock-out mice. The targeting construct is electroporated into ES cells, which are then subjected to positive and negative selection of targeted clones. The most commonly used ES cell lines originate from the 129 mouse strain. In order to maintain their pluripotency, ES cells are cultured on layers of irradiated fibroblast feeder cells or in the presence of leukaemia inhibitory factor (LIF). Blastocysts derived from a different mouse strain (usually C57Bl/6J) are injected with ES cells surviving selection and transferred to the uterus of pseudopregnant females. The chimeric offspring are then bred to produce mice carrying two copies of the null allele on a homogeneous (C57Bl/6J) genetic background.

1.5 Eukaryotic gene expression

1.5.1 The role of transcription factors as regulators of mRNA synthesis

Eukaryotic genes are transcribed by three RNA polymerases. Genes encoding ribosomal RNAs (apart from 5S rRNA) are transcribed by RNA polymerase I. Protein-encoding and small nuclear RNA (snRNA)-encoding genes are transcribed by RNA polymerase II, while RNA polymerase III is responsible for the synthesis of transfer RNA and 5S rRNA. Transcription initiation is a finely regulated step of eukaryotic gene expression. RNA polymerases do not have the inherent ability to recognise promoters and depend on interaction with accessory proteins, called general transcription factors (GTFs), for accurate initiation of transcription. Other *trans*-acting factors regulate tissue- or time-specific expression of genes, in response to extra- or intra-cellular stimuli. These factors exert their regulatory role via binding to specific DNA elements, such as activators, enhancers and repressors (Dynan and Tjian, 1985; La Thangue and Rigby, 1988; Weis and Reinberg, 1992; Pugh, 1996; Roeder, 1996).

The core promoter is the site of recruitment of the RNA polymerase II holoenzyme and comprises the minimum number of DNA elements required for accurate transcription initiation (Hernandez, 1993; Roeder, 1996; Verrijzer and Tjian, 1996). The most common feature of core promoters is the TATA-box, a short TA-rich sequence located about 30bp upstream of the transcription start site of most protein-encoding eukaryotic genes (Dynan and Tjian, 1985; Beebee and Burke, 1990; Latchman, 1998a; 1998b). Another element, called the initiator (Inr), is a pyrimidine (Y)-rich sequence ($_{-2}YYA_{+1}N^{T}/_{A}YY_{+5}$), spanning the transcription start site of several genes. The first Inr element was found in the TATA-less promoter of the gene for the terminal deoxynucleotidyl transferase. Since then, several Inr types have been described, working either in conjunction with or in the absence of a TATA-box (Weis and Reinberg, 1992; Roeder, 1996; Zhou and Chiang, 2001). The downstream promoter element (DPE) has a $^{A}/_{G}G^{A}/_{T}CGTG$ consensus sequence and is located about 30bp downstream of the transcription start site of several *Drosophila* and human TATA-less promoters (Roeder, 1996; Zhou and Chiang, 2001). Recent findings suggest that the composition of the core promoter may be important for transcriptional regulation, as specific protein activators may interact more efficiently with an Inr or a DPE or a TATA-box (Colgan and Manley, 1995; Butler and

Kadonaga, 2001). Other core promoter elements include the TFIIB-recognition element (BRE) (consensus sequence $^{G}/_{C}{}^{G}/_{C}{}^{G}/_{A}$CGCC), located just upstream of the TATA-box (Lagrange *et al.*, 1998), and the downstream core element (DCE), found 10-40bp downstream of the transcription start site of the β-globin gene (Lewis *et al.*, 2000). The DCE is distinct from the DPE, as it has a 30-nucleotide-long consensus sequence and functions with or without a TATA-box.

Additional activating elements have been found close to the core promoter of many eukaryotic genes. Most commonly encountered are the CCAAT-box and the Sp1-box, which increase activity of the core promoter above basal levels. The CCAAT box is the binding site of more than one transcription factor, including CTF (CCAAT-box transcription factor) and C/EBP (CCAAT enhancer binding protein). CTF is ubiquitously expressed and can activate DNA replication, as well as transcription. The C/EBPα factor is abundant in the liver and has been implicated in cellular differentiation processes, as it is mainly found in terminally differentiated, non-proliferating cells. Other C/EBP isoforms (e.g. C/EBPβ and δ) are activated upon stimulation by cytokines. Different sequence variants of the CCAAT-box may favour binding of either the CTF or C/EBP. Therefore, the CCAAT-box can act both as an activator and as a regulator of core promoters (La Thangue and Rigby, 1988; Johnson and McKnight, 1989; Akira *et al.*, 1990; Nerlov and Ziff, 1994; Latchman, 1998a).

The Sp1-box is a GC-rich sequence (consensus GGGCGG), initially identified as an essential component of the SV40 early promoter (Dynan and Tjian, 1985). Sp1 elements were subsequently found in numerous eukaryotic promoters, many of which are TATA-less and drive expression of housekeeping genes (e.g. the human and mouse genes for hypoxanthine phosphoribosyl transferase and dihydrofolate reductase and the human genes for adenosine deaminase and phosphoglycerate kinase) (Dynan, 1986). The Sp1-box is recognised by a family of transcription factors, including Sp1, Sp2, Sp3 and Sp4 and the more distantly related BTEB and BTEB2 factors (Lania *et al.*, 1997). The Sp1 factor, originally isolated from HeLa cells (Briggs *et al.*, 1986), is present in all cell types and is essential for the constitutive or inducible expression of many genes (Lania *et al.*, 1997). Sp1-deficient mice do not survive past day 9.5 of gestation and exhibit various abnormalities (Marin *et al.*, 1997). Other related factors are tissue-specific or may have a more complex role. For example, Sp4 expression is

restricted to the brain, while the widely expressed Sp3 factor can act as either an activator or as a repressor of transcription (Lania *et al.*, 1997).

Transcription initiation depends on the accessibility of the transcriptional apparatus, called the preinitiation complex (PIC), to the core promoter. Nucleosomal packaging and/or CpG methylation can prevent transcription factors from binding to the DNA. Acetylation of the lysine-rich N-terminal "tail" of histones significantly reduces their affinity for DNA, resulting in a more relaxed chromatin structure that may facilitate transcription factor binding. It has recently been proposed that histone acetylation and other modification patterns (e.g. phosphorylation, methylation, ubiquitination and ADP-ribosylation) may form distinct codes recognised by regulatory factors that bring about specific downstream functions (Strahl and Allis, 2000). Several transcriptional co-activators (i.e. proteins that interact with transcription factors) have been shown to possess histone acetyltransferase (HAT) activity. These include the human p300/CBP (CREB-binding protein) co-activator, which interacts with the CREB (cAMP response element binding) factor, as well as the human P/CAF (p300/CBP-associated factor) and $TAF_{II}250$ (part of TFIID) co-factors (Greenblatt, 1997; Wade *et al.*, 1997; Lee and Young, 2000). Transcriptional activators may also recruit other chromatin-remodelling factors (e.g. topoisomerases, ATPases, actin-related proteins, etc.) close to gene promoters (Lee and Young, 2000).

Once the promoter elements become accessible, formation of the PIC begins. A mechanism involving stepwise assembly of the GTFs was originally proposed, starting with binding of TFIID to the core promoter. The TFIID complex consists of the TATA-binding protein (TBP) and various TBP-associated factors (TAFs). TBP is considered as a universal transcription factor, because it is extremely well-conserved and is essential for transcription initiation by all three eukaryotic RNA polymerases. X-ray crystallography has demonstrated that TBP is a saddle-shaped molecule, the concave lower surface of which binds to the minor groove of DNA, at the site of the TATA-box. This interaction forces the DNA helix to bend and unwind locally. TBP is also required for transcription initiation from TATA-less promoters, but interaction with the DNA helix is mediated by a range of TAF-Inr co-dependent co-factors (TICs). TAFs bind to the upper surface of TBP and interact with transcriptional activators (e.g. Sp1, VP16, p53, CTF and nuclear receptors). Several TAFs have a

structure similar to that of histones, suggesting a nucleosome-like organisation of the TFIID complex (reviewed in Hernandez, 1993; Struhl, 1994; Burley, 1996; Burley and Roeder, 1996; Chao and Young, 1996; Surridge, 1996; Verrijzer and Tjian, 1996; Buratowski, 1997; Lee and Young, 2000; Lemon and Tjian, 2000).

Binding of TFIID to the core promoter is followed by recruitment of the TFIIA and TFIIB factors. The latter is important for recruitment of the pre-formed RNA polymerase II-TFIIF complex (Orphanides *et al.*, 1996; Roeder, 1996; Lee and Young, 2000; Lemon and Tjian, 2000). RNA polymerase II complexes comprise 10-12 conserved subunits (Rpb1-Rpb12), with distinct structural and catalytic roles. Analysis of the three-dimensional structure of yeast RNA polymerase II has revealed an active site cleft, accommodating a 9bp hybrid formed between the DNA template and the elongating mRNA molecule. The transcribing complex is stabilised by a clamp close to the active site, formed by the Rpb1, Rpb2 and Rpb6 subunits. Another clamp, formed by subunits Rpb1, Rpb5 and Rpb9, surrounds the DNA helix downstream of the active site cleft (Cramer *et al.*, 2000; 2001; Gnatt *et al.*, 2001). Of crucial importance is the C-terminal domain (CTD) of RNA polymerase II, which consists of tandem repeats (52 in mammals) of the heptapeptide Tyr-Ser-Pro-Thr-Ser-Pro-Ser. The CTD is hypophosphorylated in the PIC, but extensive phosphorylation is required for transition from the initiation to the elongation state of RNA polymerase II (Roeder, 1996; Orphanides *et al.*, 1996; Greenblatt, 1997; Lee and Young, 2000).

The formation of the PIC is completed by recruitment of the TFIIE and TFIIH factors, that stimulate DNA melting in the promoter region, to form a single-stranded bubble (Orphanides *et al.*, 1996; Roeder, 1996; Svejstrup *et al.*, 1996; Lemon and Tjian, 2000). The most important properties of GTFs are described in table 1.9. Co-purification of RNA polymerase II and all, but TFIID, GTFs suggests the presence of a pre-assembled holoenzyme in the nucleus, although PIC formation *in vitro* can take place in the stepwise manner, described above (Lemon and Tjian, 2000). Transcription initiation requires interaction of transcriptional activators with the components of the PIC via multisubunit complexes called "mediators". These include the nine-subunit mammalian CRSP co-factor required for activation of Sp1, the hMediator complex that binds to the activation domain of VP16 and the thyroid hormone receptor-associated protein complex TRAP (Berk, 1999; Buratowski, 2000).

Table 1.9: Properties of GTFs involved in the assembly of RNA polymerase II PIC. *(Adapted from Orphanides et al, 1996; Roeder, 1996; Lee and Young, 2000).*

GTF	Properties
TFIIA	Stabilises TBP-DNA interaction; Antirepression; Interaction with activators
TFIIB	Required for recruitment of RNA polymerase II-TFIIF complex; Selection of transcription start site; Promoter bending during transcription initiation
TFIID	Binds to TATA-box (TBP)*; HAT and serine kinase activity ($TAF_{II}250$); Interaction with transcription activators ($TAF_{II}55$ and $TAF_{II}135$); Similarity to histone H4 ($TAF_{II}18$ and $TAF_{II}80$), H3 ($TAF_{II}28$ and $TAF_{II}31$), H2A ($TAF_{II}15$ and $TAF_{II}20$)
TFIIE	Recruits and modulates activity of TFIIH; Promoter melting
TFIIF	Binds to RNA polymerase II; Prevents non-specific binding of the polymerase (homologous to the prokaryotic σ-factor)
TFIIH	Nucleotide excision repair (p34, p44, p52 and p62)*; Cyclin-dependent CTD-kinase (MAT1, cdk7 and cyclin H); 5′→3′ helicase and ATPase, required for excision repair (ERCC2); 3′→5′ helicase and ATPase, required for promoter opening (ERCC3)

** In brackets, the name of subunit or co-activator responsible for a specific function.*

Transcription factors typically consist of a DNA binding domain and at least one transcription activation domain (Mitchell and Tjian, 1989). Commonly encountered DNA binding motifs include the Cys_4- and Cys_2/His_2-zinc fingers of steroid hormone receptors and Sp1, respectively (Schwabe and Rhodes, 1991; Klug and Schwabe, 1995), the leucine zipper (bZIP) of AP1, CREB and C/EBP factors (Struhl, 1989; Lamb and McKnight, 1991), the helix-loop-helix (bHLH) of E47 and the myogenic factors MyoD and myogenin (Lamb and McKnight, 1991; Littlewood and Evan, 1995), the homeodomain of the HOX factors (Kornberg, 1993; Gehring *et al.*, 1994), the POU-domain of the Pit, Oct and Unc factors (Verrijzer and Van der Vliet, 1993; Ryan and Rosenfeld, 1997) and the Rel-domain of NF-κB (Thanos and Maniatis, 1995; Baeuerle and Baltimore, 1996). The activation domain of transcription factors can be rich in acidic, Pro, Gln or Ser/Thr residues (Gerber *et al.*, 1994). A detailed classification and description of all currently known eukaryotic transcription factors is available from the TRANSFAC database (http://transfac.gbf.de/TRANSFAC/cl/cl.html).

Many transcription factors exist in an inactive state in the cytoplasm and enter the nucleus following stimulation. Transport through the pores of the nuclear membrane depends on the presence of an amino acid sequence acting as a nuclear

localisation signal. At the inactive state, the nuclear localisation signal of transcription factors may be modified or masked by phosphorylation or interaction with other proteins (Vandromme *et al.*, 1996). For example, in the absence of a signal, NF-κB is confined to the cytoplasm, bound to an inhibitor, called IκB. Upon cell stimulation by cytokines, IκB is phosphorylated and rapidly degraded by proteolysis. The nuclear localisation signal of NF-κB is thus uncovered, allowing transport to the nucleus and binding to the DNA (Verma *et al.*, 1995; Pahl and Baeuerle, 1996).

1.5.2 Transcriptional mRNA processing

Following transcription initiation, the RNA polymerase II is released from the PIC and starts RNA synthesis. While transcription is still in progress, the primary transcript undergoes a number of modifications, involving 5′ capping, intron splicing, and 3′ cleavage and polyadenylation. The phosphorylated CTD of RNA polymerase II is of crucial importance, as it can bind processing factors and carry them to their sites of action. The CTD has also been postulated to play a more direct role in the processing reactions themselves, as its loss has been shown to prevent maturation of the transcript (Zhao *et al.*, 1999; Proudfoot, 2000; Smith and Valcárcel, 2000).

1.5.2.1 Capping

Capping is completed before the primary transcript becomes 30 nucleotides long. The procedure involves three steps; first, removal of a phosphate group from the 5′-triphosphate terminus of the elongating RNA chain, then formation of a 5′-5′ phosphodiester bond between the 5′-diphosphate and a guanine residue, and finally methylation of guanine to form a 7-methyl guanine cap (figure 1.7). The three steps are catalysed by a phosphatase, a guanyl transferase and a methylase, respectively. The latter two enzymes are recruited to the site of transcription by binding to the CTD of RNA polymerase II. Capping occurs only in eukaryotes and is thought to protect the mRNA from attack by exonucleases. The methylated cap is also crucial for binding of the mRNA to the ribosome, a prerequisite for translation initiation (Hawkins, 1996; Latchman, 1998b; Proudfoot, 2000).

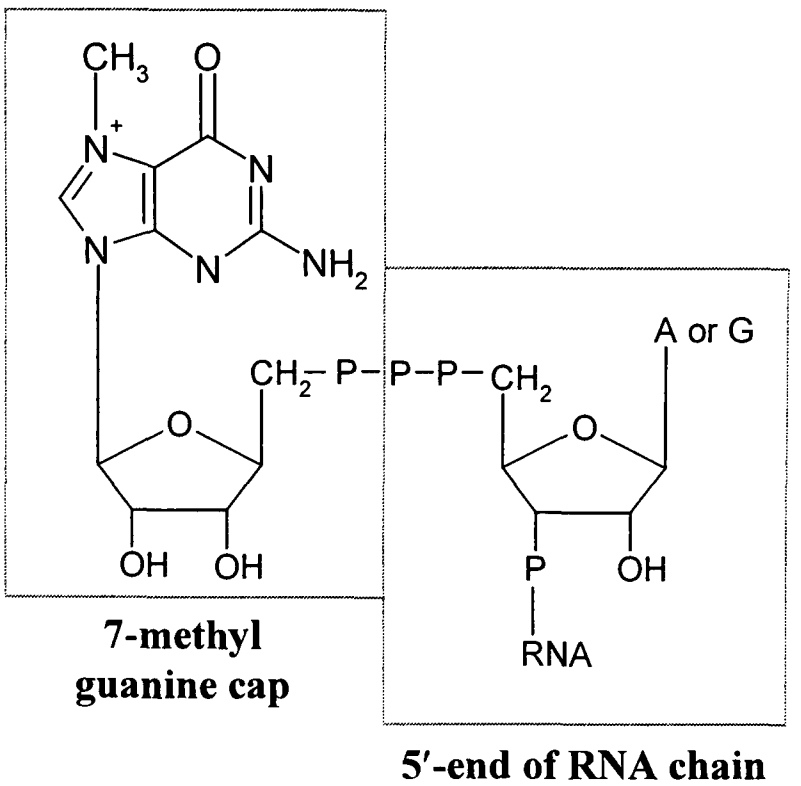


Figure 1.7: The structure of the 5′ cap. (*Adapted from Hawkins, 1996*).

1.5.2.2 Splicing

Splicing of the primary transcript is a co-transcriptional process (Proudfoot, 2000), involving precise intron removal and joining of the exons to form a continuous protein-coding mRNA sequence (Smith *et al.*, 1989). Splicing reactions take place on spliceosomes, consisting of small nuclear ribonucleoprotein particles (snRNPs), heterogeneous ribonucleoprotein particles (hnRNPs) and various protein factors. In the cell nucleus, splicing factors form clusters, visible by electron microscopy as "speckles" (Hodges and Bernstein, 1994; Elliott, 2000). Splicing is an ATP-dependent process, involving two transesterification steps. Initially, the 5′ end of the intron (the "donor splice site") is cleaved and linked to the 2′-hydroxyl group of a conserved adenosine (the "branch point"), via a 2′-5′ phosphodiester bond. This results in the formation of a lariat structure and separation of the intron from the upstream exon. The second step involves cleavage of the 3′ end of the intron (the "acceptor splice site"), followed by ligation of the two exons and release of the intervening intron as a lariat (Horowitz and Krainer, 1994; Hodges and Bernstein, 1994; Lopez, 1998).

Of great importance for accurate intron splicing is the sequence of the splice sites. The vast majority of introns start with a GU and end with an AG dinucleotide (Zaphiropoulos, 1998). Mutations at these particular sites usually prevent splicing,

leading to intron retention, exon skipping or activation of adjacent cryptic splice sites (Hodges and Bernstein, 1994). In contrast, deviation of the surrounding sequence from the consensus (AG/GURAGU for the donor and YAG/G for the acceptor splice site) is generally tolerated, but may affect the inherent strength of the splice site. The branch point sequence YNYURAY (underlined is the adenosine involved in lariat formation) is typically found 20-40 nucleotides upstream of the acceptor splice site. The branch point sequence and a downstream polypyrimidine tract are crucial for positioning of the acceptor splice site (Andreadis *et al.*, 1987; Smith *et al.*, 1989; Norton, 1994; Horowitz and Krainer, 1994).

Spliceosome assembly involves: a) binding of U1snRNP to the donor splice site, b) binding of splicing factor 1 (SF1) to the branch point and c) binding of the 65 and 35kDa subunits of U2snRNP auxiliary factor (U2AF) to the polypyrimidine tract and the acceptor splice site, respectively. Subsequent recruitment of U2snRNP to the branch point requires ATP hydrolysis. Specific binding of U1 and U2 snRNPs to the primary transcript is based on sequence complementarity between the U1 and U2 RNAs, respectively, and elements of the two splice sites. A tripartite complex of U4, U5 and U6snRNPs is then recruited to the spliceosome. U5snRNP interacts with exonic sequences adjacent to both splice sites. U6snRNP becomes catalytically active following dissociation from U4snRNP and interaction with U2snRNP, thus displacing U1snRNP (Rio, 1993; Norton, 1994; Lopez, 1998; Smith and Valcárcel, 2000). The main steps leading to intron splicing are outlined in figure 1.8.

The two splice sites are brought close to each other by a number of non-snRNP factors, called SR-proteins. These interact with both the primary transcript and the components of the splicing apparatus. The SR-proteins have a N-terminal domain containing one or two RNA recognition motifs, and a characteristic Ser/Arg-rich C-terminal domain, responsible for interaction with other SR proteins. Other splicing factors also have SR-like domains, including U1snRNP, U2AF and a group of CTD-associated SR-like proteins (CASPs). The latter link the CTD of RNA polymerase II to the spliceosome (Norton, 1994; Horowitz and Krainer, 1994; Stojdl and Bell, 1999; Smith and Valcárcel, 2000; Proudfoot, 2000).

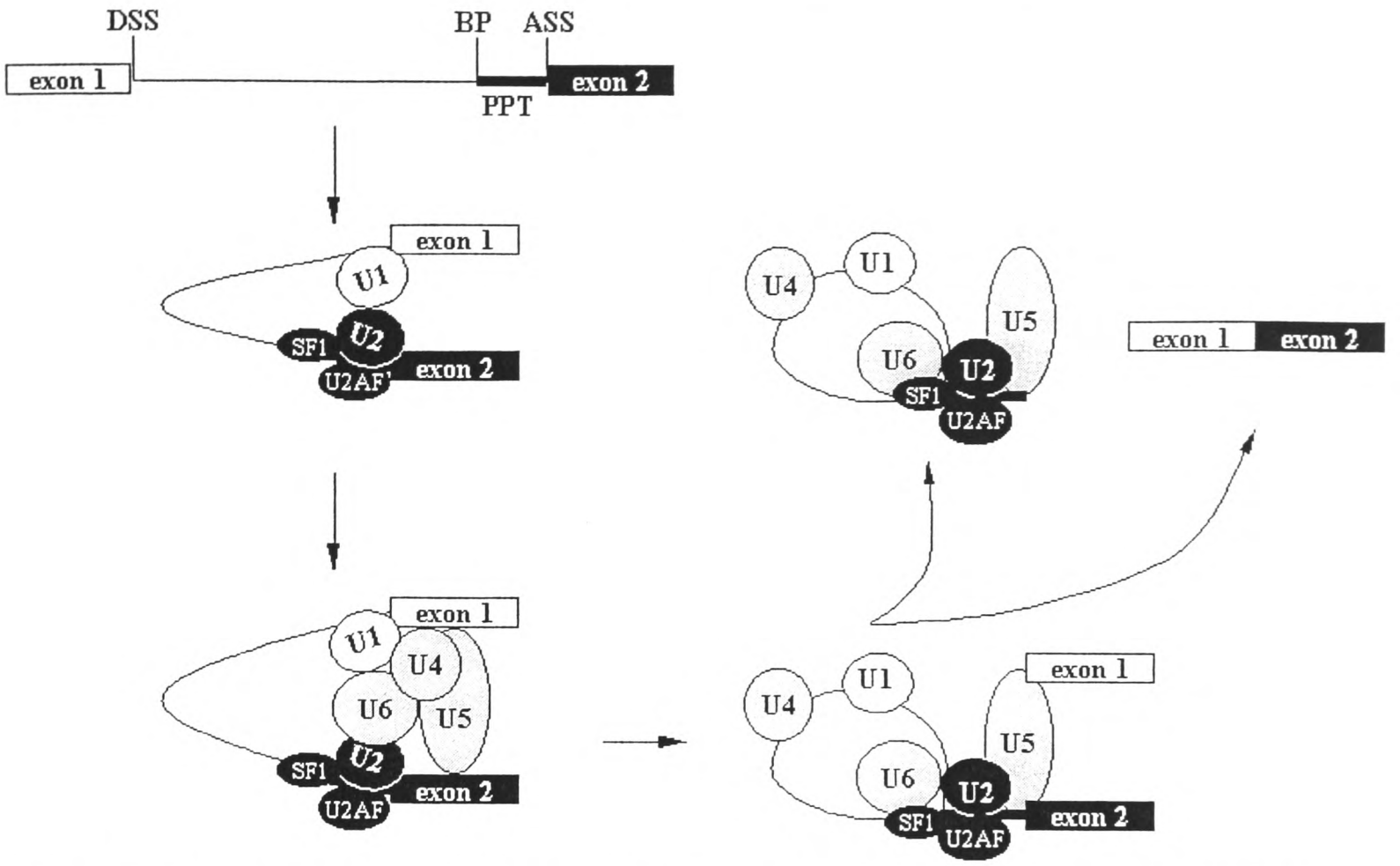


Figure 1.8: Mechanism of RNA splicing. DSS and ASS are the donor and acceptor splice site, respectively. BP is the branch point and PPT the polypyrimidine tract. Various snRNPs of the spliceosome are labelled U1-U6. (*Adapted from Hawkins, 1996*).

Alternative splicing is a very effective way of producing multiple, functionally distinct, proteins from a single gene. Based on alignment of currently known ESTs to the sequence of the human genome, it has been estimated that the primary transcript of at least 35% of the human genes is alternatively spliced. This number is likely to be a significant underestimate, as current EST collections do not represent all tissues, developmental stages or physiological conditions, and may also have missed rare transcripts (Black, 2000; Graveley, 2001). Fine regulation of splice site utilisation is of crucial importance for the production of mRNA molecules carrying the correct coding information. Up to 15% of disease mutations have been associated with defective splicing, often indirectly, i.e. by changing normal splicing patterns, rather than damaging a splice site (Philips and Cooper, 2000; Hou and Conboy, 2001).

Instead of scanning many kilobases of intronic sequence for the presence of splice sites, the splicing apparatus is capable of recognising internal exons, which are much shorter (typically 50-300 bases) and strictly confined between the splice sites of the two flanking introns. This "exon definition" model also applies to terminal exons,

recognised from the presence of a cap structure at their 5′ end or a polyA signal at their 3′ end (Chabot, 1996; Zaphiropoulos, 1998; Elliott, 2000; Smith and Valcárcel, 2000). For example, it is known that the presence of an acceptor splice site before the polyA signal can enhance polyadenylation, an effect abolished by insertion of a donor splice site between the two elements (Zhao *et al.*, 1999).

Splice site selection is influenced by several factors, such as the presence of RNA secondary structure around the splice site, the size of the adjacent exon, the inherent strength of the splice site and the presence of splicing silencers or enhancers (Horowitz and Krainer, 1994; Webster and Huang, 1999). Moreover, splice sites are not always accessible to the splicing apparatus, due to the formation of tight ribonucleoprotein complexes between the transcribed RNA and various hnRNPs (Rio, 1993; Smith and Valcárcel, 2000). Some hnRNPs (e.g. hnRNP-A1) have been shown to antagonise the activating effect of SR proteins, while others (e.g. hnRNP-H and hnRNP-I) may have either a positive or a negative effect, depending on the RNA template (Horowitz and Krainer, 1994; Elliott, 2000; Smith and Valcárcel, 2000).

Perhaps the most intriguing feature of SR-proteins is their ability to promote splice site utilisation via binding to exonic splicing enhancers (ESEs). These are purine-rich sequences within exons, favouring the selection of weak splice sites. ESE-bound SR-proteins stabilise the interaction between the splicing apparatus and the involved splice site, thus supporting its utilisation over alternative sites. The relative abundance and the phosphorylation status of different SR proteins can modulate splice site selection in a tissue- or stage-specific manner. Changes in splicing efficiency have been linked to tumourigenesis, while a number of genetic diseases have been attributed to mutations within ESEs. Although often silent, such mutations can prevent the production of a particular mRNA molecule from an alternatively spliced primary transcript (Horowitz and Krainer, 1994; Chabot, 1996; Cooper and Mattox, 1997; Stojdl and Bell, 1999; Webster and Huang, 1999; Elliott, 2000).

1.5.2.3 Cleavage and polyadenylation

Processing of the 3′ end of the primary transcript begins before splicing is completed and involves endonucleolytic cleavage of the molecule, release from RNA polymerase II and end-tailing by addition of roughly 250 adenosine residues. The

cleavage/polyadenylation apparatus consists of several factors, including the cleavage/polyadenylation specificity factor (CPSF), the cleavage stimulatory factor (CstF), the cleavage factors (CF) I_m and II_m, the polyA-polymerase (PAP) and the polyA-binding protein II (PABP II). CPSF recognises the highly conserved polyA signal (consensus sequence: AAUAAA), located 10-30 nucleotides before the cleavage site. CstF binds to another, less well-conserved, U-rich or GU-rich downstream element (DSE), usually found 10-30 nucleotides after the cleavage site. Interaction between CPSF and CstF stabilises the cleavage/polyadenylation complex and determines the site of cleavage, which is located between the polyA signal and the DSE, usually after a CA dinucleotide. It is also established that both CPSF and CstF interact with the CTD of RNA polymerase II. This association may begin before the polymerase leaves the promoter, as CPSF is known to interact with TFIID. Truncation of the CTD has been demonstrated to block cleavage and polyadenylation of the primary transcript. Other auxiliary elements (e.g. a U-rich element before the polyA signal of some viral and cellular genes, or a G-rich element after the DSE of some SV40 genes) bind factors that stabilise anchoring of the cleavage/ polyadenylation complex to the RNA and modulate the efficiency of 3′ end processing (Birnstiel *et al.*, 1985; Proudfoot and Whitelaw, 1988; Hawkins, 1996; McCracken *et al.*, 1997; Zhao *et al.*, 1999; Proudfoot, 2000).

Cleavage of the primary transcript is followed by PAP-mediated polyadenylation of the 3′ end. The first 10 adenosine residues are added slowly, but the reaction rate increases dramatically, following binding of PABP II to the growing poly-A tail (Zhao *et al.*, 1999). Cleavage and polyadenylation are coupled to transcription termination, as in the absence of an intact polyA signal the RNA polymerase II continues transcribing beyond the normal termination region (Proudfoot and Whitelaw, 1988). It is likely that transcription past the polyA signal is followed by loss of a processivity factor and random release of RNA polymerase II from the DNA template. Alternatively, cleavage of the primary transcript after the polyA signal may render the polymerase-bound RNA molecule more susceptible to exonuclease attack, due to its lack of a protective 5′ cap. Degradation may proceed faster than polymerisation, forcing the RNA polymerase II to terminate transcription (Proudfoot, 2000).

Polyadenylation is necessary for efficient export of the mature transcript to the cytoplasm, as well as for adequate mRNA protection against degradation. The site of polyadenylation may also affect translatability of the message (Hawkins *et al.*, 1996; Edwalds-Gilbert *et al.*, 1997; Zhao *et al.*, 1999). Defective polyadenylation (e.g. due to mutations within the polyA signal) is the cause of many diseases, including some forms of thalassaemia and a lysosomal storage disorder (Zhao *et al.*, 1999). On the other hand, regulation of polyadenylation efficiency is a very effective way of adjusting the amount and nature of gene products to the specific needs of the cell. This is achieved via selection of alternative polyA signals with variable inherent strengths. Genes with multiple polyA signals downstream of a single 3′ terminal exon produce mRNAs with 3′ UTRs of variable length. These messages carry the same genetic information but may be different in terms of stability and translatability. Furthermore, utilisation of alternative polyA signals, located not only in the 3′ UTR, but also within introns, can produce messages with variable 3′ terminal exons (Edwalds-Gilbert *et al.*, 1997). Therefore, alternative polyadenylation, like alternative splicing, is a means of expanding protein diversity in complex biological systems.

1.6 Project aims

The broader aim of the work described in this thesis has been to expand the current knowledge of the molecular genetics of murine NAT isoenzymes, towards development of an effective animal model for studying the role of NAT in endogenous metabolism and disease. The following objectives were pursued:

- Identification and functional characterisation of novel polymorphisms in the coding region of the *Nat* genes from different mouse strains.
- Generation of molecular tools for the production of transgenic mice lacking or overexpressing NAT.
- Analysis of the genomic region surrounding murine *Nat2* and identification of novel genetic markers for refined mapping of the *Nat* locus.
- Elucidation of the expression profile of the mouse *Nat* genes in different adult tissues and the preimplantation embryo.
- Determination of the genomic structure of murine *Nat2* and comparison with the structure of genes for other mammalian NATs, by combining data from current genome projects.
- Characterisation of core promoter elements driving transcription of the mouse *Nat2* gene and preliminary search for other regulatory sequences in the surrounding region.

CHAPTER 2

Materials and Methods

2.1 Materials and general methods

2.1.1 Water

Deionised water (dH_2O) for general use (e.g. preparation of solutions) was produced by a deioniser-linked reverse osmosis system (Purite). Double deionised water (ddH_2O) for use in molecular biology reactions was produced by an additional purification step, using the StillPLUS system (Purite). Ribonuclease (RNase)-free water was prepared by overnight treatment with diethyl pyrocarbonate (DEPC) (1ml DEPC per 1L ddH_2O), followed by autoclaving (section 2.1.7).

2.1.2 Chemicals and molecular biology reagents

All chemicals were provided by Sigma-Aldrich or Merck-BDH, unless otherwise stated. Reagents for bacterial culture were from Difco and for cell culture from GibcoBRL. Enzymes for molecular biology reactions and their buffers were from Promega, unless otherwise stated.

2.1.3 Primers

Primers were provided by Oswell, GibcoBRL or Sigma-Genosys. Lyophilised primers were dissolved in 1ml of ddH_2O. Stocks of primers, as well as working aliquots of 10pmol/µl, were stored at -20°C. Table 2.1 provides the sequence and melting temperature (Tm) of vector-specific primers used in this study. Standard primers were designed using the *Primers!* programme by R.J. Resnick, on a Macintosh computer.

Table 2.1 Vector-specific primers used for PCR amplification or sequencing from plasmid templates.

Primer name	Orientation	Sequence (5′→3′)	Tm (°C)	Vector specificity
M13F	Forward	AGGGTTTTCCCAGTCACGA	60	pBluescript, pGEM-T Easy
M13R	Reverse	ACACAGGAAACAGCTATGAC	58	pBluescript, pGEM-T Easy
T7-Promoter	Forward	TAATACGACTCACTATAGGG	47	pBluescript, pGEM-T Easy, pTargeT
pTargeT	Reverse	TTACGCCAAGTTATTTAGGTGACA	55	pTargeT

2.1.4 Cloning vectors

All cloning vectors were from Promega, apart from the pBluescript SK(-) vector, which was from Stratagene. All vectors provide ampicillin resistance to transformed bacterial cells. The graphic maps of the vectors are provided in Appendix 1.

2.1.5 Laboratory animals

Mouse inbred strains (Balb/c, C57Bl/6J, A/J) were purchased from Harlan UK. Outbred strains (PO and TO) were provided by the Department of Zoology and the Department of Experimental Psychology of the University of Oxford. The wild-derived inbred strains MSP (*Mus spretus*) and MCA (*Mus musculus castaneus*) were purchased from the MRC Mammalian Genetics Unit at Harwell, UK.

2.1.6 Tissue preparation

Mice were sacrificed either by cervical dislocation or in a CO_2 chamber. Whole organs were removed from dissected animals, quickly cut into pieces (when necessary), and frozen in liquid nitrogen for storage. Tissue was not allowed to thaw prior to homogenisation, which was carried out on ice, using a mechanical homogeniser (T25 basic Labor Technick).

2.1.7 Sterilisation

Solutions, glassware and plasticware for molecular biology procedures and bacterial or mammalian cell culture were sterilised by autoclaving (30min, 121°C,

15p.s.i.). Heat-sensitive solutions were sterilised by filtration through 0.2μm sterile acrodisc syringe filters (Whatman). Bacterial and mammalian cultures were handled in a laminar flow cabinet decontaminated with 70% (v/v) ethanol, prior to and after use.

2.2 Molecular biology methods

2.2.1 Isolation and purification of nucleic acids

2.2.1.1 Isolation of genomic DNA from tissue homogenates

For isolation of high molecular weight genomic DNA, tissue (approximately 50mg) was homogenised (section 2.1.6) in 5ml STE buffer [0.1M NaCl, 1mM ethylene-diamine-tetra-acetate (EDTA) and 40mM Tris-HCl pH 7.5] and incubated overnight at 50°C, after addition of 0.5% (w/v) sodium dodecyl sulphate (SDS) and 100μg/ml proteinase K [from a 100mg/ml stock in 10mM Tris-HCl pH 7.6, 2mM EDTA and 0.1% (w/v) SDS]. The DNA was extracted by addition of 5ml phenol:chloroform:isoamyl alcohol [24:24:1 (v/v/v)], followed by mixing on a rotator for 10min (room temperature) and incubation on ice for 30min. The mixture was centrifuged at 5000g (4°C) for 30min, and the upper (aqueous) phase was transferred to a sterile tube. The DNA was then precipitated by addition of 2 volumes of ice-cold absolute ethanol and 0.1 volume of 2M sodium acetate pH 5.5, followed by mixing and incubation at room temperature (5min). The DNA precipitate was recovered on a thin glass rod and rinsed with 70% (v/v) ethanol. Excess ethanol was removed by gently squeezing the glass rod against the walls of the tube, and the DNA was eventually released in 1ml of TE buffer (10mM Tris-HCl and 1 mM EDTA, pH 7.5), where it was allowed to dissolve overnight on a rotator (4°C). The DNA preparations were stored at 4°C.

2.2.1.2 Isolation of plasmid DNA from bacterial cultures

Isolation of plasmid DNA from overnight *E. coli* cultures (section 2.2.3.10) was performed using the Wizard Plus SV Minipreps DNA Purification System (10ml cultures for yield of up to 20μg of plasmid DNA) or the Wizard Plus Midipreps DNA

Purification System (100ml cultures for yield of up to 200μg of plasmid DNA), both by Promega. Plasmid DNA was stored at 4°C.

2.2.1.3 Purification of DNA from agarose gels and solutions

DNA was recovered from agarose gels, following electrophoresis (section 2.2.2.1), using the Gene-Clean II Kit (Bio-101), the QIAEX II Gel Extraction System (QIAGEN) or the QIAquick Gel Extraction Kit (QIAGEN). The same kits or ethanol precipitation (Sambrook *et al.*, 1989) were used for purification of DNA from enzymatic reaction mixtures.

2.2.1.4 Isolation of total RNA and mRNA from tissue homogenates and cell lysates

Special precautions were taken to protect RNA preparations from contamination with RNases, according to Promega's *Protocols and Applications Guide*. All solutions were prepared in DEPC-treated ddH_2O (section 2.1.1).

For isolation of total RNA from less than 60mg of tissue, the SV Total RNA Isolation System (Promega) was used. Deoxyribonuclease I (DNaseI) treatment was carried out during the extraction, according to the manufacturer's instructions. The system was not used with spleen, thymus and ES cells, because the tissue/cell extracts were very viscous.

Total RNA was extracted from more than 60mg of tissue or cultured cells, harvested from 75cm^2 flasks (section 2.2.4.1), using the Tri-Reagent (Sigma). Homogenisation was performed in 1ml of Tri-Reagent per 60-100mg of tissue, while cultured cells were lysed in 7.5ml of Tri-Reagent. Following a 5min incubation at room temperature, chloroform (200μl per 1ml of Tri-Reagent) was added, and the lysates were mixed and left at room temperature for 3min. Each mixture was then centrifuged at 14,000rpm for 15min in a microcentrifuge (4°C), and the upper (aqueous) phase was transferred to a sterile tube. Total RNA was precipitated by addition of 0.5ml 100% (v/v) isopropanol per 1ml of Tri-Reagent, followed by incubation at room temperature for 5-10min and centrifugation as above. The isopropanol was discarded and the RNA pellet was rinsed with 75% (v/v) ethanol. Following centrifugation and removal of the ethanol, the pellet was briefly dried and

resuspended in 50µl of RNase-free water (section 2.1.1) per 1ml of Tri-Reagent. The quality of the preparations was assessed by electrophoresis of 1-2µg of RNA product on a 1% (w/v) agarose gel (section 2.2.2.1). When DNA contamination of the preparations was suspected, treatment with DNaseI was performed (section 2.2.3.1).

Messenger RNA was isolated from total RNA, using a biotinylated oligo-dT probe and streptavidin-coated paramagnetic particles, supplied with Promega's PolyATtract mRNA Isolation System. Total RNA and mRNA preparations were stored at -70°C.

2.2.1.5 Measurement of concentration and purity of nucleic acid solutions

The concentration of DNA and RNA solutions was spectrophotometrically determined at 260nm, using quartz cuvettes of 1cm path length. The concentration was calculated based on the assumption that 1 unit of optical density (OD_{260}) corresponds to 50µg/ml of double-stranded DNA or 40µg/ml of single-stranded RNA. The ratio of OD at 260 and 280nm (OD_{260}/OD_{280}) was used as a measure of the purity of nucleic acid solutions, with solutions of high purity providing a ratio of 1.7 or greater (from Promega's *Protocols and Applications Guide*). DNA quantitation was also performed by electrophoresis (section 2.2.2.1) of samples alongside the BioMarker EXT DNA ladder (CAMBIO).

2.2.2 Gel electrophoresis of nucleic acids

2.2.2.1 Agarose gel electrophoresis

Agarose (Boehringer Mannheim or Sigma) was dissolved in 0.5xTBE (45mM Tris-borate and 1mM EDTA pH 8.0) or 1xTAE (40mM Tris-acetate and 1mM EDTA pH 8.0) buffer, and ethidium bromide was added to a final concentration of 0.2µg/ml. The gel was allowed to set and was subsequently run at 70-120V. Nucleic acids were routinely separated on 1% (w/v) gels, unless otherwise specified. For separation of fragments smaller than 200bp, gels containing 2-4% (w/v) Metaphor agarose (Flowgen), or a mixture of multi-purpose and Metaphor agarose, were used. Prior to loading, the samples were mixed with 6xSB loading dye [40% (v/v) sucrose and 0.25% (w/v) bromophenol blue in ddH_2O]. After electrophoresis, the DNA was

visualised and photographed under ultraviolet (UV) light. Molecular weight markers λ/*Hind*III and λ/*Eco*RI+*Hind*III (Promega), 1kb and 1kb Plus (GibcoBRL), and BioMarker EXT (CAMBIO) were run alongside the samples.

2.2.2.2 Polyacrylamide gel electrophoresis

The details of the different types of polyacrylamide gels used in this study are provided in table 2.2. A 40% (w/v) stock of acrylamide:bis-acrylamide (29:1) (Scotlab) was diluted in TBE buffer to the appropriate final concentration. N,N,N′,N′-tetramethyl-ethylenediamine (TEMED) [0.1% (v/v)] and ammonium persulphate (APS) [0.06% (w/v)] were then added, and the mixture was quickly poured between two glass plates. Gels were pre-run for 1h, then the samples were loaded and electrophoresis was resumed. After electrophoresis, the plates were opened and the gel was recovered on a sheet of Whatman 3MM filter paper (Sambrook *et al.*, 1989). Gels used for the analysis of non-radioactive products were stained in ethidium bromide (0.2μg/ml) for 20min, shaking gently. The Whatman paper was then removed and the DNA was visualised on the gel under UV light. Gels used for the analysis of radioactive products were dried onto the Whatman paper and exposed to X-ray films (Kodak X-OMAT) at -70°C, using intensifying screens.

Table 2.2: Description of polyacrylamide gels used for nucleic acid separation.

	Non-denaturing, ethidium bromide-stained gels	**Non-denaturing gels for analysis of ^{32}P-labelled products**	**Denaturing gels for analysis of ^{32}P-labelled products**
Acrylamide:bis-acrylamide (29:1)	6% (w/v)	4% (w/v)	6% (w/v)
Gel buffer	1xTBE	0.5xTBE	1xTBE+8M urea
Size of gel	38.5x20cm	16x18cm	38.5x20cm
Thickness of gel	1mm	1mm	0.4mm
Electrophoresis conditions	1000V for 2h in 1xTBE buffer	150V for 3h in 0.5xTBE buffer	2000V for 1h in 1xTBE buffer
Sample loading dye	6xSB[1]	10xGBX[2]	1x formamide dye[3]
Detection method	UV illumination	autoradiography	autoradiography
Size markers	10bp DNA ladder[4]	none	^{32}P-labelled DNA ladder[5]
Application	microsatellite analysis	EMSA[6]	RNase protection assay

[1]Section 2.2.2.1; [2]40% (v/v) glycerol, 0.2% (w/v) bromophenol blue and 0.2% (w/v) xylene cyanole in 250mM Tris-HCl pH 7.5; [3]95% (v/v) formamide, 0.05% (w/v) bromophenol blue and 0.05% (w/v) xylene cyanole; [4]From GibcoBRL; [5]Provided by Dr. K. Plant (Sir William Dunn School of Pathology, Oxford); [6]Electrophoretic Mobility Shift Assay.

2.2.3 Common molecular biology reactions

2.2.3.1 Treatment of RNA preparations with DNaseI

Treatment of total RNA preparations (section 2.2.1.4) with RNase-free DNaseI was performed with 1-2U of enzyme per 1μg of RNA in a 50μl total reaction volume, using the buffer supplied by the manufacturer. Incubation was carried out at 37°C for 1h and the enzyme was heat-inactivated at 65°C for 10min. The RNA was extracted with acid phenol (pH 4.5) and chloroform, followed by ethanol-precipitation (Sambrook *et al.*, 1989). The RNA pellet was dissolved in RNase-free water (section 2.1.1) and stored at -70°C.

2.2.3.2 Reverse transcription

Complementary DNA (cDNA) was generated from mRNA using 200U of Superscript II Reverse Transcriptase (GibcoBRL) per 100-200ng of template, in a reaction mixture containing the appropriate buffer, 500ng oligo-dT primer (Promega), 500μM each deoxyribonucleotide triphosphate (dNTP) (Promega), 0.01M dithiothreitol (DTT) (GibcoBRL) and 20U RNase inhibitor (Promega). A 12μl mix containing only RNA and oligo-dT primer was initially heated to 70°C for 10min, to allow template denaturation. It was then placed at 42°C and the rest of the reagents were added, to a final reaction volume of 20μl. Synthesis of cDNA was carried out at 42°C for 1h and the enzyme was finally heat-inactivated at 70°C for 15min.

2.2.3.3 Polymerase chain reaction (PCR)

PCR was performed using 500ng of genomic DNA, 50ng of plasmid DNA or 3-5μl of cDNA product (section 2.2.3.2) as template. Amplification was also carried out using bacterial or yeast colonies as template. Reactions were performed in a 50 or 100μl final volume with 2.5U *Taq*- or 3U high-fidelity *Pfu*-DNA polymerase, 200μM each dNTP, 0.25μM each primer, 2mM $MgCl_2$ and the appropriate buffer. *Pfu*-DNA polymerase was used with buffer supplemented with 2mM $MgSO_4$, so addition of $MgCl_2$ was not required. A reaction lacking template served as negative control. Unless the PCR machine possessed a heated lid, the reaction mix was overlaid with 100μl of mineral oil. The PCR cycling conditions were 95°C for 5min (10min for lysis of bacterial or yeast cells), followed by 35 cycles of: a) template denaturation at

95°C for 30sec, b) primer annealing at the appropriate temperature for 30sec, and c) product extension at 72°C, for 45sec per 1kb of target sequence. Final extension took place at 72°C for 5min.

2.2.3.4 Long and Accurate PCR (LA-PCR)

LA-PCR was performed for the amplification of fragments 3-8kb long. Reactions were performed in a 50µl mixture containing a 10:1 (5U:0.5U) ratio of *Taq:Pfu*-DNA polymerases in *Taq* enzyme buffer, plus 200µM of each dNTP, 0.25µM of each primer and 0.01% (w/v) gelatin. The optimal concentration of $MgCl_2$ ranged between 1.5-2.25mM, with lower concentrations favouring the amplification of longer sequences. The optimal amount of genomic DNA template was 250-500ng. The cycling conditions were as above (section 2.2.3.3).

2.2.3.5 Sequencing of DNA

Automated sequencing of plasmid DNA (500ng) or gel-purified (section 2.2.1.3) PCR product (100ng) was carried out using the ABI Automated Sequencer System at the Advanced Biotechnology Centre (Imperial College, Charing Cross Hospital, London) or the DNA Sequencing Facility (Department of Biochemistry, University of Oxford). The data were viewed using the EditView ABI Automated DNA sequence Viewer 1.0 and DNA Strider 1.2 programmes, on a Macintosh computer, or the Chromas 1.45 and DN*Assist* 1.02 programmes, on a PC. The GCG Winsconsin Package 10.2 was used for data analysis.

2.2.3.6 Restriction enzyme digestion

Digestions were performed with 10U of restriction enzyme per 1µg of DNA template in 10µl reaction volume, using the buffer recommended by the manufacturer, supplemented with 0.1mg/ml of acetylated bovine serum albumin (BSA). Incubations were carried out at the temperature specified by the manufacturer, for 2-16h. When simultaneous digestion with two restriction enzymes was performed, the reaction volume was doubled, so that the amount of glycerol (contained in the storage buffer of each enzyme) was not more than 10% (v/v) of the total reaction volume.

2.2.3.7 Addition of an A-tail to the 3′ ends of DNA fragments

Addition of an adenosine residue to the 3′ ends of blunt-ended DNA fragments was carried out using 2.5U *Taq*-DNA polymerase and 1μg of template, in a 100μl reaction mix, also containing 50μM dATP, 2mM $MgCl_2$ and buffer. Incubation took place at 70°C for 30min and the product was then purified, as described in section 2.2.1.3. A-tailing of restriction fragments with 5′-protruding ends was carried out as above, but all four dNTPs were added to the reaction mixture, each at a concentration of 50μM.

2.2.3.8 Dephosphorylation of 5′ ends

Dephosphorylation of the 5′ ends of vectors (section 2.1.4), linearised by restriction digestion (section 2.2.3.6) prior to cloning, was carried out in a 50μl reaction mixture, containing 0.1U calf intestine alkaline phosphatase (CIAP) per 1pmol of ends of template [pmol of ends = (μg DNA/ kb size of DNA) x 3.04, according to Promega's *Protocols and Applications Guide*], plus the appropriate buffer. Incubation took place at 37°C for 1h, then 0.5M EDTA (2μl) was added and the mixture was heated to 65°C for 20min to inactivate the enzyme.

2.2.3.9 Ligation

Ligation was carried out in a 10μl reaction mixture, containing 3-6U T4 DNA ligase and its buffer, plus 50 or 100ng of vector. Insert DNA was added to the mixture at insert:vector molar ratios ranging from 1:1 to 8:1 [ng of insert = (ng of vector x kb size of insert/kb size of vector) x insert:vector molar ratio, according to Promega's *Protocols and Applications Guide*]. Reactions were performed at 4°C overnight. Control reactions containing vector alone or vector with ligase, but no insert, were routinely performed, to determine the number of background colonies resulting from undigested or self-ligating vector.

2.2.3.10 Transformation and culture of *E. coli* cells

For transformation of *E. coli* cells with plasmid DNA, JM109 High Efficiency Competent Cells (Promega) were routinely used. Competent cells (50μl), slowly thawed on ice, were mixed with DNA (1-50ng in a volume of less than 10μl) and left

on ice for 10min. The cells were subsequently heat-shocked at 42°C for 45sec and returned to ice for 2min. Following addition of 1ml SOC medium [100ml contain 2g Bacto-Tryptone, 0.5g Bacto-Yeast Extract, 1ml 1M NaCl, 0.25ml 1M KCl, 1ml 2M glucose and 1ml 2M Mg^{2+} stock (101.5g $MgCl_2 \cdot 6H_2O$ and 123.3g $MgSO_4 \cdot 7H_2O$ per 500ml of solution)], the cells were incubated for 1-2h at 37°C on a shaker. Then, the culture (200μl) was spread on solidified LB-agar medium (10g Bacto-Tryptone, 5g Bacto-Yeast Extract, 10g NaCl and 15g Bacto-Agar per 1L of medium) supplemented with 100μg/ml ampicillin. For blue/white selection of colonies, the medium was also supplemented with 0.5mM isopropyl-β-D-thiogalactopyranoside (IPTG) and 40μg/ml 5-bromo-4-chloro-3-indoyl-β-D-galactopyranoside (x-gal) (50mg/ml stock in dimethylformamide). Incubation took place overnight. Single colonies were picked and used to inoculate 10 or 100ml of liquid LB medium (without Bacto-Agar), supplemented with 100μg/ml ampicillin. Following overnight incubation at 37°C on a shaker, liquid cultures were used for isolation of plasmid DNA (section 2.2.1.2), as well as for the preparation of 1ml stocks in 20% (v/v) glycerol (stored at -70°C).

2.2.4 Mammalian cell culture

2.2.4.1 Culture conditions

The BNL.CL2 mouse embryonic liver cell line was obtained from the American Type Culture Collection (ATCC, USA) and cultured in an atmosphere of 10% CO_2 in air, using 90% (v/v) high-glucose (4.5g/lt) Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% (v/v) heat-inactivated foetal bovine serum (FBS), 2mM L-glutamine, 0.1mg/ml kanamycin, 0.1mg/ml streptomycin and 100U/ml penicillin. Chinese hamster ovary (CHO) cells were provided by Dr. N. Johnson (Department of Pharmacology, Oxford) and cultured in an atmosphere of 5% CO_2 in air, using 90% (v/v) minimum essential medium (MEM) supplemented with 10% (v/v) FBS, 2mM L-glutamine, 1x non-essential amino acids, and the same antibiotics as above. Cells were grown at 37°C, in 75cm^2 flasks containing 14ml of medium.

Confluent cells were passaged by treatment with 2ml of 0.25% (w/v) trypsin-EDTA, at 37°C, for 10min (BNL.CL2 cells) or 2min (CHO cells). Before

trypsinisation, the culture medium was removed from the flasks and the cells were washed twice with phosphate-buffered saline (PBS) (137mM NaCl, 2.7mM KCl and 10mM sodium phosphate pH 7.4). Trypsinisation was terminated by addition of 4-6ml fresh culture medium and the cells were thoroughly mixed using a pastette. New flasks with medium were inoculated with 1ml of cell suspension. For RNA extraction or protein analysis, confluent cells were washed in PBS, as above, and then harvested in 1ml of PBS using a plastic scraper (Falcon). Cells were pelleted by centrifugation for 4min at 250g (room temperature).

Frozen stocks of cells were prepared by mixing 90% (v/v) of trypsinised cells in culture medium with 10% (v/v) dimethyl sulphoxide (DMSO). Aliquots (1ml) of cells were placed in a Cryo1°C Freezing Container (NALGENE) at -70°C overnight, to achieve a 1°C/min freezing rate, and then transferred to liquid nitrogen. Stocks were thawed rapidly in a 37°C water-bath and the cells were pelleted as above. The supernatant was discarded and 1ml of fresh culture medium was added, prior to transfer of the resuspended cells to a culture flask.

2.2.4.2 Transient transfection of mammalian cells

Fresh cultures were initiated in 75cm^2 flasks, from stocks frozen at an early passage. The day before the transfection, the cells were transferred to 100mm plates containing 10ml complete culture medium. The plating density was adjusted, so that the cells were 30-60% confluent on the day of the transfection. Three hours prior to transfection, fresh culture medium was added to the plates. Transfections were carried out by the calcium phosphate method, using the ProFection Mammalian Transfection System (Promega). A total amount of 20μg of plasmid DNA, purified by ethanol-precipitation, was used per transfection. The cells were incubated for 72h and then harvested. Mock transfections were also performed, without addition of DNA.

CHO cells were subjected to a glycerol shock 16h after transfection, by removal of the transfection medium, washing with PBS and addition of 3ml glycerol shock solution [15% (v/v) glycerol in HEPES-buffered saline (HBS), containing 25mM HEPES pH7.1, 140mM NaCl and 0.75mM Na_2HPO_4]. Following incubation at room temperature for exactly 2min, the glycerol shock solution was removed, the cells were rinsed twice with PBS, and fresh culture medium was added. BNL.CL2

cells were cultured for 40h in the transfection medium, which was then replaced by fresh medium, without a glycerol shock.

2.2.5 DNA hybridisation

2.2.5.1 Probe labelling

Probe labelling with digoxigenin (DIG) was carried out by PCR, using a 19:1 mixture of dTTP:DIG-dUTP (Boehringer Mannheim), added to the reaction mixture at a final concentration of 50μM. The PCR mixture also contained 50μM each dATP, dCTP and dGTP and was otherwise identical to that described in section 2.2.3.3. The product of the labelling reaction was gel-purified (section 2.2.1.3) and denatured at 100°C for 5min, before it was added to the hybridisation solution [50% (v/v) formamide, 0.1% (w/v) laurosylsarcosine, 0.2% (w/v) SDS, 5xSSC (prepared from a 20x stock containing 3M NaCl and 0.3M sodium citrate, pH 7.0) and 2x blocking solution (Boehringer Mannheim)] at a final concentration of 5-25ng/ml. The diluted probe mixture was stored at -20°C and used for several hybridisations. Prior to each use, it was heated to 100°C for 5min.

2.2.5.2 Hybridisation and chemiluminescence detection of the signal

DNA was fixed on pieces of positively charged nylon membrane (Boehringer Mannheim) by automatic UV-crosslinking in a UV Stratalinker 1800 (Stratagene). Membranes were pre-hybridised in 15ml of hybridisation solution (section 2.2.5.1) for 1h at 42°C, by rotating in a hybridisation oven (Hybaid). The hybridisation solution was then replaced by 15ml of probe mixture (section 2.2.5.1) and incubated overnight at 42°C. Membranes were washed twice in 2xSSC/0.1% (w/v) SDS, for 5min at room temperature, and then twice in 0.1xSSC/0.1% (w/v) SDS, at 42°C for 15min, shaking gently. Subsequent steps were carried out at room temperature on a shaker. Membranes were rinsed for 1-5min in washing buffer [0.1M maleic acid pH 7.5, 0.15M NaCl and 0.3% (v/v) Tween 20], and then incubated for 30-60min in 1x blocking solution (Boehringer Mannheim), prepared in buffer containing 0.1M maleic acid pH 7.5 and 0.15M NaCl, as described by the manufacturer. The membranes were incubated for 30min with alkaline phosphatase-conjugated anti-DIG-antibody (Boehringer Mannheim), diluted 1:10,000 in 1x blocking solution. Following two

15min washes, the membranes were equilibrated for 2-5min in detection buffer (0.1M Tris pH 9.5 and 0.1M NaCl) and incubated for 5min in CSPD® alkaline phosphatase chemiluminescence substrate (Boehringer Mannheim), diluted 1:100 in detection buffer. The membranes were then sealed in plastic bags, placed in a 37°C incubator for 15min and finally exposed to photographic films for 20-90min. For use in subsequent hybridisations, the membranes were stripped of the probe by incubation in 0.2M NaOH/0.1% (w/v) SDS at 37°C for 30min, then rinsed with 2xSSC and stored wet at 4°C.

2.2.6 Ribonuclease (RNase) protection assay

2.2.6.1 Ribo-probe labelling

The production and ^{32}P-labelling of the ribo-probe was carried out by *in vitro* transcription of a DNA fragment cloned into the pGEM-T Easy vector (Appendix 1.1a). The plasmid was linearised by restriction digestion with *Nde*I (section 2.2.3.6), and purified by phenol-chloroform extraction, followed by ethanol-precipitation under RNase-free conditions (Sambrook *et al.*, 1989). The 20µl *in vitro* transcription mixture contained 1µg of DNA template, 500µM of each rATP, rCTP and rGTP ribonucleotide triphosphates (Promega), 25µM rUTP (Promega), 1U/µl RNase inhibitor, 10mM DTT and 20U of T7 RNA polymerase and buffer. Labelling of the ribo-probe was achieved by addition of 40µCi [α^{32}P]-rUTP (Amersham) to the reaction mixture. Incubation took place at 37°C for 1.5-2h. The reaction was terminated by addition of 20µl formamide dye (table 2.2) and heating to 80°C for 5-10min. The product was electrophoresed on a 6% (w/v) denaturing polyacrylamide gel, as described in section 2.2.2.2. Following electrophoresis, one plate was removed, and the gel, still attached to the other plate, was covered with plastic wrap and exposed to a photographic film for 30sec in a dark room. Before the exposure, a fluorescent ruler was placed on the gel, to facilitate alignment of the film to the gel. The radioactive band was marked on the gel and excised with a sterile scalpel. The gel slice was then placed in a sterile tube with 400µl of elution buffer [0.5M ammonium acetate, 0.1% (w/v) SDS and 1mM EDTA], and incubated at 37°C for at least 1h, mixing thoroughly. The eluted ribo-probe was precipitated by addition of 1ml ice-cold absolute ethanol and centrifugation at 14,000rpm for 15min in a microcentrifuge

(room temperature). The pellet was air-dried briefly and resuspended in 100µl of R-loop buffer (80% v/v formamide, 40mM PIPES pH 6.4, 1mM EDTA and 400mM NaCl). An amount of ribo-probe providing about 500 radioactivity counts per second was used in each hybridisation reaction. Radio-labelled ribo-probes were stored at -20°C and used within two weeks.

2.2.6.2 RNA hybridisation and RNase treatment

Transiently transfected mammalian cells (section 2.2.4.2) were harvested from 100mm plates (section 2.2.4.1) and total RNA was extracted by the Tri-Reagent method (section 2.2.1.4). The isolated RNA (10µg) was ethanol-precipitated, following addition of 10µg of carrier tRNA to facilitate pellet formation. The air-dried pellet was resuspended in R-loop buffer and mixed with the ^{32}P-labelled ribo-probe in a final volume of 30µl. The mixture was denatured at 80°C for 10min and then hybridised at 56°C overnight. A control hybridisation with only carrier tRNA (20µg) and radiolabelled ribo-probe was included. A sample (5µl) of the control reaction was removed after hybridisation and prepared for electrophoresis (section 2.2.2.2), while the rest of the reaction was subjected to RNase digestion.

RNase mix (300µl), containing 40µg/ml RNase A and 1000U/ml RNase T_1, plus 10mM Tris pH 7.5, 5mM EDTA and 300mM NaCl, was added to the hybridisation products, which were then incubated at 18°C for 2h. Digestion was terminated by addition of 10µl 10% (w/v) SDS and 5µl 10mg/ml proteinase K, followed by incubation at 37°C for 30min. The product was then phenol-chloroform extracted and ethanol-precipitated (Sambrook *et al.*, 1989), after addition of carrier tRNA (10µg). The wet pellet was resuspended in 5-10µl of formamide dye (table 2.2) by vortexing and heating to 80°C. Samples were again heated to 80°C (10min), prior to loading to a pre-run 6% (w/v) denaturing polyacrylamide gel (section 2.2.2.2). Following electrophoresis, the gel was dried and exposed to a photographic film for 16-48h (section 2.2.2.2).

2.2.7 Electrophoretic Mobility Shift Assay (EMSA)

2.2.7.1 Probe labelling

End-labelling of PCR-amplified (section 2.2.3.3) probes was carried out in a reaction mixture containing 3.5pmoles of gel-purified (section 2.2.1.3) probe, 5-10U T4 polynucleotide kinase, 10μCi of [γ^{32}P]-dATP (Amersham) and the appropriate buffer. Incubation took place at 37°C for 10min, and the reaction was terminated by adding 1μl of 0.5M EDTA. The volume of the mixture was adjusted to 100μl with TE buffer, and unincorporated label was removed by gel filtration on a TE-Sephadex G-50 column, prepared in a 1ml syringe (Sambrook *et al.*, 1989). The eluted radioactive probe (80-100μl) was stored in -20°C and used within two weeks.

2.2.7.2 Preparation of cell extracts

Cells were harvested from 75cm^2 flasks (section 2.2.4.1) in 4ml TEN buffer (40mM Tris-HCl pH 7.4, 1mM EDTA and 150mM NaCl) and centrifuged at 250g (4°C) for 4min. The pellet was resuspended in 500μl of 40mM HEPES pH 7.9, 0.4M KCl, 1mM DTT, 10% (v/v) glycerol, 0.1mM phenylmethylsulphonyl fluoride (PMSF) and 0.1% (v/v) aprotinin, and the cells were lysed with three quick freeze-thaw cycles in a dry ice/ethanol bath. The lysate was centrifuged for 5min at 14,000rpm in a microcentrifuge (4°C), and the supernatant was stored at -70°C in aliquots.

2.2.7.3 Binding reactions

Cell extract, containing 2-4μg of total protein (section 2.3.1), was added to 1x binding buffer [the 5x stock contains 20% (v/v) glycerol, 5mM $MgCl_2$, 2.5mM EDTA, 2.5mM DTT, 150mM NaCl, 50mM Tris-HCl pH 7.5 and 0.25mg/ml poly(dI-dC)] and the mixture was incubated at room temperature for 10min. The ^{32}P-labelled probe (35fmol) was then added, and the incubation was continued for 30min. Control reactions were as follows: a) a "probe only" reaction, without cell extract, b) a "specific competition" reaction, to which a 50-fold molar excess (1.75pmol) of unlabelled probe was added, prior to the 10min incubation, and c) a "non-specific competition" reaction, to which a 50-fold molar excess (1.75pmol) of unlabelled non-specific competitor (section 6.2.9) was added. Loading buffer without dye [1μl from a stock containing 40% (v/v) glycerol in 250mM Tris-HCl pH 7.5] was added to the product of each reaction, apart from the "probe only" control, to which 1μl of GBX

loading dye (table 2.2) was added. The products were electrophoresed on a 4% (w/v) non-denaturing polyacrylamide gel and subjected to autoradiography, as described in section 2.2.2.2.

2.3 Methods for protein analysis

2.3.1 Determination of protein concentration

The protein concentration of soluble cell/tissue extracts was initially determined spectrophotometrically at 280nm and 260nm, in quartz cuvettes of 1cm path length (l). The concentration was calculated from the equation $c(mg/ml)=(1.55OD_{280}–0.76OD_{260})/l$ (Lorber and Giegé, 1999). Protein concentration was also measured by the Bradford method. Each sample was diluted in ddH_2O to an approximate concentration of 0.1-1.2mg/ml and 100µl were then mixed with 3ml of Bradford Reagent (Sigma). The OD was measured at 595nm using disposable cuvettes of 1cm path length. The blank was 100µl of ddH_2O in 3ml of Bradford Reagent. Each sample was assayed in triplicate and the average OD_{595} was calculated. The protein concentration was determined in reference to a standard curve, plotted as the OD_{595} versus the concentration of standard BSA solutions, prepared by serial dilution of a 1.4mg/ml stock in water (concentrations ranging from 0.04 to 1.4mg/ml).

2.3.2 Preparation of cell extracts

Tissue homogenates (section 2.1.6) were prepared in 2-3ml homogenisation buffer (PBS pH 7.5 with 1mM EDTA, 1mM DTT and 1mM Pefabloc) and 1ml aliquots were frozen in liquid nitrogen. Cultured cells were harvested (section 2.2.4.1) and resuspended in 1ml 20mM Tris-HCl pH 7.5, 1mM EDTA and 1mM DTT. Cells were lysed by sonication on ice (5 microns output, 3x5 bursts, 10sec intervals) in a MSE Soniprep 150 sonicator, and aliquots were stored at -70°C. Extracts were thawed immediately prior to use and centrifuged at 14,000rpm for up to 30min in a microcentrifuge (4°C). The supernatant was used for protein analysis or enzymatic activity assays. Thawed extracts were maintained on ice and were not refrozen.

2.3.3 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

Cell/tissue extracts were subjected to SDS-PAGE on 12% (w/v) polyacrylamide gels (8x10cm in size, 1mm thick). The separating gel (10ml) contained 12% (w/v) acrylamide:bis-acrylamide (29:1) in 1x separating gel buffer [the 4x stock contains 1.5M Tris-HCl pH 8.8 and 0.5% (w/v) SDS]. Polymerisation of the gel was initiated by addition of 50μl 10% (w/v) APS and 5μl TEMED. The stacking gel (10ml) contained 3.2% (w/v) acrylamide mix in 1x stacking gel buffer [the 4x stock contains 0.5M Tris-HCl pH 6.8 and 0.5% (w/v) SDS]. The stacking gel (at least 1cm high) was poured on top of the separating gel, after addition of 100μl 10% (w/v) APS and 10μl TEMED.

Each sample (up to 200μg of protein) was mixed with an equal volume of loading buffer A [0.5x stacking gel buffer, 7% (v/v) β-mercaptoethanol, 2.5% (w/v) SDS, 6mg/ml DTT (added fresh) and 0.03% (w/v) bromophenol blue] and heated to 95°C for 10min. Alternatively, samples were mixed with an equal volume of sample preparation buffer [200mM Tris-HCl pH 6.8, 8M urea, 2% (w/v) SDS and 6mg/ml DTT (added fresh)] and heated to 95°C for 10min. This was followed by addition of 0.1 volume of 83mg/ml iodoacetamide (prepared fresh), 0.1 volume of loading buffer B (5ml 4x stacking gel buffer, 2.5ml glycerol and 2.5mg bromophenol blue) and further heating to 90°C for 10min. Electrophoresis was carried out at 30mA, using 1x running buffer [the 10x stock contains 0.5M Tris-HCl pH 8.3, 0.5M glycine and 1% (w/v) SDS]. As a molecular weight standard, 5μl of pre-stained "Rainbow" ladder (Amersham) was used. The markers were treated in the same way as the samples, before loading to the gel. After electrophoresis, the gels were either blotted (section 2.3.4) or stained for 2min in Coomassie blue solution (3.75g Coomassie blue dye per 1L of 0.5M methanol and 1.87M acetic acid) and destained overnight in 0.5M methanol and 0.75M acetic acid.

2.3.4 Western blotting

Following electrophoresis, SDS-PAGE gels were blotted onto Hybond C nitrocellulose membrane (Amersham) at 4°C overnight, under 250mA constant current. Blotting was carried out in fresh 20% (v/v) methanol and 2.8g/lt Na_2HPO_4, using a BioRad Trans Blot Cell apparatus. The membrane was then removed from the

blotting apparatus and washed twice in TBST [50mM Tris-HCl pH 7.9, 150mM NaCl, and 0.05% (v/v) Tween 20] for 5min at room temperature, shaking gently. It was then placed in blocking solution [TBST containing 10% (w/v) dried skimmed milk] for 1-2h, and washed three times for 10min in TBST containing 3% (w/v) dried skimmed milk. The membrane was then incubated for 2h with primary antiserum, diluted 1:4,000 in TBST containing 3% (w/v) dried skimmed milk. The membrane was washed three times for 10min and incubated for 1h with secondary antibody, diluted 1:10,000 in TBST containing 3% (w/v) dried skimmed milk. The membrane was finally washed three times for 10min and rinsed briefly in TBST (without milk). Detection of antibody binding was performed by a chemiluminescence method, using the ECL reagents (Amersham), as described by the manufacturer. The membrane was exposed to a photographic film for 1-10min.

The primary antisera used, designated 184 and 193, are polyclonal antisera raised in rabbits against the BSA-conjugated C-terminal peptide (CVPKHGDRFFTI) of murine NAT2 and the soybean trypsin inhibitor (SBTI)-conjugated C-terminal peptide (LVPKCGNVFFTI) of murine NAT3, respectively (Ian Mills, PartII Thesis, Oxford 1996; Stanley *et al.*, 1997). The horse-radish peroxidase-conjugated anti-rabbit IgG secondary antibody was either a commercially available monoclonal antibody (Sigma), or a polyclonal antibody raised in goat (developed in the lab).

Semi-quantitative densitometry of films exposed for up to 1min was performed using the Quantity-One package by BioRad. Results were expressed relative to a fixed amount of internal standard (e.g. 2.8μg of pure BSA for antiserum 184), added to the samples prior to the SDS-PAGE.

2.3.5 Enzymatic activity assays

2.3.5.1 Dual-luciferase reporter assay

The dual-luciferase assay was performed using lysates of cells co-transfected with one pGL3 luciferase reporter vector and the control pRL-TK vector (Appendix 1.2). The procedure was carried out using the Dual-Luciferase Reporter Assay System (Promega), as described by the manufacturer. Cells were lysed on the plates with Passive Lysis Buffer (PLB) and each lysate (20μl) was then used for measurement of

the luminescence produced by *Photinus pyralis* (firefly) and *Renilla reniformis* (sea pansy) luciferases, after addition of the appropriate substrate. Luminescence was measured on a LKB Wallac 1250 or a PerkinElmer Lumat LB9507 luminometer.

2.3.5.2 NAT activity assay

NAT enzymatic activity was measured using the colourimetric method described by Andres *et al.*, 1985, with modifications by Sinclair *et al.*, 1998. The assays were performed in 20mM Tris-HCl pH 7.5/1mM DTT (added fresh), using up to 20μg/μl total protein [(contained in the cell/tissue extract (section 2.3.2)] and 0.1-0.2mM arylamine substrate. The samples were pre-warmed at 37°C for 5min and the reaction was initiated by addition of 0.4mM acetyl-CoA. The final reaction volume was 100 or 200μl. Following incubation for the appropriate length of time (up to 60min), an equal volume (100 or 200μl) of cold (4°C) 20% (w/v) trichloroacetic acid (TCA) was added, to stop the reaction. The precipitated protein was pelleted by centrifugation at 14,000rpm for 2min in a microcentrifuge (4°C). A zero time control, (TCA added prior to acetyl-CoA) and a control for spontaneous acetylation (ddH_2O instead of protein) were included. All assays were performed in triplicate.

The amount of unacetylated arylamine was determined spectrophotometrically, by mixing 200μl of each reaction product with 800μl of dimethylaminobenzaldehyde (DMAB) [5% (w/v) DMAB in 9:1 (v/v) acetonitrile:dH_2O, stored at 4°C in a dark bottle] and measurement of the OD at 450nm. The blank was 100μl 20mM Tris-HCl pH 7.5/1mM DTT, 100μl 20% (w/v) TCA, and 800μl 5% (w/v) DMAB.

The mean OD_{450} measured for each triplicate was plotted versus the assay time, and the rate of OD_{450} reduction [$\Delta(OD_{450})$/min] was calculated for the linear part of the curve. This rate was translated into substrate conversion rate (nmoles of N-acetylated substrate per min), using the appropriate standard curve (figure 2.1). The NAT specific activity (nmoles of N-acetylated substrate/min/mg of total protein) was then determined.

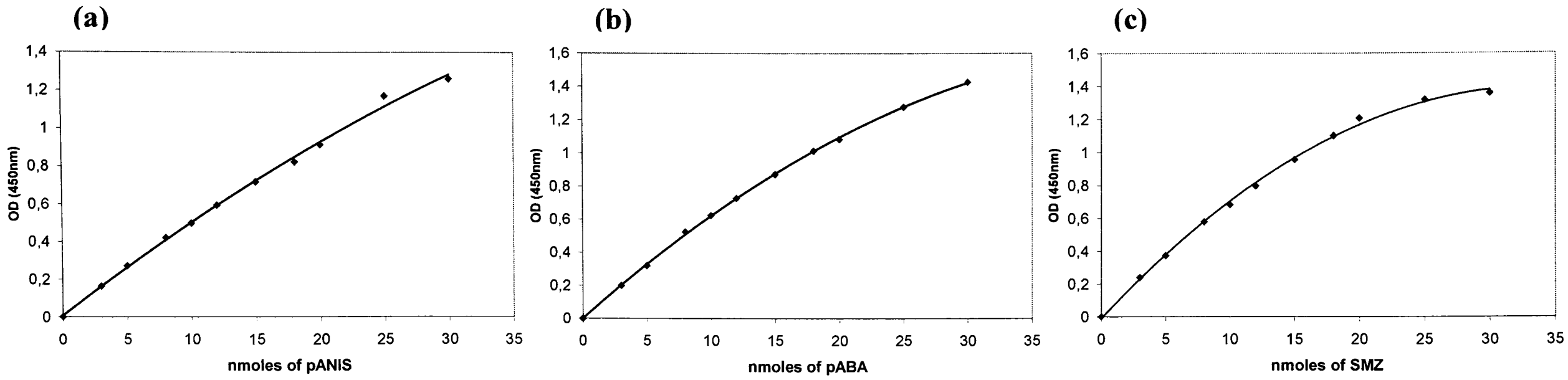


Figure 2.1: Standard curves for substrates used in NAT enzymatic activity assays. Standard solutions were prepared by addition of 0, 3, 5, 8, 10, 12, 15, 18, 20, 25 and 30nmoles of substrate in 100µl of 20mM Tris-HCl pH 7.5. To these, 100µl of 20% (w/v) TCA and 800µl of 5% (w/v) DMAB were added and the OD was measured at 450nm using the first of the standard solutions (0 nmoles of substrate) as blank. The change in OD was plotted against the amount of a) pANIS, b) pABA, c) SMZ, d) pAS and e) 5-AS substrate in the standard solutions.

CHAPTER 3

Identification and functional analysis of novel alleles for arylamine N-acetyltransferases in mouse strains

3.1 Introduction

Polymorphisms in the *NAT1* and *NAT2* genes, encoding for human NAT isoenzymes (Blum *et al.*, 1990a), have been associated with variable susceptibility to drug toxicity and cancer. Rodents have been extensively used as models for studying the human acetylation polymorphism, as inbred strains of rat, hamster and mouse have been identified to carry SNPs in their *Nat2* locus, responsible for the slow acetylator phenotype (reviewed in Hein *et al.*, 1997; sections 1.3.2 and 1.3.3). The classic acetylation polymorphism in mice, caused by an $Asn^{99}\rightarrow$Ile amino acid change in the NAT2 isoenzyme of the slow acetylating A/J and A/HeJ inbred strains (Martell *et al.*, 1991; Fretland *et al.*, 1997), has previously been employed to investigate the link between NAT activity and the toxic or carcinogenic properties of various xenobiotic compounds (reviewed in Levy *et al.*, 1992; section 1.3.3). To date, no polymorphism has been identified in the mouse *Nat1* and *Nat3* genes, which have, therefore, been considered as monomorphic (Martell *et al.*, 1991; Kelly and Sim, 1994; Fretland *et al.*, 1997; Estrada-Rodgers *et al.*, 1998b).

The aim of the work described in this Chapter has been to look for additional polymorphisms in the coding region of the three murine *Nat* genes and to assess the phenotypic effect of these polymorphisms, using enzymatic activity assays with known NAT substrates and Western blot analysis. Two inbred (129/Ola and CBA) and two outbred (PO and TO) mouse strains were used, together with the wild-derived inbred strains MCA and MSP, which are of *Mus musculus castaneus* and *Mus spretus* origin, respectively. The *Nat* genes of the fast acetylating C57Bl/6J and Balb/c, and

the slow acetylating A/J inbred mouse strains have previously been sequenced (Martell *et al.*, 1991; Kelly and Sim, 1994; Fretland *et al.*, 1997; Estrada-Rodgers *et al.*, 1998b) and were used here as reference. The experiments described in sections 3.2.1-3.2.3 were carried out in collaboration with Naomi Price (Biochemistry Part II student), under the supervision of the author.

3.2 Results

3.2.1 Genotyping of mouse strains for the *Nat2*8* and *Nat2*9* alleles

Strains PO, TO, MCA and MSP were genotyped for the *Nat2*8* and *Nat2*9* alleles, alongside the previously genotyped Balb/c, C57Bl/6J, 129/Ola and CBA strains, carrying the *Nat2*8* allele, and the A/J strain, carrying the *Nat2*9* allele (Payton *et al.*, 1999b). Genotyping was carried out using the PCR-RFLP method described in Payton *et al.* (1999b). A 133bp fragment of the *Nat2* coding region (nucleotide position 277 to 409) was amplified (section 2.2.3.3) from genomic DNA, using primers mNAT2-278 and mNAT2-409 (table 3.2.1). The *Nat2*9* allele contains a single *Tru*9I restriction site within the amplified region, created by the $A^{296}\rightarrow T$ nucleotide substitution responsible for the slow acetylator phenotype. The *Nat2*8* allele does not contain this restriction site. As a consequence, following incubation with *Tru*9I (section 2.2.3.6), the PCR product generated from the *Nat2*9* allele is digested into two fragments, 113 and 20bp in size, while the product amplified from the *Nat2*8* allele remains undigested.

Following treatment with *Tru*9I, the amplification product from all strains remained undigested, except that generated from the A/J strain which was used as positive control for the *Nat2*9* allele (figure 3.2.1). This confirmed that none of the examined strains carried the $A^{296}\rightarrow T$ polymorphism, characteristic of the *Nat2*9* allele.

Table 3.2.1: Primers used for PCR amplification and sequencing in Chapter 3. The nucleotide sites of primer annealing are indicated relative to the beginning of the coding region of each *Nat* gene, using adenosine of the ATG codon as position 1.

Primer name	Orientation	Sequence (5′→3′)	Tm (°C)	Specificity
mNAT2-278[1]	forward	277 GTCTTTAACACTCCAGCC 294	54	Mouse *Nat2*
mNAT2-409[1]	reverse	409 ATTCCAGAGGCTCCCAC 393	54	
Mus12[2]	forward	-14 TGCCTTAGGGACATATGGACAT 8	61.2	Mouse *Nat1*, *Nat2* & *Nat3*
Mus13[2]	reverse	894 AGAAGAATTCTGCTCCTTACCC 873	61.2	Mouse *Nat1*
Mus15[2]	reverse	893 AGATCGGATCCCCTTATTACTC 872	61.2	Mouse *Nat3*
mNAT2-1[3]	forward	1 ATGGACATCGAAGCGTACTTTG 22	61.2	Mouse *Nat2*
mNAT2-910[3]	reverse	910 TTCCAAGTACATGGAAGGACACC 888	63.5	
mNAT2-691[4]	reverse	691 ACTCCTTCTGGGGTCTGCA 673	61.4	

[1]*Payton et al., 1999b;* [2]*Kelly and Sim, 1994;* [3]*Fakis et al., 2000;* [4]*This study*

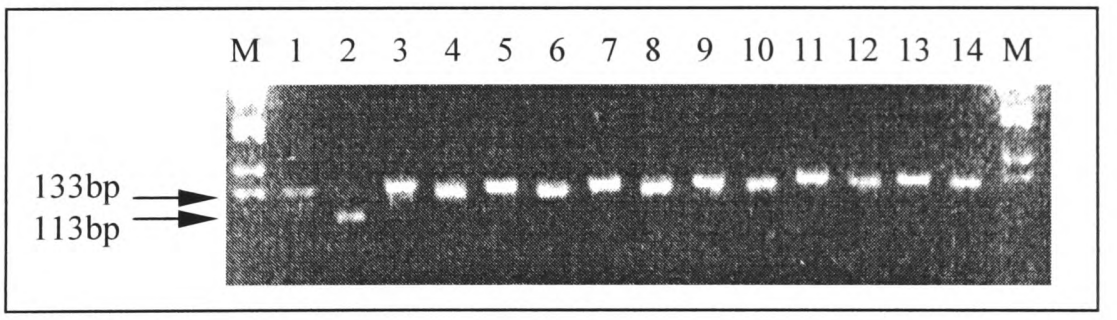


Figure 3.2.1: Genotyping of mouse strains for the *Nat2*8* and *Nat2*9* alleles. Genomic DNA (500ng) from liver tissue was used as template for amplification with primers mNAT2-278 and mNAT2-409 (table. 3.2.1). Each PCR product (10µl) was subsequently incubated with *Tru*9I at 65°C for 3h. Digests were separated on a 3% (w/v) Metaphor/ 1% (w/v) agarose gel, alongside the corresponding undigested PCR products. Lanes 1, 3, 5, 7, 9, 11 and 13 were loaded with undigested PCR product from strains A/J, CBA, 129/Ola, PO, TO, C57Bl/6J and Balb/c, respectively. Lanes 2, 4, 6, 8, 10, 12 and 14 were loaded with the same PCR products, after digestion with *Tru*9I. Lanes M were loaded with 1µg of 1kb DNA ladder. Strains CBA, 129/Ola, PO, TO, C57Bl/6J and Balb/c provided the same 133bp band before and after digestion, suggesting that they are all homozygous for the *Nat2*8* allele. The amplification product from A/J strain was digested as expected, producing a band 113bp in size (the smaller 20bp band is not visible on the gel) indicative of the *Nat2*9* allele.

3.2.2 Screening for polymorphisms in the coding region of mouse *Nat1*, *Nat2* and *Nat3* genes

To search for novel polymorphisms in the murine *Nat* genes, genomic DNA from 129/Ola, CBA, PO, TO, MCA and MSP strains was used as template for PCR amplification of the entire coding region of each *Nat* gene (figure 3.2.2). Forward primer Mus12 was used with reverse primers Mus13, mNAT2-910 or Mus15, to amplify the coding region of the *Nat1*, *Nat2* and *Nat3* genes, respectively (table 3.2.1). The PCR products were subsequently gel-purified (section 2.2.1.3) and sequenced (section 2.2.3.5) from both directions with the amplification primers, apart from the *Nat2* products which were sequenced with the two reverse primers mNAT2-691 and mNAT2-910 (table 3.2.1). The obtained sequences were aligned to the consensus sequences available for the C57Bl/6J (Martell *et al.*, 1991; Estrada-Rodgers *et al.*, 1998b) and Balb/c (Kelly and Sim, 1994) strains, for comparison. Detected polymorphisms were confirmed by sequencing the amplification product from a second individual of the same strain.

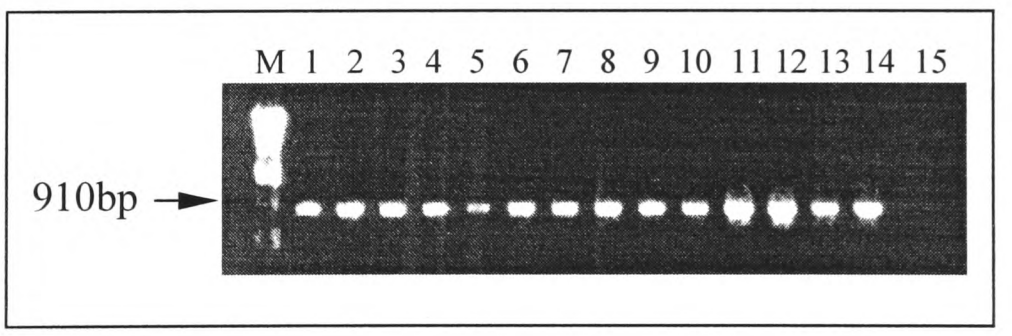


Figure 3.2.2: Amplification of the *Nat2* coding region from mouse strains for sequencing analysis. Genomic DNA (500ng) from strains 129/Ola, CBA, PO, TO, MCA and MSP was used as template for PCR amplification of the entire coding region of the *Nat2* gene, using primers Mus12 and mNAT2-910 (table 3.2.1). The gel was loaded with 10μl of PCR product from 129/Ola (lanes 1 and 2), CBA (lanes 3 and 4), PO (lanes 5 and 6) and TO (lanes 4 and 8) strains. Lanes 9, 10 and 11 were loaded with PCR product from three MCA individuals, while lanes 12, 13 and 14 with products from three MSP individuals. Lane 15 is the PCR negative control (no DNA template), while lane M was loaded with 1μg of 1kb DNA ladder.

Polymorphisms were not detected in any of the *Nat* genes of 129/Ola, CBA, PO and TO mouse strains (figures 3.2.3a, 3.2.4a and 3.2.5a). The *Nat1* gene of MCA does not contain any polymorphisms either (figure 3.2.3a), but the *Nat2* gene contains three conservative SNPs at positions 537, 747 and 834 of the coding region (figure 3.2.4b). The *Nat3* gene of MCA contains three non-conservative SNPs at positions 238, 607 and 608 of the coding region, resulting in two amino acid substitutions at positions 80 (Ala→Thr) and 203 (Trp→Gln) of the NAT3 polypeptide chain (figure 3.2.5b). The *Nat1* gene of MSP was found to contain four conservative SNPs at nucleotide positions 642, 684, 699 and 795, as well as one non-conservative SNP at position 695, which results in a His→Arg amino acid substitution at position 232 of the NAT1 protein (figure 3.2.3b). The *Nat2* gene of MSP contains five conservative SNPs at nucleotide positions 117, 480, 537, 690 and 807 of the coding region, as well as two non-conservative SNPs at nucleotide positions 78 and 244, which lead to amino acid changes at positions 26 (Glu→Asp) and 82 (Leu→Met) of the NAT2 protein (figure 3.2.4c). The *Nat3* gene of MSP contains eleven SNPs, of which only two (nucleotide positions 147 and 162) are conservative. The rest (nucleotide positions 238, 295, 413, 511, 607, 608, 638, 796 and 857) are non-conservative and lead to a total of eight amino acid changes in the NAT3 polypeptide chain (Ala^{80}→Thr, Cys^{99}→Arg, Thr^{138}→Ile, Ser^{171}→Pro, Trp^{203}→Gln, Arg^{213}→Gln, Val^{266}→Ile and Val^{286}→Ala) (figure 3.2.5c).

The sequencing data generated for the *Nat* genes of 129/Ola, CBA, PO, TO, MCA and MSP strains have been submitted to the EMBL Database and assigned the accession numbers shown in table 3.2.2. The identified *Nat* alleles were named according to the guidelines provided by the NAT Nomenclature Committee (www.louisville.edu/medschool/pharmacology/NAT.html; Vatsis *et al.*, 1995; Hein *et al.*, 2000b). The names of the identified alleles and the corresponding protein products are provided in table 3.2.2.

Figure 3.2.3: Screening for polymorphisms in the coding region of the mouse *Nat1* gene. Sequencing of the entire coding region of *Nat1* gene from 129/Ola, CBA, PO, TO, MCA and MSP mouse strains allowed comparison with the sequence previously published for the C57Bl/6J and Balb/c strains (Martell *et al.*, 1991; Kelly and Sim, 1994; Estrada-Rodgers *et al.*, 1998b). Each blue horizontal line represents the coding region of the *Nat1* gene, with position 1 corresponding to the adenosine of the initiation codon and position 873 to the last nucleotide of the stop codon. The type and position of each SNP and the encoded amino acid are indicated relative to the consensus sequence. Amino acids are coloured blue if polymorphism is conservative and red if non-conservative. The sequence of *Nat1* coding region from 129/Ola, CBA, PO, TO and MCA is identical to the consensus (a), while the sequence of MSP contains a number of SNPs (b).

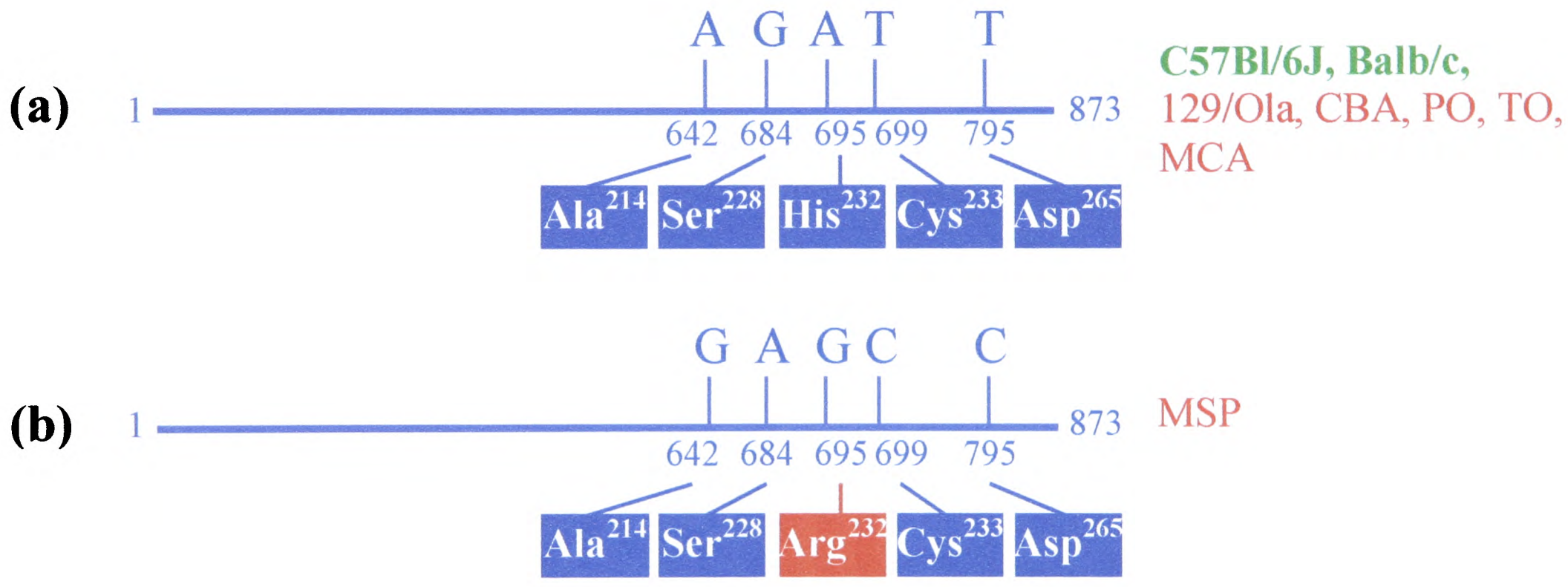


Figure 3.2.4: Screening for polymorphisms in the coding region of the mouse *Nat2* gene. The sequencing data are represented as described for the *Nat1* gene in figure 3.2.3. The sequence of *Nat2* coding region from 129/Ola, CBA, PO and TO strains is identical to the consensus (a), while the sequence of MCA (b) and MSP (c) strains, each contains a number of SNPs.

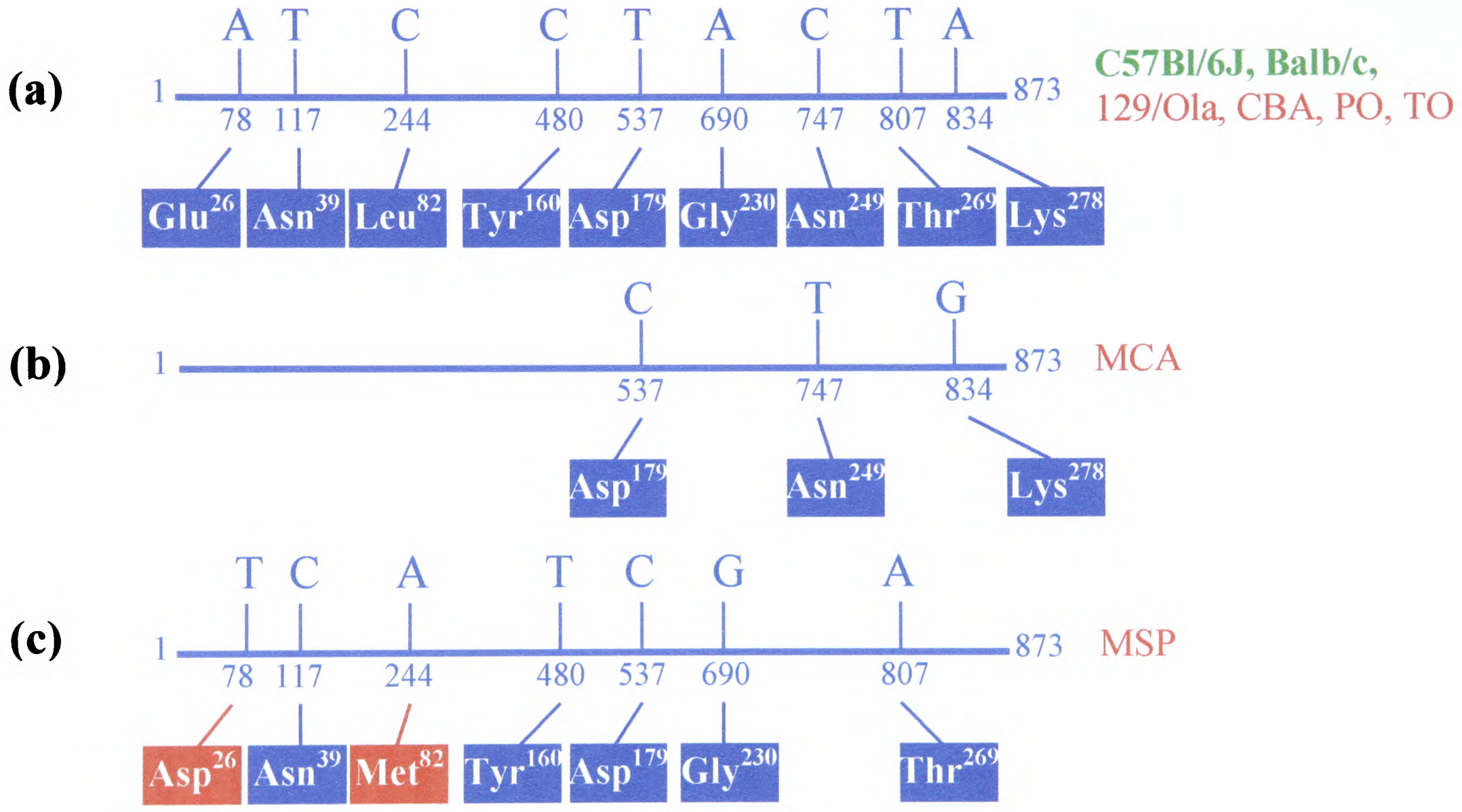


Figure 3.2.5: Screening for polymorphisms in the coding region of the mouse *Nat3* gene. The sequencing data are represented as described for the *Nat1* gene in figure 3.2.3. The sequence of *Nat3* coding region from 129/Ola, CBA, PO and TO strains is identical to the consensus (a), while the sequence of MCA (b) and MSP (c) strains, each contains a number of SNPs.

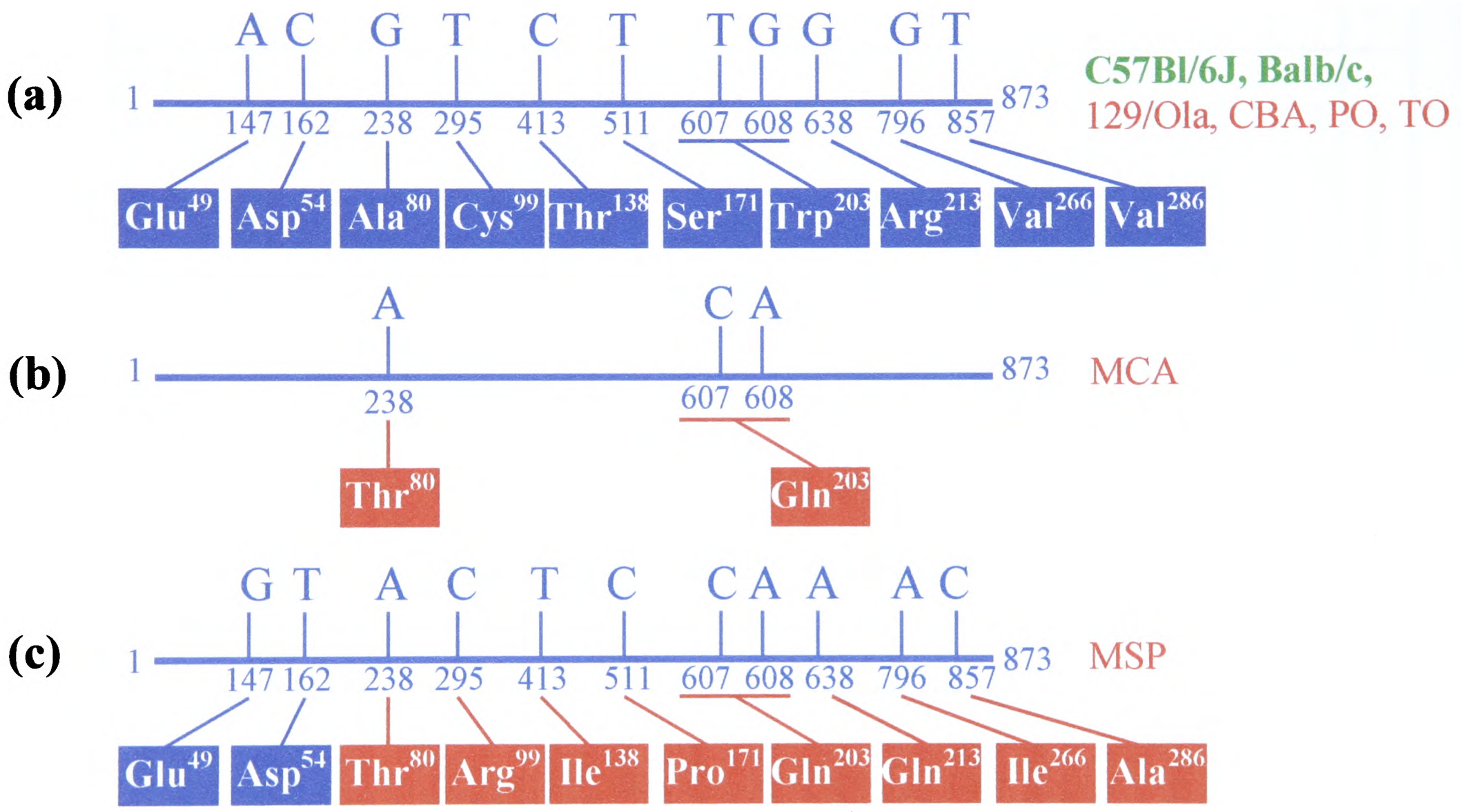


Table 3.2.2: Summary of the *Nat* gene sequencing data submitted to the EMBL database. A description for each entry and the corresponding accession number are provided, together with the symbol designated for each allele and its protein product. The novel alleles and the corresponding protein variants were assigned new symbols.

Accession no.	Description of database entry	Allele	Protein
AJ314650	*M. musculus Nat1* gene for arylamine N-acetyltransferase 1, strain 129/Ola.	*Nat1*6*[2]	NAT1 6
AJ314651	*M. musculus Nat1* gene for arylamine N-acetyltransferase 1, strain CBA.	*Nat1*6*	NAT1 6
AJ314652	*M. musculus Nat1* gene for arylamine N-acetyltransferase 1, strain PO.	*Nat1*6*	NAT1 6
AJ314653	*M. musculus Nat1* gene for arylamine N-acetyltransferase 1, strain TO.	*Nat1*6*	NAT1 6
AJ314654	*M.m. castaneus Nat1* gene for arylamine N-acetyltransferase 1.	*Nat1*6*	NAT1 6
AJ314655	*M. spretus Nat1* gene for arylamine N-acetyltransferase 1.	*Nat1*30*	NAT1 30
AJ250123[1]	*M. musculus Nat2* gene for arylamine N-acetyltransferase 2, strain 129/Ola.	*Nat2*8*[2]	NAT2 8
AJ314656	*M. musculus Nat2* gene for arylamine N-acetyltransferase 2, strain CBA.	*Nat2*8*	NAT2 8
AJ314657	*M. musculus Nat2* gene for arylamine N-acetyltransferase 2, strain PO.	*Nat2*8*	NAT2 8
AJ314658	*M. musculus Nat2* gene for arylamine N-acetyltransferase 2, strain TO.	*Nat2*8*	NAT2 8
AJ314659	*M.m. castaneus Nat2* gene for arylamine N-acetyltransferase 2.	*Nat2*22*	NAT2 22
AJ314660	*M. spretus Nat2* gene for arylamine N-acetyltransferase 2.	*Nat2*23*	NAT2 23
AJ314661	*M. musculus Nat3* gene for arylamine N-acetyltransferase 3, strain 129/Ola.	*Nat3*1*[2]	NAT3 1
AJ314662	*M. musculus Nat3* gene for arylamine N-acetyltransferase 3, strain CBA.	*Nat3*1*	NAT3 1
AJ314663	*M. musculus Nat3* gene for arylamine N-acetyltransferase 3, strain PO.	*Nat3*1*	NAT3 1
AJ314664	*M. musculus Nat3* gene for arylamine N-acetyltransferase 3, strain TO.	*Nat3*1*	NAT3 1
AJ314665	*M.m. castaneus Nat3* gene for arylamine N-acetyltransferase 3.	*Nat3*2*	NAT3 2
AJ314666	*M. spretus Nat3* gene for arylamine N-acetyltransferase 3.	*Nat3*3*	NAT3 3

[1]*This database entry describes a 14.3kb plasmid clone of 129/Ola mouse strain origin (clone A), previously identified to contain the entire coding region of the Nat2 gene (Fakis et al., 2000), and subsequently sequenced as will be described in Chapter 4.*
[2]*Nat1*6, Nat2*8 and Nat3*1 are the Nat1, Nat2 and Nat3 alleles, respectively, of C57Bl/6J and Balb/c mouse strains (Vatsis et al., 1995).*

3.2.3 NAT enzymatic activities in different mouse strains

To assess the phenotypic effect of the polymorphisms described in the previous section (3.2.2), NAT enzymatic activity was measured in liver homogenates (section 2.3.2) from five MCA and five MSP mice, in comparison with four fast acetylating Balb/c and five slow acetylating A/J mice, matched for age and sex. A freshly thawed aliquot of liver homogenate from another Balb/c mouse was used as a standard for comparison between different sets of assays.

Two arylamine substrates were used to measure NAT activity in liver homogenates, namely p-anisidine (pANIS), which is N-acetylated by both NAT1 and NAT2 isoenzymes (Kelly and Sim, 1994), and p-aminobenzoic acid (pABA), which is a probe substrate for mouse NAT2 activity (Martell *et al.*, 1992). The approximate concentration of protein in each reaction was 1.5µg/µl with pANIS and 7.5µg/µl with pABA. Each liver homogenate was assayed for 0, 5, 10, 15 and 20min, in order to determine the time period during which the substrate conversion rate was linear (section 2.3.5.2). The results were expressed as pmoles of N-acetylated substrate per minute per mg of total protein (specific activity), relative to the specific activity of the standard sample (table 3.2.3). The interindividual variation in NAT specific activity for each strain is shown in figure 3.2.6. The mean specific activities calculated for the four analysed strains are provided in table 3.2.4 and in figure 3.2.7.

Table 3.2.3: NAT specific activity of mouse liver homogenates. Specific activity (pmoles of N-acetylated substrate/min/mg of total protein) in fresh liver homogenates prepared from MSP, MCA, A/J and Balb/c mice was determined relative to the specific activity of a standard sample included in each set of assays.

Sample name	Specific activity with pANIS	Specific activity with pABA
MSP-1	234	97
MSP-2	127	73
MSP-3	156	103
MSP-4	164	79
MSP-5	175	123
MCA-1	439	168
MCA-2	500	232
MCA-3	448	269
MCA-4	433	121
MCA-5	199	214
A/J-1	165	77
A/J-2	151	94
A/J-3	93	84
A/J-4	150	88
A/J-5	69	31
Balb/c-1	516	229
Balb/c-2	300	256
Balb/c-3	512	191
Balb/c-4	340	185
Standard[1]	**134.5**	**139.3**

[1]*Mean specific activity for the four sets of assays*

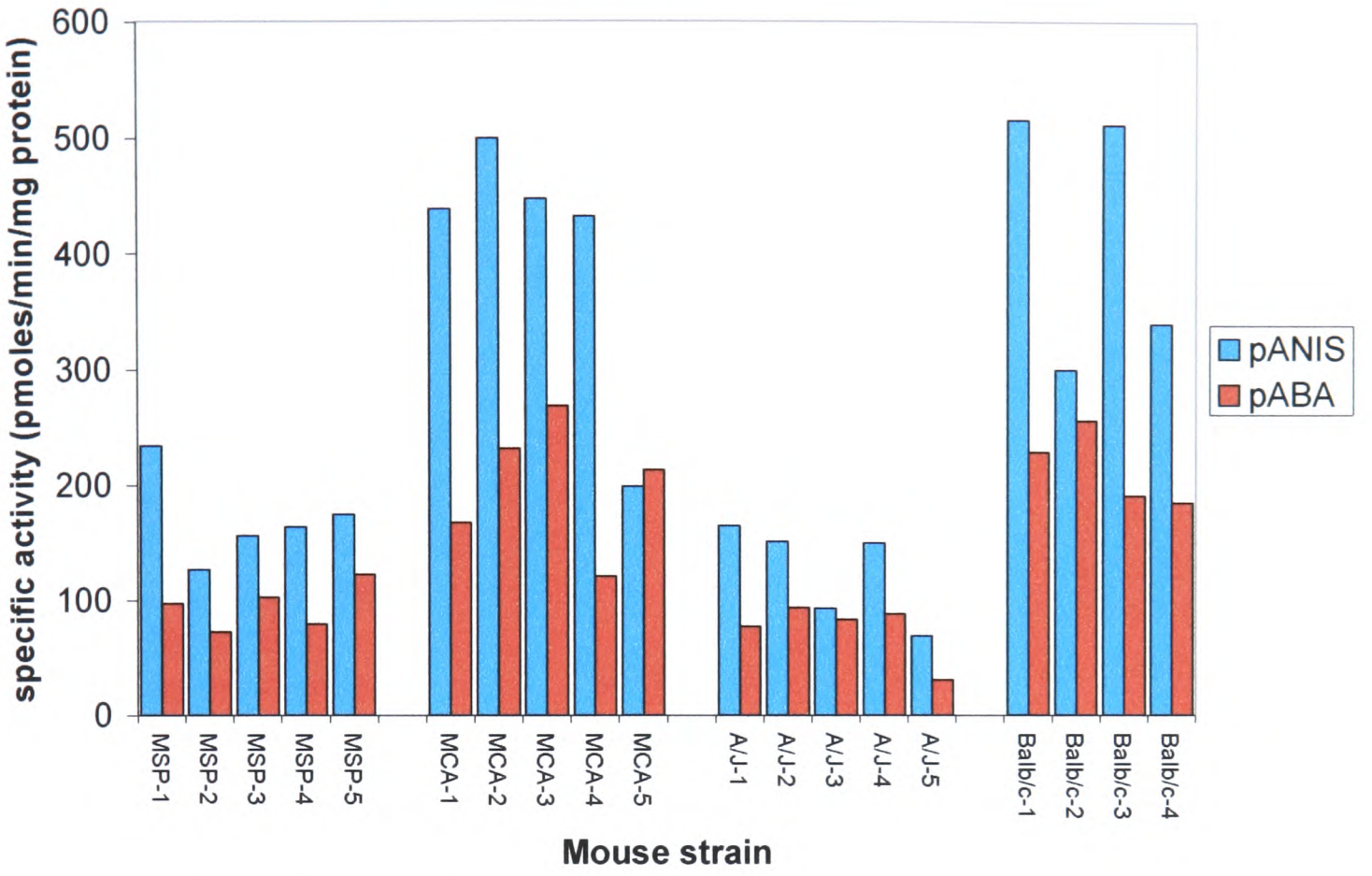
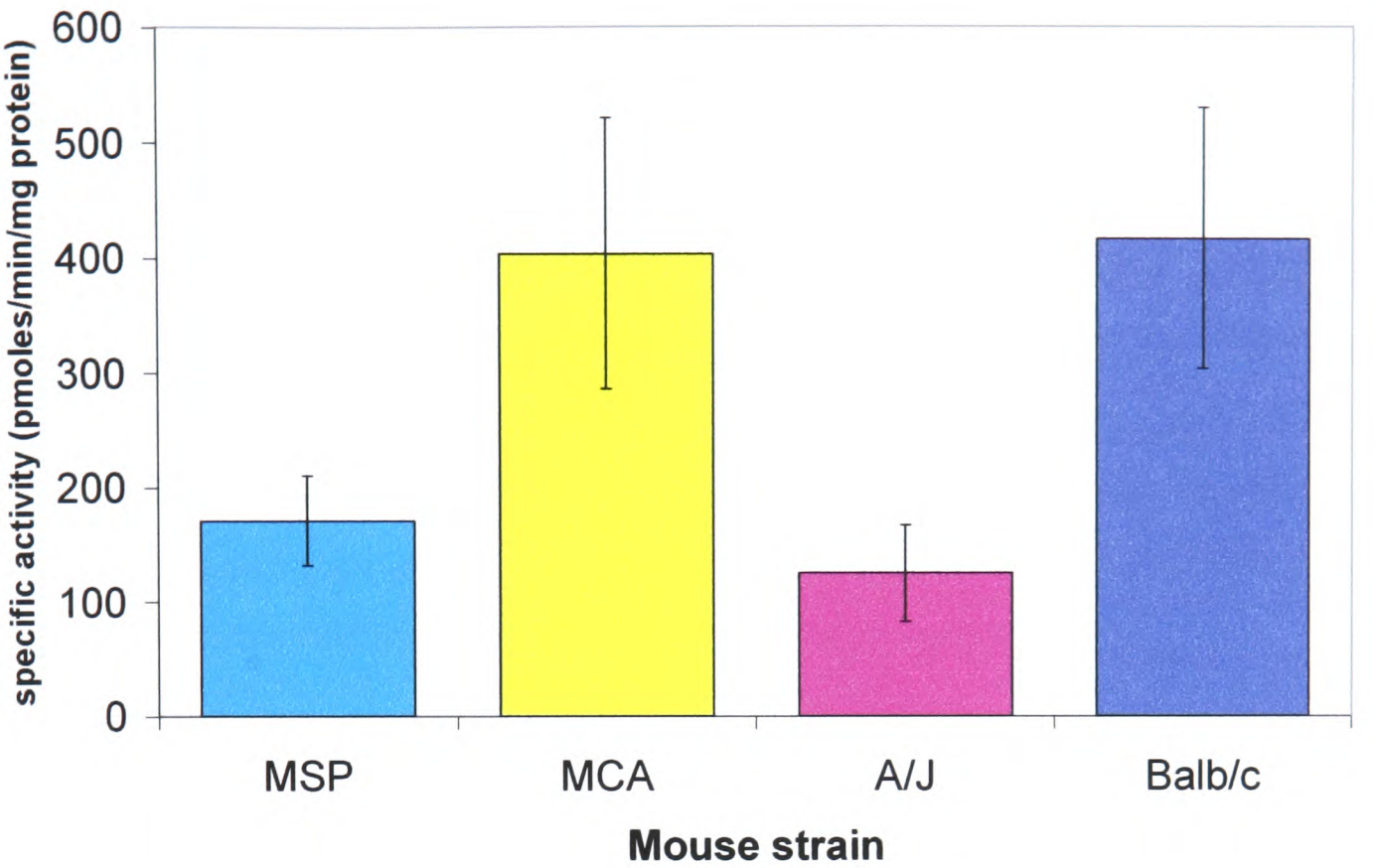


Figure 3.2.6: Variation in NAT specific activity between individual mice of the same strain. The combined NAT1 and NAT2 specific activity measured in liver homogenates with pANIS (table 3.2.3) varied 1.8-fold between MSP, 2.5-fold between MCA, 2.4-fold between A/J and 1.7-fold between Balb/c mice. The NAT2 specific activity measured with pABA (table 3.2.3) varied 1.7-fold between MSP, 2.2-fold between MCA, 3.0-fold between A/J and 1.4-fold between Balb/c mice.

Table 3.2.4: Mean specific activity of NAT isoenzymes in mouse liver homogenates. The data in table 3.2.3 were used to determine the mean specific activity with pANIS and pABA substrates for each strain. The corresponding standard deviations are also provided.

Mouse strain (n)	**Mean specific activity ± standard deviation (pmoles/min/mg of protein)**	
	pANIS	**pABA**
MSP (5)	171 ± 39	95 ± 20
MCA (5)	404 ± 117	201 ± 57
A/J (5)	126 ± 42	75 ± 25
Balb/c (4)	417 ± 113	215 ± 33



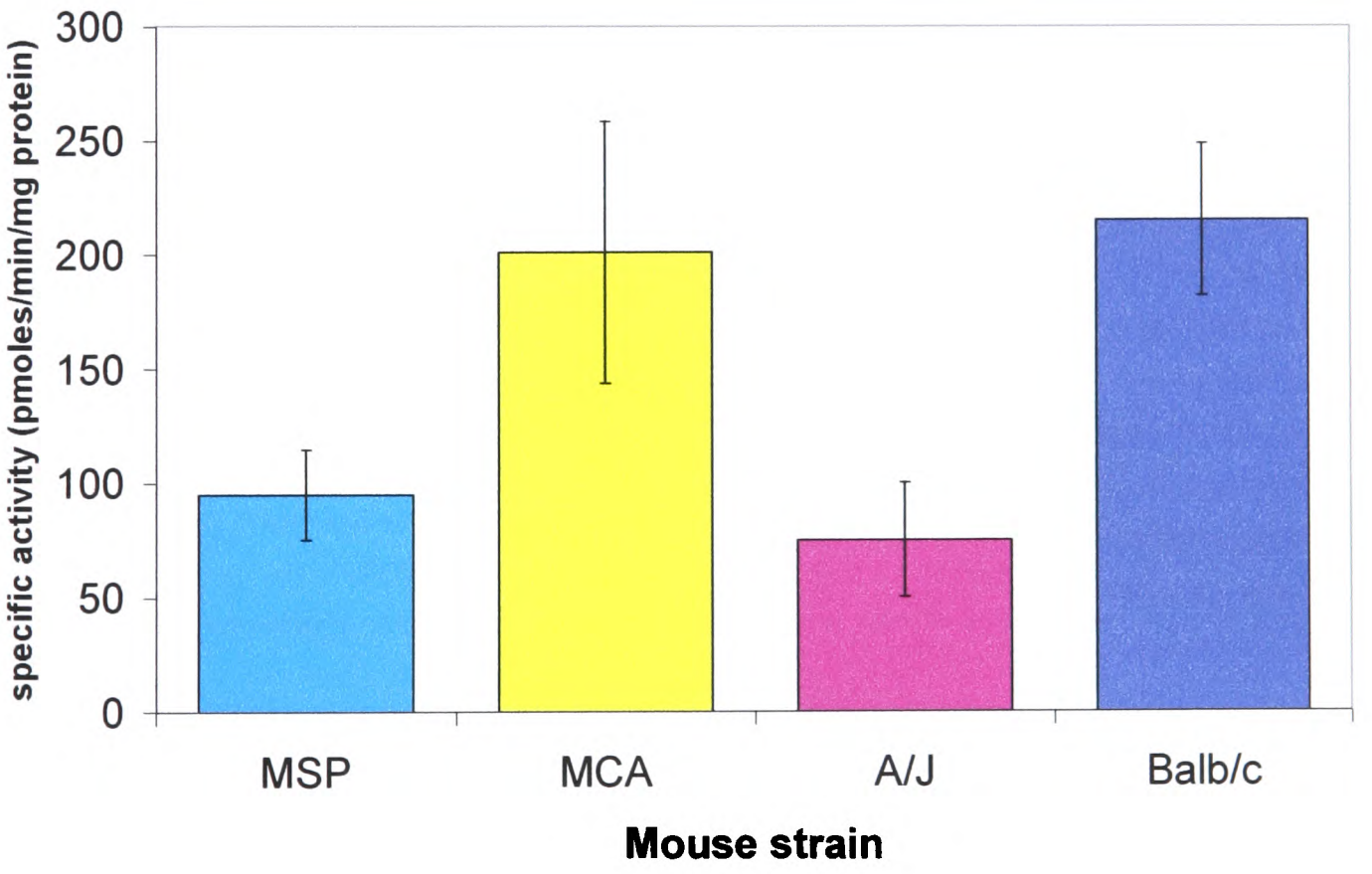


Figure 3.2.7: Comparison of hepatic NAT activities between mouse strains. a) Combined NAT1 and NAT2 mean specific activity with pANIS as substrate. b) NAT2 mean specific activity with pABA as substrate. The error bars are indicative of the interindividual variation from the calculated mean within each strain.

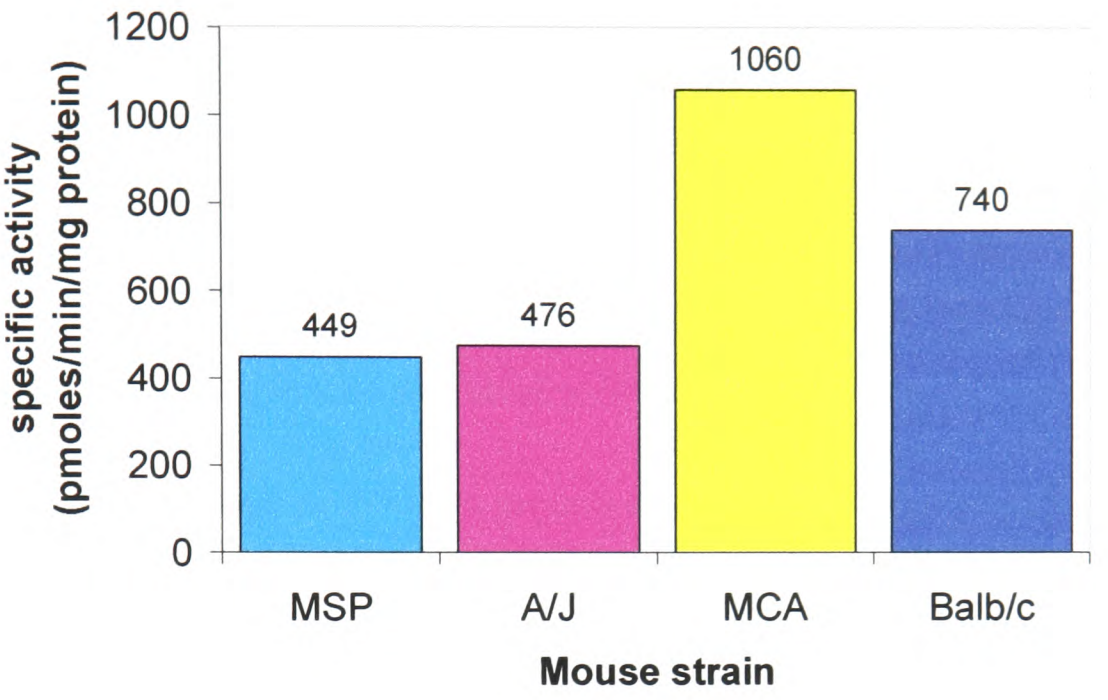
In order to investigate whether the difference in mean specific activity between the examined strains was statistically significant, a two-sided *t*-test statistical analysis was performed. The test was adapted for use with small samples (n=5 or n=4) with unequal standard deviations (more than 2-fold different) and required square root transformation of the data (Kirkwood, 1988). The results of the analysis, summarised in table 3.2.5, were similar for both pANIS and pABA substrates. The NAT mean specific activity of MSP was significantly lower than that of Balb/c and MCA, but similar to that of the A/J mouse. The NAT mean specific activity of MCA was significantly higher than that of A/J and MSP, and similar to that of the Balb/c mouse. The results indicate that MCA is a fast and MSP a slow acetylating strain.

Table 3.2.5: Statistical analysis (*t*-test) for comparison of NAT mean specific activity between mouse strains. The p value calculated for each pair of data is presented. Values coloured blue and green were calculated using data generated from pANIS and pABA assays, respectively. The difference in each pair of means was considered statistically significant if $p<0.05$.

	MSP	A/J	MCA	Balb/c
MSP		$p>0.1$	$p<0.005$	$p<0.002$
A/J	$p>0.2$		$p<0.001$	$p<0.001$
MCA	$p<0.005$	$p<0.002$		$p>0.5$
Balb/c	$p<0.002$	$p<0.001$	$p>0.5$	

A similar acetylation pattern was observed when pooled liver homogenates were assayed with pANIS and pABA (figure 3.2.8). The pooled homogenates were prepared by mixing equal volumes of liver homogenate from all individuals of the same strain. The NAT specific activity determined for each pool is presented in figure 3.2.8.

(a)



(b)

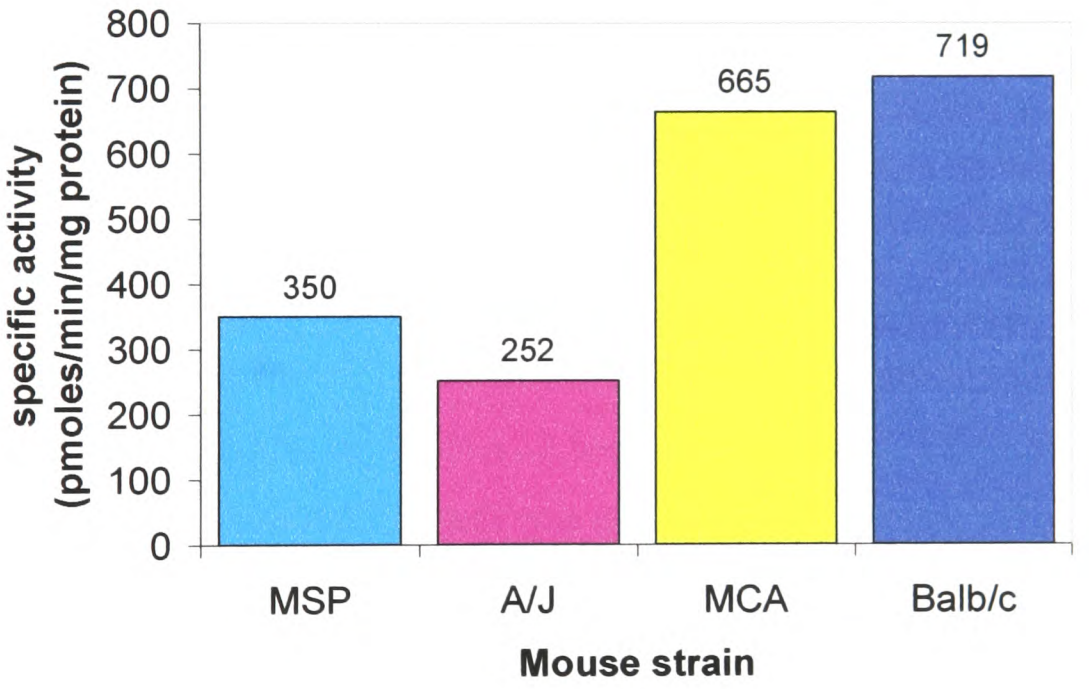


Figure 3.2.8: Comparison of NAT specific activity in pooled liver homogenates between mouse strains. a) Combined NAT1 and NAT2 specific activity with pANIS as substrate. b) NAT2 specific activity with pABA as substrate. Pooled liver homogenates were prepared from five MSP, five MCA, five A/J and four Balb/c individuals (figure 3.2.6).

3.2.4 Detection of NAT2 protein in different mouse strains by Western blot analysis

To assess whether the difference in hepatic NAT2 activity between the fast and the slow acetylating mouse strains was due to lower amounts of NAT2 protein or to decreased catalytic activity of the enzyme, the pooled liver homogenates of MSP, MCA and the control Balb/c and A/J strains were subjected to Western blot analysis with NAT2-specific antiserum 184 (Stanley *et al.*, 1997).

The concentration of the total protein in each one of the pooled homogenates was determined by the Bradford assay (section 2.3.1). The same amount (100 or 200μg) of total protein from each sample was subjected to SDS-PAGE (section 2.3.3), following addition of 2.8μg of pure BSA protein to each sample, to serve as an internal standard. Western blot analysis was carried out with antiserum 184, which was raised against a BSA-conjugated NAT2 peptide and is, therefore, capable of binding to the BSA added to the samples (section 2.3.4)

An example of semi-quantitative Western blot is shown in figure 3.2.9. Five independent experiments of this type were performed on three different occasions. The films were subjected to densitometric analysis (section 2.3.4), allowing comparison of the intensities of the NAT2 bands from Balb/c, MSP, MCA and A/J strains. The density of each NAT2 band was initially modified according to the BSA standards, to take into account variation due to small differences in sample handling, and expressed relative (%) to the mean density of all NAT2 bands on the same blot (relative density). The results of the five independent experiments were used to calculate the mean relative density of the NAT2 bands for each strain (table 3.2.6 and figure 3.2.10).

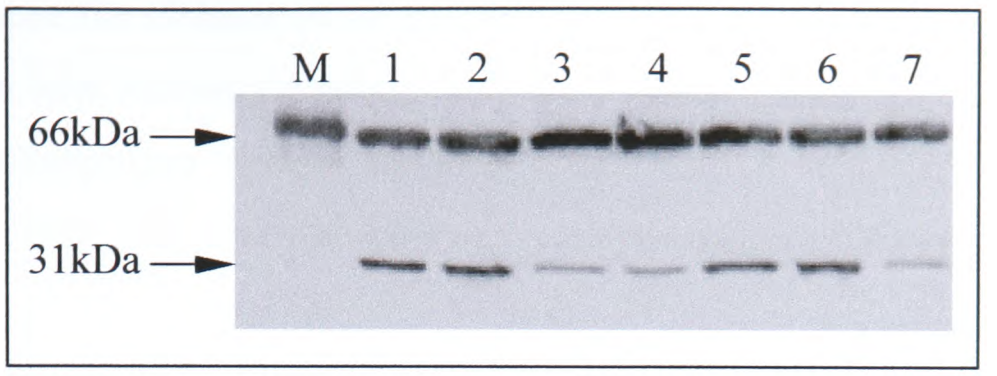


Figure 3.2.9: Semi-quantitative Western blot analysis of pooled liver homogenates. Western blot analysis (section 2.3.4) of a 12% (w/v) SDS-PAGE gel (section 2.3.3), loaded with pooled liver homogenate from Balb/c (lanes 1, 2), MSP (lanes 3, 4), MCA (lanes 5, 6) and A/J (lane 7) mice. To each sample (200µg of hepatic protein), 2.8µg of pure BSA protein were added, as an internal standard. Lane M was loaded with approximately 5µg of Rainbow markers. Antiserum 184 (section 2.3.4) was used to detect NAT2 at 31kDa and BSA at 66kDa molecular weight. The BSA contained in the markers was also detected.

Table 3.2.6: Densitometric analysis of semi-quantitative Western blots. The mean relative density of the NAT2 bands and the corresponding standard deviations were calculated for the five independent measurements taken for each strain.

Mouse strain	Mean relative density ± standard deviation (%)
Balb/c	181±27
MSP	50±24
MCA	138±21
A/J	31±11

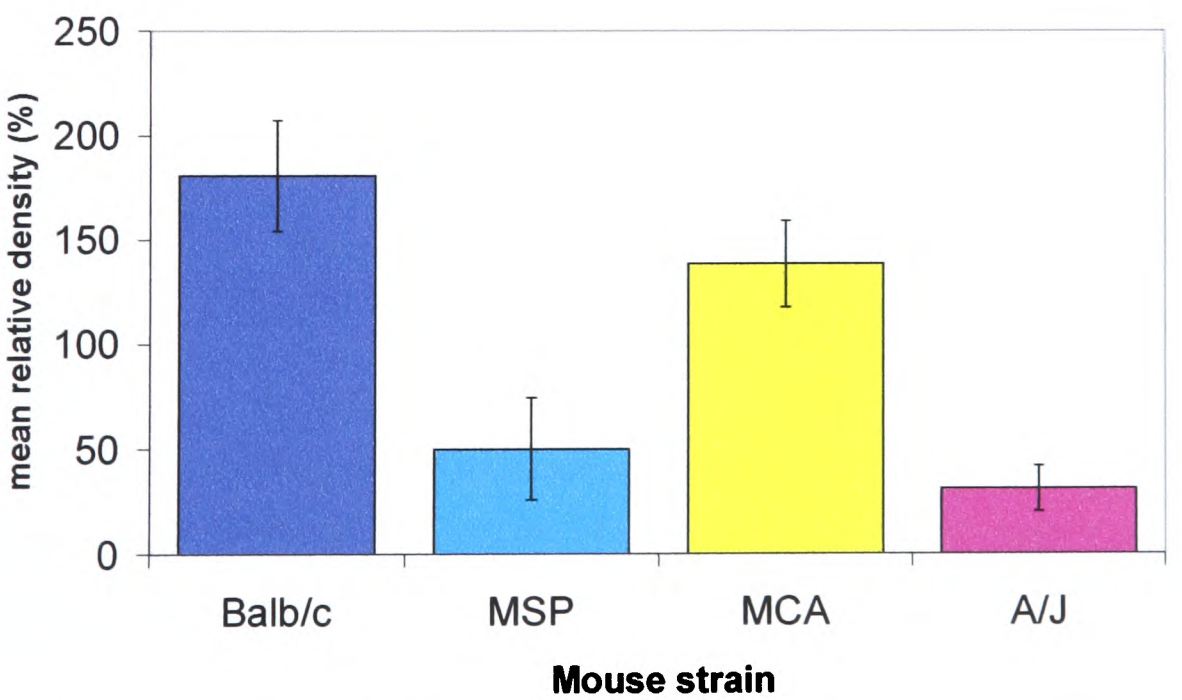


Figure 3.2.10: Comparison of the mean relative density of NAT2 bands between mouse strains. The mean relative density of the MSP NAT2 bands is 3.6- and 2.7-fold lower than that of Balb/c and MCA, respectively. The mean relative density of the A/J NAT2 bands is 5.8- and 4.5-fold lower than that of Balb/c and MCA, respectively. The mean relative densities of MSP and A/J NAT2 bands are only 1.6-fold different, while those of Balb/c and MCA are only 1.3-fold different.

To ensure the absence of an endogenous BSA-like protein in the liver, that might interfere with antibody binding, a 12% (w/v) SDS-PAGE gel was loaded with pooled liver homogenates, without additional BSA, and subjected to Western blotting (figure 3.2.11). No endogenous BSA-like protein was detected with antiserum 184.

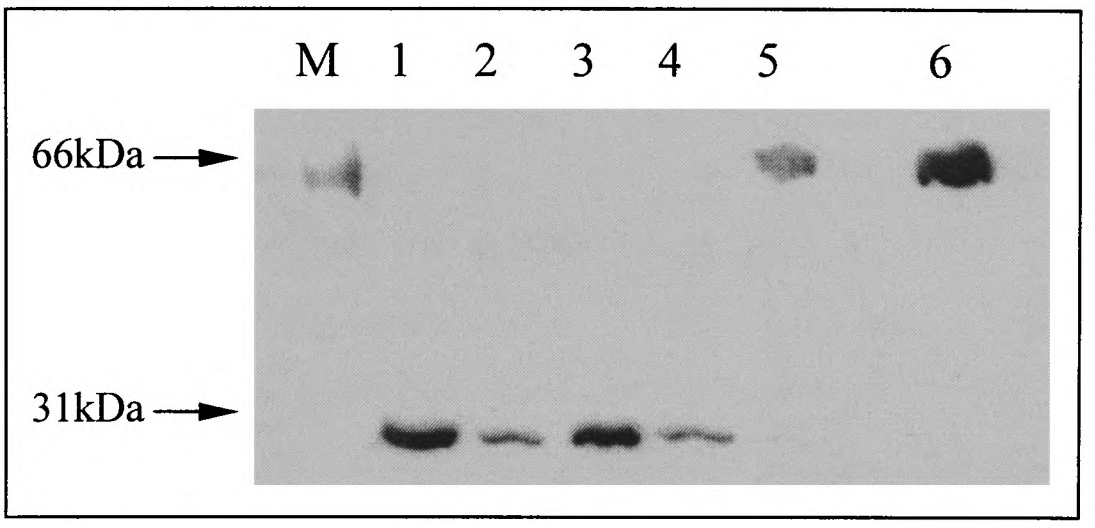


Figure 3.2.11: Western blot demonstrating absence of endogenous BSA in the liver. Pooled liver homogenate (200μg of total protein) from Balb/c, MSP, MCA and A/J strains (lanes 1, 2, 3 and 4, respectively) was subjected to Western blot analysis with antiserum 184, without prior addition of BSA. Pure BSA protein was loaded in lanes 5 (2.8μg) and 6 (5.6μg), as positive control. Lane M was loaded with approximately 5μg of Rainbow markers, containing BSA. The 66kDa BSA band was visible in the marker lane, as well as in the lanes of the positive controls. No BSA was detected in any of the pooled liver homogenates, while NAT2 was detected at 31kDa, as expected.

Statistical analysis of the results was performed using the two-sided *t*-test for small samples (n=5) with similar (up to approximately 2-fold different) standard deviations (Kirkwood, 1988). The difference in the amount of immunoreactive NAT2 protein was highly significant between Balb/c and MSP ($p<0.001$), as well as between Balb/c and A/J ($p<0.001$) strains. The difference between the Balb/c and MCA strains was not as pronounced, but still statistically significant ($0.05>p>0.02$). The difference between MCA and MSP, as well as between MCA and A/J was significant ($p<0.001$ in both cases), unlike MSP and A/J strains which did not differ ($0.5>p>0.1$) in their amounts of immunoreactive NAT2 protein. The results indicate that NAT2 is less abundant in the slow acetylating A/J and MSP strains, compared with the fast acetylating Balb/c and MCA strains, which may have up to six times more NAT2 protein in the liver (table 3.2.6).

3.2.5 Cloning of the *Nat* gene variants into the pTargeT mammalian expression vector

In order to study the enzymatic properties of NAT variants (allozymes) from different mouse strains, the corresponding genes were cloned into a mammalian expression vector and introduced to a suitable cell line for the production of recombinant protein. The coding region of *Nat1* from MCA and MSP, the coding region of *Nat2* from Balb/c, MCA, MSP and A/J and the coding region of *Nat3* from Balb/c, MCA and MSP strains was amplified from genomic DNA, using the high-fidelity *Pfu*-DNA polymerase (section 2.2.3.3). The primers used were Mus12/Mus13 for *Nat1*, mNAT2-1/mNAT2-910 for *Nat2* and Mus12/Mus15 for *Nat3* (table 3.2.1). The PCR products (figure 3.2.12) were gel-purified (section 2.2.1.3) and A-tailed (section 2.2.3.7), then ligated (section 2.2.3.9) into the T-tailed pTargeT vector (Appendix 1.1b). The ligation products were introduced to JM109 High Efficiency Competent Cells (*E. coli*) (section 2.2.3.10), which were then grown overnight on LB-agar plates supplemented with 100μg/ml ampicillin, 0.5mM IPTG and 40μg/ml x-gal. Five white colonies were selected from each plate and grown overnight in 10ml LB medium with 100μg/ml ampicillin. Plasmid DNA was extracted as described in section 2.2.1.2.

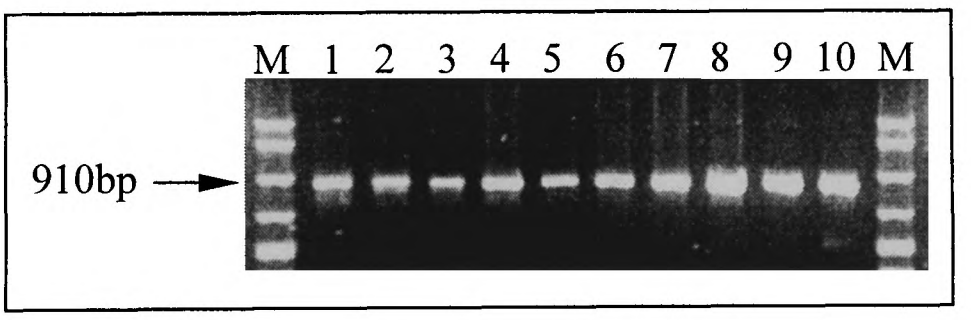


Figure 3.2.12: Amplification of *Nat* gene variants for cloning into the pTargeT vector. Genomic DNA (500ng) from MCA, MSP, Balb/c and A/J mouse strains was used to amplify the entire coding region of the *Nat* genes, using primers Mus12/Mus13 for *Nat1*, mNAT2-1/mNAT2-910 for *Nat2* and Mus12/Mus15 for *Nat3*. Following gel-purification (section 2.2.1.3), each PCR product (5μl) was subjected to agarose gel electrophoresis. Lanes 1 and 2 were loaded with *Nat1* amplification product from MSP. Lane 3 is the *Nat1* amplification product from MCA. Lanes 4, 5, 6 and 7 are the *Nat2* amplification products from Balb/c, MSP, MCA and A/J, respectively. Lanes 8, 9 and 10 are the *Nat3* amplification products from Balb/c, MSP and MCA, respectively. Lanes M were loaded with 5μl of BioMarker EXT ladder.

The identity and correct orientation of the insert in the pTargeT vector was verified by restriction digestion (section 2.2.3.6) of plasmid DNA isolated from the selected clones. Those clones expected to carry the coding region of *Nat1* were digested with *Nde*I, those carrying the coding region of *Nat2* with *Sal*I and *Bst*EII, and those carrying the coding region of *Nat3* with *Bam*HI. The expected digestion patterns are described in figure 3.2.13. The orientation of the insert was correct, if digestion confirmed that the initiation codon of the *Nat* coding region had ligated close to the CMV early enhancer/promoter of the pTargeT vector (Appendix 1.1b). *Nde*I cuts within primer Mus12 (Kelly and Sim, 1994) and at position 387 of the vector (Appendix 1.1b). Correct orientation of the *Nat1* insert would provide two fragments 910 and 5669bp in size, while the wrong orientation would provide fragments 1793 and 4786bp in size (figures 3.2.13 and 3.2.14a). *Bst*EII cuts at position 333 of the *Nat2* coding region (Martell *et al.*, 1991) and *Sal*I at position 1303 of the vector (Appendix 1.1b). Correct orientation of the *Nat2* insert would provide fragments 596 and 5984bp in size, while wrong orientation would provide fragments 352 and 6228bp in size (figures 3.2.13 and 3.2.14b). *Bam*HI cuts in primer Mus15 (Kelly and Sim, 1994) and at position 1256 of the vector (Appendix 1.1b). Correct orientation of the *Nat3* insert would provide fragments 926 and 5652bp in size, while wrong orientation would provide fragments 38 and 6540bp in size (figures 3.2.13 and 3.2.14c).

All clones identified to contain the insert in the correct orientation were sequenced with vector-specific primers T7-promoter and pTargeT (table 2.1). The generated data were compared with the sequences of the *Nat* alleles described in section 3.2.2. Positive clones without polymerase-introduced mutations were used in subsequent experiments. Cloning of the *Nat2* coding region from Balb/c was not successful. However, this was not a problem, as the protein products of Balb/c *Nat2* (*Nat2*8* allele) and MCA *Nat2* (*Nat2*22* allele) are the same, the *Nat2*22* allele containing only conservative SNPs (section 3.2.2). Moreover, A/J *Nat2* (*Nat2*9* allele) was also used as reference for comparison amongst the different NAT2 variants.

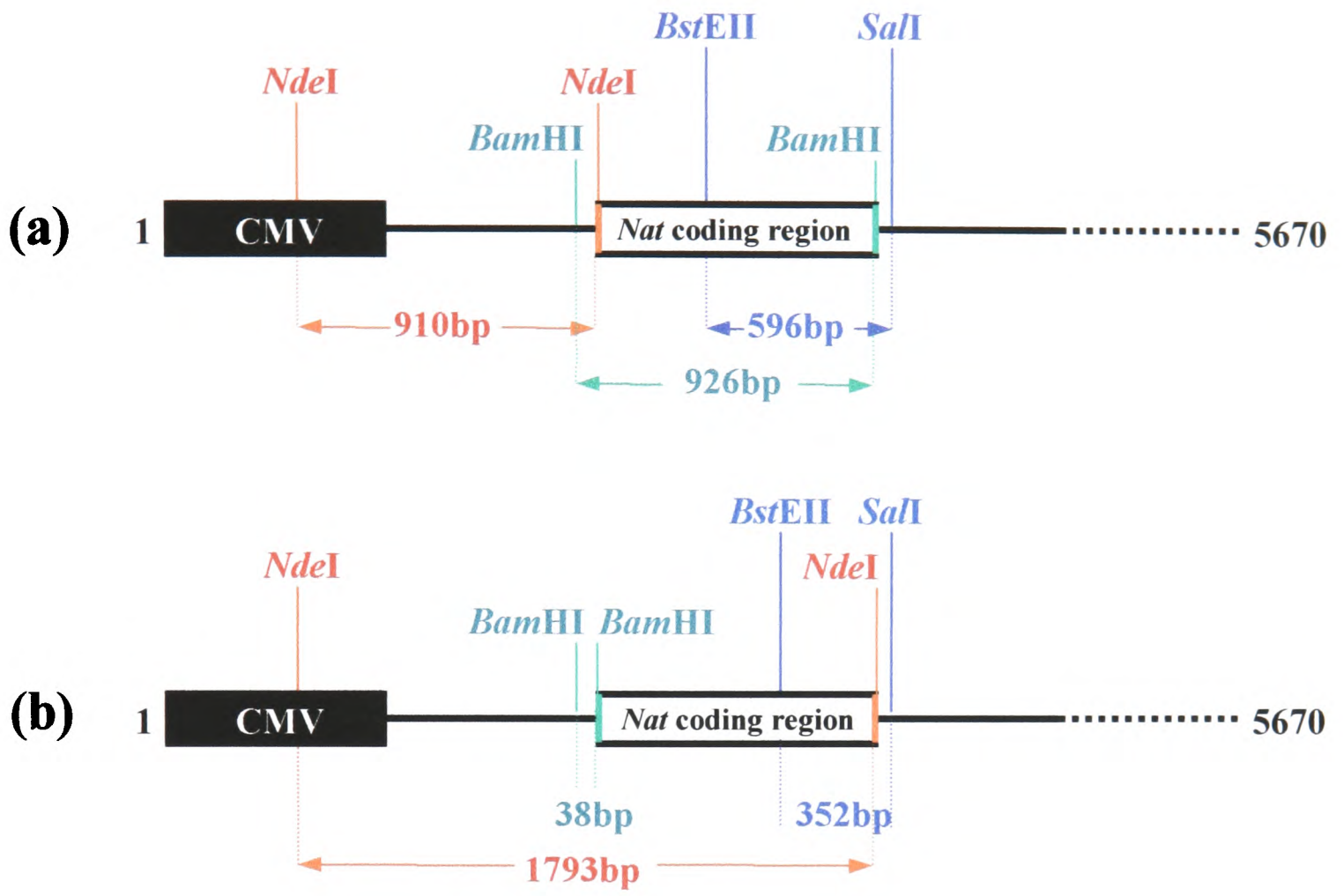


Figure 3.2.13: Strategy for determining the orientation of the *Nat* inserts in the pTargeT vector. The black horizontal line is the pTargeT vector, the black box is the CMV early enhancer/promoter (Appendix 1.1b) and the white box is the *Nat* insert. The restriction sites used to check the orientation of incorporation of the *Nat1*, *Nat2* and *Nat3* inserts, as well as the size of the corresponding fragments, are shown in red, blue and green, respectively. The forward primer Mus12, used to amplify the *Nat1* coding region, contains a *Nde*I site and its position is indicated in red. The reverse primer Mus15, used to amplify the *Nat3* coding region, contains a *Bam*HI site and its position is indicated in green. The expected restriction pattern is shown for the correct (a) and the wrong (b) orientation of insert incorporation.

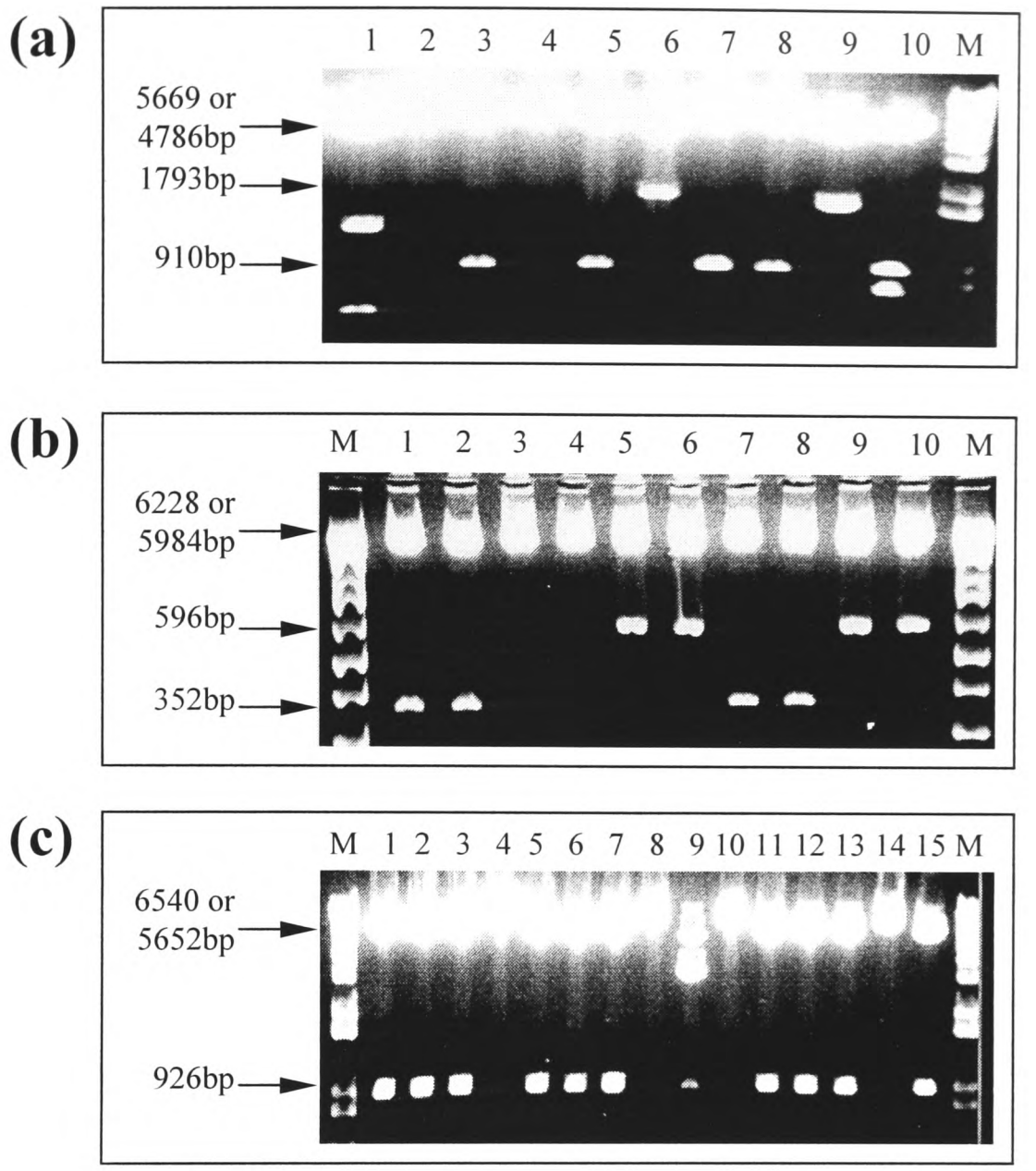


Figure 3.2.14: Checking for the orientation of the *Nat* inserts in the pTargeT vector. a) Digestion with *Nde*I of plasmid clones expected to carry the *Nat1* gene. The reactions took place at 37°C for 4h. Lanes 1 to 5 were loaded with digested clones of MSP origin and lanes 6 to 10 with digested clones of MCA origin. The 910bp band indicating correct orientation of the insert is present in lanes 3, 5 (MSP), 7 and 8 (MCA). b) Digestion with *Bst*EII and *Sal*I of plasmid clones expected to carry the *Nat2* gene. The reactions took place with *Sal*I alone at 37°C for 4h, then *Bst*EII was added and incubation was continued for 4h at 60°C. The products were separated on a 2.5% (w/v) agarose gel. Lanes 1 to 5 were loaded with digested clones of MCA origin and lanes 6 to10 with digested clones of A/J origin. The 596bp band indicating correct orientation of the insert is present in lanes 5 (MCA), 6, 9 and 10 (A/J). c) Digestion with *Bam*HI of plasmid clones expected to carry the *Nat3* gene. The reactions took place at 37°C for 4h. Lanes 1 to 5 were loaded with digested clones of Balb/c origin, lanes 6 to 10 with digested clones of MSP origin and lanes 11 to 15 with digested clones of MCA origin. The 926bp band indicating correct orientation of the insert is present in lanes 1, 2, 3, 5 (Balb/c), 6, 7 (MSP), 11, 12, 13 and 15 (MCA). All restriction digestions were carried out using approximately 1μg of plasmid DNA as template. Lanes M were loaded with 1μg of 1kb Plus DNA ladder.

3.2.6 Transfection of CHO cells with pTargeT constructs and preparation of cell lysates for analysis

Eight pTargeT clones (constructs) were selected for transfection of CHO cells. Constructs carrying the *Nat1*6* allele of MCA and the *Nat1*30* allele of MSP (table 3.2.2) were named MCA-*Nat1* and MSP-*Nat1*, respectively. Constructs carrying the *Nat2*22* allele of MCA, the *Nat2*23* allele of MSP (table 3.2.2) and the *Nat2*9* allele of A/J (Vatsis *et al.*, 1995) were named MCA-*Nat2*, MSP-*Nat2* and A/J-*Nat2*, respectively. Constructs carrying the *Nat3*1* allele of Balb/c (Vatsis *et al.*, 1995), the *Nat3*2* allele of MCA and the *Nat3*3* allele of MSP (table 3.2.2) were named Balb/c-*Nat3*, MCA-*Nat3* and MSP-*Nat3*, respectively.

Purified plasmid DNA from each construct (10μg) was co-transfected into CHO cells (section 2.2.4.2) with an equal amount of purified pGL3-Control luciferase reporter vector (Appendix 1.2d). A control co-transfection was performed with 10μg of pTargeT vector, lacking insert, and the same amount of pGL3-Control vector. Cells were also transfected with 10μg of either only pTargeT vector, lacking insert, or only pGL3-Control vector. Mock transfections were also performed (section 2.2.4.2). The transfected cells were harvested at confluency (section 2.2.4.2) and lysed by sonication (section 2.3.2).

The luminescence produced by the firefly luciferase expressed from the pGL3-Control vector was measured in each cell lysate as described in section 2.3.5.1. The luminometer readings are shown in table 3.2.7. The luminescence measured for all cultures should be similar, as the same amount of pGL3-Control vector was used in all co-transfections. Therefore, any variation observed in luciferase activity should reflect variation in the transfection efficiency. This variation was taken into account in subsequent experiments measuring NAT enzymatic activity in lysates of transfected CHO cells.

Table 3.2.7: Luciferase activity in lysates of transfected CHO cells. Cells were co-transfected with equal amounts (10μg) of each pTargeT contruct and the reporter vector pGL3-Control. Plates 9-12 are control transfections.

Plate no.	Construct/vector used for transfection	Luminescence (10^6)
1	MCA-*Nat1* + pGL3-Control	5.5
2	MSP-*Nat1* + pGL3-Control	7.2
3	MCA-*Nat2* + pGL3-Control	6.5
4	MSP-*Nat2* + pGL3-Control	7.5
5	A/J-*Nat2* + pGL3-Control	6.8
6	Balb/c-*Nat3* + pGL3-Control	2.8
7	MCA-*Nat3* + pGL3-Control	3.7
8	MSP-*Nat3* + pGL3-Control	4.8
9	pTargeT (no insert) + pGL3-Control	10.2
10	pGL3-Control	13.8
11	pTargeT (no insert)	<0.01
12	Mock transfected cells	<0.001

To investigate whether the pTargeT vector could drive expression of the *Nat* genes, lysates of CHO cells transfected with constructs MCA-*Nat2*, MSP-*Nat2* and A/J-*Nat2* were subjected to Western blot analysis with antiserum 184, as described in sections 2.3.3 and 2.3.4. CHO cells have been demonstrated to lack endogenous NAT2 protein both by enzymatic activity assay (section 3.2.7) and Western blotting with antiserum 184 (Nichola Johnson, D.Phil. Thesis, Oxford 2001). Lysates were shown to contain recombinant NAT2 protein (figure 3.2.15), indicating successful expression of the *Nat* genes from the pTargeT vector.

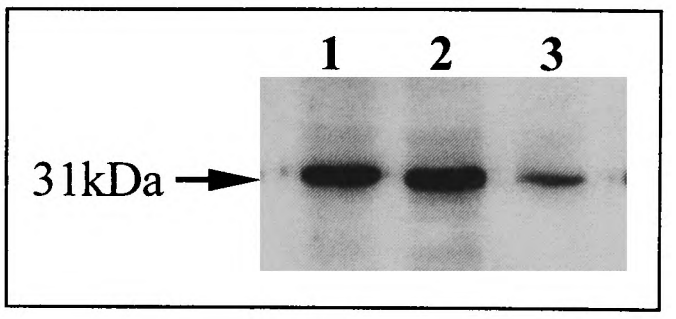


Figure 3.2.15: Western blot analysis showing murine NAT2 expression from the pTargeT constructs in CHO cells. Lysates of CHO cells (20µl), transfected with constructs MCA-*Nat2* (lane 1), MSP-*Nat2* (lane 2) and A/J-*Nat2* (lane 3), were loaded to a 12% (w/v) SDS-PAGE gel (section 2.3.3), which was then blotted overnight (section 2.3.4). Detection of the 31kDa NAT2 protein was carried out with antiserum 184. Expression of recombinant NAT2 protein from the pTargeT constructs was confirmed for all cell lysates.

3.2.7 Comparison between NAT allozymes expressed in CHO cells by enzymatic activity assay

Enzymatic activity of individual NAT allozymes was measured in lysates of CHO cells, transfected with pTargeT constructs (section 3.2.6), using p-anisidine (pANIS), p-aminobenzoic acid (pABA), sulphamethazine (SMZ), p-aminosalicylic acid (pAS) and 5-aminosalicylic acid (5-AS) as substrates. Assays were performed as described in section 2.3.5.2, using 0.2mM of each substrate. The initial reaction rates (pmoles of N-acetylated substrate per minute) were expressed relative to the luminescence measured for each cell lysate (table 3.2.7), to provide the NAT specific activity (pmoles of N-acetylated substrate per minute per unit of luminescence). Acetylation activity with both pANIS and pABA was absent in mock-transfected CHO cells. The results are presented in table 3.2.8 and figure 3.2.16.

Recombinant NAT1 and NAT2 allozymes were all capable of N-acetylating pANIS (table 3.2.8; figure 3.2.16a). The specific activity of MCA NAT2 was 2.4-fold higher than that of MCA NAT1, while the specific activity of MSP NAT2 was 1.8-fold higher than that of MSP NAT1. This indicates that pANIS is a better substrate for NAT2 than NAT1 in MCA and MSP mice. In contrast, the specific activity of the NAT1 6 allozyme (the same in MCA and A/J strains) with pANIS was 2.7-fold higher than that of the NAT2 9 allozyme of the A/J strain, indicating that NAT1 acetylates pANIS more efficiently than NAT2 in the A/J mouse.

In transfected CHO cells, pABA was a substrate for all NAT2 allozymes, as expected (table 3.2.8; figure 3.2.16b). Activity with pAS was detectable in cells expressing MCA NAT1 and MSP NAT1, but it was significantly lower than that measured in cells expressing MCA NAT2 and MSP NAT2 (table 3.2.8; figure 3.2.16c). The specific activity of MCA NAT2 was 3.6-fold higher than that of MCA NAT1, while the specific activity of MSP NAT2 was 5.5-fold higher than that of MSP NAT1. This indicated that pAS is a much better substrate for NAT2, compared with NAT1, in both MCA and MSP strains. In the A/J mouse, however, the NAT1 and NAT2 isoenzymes N-acetylated pAS at approximately the same rate, as indicated by the specific activities measured for the NAT1 6 (MCA and A/J) and NAT2 9 (A/J) allozymes.

Activity with 5-AS was virtually undetectable in cells expressing MCA NAT1 and MSP NAT1. However, 5-AS was a very good substrate for the NAT2 allozymes (table 3.2.8; figure 3.2.16d). SMZ was a poor substrate for all NAT1 and NAT2 allozymes (table 3.2.8), while none of the five compounds used was a good substrate for any of the NAT3 allozymes studied (table 3.2.8).

When pANIS was used as substrate, the specific activity of MCA NAT1 was 1.5-fold higher than that of MSP NAT1 (table 3.2.8; figure 3.2.16a). When pAS was used as substrate, MCA NAT1 provided a 2.8-fold higher specific activity than MSP NAT1 (table 3.2.8; figure 3.2.16c). These results indicate that MSP NAT1 is a slow acetylator of pANIS and pAS, compared with MCA NAT1.

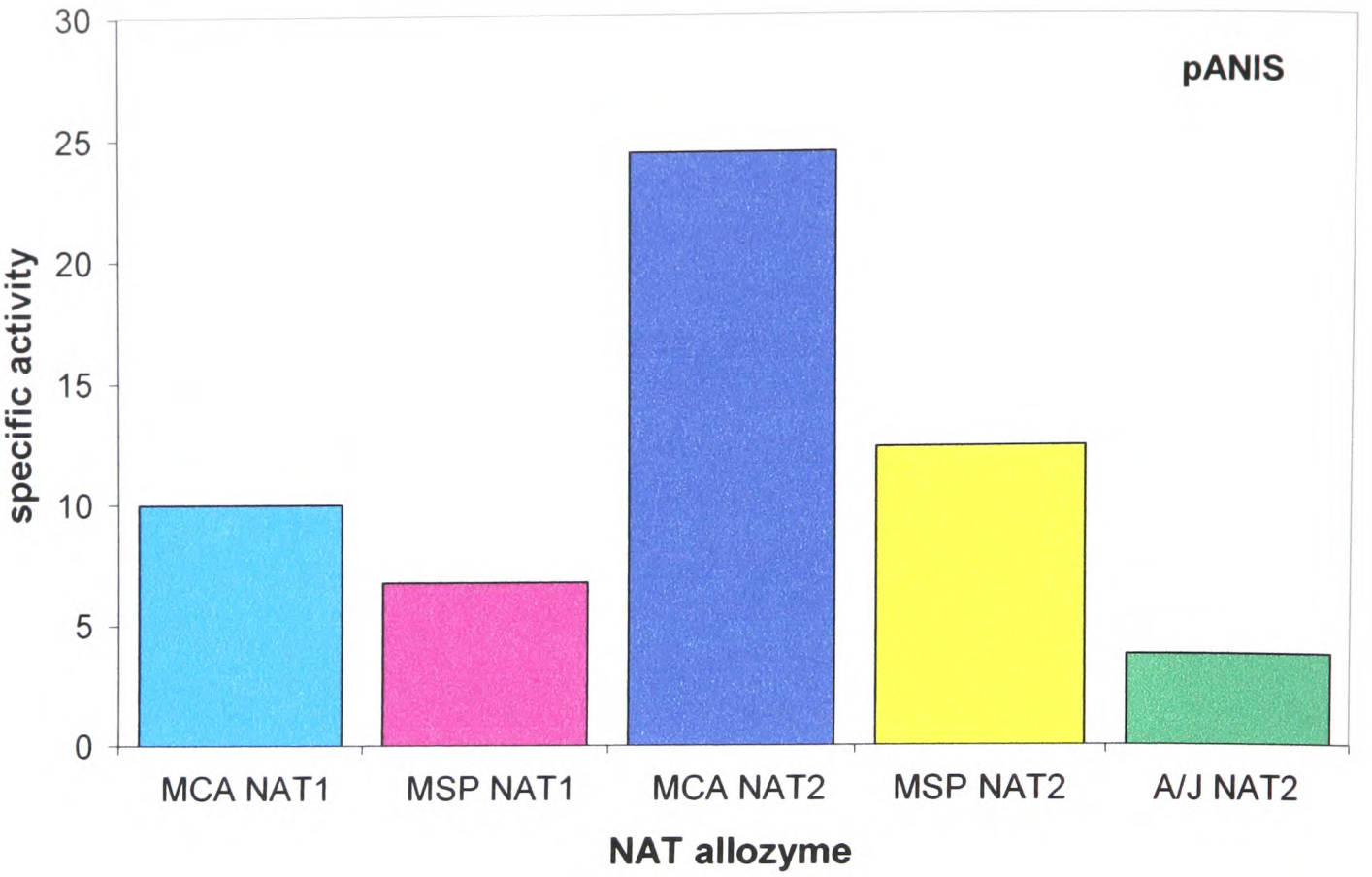
When compared with MSP NAT2, MCA NAT2 provided about 2-fold higher specific activity with pANIS, pABA, pAS and 5-AS substrates. When compared with A/J NAT2, MCA NAT2 provided 6.6-, 3-, 3.2- and 1.8-fold higher specific activity with these substrates (table 3.2.8; figure 3.2.16a-d). Specific activity of MSP NAT2 was 3.3-, 1.5- and 1.8-fold higher than that of A/J NAT2, when pANIS, pABA and pAS were used as substrates, respectively (table 3.2.8; figure 3.2.16a-c). However, no difference was observed between the specific activity of MSP NAT2 and A/J NAT2, with 5-AS as substrate (table 3.2.8; figure 3.2.16d). These results indicate that MCA NAT2, is about two times as efficient as MSP NAT2 with all substrates used. A/J NAT2 had the lowest acetylation capacity, although this varied amongst substrates.

NAT allozyme	Specific activity (pmol/min/unit of luminescence)				
	pANIS	pABA	pAS	5-AS	SMZ
MCA NAT1	10.0	—[a]	9.7	1.2	0.09
MSP NAT1	6.8	—	3.5	n/d[b]	n/d
MCA NAT2	24.4	50.3	35.2	43.3	n/d
MSP NAT2	12.3	25.4	19.2	23.0	0.3
A/J NAT2	3.7	17.05	10.6	23.5	0.15
Balb/c NAT3	0.6	2.2	2.3	n/d	n/d
MCA NAT3	0.8	1.6	2.5	1.1	n/d
MSP NAT3	1.4	2.2	0.5	n/d	0.4

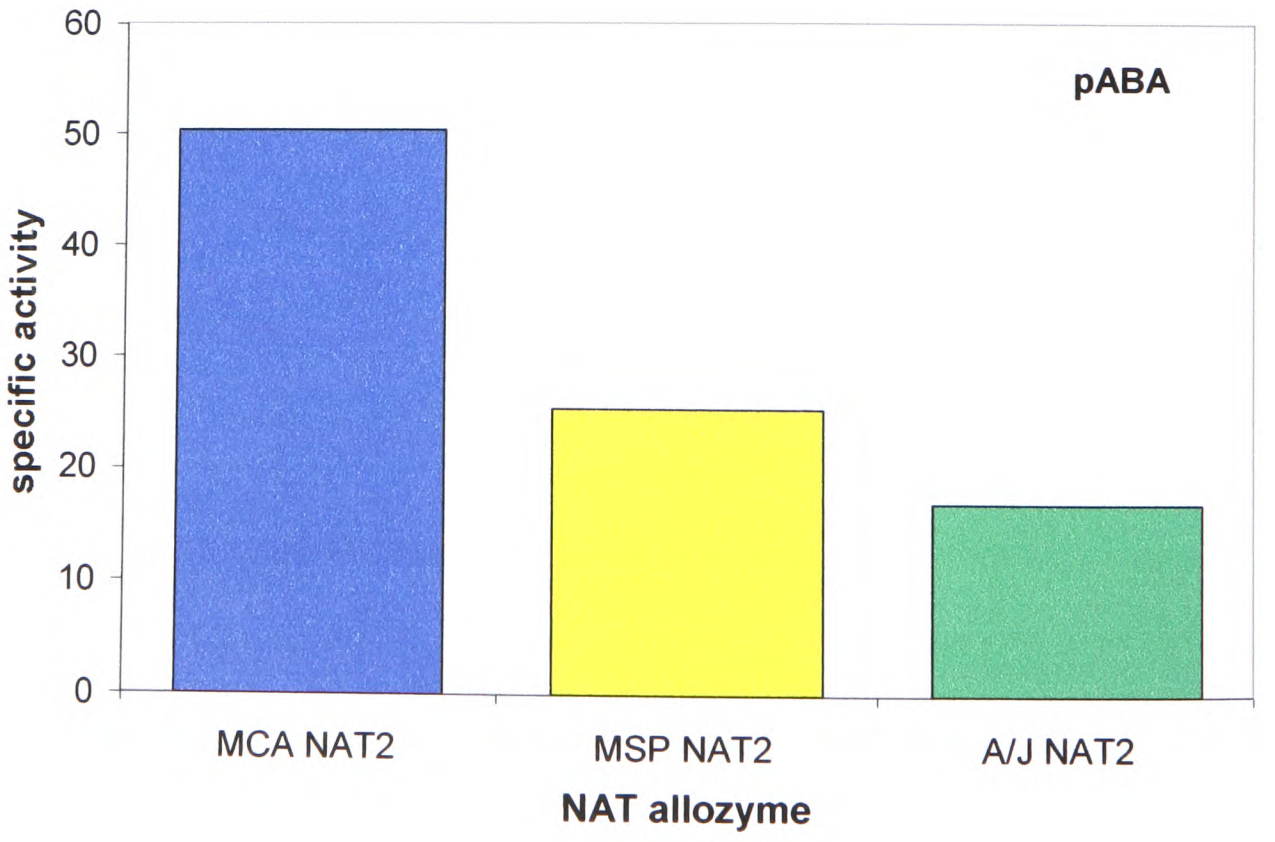
[a]*Not performed;* [b]*Not detected*

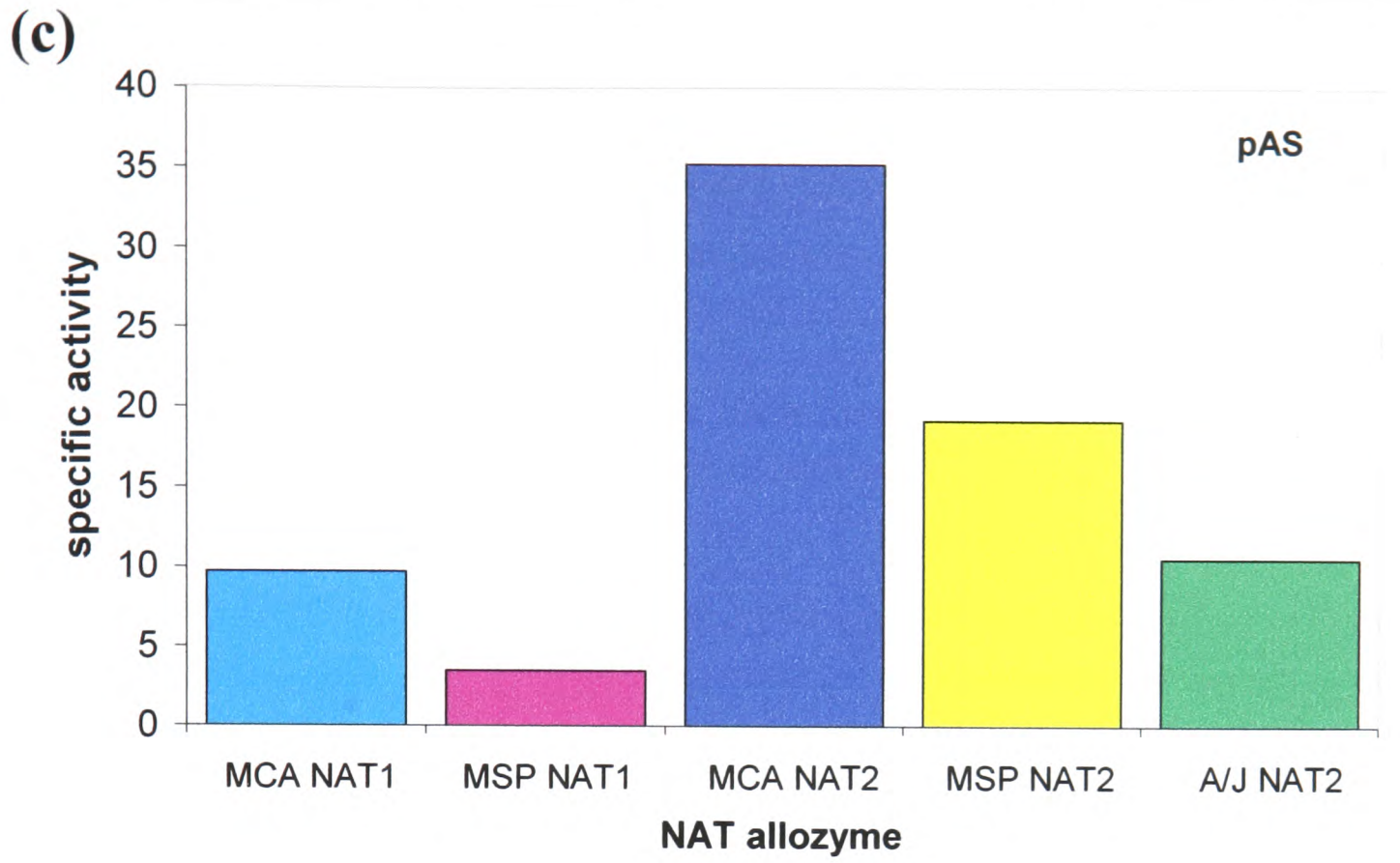
Table 3.2.8: Specific activity of murine NAT allozymes expressed in CHO cells. Enzymatic activity assays (section 2.3.5.2) were performed using 0.2mM pANIS, pABA, pAS, 5-AS or SMZ as substrate. Lysates of CHO cells (section 2.3.2), harvested at confluency (section 2.2.4.2), were diluted 1:10 in the reaction mixture and assayed over a 30 min time course, during which initial reaction rates were measured. Assays with SMZ were performed for 1h with lysates diluted 1:2 in the reaction mixture. Activity with all substrates in cells expressing the NAT3 allozymes was also measured for 1h in lysates diluted 1:2. NAT specific activity was calculated in pmoles of N-acetylated substrate per minute per unit of luminescence, produced by the pGL3-Control luciferase reporter vector.

(a)


pANIS
specific activity
30
25
20
15
10
5
0
MCA NAT1
MSP NAT1
MCA NAT2
MSP NAT2
A/J NAT2
NAT allozyme

(b)


pABA
specific activity
60
50
40
30
20
10
0
MCA NAT2
MSP NAT2
A/J NAT2
NAT allozyme



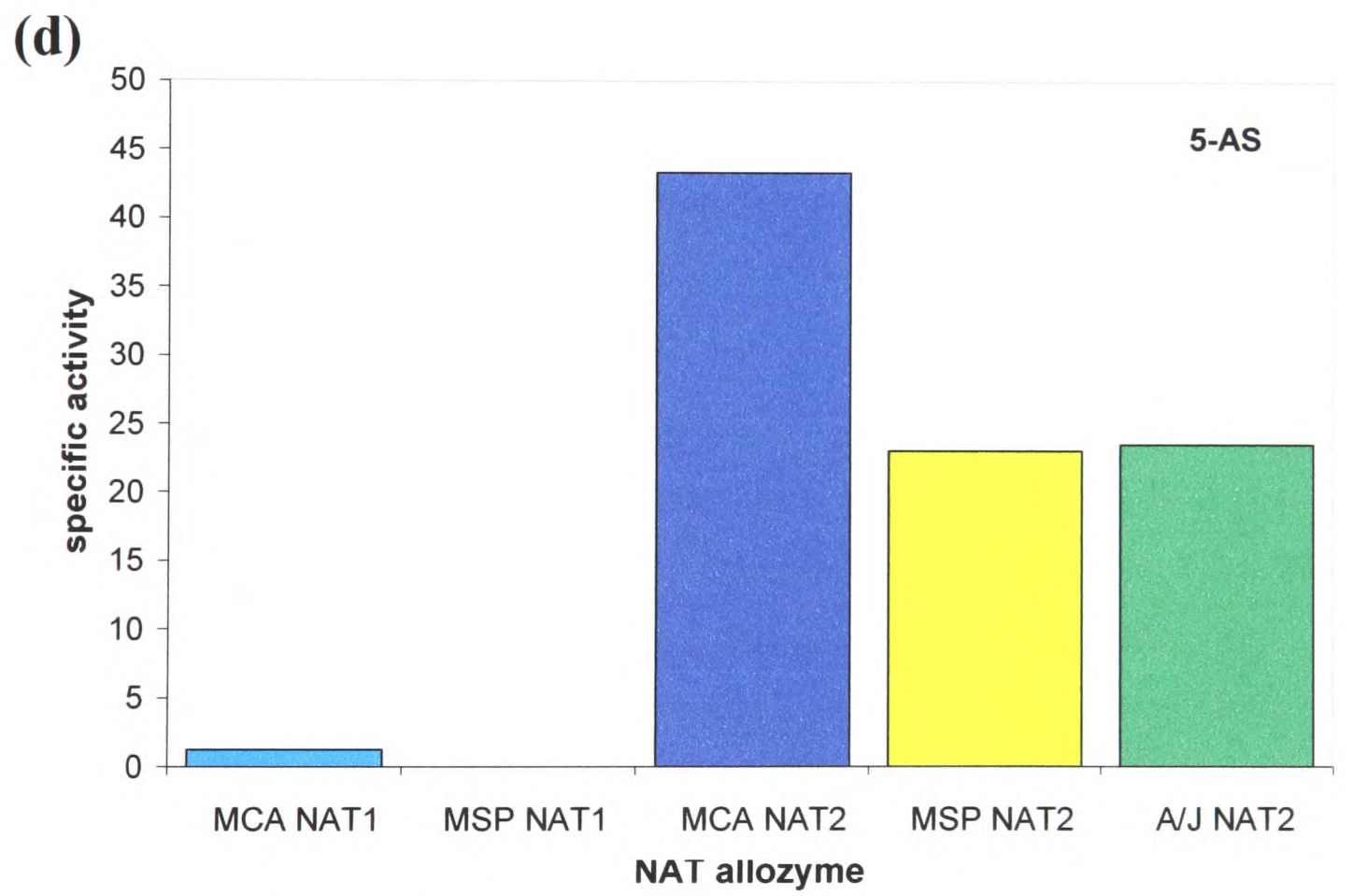


Figure 3.2.16: Comparison of specific activities of murine NAT allozymes expressed in CHO cells. NAT specific activity (pmoles of N-acetylated substrate/min/unit of luminescence) in CHO cells expressing MCA NAT1, MSP NAT1, MCA NAT2, MSP NAT2 and A/J NAT2, with a) pANIS, b) pABA, c) pAS and d) 5-AS as substrate.

3.3 Discussion

The value of the laboratory mouse as a model for studying the role of NAT has been limited by the scarcity of genetic polymorphism in the *Nat* genes among strains. Previous sequencing of the *Nat* genes from three fast acetylating (C57Bl/6J, Balb/c and C3H/HeJ) and two slow acetylating (A/J and A/HeJ) inbred strains did not reveal any polymorphisms, other than the classical $A^{296}\rightarrow T$ SNP associated with the *Nat2*9* allele of the slow strains (Martell *et al.*, 1991; Kelly and Sim, 1994; Fretland *et al.*, 1997; Estrada-Rodgers *et al.*, 1998b). Mice belonging to the CD1 outbred strain were also demonstrated to carry the *Nat2*8* and *Nat2*9* alleles, but new polymorphisms were not found (Estrada-Rodgers *et al.*, 1998b; Estrada *et al.*, 2000). In the study presented in this Chapter, two additional inbred (129/Ola and CBA), two outbred (PO and TO) and two wild-derived inbred (MCA and MSP) strains were screened for polymorphisms in the coding region of the three *Nat* genes. No polymorphisms were found in the inbred and outbred strains, which carry the previously described *Nat1*6*, *Nat2*8* and *Nat3*1* alleles (Vatsis *et al.*, 1995). The *Nat2*8* allele is responsible for the fast acetylator phenotype (Vatsis *et al.*, 1995), therefore, it was assumed that these strains are also fast acetylators.

The most common inbred mouse strains belong to the *Mus musculus* species, including those used in the present and previous studies of murine NAT (figure 3.3.1). In the beginning of the 20th century, W.E. Castle used Lathrop's fancy mice to initiate lines that led to a number of popular inbred strains, such as Strong's CBA, C3H and A or Snell's Balb/c mice. The first 129 strain was also developed from Castle's mice, but via a different route of inbreeding. Following an independent line of crossing, C.C. Little generated the C57Bl strain, again starting from Lathrop's mice. The A/J and A/HeJ strains diverged from Strong's A strain more recently and are, therefore, genetically very similar (reviewed in Nishioka, 1995; Beck *et al.*, 2000).

The value of the wild-derived inbred strains is that they are evolutionarily closer to their wild ancestors, therefore, they are genetically more diverse than the commonly used inbred strains. MCA has been derived from *M.m. castaneus*, a *M. musculus* subspecies found in southeastern Asia. MSP has been derived from *M. spretus*, a species found in southern Spain and northern Africa (Taylor, 2000). These

two strains, as well as other inbred lines derived from the same wild species (e.g. the CAST/Ei and SPRET/Ei strains maintained by the Jackson Laboratory in USA), have been extensively used for genetic mapping (Brady *et al.*, 1997; Avner, 1998; Rhodes *et al.*, 1998).

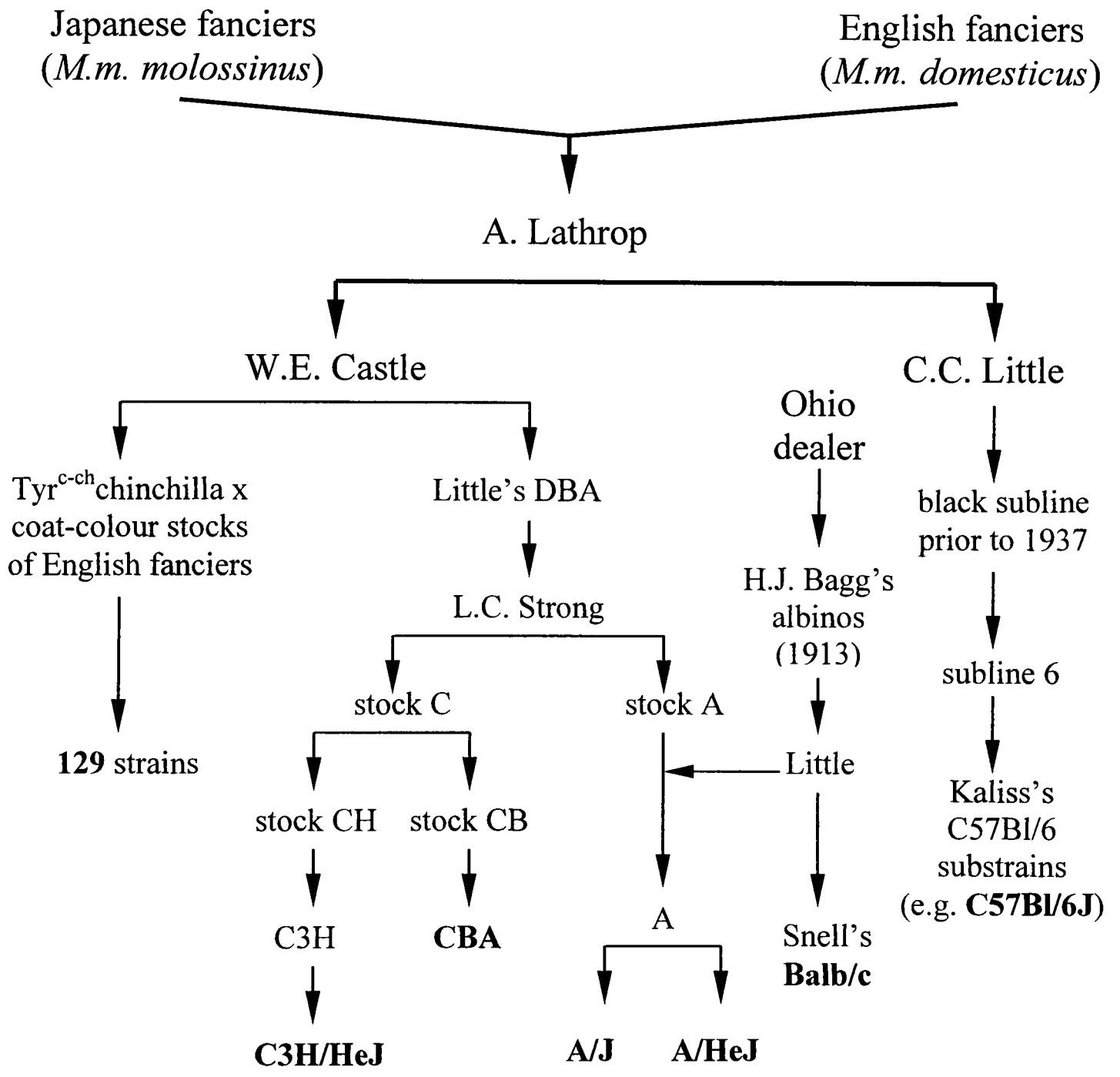


Figure 3.3.1: Genealogy of mouse inbred strains used in NAT studies. Japanese fancy mice (*M.m. molossinus*, a hybrid between the central Asian *M.m. musculus* and the southeastern Asian *M.m. castaneus* subspecies) were taken to Europe by British traders in the 19th century and mixed with English fanciers' colonies (*M.m. domesticus*). In USA, A. Lathrop created a big collection of these mice, which were then introduced to W.E. Castle's laboratory at Harvard. Castle and his student, C.C. Little, developed the ancestors of many commonly used inbred strains. In the 1920s, L.C. Strong used Bagg's albino mice to generate the A strain at the Cold Spring Harbor Laboratory. Bagg's albinos were also ancestors of the Balb/c strain, developed in 1935 by G.D. Snell at the Jackson Laboratory (adapted from Nishioka, 1995; Beck *et al.*, 2000).

Several SNPs were identified in the *Nat* genes of the wild-derived inbred strains, the MSP strain being more polymorphic than the MCA. The *Nat1* sequence of MCA is identical to that previously described for the *Nat1*6* allele (Vatsis *et al.*, 1995), while the *Nat1* sequence of MSP contains five SNPs, one of which causes a $His^{232} \rightarrow Arg$ amino acid substitution. Additional SNPs were found in the *Nat2* gene of the wild-derived inbred strains, namely three conservative SNPs in MCA and seven SNPs in MSP, two of which cause amino acid substitutions ($Glu^{26} \rightarrow Asp$ and $Leu^{82} \rightarrow Met$). The genetic diversity between strains is most evident for the *Nat3* gene, which contains three non-conservative SNPs in MCA and eleven SNPs in MSP, only two of which are conservative. Interestingly, one SNP in *Nat2* ($T^{537} \rightarrow C$) and three SNPs in *Nat3* ($G^{238} \rightarrow A$, $T^{607} \rightarrow C$ and $G^{608} \rightarrow A$) are the same between MSP and MCA strains. It is likely that the ancestral species had a C at position 537 of *Nat2*, and A, C and A at positions 238, 607 and 608 of *Nat3*, respectively. Following the divergence of *M. musculus* and *M. spretus* species, the accumulation of SNPs has produced distinct *Nat* alleles. The *Nat1*6* allele is conserved among *M. musculus* strains, but the *Nat2*8* and *Nat3*1* alleles probably emerged during early breeding of fancy mice. It is also likely that the $A^{296} \rightarrow T$ SNP present in the *Nat2*9* allele of the A/J and A/HeJ strains arose spontaneously during inbreeding of the A strain. Genotyping of additional wild-derived inbred lines (Beck *et al.*, 2000) may provide more information about the evolution of murine *Nat* genes and allow identification of the oldest alleles.

Measurement of NAT1 and NAT2 activities in liver homogenates indicated that MSP is a slow and MCA a fast acetylating strain. A change in NAT activity of the MSP strain would be expected, as both *Nat1* and *Nat2* genes contain non-conservative SNPs. A His at position 232 is conserved in all mouse and rabbit NATs, as well as in hamster and rat NAT2 and human NAT1 (table 3.3.1). Although the $His^{232} \rightarrow Arg$ substitution in MSP NAT1 would not be detrimental in terms of hydrophobicity and charge conservation, it could cause conformational changes. Mammalian His^{232} is homologous to Lys^{214} of *S. typhimurium* NAT (*St*NAT) (figure 3.3.2), located in the third domain of the molecule, at the beginning of sheet β12. The third domain forms an α/β "lid" and is involved in the formation of the active site cleft (Sinclair *et al.*, 2000). The considerable decrease in the activity of recombinant MSP NAT1 could, therefore, suggest a role for His^{232} in stabilisation of the region.

	AMINO ACID POSITION										
	232	**26**	**82**	**80**	**99**	**138**	**171**	**203**	**213**	**266**	**286**
Balb/c NAT1	His	Glu	Met	Thr	Ser	Thr	Pro	Glu	Pro	Val	Leu
MCA NAT1	His	Glu	Met	Thr	Ser	Thr	Pro	Glu	Pro	Val	Leu
MSP NAT1	**Arg**	Glu	Met	Thr	Ser	Thr	Pro	Glu	Pro	Val	Leu
Balb/c NAT2	His	Glu	Leu	Thr	Ile	Thr	Pro	Glu	Pro	Val	Arg
MCA NAT2	His	Glu	Leu	Thr	Ile	Thr	Pro	Glu	Pro	Val	Arg
MSP NAT2	His	**Asp**	**Met**	Thr	Ile	Thr	Pro	Glu	Pro	Val	Arg
Balb/c NAT3	His	Glu	Ile	Ala	Cys	Thr	Ser	Trp	Arg	Val	Val
MCA NAT3	His	Glu	Ile	**Thr**	Cys	Thr	Ser	**Gln**	Arg	Val	Val
MSP NAT3	His	Glu	Ile	**Thr**	**Arg**	**Ile**	**Pro**	**Gln**	**Gln**	**Ile**	**Ala**
Rabbit NAT1	His	Asp	Thr	Thr	Asp	Ile	Pro	Glu	Pro	Val	Leu
Rabbit NAT2	His	Asp	Thr	Thr	Asp	Ile	Pro	Glu	Pro	Val	Leu
Hamster NAT1	Cys	Glu	Met	Thr	Ser	Ala	Pro	Glu	Pro	Val	Leu
Hamster NAT2	His	Glu	Met	Thr	Asn	Val	Pro	Glu	Pro	Val	Arg
Rat NAT1	Cys	Glu	Met	Thr	Asn	Thr	Pro	Glu	Pro	Val	Leu
Rat NAT2	His	Glu	Met	Thr	Asn	Thr	Pro	Glu	Pro	Val	Arg
Human NAT1	His	Asp	Ile	Thr	Lys	Ile	Pro	Glu	Pro	Val	Arg
Human NAT2	Tyr	Asp	Ile	Thr	Asn	Ile	Thr	Glu	Pro	Val	Gly

Table 3.3.1: Species comparison of NAT proteins at amino acid positions substituted in wild-derived inbred mouse strains. The amino acid sequences of NAT1, NAT2 and NAT3 isoenzymes from the Balb/c mouse are compared with those of the wild-derived inbred strains MSP and MCA, as well as with the NAT isoenzymes of rabbit, Syrian hamster, rat and human. Amino acid position 232 is substituted in MSP NAT1, positions 26 and 82 in MSP NAT2, positions 80 and 203 in both MCA and MSP NAT3, while positions 99, 138, 171, 213, 266 and 286 in MSP NAT3 only. These substitutions are indicated in bold.

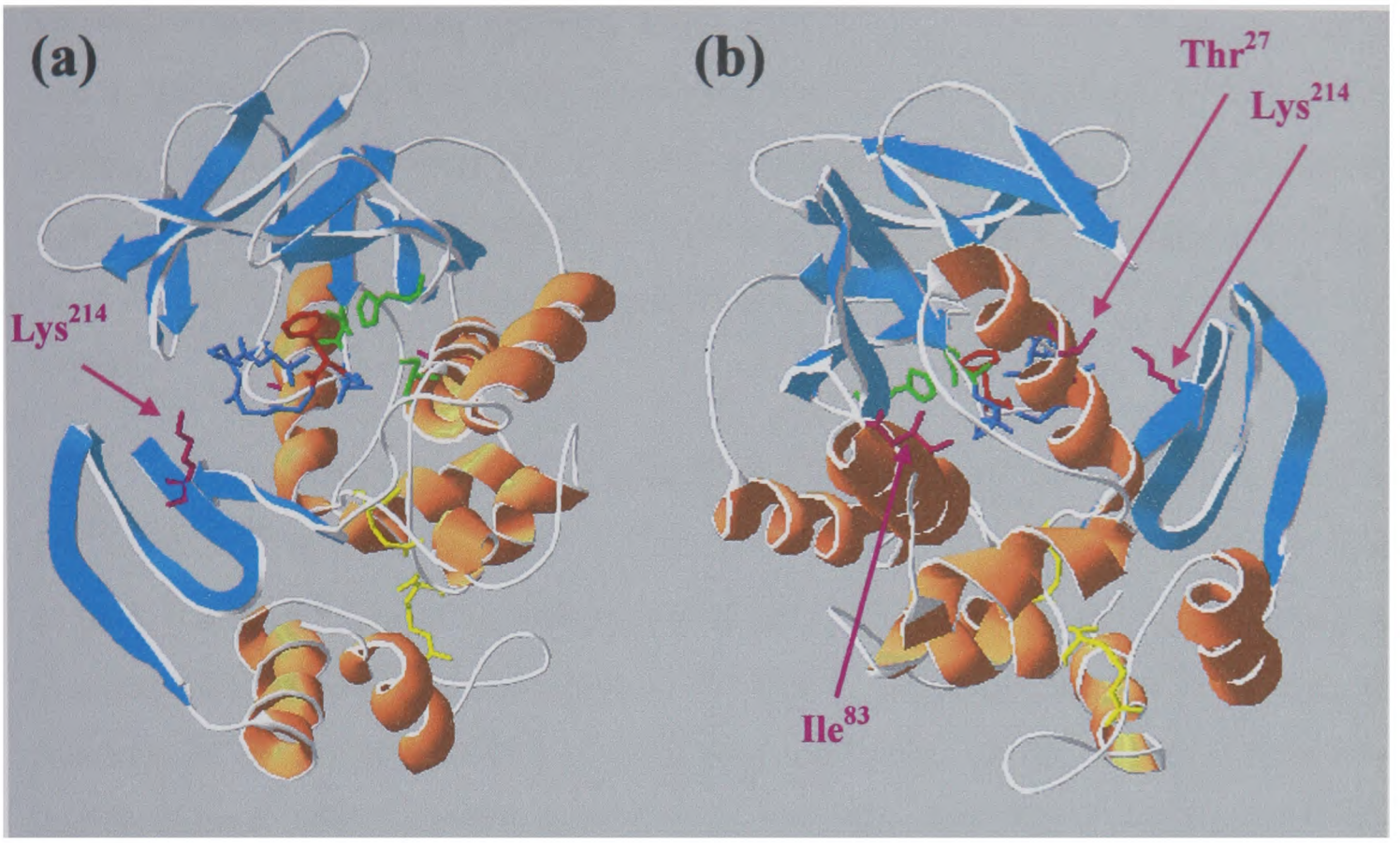


Figure 3.3.2: Structure of *S. typhimurium* NAT, showing the positions of amino acids substituted in the NAT1 and NAT2 isoenzymes of the MSP mouse strain. Analysis of *St*NAT structure (PDB accession no. 1E2T) was carried out using the Swiss-PdbViewer. The two images (a and b) are rotated relative to each other by 180° around the z axis. The Cys^{69}-His^{107}-Asp^{122} catalytic triad (Sinclair *et al.*, 2000) is shown in green. The backbone of the postulated active site loop, spanning amino acid positions 123-129 (Rodrigues-Lima *et al.*, 2001), is indicated in blue. Phe^{125}, shown in red, is part of the active site loop and appears to be a key element for substrate selectivity of mammalian NATs (Goodfellow *et al.*, 2000). Conserved Arg^{64} and Glu^{39} residues, important for conformational stabilisation of the active site (Sinclair *et al.*, 2000), are shown in yellow. Amino acids at positions 27, 83 and 214 of *St*NAT, shown in purple, are homologous to positions 26, 82 and 232 of murine NATs, substituted in the MSP strain (table 3.3.1).

Position 26 bears a negatively charged amino acid in all mammalian NAT proteins characterised to date (table 3.3.1). All inbred mouse strains analysed here and in previous studies (Martell *et al.*, 1991; Kelly and Sim, 1994; Fretland *et al.*, 1997; Estrada-Rodgers *et al.*, 1998b), as well as the wild-derived MCA strain, possess a Glu at position 26 of all three NAT isoenzymes. This is substituted by an Asp in MSP NAT2 (table 3.3.1). An Asp at position 26 is also present in rabbit and human NATs, while the isoenzymes of hamster and rat possess a Glu, similarly to the mouse strains of *M. musculus* origin.

Amino acid position 82, bearing a Leu→Met substitution in MSP NAT2, is less well conserved among different NATs (table 3.3.1). The most common amino acid at this position is Met, found in murine NAT1, in both isoenzymes of hamster and rat, as well as in chicken NAT. Among mammalian NATs, Leu is only present at position 82 of murine NAT2 (table 3.3.1), but is also found at the equivalent position 83 of *E. coli* NAT, and 84 of *M. smegmatis* and *M. tuberculosis* NATs (Payton *et al.*, 1999a).

Amino acids at positions 26 and 82 of mammalian NATs are homologous to Thr^{27} and Ile^{83} of *St*NAT (figure 3.3.2). Thr^{27} is part of the α2 helix of the first domain, while Ile^{83} is the last amino acid of helix α5, located at the boundary of the first and the second domain (Sinclair *et al.*, 2000). Although amino acid changes at positions 26 and 82 of MSP NAT2 are conservative, they cause an approximately 2-fold decrease in NAT2 specific activity in liver homogenates and in transfected CHO cells, suggesting that they may affect sites important for maintaining the activity and/or stability of the enzyme. Semi-quantitative Western blot analysis indicated that MSP resembles A/J in that hepatic NAT2 is less abundant than in Balb/c and MCA liver homogenates. Polymorphisms in the coding region of a gene would not be expected to affect levels of mRNA or protein synthesis. Therefore, it seems more likely that the produced MSP NAT2 protein is less stable and/or more susceptible to proteolytic degradation. A similar effect has been observed for the A/J mouse, where the Asn^{99}→Ile substitution in NAT2 reduces stability of the enzyme and may also compromise its catalytic efficiency (De Leon *et al.*, 1995). A decrease in the amount of immunoreactive protein has also been observed for human NAT2 variants with amino acid changes at positions 114 (Ile→Thr), 145 (Gln→Pro) and 197 (Arg→Gln) (Fretland *et al.*, 2001).

Variation in the acetylation activity among inbred and wild-derived strains, especially of such diverse origin as the *M. spretus*, could be attributed to differences in their genetic background, rather than to polymorphisms affecting their NAT isoenzymes. Differences of this kind have been described in earlier studies, comparing *Nat2* congenic mice with the parental C57Bl/6J and A/J inbred strains (Levy and Weber, 1989; Levy *et al.*, 1994). To overcome this problem, the NAT1 and NAT2 allozymes of MCA, MSP and A/J strains were expressed in CHO cells and

enzymatic activity was subsequently measured with different substrates. The results with pANIS and pABA matched those obtained with liver homogenates.

Activity measurements with cell lysates were normalised to the luminescence generated by firefly luciferase, the product of the pGL3-Control vector co-transfected in CHO cells with each *Nat* construct. Similar co-transfections of COS-1 cells, using a β-galactosidase reporter vector as internal standard, have previously been described by Martell *et al.* (1992). Such protocols allow a more accurate estimation of the variation in transfection efficiency, compared with previous measurements of total protein in lysates of transfected cells (Blum *et al.*, 1990a; Ohsako and Deguchi, 1990; Sasaki *et al.*, 1991; Deguchi, 1992; Estrada-Rodgers *et al.*, 1998b). Small differences in transfection efficiency could affect levels of recombinant NAT expression, without necessarily changing the total amount of protein to a detectable degree. Other investigators have normalised enzymatic activity on the amount of NAT protein detected by Western blot analysis (Doll and Hein, 1995; Fretland *et al.*, 1997). However, differences in the amount of immunoreactive protein may be due to instability of the expressed molecule rather than variation in the transfection efficiency.

All NAT2 allozymes expressed in CHO cells acetylated pAS and 5-AS quite efficiently, although there was a difference in the acetylation activity between allozymes from fast and slow acetylating strains. Aminosalicylates are used to treat bowel diseases, including ulcerative colitis and Crohn's disease (Hanauer and Dassopoulos, 2001). The front-line drug 5-AS is subject to N-acetylation by human NAT1 (reviewed in Weber and Hein, 1985), the functional equivalent of murine NAT2. It is demonstrated here that murine NAT2 is also capable of N-acetylating 5-AS, in agreement with a previous study (Estrada-Rodgers *et al.*, 1998b). Both murine NAT2 and human NAT1 are expressed along the full length of the intestine (Chung *et al.*, 1993; Ware and Svensson, 1996; Stanley *et al.*, 1997; Hickman *et al.*, 1998), where they may be involved in drug inactivation. Moreover, a recent study by Deloménie *et al.* (2001) has demonstrated high acetylation activity towards 5-AS in bacteria of the intestinal microflora, including species over-represented in inflammatory bowel diseases. The combined actions of endogenous and bacterial NATs may determine the rate of 5-AS inactivation in the intestine and, therefore, the

efficacy of treatment. The use of fast and slow acetylating mice as models for studying the metabolism of 5-AS in the intestine, may help estimate the contribution of enteric bacteria and the host to 5-AS acetylation.

In agreement with previous studies (Martell *et al.*, 1992; Kelly and Sim, 1994; Esrada-Rodgers *et al.*, 1998b), acetylation of the sulphonamide drug SMZ was almost undetectable with all NAT1 and NAT2 allozymes expressed in CHO cells. Earlier studies have also failed to demonstrate significant activity of mouse NAT3 towards common NAT substrates (Kelly and Sim, 1994; Fretland *et al.*, 1997; Estrada-Rodgers *et al.*, 1998b). As it was possible that the *Nat3*1* allele from previously analysed inbred strains might contain deleterious mutations, the *Nat3*2* allele from MCA and the *Nat3*3* allele from MSP were expressed in CHO cells and their recombinant products tested for activity with a series of substrates. Activity was still very low or absent. It appears that the NAT3 allozymes examined here are either non-functional or have substrate specificities which do not overlap with those of other NATs. Many sites of amino acid change (e.g. 80, 171, 203, 213 and 266) in the NAT3 allozymes described in this study are very well conserved among mammalian NATs (table 3.3.1), suggesting that substitutions at these sites might compromise enzymatic function. Other amino acids, which are strictly conserved among mammalian NATs, but are different in murine NAT3 allozymes, are located at positions 57 (Asp→His), 126 (Gly→Pro), 131 (Met→Leu), 162 (Asp→Glu), 164 (Ile→Thr), 165 (Arg→Lys), 168 (Gln→Glu), 180 (Leu→Phe), 206 (Asn→Ser), 209 (Leu→Tyr), 217 (Phe→Met), 226 (Gln→His), 233 (Cys→Gly), 242 (Arg→Lys) and 274 (Ser→His) (figure 1.5). Many of these substitutions would be expected to have a severe effect on protein structure and function, therefore, their accumulation during evolution may have rendered NAT3 inactive.

Although there are fast and slow acetylating inbred strains for many laboratory species, there is currently no ideal model for the human NAT polymorphism. In rodents, polymorphism is always associated with NAT2, which is functionally related to human NAT1 (Ozawa *et al.*, 1990; Martell *et al.*, 1992; Doll and Hein, 1995). For example, a nonsense mutation at position 727 of the *Nat2*16* allele of the Syrian hamster leads to the production of a truncated protein, of only 242 amino acids (Nagata *et al.* 1994; Ferguson *et al.*, 1996). In humans, *NAT1*15* and *NAT1*19* alleles

also contain nonsense mutations, leading to the production of NAT1 proteins with only 186 and 32 amino acids, respectively (Hubbard *et al.*, 1998; Hughes *et al.*, 1998a; Lin *et al.*, 1998). Therefore, the slow acetylating Syrian hamster strains can be good models for studying the effect of NAT1 null phenotype in humans. Other human *NAT1* alleles have been associated with lower acetylation capacities (section 1.2.3.1). Slow acetylating mouse and rat strains could, therefore, be useful models for studying the impact of this type of mutations *in vivo*.

Establishing a model for human NAT2 polymorphism has been a more difficult task. Rabbit is polymorphic at the *Nat2* locus, but the slow acetylator phenotype is caused by deletion of the entire *Nat2* gene, rather than presence of point mutations in the coding region (Blum *et al.*, 1989b). To date, deletion of human *NAT2* has not been reported, suggesting that the molecular mechanism underlying the NAT2 polymorphism in rabbits may be entirely different from that in humans. The novel *Nat1*30* allele of the MSP strain, described in the present study, contains a non-conservative SNP, leading to a decrease in NAT1 activity. Therefore, the MSP strain represents a good model for studying human NAT2 polymorphism. MCA and MSP strains can also be used alongside the already popular C57Bl/6J, Balb/c and A/J strains for studying the role of murine NAT2 in the metabolism of xenobiotics or putative endogenous substrates. The novel polymorphisms described here are anticipated to promote our understanding of the function of murine NAT isoenzymes and provide useful models for studying the acetylation polymorphism of humans.

CHAPTER 4

Study of the genomic region around the murine *Nat* loci; towards the production of a *Nat2* knock-out mouse

4.1 Introduction

Despite their important role in the metabolism of environmental and synthetic chemicals, the xenobiotic metabolising enzymes are thought to have originally evolved to catalyse reactions of endogenous metabolism (Nebert, 1997). For example, the P450 enzymes of phase I metabolism are known to act upon a wide range of endogenous substrates, including steroids, fatty acids, bile acids, prostaglandins, leukotrienes, retinoids and biogenic amines (Nelson *et al.*, 1996; Wong, 1998). The sulphotransferases of phase II metabolism play a key role in the conversion of steroids, thyroid hormones and catecholamines (Coughtrie *et al.*, 1998; Glatt *et al.*, 2000), while the UDP-glucuronosyltransferases are involved in the metabolism of steroids and bilirubin (Green and Tephly, 1998; McGurk *et al.*, 1998).

Very little is known about the possible involvement of NAT in the metabolism of endogenous substrates. NATs are a group of proteins conserved in both prokaryotes and eukaryotes (section 1.3), suggesting that they may be responsible for important life functions. As will be described in the following Chapters, human NAT1 and its functional equivalent murine NAT2 might play a role in endogenous metabolism, because of their widespread tissue distribution and expression very early in development.

The most widely accepted hypothesis today is that NAT may be involved in the metabolism of folate. This was based on the earlier finding that recombinant human NAT1 is capable of N-acetylating the folate catabolite p-aminobenzoyl glutamate (pABGlu) (Minchin, 1995; Ward *et al.*, 1995). More recent studies using

whole tissue homogenates from mice have demonstrated a good correlation between the acetylation of pABGlu and the NAT2-specific substrate pABA, with the slow acetylating A/J mouse showing lower activity levels than the fast acetylating C57Bl/6J mouse (Payton *et al.*, 1999b). A similar correlation between pABA and pABGlu acetylation has also been observed with placental homogenates from human subjects (Smelt *et al.*, 2000). Human NAT1 has been identified, by specific inhibition with 5-iodosalicylic acid, as the major enzymatic activity responsible for the conversion of pABGlu. Formation of N-acetyl pABGlu (N-apABGlu), as the end-product of this reaction, has been confirmed by mass spectrometry (Upton *et al.*, 2000).

"Folate" (figure 4.1.1) is the generic term for a large family of water-soluble B-vitamin coenzymes, responsible for the transfer of one-carbon units in metabolic reactions leading to the biosynthesis of methionine, purines and pyrimidines (Lucock, 2000). Mammals cannot synthesise folate and depend on a variety of dietary sources (mainly leafy green vegetables) for this essential nutrient. However, the bacteria of the intestinal microflora can synthesise folate and may contribute to the total intake of the vitamin by the body (Camilo *et al.*, 1996). Much of the folate absorbed by the intestine is taken up by the liver or stored in the erythrocytes. Folate is also found in the plasma, associated with low- (mainly albumin) or high-affinity binding proteins. Folate coenzymes enter the cell via binding to specialised receptors (FR-α, β and γ) and are distributed either in the cytosol or the mitochondria (Bailey and Gregory, 1999a; Barber *et al.*, 1999). The catabolism of folate involves non-catalytic cleavage of the molecule at the C9-N10 bond, producing a variety of pteridines and pABGlu (figure 4.1.1). The latter undergoes N-acetylation prior to excretion, N-apABGlu being a major folate catabolite in urine (McPartlin *et al.*, 1992; Geoghegan *et al.*, 1995).

Folate is essential for normal embryonic development, as it is required for the biosynthesis of nucleic and amino acids. Folate deficiency is an established risk factor for neural tube defects (NTDs) (Barber *et al.*, 1999) and has also been implicated in the aetiology of cleft lip and palate, congenital heart, limb and urinary tract defects, recurrent early pregnancy loss, low birth weight and preeclampsia (Lucock, 2000). Low folate levels, associated with elevated plasma homocysteine levels, have also

been linked to various diseases of the adult life, such as occlusive vascular disease, colorectal cancer and Alzheimer's disease (Clarke *et al.*, 1998; Duthie, 1999; Lucock, 2000). Severe folate deficiency leads to megaloblastic anaemia, a condition characterised by abnormal cell division of red cell precursors and accumulation of large, nucleated, defective cells (megaloblasts) in the bone marrow (Bailey and Gregory, 1999a).

Impaired function of enzymes of the folate cycle can lead to folate deficiency and disease (Harris, 2001). A thermolabile, low-activity variant of methylene-tetrahydrofolate reductase (MTHFR) has been implicated in up to 15% of folate-sensitive NTD cases (Bailey and Gregory, 1999b; Ueland *et al.*, 2001). However, maternal folic acid supplementation during pregnancy has been demonstrated to reduce the risk for NTDs by up to 70% (MRC Vitamin Study Research Group, 1991; Czeizel and Dudás, 1992), suggesting that factors other than the *MTHFR* mutation should be involved in the aetiology of the disease. NAT could be important in maintaining the bioavailability of folate, by regulating its rate of catabolism via pABGlu acetylation. During pregnancy, excessive N-acetylation in the placenta might limit the amount of biologically active vitamin reaching the developing foetus, providing a plausible explanation for the beneficial effect of folate supplementation.

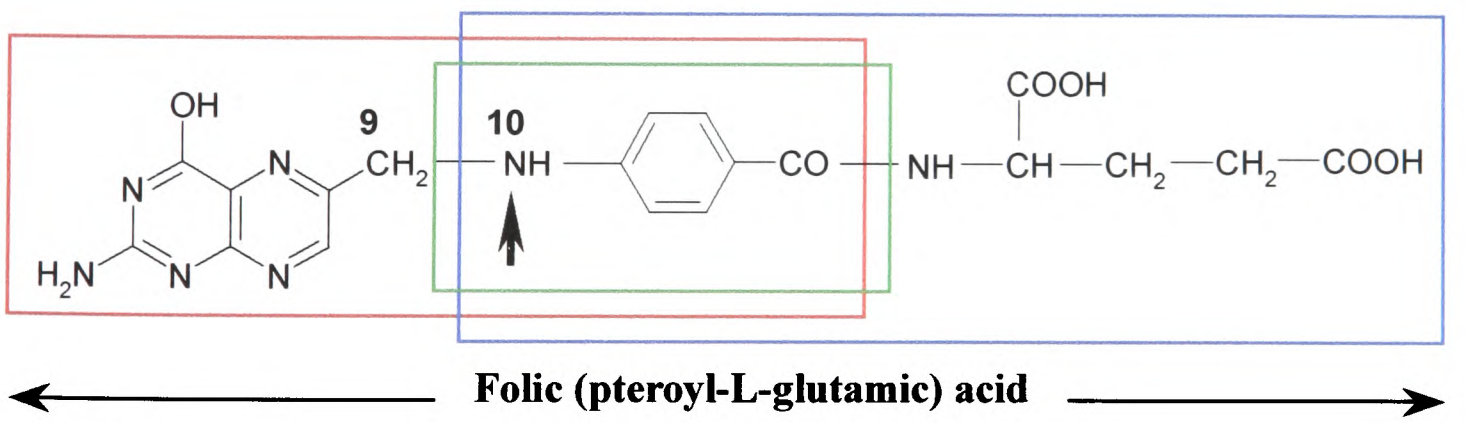


Figure 4.1.1: The chemical structure of folic (pteroyl-L-glutamic) acid. This is a stable synthetic analogue, the parent form of the large family of folate coenzymes. The pteridine moiety is enclosed in the red box and the pABGlu moiety in the blue box. The structure in the green box is the known NAT substrate pABA. The most prevalent dietary forms of the vitamin are polyglutamyl folates (mainly polyglutamyl 5-methyltetrahydrofolate and formyltetrahydrofolate), which undergo hydrolysis into their monoglutamyl equivalents prior to absorption by the intestine. Folate catabolism involves cleavage at the indicated C9-N10 bond, giving rise to pteridines and pABGlu. The latter is then N-acetylated (N indicated with an arrow), probably by NAT1 in humans and NAT2 in mice (modified from Gregory, 2001).

The study of the postulated role of NAT in endogenous metabolism can be greatly facilitated by the production of transgenic mice, lacking or overexpressing the *Nat2* gene. The task of producing a *Nat2* knock-out mouse was undertaken in collaboration with the laboratory of Dr. Andrew Smith at the Centre for Genome Research of the University of Edinburgh, and has involved the work of many researchers. The work presented in this Chapter (section 4.2.1 and 4.2.2) has provided the basis for the generation of a targeting construct, used to inactivate the *Nat2* gene. To facilitate this process, an extended region around the *Nat2* gene was sequenced and subjected to computational analysis. Additional tools were developed, to allow screening of transformed mouse ES cells for targeted replacement of the endogenous *Nat2* locus via homologous recombination with the construct.

Early work with recombinant inbred and congenic mouse strains has suggested linkage between the slow acetylator phenotype and high susceptibility of A/J mice to orofacial clefting induced by teratogens, such as glucocorticoids, 6-aminonicotinamide and possibly phenytoin (Karolyi *et al.*, 1987; 1988; 1990). Although murine *Nat2* could be involved in early embryonic development via its postulated role in folate catabolism, it is possible that another nearby polymorphic locus is responsible for the increased susceptibility of A/J mice to cleft lip and palate. To explore this possibility, preliminary work was carried out (sections 4.2.3 and 4.2.4), towards physical localisation of the mouse *Nat* genes and characterisation of proximal polymorphic genetic markers, to facilitate refined linkage analysis between *Nat* and other disease loci in the region.

4.2 Results

4.2.1 Towards the generation of a *Nat2* knock-out mouse

4.2.1.1 Cloning of the positive and negative selection cassettes

Targeted inactivation of murine genes is based on the intrinsic ability of ES cells to mediate recombination between homologous DNA sequences (Capecchi, 1989). Because the frequency of homologous recombination in higher eukaryotes is extremely low, it is necessary to enrich ES cells that have correctly integrated the targeting construct into their genome. This can be achieved via "positive" and "negative" selection of the transformed cells, provided that specially designed cassettes have been cloned into the targeting construct (Galli-Taliadoros *et al.*, 1995; Moreadith and Radford, 1997; Müller, 1999).

The selection cassettes (Nehls *et al.*, 1996) used for the generation of a mouse *Nat2* targeting construct were obtained from the Centre for Genome Research in Edinburgh. The 5.2kb positive selection cassette (figure 4.2.1a) was supplied as a *Bam*HI fragment and the 4kb negative selection cassette (figure 4.2.1b) as a *Xho*I/*Sal*I fragment. Both cassettes were ligated into the pBluescript (pBS) vector (Appendix 1.3), then transfected into *E. coli* cells (section 2.2.3.10), to create an unlimited source of cloned material (figure 4.2.2) for subsequent production of the *Nat2* targeting construct.

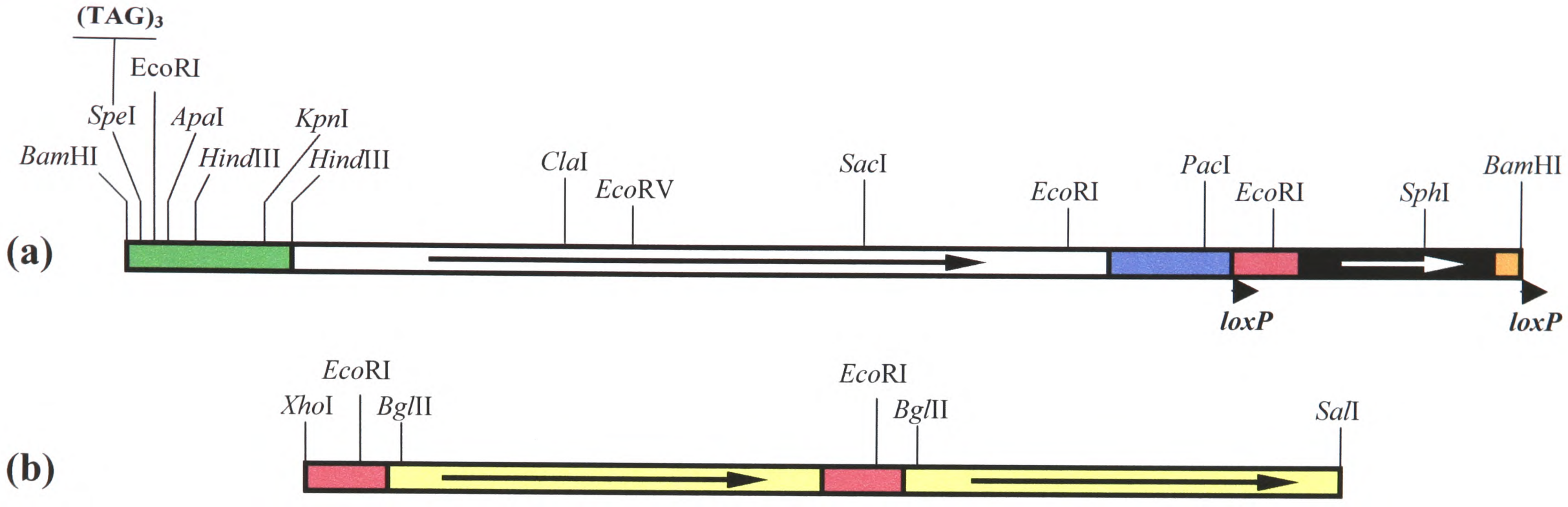


Figure 4.2.1: The selection cassettes used for the generation of a mouse *Nat2* targeting construct. a) The 5.2kb positive selection cassette was used to inactivate the *Nat2* gene via interruption of its coding region. It carries the *neo*^{r} gene (black box), which provides resistance to the antibiotic G418. Flanking the *neo*^{r} gene, there are two *loxP* sites (black triangles), recognised by *Cre* recombinase. Transcription of the *neo*^{r} gene is initiated by the synthetic polyoma enhanced HSV thymidine kinase (*tk*) promoter (MC1) (red box) and terminated after a HSV *tk* polyA signal (orange box). The cassette also contains the *lacZ* gene for β-galactosidase (white box), preceded by an internal ribosomal entry site (IRES) (green box) and three stop codons [$(TAG)_3$], one for each reading frame. The IRES element creates a bicistronic transcription unit, which consists of the truncated *Nat2* coding region, the entire *lacZ* gene and its 3′ UTR containing a SV40 polyA signal (blue box). In the knock-out mouse, transcription of the *lacZ* gene is placed under the control of the *Nat2* promoter, leading to β-galactosidase-positive blue staining of cells naturally expressing the *Nat2* gene. b) The 4kb negative selection cassette is incorporated at the end of the targeting construct and contains two copies of the HSV *tk* gene (yellow boxes), providing sensitivity to ganciclovir (GCV). Transcription of the *tk* genes is initiated by MC1 promoters (red boxes). When replacement of the endogenous locus is correct, the cassette is removed, allowing survival of the ES cells in GCV. In both figures, the arrows indicate the orientation of the coding sequences. *(Related references: Hasty and Bradley, 1992; Galli-Taliadoros et al., 1995; Nehls et al., 1996; Müller, 1999).*

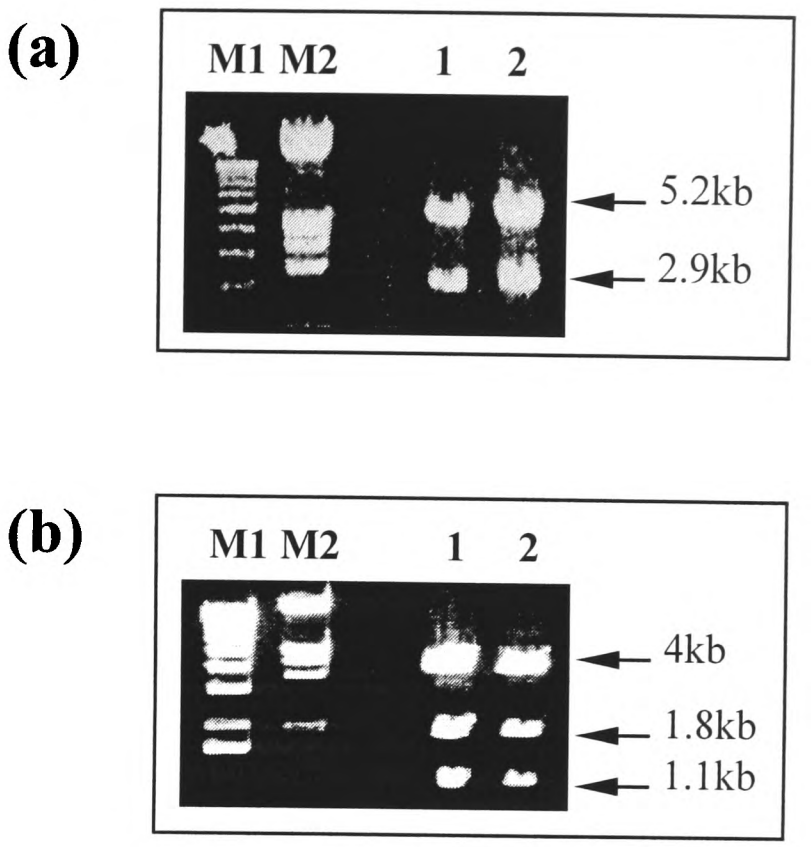


Figure 4.2.2: Cloning of the selection cassettes into the pBS vector. The positive and negative selection cassettes were ligated (section 2.2.3.9) to *Bam*HI- and *Xho*I+*Sal*I-digested pBS vector, respectively. Prior to ligation, the digested vector was dephosphorylated (section 2.2.3.8), to prevent self-ligation, and gel purified (section 2.2.1.3). Transformed JM109 High Efficiency Competent Cells (*E. coli*) were grown on LB-agar plates, supplemented with 100μg/ml ampicillin, 0.5mM IPTG and 40μg/ml x-gal for blue/white selection of colonies (section 2.2.3.10). Plasmid DNA was isolated from liquid cultures of selected white colonies (section 2.2.1.2). a) Excision of the positive selection cassette (5.2kb band) from the pBS vector (2.9kb band) by *Bam*HI digestion. b) Excision of the negative selection cassette from the pBS vector by *Xho*I and *Sal*I digestion. Additional digestion with *Sca*I, cutting the pBS vector at position 2526 (Appendix 1.3), allowed separation of the cassette (4kb band) from the vector (1.8 and 1.1kb bands). Lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively.

4.2.1.2 Cloning of the *Nat2* homology region

To achieve targeted replacement of the endogenous *Nat2* locus, the construct should contain two homology arms (typically 2-10kb in size), flanking the coding region of the *Nat2* gene. It has been estimated that the frequency of homologous recombination in ES cells increases exponentially with the length of these arms (Hasty and Bradley, 1992; Galli-Taliadoros *et al.*, 1995; Müller, 1999). Screening of the E14 TG2a mouse genomic library (Nehls *et al.*, 1994) with a *Nat2*-specific probe has previously provided a 14.3kb plasmid clone (clone A), containing part of the *Nat1* and the entire *Nat2* coding region (Fakis *et al.*, 2000). Clone A and the ES cells used to target the *Nat2* gene were both derived from the 129/Ola mouse strain. Transfection of ES cells with isogenic constructs has been shown to greatly improve the rate of homologous recombination (Galli-Taliadoros *et al.*, 1995).

A previously generated restriction map of 129/Ola clone A (Fakis *et al.*, 2000) is shown in figure 4.2.3a. Restriction sites on this map were used to put together the different parts of the targeting construct. The first step involved isolation of a 8.6kb *Hind*III fragment, spanning the *Nat2* coding region (figure 4.2.3b). Digestion of 129/Ola clone A DNA with *Hind*III provided a mixture of two products of indistinguishable size on agarose gels. One was the 8.6kb fragment of interest and the other was a 7.2kb fragment, containing the 5′ end of 129/Ola clone A (4.3kb) and the entire cloning vector (2.9kb) (figure 4.2.3a). Cloning this mixture of products into the pBS vector was followed by screening of fifty transformed *E. coli* colonies by both PCR and restriction digestion (figure 4.2.4). The selected clone (clone 13) provided the upstream (6kb) and the downstream (2.5kb) homology arms of the *Nat2* targeting construct and served as the backbone for subsequent cloning of the positive and negative selection cassettes (figure 4.2.3c).

Insertion of the cassettes into the targeting construct was carried out by Dr Katalin Pinter (Department of Pharmacology, Oxford). The completed construct (figure 4.2.3c) was then sent to the Centre for Genome Research in Edinburgh, where it was introduced to 129/Ola ES cells.

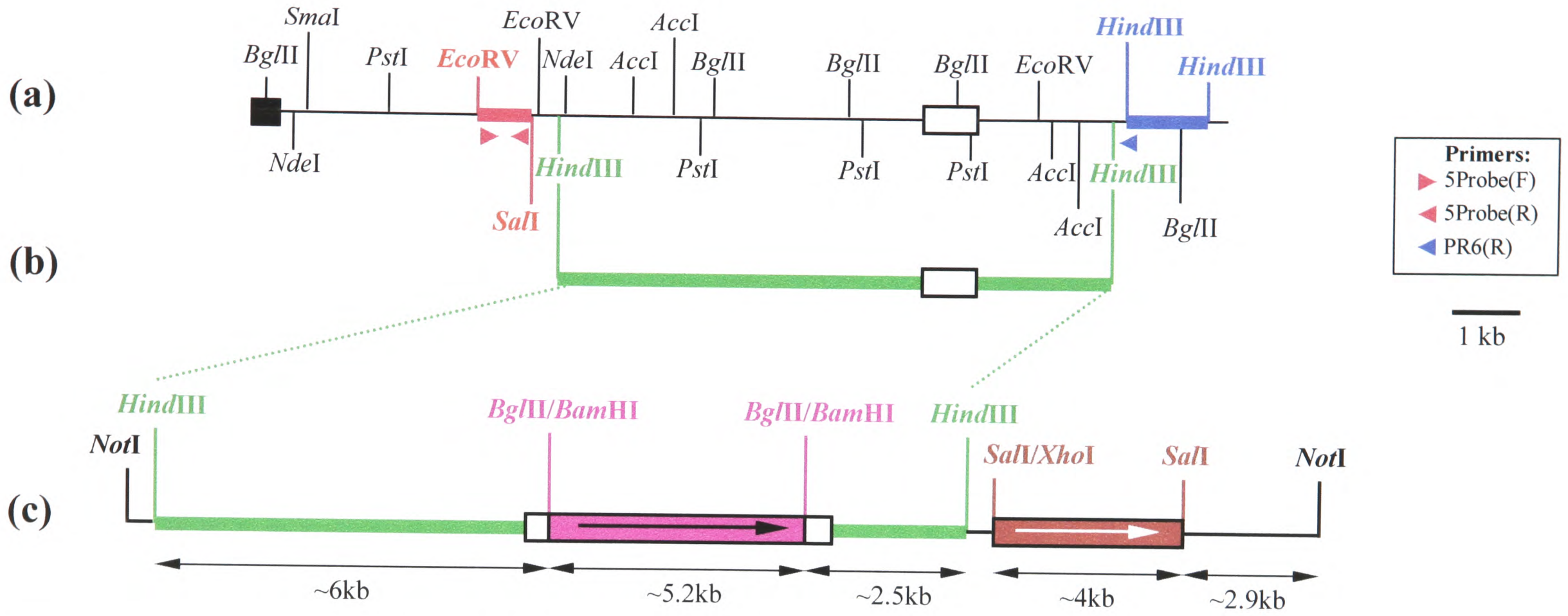


Figure 4.2.3: Summary of the cloning steps leading to the generation of a replacement targeting construct for murine *Nat2*. a) Previously generated restriction map of 129/Ola clone A (Fakis *et al.*, 2000). The clone is approximately 14.3kb in size and contains part of *Nat1* (black box) and the entire coding region of *Nat2* (white box). The relative position of the 5′ (*Eco*RV/*Sal*I fragment, coloured red) and the 3′ (*Hind*III fragment, coloured blue) flanking probes (section 4.2.1.3), is also indicated. b) Relative position of a 8.6kb *Hind*III fragment (thick green line) spanning the *Nat2* coding region (white box). The fragment was isolated from 129/Ola clone A and subcloned into the pBS vector (clone 13 in figure 4.2.4). It was then used to build the *Nat2* targeting construct. c) The complete *Nat2* replacement targeting construct, generated by Dr. Katalin Pinter (Department of Pharmacology, Oxford) using clone 13. The 5.2kb *Bam*HI positive selection cassette (pink box) was incorporated at the indicated compatible *Bgl*II site, disrupting the *Nat2* coding region (white boxes) at nucleotide position 533. The 4kb *Xho*I/*Sal*I negative selection cassette (bordeaux box) was incorporated at the compatible *Sal*I site of the 2.9kb pBS vector (thin black line). The coding sequences on the selection cassettes (figure 4.2.1) have the same 5′→3′ orientation (indicated by single-headed arrows) as the *Nat2* gene. The vector was linearised at the *Not*I site prior to electroporation into ES cells.

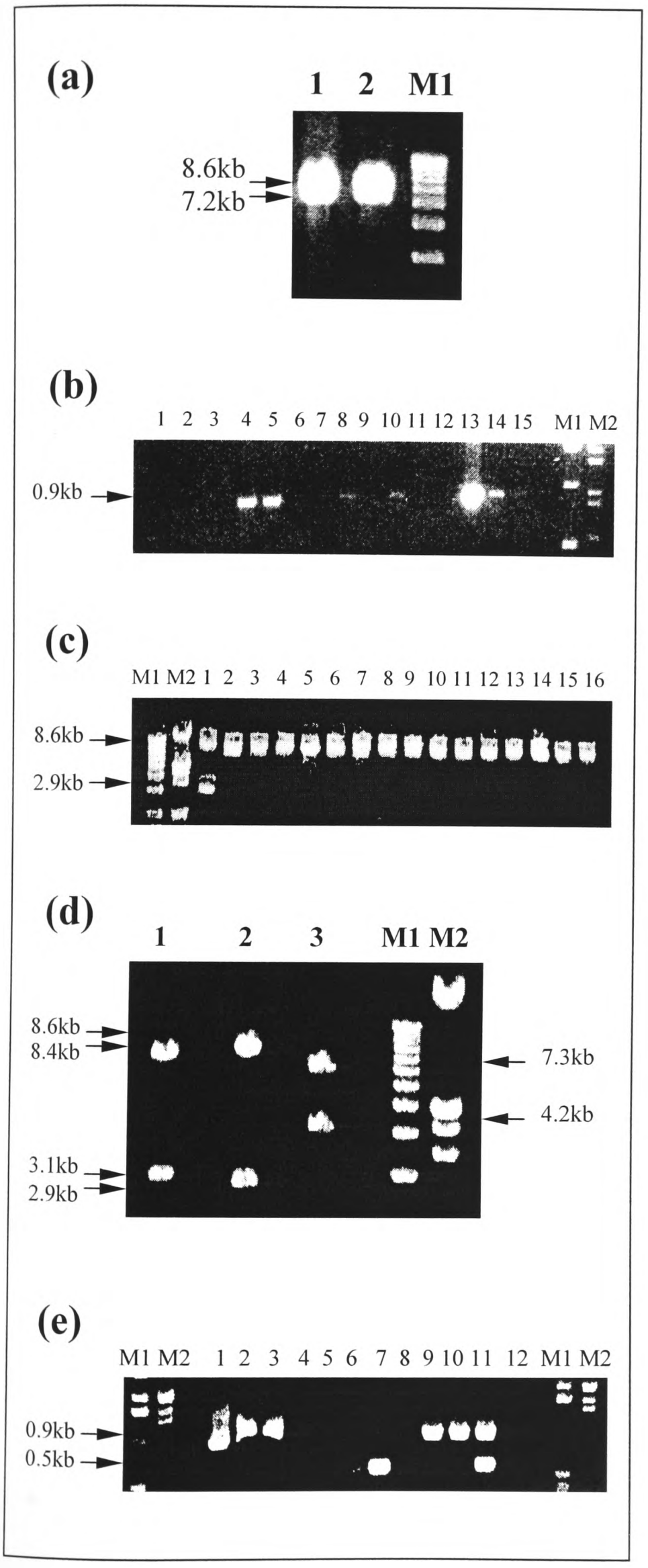


Figure 4.2.4: Subcloning of the 8.6kb *Hind*III fragment. a) Plasmid DNA from 129/Ola clone A was digested with *Hind*III and two fragments, 8.6kb and 7.2kb in size, were co-purified and subcloned into the pBS vector. b) White transformed *E. coli* colonies were screened by PCR, using the *Nat2*-specific primers mNAT2-1/mNAT2-910 (table 3.2.1). Clones 4, 5, 13 and 14 provided the expected 0.9kb product. c) Plasmid DNA from 16 positive colonies was digested with *Hind*III. Only clone 13 (lane 1) contains the correct 8.6kb insert. d) The presence of an intact 8.6kb insert was further confirmed by digestion of clone 13 with *Sal*I+*Nde*I (lane 1) and *Eco*RV (lane 3). A *Hind*III-digestion is also shown (lane 2) for comparison. e) Clone 13 was checked for the presence of only the desired 8.6kb insert by PCR with primers mNAT2-1/mNAT2-910. These provided the expected 0.9kb product (lanes 1 and 2). Primers 5Probe(F)/5Probe(R) (table 4.2.1 and figure 4.2.3a), specific for the 7.2kb insert described in (a), did not provide any product (lanes 5 and 6). PCR with both sets of primers (lanes 9 and 10) provided only the 0.9kb product, expected with primers mNAT2-1/mNAT2-910. Lanes 3, 7 and 11 are PCR products from 129/Ola clone A DNA, used as positive control. Lanes 4, 8 and 12 are the corresponding PCR negative controls. Lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively.

Table 4.2.1: PCR primers used for screening procedures during the generation of a *Nat2* knock-out mouse. Primers 5Probe(F) and 5Probe(R) amplify a 548bp region on 129/Ola clone A, between the *Eco*RV and *Sal*I restriction sites coloured red in figure 4.2.3a. Primers IRES-1(R) and Neo-T bind to the positive selection cassette, at the IRES element and the *neo*^{r} gene, respectively. Primer PR6(R) anneals to 129/Ola clone A, just downstream of the homology region of the construct, coloured blue in figure 4.2.3a.

Primer name	Orientation	Sequence (5′→3′)	Tm (°C)
5Probe(F)	forward	AAGCCCTACAACTACATTTCCC	61.2
5Probe(R)	reverse	AAGAGTCCCAGAGGGAACAA	60
IRES-1(R)	reverse	AGAGGGGCGGAATTCCTCTAG	62
Neo-T	forward	CATCGCCTTCTATCGCCTTCT	60
PR6(R)	reverse	CTTGTCACAGTTCTTTTATCTGCAT	58.5

4.2.1.3 Generation of probes flanking the homology region of the *Nat2* targeting construct

Two flanking probes were generated for Southern blotting screening of transfected ES cells for homologous recombination between the endogenous *Nat2* locus and the introduced targeting construct. The 5′ flanking probe was isolated from 129/Ola clone A as a 746bp *Eco*RV/*Sal*I fragment, and the 3′ flanking probe as a 986bp *Hind*III fragment (figure 4.2.5a). The relative position of the two probes on 129/Ola clone A is shown in figure 4.2.3a. Both restriction fragments were cloned into the T-tailed pGEM-T Easy vector (Appendix 1.1a), to provide an unlimited source of probe for subsequent Southern blot analyses. Successful cloning of the probes was confirmed by digestion of plasmid DNA with *Eco*RI, cutting the pGEM-T Easy vector on both sides of the insert (figure 4.2.5b). The probes were also sequenced from both ends using the vector-specific primers M13F and M13R (table 2.1). Standard GeneBank search, using BLAST, confirmed specificity of the 5′ flanking probe to the region just outside the upstream homology arm of the *Nat2* targeting construct (figure 4.2.3a). The cloned probe was subsequently sent to the Centre for Genome Research in Edinburgh for use in Southern hybridisations. In

contrast, the 3′ flanking probe was identified to contain a number of repeat elements [e.g. T-cell receptor (TCR) α- and β-like sequences, and LTRs] and was, therefore, excluded from use in Southern hybridisations. However, the sequence of the 3′ flanking probe served as a starting point for sequencing the entire 129/Ola clone A, as will be described in section 4.2.2.

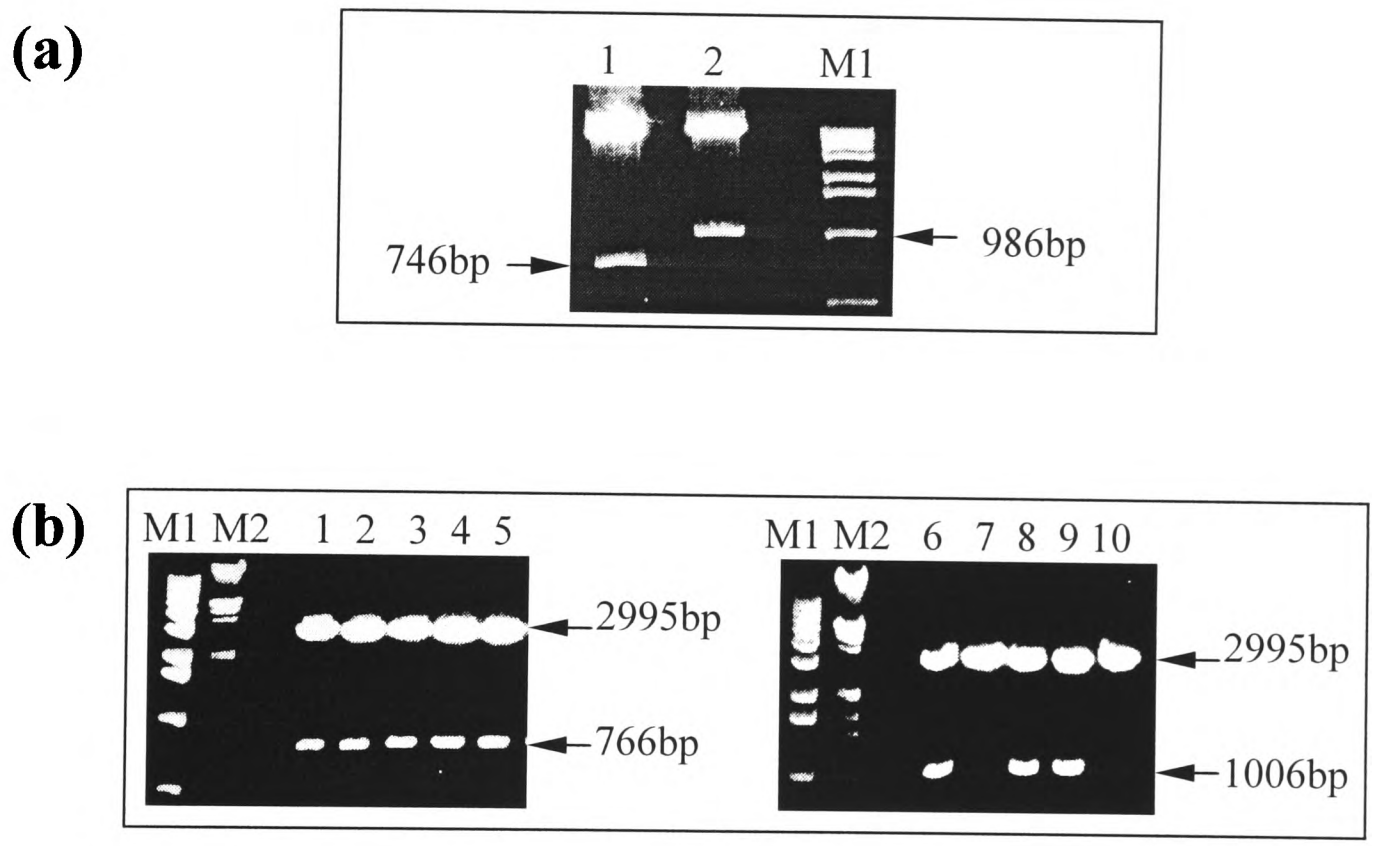


Figure 4.2.5: Cloning of probes flanking the homology region of the *Nat2* targeting construct. a) The 5′ and 3′ flanking probes were isolated from 129/Ola clone A DNA by digestion with *Eco*RV+*Sal*I (lane 1) and *Hind*III (lane 2), respectively. The generated 746bp (lane 1) and 986bp (lane 2) fragments were gel-purified (section 2.2.1.3), blunt-ended and A-tailed (section 2.2.3.7), to allow ligation (section 2.2.3.9) into the pGEM-T Easy vector (Appendix 1.1a). Transformed JM109 High Efficiency Competent Cells (*E. coli*) were grown on LB-agar plates containing 100μg/ml ampicillin, 0.5mM IPTG and 40μg/ml x-gal (section 2.2.3.10). b) Plasmid DNA was extracted from selected white colonies (section 2.2.1.2) and the presence of the correct size insert was checked by digestion with *Eco*RI, cutting the pGEM-T Easy vector on either side of the insert (Appendix 1.1a). Lanes 1-5 are *Eco*RI-digested clones expected to carry the 5′ flanking probe. All clones contain the correct size insert. Lanes 6-10 are *Eco*RI-digested clones expected to carry the 3′ flanking probe. Only clones 6, 8 and 9 contain the correct insert. Lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively.

4.2.1.4 Screening of ES cells for homologous recombination by long and accurate PCR (LA-PCR)

Gene targeting events, identified after initial screening of ES cell clones by Southern hybridisation, were further confirmed by PCR. This was especially important in order to ensure targeted incorporation of the short (2.5kb) downstream homology arm of the construct, for which a 3′ flanking probe was not available (section 4.2.1.3). The adopted PCR method (Herrup, 1994; Plagge *et al.*, 2000) used one primer specific for the positive selection cassette and a second primer binding to the genomic region just outside the boundary of the construct. An outline of the screening strategy is provided in figure 4.2.6.

Due to the large size of the amplified fragments (about 7.2kb for the upstream homology arm and 3kb for the downstream homology arm), it was necessary to optimise a LA-PCR protocol to increase the efficiency of reactions (Barnes, 1994; Cheng *et al.*, 1994; Cheng, 1995). Amplification was carried out with a mixture of *Taq* and *Pfu*-DNA polymerases, to combine the efficiency of the former with the proofreading activity of the latter (Barnes, 1994). Reaction conditions were optimised as described in figure 4.2.7. Optimal amplification was with a 10:1 mixture of *Taq:Pfu*-polymerases, in *Taq*-polymerase buffer supplemented with 1.5-2.25mM $MgCl_2$ and 0.01% (w/v) gelatin (section 2.2.3.4).

Targeted replacement of the endogenous *Nat2* locus was detected in a total of 11 ES cell clones by Southern blot analysis. DNA from these clones was provided by the Centre for Genome Research and subjected to LA-PCR screening. Amplification of the homologous regions flanking the endogenous *Nat2* locus was also performed as a positive control. Targeted incorporation of the downstream homology arm was confirmed for 6 of the 11 examined clones (figure 4.2.8). Correct incorporation of the upstream homology arm was then confirmed for 3 of these clones, namely 2/C3, 2/E5 and 2/G3 (figure 4.2.9).

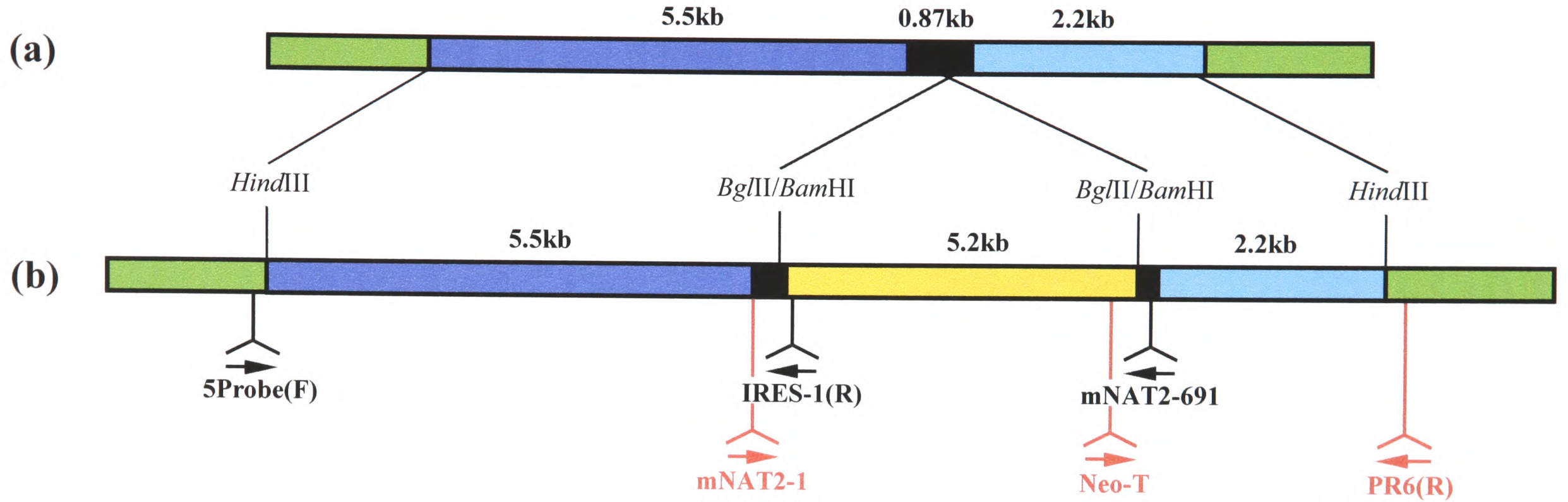


Figure 4.2.6: Strategy for LA-PCR screening of 129/Ola ES cells for targeted incorporation of the *Nat2* construct. The diagram shows the structure of the *Nat2* locus before (a) and after (b) homologous recombination with the *Nat2* targeting construct. Genomic DNA flanking the region of homology is coloured green. The upstream and downstream homology arms are shown in dark and light blue, respectively. The positive selection cassette is coloured yellow, while black is the *Nat2* coding region. Two screening strategies were adopted: i) Screening for targeted insertion of the upstream homology arm of the construct with primers coloured black. Primer pair 5Probe(F)/mNAT2-691 amplifies only from the endogenous *Nat2* locus, while pair 5Probe(F)/IRES-1(R) amplifies from the targeted *Nat2* locus. ii) Screening for targeted insertion of the downstream homology arm of the construct with primers coloured red. Primer pair mNAT2-1/PR6(R) amplifies only from the endogenous *Nat2* locus, while pair Neo-T/PR6(R) amplifies from the targeted *Nat2* locus. Targeted incorporation of the construct should result in successful amplification with all four sets of primers, as the transfected ES cells are heterozygous for both the endogenous and the recombinant *Nat2* locus.

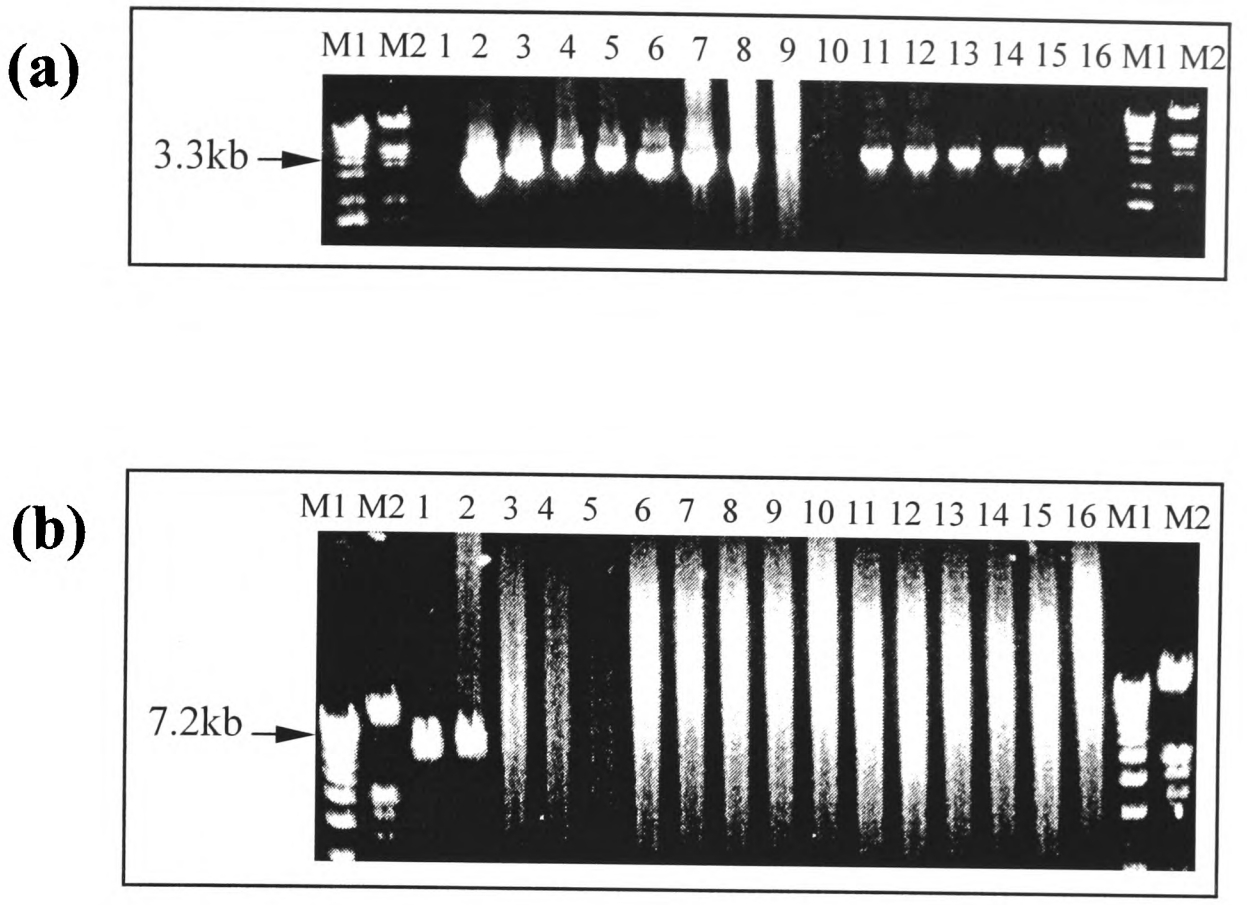


Figure 4.2.7: Optimisation of LA-PCR. a) Amplification of the 3.3kb region between primers mNAT2-1 and PR6(R) (figure 4.2.6), using a 10:1 ratio of *Taq:Pfu*-DNA polymerases. Lanes 1-5: Reactions in *Taq*-buffer with 0.75mM (lane 1), 1.5mM (lane 2), 2.25mM (lane 3), 3mM (lane 4) or 3.75mM (lane 5) $MgCl_2$. Lanes 6-10: Reactions in *Pfu*-buffer (containing 2mM $MgSO_4$), without extra $MgCl_2$ (lane 6) or with 0.25mM (lane 7), 0.5mM (lane 8), 1mM (lane 9) or 1.75mM (lane 10) $MgCl_2$. In lanes 1-10, the template was 0.5µg of genomic DNA. Lanes 11-15: Reactions in *Taq*-buffer with 2.25mM $MgCl_2$. The template was 0.1µg (lane 11), 0.25µg (lane 12), 0.5µg (lane 13), 0.75µg (lane 14) or 1µg (lane 15) of genomic DNA. Lane 16 is the PCR negative control. b) Amplification of the 7.2kb region between primers 5Probe(F) and mNAT2-691 (figure 4.2.6). Lanes 1-5: Reactions in *Taq*-buffer with 0.75mM (lane 1), 1.5mM (lane 2), 2.25mM (lane 3), 3mM (lane 4) or 3.75mM (lane 5) $MgCl_2$. A 10:1 ratio of *Taq:Pfu*-DNA polymerases was used. Lanes 6-10: Reactions in *Pfu*-buffer, using 10:1 (lane 6), 5:1 (lane 7), 1:1 (lane 8), 1:2 (lane 9) or 1:5 (lane 10) ratios of *Taq:Pfu*-DNA polymerases. In lanes 1-10, the template was 0.5µg of genomic DNA. Lanes 11-15: Reactions in *Taq*-buffer with 2.25mM $MgCl_2$. The template was 0.1µg (lane 11), 0.25µg (lane 12), 0.5µg (lane 13), 0.75µg (lane 14) or 1µg (lane 15) of genomic DNA. Lane 16 is the PCR negative control. A 10:1 ratio of *Taq:Pfu*-DNA polymerases was used. The observed smearing (even in the negative control) is a common feature of unsuccessful LA-PCR reactions (Hengen, 1994). Lanes M1 and M2 are 1µg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively.

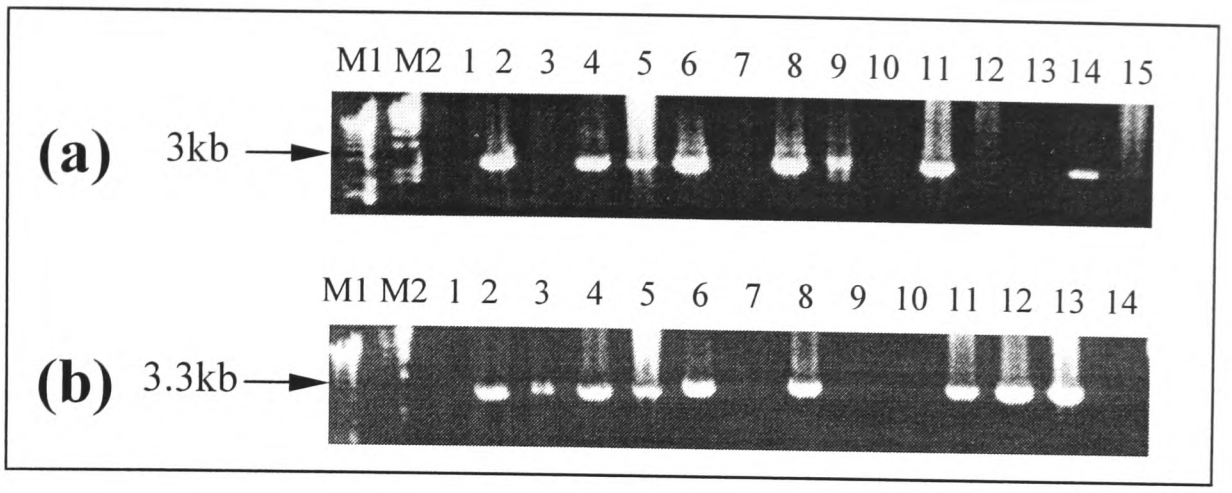


Figure 4.2.8: LA-PCR screening of 129/Ola ES cells for targeted insertion of the downstream homology arm of the *Nat2* construct. a) Screening of clones 2/B5 (lane 1), 2/C3 (lane 2), 2/C4 (lane 3), 2/C9 (lane 4), 2/C10 (lane 5), 2/D6 (lane 6), 2/D11 (lane 7), 2/E5 (lane 8), 2/E6 (lane 9), 2/F3 (lane 10) and 2/G3 (lane 11) with primers Neo-T and PR6(R) (figure 4.2.6). Lane 12 shows amplification from non-transfected 129/Ola ES cells with primers Neo-T and PR6(R), while lane 14 shows amplification from the targeting construct with primers Neo-T and M13F. Lanes 13 and 15 are PCR negative controls for primer sets Neo-T/PR6(R) and Neo-T/M13F, respectively. b) PCR with primers mNAT2-1 and PR6(R) (figure 4.2.6). The order of the clones in lanes 1-11 is the same as in (a). Lanes 12 and 13 show amplification from non-transfected 129/Ola ES cells, while lane 14 is the PCR negative control. Lanes M1 and M2 are 1μg of 1kb and λ/*Hind*III size markers, respectively. Targeted insertion of the downstream homology arm was confirmed for clones 2/C3, 2/C9, 2/C10, 2/D6, 2/E5 and 2/G3.

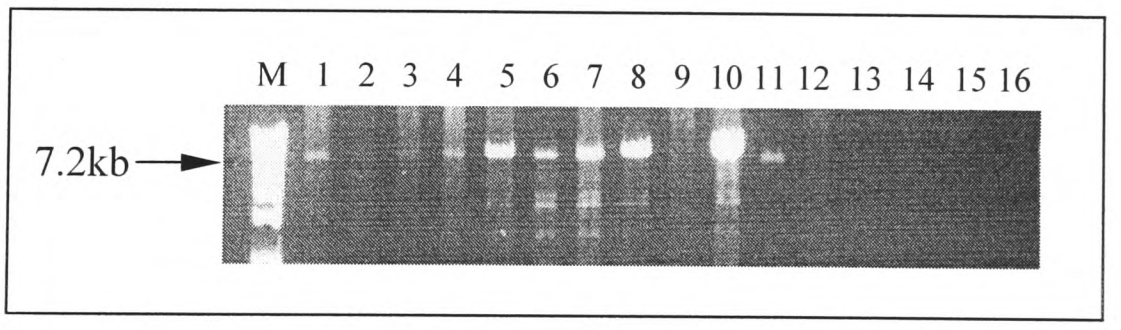


Figure 4.2.9: LA-PCR screening of 129/Ola ES cells for targeted insertion of the upstream homology arm of the *Nat2* construct. Screening of clones 2/C3, 2/C9, 2/E5 and 2/G3 with primer pairs 5Probe(F)/IRES-1(R) (lanes 1-4) and 5Probe(F)/mNAT2-691 (lanes 5-8). Lanes 9 and 10 show amplification of non-transfected 129/Ola ES cells with primer pairs 5Probe(F)/IRES-1(R) and 5Probe(F)/mNAT2-691, respectively. Lanes 11 and 12 show amplification from the targeting construct with primer pairs M13R/IRES-1(R) and M13R/mNAT2-691, respectively. Lanes 13-16 are PCR negative controls for the four sets of primers. Lane M is 1μg of 1kb Plus DNA ladder. Targeted insertion of the upstream homology arm was confirmed for clones 2/C3, 2/E5 and 2/G3.

4.2.2 Sequencing analysis of 129/Ola clone A

4.2.2.1 Sequencing strategy

Sequencing of the 14.3kb 129/Ola clone A was undertaken to facilitate the generation of the *Nat2* targeting construct, as well as Southern blot analysis of the transformed 129/Ola ES cells. A "primer walking" strategy was adopted (figure 4.2.10), to produce a series of partially overlapping pieces of novel sequence, starting from sites of previously known sequence (major starting points). The primers used in sequencing reactions (section 2.2.3.5) are listed in table 4.2.2.

As a first step, the ends of both 129/Ola clone A and clone 13 (section 4.2.1.2) were sequenced with the vector-specific primers M13F and M13R (figure 4.2.10). The sequences of the previously analysed 5′ and 3′ flanking probes (section 4.2.1.3) were used as additional major starting points (6 and 22 in figure 4.2.10, respectively). Reactions were also initiated from the 3′ end of the *Nat1* coding region, to the forward direction (major starting point 1 in figure 4.2.10), as well as from the *Nat2* coding region, towards opposite directions (major starting point 18 in figure 4.2.10). A short gap between the end of the 5′ flanking probe and the beginning of clone 13 was bridged with a sequencing reaction initiated from within clone 13 (starting point 7 in figure 4.2.10) and proceeding to the reverse direction. Sequencing of the region between two major starting points was complete when reactions proceeding from opposite directions provided sufficient sequence overlap.

Sequencing beyond starting points 9 and 13 has not been possible using plasmid DNA from either 129/Ola clone A or clone 13, possibly due to increased secondary structure of the template. To overcome this problem, the 3.4kb region between the two starting points was amplified by LA-PCR, using a pair of flanking primers (figure 4.2.11a). The amplified fragment was then A-tailed and cloned into the pGEM-T Easy vector. Successful incorporation of the insert was confirmed by digestion of plasmid DNA with *Eco*RI and *Bgl*II (figure 4.2.11b). A positive clone (UR) was then sequenced from both ends with the vector-specific primers M13F and M13R (figure 4.2.10). The rest of the clone was finally sequenced by "primer walking" from opposite directions (major starting points 9 and 13 in figure 4.2.10), until sequence overlap was established.

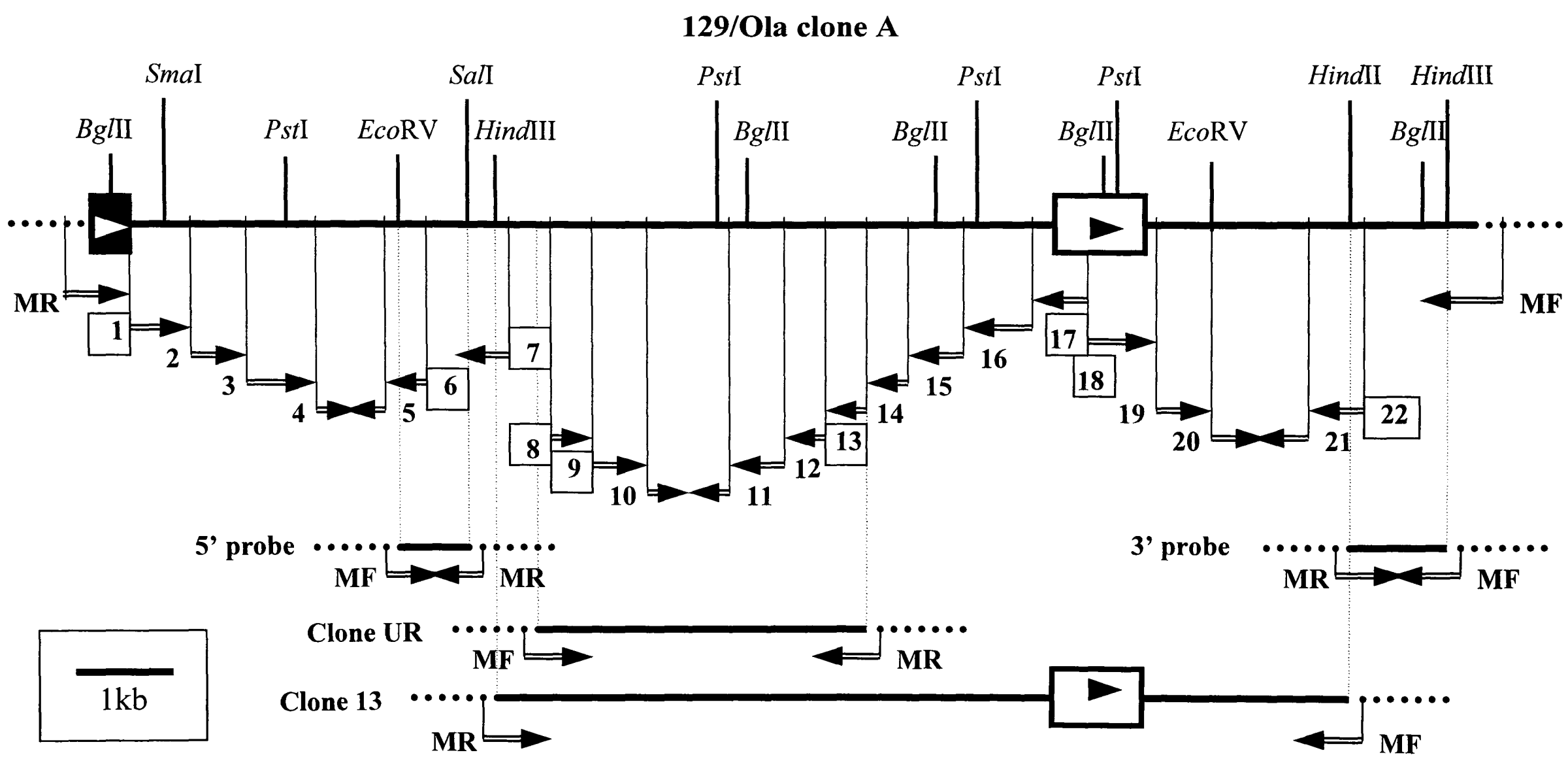


Figure 4.2.10: Strategy for sequencing the 14.3kb 129/Ola clone A. A partial restriction map of 129/Ola clone A (Fakis *et al.*, 2000) is shown in the top half of the figure, with the *Nat1* and *Nat2* coding region represented by a black and a white box, respectively. Arrowheads within the boxes indicate the 5′→3′ orientation of the *Nat* genes. The starting points of sequencing reactions, i.e. the sites of primer annealing, are numbered 1-22 and their relative position on 129/Ola clone A is indicated with thin vertical lines. Boxed numbers correspond to major starting points, i.e. primer annealing sites within regions of previously known sequence. The arrows show the direction of each reaction and the distance covered on 129/Ola clone A. The previously generated clone 13 (section 4.2.1.2), as well as the cloned 5′ and 3′ flanking probes (section 4.2.1.3), are shown in the bottom half of the figure. Their position relative to 129/Ola clone A is indicated with dotted vertical lines. Also shown is the relative position of clone UR, used to bridge a gap between starting points 9 and 13 (figure 4.2.11). The ends of all five clones were initially sequenced from the vector (flanking dotted lines), using the vector-specific primers M13F (MF) and M13R (MR).

Primer Name	Orientation	Sequence (5′→3′)	Tm (°C)	Specificity
PR1(F)	forward	560- TGACTTGGTGTATCTTCTATTTCC -583	58	Starting point 1
PR1.1(F)	forward	1147- CTACAATATATGTTACCTGAAAAA -1170	53	Starting point 2
PR1.2(F)	forward	1739- TGTAAATGTATGGCTTGTGCC -1759	56	Starting point 3
PR1.3(F)	forward	2321- TATTCCTCTCAGTCCTCGTCT -2341	58	Starting point 4
PR2.1(R)	reverse	2940- AATGCTTGCTGTTTCTTAGC -2921	53.5	Starting point 5
PR2(R)	reverse	3405- CGCCCCTTATGCTTAGTTT -3387	54.5	Starting point 6
PR-A	reverse	4464- ATCCAAGGCTGCGAGTAAA -4446	54.5	Starting point 7
PR-B	forward	4834- TGTTGGCTTCTTGTTCACA -4852	52.5	Starting point 8
UR-1(F)	forward	5111- CTTCAGAAAGCAGCAAGACT -5130	55.5	Starting point 9
UR-2(F)	forward	5617- ACAGGATTATCACAAGCTCAC -5637	60	Starting point 10
UR-3(R)	reverse	6580- GGCATCTTCTGACTCTCATCTA -6559	58.5	Starting point 11
UR-2(R)	reverse	7143- ATAACCGGAACCCTGTTTAC -7124	55.5	Starting point 12
UR-1(R)	reverse	7710- CATTGTGCTTTCTGCAATGA -7691	53.5	Starting point 13
PR4.3(R)	reverse	8140- AAAGTGGTATGTGCAGTGGC -8121	57.5	Starting point 14
PR4.2(R)	reverse	8466- AACATTACATTTCTTGAAAGC -8446	50	Starting point 15
PR4.1(R)	reverse	8984- CATACCTGGGTATTCTGTGACTAA -8961	59.5	Starting point 16
PR4(R)	reverse	9707- CCACGATTATATGAACTGACTACAA- 9683	58.5	Starting point 17
BstEII(R)	reverse	10266- AGATGTTAATTCCAGAGGC -10248	52.5	Starting point 18
BstEII(F)	forward	10144- AATAAGTACAGCAGTGGCATGA -10165	57	Starting point 18
PR5(F)	forward	10923- CAACCAGCATTATCCATATCAA -10944	55	Starting point 19
PR5.1(F)	forward	11552- TGTGACCAGATGCTTTATG -11570	52.5	Starting point 20
PR6.1(R)	reverse	12388- GAGGTAGAGAAGAGTCAACATT -12367	57	Starting point 21
PR6(R)	reverse	13216- CTTGTCACAGTTCTTTTATCTGCAT -13192	58.5	Starting point 22
PR1-gap	forward	428- GGTGAACTAGTTTTTACGATTTAG -451	54.5	Just upstream of starting point 1
CORR-1	forward	4740- CCCTAATGGTGAATCCTCCT -4759	57.5	Just upstream of starting point 8
PR3(F)	forward	5187- TATTGCAGTTCTATGGAATG -5206	51.5	Just downstream of starting point 9
PR4.4(R)	reverse	7826- CTAGGAGATGAGCAAGCATACC -7805	60.5	Just downstream of starting point 13
CORR-2	forward	8388- CCGTGTTTGTGATGTTCTGT -8407	55.5	Just upstream of starting point 15
PR5-gap	forward	10818- ATGTGTGCTAGGAAATACCC -10837	55.5	Just upstream of starting point 19
PR6-gap	forward	13030- AAATGTATATAGCCCTCCAT -13049	51.5	Just upstream of starting point 22

Table 4.2.2: Primers used for sequencing of 129/Ola clone A. The exact position of each primer annealing site was determined following sequencing of the entire 129/Ola clone A (Appendix 2) and is provided relative to the beginning of the clone. The numbering of sequencing starting points is the same as in figure 4.2.10. Primers PR1(F) to PR6(R) were used to generate the bulk of the sequence, while primers PR-1gap to PR6-gap were used to fill in short gaps of sequence or check regions of uncertainty.

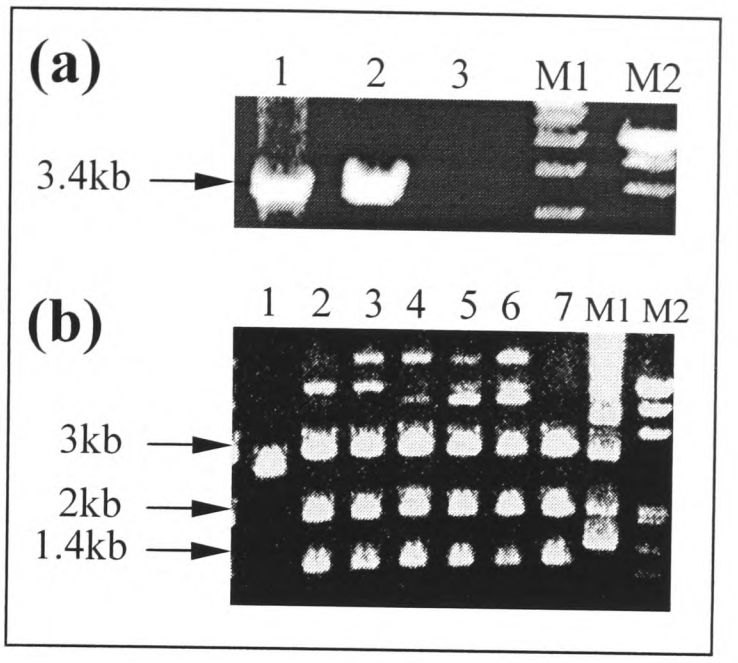


Figure 4.2.11: Amplification and cloning of a region difficult to sequence directly from 129/Ola clone A. a) The region of 129/Ola clone A between starting points 9 and 13 (figure 4.2.10) was amplified by LA-PCR (section 2.2.3.4), using flanking primers CORR-1 and PR4.3(R) (table 4.2.2). Lanes 1 and 2 are the 3.4kb amplification product, while lane 3 is the reaction negative control. b) The LA-PCR product was gel-purified (section 2.2.1.3), A-tailed (section 2.2.3.7) and ligated (section 2.2.3.9) to the pGEM-T Easy vector. Plasmid DNA was isolated from seven white colonies of transformed *E. coli* cells, and digested with *Eco*RI, which cuts the vector on either side of the insert (Appendix 1.1a), and *Bgl*II, which cuts within the insert (figure 4.2.10). In lanes 1-7, the 2kb and 1.4kb bands are the two portions of the *Bgl*II-digested insert, while the 3kb band is the vector. The larger bands correspond to undigested or partially digested plasmid. All clones, apart from that shown in lane 1, contain the right insert. The clone in lane 7 (named UR) was used for sequencing (figure 4.2.10). Lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively.

The assembled 14357bp sequence of 129/Ola clone A was deposited to the EMBL database (accession no. AJ250123). The complete sequence and restriction map of the clone is provided in Appendix 2. The exact distance between the end (stop codon) of the *Nat1* and the beginning (translation initiation codon) of the *Nat2* coding region is 9400 nucleotides. The portion of the *Nat1* gene contained in the clone is 451 nucleotides long and extends from nucleotide position 423 to the end of the coding region. The coding regions of the *Nat1* and *Nat2* genes have the same 5′→3′ orientation.

4.2.2.2 Computational analysis of the sequencing data

The sequence of 129/Ola clone A (Appendix 2) was subjected to computational analysis, using the NIX programme available from the HGMP (www.hgmp.mrc.ac.uk/Registered/Webapp/nix). NIX is a web-based tool designed to aid the identification of interesting features on DNA or RNA sequences submitted for analysis. This is achieved via BLAST search against many of the currently available databases (section 1.4.3), as well as by application of different prediction programmes. NIX also provides a convenient platform for viewing and comparing the results of different analyses. For simplicity purposes, the quality of predictions is referred to as "excellent", "good" or "marginal".

The results of the NIX analysis performed for the 14.3kb sequence of 129/Ola clone A are presented in figure 4.2.12. It was initially established by BLAST/ecoli and BLAST/vector database searches that the assembled sequence was devoid of contamination by vector or bacterial (*E. coli*) sequences.

Screening for sequences matching RNA-polymerase II promoters (GRAIL/polIIprom, TSSW/Promotor, Genscan/Prom and FGenes/Prom programmes) provided 11 candidate elements, 9 of which exhibited only "marginal" sequence similarity to the consensus. A "good" promoter was predicted by the GRAIL/polIIprom programme at position 9221 of the reverse DNA strand (position 5136 of the forward strand), while the TSSW/promoter programme predicted an "excellent" promoter at position 3378 of the forward strand (enclosed in a green circle in figure 4.2.12). No CpG islands were predicted on either DNA strand of 129/Ola clone A by the GRAIL/cpg programme.

Several polyadenylation signals were predicted by the GENSCAN/polya, FGenes/polya and GRAIL/polya programmes. Two of them (one predicted twice) were found on the forward DNA strand of 129/Ola clone A. One is located downstream of the *Nat1* and the other downstream of the *Nat2* coding region. Additional putative signals were found on the reverse strand of 129/Ola clone A.

Different exon-prediction programmes (Fex, Hexon, MZEF, Genemark, and GRAIL/exons) identified a total of 11 candidate exons on the forward DNA strand, those corresponding to the coding region of the *Nat1* and *Nat2* genes (enclosed in

white boxes in figure 4.2.12) providing the best quality of prediction. Of the candidate exons, one located at position 3445-3507 of 129/Ola clone A (enclosed in a purple box in figure 4.2.12) is of special interest, because it is adjacent to the above mentioned putative promoter at position 3378 and was predicted by more than one programme (Fex, Hexon and MZEF). At least 12 additional exons were predicted on the reverse strand of 129/Ola clone A. Despite the relatively large number of candidate exons, a variety of gene-finding programmes (GRAIL/gap2, Genefinder, FGene, GENSCAN, FGenes, and HMMGene) did not predict any genes on 129/Ola clone A, other than the already known *Nat1* and *Nat2* genes. It may be worthwhile investigating whether any of the numerous predicted exons could be non-coding exons of the two *Nat* genes.

The search of various databases further supported the prediction that *Nat1* and *Nat2* must be the only genes in the analysed region. Search of the TREMBL and SWISS-PROT databases (BLAST/trembl and BLAST/swissprot, respectively) identified the *Nat1* and *Nat2* coding regions as the only protein-coding sequences on 129/Ola clone A. Moreover, search of the EMBL databases (BLAST/unigene, BLAST/mRNA, BLAST/est and BLAST/embl-) provided numerous matches to the coding region of the two genes, including 54 genomic sequences and 16 ESTs from a variety of mammalian species (e.g. mouse, rat, hamster, rabbit and human).

The total number of BLAST/embl- and BLAST/est hits was 112 and 54, respectively. The mRNA sequences recovered by BLAST/mRNA search are identical to the genomic sequences previously described for the *Nat1* and *Nat2* genes by Martell *et al.* (1991) and are likely to represent incorrect database entries. The region at position 3337-3376 of 129/Ola clone A (Apendix 2) was identified as 90% homologous to the previously described upstream non-coding exon of the human *NAT2* gene (Ebisawa and Deguchi, 1991), while a nearby region (position 3512-3853 of 129/Ola clone A) provided a perfect match to an EST with accession no. AI006867 (enclosed in a yellow circle in figure 4.2.12). These regions are of special interest, as they are adjacent to the predicted promoter at position 3378 and could constitute an upstream non-coding exon for murine *Nat2.* The rest of the sequences recovered by search of the EMBL databases mostly aligned with the 3′ end of 129/Ola clone A, but the degree of homology was low ("marginal" or "good").

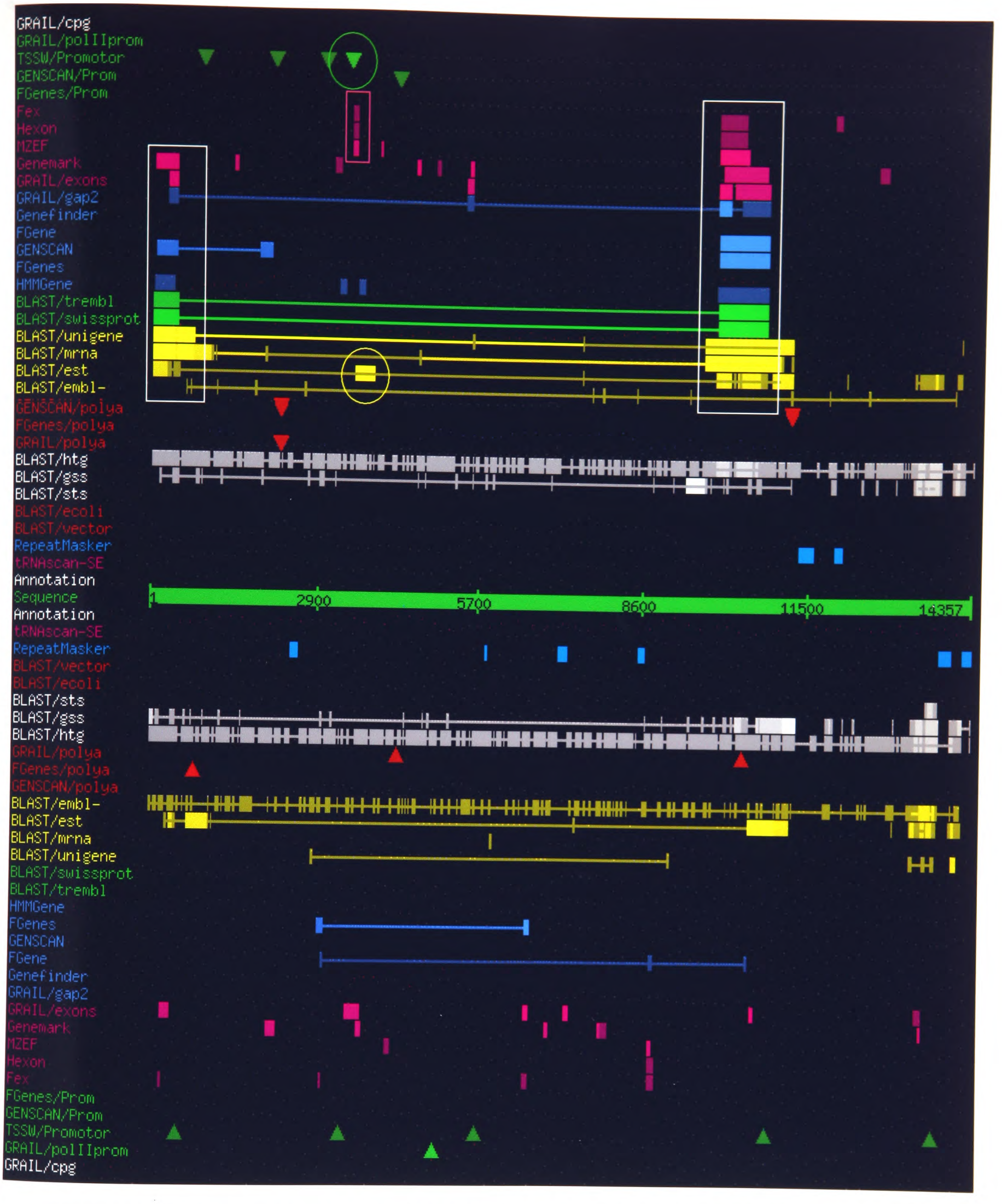


Figure 4.2.12: NIX analysis of the complete sequence of 129/Ola clone A. The 14.3kb sequence of the clone was submitted to HGMP (http://menu.hgmp.mrc.ac.uk/menu-bin/Nix/Nix.pl) for NIX analysis, the results of which were provided as a graphic overview. Detailed information about each of the predicted elements was recovered by "clicking" on the relevant feature. Information about the programmes and databases used for analysis was provided by "clicking" on the labels on the left-hand side. The results of programmes with similar functions (e.g. exon-prediction or EMBL-search) have been grouped together and illustrated with the same colour (e.g. the results of all gene-finding programmes are coloured blue). The shade of each colour indicates the quality of prediction; the brighter the shade, the better the prediction. The central green line represents the sequence of 129/Ola clone A. Features above or below this line are found on the forward or reverse DNA strand, respectively. Features that share common properties (e.g. a high degree of sequence similarity) are linked with a horizontal line. Features of special interest (described in the text) are enclosed in coloured boxes or circles.

BLAST search of sequences deposited to the HTG and GSS divisions of the EMBL database (section 1.4.3) provided a total of 402 matches, the bulk of which were short and aligned poorly with the sequence of 129/Ola clone A. Only two matches from the GSS database (positions 9330-9632 and 11099-11293) showed perfect homology to the analysed sequence. The data suggest that the sequencing information available from the public Mouse Genome Project for the region surrounding the *Nat* genes is currently very limited.

BLAST/sts search did not identify any previously mapped STSs on 129/Ola clone A. However, nine repeat elements were identified by RepeatMasker, which could serve as markers for accurate genetic and physical mapping of the mouse *Nat* locus. A description of these repeat elements is presented in table 4.2.3.

Table 4.2.3: Repeat elements predicted by NIX analysis of 129/Ola clone A. Screening of the analysed sequence against a library of repeat elements was performed by the RepeatMasker programme. A number of SINEs, LINEs and LTRs were identified on either the forward (+) or the reverse (-) DNA strand of 129/Ola clone A. The position of each element is provided relative to the beginning (forward direction) of the clone (Appendix 2).

Type of repeat element	DNA Strand	Position on 129/Ola clone A	Quality of prediction
SINE/Alu (B1F)	(-)	2431-2448	excellent
SINE/B4 (ID_B1)	(-)	2449-2560	excellent
SINE/B4 (RSINE1)	(-)	5825-5862	excellent
LINE/L1 (HAL1b)	(-)	7117-7279	excellent
SINE/B4 (RSINE1)	(-)	8525-8630	excellent
LTR/MaLR (MTE)	(+)	11322-11578	excellent
SINE/Alu (B1_MM)	(+)	11948-12077	excellent
LTR/ERVK (BGLII)	(-)	13800-13996	excellent
LTR/ERVK (MYSERV)	(-)	14189-14337	excellent

4.2.3 Towards physical localisation of the mouse *Nat* genes

4.2.3.1 Screening of mouse chromosome 8 YAC contigs for the *Nat* genes

In this section, preliminary work towards identification of the physical map position of the mouse *Nat* locus is described, aiming to open the way for future studies of the surrounding chromosomal region, identification of neighbouring genes and comparison with the syntenic 8p22 region harbouring the *NAT* locus in humans (section 1.2.2). The PCR screening described in the following sections (4.2.3 and 4.2.4) was carried out in collaboration with Naomi Price, under the supervision of the author.

A series of YAC contigs, spanning the region 15.7-41.5cM on mouse chromosome 8, were screened for the presence of the *Nat* genes, based on previous mapping of the *Nat* locus to the middle portion of chromosome 8, around the 31cM genetic position (Mattano *et al.*, 1988; Fakis *et al.*, 2000). The analysed contigs are part of the physical map recently constructed for the entire mouse genome by the Centre for Genome Research of the Whitehead Institute at MIT (Nusbaum *et al.*, 1999). The YAC clones of each contig have been aligned to the genetic map of the mouse genome (Rhodes *et al.*, 1998), using different types of STSs, mostly PCR-amplifiable SSLPs (Nusbaum *et al.*, 1999). The integrated cytogenetic, genetic and physical map of the analysed region on mouse chromosome 8 is shown in figure 4.2.13. Coverage by the YAC contigs is not continuous, with four gaps spanning regions 25.1-27.3cM, 28.4-29.5cM, 31.7-32.8cM and 32.8-33.9cM.

The individual clones comprising the 16 YAC contigs of interest were obtained from the HGMP-RC (www.hgmp.mrc.ac.uk) and screened for the presence of the *Nat1* gene by PCR with primers Mus12 and Mus13 (table 3.2.1). To test the reliability of the screening method, parallel amplifications of SSLP markers of known position on the analysed contigs were included in each set of reactions. Detailed information about the exact position of the contigs, the number of examined YAC clones and the control SSLP markers, is provided in table 4.2.4.

Figure 4.2.13: Integrated cytogenetic, genetic and physical mapping data for the mouse *Nat* genes. The cytogenetic banding pattern of murine chromosome 8 is shown on the left, with bands B3.1-3.3 harbouring the *Nat* locus (Fakis *et al.*, 2000) enclosed in a box. According to the 2000 Chromosome 8 Committee Report (www.informatics.jax.org/ccr), these bands align with the 33cM genetic position on chromosome 8 (middle). The *Nat* locus is estimated to lie nearby, at the 31cM position (Mattano *et al.*, 1988). YAC contigs WC8.11 to WC8.27, screened for the presence of the *Nat* locus, are represented by coloured vertical lines of variable length (far right). Contig WC8.15 was not included in this analysis, due to its small size and complete overlap with contig WC8.14. Each contig has been anchored to the genetic map of the 15.7-41.5cM region (graded yellow bar) via its first and last STS (same colour as the contig). Coverage of the analysed genomic region is not complete, with one 2.2cM and three 1.1cM gaps (coloured red on the genetic map) located between contigs WC8.12-WC8.13, WC8.13-WC8.14, WC8.16-WC8.17 and WC8.18-WC8.20. The figure was assembled, based on maps available from the MGD (www.informatics.jax.org).

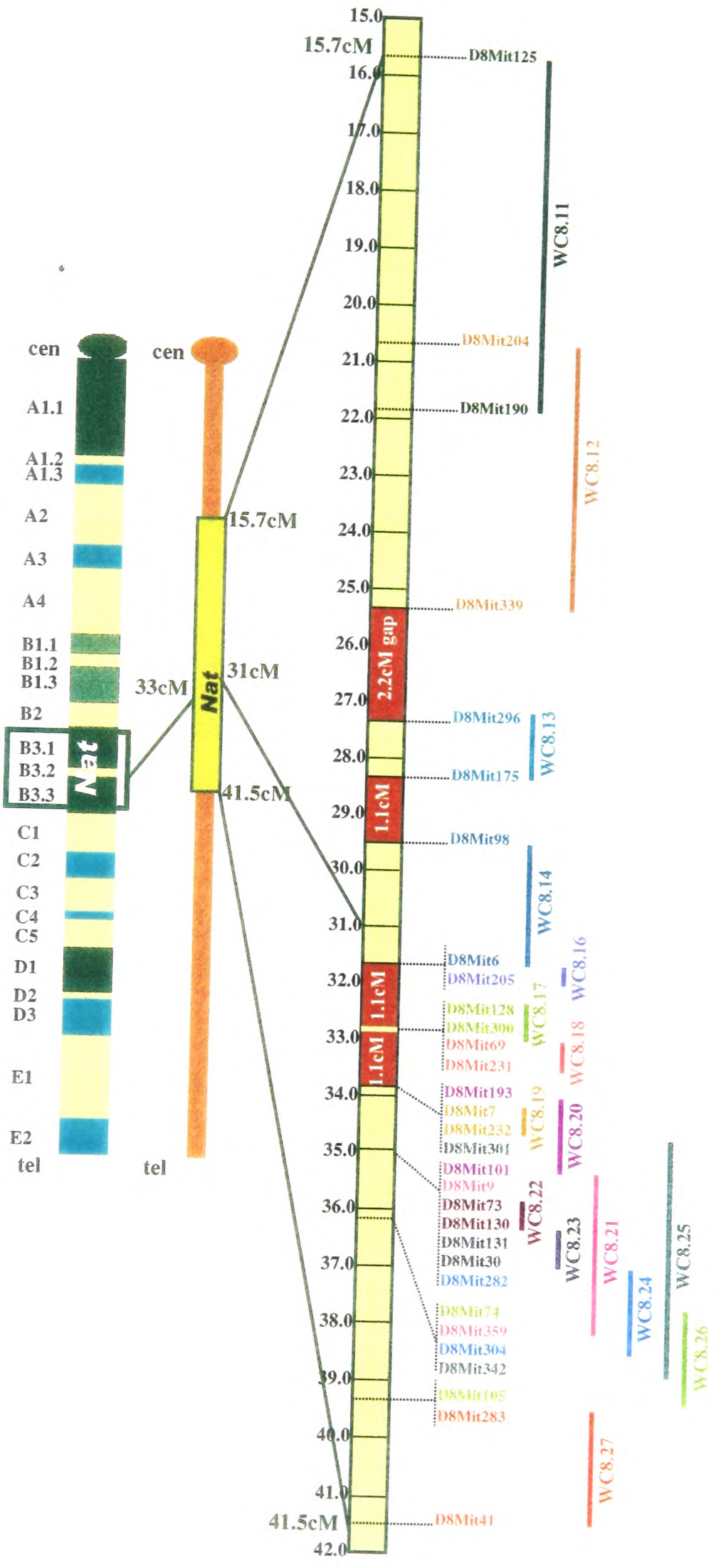

cen
A1.1
A1.2
A1.3
A2
A3
A4
B1.1
B1.2
B1.3
B2
B3.1
B3.2
B3.3
C1
C2
C3
C4
C5
D1
D2
D3
E1
E2
tel
cen
15.7cM
33cM
Nat
31cM
41.5cM
tel
15.0
15.7cM
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
27.0
28.0
29.0
30.0
31.0
32.0
33.0
34.0
35.0
36.0
37.0
38.0
39.0
40.0
41.0
41.5cM
42.0
2.2cM gap
1.1cM
1.1cM
D8Mit125
D8Mit204
D8Mit190
D8Mit339
D8Mit296
D8Mit175
D8Mit98
D8Mit6
D8Mit205
D8Mit128
D8Mit300
D8Mit69
D8Mit231
D8Mit193
D8Mit7
D8Mit232
D8Mit301
D8Mit101
D8Mit9
D8Mit73
D8Mit130
D8Mit131
D8Mit30
D8Mit282
D8Mit74
D8Mit359
D8Mit304
D8Mit342
D8Mit105
D8Mit283
D8Mit41
WC8.11
WC8.12
WC8.13
WC8.14
WC8.16
WC8.17
WC8.18
WC8.19
WC8.20
WC8.22
WC8.23
WC8.21
WC8.24
WC8.25
WC8.26
WC8.27

Table 4.2.4: Summary of PCR screening of the YAC contigs. The precise map position of each contig is provided, as well as the number of its constituent YAC clones. Of the 495 clones composing the 16 contigs, 290 were screened for the presence of the *Nat1* gene. When only a subset of clones were analysed from a specific contig, these were chosen to provide continuous overlap over the region covered by the complete contig. Amplification of selected SSLP markers of known position on the contigs was carried out from two positive and one negative YAC clone. The expected size of the PCR products is provided for the C57Bl/6J strain, whose DNA was used for the construction of the YAC clones. Detailed information about the structure of the YAC contigs, as well as the sequence of the primers used for SSLP amplification is available from the MGD (www.informatics.jax.org).

YAC Contig	Genetic map position (cM)	No. of clones (no. screened)	Control marker	Control YAC clones	PCR product size (bp)
WC8.11	15.7-21.9	19 (5)	n.u.*	—	—
WC8.12	20.8-25.1	95 (13)	n.u.	—	—
WC8.13	27.3-28.4	34 (34)	*D8Mit229*	432-E3 (+), 177-A11 (+), 324-H9 (-)	162
WC8.14	29.5-31.7	38 (38)	*D8Mit6*	124-G5 (+), 420-E2 (+), 178-B1 (-)	170
WC8.16	31.7-31.7	6 (5)	*D8Mit205*	456-F2 (+), 462-E11 (+), 328-G2 (-)	146
WC8.17	32.8-32.8	14 (13)	*D8Mit160*	153-D6 (+), 133-A3 (+), 453-E5 (-)	149
WC8.18	32.8-32.8	22 (22)	*D8Mit68*	135-G9 (+), 321-C3 (+), 293-H6 (-)	142
WC8.19	33.9-33.9	16 (16)	*D8Mit7*	398-B11 (+), 405-C3 (+), 458-H11 (-)	178
WC8.20	33.9-35.0	17 (17)	*D8Mit193*	311-F11 (+), 382-A5 (+), 187-C1 (-)	146
WC8.21	35.0-36.1	12 (12)	*D8Mit233*	308-D7 (+), 406-G8 (+), 136-D5 (-)	313
WC8.22	35.0-35.0	24 (24)	*D8Mit29*	445-F2 (+), 341-A6 (+), 426-G2 (-)	110
WC8.23	35.0-35.0	16 (16)	*D8Mit131*	151-B10 (+), 444-E12 (+), 295-A5 (-)	133
WC8.24	35.0-36.1	50 (46)	*D8Mit282*	172-D3 (+), 423-H12 (+), 361-D11 (-)	118
WC8.25	33.9-36.1	11 (11)	n.u.	—	—
WC8.26	36.1-39.3	79 (12)	n.u.	—	—
WC8.27	39.3-41.5	42 (6)	n.u.	—	—

**None used*

None of the 290 YAC clones examined was positive for the *Nat1* gene, although amplification of all control markers was successful (figure 4.2.14). The substantial overlap between the clones of each contig and the fact that the *Nat* locus is known to extend over less than 130kb of genomic sequence on mouse chromosome 8 (Fakis *et al.*, 2000) make the presence of the other two *Nat* genes on any of the

analysed clones extremely unlikely. It was, therefore, concluded that the *Nat* genes must be located within one of the above-mentioned gaps between the contigs.

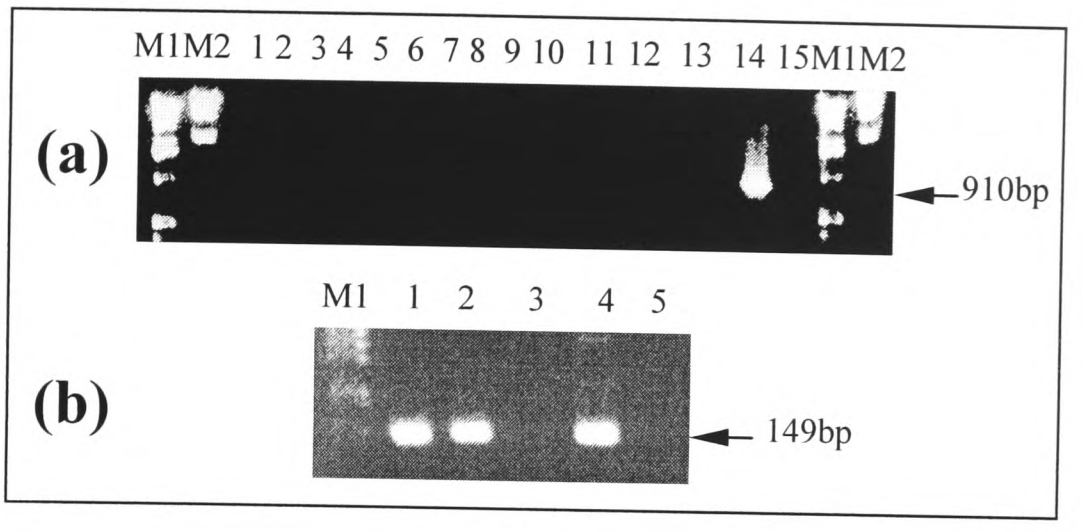


Figure 4.2.14: Screening of YAC contig WC8.17 for the presence of the murine *Nat1* gene. a) Thirteen YAC clones were used in PCR reactions with *Nat1*-specific primers Mus12 and Mus13 (table 3.2.1). Lanes 1 to 13 were loaded with amplification products from clones 153-D6, 133-A3, 188-C10, 415-A8, 120-D6, 432-H2, 377-B10, 453-E5, 430-A4, 376-F12, 430-B4, 362-G5 and 335-B5. Lane 14 shows amplification from C57Bl/6J genomic DNA (positive control), while lane 15 is the PCR negative control. None of the examined clones was positive for the *Nat1* gene. b) Amplification of SSLP marker *D8Mit160* (table 4.2.4) was performed as a PCR control with YAC templates. Lanes 1, 2 and 3 show products from two positive (153-D6 and 133-A3) and one negative (453-E5) clone, respectively. Lane 4 is the PCR product from C57Bl/6J genomic DNA and lane 5 is the reaction negative control. Products were analysed on a 3% (w/v) agarose gel. Lanes M1 and M2 are 1μg of 1kb and λ/*Hind*III size markers, respectively.

4.2.3.2 Screening of *Nat*-positive mouse YAC genomic clones for chromosome 8 specific markers

As an alternative approach towards physical localisation of the murine *Nat* genes, four previously identified *Nat*-positive YAC clones (Giannoulis Fakis, D.Phil. Thesis, Oxford 2000) were screened for the presence of markers of known genetic position on chromosome 8. These positive clones contain all three *Nat* genes and were isolated by PCR screening of the WI/MIT mouse YAC genomic library, created at the Whitehead Institute (MIT) with genomic DNA from the C57Bl/6J strain (Haldi *et al.*, 1996). The library consists of about 40,000 YAC clones, 24,000 of which have been

used for the construction of the YAC contigs described above (section 4.2.3.1). The average insert size of the clones is 820kb.

The *Nat*-positive YAC clones 107-A2, 110-G4, 156-A8 and 435-B7 were provided by HGMP-RC and screened by PCR with primers specific for 21 markers previously mapped within region 25.1-33.9cM on mouse chromosome 8 (table 4.2.5). The markers were chosen on the basis of their estimated physical map position near or within the gaps between the contigs (table 4.2.5). None of the selected markers is present on any of the four *Nat*-positive YAC clones. Additional markers need to be examined in order to establish co-localisation with the *Nat* genes.

Table 4.2.5: List of markers examined for co-localisation with the mouse *Nat* genes on the same YAC clone. All markers are of the SSLP type, apart from marker *Ms10b*. The established genetic and the estimated physical map position of each marker is provided, as well as the size of the corresponding PCR product from genomic DNA of the C57Bl/6J strain. Information about the markers and the primers used for their amplification is available from the MGD (www.informatics.jax.org).

Marker	Genetic map position (cM)	Estimated physical localisation	PCR product size (bp)
D8Mit339	25.1	End of WC8.12	123
Ms10b	26.0	Between WC8.12 & WC8.13	n.p.*
D8Mit227	26.2	Between WC8.12 & WC8.13	133
D8Mit5	27.3	Between WC8.12 & WC8.13	161
D8Mit296	27.3	Beginning of WC8.13	81
D8Mit175	28.4	End of WC8.13	129
D8Mit126	29.5	Between WC8.13 & WC8.14	147
D8Mit98	29.5	Beginning of WC8.14	148
D8Ertd268e	30.0	Between WC8.14 & WC8.16	n.p.
D8Ertd124e	31.5	Between WC8.14 & WC8.16	n.p.
D8Mit205	31.7	Within WC8.16	141
D8Ertd252e	32.0	Between WC8.16 & WC8.17	n.p.
D8Mit129	32.8	Between WC8.16 & WC8.17	111
D8Mit299	32.8	Between WC8.16 & WC8.17	114
D8Mit358	32.8	Between WC8.16 & WC8.17	120
D8Mit144	32.8	Between WC8.16 & WC8.17	148
D8Mit128	32.8	Beginning of WC8.17	146
D8Mit231	32.8	End of WC8.18	137
D8Mit8	33.9	Between WC8.18 & WC8.20	121
D8Mit70	33.9	Between WC8.18 & WC8.20	173
D8Mit193	33.9	Beginning of WC8.20	141

**Not provided by the MGD*

4.2.4 Identification of novel polymorphic microsatellite markers close to the mouse *Nat* genes

Sequencing analysis of 129/Ola clone A using the RepeatMasker programme (section 4.2.2.2) revealed a number of LINEs, SINEs and LTRs in the vicinity of the *Nat1* and *Nat2* genes (table 4.2.3). These elements are often associated with dinucleotide $(CA)_n$ or trinucleotide $(CAA)_n$ microsatellite repeats (Curran, 1997). Inspection of 129/Ola clone A sequence (Appendix 2) identified one trinucleotide repeat (n=6) located just downstream of SINE/B4 at position 2449-2560, one dinucleotide repeat (n=18) preceding SINE/B4 at position 5825-5862, one dinucleotide repeat (n=30) just downstream of LINE/L1 at position 7117-7279 and one mixed dinucleotide repeat (n=64) flanked by LTR/MaLR at position 11322-11578 and SINE/Alu at position 11948-12077. The first three of these microsatellite loci are located between the *Nat1* and *Nat2* coding regions, while the fourth is located downstream of *Nat2*. The loci were named *D8Srb1*, *D8Srb2*, *D8Srb3* and *D8Srb4*, according to the specifications of the International Committee on Standardised Genetic Nomenclature for Mice (www.informatics.jax.org/mgihome/nomen/gene.shtml). The symbol *Srb* was registered as a laboratory code via the Institute for Laboratory Animal Research of the National Academy of Sciences, USA (www4.nas.edu/cls/ilarhome.nsf). The exact position of the four loci on 129/Ola clone A is shown in figure 4.2.15.

Amplification across each microsatellite locus with specially designed flanking primers (figure 4.2.15 and table 4.2.6) revealed SSLP among 129/Ola, A/J, CBA, C57Bl/6J, Balb/c, PO, TO, MSP and MCA mouse strains (figure 4.2.16 and table 4.2.7). The wild-derived inbred strains MSP and MCA exhibited the highest degree of polymorphism, providing distinct alleles for all four loci (table 4.2.7). The inbred CBA strain provided a distinct *D8Srb2* allele ($D8Srb2^b$), while the outbred TO strain provided a distinct *D8Srb3* allele ($D8Srb3^b$) (table 4.2.7). The results of the PCR analysis were confirmed, whenever possible, by sequencing analysis of the amplification products. This provided the exact number of repeats contained in each one of the identified alleles (table 4.2.7). The results of the analysis, as well as the sequences of the primers were deposited to the MGD (accession no. J:69675).

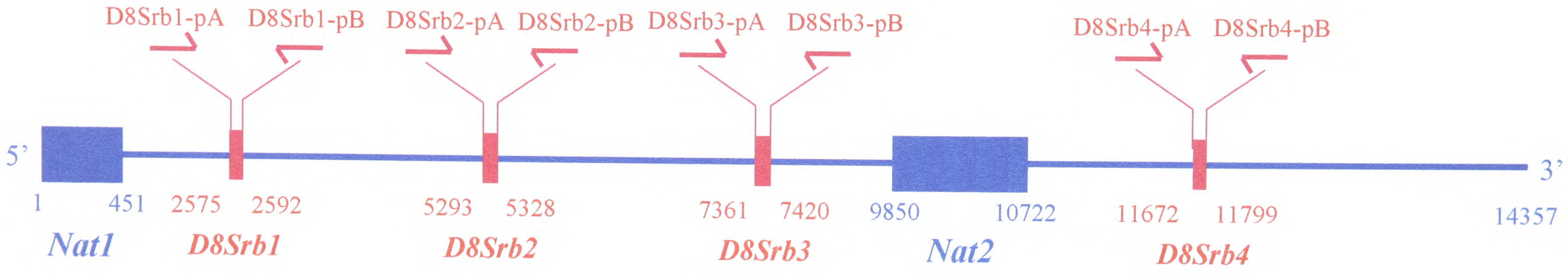


Figure 4.2.15 Map showing the position of four microsatellite loci on 129/Ola clone A. The blue boxes correspond to the coding region of the *Nat1* and *Nat2* genes and the red boxes mark the position of microsatellite loci *D8Srb1*, *D8Srb2*, *D8Srb3* and *D8Srb4*. The flanking primers (arrows) used for amplification across each microsatellite locus (table 4.2.6) are also shown. The numbering is relative to the beginning of 129/Ola clone A (Appendix 2).

Primer name	Orientation	Sequence (5′→3′)	Tm (°C)	Specificity
D8Srb1-pA D8Srb1-pB	forward reverse	2541- GGATTGTGGGTATCAGCCA -2559 2718- CTGCATGGTTCTGAGGGATT -2699	58 60	*D8Srb1* locus
D8Srb2-pA D8Srb2-pB	forward reverse	5252- GGCAAAGTTACATGACTGAGTGAAC -5276 5404- ACAAGGCCTTGAAGGAATAGCTTC -5381	72 70	*D8Srb2* locus
D8Srb3-pA D8Srb3-pB	forward reverse	7333- TTTTCTGGTAGGATTTCTGCAA -7354 7509- TGACAGGAATATAAAGGAACCAAAG -7485	60 68	*D8Srb3* locus
D8Srb4-pA D8Srb4-pB	forward reverse	11639- TTATCTGTAACAAATCAGCCACAGA -11663 11847- GGTGGCAACCTCAGGCTAAATAT -11827	68 68	*D8Srb4* locus

Table 4.2.6: Primers used for the amplification of microsatellite loci *D8Srb1*, *D8Srb2*, *D8Srb3* and *D8Srb4*. The sites of primer annealing are indicated relative to the beginning of 129/Ola clone A (Appendix 2). Each set of primers was designed to flank a specific microsatellite locus. The primers were named according to the specifications of the MGD and their sequences were assigned MGD accession no. J:69675.

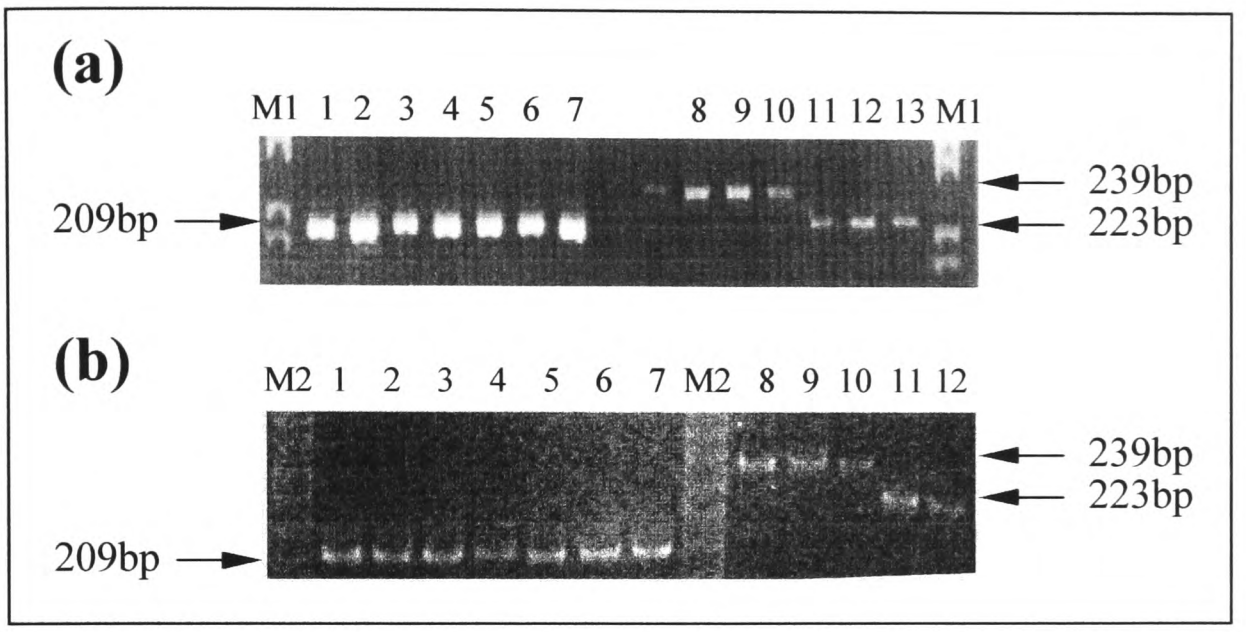


Figure 4.2.16: SSLP at microsatellite locus *D8Srb4*, between mouse strains. Amplification was carried out with flanking primers D8Srb4-pA and D8Srb4-pB (table 4.2.6). Products were initially analysed on a 3% (w/v) Metaphor/1% (w/v) agarose gel to check specificity of amplification (a). Successfully amplified products were then run on an ethidium bromide-stained 6% (w/v) non-denaturing polyacrylamide gel (section 2.2.2.2), to distinguish between the different alleles (b). Lanes 1-7 show amplification from genomic DNA of A/J, CBA, 129/Ola, PO, TO, C57Bl/6J and Balb/c mice, respectively. Lanes 8-10 show amplification from genomic DNA of individual mice of the MSP strain, while lanes 11-13 show amplification from genomic DNA of individual mice of the MCA strain. The numbering of lanes is the same in (a) and (b). Lanes M1 and M2 were loaded with 1µg of 1kb and 10bp DNA ladder, respectively. Amplification from the Balb/c, C57Bl/6J, TO, PO, 129/Ola, CBA and A/J strains provided the same 209bp product (*D8Srb4*^{a} allele). Amplification from the MSP strain provided a 239bp product (*D8Srb4*^{b} allele) and from the MCA strain a 223bp product (*D8Srb4*^{c} allele). The exact size of the amplified fragments was determined by sequencing with the forward D8Srb4-pA amplification primer.

Locus	Mouse strain	PCR product size (bp)	No. (type) of repeats	Allele
D8Srb1	129/Ola, A/J, CBA, C57Bl/6J, Balb/c, PO, TO	178	6 (GTT)	a
	MSP	184	8 (GTT)	b
	MCA	181	7 (GTT)	c
D8Srb2	129/Ola, A/J, C57Bl/6J, Balb/c, PO, TO	153	18 (GT)	a
	CBA	151	17 (GT)	b
	MSP	145	14 (GT)	c
	MCA*	161	22 (GT)	d
D8Srb3	129/Ola, A/J, CBA, C57Bl/6J, Balb/c, PO	177	30 (CA)	a
	TO*	179	31 (CA)	b
	MSP	159	21 (CA)	c
	MCA	161	22 (CA)	d
D8Srb4	129/Ola, A/J, CBA, C57Bl/6J, Balb/c, PO, TO	209	64 (mixed dinucleotides)	a
	MSP	239	79 (mixed dinucleotides)	b
	MCA	223	71 (mixed dinucleotides)	c

**Sequencing not possible*

Table 4.2.7: Combined results of the analysis of microsatellite loci *D8Srb1*, *D8Srb2*, *D8Srb3* and *D8Srb4* for SSLP between different mouse strains. The size of the PCR product generated for each mouse strain by specific amplification across each microsatellite locus is provided. Sequencing analysis of the PCR products indicated the exact number and type of repeats contained in each allele. When sequencing was not possible (e.g. due to failure of the DNA polymerase to read through the repetitive sequence), the number of repeats was estimated based only on the available electrophoretic data. Small letters denote different alleles at the same locus, according to the guidelines of the International Committee on Standardised Genetic Nomenclature for Mice. The above data were submitted to the MGD and assigned accession no. J:69675.

4.3 Discussion

Since the development of the first knock-out mice, lacking the *Hprt* gene for hypoxanthine phosphoribosyl transferase (Hooper *et al.*, 1987; Kuehn *et al.*, 1987; Thomas and Capecchi, 1987), a vast amount of information has poured into the scientific literature about the effect of targeted mutations in more than 3000 murine genes (Hardouin and Nagy, 2000). Although the basic procedure for creating knock-out strains has not changed much over the past fifteen years, new technologies have expanded the potential for genetic manipulation of the mouse (section 1.4.4). The Mouse Knock-out and Mutation Database (http://research.bmn.com/mkmd/) currently contains at least 7000 entries, a number increasing rapidly.

The work described in this Chapter (sections 4.2.1 and 4.2.2) has opened the way towards the generation of a *Nat2* knock-out mouse. The targeted 129/Ola ES cell clones, identified by a combination of positive/negative selection, Southern blot analysis and LA-PCR screening, were injected into blastocysts of the C57Bl/6J strain and transferred to pseudopregnant females for the production of chimeras. The pups showed high level of chimerism and were apparently normal. Of the nine chimeras generated, two provided germline transmission of the inactive *Nat2* allele and were intercrossed to produce heterozygous and homozygous offspring. These were normal, but had a wide variety of coat colours, probably reflecting the variable contribution of the 129/Ola and C57Bl/6J genetic backgrounds to the phenotype of different individuals. The litter sizes were within normal range and the male:female ratio was balanced. Tissues from homozygous *Nat2*-null mice were confirmed to lack *Nat2* enzymatic activity, using the specific substrate pABA. Backrossing of the knock-outs to C57Bl/6J mice is currently in progress, to place the *Nat2*-null allele within a pure C57Bl/6J genetic background (Cornish *et al.*, 2001). This is particularly important, as the phenotypic impact of a mutation can vary, depending on the effect of modifier loci associated with a specific genetic background (Müller, 1999).

Earlier attempts to produce transgenic mice overexpressing human *NAT1*, by pronuclear injection of a vector carrying the coding region of the gene under the control of the strong cytomegalovirus (CMV) promoter, have been unsuccessful. The transgene was not detected in any of the pups born and the size of litters was

significantly smaller than normal, indicating that random integration of multiple copies of the human *NAT1* gene might be lethal (Nichola Johnson, D.Phil. Thesis, Oxford 2001). To overcome these problems, an alternative "knock-in" approach was undertaken, to introduce the CMV-human *NAT1* complex into the same targeting vector used for the production of the *Nat2* knock-out mouse (figure 4.2.3). Homologous recombination between the endogenous *Nat2* locus and the new targeting construct in 129/Ola ES cells led to simultaneous inactivation of the murine *Nat2* gene and targeted insertion of a highly expressed copy of the human *NAT1* gene. The phenotypic effect of the transgene expression was dramatic, even at the stage of the chimeras. Embryos carrying the transgene had abnormal phenotypes and died *in utero* or shortly after birth. The surviving pups either did not carry the transgene or exhibited very limited chimerism. The latter were smaller than normal and had a "kinked" tail (Pinter *et al.*, 2001).

Abnormalities of the tail are common in mice heterozygous for mutations in genes involved in the process of neurulation. For example, heterozygotes for the *Loop-tail* (*Lp*), *Curly-tail* (*Ct*) and *Crooked-tail* (*Cd*) mutations have characteristically abnormal tails, compared with the homozygotes which suffer from severe abnormalities of the neural tube, such as craniorachischisis (*Lp*), exencephaly (*Ct*, *Cd*) and spina bifida (*Ct*) (Corcoran, 1998; Greene *et al.*, 1998; Carter *et al.*, 1999; Juriloff and Harris, 2000). Although NTDs were not seen among the embryos carrying the human *NAT1* transgene, the reduced litter size could indicate early death and subsequent resorption of the abnormal siblings. It appears that overproduction of the NAT protein is not tolerated by the early embryo, suggesting an important role for the enzyme in development. The findings are in support of the postulated involvement of NAT in folate catabolism (section 4.1); the observed phenotypic anomalies could be the result of rapid loss of the vitamin, due to excessive N-acetylation of pABGlu in the transgenic embryos.

In contrast, mice lacking the *Nat2* gene appear to be phenotypically normal. This is not a contradiction to the hypothesised role of murine NAT2 in folate catabolism, since complete lack of the enzyme would not cause folate deficiency in the developing embryo. Functional redundancy or gene compensation are the most likely reasons for the lack of abnormal phenotype in the *Nat2* knock-out mice (Lobe

and Nagy, 1998; Müller, 1999; Williams and Wagner, 2000). A minor residual N-acetylating activity towards pABGlu has been identified in human placental homogenates, following specific inhibition of human NAT1 (Upton *et al.*, 2000), suggesting that other metabolic pathways may compensate for the loss of the functionally equivalent NAT2 isoenzyme in the *Nat2* knock-out mice. Alternatively, other *Nat* genes could modify their expression, to provide the cell with sufficient amount of NAT protein. Overlapping or redundant enzymatic functions are often seen in xenobiotic metabolism, where biotransformation of a specific compound can be carried out by different isoenzymes or via distinct metabolic routes (Gonzalez, 1998).

Knock-out mice deficient in enzymes of phase I metabolism, such as various P450s (CYP1A1, CYP1A2, CYP1B1 and CYP2E1) and microsomal epoxide hydrolase (mEH), have been reported to be viable, fertile and phenotypically normal. However, these mice respond differently to their wild type littermates, when exposed to toxic or carcinogenic compounds (Gonzalez and Kimura, 1999; Kimura and Gonzalez, 2000; Gonzalez and Kimura, 2001; Gonzalez, 2001). For example, CYP2E1-deficient mice are highly resistant to the hepatotoxic effect of paracetamol (Lee *et al.*, 1996) and carbon tetrachloride (Wong *et al.*, 1998), as well as to the myelotoxic effect of benzene (Valentine *et al.*, 1996). CYP1B1-deficient mice are less prone to carcinogenicity caused by 7,12-dimethylbenz[α]anthracene (DMBA), and so are mEH-null mice (Buters *et al.*, 1999; Miyata *et al.*, 1999). In contrast, lack of CYP1A2 does not affect sensitivity to arylamine and heterocyclic amine carcinogens, as would be expected, suggesting the involvement of additional enzymes in the bioactivation of these compounds (Kimura *et al.*, 1999). The CYP1A2-, CYP1B1- and CYP2E1-deficient mice are currently being used to generate transgenic lines expressing the corresponding human genes (Kimura and Gonzalez, 2000).

Targeted inactivation of genes encoding for enzymes of phase II metabolism has also been reported to affect sensitivity to various toxic or carcinogenic agents. Mice lacking NADPH:quinone oxidoreductase 1 (NQO1) are more sensitive to the toxic effects of menadione, a quinone normally detoxified by this enzyme (Radjendirane *et al.*, 1998). Similarly, mice deficient in glutathione S-transferase P1 (GSTP1) are more prone to DMBA-induced carcinogenesis (Henderson *et al.*, 1998), Transgenic mice overexpressing the human NAT2 isoenzyme specifically in the

prostate have also been produced, with the aim to investigate the role of dietary heterocyclic amines, such as 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP), in prostate carcinogenesis. Although prostate-specific overexpression of human NAT2 was achieved, the transgenic animals did not differ from the wild type controls in their capacity to bioactivate a range of carcinogens, including N-hydroxy-PhIP (Leff *et al.*, 1999a). The significance of these results is uncertain, since it is currently unknown whether NAT2 is expressed in the human prostate. The transgenic and knock-out strains described in the present study should serve as ideal models for studying the role of NAT in chemical toxicity and carcinogenesis. They may also be useful models for testing the efficacy and toxic effects of drugs metabolised by NAT.

Numerous epidemiological studies have investigated the involvement of human NAT isoenzymes in various types of cancer, identifying – at least in the case of bladder cancer – a link between the acetylator status and predisposition to the disease (section 1.2.4.1). However, human *NAT* is not the only locus on chromosomal region 8p22 implicated in carcinogenesis. Many investigators have detected loss of heterozygosity (LOH) at several polymorphic sites of region 8p22 in colorectal, hepatocellular, mammary, urothelial, prostatic and non-small cell lung carcinomas (Fujiwara *et al.*, 1993; Knowles *et al.*, 1993; Farrington *et al.*, 1996; Lerebours *et al.*, 1999; Van Alewijk *et al.*, 1999; Wang *et al.*, 1999). The data support the presence of at least two tumour suppressor genes (TSGs), lying within critical intervals near the *LPL* gene for lipoprotein lipase. The *NAT* locus is located close to the *LPL* gene (section 1.2.2), and its loss has often been detected in tumours of the bladder by FISH and LOH (Schnakenberg *et al.*, 1998; Stacey *et al.*, 1999; Thygesen *et al.*, 1999). However, the *NAT* genes themselves are not believed to act as TSGs, as retention of the *NAT* locus has been observed in a large proportion of colorectal and bladder carcinomas with 8p deletions (Hubbard *et al.*, 1997; Schnakenberg *et al.*, 1998), and mutation analysis has failed to reveal any deleterious mutations in the *NAT* genes of bladder cancer cells (Knowles, 1999). Nevertheless, the large number of SNPs associated with the human *NAT* genes (section 1.2.3), as well as a series of *NAT*-specific cosmid probes, recently developed for FISH analysis, have established *NAT* as a useful marker for loss of region 8p22 in humans (Stacey *et al.*, 1996; 1998; 1999).

Extensive physical maps have been generated for the human 8p22 region (Bookstein *et al.*, 1994; Fujiwara *et al.*, 1994; Wang *et al*, 1999; Arbieva *et al.*, 2000) and the available information has dramatically increased since the announcement of the draft sequence of the human genome (IHGSC, 2001; Venter *et al.*, 2001). Several candidate TSGs have been described in the region, including the *CTSB* gene for cathepsin B (Hughes *et al.*, 1998b), the *FEZ1* gene encoding for a leucine-zipper protein (Ishii *et al.*, 1999), the dematin gene for a cytoskeletal protein (Lutchman *et al.*, 1999), and the *REAM* and *CSMD1* genes of unknown function (Oyama *et al.*, 2000; Sun *et al.*, 2001). However, a consensus has not been reached as to the exact identity of the TSGs on 8p22. Even the position of the critical intervals may vary between studies, depending on the markers and the type of cancerous material used.

Comparative mapping, i.e. the search for homologous genes within chromosomal regions of conserved synteny (Carver and Stubbs, 1997), could help identify the TSGs of interest, first in the mouse and then in the human. The 8 B3.1-3.3 chromosomal region, where the murine *Nat* genes have been localised (Fakis *et al.*, 2000), corresponds to the 33cM genetic position on mouse chromosome 8 (2000 Chromosome 8 Committee Report; www.informatics.jax.org/ccr). This region is syntenic with human 8p22, based on the presence of several conserved ("anchor") loci, such as the *Lpl* gene for lipoprotein lipase, the *Atp6b1* gene for a lysosomal ATPase and the *Slc18a1* gene for a solute carrier (MGD Mouse Chromosome 8 Linkage Map with human homologies; www.informatics.jax.org). Given the close proximity of *NAT* and *LPL* on human chromosome 8 (figure 1.3), it is highly possible that the two loci are conserved on the same linkage group, at around 33cM from the centromere of mouse chromosome 8.

Previous genetic mapping has placed the *Nat* genes at 28.2-33.8cM (31±2.8cM) on mouse chromosome 8 (Mattano *et al.*, 1988), i.e. close to the *Lpl* locus at 33cM. However, screening of YAC contigs spanning this critical interval (table 4.2.4) did not provide a *Nat*-positive clone (section 4.2.3.1). It is most likely that *Nat* is located in one of the gaps between the contigs, an assumption further supported by screening of flanking uninterrupted contigs, covering region 15.7-25.1cM on the centromeric side and 33.9-41.5cM on the telomeric side of the above-mentioned interval. Although the preliminary work described in this Chapter has not identified

the exact physical map position of the murine *Nat* locus, it has narrowed down the critical interval to chromosome 8 regions 28.4-29.5cM, 31.7-32.8cM or 32.8-33.9cM, corresponding to the three gaps between contigs WC8.13-WC8.19 (section 4.2.3.1).

Accurate physical mapping of the murine *Nat* genes will serve as a starting point towards screening of the syntenic 8 B3.1-3.3 region for disease loci, including the homologues of the above-mentioned TSGs, and the long suspected gene(s) responsible for high predisposition of the A/J strain to teratogen-induced orofacial clefting (section 4.1). The latter gene(s), would be expected to lie close to *Nat*, as the disease phenotype is co-inherited with the slow NAT2 phenotype in congenic strains carrying the *Nat2*9* allele of the A/J mouse on a C57Bl/6J genetic background (Karolyi *et al.*, 1990). It has been estimated that, in these congenic strains, the chromosomal segment of A/J origin does not extend further than 12cM to the telomeric side of the *Nat2* gene (Mattano *et al.*, 1988). The corresponding boundary on the centromeric side of *Nat2* has not been mapped yet, but it would be expected to be of comparable size. The SSLP markers identified in section 4.2.4 should be useful tools for refined linkage analysis between the *Nat* locus and polymorphisms in neighbouring disease loci on mouse chromosome 8. These analyses will be further facilitated by the use of the wild-derived inbred strains MSP and MCA, which are known to differ from the common inbred strains at about 90% of microsatellite loci across their genome (Meisler, 1996; Rhodes *et al.*, 1998).

Despite the enormous recent progress of the Mouse Genome Project (section 1.4.1.2), no sequencing information is currently available about the broader genomic region harbouring the murine *Nat* genes, as demonstrated by search of various genomic databases (e.g. EMBL, HTG, GSS) against the sequence of 129/Ola clone A (section 4.2.2.2). Moreover, search of the dbSTS database did not identify any previously mapped STSs on the same clone. Combined with the above-mentioned lack of coverage by YAC contigs, the absence of defined markers within the *Nat* region makes localisation of the *Nat* genes even more complicated. On the other hand, analysis of the sequencing data generated for 129/Ola clone A has revealed a number of interesting features, including a candidate promoter and polyA signal, as well as various putative exons for the *Nat2* gene. The significance of these elements will be the focus of the following two Chapters.

CHAPTER 5

Genomic structure and tissue-specific expression of the murine genes for arylamine N-acetyltransferases

5.1 Introduction

The liver is the main site of xenobiotic metabolism, although detoxification or bioactivation of toxic and carcinogenic compounds may also take place in other tissues (Caldwell *et al.*, 1995). Study of the tissue expression profiles of xenobiotic metabolising enzymes is important, since their specific interactions can determine the local effects of potentially hazardous foreign compounds (Windmill *et al.*, 1997). Furthermore, many xenobiotic metabolising enzymes have endogenous substrates, indicating that their tissue-specific expression may be linked to fundamental cellular functions in adult life, but also during development (Nebert, 1997; Nebert and Duffy, 1997).

The main site of human NAT2 expression is the liver, as has previously been demonstrated by a variety of methods, including enzymatic activity assays with specific NAT2 substrates (Jenne, 1965; Grant *et al.*, 1989b), cDNA cloning (Ohsako and Deguchi, 1990) and *in situ* hybridisation with specific riboprobes (Windmill *et al.*, 1997; Debiec-Rychter *et al.*, 1999). Human NAT2 enzymatic activity and mRNA have also been detected in normal and hepatoma (HepG2 and HepG3) liver cell lines (Coroneos *et al.*, 1991; Coroneos and Sim, 1993). Another important site of human NAT2 expression is the small and large intestine. Although the predominant isoenzyme in the intestine is NAT1, the contribution of NAT2 to the local bioactivation of dietary carcinogens could still be significant (Jenne, 1965; Ilett *et al.*, 1994; Windmill *et al.*, 1997; Hickman *et al.*, 1998; Debiec-Rychter *et al.*, 1999). Human NAT2 activity or immunoreactive protein have not been detected in any other

tissues. However, the recent application of sensitive molecular techniques, such as RT-PCR and *in situ* hybridisation, have allowed detection of *NAT2* mRNA in several tissues, including the bladder, breast, esophagus, stomach, kidney tubules, pineal gland and lung (Kloth *et al.*, 1994; Sadrieh *et al.*, 1996; Windmill *et al.*, 1997; Macé *et al.*, 1998; Debiec-Rychter *et al.*, 1999; Williams *et al.*, 2001).

The above-mentioned and other (Pacifici *et al.*, 1986; Stanley *et al.*, 1996; Lee *et al.*, 1997; Vaziri *et al.*, 2001) investigators have also studied expression of the human NAT1 isoform in the liver and various extrahepatic tissues. NAT1 is the predominant -if not the only- human NAT isoform expressed in all extrahepatic tissues analysed. In contrast, NAT1 is less abundant than NAT2 in the liver (Jenne, 1965; Grant *et al.*, 1989b; Ohsako and Deguchi, 1990). NAT1 has been shown to be exclusively expressed in white and red blood cells (Cribb *et al.*, 1991; Ward *et al.*, 1992; Risch *et al.*, 1996), as well as in various non-hepatic normal and carcinogenic cell lines (Coroneos and Sim, 1991; Coroneos *et al.*, 1991; Kelly and Sim, 1991; Coroneos and Sim, 1993; Stacey *et al.*, 1998). It is also found in foetal tissues (Pacifici *et al.*, 1986), as well as in the placenta (Derewlany *et al.*, 1994; Smelt *et al.*, 1997).

The mouse NAT2 isoenzyme shows a wide tissue distribution pattern, similar to its functional equivalent human NAT1. This has previously been demonstrated by enzymatic activity assay with pABA, as well as by Western blot analysis and immunohistochemistry with specific antisera (Chung *et al.*, 1993; Stanley *et al.*, 1997). Similarly to human NAT2, the mouse NAT1 isoenzyme shows a more restricted tissue distribution pattern. It has been detected only in the liver and the intestine by enzymatic activity assay with procainamide and isoniazid, as well as by Northern blotting (Hein *et al.*, 1988; Martell *et al.*, 1992; Levy *et al.*, 1993; Ware and Svensson, 1996). Other extrahepatic tissues of the mouse have not yet been examined for murine NAT1 expression and there is also no information regarding the tissue distribution of the murine NAT3 isoenzyme.

As very little is known about the expression of murine *Nat* genes at the level of transcription, the aim of the work presented here has been to look for *Nat1*, *Nat2* and *Nat3* transcripts in a variety of tissues of the adult mouse. In view of the potential role of NAT in folate metabolism and early embryonic development (section 4.1),

another major aim of this work has been to search for expression of the human and murine genes for NAT in the preimplantation embryo. Interest has further been focused on the elucidation of the genomic structure of murine *Nat2*, using the sequence information generated in Chapter 4 and resources available from the Mouse Genome Project.

5.2 Results

5.2.1 Tissue-specific expression of murine *Nat* genes

A number of tissues, namely liver, spleen, thymus, heart, brain, lung, kidney, submaxillary gland and placenta, were examined for murine *Nat* gene expression by RT-PCR. Whole organs were obtained from Balb/c mice (section 2.1.6) and total RNA and mRNA was isolated as described in section 2.2.1.4 (figure 5.2.1). The generated cDNA (section 2.2.3.2) was subsequently used for PCR amplification (section 2.2.3.3) of the entire coding region of *Nat1*, *Nat2* and *Nat3* genes, with primer pairs Mus12/Mus13, mNAT2-1/mNAT2-910 and Mus12/Mus15 (table 3.2.1), respectively. Amplification of the housekeeping gene for murine glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*) (accession no. NM008084; Hanauer and Mandel, 1984; Sabath *et al.*, 1990) was performed as positive control of PCR with cDNA templates. Primers GAPDH-S and GAPDH-AS (table 5.2.1), used for amplification of the *Gapdh* coding region, do not span any introns, providing the same 910bp PCR product from genomic DNA and cDNA. Treatment of total RNA with DNaseI (section 2.2.3.1) was performed to remove genomic DNA from the RNA preparations, and the product of a mock reverse transcription without reverse transcriptase in the reaction mixture was included in all sets of PCR amplifications, as a control for genomic DNA contamination of the preparations. Lack of amplification in this reaction indicated that mRNA preparations were free of genomic DNA.

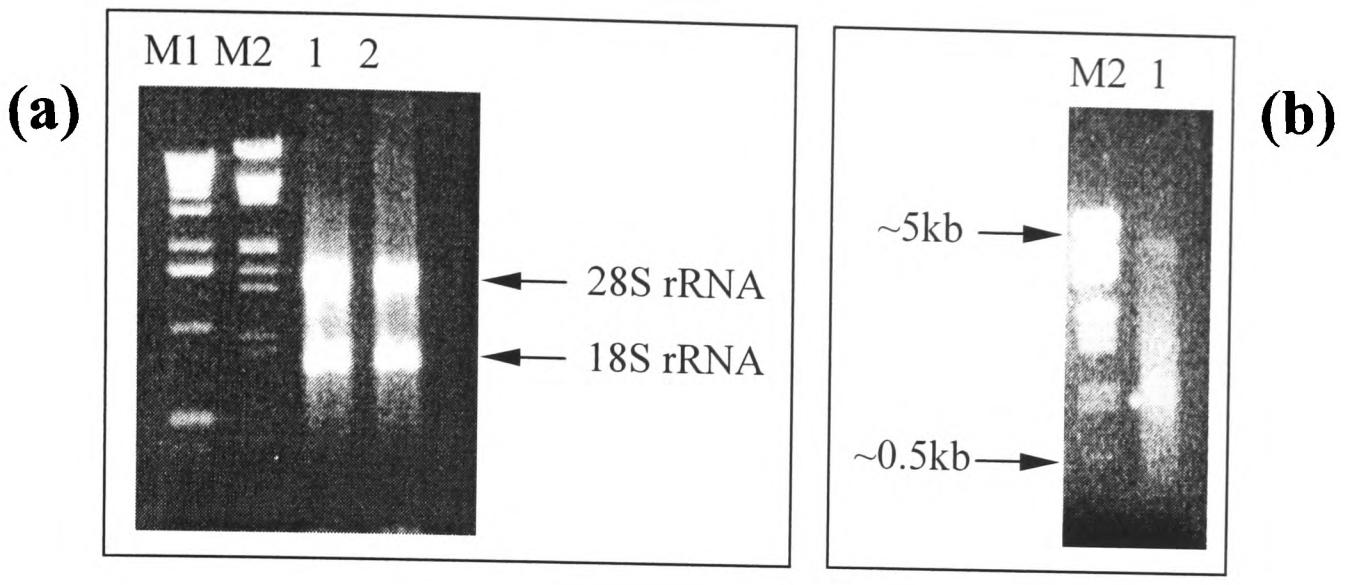


Figure 5.2.1: Isolation of total RNA and mRNA from tissues of the Balb/c mouse. a) Total RNA was extracted from kidney (lane 1) and heart (lane 2), using the Tri-Reagent method (section 2.2.1.4). Approximately 1.5μg of total RNA was then loaded to an agarose gel. Good quality total RNA appears as a smear, extending from approximately 8kb to 0.5kb, with two thick bands corresponding to 28S and 18S ribosomal RNA. b) mRNA (lane 1) was isolated from liver total RNA, as described in section 2.2.1.4. The same size range smear is present, but the ribosomal RNA bands are absent. Lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively.

Expression of murine *Nat2* gene was evident in all of the examined tissues (figures 5.2.2 and 5.2.3), while *Nat1* expression was detected only in the liver (figure 5.2.2). Expression of *Nat3* was detected in the spleen (figure 5.2.3), but not in any other of the analysed tissues. A second round of PCR was performed for *Nat1* and *Nat3*, using the sets of primers described above, to allow detection of transcript, even if present at very low amounts in the analysed tissue. The results are summarised in table 5.2.2.

Specific amplification from the *Nat1*, *Nat2* and *Nat3* coding region was confirmed by restriction digestion of the RT-PCR products with *Ava*I, *Pst*I and *Hae*III (Kelly and Sim, 1994). *Ava*I cuts only within the *Nat1* coding region (nucleotide position 587), providing fragments 603 and 307bp in size. *Pst*I cuts only within the *Nat2* coding region (nucleotide position 673), providing fragments 673 and 237bp in size, while *Hae*III cuts only within the *Nat3* coding region (nucleotide position 237), providing fragments 659 and 251bp in size (figures 5.2.2d and 5.2.3d).

Primer name	Orientation	Sequence (5′→3′)	Tm (°C)	Specificity
UPSTREAM[1]	forward	3339- AAGCCCTACAACTACATTTCC -3359	58.6	Region upstream of mouse *Nat2* non-coding exon
NCE-F[1]	forward	3553- TCTTCCAGAAGGAAGACCGG -3572	59.5	Mouse *Nat2* non-coding exon
polyA-1R[1]	reverse	10860- AAATTTAGCTGTAAGCCAATGCTG -10837	58	Region upstream of first, second and third putative polyA signals of mouse *Nat2*, respectively
polyA-2R[1]	reverse	11131- ACTTACTCAATTGGTGTTTGTTCCC -11107	60	
polyA-3R[1]	reverse	11386- GCCTAGGAATGTGTCTTCTATGTGG -11362	63.5	
GAPDH-S[2]	forward	105- CCATTTGCAGTGGCAAA -121	52.8	Murine *Gapdh* gene
GAPDH-AS[2]	reverse	1015- CACCACCCTGTTGCTGTAG -997	61.4	
NatHu-7[3]	forward	32- GGCTATAAGAACTCTAGGAAC -52	60	Human *NAT2* gene
NatHu-8[3]	reverse	862- AATAGTAAGGGATCCATCACC -842	60	
N-376[4]	forward	(-376) - TATTGCATGATTCTCCTGCCTA - (-355)	62	Region upstream of human *NAT1* gene
N 1177[4]	reverse	1177- GGAATTCAACAATAAACCAACAT -1155	60	Region downstream of human *NAT1* gene
HPRT-S[5]	forward	162- AATTATGGACAGGACTGAACGTC -184	59	Human *HPRT1* gene
HPRT-AS[5]	reverse	748- GGCGATGTCAATAGGACTCCAGATG -724	65	

[1]This study; [2]Provided by Dr. Valerie Smelt (Department of Pharmacology, Oxford); [3]Hickman and Sim, 1991; [4]N. Johnson, D.Phil. Thesis, Oxford 2001; [5]Smelt et al., 2000

Table 5.2.1: Primers used for PCR amplification and sequencing in Chapter 5. The nucleotide sites of primer annealing are indicated relative to the beginning of 129/Ola clone A (Appendix 2) for primers UPSTREAM, NCE-F, polyA-1R, polyA-2R and polyA-3R; relative to the beginning of the coding region of the human *NAT2* gene for primers NatHu-7 and NatHu-8; relative to the beginning of the coding region of the human *NAT1* gene for primers N-376 and N 1177; relative to the beginning of the murine *Gapdh* mRNA (accession no. NM008084) for primers GAPDH-S and GAPDH-AS; relative to the beginning of the human *HPRT1* mRNA (accession no. XM040683) for primers HPRT-S and HPRT-AS. Primers Mus12, Mus13, Mus15, mNAT2-1, mNAT2-910 and mNAT2-691, also used in this Chapter, have been described previously in Chapter 3 (table 3.2.1).

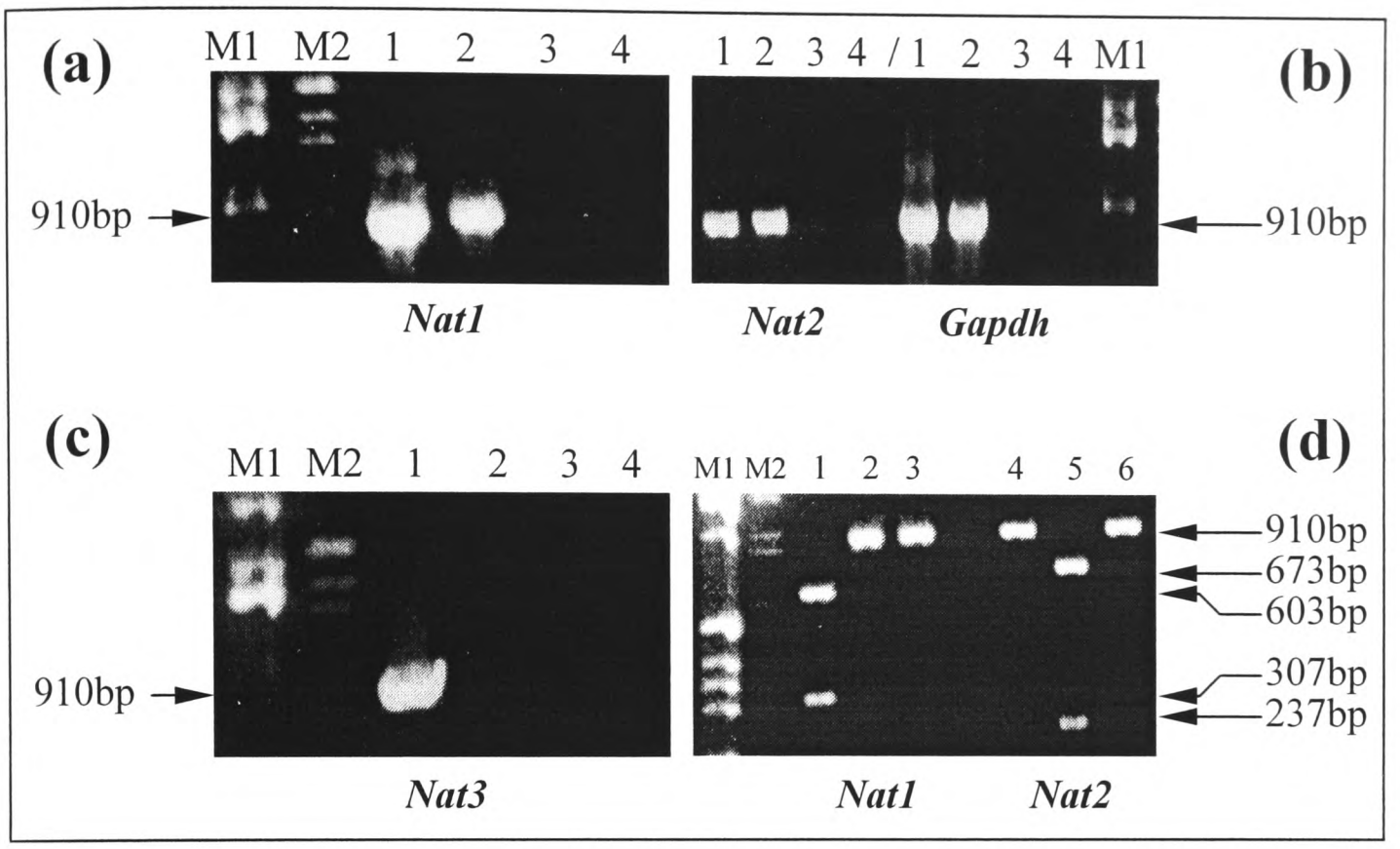


Figure 5.2.2: Investigation of *Nat* gene expression in the liver of Balb/c mouse by RT-PCR. a-c) Amplification from genomic DNA (lane 1) and cDNA (lane 2), as well as from the product of a mock reverse transcription without reverse transcriptase in the reaction mixture (lane 3). Lane 4 is the PCR negative control. a) RT-PCR with primers Mus12 and Mus13, demonstrating *Nat1* expression in the liver. b) RT-PCR with primers mNAT2-1 and mNAT2-910 (left), demonstrating expression of *Nat2* in the liver. Amplification of *Gapdh* with primers GAPDH-S and GAPDH-AS (right) was used as positive control. c) RT-PCR with primers Mus12 and Mus15, indicating lack of *Nat3* expression in the liver. d) Restriction digestions of *Nat1* (lanes 1-3) and *Nat2* (lanes 4-6) RT-PCR products, to confirm specific primer binding. Digestions were carried out with *Ava*I (lanes 1 and 4), *Pst*I (lanes 2 and 5) and *Hae*III (lanes 3 and 6), and the products were analysed on a 2% (w/v) agarose gel. *Nat1*-specific RT-PCR products were digested only by *Ava*I, providing the expected 603 and 307bp fragments. *Nat2*-specific amplification products were digested only by *Pst*I, providing the expected 673 and 237bp fragments. Lanes M1 and M2 are 1µg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively.

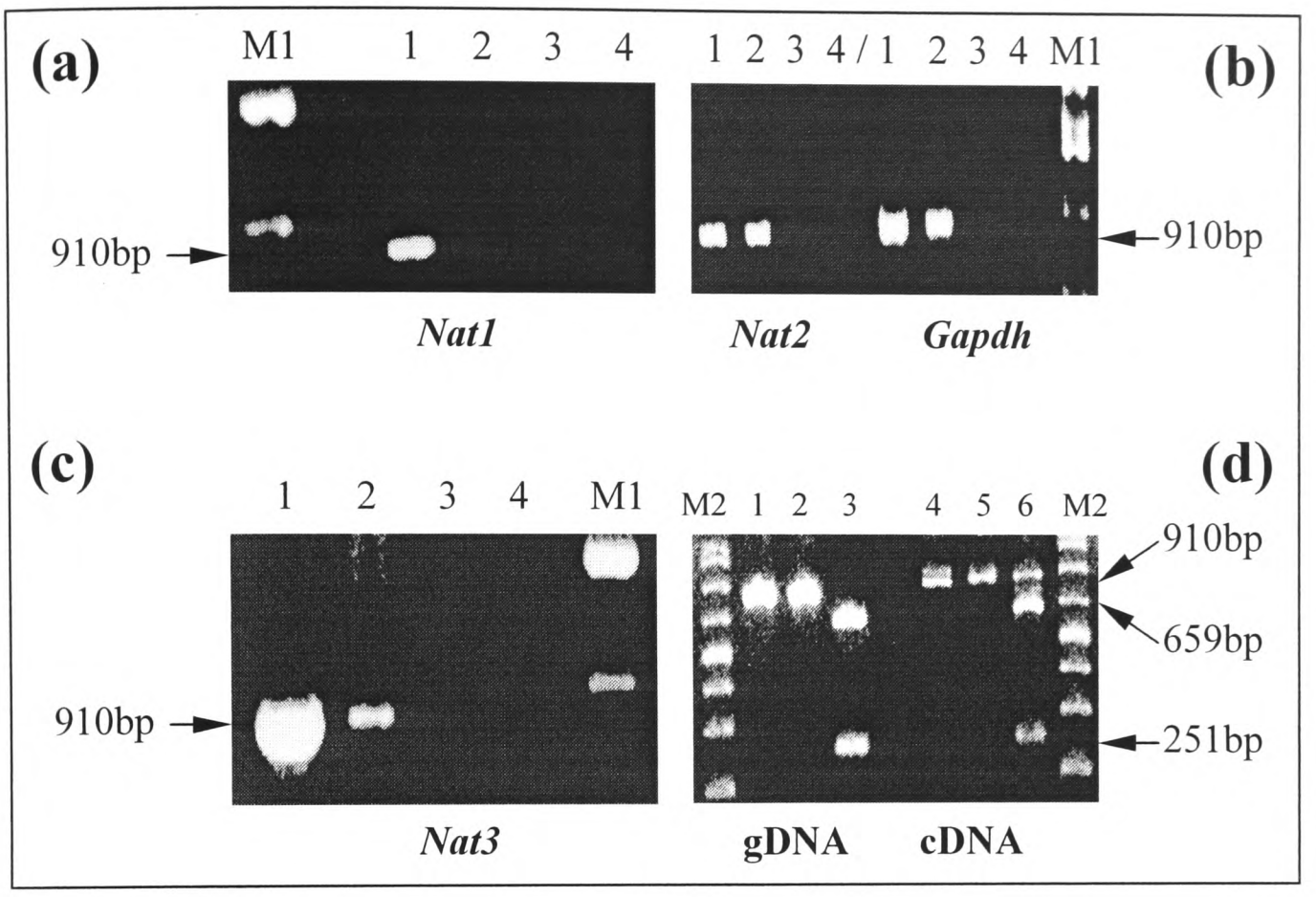


Figure 5.2.3: Investigation of *Nat* gene expression in the spleen of Balb/c mouse by RT-PCR. a-c) Amplification from genomic DNA (lane 1) and cDNA (lane 2), as well as from the product of a mock reverse transcription without reverse transcriptase in the reaction mixture (lane 3). Lane 4 is the PCR negative control. Lanes M1 were loaded with 1µg of 1kb DNA ladder. a) RT-PCR with primers Mus12 and Mus13, indicating lack of *Nat1* expression in the spleen. b) RT-PCR with primers mNAT2-1 and mNAT2-910 (left), demonstrating expression of *Nat2* in the spleen. Amplification of *Gapdh* with primers GAPDH-S and GAPDH-AS (right) was used as positive control. c) RT-PCR with primers Mus12 and Mus15, demonstrating *Nat3* expression in the spleen. d) Restriction digestions of *Nat3* amplification products from genomic DNA (lanes 1-3) and cDNA (lanes 4-6), to confirm specific primer binding. Digestions were carried out using *Ava*I (lanes 1 and 4), *Pst*I (lanes 2 and 5) and *Hae*III (lanes 3 and 6), and the products were analysed on a 2% (w/v) Metaphor gel. Amplification products were digested only by *Hae*III, providing the expected 659 and 251bp fragments. A small amount of undigested RT-PCR product (910bp) is visible in lane 6. Lanes M2 are 5µl of BioMarker EXT size markers.

Following detection of *Nat3* transcript in the spleen, Western blot analysis was performed with antiserum 193 (section 2.3.4), to detect NAT3 protein in pooled spleen homogenates prepared from three Balb/c mice. Antibody binding to pure SBTI protein, used as positive control, was evident, but immunoreactive NAT3 protein was not detected (figure 5.2.4).

Table 5.2.2: Tissue-specific expression of *Nat* genes in the Balb/c mouse. RT-PCR was used to detect *Nat* gene transcripts, as described in figures 5.2.2 and 5.2.3. Presence of a *Nat* transcript is indicated with a plus (+) and absence with a minus (-).

Tissue	Expressed *Nat* gene		
	Nat1	***Nat2***	***Nat3***
Liver	+	+	-
Spleen	-	+	+
Thymus	-	+	-
Heart	-	+	-
Brain	-	+	-
Lung	-	+	-
Kidney	-	+	-
Submaxillary gland	-	+	-
Placenta	-	+	-

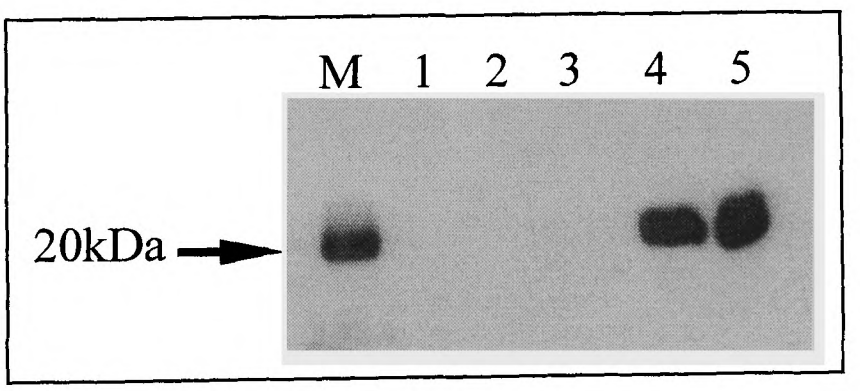


Figure 5.2.4: Western blot analysis of pooled spleen homogenate from three Balb/c mice. Lanes 1, 2 and 3 were loaded with 20, 15 and 10µl of pooled spleen homogenate, respectively. Lanes 4 and 5 were loaded with 1.5 and 3µg of pure SBTI protein, as positive control. Lane M was loaded with approximately 5µg of Rainbow markers, also containing SBTI. Antiserum 193, raised against the SBTI-conjugated C-terminal peptide of murine NAT3, bound to the 20kDa SBTI protein of the markers and the controls. No immunoreactive NAT3 protein was detected in the spleen homogenate. At the time of this experiment, recombinant NAT3 protein was not available for use as positive control. However, binding of antiserum 193 to recombinant NAT3 has previously been demonstrated (Ian Mills, Part II Thesis, Oxford 1996).

5.2.2 Expression of the genes for NAT in human and mouse preimplantation embryos

Mouse ES cells of 129/Ola strain origin were provided by the Centre for Genome Research of the University of Edinburgh. The cells were used to isolate total RNA and mRNA (section 2.2.1.4), and expression of the murine *Nat* genes was investigated by RT-PCR, as described in section 5.2.1. Expression of the *Nat2* gene was evident, confirming a previous study (Payton *et al.*, 1999b). Furthermore, it was established that the *Nat1* and *Nat3* genes are not expressed in ES cells of the preimplantation mouse embryo (figure 5.2.5).

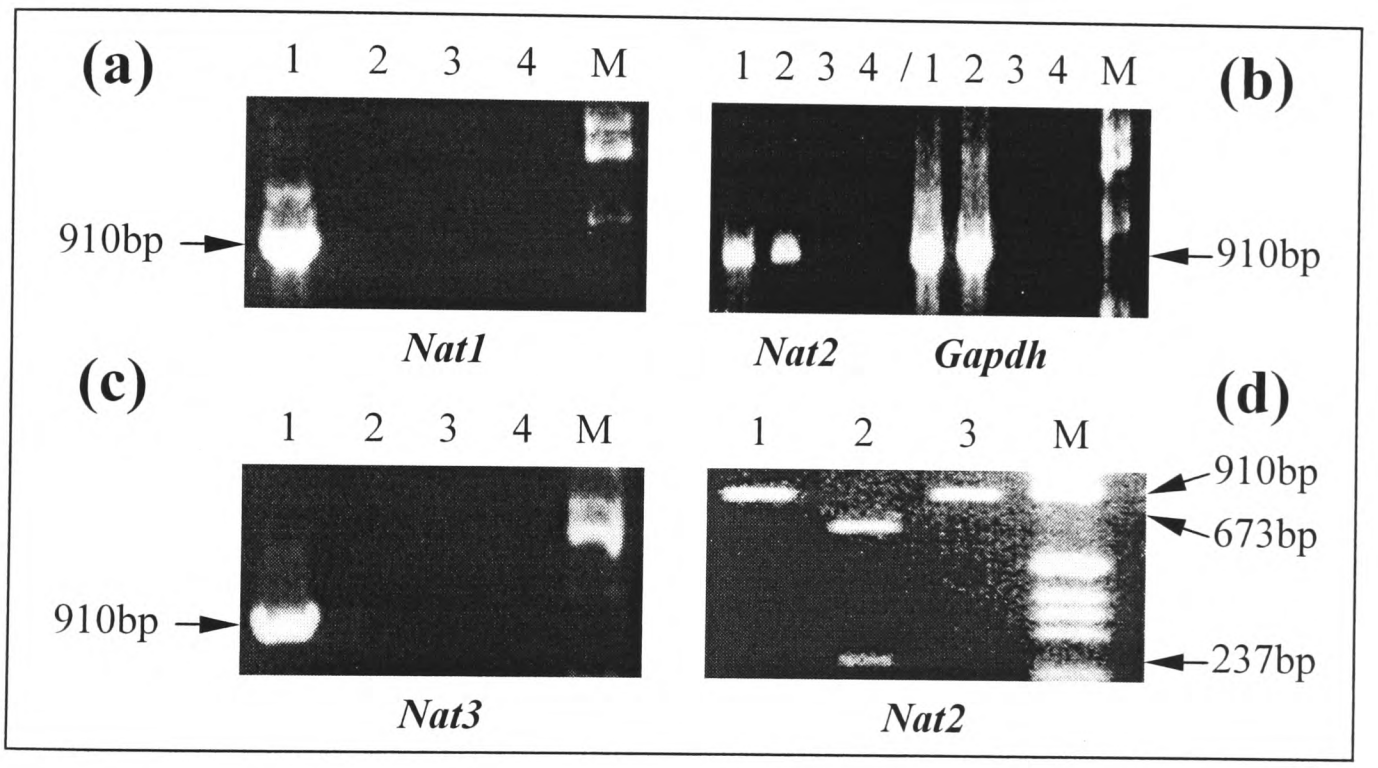
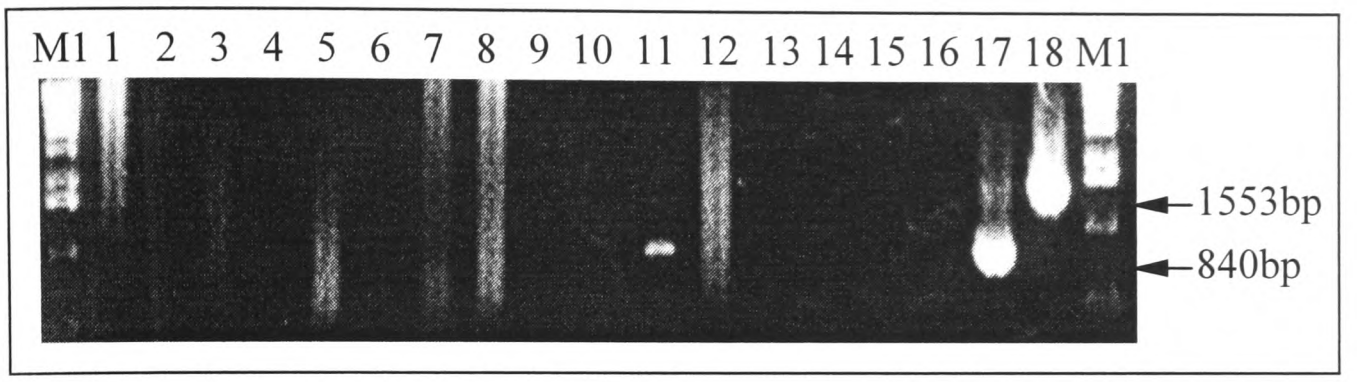


Figure 5.2.5: Investigation of *Nat* gene expression in ES cells of 129/Ola mouse by RT-PCR. a-c) Amplification from genomic DNA (lane 1) and cDNA (lane 2), as well as from the product of a mock reverse transcription without reverse transcriptase in the reaction mixture (lane 3). Lane 4 is the PCR negative control. a) RT-PCR with primers Mus12 and Mus13, indicating lack of *Nat1* expression. b) RT-PCR with primers mNAT2-1 and mNAT2-910 (left), demonstrating expression of *Nat2*. Amplification of *Gapdh* with primers GAPDH-S and GAPDH-AS (right) was used as positive control. c) RT-PCR with primers Mus12 and Mus15, indicating lack of *Nat3* expression. d) Restriction digestions of the *Nat2* RT-PCR product, to confirm specific primer binding. Digestions were carried out with *Ava*I (lane 1), *Pst*I (lane 2) and *Hae*III (lane 3), and the products were analysed on a 2% (w/v) agarose gel. The *Nat2*-specific amplification product was digested only by *Pst*I, providing the expected 673 and 237bp fragments. Lanes M are 1µg of 1kb DNA ladder.

Previous PCR screening of cDNA libraries generated from human oocytes and single preimplantation embryos (Adjaye *et al.*, 1997; 1998; 1999) has demonstrated expression of the human *NAT1* gene at the blastocyst stage (Smelt *et al.*, 2000). To investigate whether the human *NAT2* gene is also expressed in the preimplantation embryo, the same cDNA libraries were screened by PCR with primers NatHu-7 and NatHu-8 (table 5.2.1), specific for the intronless coding region of human *NAT2*. To detect possible contamination of the cDNA preparations with genomic DNA, a parallel amplification was carried out with human *NAT1*-specific primers N-376 and N 1177 (table 5.2.1). Forward primer N-376 anneals upstream of the transcribed region of human *NAT1* gene (Ohsako and Deguchi, 1990), therefore, it should provide amplification product only from genomic DNA with reverse primer N 1177. As a positive control for amplification from cDNA templates, part of the housekeeping gene for human hypoxanthine phosphoribosyl transferase (*HPRT1*) (accession no. XM040683 and M26434, for the mRNA and genomic sequence, respectively; Jolly *et al.*, 1983; Kim *et al.*, 1986) was amplified with primers HPRT-S and HPRT-AS (table 5.2.1). Forward primer HPRT-S anneals to the junction between exons 2 and 3, while reverse primer HPRT-AS anneals to exon 9 of the *HPRT1* gene (Daniels *et al.*, 1998). Therefore, these primers should provide amplification product only from cDNA.

When genomic DNA was used as template, PCR efficiency with both sets of *NAT*-specific primers was high (figure 5.2.6a). Genomic DNA contamination was not observed (figure 5.2.6a) and amplification of the *HPRT1* gene was successful (figure 5.2.6b), indicating good quality cDNA template. Following two rounds of amplification, human *NAT2* transcript was not detected in the oocyte or in 2-cell, 4-cell, 8-cell and blastocyst stage (preimplantation) embryos. *NAT2* expression was evident only in the 10-week (post-implantation) embryo (figure 5.2.6a). It was established that the only *NAT* gene expressed in human blastocyst is *NAT1*, as previously demonstrated by Smelt *et al.* (2000).

(a)



(b)

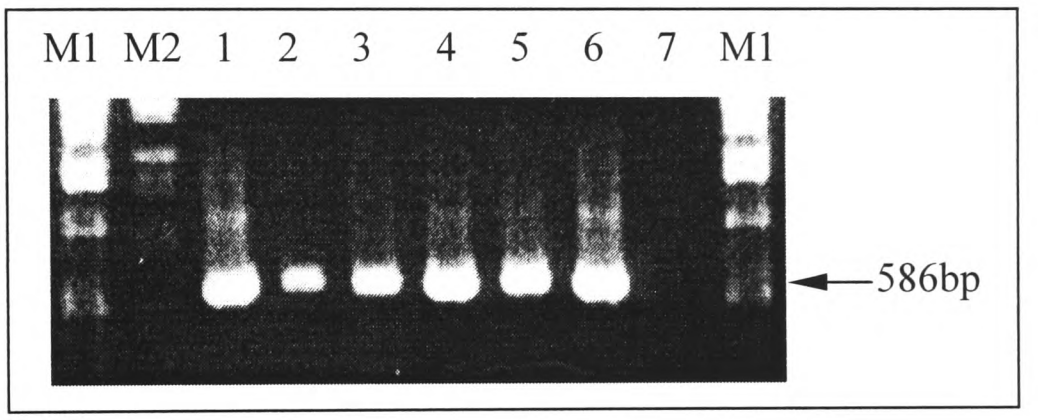


Figure 5.2.6: PCR screening of cDNA libraries from human oocytes and single preimplantation embryos for *NAT2* gene expression. a) Second round of amplification with *NAT2*-specific primers NatHu-7 and NatHu-8 from oocyte (lane 1), 2-cell (lane 3), 4-cell (lane 5), 8-cell (lane 7), blastocyst (lane 9) and 10-week (lane 11) embryo cDNA. Lanes 2, 4, 6, 8, 10 and 12 are control amplifications with *NAT1*-specific primers N-376 and N 1177, using cDNA templates in the same order. Negative controls from the first PCR round were subjected to a second round of amplification and loaded to lanes 13 (NatHu-7 and NatHu-8 primers) and 14 (N-376 and N 1177 primers). The corresponding negative controls from the second PCR round were loaded to lanes 15 and 16. Amplification products from genomic DNA with primer pairs NatHu-7/NatHu-8 (lane 17) and N-376/N 1177 (lane 18) are also shown. The *NAT2* gene was amplified only from 10-week embryo cDNA (lane 11), as well as from genomic DNA (lane 17). Primers N-376 and N 1177 provided amplification product only from genomic DNA (lane 18), indicating that the cDNA preparations were free of genomic DNA. b) Amplification of the human *HPRT1* gene with primers HPRT-S and HPRT-AS, used as positive control for PCR efficiency with cDNA templates. The expected 586bp product was amplified from oocyte (lane 1), 2-cell (lane 2), 4-cell (lane 3), 8-cell (lane 4), blastocyst (lane 5) and 10-week (lane 6) embryo cDNA. Lane 7 is the PCR negative control. Lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively.

5.2.3 Screening of the mouse embryonic region cDNA library ER-IV for *Nat*-positive clones

The mouse embryonic region cDNA library ER-IV (Harrison *et al.*, 1995) was screened for *Nat*-positive clones, as a first step towards isolation of full length cDNAs encoding for murine NAT isoenzymes. The library was constructed using a whole 7.5 days post coitum gastrulating mouse embryo, devoid of extraembryonic tissue, and is available from HGMP-RC (www.hgmp.mrc.ac.uk) in the form of three high-density gridded filters, suitable for screening by a single hybridisation step.

DIG-labelled probes (section 2.2.5.1) were prepared by amplification of the entire coding region of murine *Nat1*, *Nat2* and *Nat3* genes (section 5.2.1). The three *Nat*-specific probes were pooled together and hybridised (section 2.2.5.2) to the library filters. A PCR-amplified (section 5.2.1) and DIG-labelled (section 2.2.5.1) mouse *Gapdh*-specific probe was also used in control hybridisations of the library filters.

Probe labelling efficiency was assessed by a series of mock hybridisations (figure 5.2.7). Optimally labelled probes provided a strong specific signal (figure 5.2.7a and b), combined with the lowest possible background signal (figure 5.2.7c). Hybridisation of the ER-IV library filters with the *Nat*-specific probe mixture did not provide any positive cDNA clones (figure 5.2.8a). However, hybridisation with the *Gapdh*-specific probe provided at least 90 positive cDNA clones (figure 5.2.8b), ensuring reliability of the technique. The small size of the library (only 43,000 clones) may account for the absence of *Nat*-positive cDNA clones.

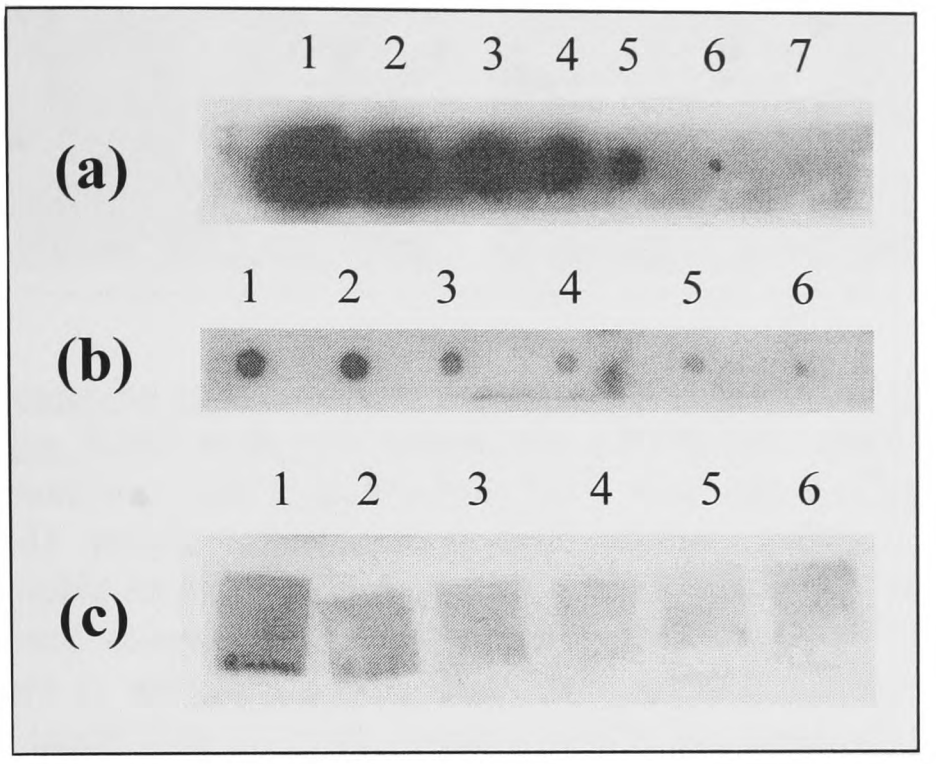


Figure 5.2.7: Mock hybridisation using the DIG-labelled *Nat*-specific probe mixture prepared for screening of the ER-IV cDNA library. a) Mock hybridisation to estimate efficiency of the probe labelling reaction. Equal amounts of gel-purified DIG-labelled PCR products with *Nat1*-specific primers Mus12/Mus13, *Nat2*-specific primers mNAT2-1/mNAT2-910 and *Nat3*-specific primers Mus12/Mus15 were pooled together, to generate a *Nat*-specific probe mixture. Spots 1-7 are 1μl of this mixture, spotted and fixed on a strip of nylon membrane (section 2.2.5.2), diluted in water as follows: "neat" (1), 10^{-1} (2), 10^{-2} (3), 10^{-3} (4), 10^{-4} (5), 10^{-5} (6) and 10^{-6} (7). The probe mixture was sufficiently labelled for use in hybridisations, as it provided a signal even at the 10^{-5} dilution. b-c) Mock hybridisations to determine the optimal concentration of probe in hybridisation reactions (section 2.2.5.2). The probe was initially added to the hybridisation buffer at a concentration of 25ng/ml (1) and then diluted 1:2 (2), 1:5 (3), 1:10 (4), 1:20 (5) and 1:50 (6). b) Spots 1-6 are 1μl of neat (1) or diluted (2-6) probe, spotted and fixed on a strip of nylon membrane (section 2.2.5.2). c) Membrane pieces, lacking DNA, were hybridised overnight with 1ml of neat (1) or diluted (2-6) probe. Development of all membranes was carried out as described in section 2.2.5.2. The optimal dilution of the probe was 1:2 in hybridisation buffer, combining a high specific (spot 2 in b) with a low background (membrane 2 in c) signal in mock hybridisations.

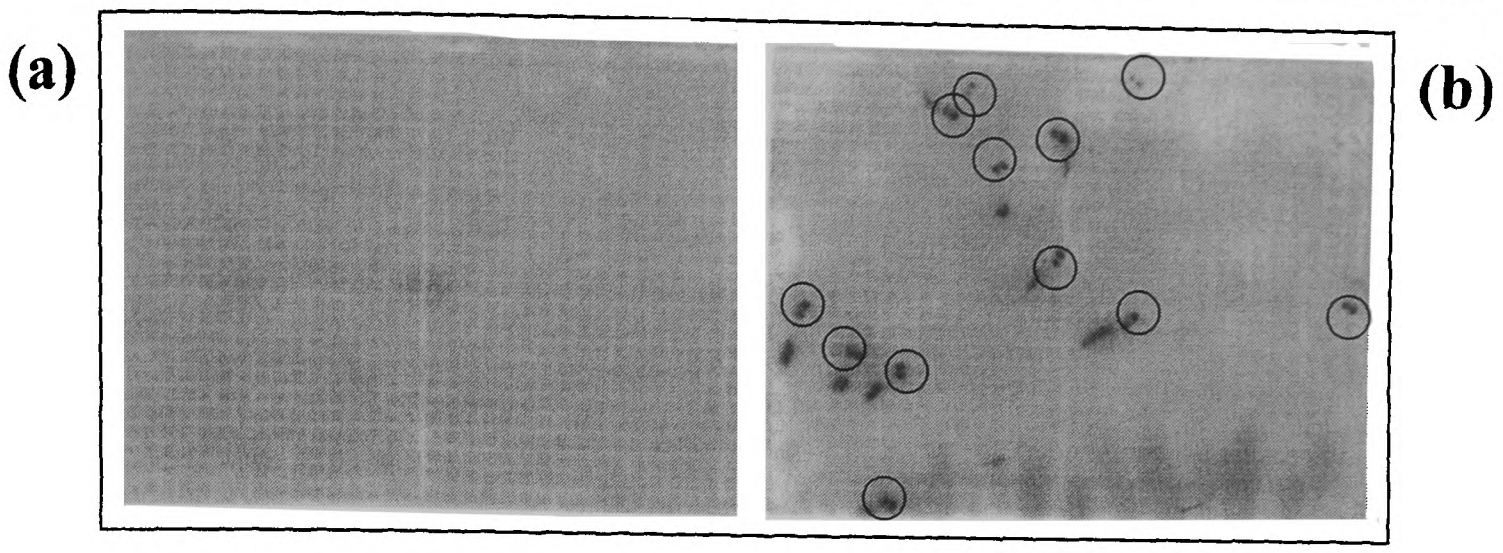


Figure 5.2.8: Screening of the mouse embryonic region cDNA library ER-IV by DNA hybridisation. Entire embryonic tissue from a C57Bl/6J x DBA mouse mating at 7.5 days post coitum was used to isolate total RNA. First strand cDNA was generated using an oligo-dT primer adjacent to a *Not*I adaptor. Following second strand synthesis, the double-stranded cDNA products were ligated to *Sal*I adaptors. The cDNA inserts were cloned into double-digested (*Not*I+*Sal*I) pSPORT vector and electroporated into *E. coli* DH12S cells. The library consists of 43,000 cDNA clones, with an average insert size of 1.3kb (range 0.3-3kb) and has been gridded in a 4x4 array on three nylon membranes, allowing easy determination of the co-ordinates of positive clones. Each clone has been spotted twice, to distinguish from false positives (Harrison *et al.*, 1995). The figure shows part of filter 1 (panels 4 and 6), following hybridisation with a) the *Nat*-specific probe mixture described in figure 5.2.7 and b) a control *Gapdh*-specific probe. No *Nat*-positive clones were identified, as opposed to 12 *Gapdh*-positive clones (in circles).

5.2.4 Analysis of the 5′ untranslated region of the mouse *Nat2* gene

Computational analysis of the sequencing data generated for 129/Ola clone A (section 4.2.2.2) has suggested the presence of a non-coding exon (NCE) located 6.0-6.5kb upstream of the murine *Nat2* coding region. A standard GeneBank search indicated 90% homology between the upstream NCE of the human *NAT2* gene (Ebisawa and Deguchi, 1991; accession no. M75162.1) and region 3337-3376 of 129/Ola clone A (Appendix 2), which is located 6513-6474bp upstream of the mouse *Nat2* coding region. Primer UPSTREAM (table 5.2.1), designed to bind to that region of homology, was used with the *Nat2*-specific primer mNAT2-691 (table 3.2.1) for LA-PCR amplification (section 2.2.3.4) with genomic DNA and cDNA prepared from ES cells of the 129/Ola mouse strain. The expected 7.2kb product was amplified from genomic DNA, but amplification from cDNA was not evident (figure 5.2.9). This indicated that the forward primer UPSTREAM does not bind to a transcribed region of the mouse *Nat2* gene.

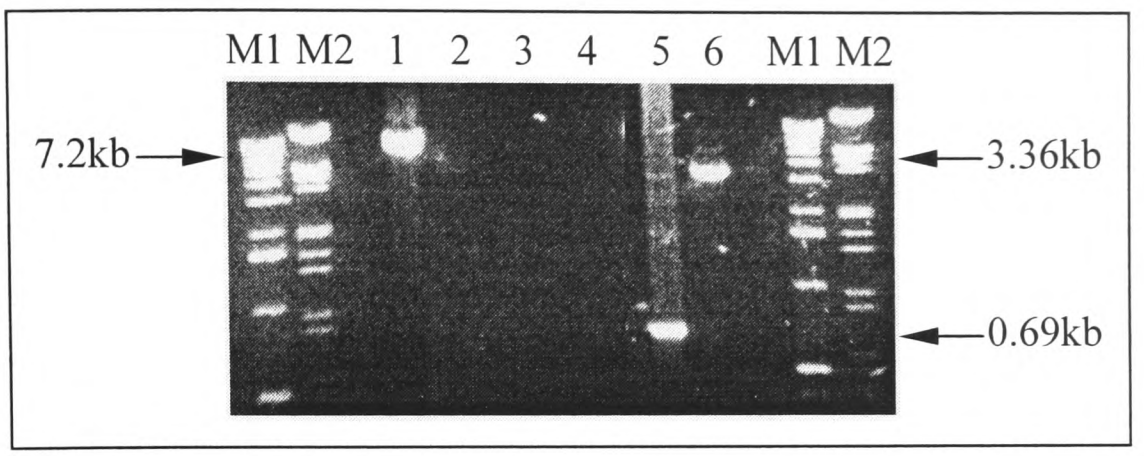


Figure 5.2.9: Search for a predicted NCE located 6513-6474bp upstream of mouse *Nat2* coding region. LA-PCR was performed with forward primer UPSTREAM (table 5.2.1), specific for the upstream region of interest, and reverse primer mNAT2-691 (table 3.2.1), specific for the *Nat2* coding region. Reactions were carried out as described in section 2.2.3.4, using 1.5mM $MgCl_2$. Lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively. Lane 1 shows amplification from 129/Ola genomic DNA, lane 2 from 129/Ola ES cell cDNA and lane 3 from the product of a mock reverse transcription without reverse transcriptase in the reaction mixture. Lane 4 is the PCR negative control. Amplifications from genomic DNA were carried out with primer pairs mNAT2-1/mNAT2-691 (lane 5) and mNAT2-1/PR6(R) (lane 6), as positive controls. Amplification from cDNA was not observed, indicating that the analysed region is not part of the *Nat2* transcript.

Search of the dbEST database identified an EST (accession no. AI006867), showing 99.7% homology with region 3512-3853 of 129/Ola clone A (Appendix 2), which is located 6338-5997bp upstream of the *Nat2* coding region. The EST has originated from clone 1363318 of Soares 2NbMT cDNA library, constructed with RNA isolated from the thymus of a 4-week old male C57Bl/6J mouse. Primer NCE-F (table 5.2.1), specific for the EST, was used with *Nat2*-specific primer mNAT2-691 (table 3.2.1) for LA-PCR with genomic DNA and cDNA prepared from ES cells of the 129/Ola mouse strain. A 7kb product was amplified from genomic DNA, as expected, and a much smaller 0.8kb product from cDNA (figure 5.2.10a), indicating splicing of an approximately 6.1kb intron between the analysed upstream region and the intronless *Nat2* coding region. Sequencing of the RT-PCR product (EMBL accession no. AJ251710), confirmed presence of a NCE, at least 126bp in size, and identified the splice sites of the 6.1kb intron. The donor splice site (CAG/GTGATT) is located at position 3678 of 129/Ola clone A, i.e. 6172bp upstream of the *Nat2* coding region, while the acceptor splice site (TAG/G) is located just 6bp upstream of

the *Nat2* translation initiation codon (figure 5.2.10b). The sequence of both splice sites matches that of the consensus (donor: $^{C}/_{A}$AG/GTRAGT; acceptor: YAG/G). A branch point sequence (consensus: YNYTRAY) and a polypyrimidine stretch (section 1.5.2.2) are present at positions 9763-9769 (TTTTAAC) and 9813-9840 (75% pyrimidine-rich) of 129/Ola clone A, respectively, i.e. within the 90bp intronic region immediately upstream of the acceptor splice site (Nevins, 1983; Smith *et al.*, 1989; Hodges and Bernstein, 1994; Norton, 1994; Elliott, 2000).

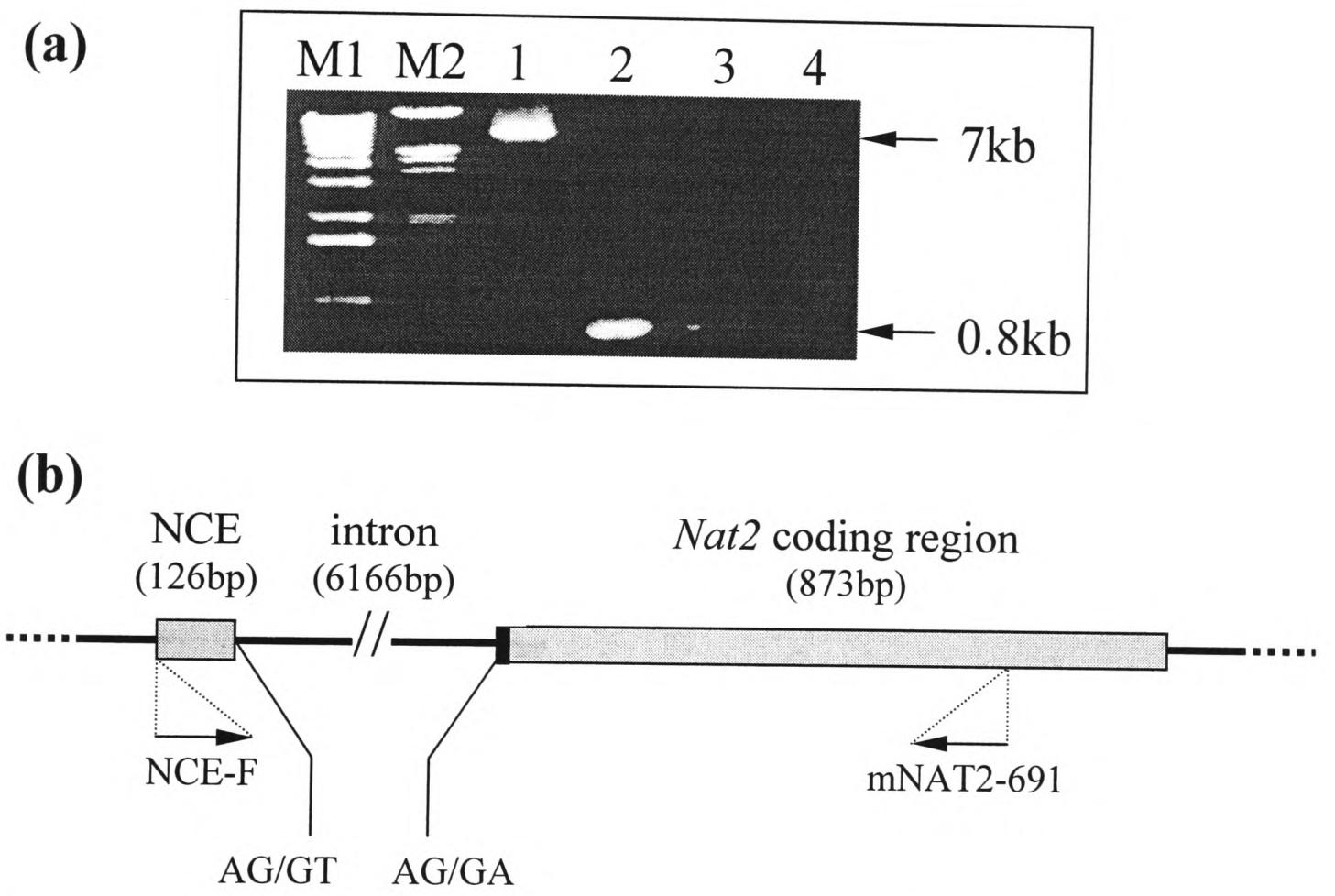


Figure 5.2.10: Identification of an upstream NCE for the mouse *Nat2* gene. a) LA-PCR was performed with the EST-specific forward primer NCE-F (table 5.2.1) and the *Nat2*-specific reverse primer mNAT2-691 (table 3.2.1). Reactions were carried out as described in section 2.2.3.4, using 1.5mM $MgCl_2$. Lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively. Lane 1 shows amplification from 129/Ola genomic DNA, lane 2 from 129/Ola ES cell cDNA and lane 3 from the product of a mock reverse transcription without reverse transcriptase in the reaction mixture. Lane 4 is the PCR negative control. A 7kb product was amplified from genomic DNA, as expected. Amplification from cDNA provided a 0.8kb product, indicating presence of an upstream NCE separated from the *Nat2* coding region by a 6.1kb intron. b) Exon-intron structure of the mouse *Nat2* gene, as confirmed by sequencing analysis of the above RT-PCR product (EMBL accession no. AJ251710). The position of the upstream NCE, as well as of the donor and acceptor splice sites is shown relative to the *Nat2* coding region. The position of the primers used for amplification is also indicated.

A number of tissues of the Balb/c mouse were subsequently screened for splicing of the *Nat2* transcript, using the same RT-PCR approach. Presence of the *Nat2* transcript described in figure 5.2.10 was confirmed in the liver, spleen, submaxillary gland, kidney, brain, thymus, lung and placenta, by sequencing of the major RT-PCR products shown in figures 5.2.11 and 5.2.12. In the heart, RT-PCR amplification was not evident with primer pair NCE-F/mNAT2-691, designed to detect the *Nat2* NCE, although a positive reaction was obtained with primer pair mNAT2-1/mNAT2-910, specific for the *Nat2* coding region (figure 5.2.11). It seems likely that the *Nat2* transcript, although present in the heart, does not contain the upstream NCE described in figure 5.2.10. A number of less intense bands on the RT-PCR gels (figure 5.2.12) have not been sequenced yet, but may correspond to differentially spliced *Nat2* transcripts. The identified EST from Soares thymus cDNA library (accession no. AI006867) extends at least 175bp downstream of the NCE described in figure 5.2.10 and should, therefore, correspond to an alternative *Nat2* transcript. The approximately 1.4kb product, amplified from the cDNA of the thymus and other tissues (figure 5.2.12b), could be the derivative of such alternative splicing. Sequencing of all RT-PCR products from the above tissues may allow identification of additional transcripts for murine *Nat2*.

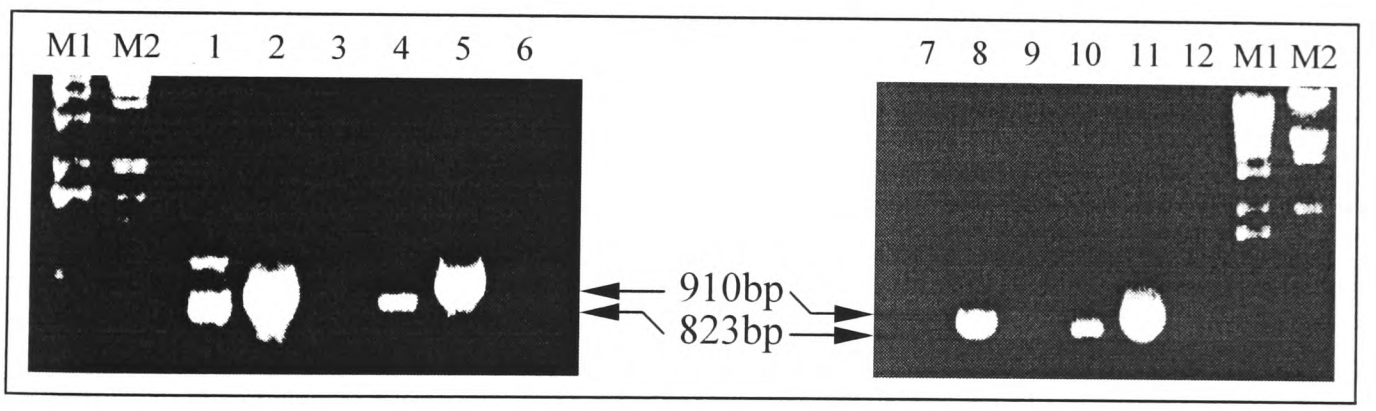


Figure 5.2.11: Screening of Balb/c mouse tissues for presence of the *Nat2* NCE. RT-PCR was performed with cDNA template from liver (lanes 1-3), spleen (lanes 4-6), heart (lanes 7-9) and placenta (lanes 10-12). Lanes 1, 4, 7 and 10 are the RT-PCR products with primers NCE-F and mNAT2-691, designed to detect the upstream NCE of mouse *Nat2* (figure 5.2.10). Lanes 2, 5, 8 and 11 are the RT-PCR products with primers mNAT2-1 and mNAT2-910, specific for the intronless coding region of *Nat2*. Lanes 3, 6, 9 and 12 show amplification with primers mNAT2-1 and mNAT2-910 from the product of a mock reverse transcription. Lanes M1 and M2 are 1kb and λ/*Eco*RI+*Hind*III size markers, respectively. The *Nat2* gene is expressed in all tissues analysed, as indicated by the 910bp band in lanes 2, 5, 8 and 11. Transcription of the *Nat2* NCE was evident in the liver, spleen and placenta, as indicated by the 823bp band in lanes 1, 4 and 10, respectively. The *Nat2* NCE is not transcribed in the heart, as indicated by absence of the 823bp band in lane 7.

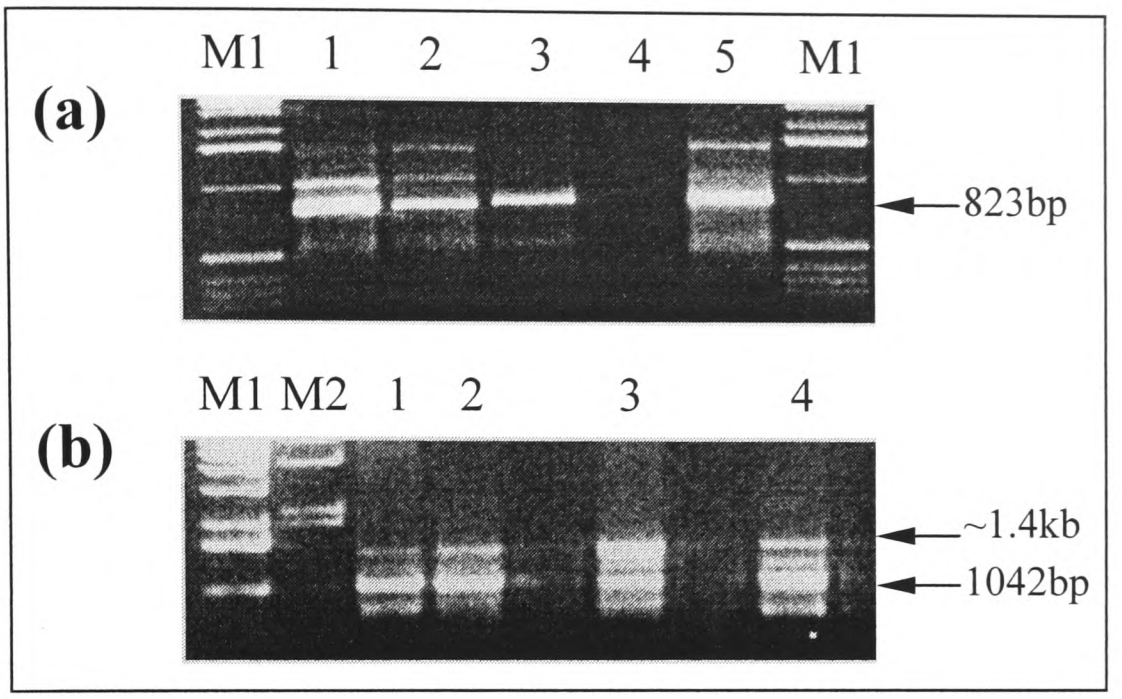


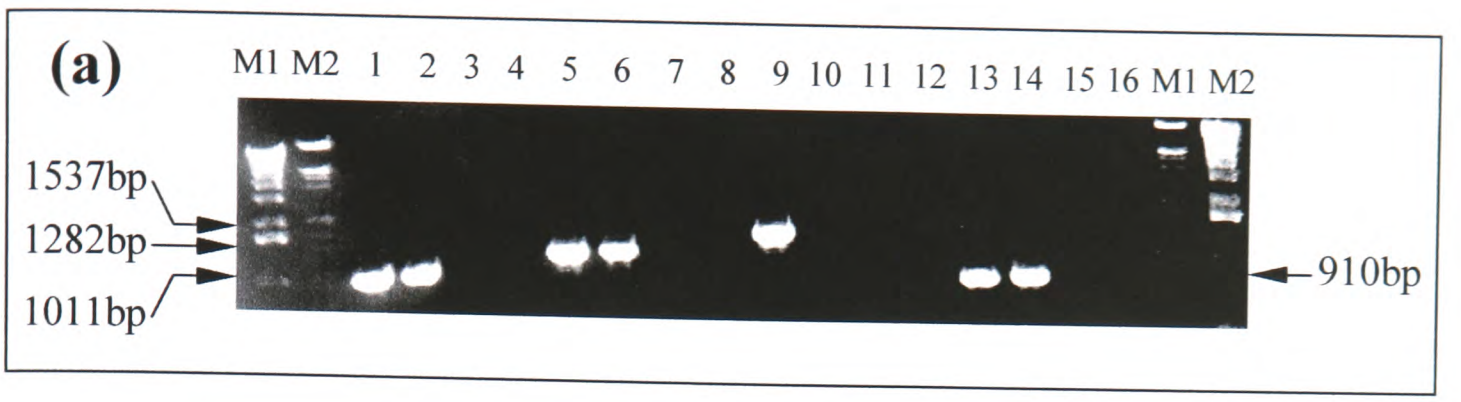
Figure 5.2.12: Investigation for putative alternative transcripts of the mouse *Nat2* gene. RT-PCR was performed as described in figure 5.2.11, and the entire NCE-specific products were loaded to agarose gels for subsequent purification (section 2.2.1.3) and sequencing (section 2.2.3.5). a) RT-PCR product with primers NCE-F and mNAT2-691 from liver (lane 1), spleen (lane 2), submaxillary gland (lane 3), heart (lane 4) and placenta (lane 5). b) RT-PCR product with primers NCE-F and mNAT2-910 from kidney (lane 1), brain (lane 2), thymus (lane 3) and lung (lane 4). Presence of the *Nat2* transcript, shown in figure 5.2.10, was confirmed by sequencing of the 823bp (a) and 1042bp (b) RT-PCR products. Other bands visible on the gels (e.g. the approximately 1.4kb band in (b) could be amplification products from alternative *Nat2* transcripts, but have not been sequenced yet.

5.2.5 Analysis of the 3′ untranslated region of the mouse *Nat2* gene

Analysis of the sequence up to 1000bp downstream of the mouse *Nat2* coding region indicated the presence of six putative polyA signals, located 193bp (polyA-1), 438bp (polyA-2), 442bp (polyA-3), 512bp (polyA-4), 513bp (polyA-5) and 767bp (polyA-6) downstream of the stop codon of the *Nat2* gene, i.e. at positions 10915, 11160, 11164, 11234, 11235 and 11489 of 129/Ola clone A (Appendix 2). The first and fifth putative polyA signals have the less typical ATTAAA sequence (found in about 10% of polyA sites), the second and fourth signals have the rare AATTAA

sequence (found in less than 2% of polyA sites), while the third and sixth signals have a sequence identical to the consensus AATAAA (Birnstiel *et al.*, 1985; Proudfoot and Whitelaw, 1988; Humphrey and Proudfoot, 1988; Edwalds-Gilbert *et al.*, 1997; Zhao *et al.*, 1999).

A RT-PCR approach was undertaken, to determine the approximate size of the *Nat2* 3'-UTR (figure 5.2.13a). Reverse primer polyA-1R is specific for the region just upstream of the first putative polyA signal (polyA-1), while reverse primer polyA-2R is specific for the region just upstream of polyA-2, polyA-3, polyA-4 and polyA-5 putative signals, clustered together within a 80bp region (position 11160-11240 of 129/Ola clone A). Reverse primer polyA-3R binds to the region just upstream of putative signal polyA-6 (table 5.2.1; figure 5.2.13b). The forward *Nat2*-specific primer mNAT2-1 (table 3.2.1) was used with each one of the above reverse primers in RT-PCR amplifications from 129/Ola ES cell cDNA. Amplification product was generated with the first two sets of primers, but not with the third, indicating that transcription of the *Nat2* gene stops between signals polyA-2 and polyA-6 (figure 5.2.13). Putative signals polyA-2 and polyA-3 are overlapping, as well as signals polyA-4 and polyA-5. Of these signals, only polyA-3 is identical to the consensus (AATAAA), so it is likely to dominate over the three nearby signals (Birnstiel *et al.*, 1985; Proudfoot and Whitelaw, 1988). Cleavage and polyadenylation of the *Nat2* transcript would, therefore, be expected to occur about 10-30bp downstream of polyA-3 signal (Proudfoot and Whitelaw, 1988; Humphrey and Proudfoot, 1988; Hawkins, 1996; Zhao *et al.*, 1999).



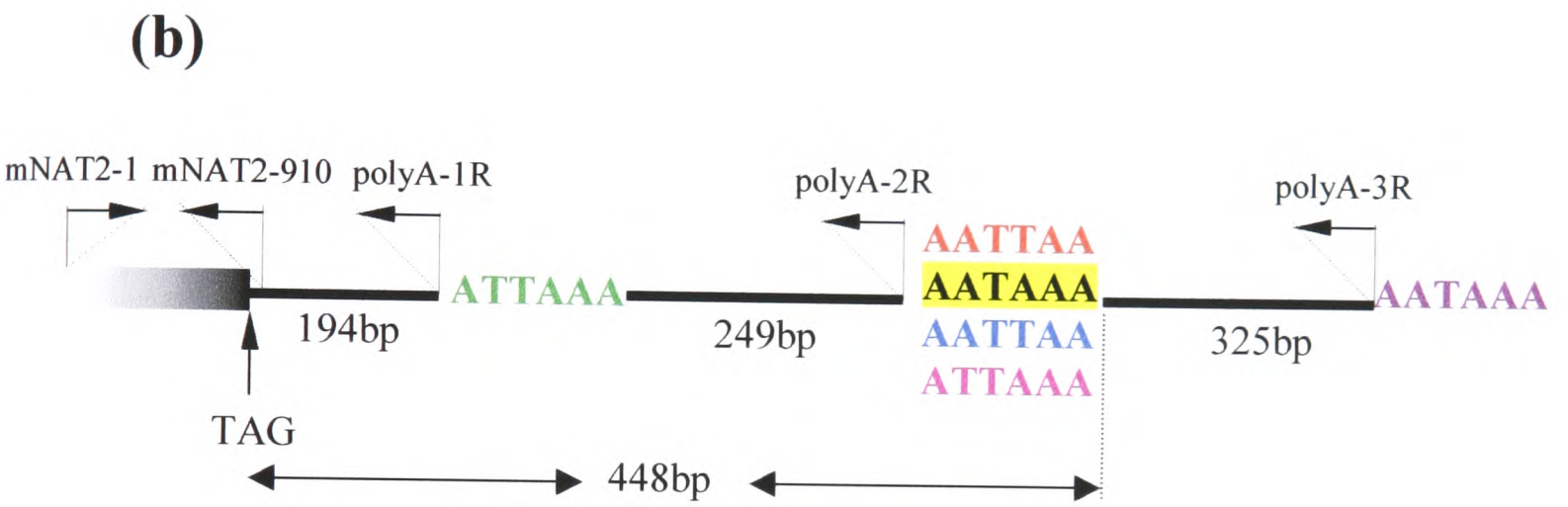


Figure 5.2.13: Identification of a polyadenylation signal for the mouse *Nat2* gene. a) Amplification was carried out using the *Nat2*-specific forward primer mNAT2-1 and reverse primers polyA-1R (lanes 1-4), polyA-2R (lanes 5-8), polyA-3R (lanes 9-12) and the *Nat2*-specific primer mNAT2-910 (lanes 13-16). Genomic DNA (lanes 1, 5, 9 and 13) and cDNA (lanes 2, 6, 10 and 14) from 129/Ola ES cells was used as template, as well as the product of a mock reverse transcription (lanes 3, 7, 11 and 15). Lanes 4, 8, 12 and 16 are PCR negative controls, while lanes M1 and M2 are 1μg of 1kb and λ/*Eco*RI+*Hind*III size markers, respectively. The expected 1011 and 1282bp products were amplified from genomic DNA and cDNA with primers mNAT2-1 and polyA-1R or polyA-2R, respectively, while the 1537bp product expected with primers mNAT2-1 and polyA-3R was amplified only from genomic DNA. Control amplifications shown in lanes 13-16 provided the expected 910bp product from genomic DNA and cDNA. b) Diagram showing the position of putative polyA signals polyA-1 (green), polyA-2 (red), polyA-3 (highlighted in yellow), polyA-4 (blue), polyA-5 (pink) and polyA-6 (violet), located downstream of the *Nat2* coding region (shaded box). The position of the primers used for amplification is indicated with single-headed horizontal arrows. The polyA-3 signal (highlighted in yellow) is likely to be responsible for polyadenylation of the *Nat2* transcript.

5.2.6 Structure of the genes for NAT in different mammalian species

Following the identification of an upstream NCE for murine *Nat2* (section 5.2.4), it was investigated whether other mammalian genes for NAT have a similar exon-intron structure, possibly associated with alternative splicing. Exhaustive search of the GeneBank and dbEST databases (section 1.4.3) provided a number of ESTs, whose sequence matched that of many genes for mammalian NAT. Of the 73 ESTs identified, 32 matched human *NAT1*, 18 human *NAT2*, 11 mouse *Nat2*, 7 rat *Nat1*, 3 bovine *Nat* and 2 porcine *Nat*. At the time of writing, ESTs matching the genomic sequences of murine *Nat1* and *Nat3*, as well as rabbit and hamster *Nat1* and *Nat2*, were not available.

Mouse *Nat2* ESTs (figure 5.2.14) were aligned to the genomic sequence of 129/Ola clone A (Appendix 2). Consistent with the wide tissue distribution observed for murine *Nat2* (section 5.2.1; Chung *et al.*, 1993; Stanley *et al.*, 1997), these ESTs are of variable tissue origin. ESTs BG972080, BF119463, BE633877 and BF164333 partially cover the coding region of the *Nat2* gene. EST BF164333 also contains a short NCE, the splice sites of which are located 6344 (position 3506 of 129/Ola clone A) and 6bp upstream of the *Nat2* coding region. Together, this EST and EST AI006867 (described in section 5.2.4, also shown in figure 5.2.14) provide two donor splice sites, alternative to that identified by RT-PCR and sequencing in section 5.2.4. The acceptor splice site appears to be the same for all of the above *Nat2* transcripts. Five of the identified mouse ESTs (BG094023, BB100623, AV052552, AV377607 and BB436983) end 15-23bp after polyA-3, consistent with the observation in section 5.2.5 that the *Nat2* transcript should be cleaved downstream of this polyA signal.

Genomic (Blum *et al.*, 1990a) and liver cDNA (Ohsako and Deguchi, 1990) sequences have previously been reported for the human *NAT* genes, and a NCE has been found approximately 8kb upstream of the *NAT2* coding region (Ebisawa and Deguchi, 1991). To investigate the possibility that the human *NAT* genes might possess additional NCEs, the identified human *NAT* ESTs were aligned with genomic sequences from the contiguous BAC clones AB020868, AC025062 and AC025648 (figure 1.3), which contain the *NAT1* and *NAT2* genes, plus their surrounding genomic region.

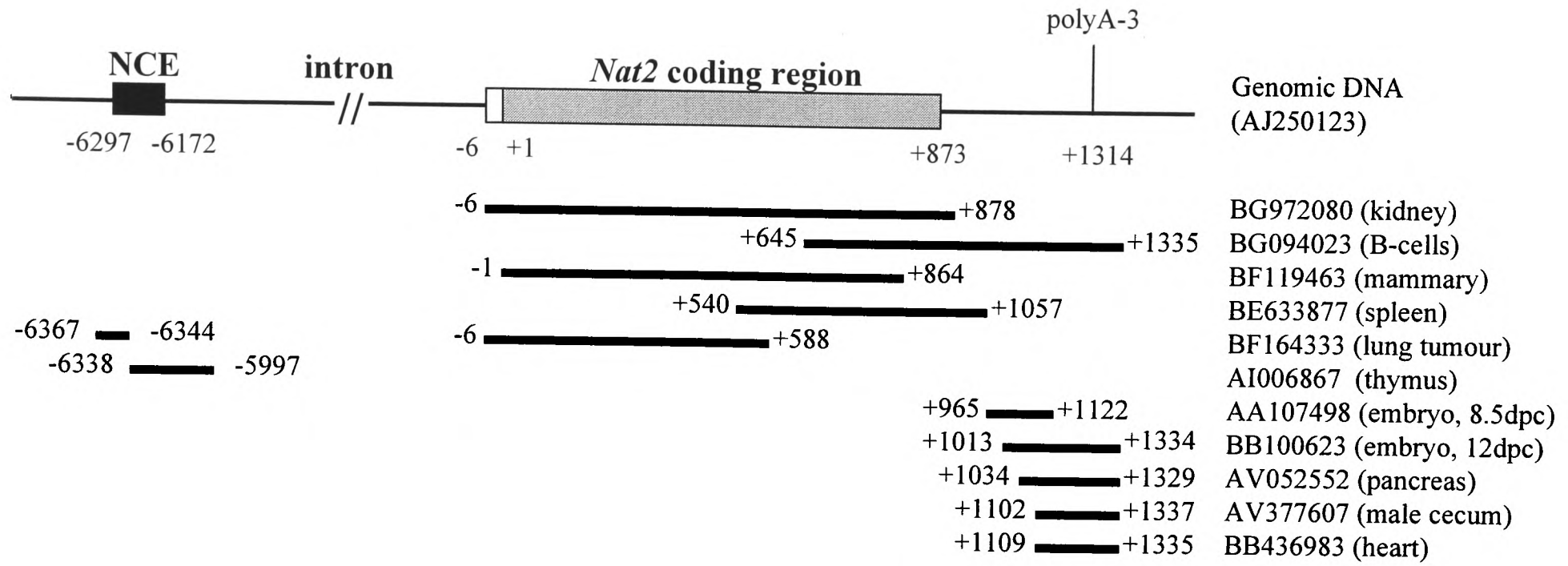


Figure 5.2.14: Alignment of mouse *Nat2* genomic sequence with ESTs deposited in electronic databases. Search of GeneBank and dbEST databases using the sequence of 129/Ola clone A (Appendix 2) as probe, provided 11 ESTs corresponding to parts of the *Nat2* gene. The diagram shows the exact position of each EST, relative to the *Nat2* translation initiation codon (+1). The position of the polyA-3 signal (section 5.2.5) is also indicated. The accession numbers and tissue origin of presented ESTs are listed on the right of the figure.

Of the human *NAT2* ESTs (figure 5.2.15), 10 have 5′ UTRs matching the sequence of the upstream NCE, described by Ebisawa and Deguchi (1991), including both splice sites of the adjacent intron. It was determined that the donor splice site is located exactly 8661bp upstream of the *NAT2* coding region, while the position of the donor splice site is 6bp upstream of the coding region, as previously described (Ohsako and Deguchi, 1990; Ebisawa and Deguchi, 1991). ESTs AI460128, AI733799, BG204539 and BG617259 end 10-48bp after a polyA signal (polyA-2) located 320bp downstream of the *NAT2* stop codon (figure 5.2.15). This is a signal alternative to that described by Ohsako and Deguchi (1990), which is located 198bp downstream of the *NAT2* stop codon (polyA-1 in figure 5.2.15). With the exception of ESTs AU105863, AU099534, AU105864 and BG204539, which are of unspecified tissue origin, all analysed *NAT2* ESTs have originated either from liver or colon tissue, in agreement with previous observations on the tissue expression profile of the NAT2 isoenzyme (section 5.1).

Of the human *NAT1* ESTs (figure 5.2.16), only 2 (BG185950 and BG201534) match the sequence of the *NAT1* cDNA clone, previously isolated by Ohsako and Deguchi (1990), in that they both lack upstream NCEs. However, these ESTs have a longer 5′ UTR, compared with the previously described liver cDNA clone (Ohsako and Deguchi, 1990), probably indicating that the latter does not contain a full-length message. A total of 3 upstream NCEs were identified by comparison of human *NAT1* ESTs with the genomic sequences obtained from BAC clones AB020868 and AC025062. One NCE (NCE3), found in 9 ESTs (figure 5.2.16), is located 2637-2558bp upstream of the *NAT1* coding region and has very well defined splice sites at the 5′ and 3′ ends. One of these ESTs (BG941635) contains two more NCEs, located 10855-10706bp (NCE2) and 11916-11870bp (NCE1) upstream of the *NAT1* coding region (figure 5.2.16). NCE1 is also present in ESTs AV754344, AW956027, BF247489, BG529392, BG655073, R91802 and T29485, and has a well-defined donor splice site at the 3′ end. The size of NCE1 ranges from 17 to 86bp, suggesting that it must form the very 5′ end of the analysed transcripts (figure 5.2.16). The acceptor splice site proximal to the *NAT1* coding region is located 6bp upstream of the translation initiation codon (figure 5.2.16). Furthermore, 10 of the identified *NAT1* ESTs, one of which (AW242986) has a poly-A tail indicating that it must be full-length at the 3′ end, terminate 11-62bp after a polyA signal (polyA-2) located 330bp

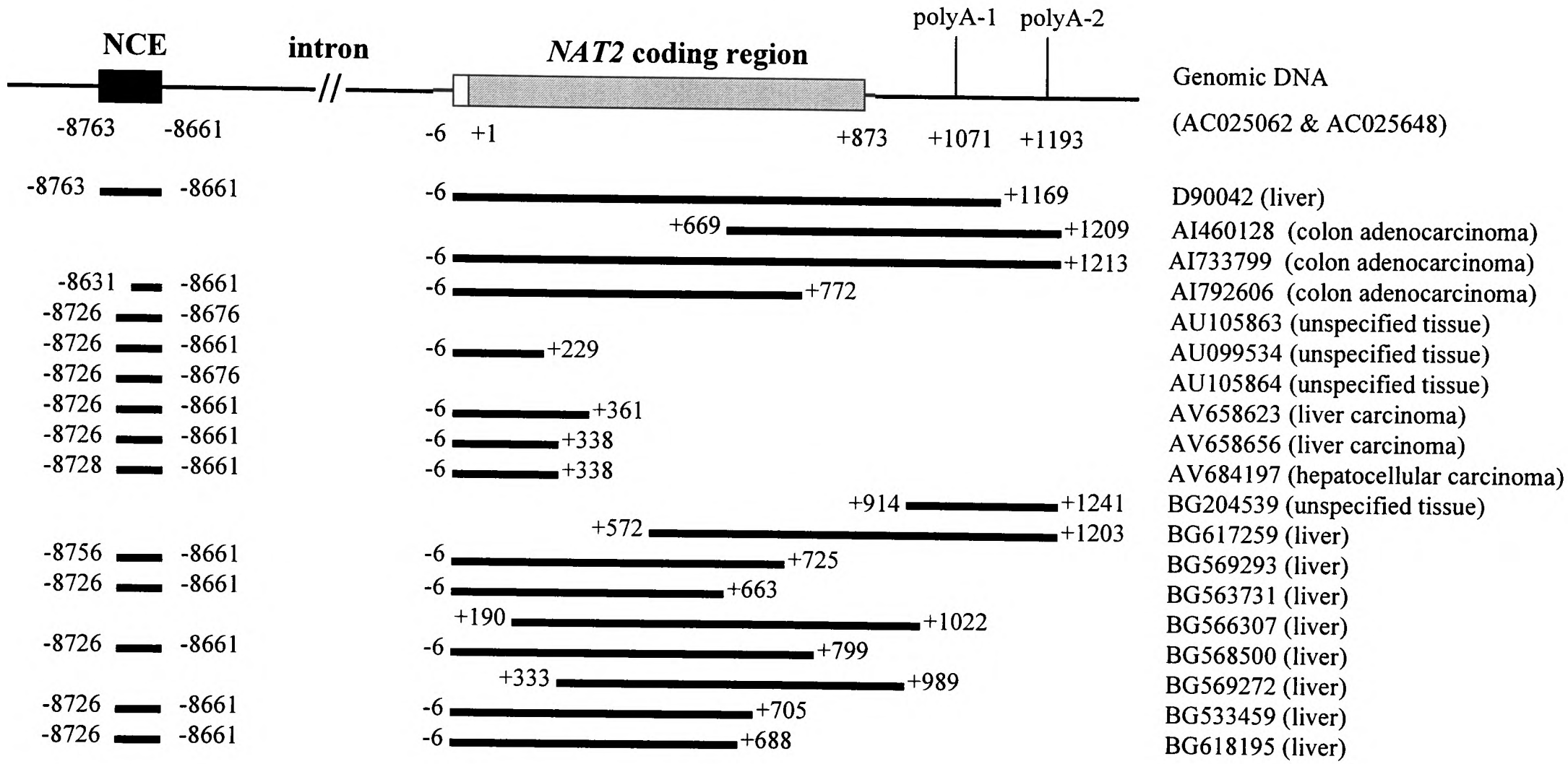
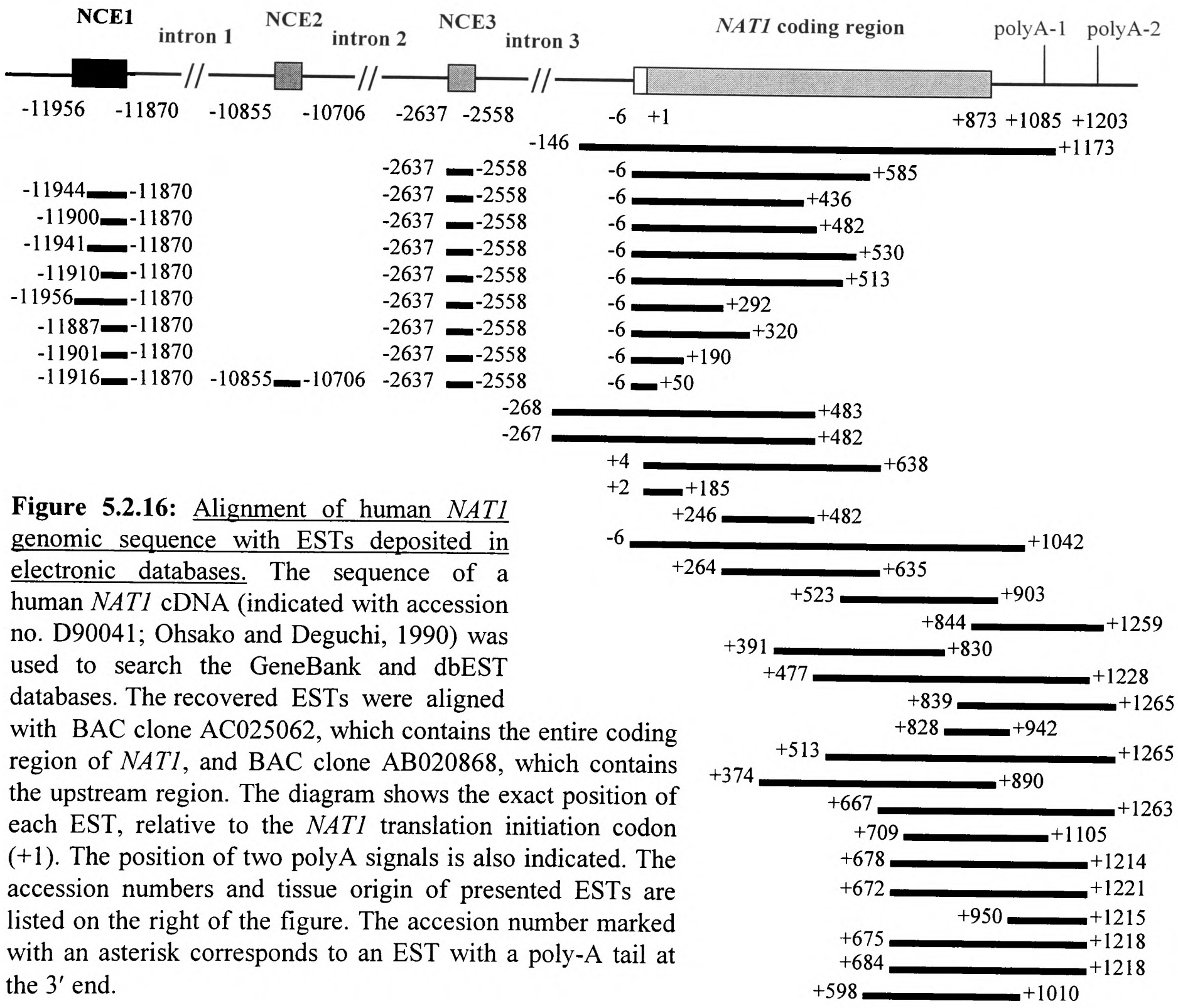


Figure 5.2.15: Alignment of human *NAT2* genomic sequence with ESTs deposited in electronic databases. The sequence of a human *NAT2* cDNA (indicated with accession no. D90042; Ohsako and Deguchi, 1990) was used to search the GeneBank and dbEST databases. The recovered ESTs were aligned with genomic sequences from BAC clones AC025062 and AC025648, which contain the entire *NAT2* coding region. The diagram shows the exact position of each EST, relative to the *NAT2* translation initiation codon (+1). The position of two polyA signals is also indicated. The accession numbers and tissue origin of presented ESTs are listed on the right of the figure.



Genomic DNA
(AB020868 & AC025062)

D90041 (liver)
AL554585 (placenta)
AV754344 (pituitary tumour)
AW956027 (colon tumour)
BF247489 (kidney)
BG529392 (testis carcinoma)
BG655073 (pancreatic islet)
R91802 (fetal liver & spleen)
T29485 (prostate gland)
BG941635 (erythroid cells)
BG185950 (unspecified tissue)
BG201534 (unspecified tissue)
BF126688 (liver)
BG656220 (pancreatic islet)
BG212373 (unspecified tissue)
AI204592 (testis)
T67129 (fetal liver & spleen)
AI861954 (colon tumour)
AI022585 (parathyroid gland)
AI909313 (mammary)
AL554525 (placenta)
AI888631 (stomach carcinoma)
BE775188 (uterus)
*AW242986 (kidney)
AW390667 (stomach)
AW977058 (colon tumour)
AW071587 (mammary)
BF594441 (colon)
BF431254 (pancreatic islet)
BF591924 (kidney)
BG656530 (pancreatic islet)
BG654454 (pancreatic islet)
R79400 (placenta)

Figure 5.2.16: Alignment of human *NAT1* genomic sequence with ESTs deposited in electronic databases. The sequence of a human *NAT1* cDNA (indicated with accession no. D90041; Ohsako and Deguchi, 1990) was used to search the GeneBank and dbEST databases. The recovered ESTs were aligned with BAC clone AC025062, which contains the entire coding region of *NAT1*, and BAC clone AB020868, which contains the upstream region. The diagram shows the exact position of each EST, relative to the *NAT1* translation initiation codon (+1). The position of two polyA signals is also indicated. The accession numbers and tissue origin of presented ESTs are listed on the right of the figure. The accesion number marked with an asterisk corresponds to an EST with a poly-A tail at the 3′ end.

downstream of the stop codon of the gene (figure 5.2.16). This is a signal alternative to that previously localised 212bp downstream of the *NAT1* stop codon (polyA-1) by Ohsako and Deguchi (1990). In agreement with previous studies (section 5.1), the identified *NAT1* ESTs are of diverse tissue origin.

Rat ESTs (figure 5.2.17) BF566785 and AW916303 cover part of the *Nat1* coding region (nucleotide position 1-465 and 605-733, respectively), but the former also contains a 31bp NCE, the splice junction of which is located 6bp upstream of the translation initiation codon. This is a novel upstream NCE, since it does not align with any of the sequences previously described for five NCEs of the rat *Nat* genes (Ebisawa *et al.*, 1995). ESTs BF282319, BI301163, AI180295 and AW434389 cover region 173-698 downstream of the *Nat1* stop codon. ESTs BI301163 and AW434389 have a poly-A tail, indicating that they must be full-length at the 3′ end. However, EST AA892090 extends 760bp downstream of the *Nat1* stop codon, suggesting the presence of an alternative polyadenylation site.

The 2 identified porcine ESTs (BI344374 and BE014631) cover part of the coding region of an unknown *Nat* gene and do not provide information about any potential upstream NCEs. The former EST is of pooled adult tissue origin (testis, ovary, endometrium, placenta, hypothalamus and pituitary), while the latter is of pooled embryonic tissue origin (11-, 13-, 15-, 20- and 30-day embryos).

The 3 identified bovine ESTs (figure 5.2.18) show perfect match with the partially sequenced coding region of a previously described *Nat* gene (Upton and Sim, 1999). ESTs BG689007 and AW484954 also contain parts of the 5′ UTR, allowing identification of three novel NCEs (figure 5.2.18). EST BG689007 has two upstream NCEs. EST AW484954 shares the same NCE proximal to the coding region, but has a different distal NCE. In both cases, the junction between the proximal non-coding and the coding exon is located 5bp upstream of the translation initiation codon. The junction between the proximal and distal NCEs is located 88bp upstream of the translation initiation codon.

Additional NCEs have previously been reported for rabbit *Nat2*, namely two found on a single cDNA clone described by Blum *et al.* (1989a; 1990b) and a third one described by Sasaki *et al.* (1991). NCEs have also been reported for the *Nat1* and

Nat2 genes of the hamster (Land *et al.*, 1994; Nagata *et al.*, 1994), suggesting that the 5′ UTR of potentially all mammalian *Nat* transcripts may be created via alternative splicing.

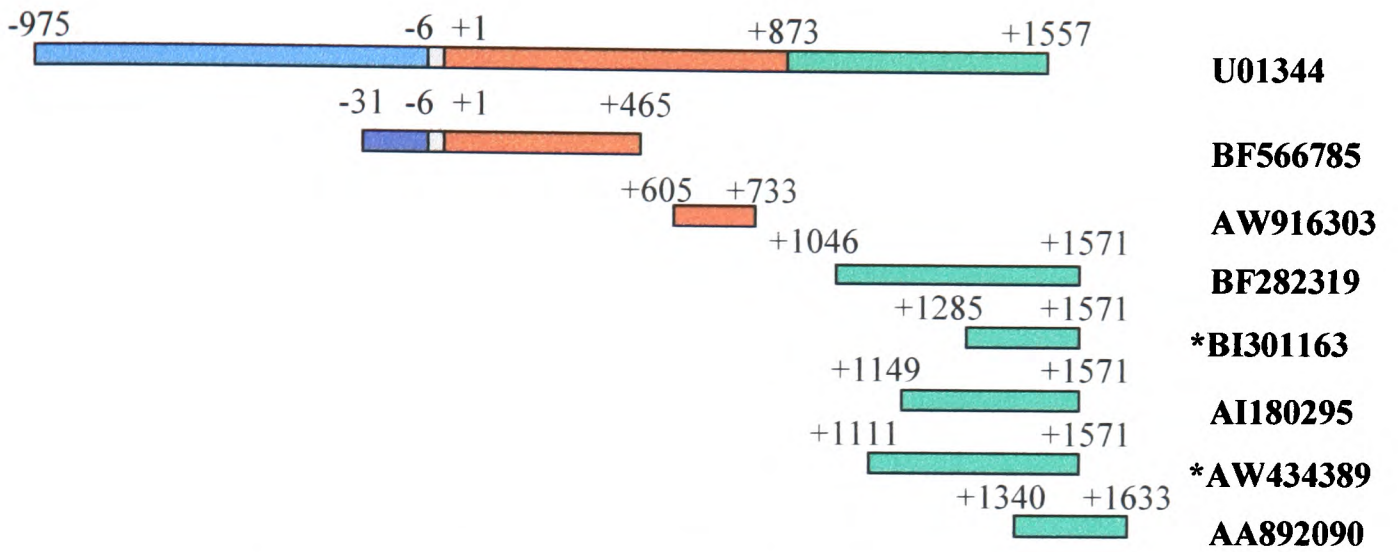


Figure 5.2.17: Alignment of a rat *Nat1* cDNA sequence with ESTs deposited in electronic databases. Sequences of five *Nat1* and *Nat2* cDNA clones from rat pineal gland (Ebisawa *et al.*, 1995) were used as probes to search the GeneBank and dbEST databases. The recovered ESTs are shown in comparison with one of the five cDNA clones (U01344), and their position is indicated relative to the *Nat1* translation initiation codon (+1). The accession numbers of ESTs are listed on the right of the figure. Accession numbers marked with an asterisk correspond to ESTs with a poly-A tail at their 3′ end. Blue is the 5′ UTR, red is the *Nat1* coding region and green is the 3′ UTR. Distinct sequences of a specific region are indicated with different shades of the same colour. ESTs BF566785, BI301163, AI180295 and AW434389, are of unspecified tissue origin, while ESTs AW916303 and BF282319 are of mixed tissue origin (brain, ovary, placenta, kidney, lung, liver, heart, muscle, spleen and embryo), and EST AA892090 has been derived from the kidney.

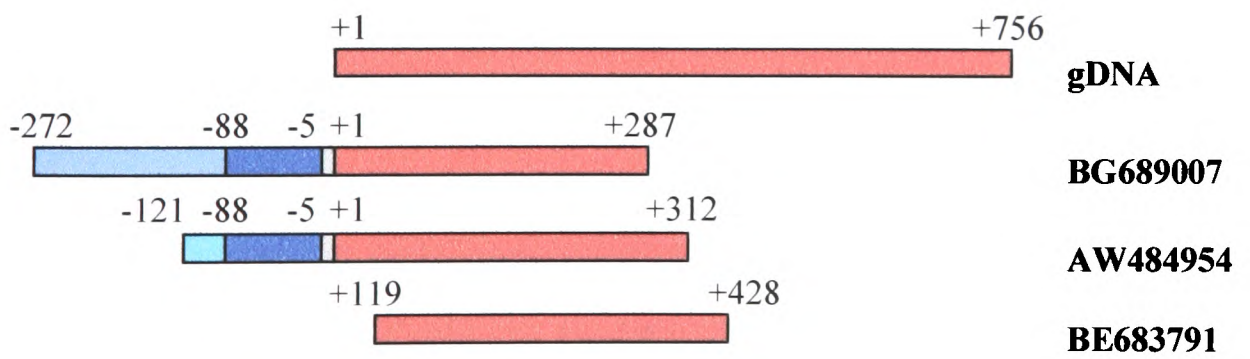


Figure 5.2.18: Alignment of a bovine *Nat* genomic sequence with ESTs deposited in electronic databases. The partial coding sequence of a previously identified bovine *Nat* gene (Upton and Sim, 1999) was used to search the GeneBank and dbEST databases. The relative position of the recovered ESTs is shown in reference to the translation initiation codon (+1) of the *Nat* coding region. The accession numbers of the ESTs are listed on the right of the figure. Blue is the 5′ UTR and red is the *Nat* coding region. Distinct sequences of a specific region are indicated with different shades of the same colour. EST BG689007 is of mammary tissue origin, AW484954 is of mixed tissue origin (marrow, macrophage, ovary and fetal muscle) and BE683791 has derived from 20- and 40-day embryos.

5.3 Discussion

It is generally accepted that biotransformation of xenobiotic compounds is mediated by enzymes of the liver, and that reactive metabolites are then transported to extrahepatic tissues, where they can exert their toxic effects (Hirvonen, 1999). For example, the slow NAT2 phenotype in humans has been associated with an increased risk for arylamine-induced bladder cancer (section 1.2.4.1), although NAT2 protein is not readily detected in the bladder (Stanley *et al.*, 1996) or in cultured urothelial cells (Coroneos and Sim, 1991). On the other hand, recent analyses of expression profiles have demonstrated presence of many xenobiotic metabolising enzymes in extrahepatic tissues, indicating that full activation of carcinogens may take place directly in the tissue undergoing transformation (Windmill *et al.*, 1997; Hirvonen, 1999). It has been postulated, for instance, that expression of both human NAT isoenzymes in the colon (Ilett *et al.*, 1994; Hickman *et al.*, 1998) may be responsible for local bioactivation of dietary pro-carcinogens, leading to colorectal carcinomas (Minchin *et al.*, 1993; Bell *et al.*, 1995b). Several studies have also investigated expression of human NAT1 in other tissues prone to chemical carcinogenesis, such as the mammary gland (Sadrieh *et al.*, 1996; Lee *et al.*, 1997; Williams *et al.*, 2001) and lung (Macé *et al.*, 1998).

The mouse has been a popular model in studies investigating involvement of NAT in chemical carcinogenesis (Levy and Weber, 1989; Flammang *et al.*, 1992; Levy and Weber, 1992; Levy *et al.*, 1993; 1994; Leff *et al.*, 1999a; Li *et al.*, 2001) and toxicity (Karolyi *et al.*, 1987; 1988; 1990; Goebel *et al.*, 1999). The RT-PCR analysis presented in section 5.2.1 demonstrated widespread expression of the murine *Nat2* gene, in agreement with previous studies measuring NAT2 enzymatic activity and immunoreactive NAT2 protein (Chung *et al.*, 1993; Stanley *et al.*, 1997). In contrast, the *Nat1* gene is expressed only in the liver, possibly at a lower rate than *Nat2*, as a second PCR round was necessary for sufficient amplification of the *Nat1* transcript. Although PCR is a very sensitive technique, two rounds of amplification are often needed for detection of transcripts under-represented in the mRNA population of complex tissues (Sosinsky *et al.*, 2000; Fletcher *et al.*, 2001). Previous Northern blot analysis has also indicated that the *Nat1* transcript is less abundant than *Nat2* in the liver (Martell *et al.*, 1992).

Ubiquitous expression of the mouse *Nat2* gene is supportive of the hypothesis that its protein product may be involved in endogenous metabolism (Payton *et al.*, 1999b). Expression of the *Nat2* gene has been detected in embryos (gestational day 10 to term) by RT-PCR (Mitchell *et al.*, 1999). NAT2 activity then increases steadily until postnatal day 80 (Estrada *et al.*, 2000). More importantly, immunohistochemical staining with antiserum 184 (Stanley *et al.*, 1997) has detected NAT2 protein in the developing neural tube of mouse embryos around the time of closure (9.5-13.5 days of gestation) (Stanley *et al.*, 1998). This is consistent with the potential involvement of the NAT2 isoenzyme in folate metabolism (Payton *et al.*, 1999b). Similar observations have been made for humans, with NAT1 activity present in tissues of second-trimester embryos (Pacifici *et al.*, 1986).

Early embryonic expression of murine *Nat2* is of key importance, if the isoenzyme is to play a role early in development, during neurulation. It has been demonstrated here (section 5.2.2) and in a previous study (Payton *et al.*, 1999b) that mouse *Nat2* is transcribed in ES cells of the preimplantation mouse embryo. Substantial NAT2 enzymatic activity and immunoreactive protein has also been detected in mouse ES cells (Nichola Johnson, D.Phil. Thesis, Oxford 2001). In contrast, the *Nat1* and *Nat3* genes are not expressed at this early developmental stage (section 5.2.2). The situation is analogous to that in humans, where cDNA library screening has demonstrated expression of human *NAT1* (Smelt *et al.*, 2000), but not *NAT2* (section 5.2.2) at the blastocyst stage. Placental expression of human *NAT1* (Pacifici *et al.*, 1986; Derewlany *et al.*, 1994; Smelt *et al.*, 1997; 1998; 2000; Upton *et al.*, 2000) and murine *Nat2* (section 5.2.1) may also modulate folate availability to the developing foetus, as placental NAT1 in humans has been shown to be metabolically active towards pABGlu (Smelt *et al.*, 2000; Upton *et al.*, 2000). Moreover, placental acetylation may protect or compromise embryonic development via detoxification or bioactivation, respectively, of potentially teratogenic compounds found in drugs, food or cigarette smoke (Derewlany and Koren, 1994).

Transcription of the mouse *Nat3* gene was demonstrated in the adult spleen (section 5.2.1). However, NAT3 protein was not detected in spleen homogenates by Western blot analysis (section 5.2.1). It is possible that NAT3 protein is present in the spleen, but at levels below the detection limit of the polyclonal antiserum used. A

recent preliminary study (Windridge *et al.*, 2001) has reported presence of NAT3 protein in the developing neural tube of mouse embryos by immunohistochemistry. Since recombinant NAT3 from Balb/c, MCA and MSP mouse strains has very low or no activity towards common NAT substrates (section 3.2.7), it is worth investigating whether it has evolved to perform functions other than the acetylation of arylamines. This would be a situation analogous to the S-crystallins of the lens, which have diverged from glutathione S-transferases (Tomarev and Piatigorsky, 1996).

Genes with an intronless coding region, like the genes for NAT, are rare in higher eukaryotes and are estimated to represent less than 5% of all genes in humans (Rampazzo *et al.*, 2000). Most genes with an intronless coding region belong to the G-protein-coupled receptor (Mummidi *et al.*, 1997; Sawzdargo *et al.*, 1999; Schöneberg *et al.*, 1999; Nilsson *et al.*, 2000) and olfactory receptor (Sosinsky *et al.*, 2000) families. Other examples of genes with intronless coding region include the human insulinoma-associated *IA1* gene (Lan *et al.*, 1994), the mouse and human genes for necdin (Nakada *et al.*, 1998), the *Amd-2* gene for murine S-adenosylmethionine decarboxylase (Persson *et al.*, 1999), the human choroideremia-like Rab escort protein 2 *CHML* gene (Halford *et al.*, 2001), as well as the human *C14orf4* (Rampazzo *et al.*, 2000) and the murine *RP42* genes (Mas *et al.*, 2000), as yet of unknown function. The histone genes are also intronless, but differ in that they possess their own unique elements regulating transcription and RNA-processing (Huang and Carmichael, 1997). Many gene families can be divided into subgroups on the basis of the presence or absence of introns. For example, type I interferon genes are intronless, in contrast with type II genes which possess introns (Roberts *et al.*, 1998). The *myc* family also contains genes with or without introns in their coding region (Sugiyama, *et al.*, 1999). Similarly, the gene for hsp90α heat-shock protein contains 10 introns, while the gene for hsp70 is intronless (Jolly *et al.*, 1999).

The majority of intronless genes have derived from retrotransposition, which involves incorporation of the processed transcript of a gene back into the genome, via reverse transcription (Rogalla *et al.*, 2000). Retrotransposition usually leads to the formation of pseudogenes (Jeong *et al.*, 1998; Fletcher *et al.*, 2001), but in rare cases the reintegrated transcript may continue to be active (Chiang *et al.*, 1998; Makeyev *et al.*, 1999; Sedlacek *et al.*, 1999; Gray *et al.*, 2000). Intronless retrosposons usually

retain part of the poly-A tail of the original message and are flanked by elements, such as LINEs, SINEs, LTRs and *Alu* repeats (Herbert and Rich, 1999; Rogalla *et al.*, 2000; Sosinsky *et al.*, 2000). Murine *Nat2* has the characteristics of a retrosposon, including an intronless coding region and the remnants of a long poly-A tail (located 1346bp downstream of the stop codon), as well as flanking SINE/B4, LTR/MaLR and SINE/*Alu* repeat elements (section 4.2.2.2). The human *NAT* genes share similar characteristics, and this may also be the same for other eukaryotic *Nat* genes, for which extensive sequencing information is not yet available. However, search of current genome databanks (data not shown) has failed to indicate a gene that could have been the intron-containing ancestor of eukaryotic *Nat* genes. Given their very high degree of conservation in both prokaryotic and eukaryotic genomes (section 1.3), the genes for NAT may have evolved from a common intronless ancestor.

Upstream NCEs are common among mammalian genes. Examples are the rabbit gene for protein inhibitor of neuronal nitric oxide synthase (*PIN*) (Jeong *et al.*, 1998), the rat glucocorticoid receptor (*GR*), estrogen receptor (*ER*) and androgen-regulated acidic epididymal glycoprotein (*Crisp-1*) genes (Gearing *et al.*, 1993; Hirata *et al.*, 1996; Klemme *et al.*, 1999), the murine TATA-binding protein (*TBP*) and serotonin transporter (*5-HTT*) genes (Ohbayashi *et al.*, 1996; Bengel *et al.*, 1997), as well as the human genes for aromatic L-amino acid decarboxylase (*AADC*) (Le Van Thai *et al.*, 1993), parathyroid hormone-related peptide (*PTHrP*) (Bui *et al.*, 1996), β-galactoside α2,6-sialyltransferase (*SIAT1*) (Lo and Lau, 1996) and several serine protease inhibitors (Hayashi and Suzuki, 1993). Intronless genes with upstream NCEs include members of the G-protein coupled receptor (Mummidi *et al.*, 1997; Nilsson *et al.*, 2000) and olfactory receptor (Sosinsky *et al.*, 2000) gene families, as well as the human connexin-26 (*CX26*) and CC chemokine receptor 3 (*CCR3*) genes (Kiang *et al.*, 1997; Zimmermann *et al.*, 2000).

The work described in this Chapter shows that the genes for mammalian NAT isoenzymes have a conserved structure, consisting of a single coding exon and one or more upstream NCEs. This was demonstrated both experimentally (section 5.2.4) and by analysis of ESTs deposited in electronic databases (section 5.2.6). Sequencing of a RT-PCR product (section 5.2.4) provided the exact position of an upstream NCE for the mouse *Nat2* gene. Although a donor splice site was identified by analysis of this

particular transcript (section 5.2.4), there appear to be additional donor splice sites, as indicated by comparison of the experimentally obtained cDNA sequence (EMBL accession no. AJ251710; figure 5.2.10) with ESTs BF164333 and AI006867 (figure 5.2.14). This was further supported by the presence of multiple bands on the RT-PCR gels (figure 5.2.12), possibly corresponding to differentially spliced transcripts of murine *Nat2* in adult tissues. Moreover, although *Nat2* is expressed in the heart, its transcription does not appear to begin from the same NCE as in other tissues (figure 5.2.11). It is possible that transcription of the *Nat2* gene in the heart involves a different NCE. Alternatively, transcription may be initiated from a site close to the *Nat2* coding region, as suggested by recent RNase mapping of several transcription initiation sites within 112-151bp upstream of the *Nat2* coding region (Estrada-Rodgers *et al.* 1998a).

Comparison of ESTs with the genomic region upstream of the human *NAT2* gene has provided the exact position (figure 5.2.15) of the previously described upstream NCE (Ohsako and Deguchi, 1990; Ebisawa and Deguchi, 1991). It appears that human *NAT2* makes use only of this particular NCE, although alternative donor splice sites may be involved in the splicing of the intervening 8.6kb intron (figure 5.2.15). In contrast, human *NAT1* appears to produce several transcripts, which are initiated either immediately upstream of the coding region or from at least three upstream NCEs with well-defined splice sites (figure 5.2.16). It is currently unclear whether this complex transcription pattern of human *NAT1* is associated with tissue- or developmental- specific expression of the gene, but these are possibilities worth investigating.

Comparison of the rabbit *Nat2* cDNA sequences, previously reported by Blum *et al.* (1990b) and Sasaki *et al.* (1991), indicates differential utilisation of at least three NCEs. Alternative splicing of the 5′ UTR has also been observed for the hamster *Nat1* and *Nat2* genes (Land *et al.*, 1994; Nagata *et al.*, 1994), as well as for a bovine *Nat* gene (section 5.2.6). Ebisawa *et al.* (1995) have described four alternatively spliced upstream NCEs for rat *Nat1* and another four for rat *Nat2* gene. Interestingly, three out of four of these NCEs are shared between rat *Nat1* and *Nat2* transcripts, indicating that the two *Nat* genes are nested and probably regulated by the same promoter. The coding regions of other rodent *Nat1* and *Nat2* genes are relatively close to each other,

separated by 9.4kb in the mouse (section 4.2.2.1; Martell *et al.*, 1991; Fakis *et al.*, 2000) and 5kb in the hamster (Nagata *et al.*, 1994). It is, therefore, possible that the *Nat1* and *Nat2* transcripts of rodent species may be subject to alternative splicing of shared upstream NCEs. Of particular interest is also the fact that in all genes for NAT, analysed in this Chapter, the acceptor splice site adjacent to the coding exon has a strictly conserved position (section 5.2.6), while the position of the donor splice site varies considerably. A comparison of the currently available information about the exon-intron organisation of different genes for mammalian NATs is attempted in table 5.3.1.

Several lines of evidence suggest that splicing of a transcript can be of great importance for eukaryotic gene expression. Introduction of either orthologous or heterologous introns in transiently transfected gene constructs has been demonstrated to enhance accumulation of recombinant mature transcripts (Gruss and Khoury, 1980; Korb *et al.*, 1993; Bardonaro and Nordstrom, 1994; Rafiq *et al.*, 1997). Moreover, efficient expression of transgenes in fungi (Lugones *et al.*, 1999), plants (Tanaka *et al.*, 1990) and animals (Bruce *et al.*, 1991; Palmiter *et al.*, 1991; Clark *et al.*, 1993) largely depends on the presence of introns. It has been established that splicing is essential for transport of mature mRNA from the nucleus to the cytoplasm (Palmiter *et al.*, 1991; Haselbeck and Greer, 1993; Elliott *et al.*, 1994; Fu *et al.*, 1997). Transgene expression from intronless cDNA sequences leads to retention of large amounts of transcript in the nucleus (Fu *et al.*, 1997), while introduction of a heterologous intron between the promoter and the intronless coding region enhances accumulation of the spliced message in the cytoplasm (Palmiter *et al.*, 1991). It is, therefore, possible that genes with intronless coding regions, like those for eukaryotic NATs, have a requirement of upstream NCEs to achieve sufficient nucleo-cytoplasmic transport of their transcript, via splicing. The transcripts of naturally intronless eukaryotic genes, such as the histone genes, have special *cis*-elements to facilitate their migration to the cytoplasm (Huang and Carmichael, 1997; Huang *et al.*, 1999b). Many viral genes transcribed in infected eukaryotic cells also employ special mechanisms to promote nucleo-cytoplasmic transport of unspliced messages (Gruss *et al.*, 1981; Kiyokawa *et al.*, 1997; Lee and Johnson, 1998; Sandri-Goldin, 1998; Trubetskoy *et al.*, 1999).

1	2	3	4	5	6	7	8	9	10
Mammal	**Gene**	**No. of NCEs**	**Position of NCEs**	**Variation in position of splice sites**	**Position of acceptor splice site proximal to the coding region**	**Transcription not involving NCEs**	**Position of polyA signals**	**Alternative 5′ UTRs**	**Alternative polyadenylation**
Human[1]	*NAT1*	3	-11.9kb, -10.8kb, -2.6kb	No	-6bp	Yes	1085bp, 1203bp	Yes	Yes
	NAT2	1	-8.7kb	Possibly (donor)	-6bp	No	1071bp, 1193bp	Possibly	Yes
Mouse[2]	*Nat2*	1	-6.3kb	Yes (donor)	-6bp	Yes[9]	1314bp	Yes	No
Rat[3]	*Nat1*	5[7]	?	?	-6bp	?	1539bp, 1606bp	Yes	Yes
	Nat2	4[7]	?	?	-6bp	?	?	Yes	?
Hamster[4]	*Nat1*	1	?	?	-6bp	?	184bp[10]	No	Yes[11]
	Nat2	multiple[8]	?	?	?	?	?	Yes	Yes[11]
Rabbit[5]	*Nat2*	3	-3.3kb, -1.9kb, -1.7kb	?	-6bp	?	95bp, 521bp	Yes	Yes
Cow[6]	*Nat*	3	?	?	-5bp	?	?	Yes	?

[1]Combined data from figures 5.2.15 and 5.2.16, as well as from Ohsako and Deguchi (1990); [2]Data from figure 5.2.14; [3]Combined data from figure 5.2.17, as well as from Ebisawa et al. (1995); [4]Combined data from Abu-Zeid et al. (1991), Land et al. (1994) and Nagata et al. (1994); [5]Combined data from Blum et al. (1989a and 1990b), as well as Sasaki et al. (1991); [6]Data from figure 5.2.18. [7]Three of the NCEs are shared between the Nat1 and Nat2 genes; [8]According to Land et al. (1994), although the exact number of NCEs is not specified. [9]According to Estrada-Rodgers et al. (1998a); [10]According to Nagata et al. (1994); [11]According to Land et al. (1994), although the position of the alternative polyA signals is not specified.

Table 5.3.1: Structure comparison of different genes for mammalian NAT enzymes. Combined information from current and previous experimental work, as well as from the analysis of ESTs deposited in databases (section 5.2.6), was used to compare the genomic organisation of nine mammalian genes for NAT (columns 1 and 2). Genes for which only genomic sequence is available (e.g. mouse *Nat1* and *Nat3* or rabbit *Nat1*) were not included in the table. The positions of the NCEs (column 4), splice sites (column 6) and polyadenylation signals (column 8) are indicated relative to the start of the coding region. The question marks indicate lack of sufficient data. The NCEs of mouse *Nat2* and possibly human *NAT2* genes have more than one differentially used donor splice sites (column 5). Transcription does not always involve a NCE, but may be initiated from sites adjacent to the coding region (column 7). Information on the alternative splicing of the 5′ UTR and the alternative use of more than one polyadenylation signal is summarised in columns 9 and 10.

Differential utilisation of upstream NCEs is often linked to tissue- or developmental-specific expression of genes from more than one promoter (Jitrapakdee *et al.*, 1997; Hu *et al.*, 1998; 1999; Madiai *et al.*, 1999; Geiger *et al.*, 2000). Transcription initiation from NCE1, as well as upstream of the coding region of human *NAT1* (figure 5.2.16) is likely to involve at least two active promoters. This may also be the case for murine *Nat2*, as evidence supports differential initiation of transcription from either upstream of the NCE (section 5.2.4) or upstream of the coding region (Estrada-Rodgers *et al.*, 1998a) of the gene. In contrast, only one promoter, located upstream of the single NCE (figure 5.2.15), is likely to drive expression of the human *NAT2* gene.

Heterogeneity within the 5′ UTR of a gene, via alternative splicing, may also modulate translation efficiency (Smith *et al.*, 1989). Binding of the mRNA to the ribosome is the rate-limiting step in translation initiation. Secondary structure formation at the 5′ end of a mRNA or specific binding of activating or inhibitory factors can adjust translation rates to the specific requirements of a cell type or developmental stage (Sonenberg, 1994). This could be an important role of the differentially utilised NCEs, associated with the ubiquitously expressed human *NAT1* and murine *Nat2* genes.

Splicing of long upstream introns is a common feature of retrosposons. Insertion of a retrogene to the genome could be followed by the generation of splice sites, via mutation, to bridge the distance between the coding exon and the closest available promoter (Sosinsky *et al.*, 2000). During this process, generation of multiple splice sites of different strengths may lead to the non-specific production of transcripts with variable 5′ UTRs, without any functional consequences (Chabot, 1996; Elliott, 2000; Graveley, 2001). This could explain, for instance, why the donor splice site of the mouse *Nat2* NCE may vary in different tissues (sections 5.2.4 and 5.2.6). However, NCEs have also been shown to play key roles in the regulation of splicing, as they may contain ESEs involved in exon definition and splice site selection (Chabot, 1996; Cooper and Mattox, 1997). The NCEs of several genes have been reported to contain polymorphisms, often with a phenotypic effect. For example, mutations within the NCE1 of human nuclear respiratory factor 1 (*NRF-1*) gene have been shown to reduce translational efficiency in transfected cells and *in vitro*

translation systems (Huo and Scarpulla, 1999). Mutations in the NCEs of the tuberous sclerosis complex 1 (*TSC1*) and connexin-26 (*CX26*) genes in humans may also have pathogenic significance (Dabora *et al.*, 1998; Denoyelle *et al.*, 1999). Screening of different mouse strains for putative polymorphisms in the *Nat2* NCE might, therefore, be worthwhile. Furthermore, genotype-based prediction of the acetylator status of human individuals should also take into account the possible presence of polymorphisms in the NCEs of the human *NAT* genes.

The differential utilisation of tandem polyA signals, present in the 3′ UTR of many eukaryotic genes, is a common phenomenon. The selection of a polyadenylation site can be a regulated process or a random event, depending on the inherent strength of the alternative polyA signals (section 1.5.2.3). The experimental data and EST analysis described in this Chapter (sections 5.2.5 and 5.2.6) suggest operation of only one polyA signal for the murine *Nat2* gene. Other elements in the surrounding region, which could also serve as polyA signals, appear to have no function. On the other hand, comparison of published human *NAT* cDNA sequences (Ohsako and Deguchi, 1990) with a series of ESTs (figures 5.2.15 and 5.2.16) revealed that the *NAT1* and *NAT2* genes must have at least two active polyA signals. Analysis of ESTs and previously published work further suggested alternative polyadenylation of the *Nat* transcripts in tissues of the rat, hamster and rabbit (table 5.3.1). It might be of interest to investigate whether alternative polyadenylation of these transcripts is the result of a regulated or random selection of the available polyA signals.

Although the molecular genetics of mammalian NATs has been extensively studied in a variety of species, many important questions remain unanswered. Understanding of the structure of the genes for NAT will shed light on their evolutionary origins, their role in endogenous metabolism and the mechanisms regulating their tissue- and developmental-specific expression. Search for elements regulating murine *Nat2* expression will be the topic of the following Chapter.

CHAPTER 6

Control of transcription of the mouse *Nat2* gene

6.1 Introduction

The study of elements regulating expression of the genes for mammalian NAT isoenzymes has been hindered by the lack of sufficient knowledge about the structure of the genes and the sequence of the surrounding genomic DNA. Although the mouse *Nat2* gene appears to be ubiquitously expressed (section 5.2.1; Chung *et al.*, 1993; Stanley *et al.*, 1997), its rate of transcription may be modulated to meet the requirements of a specific cell type or developmental stage. Differences in the amount of murine NAT2 protein and enzymatic activity have been observed between tissues of either embryonic (Stanley *et al.*, 1998) or adult origin (Chung *et al.*, 1993; Stanley *et al.*, 1997; Ware and Svensson, 1996; Payton *et al.*, 1999b), and could be the result of transcriptional regulation of the *Nat2* gene in a tissue-specific manner.

There is also a well documented difference in renal NAT2 enzymatic activity between male and female mice (Glowinski and Weber, 1982b; Smolen *et al.*, 1993; Estrada *et al.*, 2000), reflecting a difference in the amount of synthesised NAT2 protein (Estrada *et al.*, 2000). It has been demonstrated that this gender-related difference in renal NAT2 activity is androgen-dependent, as exposure of female animals to testosterone increased activity to male levels, while exposure of male animals to oestrogens decreased activity to female levels. NAT2 activity was low in the kidney of castrated males, but did not change in the kidney of ovariectomised females. A hormone response element, located 576-561bp upstream of the *Nat2* coding region (Estrada-Rodgers *et al.*, 1998a) could be responsible for the observed differences in renal *Nat2* expression between male and female mice. This element positively regulated a heterologous promoter (HSV *tk1*) in reporter gene assays, following exposure of transiently transfected CV1 cells to glucocorticoid hormones,

including androgens. The positive regulation of HSV *tk1* promoter by the hormone response element was entirely dependent on the presence of glucocorticoid or androgen receptors in the transfected cells, and would, therefore, be expected to take place in a tissue-specific manner. Glucocorticoids have also been demonstrated to increase hepatic NAT activity in rats (Zaher and Svensson, 1994) and rabbits (Reeves *et al.*, 1988; 1989), however, the biological significance of this regulation remains unclear.

Using RNase protection assays and primer extension analysis, Estrada-Rodgers *et al.* (1998a) have also identified six transcription initiation sites for murine *Nat2*, all of which are located 112-151bp upstream of the coding region of the gene. However, the identification of a NCE (section 5.2.4) points to the presence of at least one alternative transcription initiation site, located more than 6.3kb upstream of the coding region. This finding has important implications in the search for regulatory elements, such as promoters and enhancers or repressors of gene expression.

The major aim of the work presented in this Chapter has been to look for core promoter elements upstream of the NCE, as well as within the 6.1kb intron of the *Nat2* gene, by reporter gene assay. Moreover, the transcription initiation site marking the beginning of the *Nat2* NCE was accurately mapped by RNase protection assay. Preliminary work was also carried out towards identification of candidate regions harbouring *cis*-elements that might control activity of the core *Nat2* promoter. The regulatory potential of these regions, as well as their capacity to bind transcription factors, was assessed by reporter gene assay and electrophoretic mobility shift assay (EMSA), respectively. Elucidation of the mechanisms that control transcription of the mouse *Nat2* gene is central to an understanding of its role in endogenous metabolism and development.

6.2 Results

6.2.1 Characterisation of the BNL.CL2 mouse embryonic liver cell line

Mapping of the transcription initiation site of the murine *Nat2* gene by RNase protection assay and subsequent characterisation of transcription regulatory elements by reporter gene assay, require identification of a suitable cell line expressing the *Nat2* gene and its upstream NCE (section 5.2.4). The BNL.CL2 cell line, developed from normal hepatic cells of the Balb/c mouse embryo (Patek *et al.*, 1978; Boussif *et al.*, 1996), was examined for *Nat2* expression by enzymatic activity assay, Western blot analysis and RT-PCR.

NAT2 enzymatic activity was measured in BNL.CL2 cell lysates, prepared as described in section 2.3.2, using 0.2mM pABA as substrate (section 2.3.5.2). The calculated specific activity was 1.2 nmoles of N-acetylated pABA/min/mg of total protein. The presence of NAT2 protein in BNL.CL2 cells was further confirmed by Western blot analysis of cell lysates with antiserum 184 (section 2.3.4) (figure 6.2.1).

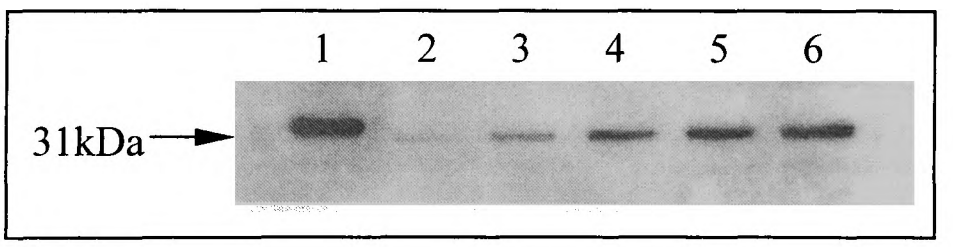


Figure 6.2.1: Detection of murine NAT2 protein in BNL.CL2 cell lysates by Western blotting. Lane 1 is 10μg of total protein from CHO cells expressing recombinant human NAT1, used as positive control (provided by N. Johnson, Department of Pharmacology, Oxford). Lanes 2-6 are 10, 25, 50, 75 and 100μg of total protein contained in BNL.CL2 cell lysates. The 31kDa murine NAT2 protein was detected in BNL.CL2 cell lysates by antiserum 184. The human NAT1 protein, contained in the positive control (lane 1), has a C-terminal amino acid sequence identical to that of murine NAT2, therefore, it can also bind antiserum 184 (Stanley *et al.*, 1996).

Total RNA and mRNA was extracted from BNL.CL2 cells (section 2.2.1.4) and analysed for expression of the three *Nat* genes by RT-PCR (section 5.2.1). Only *Nat2* transcript was detected (figure 6.2.2a-c), and this was further confirmed to contain the upstream NCE described in section 5.2.4 (figure 6.2.2d). The BNL.CL2 cell line is, therefore, suitable for studying the role of the *Nat2* upstream NCE, and the transcriptional control of the *Nat2* gene, independently of the *Nat1* and *Nat3* genes.

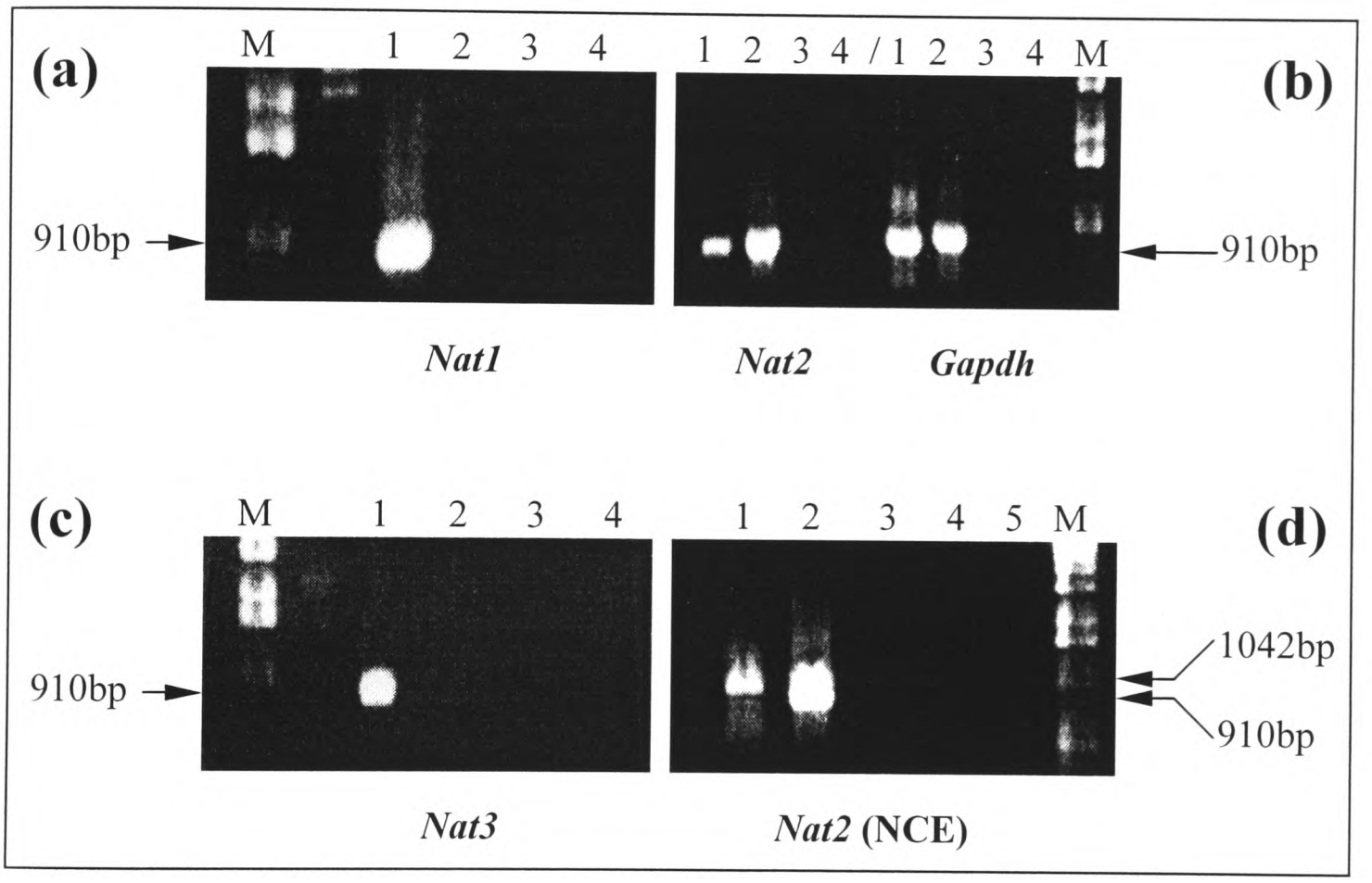


Figure 6.2.2: Investigation of *Nat* gene expression in BNL.CL2 cells by RT-PCR. a-c) Amplification from genomic DNA (lane 1) and cDNA (lane 2), as well as from the product of a mock reverse transcription without reverse transcriptase in the reaction mixture (lane 3). Lane 4 is the PCR negative control. a) RT-PCR with primers Mus12 and Mus13, showing lack of *Nat1* expression in BNL.CL2 cells. b) RT-PCR with primers mNAT2-1 and mNAT2-910 (left), demonstrating expression of *Nat2* in BNL.CL2 cells. Amplification of *Gapdh* with primers GAPDH-S and GAPDH-AS (right) was used as positive control. c) RT-PCR with primers Mus12 and Mus15, showing lack of *Nat3* expression in BNL.CL2 cells. d) Amplification from BNL.CL2 cDNA with primer pairs NCE-F/mNAT2-910 (lane 1) and mNAT2-1/ mNAT2-910 (lane 2). Lane 3 shows amplification from the product of a mock reverse transcription with primers mNAT2-1 and mNAT2-910, while lanes 4 and 5 are the PCR negative control for each set of primers. Successful amplification in lane 1 confirmed transcription of the upstream NCE of the *Nat2* gene (section 5.2.4) in BNL.CL2 cells. Lanes M are 1μg of 1kb DNA ladder.

6.2.2 Generation of a *Nat2*-specific ribo-probe for use in RNase protection assays

A ribo-probe, covering the region immediately upstream of the mouse *Nat2* NCE (section 5.2.4), was generated for mapping of the transcription initiation site of the *Nat2* gene by RNase protection assay. A 247bp fragment spanning region 3598 (within the NCE) to 3351 (202bp upstream of the NCE) of 129/Ola clone A (Appendix 2) was amplified with primers Ribo-1F and Ribo-1R (table 6.2.1 and figure 6.2.3) and cloned into the pGEM-T Easy vector (Appendix 1.1a), which contains T7 and SP6 promoters for *in vitro* transcription by RNA polymerase. White colonies of JM109 *E. coli* cells (section 2.2.3.10) were screened for the presence of the correct size insert by restriction digestion of plasmid DNA with *Eco*RI (figure 6.2.4a), which cuts the pGEM-T Easy vector on either side of the insert (Appendix 1.1a). The orientation of the insert was determined by digestion of plasmid DNA from positive clones with *Sac*I (figure 6.2.4b), which cuts the insert at position 77 (position 3428 of 129/Ola clone A; Appendix 2) and the vector at position 109 (Appendix 1.1a). Incorporation of the insert in the T7→SP6 orientation provided fragments 226 and 3050bp in size, while the opposite (SP6→T7) orientation provided fragments 133 and 3143bp in size (figure 6.2.4b).

Table 6.2.1: Primers used for amplification of the *Nat2*-specific ribo-probe. The sites of primer annealing are indicated relative to the beginning of 129/Ola clone A (Appendix 2).

Primer name	Orientation	Sequence (5′→3′)	Tm (°C)
Ribo-1F	forward	3351-TACATTTCCCAGAGATCC -3368	52
Ribo-1R	reverse	3598-TCAGGACTGTGTTGTGCT -3581	54

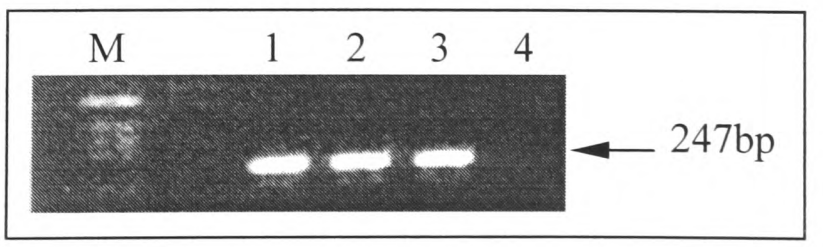


Figure 6.2.3: PCR amplification of the *Nat2*-specific ribo-probe. Plasmid DNA (50ng) from 129/Ola clone A was used for amplification of a 247bp fragment (lanes 1-3) with primers Ribo-1F and Ribo-1R (table 6.2.1). The amplified fragment was then cloned into the pGEM-T Easy vector. Lane 4 is the PCR negative control. Lane M is 1μg of 1kb DNA ladder. Amplification products were analysed on a 2.5% (w/v) agarose gel.

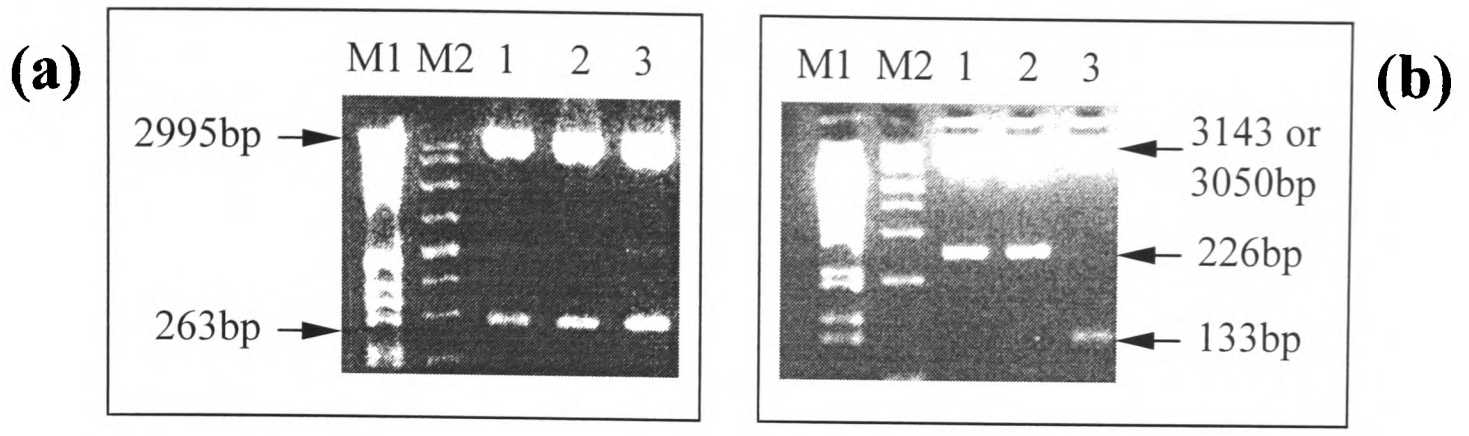


Figure 6.2.4: Checking for incorporation of the *Nat2*-specific ribo-probe insert into the pGEM-T Easy vector. a) Digestion with *Eco*RI of plasmid clones expected to carry the 247bp ribo-probe insert. Products were analysed on a 2.5% (w/v) agarose gel. All three clones (lanes 1, 2 and 3) carry the expected size insert. b) Digestion of the positive plasmid clones with *Sac*I, to determine the orientation of the ribo-probe insert into the pGEM-T Easy vector. Products were analysed on a 2% (w/v) Metaphor/2% (w/v) agarose gel. The 226bp band, indicating incorporation of the insert in the T7→SP6 orientation, is present in lanes 1 and 2 (clones 1 and 2, respectively). The 133bp band, indicating opposite orientation of the insert, is present in lane 3 (clone 3). Lanes M1 and M2 are 1kb and BioMarker EXT size markers, respectively.

For successful hybridisation to occur, the generated ribo-probe needs to have a sequence complementary to the target *Nat2* mRNA. According to the scheme depicted in figure 6.2.5, only positive clone 3 (shown in figure 6.2.4) has the SP6→T7 insert orientation required for *in vitro* transcription of a complementary ribo-probe by T7-RNA polymerase. Transcription by SP6-RNA polymerase would require opposite orientation of the insert. Prior to *in vitro* transcription, plasmid DNA from clone 3 was linearised with *Nde*I, which cuts the pGEM-T Easy vector at position 97, i.e. just before the SP6 promoter (Appendix 1.1a). *Nde*I is an appropriate restriction enzyme, because it creates 5′ overhangs. Templates with 3′ overhangs have been demonstrated to drive synthesis of long extraneous transcripts hybridising to the vector (Schenborn and Mierendorf, 1985) and are not suitable for *in vitro* transcription.

A clone in pUC18 vector, containing a ribo-probe hybridising to the 3′ end of the adenoviral VAI transcript (Akusjärvi *et al.*, 1980), was provided by Dr. Kathryn Plant (Sir William Dunn School of Pathology, Oxford) and used as an internal control for assessing transfection efficiencies.

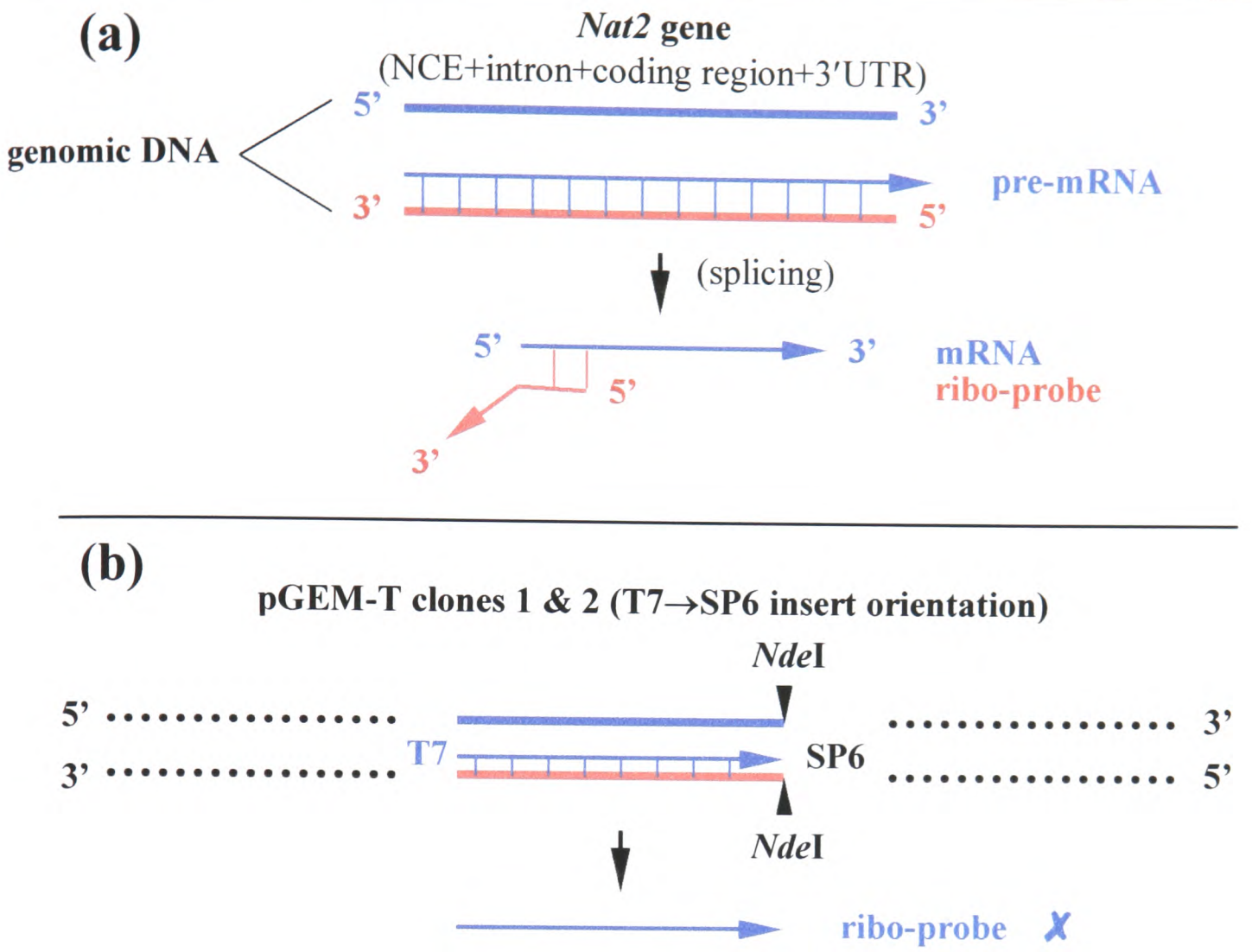


Figure 6.2.5: Insert orientation in pGEM-T Easy vector, allowing *in vitro* transcription by T7-RNA polymerase of the *Nat2*-specific ribo-probe. a) The "red" strand of the *Nat2* gene is transcribed to provide a pre-mRNA molecule (blue arrow), containing the NCE, the 6.1kb intron, the coding region and the 3′ UTR of the gene. Following splicing of the intron, the produced mRNA can hybridise only to a "red" ribo-probe molecule. b) The "blue" strand of the insert in pGEM-T Easy vector (black dotted lines) is identical to the "blue" DNA strand in (a). The "red" strand of the insert is identical to the transcribed "red" DNA strand in (a). The position of the *Nde*I site used for linearisation of the plasmid is also indicated. Incorporation of the insert in the T7→SP6 orientation (clones 1 and 2 in figure 6.2.4) leads to the production of a "blue" ribo-probe, following *in vitro* transcription by T7-RNA polymerase, which does not hybridise to the *Nat2* mRNA. Incorporation of the insert in the SP6→T7 orientation (clone 3 in figure 6.2.4) leads to the production of a "red" ribo-probe, which hybridises to the *Nat2* mRNA. Therefore, the latter is the insert orientation of choice for *in vitro* transcription by T7-RNA polymerase.

6.2.3 Mapping of the transcription initiation site of the murine *Nat2* gene by RNase protection assay

Mouse embryonic liver BNL.CL2 cells were co-transfected with 15μg of purified 129/Ola clone A DNA and 5μg of pVA clone, containing the entire VAI gene in pUC18 vector (section 6.2.2). Transfections (section 2.2.4.2) were carried out in triplicate. About 16h following transfection, one plate was subjected to a glycerol shock (section 2.2.4.2), followed by change of the culture medium. The second plate was not subjected to a glycerol shock, but had its medium changed after 16h. The medium of the third plate was changed about 40h following transfection, without a glycerol shock. Triplicate plates of mock-transfected cells (no DNA in the transfection mixture) were treated in exactly the same way. Cells were harvested 72h after transfection and immediately used for extraction of total RNA (section 2.2.1.4).

The RNase protection assays (section 2.2.6) were carried out in the laboratory of Professor N. Proudfoot (Sir William Dunn School of Pathology, Oxford), with the guidance of Dr. Kathryn Plant. *In vitro* transcription and radioactive labelling of the *Nat2*-specific and the VA-specific ribo-probe (section 6.2.2) was carried out as described in section 2.2.6.1. The RNA polymerase III-transcribed adenoviral VAI gene is not present in eukaryotic genomes (Akusjärvi *et al.*, 1980), therefore, the amount of transcript detected by the VA-specific ribo-probe in RNA peparations from transfected BNL.CL2 cells was exclusively the product of the pVA vector. The VA-specific ribo-probe was used to determine optimal transfection conditions (figure 6.2.6). Maximum transfection efficiency was achieved with cells incubated in the transfection medium for a prolonged period of time (40h), without the need for glycerol shock. Mock-transfected cells did not produce any VAI RNA, neither did untransfected cells, as expected. Digestion of free (unprotected) ribo-probe was complete, while contamination of the RNA preparations with genomic DNA was minimal. Degradation of the undigested ribo-probes due to radiolysis or introduction of exogenous RNases was also limited (figure 6.2.6).

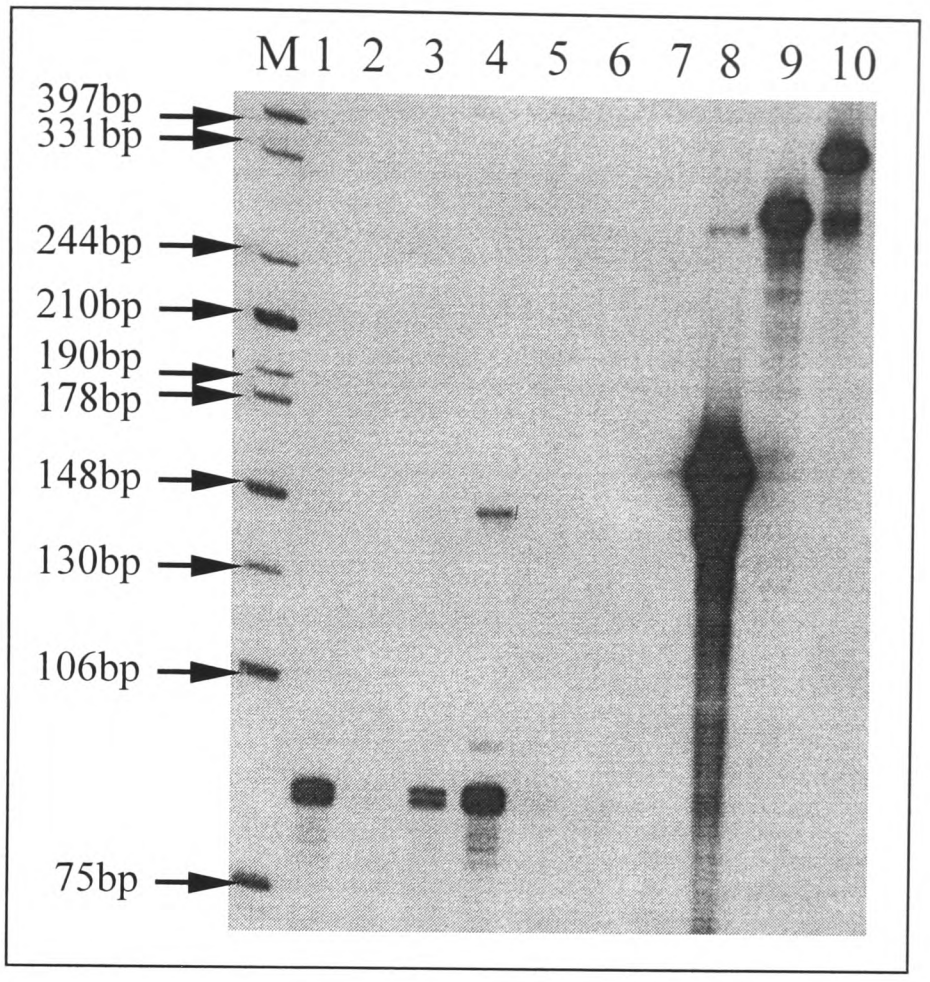


Figure 6.2.6: Optimisation of transfection and RNase protection assay conditions, using the VA-specific ribo-probe. BNL.CL2 cells were co-transfected with 15μg of 129/Ola clone A plasmid DNA and 5μg of pVA vector (section 2.2.4.2). Total RNA was extracted from cells (section 2.2.1.4) and incubated with [α^{32}P]-rUTP-labelled VA-specific ribo-probe (section 6.2.2), followed by digestion with a mixture of RNaseA and RNaseT_1 ribonucleases (section 2.2.6). Products were analysed on a 6% (w/v) denaturing polyacrylamide gel (section 2.2.2.2). RNA sequences hybridising to the ribo-probe were protected from RNase digestion and are visible on the gel. Lanes 1-6 show the protected RNA product from BNL.CL2 cells, which were either subjected to a glycerol shock 16h after tansfection (lane 1), or had their medium replaced 16h (lane 3) or 40h (lane 4) after transfection, without a glycerol shock. Mock-transfected cells were either subjected to a glycerol shock (lane 2) or to a 40h incubation in the transfection medium (lane 5), as above. Lane 6 is the product of non-transfected cells, used as negative control. Lanes 7 and 8 are RNase-digested and intact VA-specific ribo-probe, respectively, while lane 10 is the intact *Nat2*-specific ribo-probe. Lane 9 is a miscellaneous ribo-probe, not used in this study. DNA size markers in lane M (generated by *Hinf*I+*Eco*RI-digestion of a pUC clone containing the human β-globin gene) were end-labelled with [α^{32}P]-dATP, following incubation with DNA polymerase I large (Klenow) fragment (provided by Dr. K. Plant).

RNase protection using the same total RNA preparations and the *Nat2*-specific ribo-probe (section 6.2.2), provided a major product co-migrating on the gel with the 210bp DNA size marker (figure 6.2.7). The intact ribo-probe is 358bp in size (figure 6.2.6), due to transcribed flanking sequences of the vector (linkers). The linkers and at

least 37bp at the 5′ end of the ribo-probe were RNase-digested, while the rest of the molecule hybridised to *Nat2* mRNA and escaped digestion. From the size of the major protected product (figure 6.2.7), the transcription initiation site of the *Nat2* gene can be mapped to position 3388 of 129/Ola clone A (Appendix 2). Other less intense bands on the autoradiogram, approximately 175 and 135bp in size (figure 6.2.7), may indicate additional transcription initiation sites at around positions 3423 and 3463 of 129/Ola clone A, respectively. Initiation of transcription from this genomic region is further supported by computational NIX-analysis (section 4.2.2.2), which predicts a transcription initiation site at position 3420 of 129/Ola clone A.

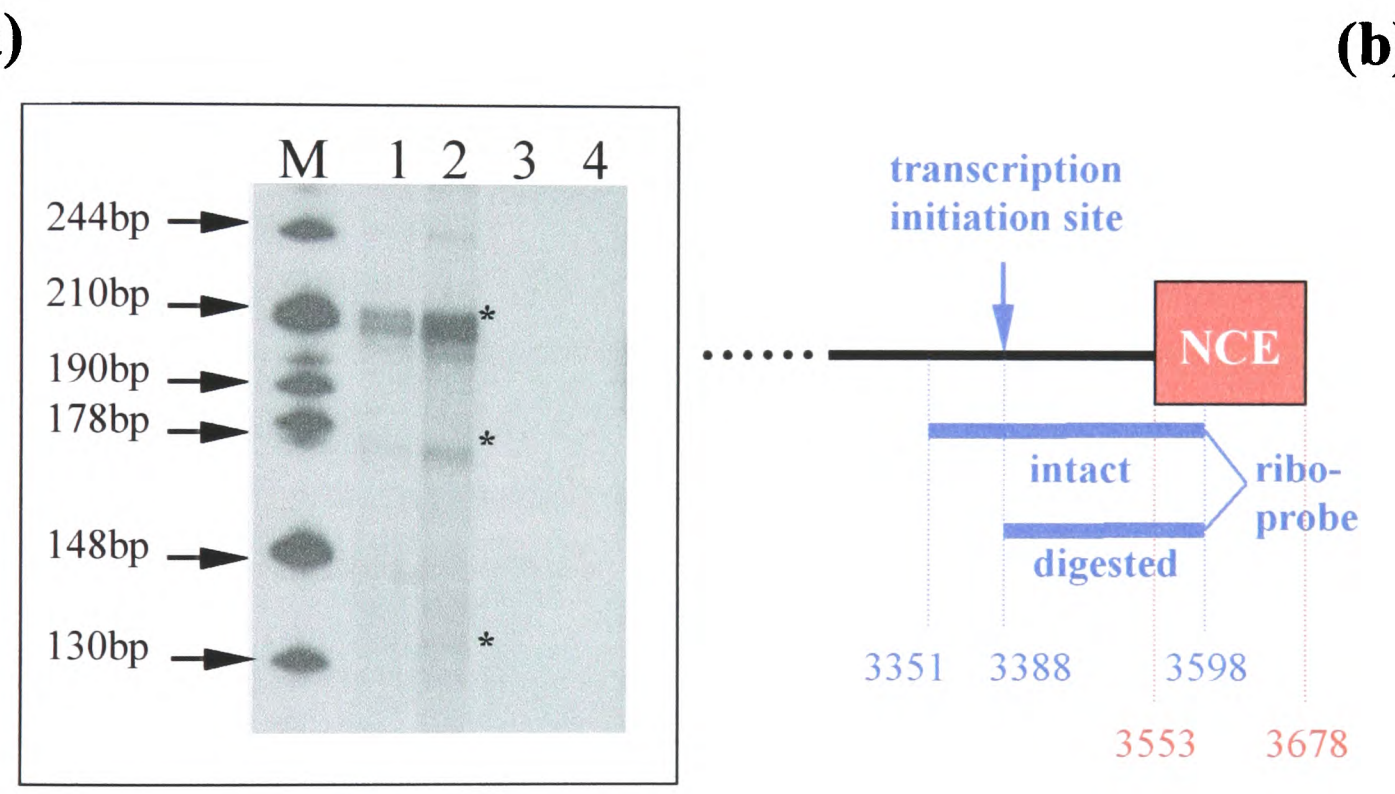


Figure 6.2.7: Mapping of the transcription initiation site of the *Nat2* gene by RNase protection assay. a) RNase protection assays were performed as described for figure 6.2.6. Lanes 1 and 2 show the protected RNA products (marked with an asterisk) from BNL.CL2 cells, which were either subjected to a glycerol shock 16h after tansfection (lane 1) or had their culture medium replaced after 40h (lane 2). Lane 3 is the product of non-transfected cells, while lane 4 is RNase-digested free (unprotected) ribo-probe. Markers in lane M are the same as in figure 6.2.6. b) Diagram showing the position and relative size of the intact (247bp) and RNase-digested (210bp) ribo-probe, spanning part of the *Nat2* NCE (section 5.2.4) and the 202bp upstream region. The numbering is relative to the beginning of 129/Ola clone A (Appendix 2). The approximate size (RNA products may migrate slightly differently than the DNA markers on a polyacrylamide gel) of the digested ribo-probe indicates the major transcription initiation site of the *Nat2* gene.

6.2.4 Subcloning of 129/Ola clone A into pGL3 luciferase reporter vectors

Identification of elements regulating expression of mouse *Nat2* by reporter gene assay requires subcloning of the surrounding genomic region into reporter vectors. The sequencing information generated for the 14.3kb 129/Ola clone A in section 4.2.2.1 was exploited to design primers (table 6.2.2) suitable for the amplification of 21 contiguous segments covering the entire length of the clone, apart from the *Nat1* and *Nat2* coding region and part of the *Nat2* NCE (figure 6.2.8). By incorporating 2-3 mismatches, the primer sequences were modified to contain *Mlu*I or *Xho*I restriction sites (table 6.2.2), allowing cloning of the amplified segments into the reporter vectors. The exact position and size of the segment amplified from 129/Ola clone A with each set of primers is provided in table 6.2.3. Amplification products were digested with *Mlu*I and *Xho*I (section 2.2.3.6) and ligated (section 2.2.3.9) into similarly digested and dephosphorylated (section 2.2.3.8) firefly luciferase (pGL3) reporter vectors (Appendix 1.2). All segments were ligated to pGL3-Basic and pGL3-Promoter vectors, apart from segments 9, 11, 13, 15 and 17, which were ligated only to the pGL3-Basic vector (table 6.2.3). Fragment 7, predicted by NIX analysis to contain a putative promoter (section 4.2.2.2), was ligated to pGL3-Basic and pGL3-Enhancer reporter vectors (table 6.2.3).

Transformed JM109 High Efficiency Competent Cells (*E. coli*) (section 2.2.3.10) were grown overnight on LB-agar plates, containing 100μg/ml ampicillin. Five colonies from each plate were split into two halves, one of which was used for PCR screening with the insert-specific primers described in tables 6.2.2 and 6.2.3. The second half of positive-testing colonies was then used to inoculate 10ml of LB medium containing 100μg/ml ampicillin (section 2.2.3.10). Plasmid DNA was extracted from these cultures (section 2.2.1.2) and each clone was examined for the presence of the expected size insert by restriction digestion with vector-specific enzymes (table 6.2.3 and figure 6.2.9). The identity and orientation of each insert was further confirmed by digestion of plasmid DNA with enzyme combinations cutting both the insert and the reporter vector (table 6.2.3 and figure 6.2.9).

Table 6.2.2: Primers used for PCR amplification and sequencing of DNA segments cloned into the pGL3 luciferase reporter vectors. The nucleotide sites of primer annealing are indicated relative to the beginning of 129/Ola clone A (Appendix 2). Primers marked with a (F) are forward, while those marked with a (R) are reverse. All primers were designed to contain restriction sites suitable for cloning into the pGL3 vectors. *Mlu*I sites are highlighted in yellow, *Xho*I sites in green, *Nde*I sites in purple, *Not*I sites in blue and *Xba*I sites in red.

Primer name	Sequence (5′→3′)	Tm (°C)
REP1(F)	418-GCCCAAACATGGTGAACGCGTTTTTACTATTTAGG-452	76.3
REP2(R)	988-CGGGTGCACTGGCTCGAGTTCTGTTTTGAGGACTG-954	83.4
REP3(F)	938-GATTCAGGTAACTTACGCGTCCTCAAAACAG-968	72.2
REP4(R)	1464-ACTGTTTTACATCTCGAGTCAGATCCACAAG-1434	70.2
REP5(F)	1424-ACTGGGTAAACGCGTGGATCTGATTCAAG-1452	74.6
REP6(R)	1919-TGGGAATCCTGATAGCTCGAGACTCTCCGATG-1888	78.2
REP7(F)	1850-TCTCCATAGCTACGCGTACTGCTGGACAGG-1879	76.3
REP8(R)	2316-CTATGCAGTGCAAACTCGAGGCCTGGCATTGG-2285	81.9
REP9(F)	2283-TTCCAATGCCACGCGTCTAGTTTGCACTG-2311	78.5
REP10(R)	2821-TAGTCGGGTGTGGCTCGAGTGAGGAGAAGTGAG-2789	79.8
REP11(F)	2720-GAGGTATTGCTGTGCACGCGTCACAGTG-2747	77.8
REP12(R)	3286-TGATATTTTTGCATTTCTCGAGATAGGTTTG-3256	69.3
REP13(F)	2876-CAGTGTGTGCCAACGCGTCTTTAAATTATAG-2906	71.8
REP15(F)	3113-TAGTGGAAGGAAACGCGTACATTAGGAATG-3142	71.6
REP17(F)	3305-ACGGACTAGAGAACGCGTTCACCAGGGTTAAG-3336	76.7
PROM-1(F)	3255-GCAAACCTATCACGCGTAATGCAAAAATATC-3285	65.0
EXON-1(R)	3620-AAAAGAACTTGGGCCTCGAGAGTCAGGACTG-3590	70.0
RA-1(F)	3663-ACTTTGCAGTACACGCGTGATTCTGACAA-3691	74.0
RA-1(R)	4776-ACAGTAGTCAGACTCGAGGGAGGATTCACC-4747	73.1
REP20(R)	4142-CCAAAGTACAGACTCGAGTTCATGATATCAG-4112	68.8
RA-2(F)	4733-TATCGGCCCCTAACGCGTGAATCCTCCTTC-4761	79.4
RA-2(R)	5772-TGTAAAATACACCTCGAGTCACACGAAACC-5743	71.3
REP22(R)	5226-AGTTTGGTCAGCCTCGAGTTCATTCCATAG-5197	73.2
RA-3(F)	5728-GACAGATGTGTAACGCGTTTCGTGTGAGTC-5757	74.7
RA-3(R)	6880-GCATATTATGGCTCGAGAACAAGCCTATTC-6851	70.7
REP24(R)	6247-TTGATGGCACACCTCGAGTGATGCACTTGTTC-6216	80.0
RA-4(F)	6843-TATAAGAGGAATACGCGTGTTCAGGAGCCA-6872	73.8
RA-4(R)	8009-CCACTAAAGGATCTCGAGTAGTATTTGATTC-7979	66.3
REP26(R)	7319-CATTTTAAGAAAAACTCGAGTTCTACATACATG-7288	65.9
RA-5(F)	7830-CACTCCCCTGGTACGCGTTCATTCTCTATTTTC-7862	75.9
RA-5(R)	8939-ATAATGATGGGGCTCGAGGGAAAATTCTGTG-8909	75.8
REP28(R)	8443-AGGAGTAGTTATCTCGAGCGTACATTATGAG-8413	66.8
RA-6(F)	8838-TCTCTGTGAGCTACGCGTGAATCTATAGTATTC-8870	69.8
INTRON-1(R)	9850-TGGTTTCCTCGAGGCAAGAAAAACAATCC-9823	66.0
RA-7(F)	11177-TATATATCTAAAACGCGTTTTCAAAGTGG-11205	65.3
RA-7(R)	12183-AATGTCACAATTCTCGAGTTAATGACATTAC-12153	66.4
RA-8(F)	12132-ATCTTTTAGATAAACGCGTGGGTAATGTCATTAAG-12166	70.5
RA-8(R)	13280-TGTGGTTTCCTCTCGAGAATTAAAAGGGGAAC-13249	74.3
RA-9(F)	13194-GCAGATAAAAGAACGCGTGACAAGTCCCAATG-13224	77.0
RA-9(R)	14286-ATTCTTGGATTTCTCGAGGGGTTGAGAATTG-14256	74.0
DEL(F)	3409-CAACTTGAACAGCCCATATGAGCTCTTCC-3437	73.0
DEL(R)	3392-TAGTTTTTTTTTAAACCATATGAAGGATCTC-3362	64.7
SP1(F)	3388-AACTAAGCATAAGCGGCCGCGCAACTTGAACA-3419	80.3
SP1(R)	3419-TGTTCAAGTTGCGCGGCCGCTTATGCTTAGTT-3388	80.3
TATA(F)	3366-TCCTTGAGATGGTCTAGAAAAAAACTAAGC-3395	66.9
TATA(R)	3395-GCTTAGTTTTTTTCTAGACCATCTCAAGGA-3366	66.9

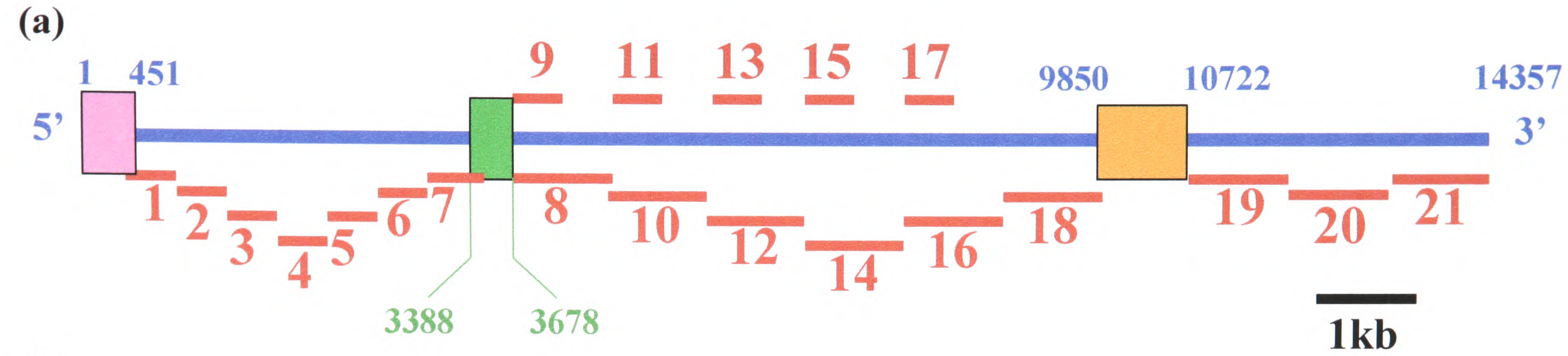


Figure 6.2.8: Amplification of contiguous segments, spanning 129/Ola clone A, for subcloning into the pGL3 luciferase reporter vectors. a) Relative position of 21 segments (red lines) amplified from 129/Ola clone A (blue line). The position of the *Nat1* (pink box) and the *Nat2* (orange box) coding region, as well as of the *Nat2* NCE (green box), is indicated relative to the beginning of 129/Ola clone A (Appendix 2). b) Amplification of the contiguous segments shown in (a), using a series of forward and reverse primers, containing *Mlu*I and *Xho*I restriction sites, respectively (table 6.2.2). Plasmid DNA (50ng) from 129/Ola clone A was used as template for PCR with *Pfu*-DNA polymerase. The numbering of the lanes is the same as that of the amplified segments shown in (a). Product 7 was analysed on a 2.5% (w/v) agarose gel. Lanes M1, M2 and M3 are 1µg of 1kb ladder, 5µl of BioMarker EXT ladder and 1µg of 1kb Plus ladder, respectively.

Table 6.2.3: Summary of the strategies used for amplification and cloning of 21 segments, spanning 129/Ola clone A, into the pGL3 luciferase reporter vectors. The position and size of each segment (numbered 1-21 as in figure 6.2.8) are provided relative to the beginning of 129/Ola clone A (Appendix 2). The type of vector used for cloning of each segment and the name of the corresponding construct are also indicated. The identity and orientation of the insert in each clone was confirmed by digestion of plasmid DNA with two sets of restriction enzymes. Enzymes marked with an asterisk cut the insert (Appendix 2). *Acc*I and *Bgl*II cut both the vector and the corresponding insert (Appendix 1.2 and Appendix 2), while *Mlu*I cuts at the insert-vector junction. Other enzymes cut only the vector (Appendix 1.2). The insert of clones B7 and B18 was also sequenced, using the amplification primers. It has not been possible to identify a P2 positive clone, even after repeated cloning and extensive colony screening.

Amplified fragment	Name of construct	Amplification primers	Amplified region of 129/Ola clone	Size of amplified segment (insert)	Type of pGL3 reporter vector	Checking for presence of insert		Checking for orientation of insert	
						Enzymes	Size of restriction fragments	Enzymes	Size of restriction fragments
1	**B1**	REP1(F) REP2(R)	418-988	570bp	pGL3-Basic	*Kpn*I *Hind*III	575bp+4770bp	*Nde*I* *Bgl*II	474bp+4870bp
	P1				pGL3-Promoter		767bp+4770bp		474bp+5062bp
2	**B2**	REP3(F) REP4(R)	938-1464	526bp	pGL3-Basic	*Kpn*I *Hind*III	531bp+4770bp	*Xma*I* *Bgl*II	471bp+4830bp
	P2				pGL3-Promoter		723bp+4770bp		471bp+5022bp
3	**B3**	REP5(F) REP6(R)	1424-1919	495bp	pGL3-Basic	*Kpn*I *Hind*III	502bp+4770bp	*Mlu*I* *Bgl*II	475bp+4797bp
	P3				pGL3-Promoter		694bp+4770bp		475bp+4989bp
4	**B4**	REP7(F) REP8(R)	1850-2316	466bp	pGL3-Basic	*Kpn*I *Hind*III	472bp+4770bp	*Pst*I* *Bgl*II	281bp+4961bp
	P4				pGL3-Promoter		664bp+4770bp		281bp+5153bp
5	**B5**	REP9(F) REP10(R)	2283-2821	538bp	pGL3-Basic	*Kpn*I *Hind*III	546bp+4770bp	*Mlu*I* *Bgl*II	519bp+4797bp
	P5				pGL3-Promoter		738bp+4770bp		519bp+4989bp
6	**B6**	REP11(F) REP12(R)	2720-3286	566bp	pGL3-Basic	*Kpn*I *Hind*III	566bp+4770bp	*Bst*EII* *Bgl*II	119bp+5217bp
	P6				pGL3-Promoter		758bp+4770bp		119bp+5409bp
7	**B7**	PROM-1(F) EXON-1(R)	3255-3620	365bp	pGL3-Basic	*Kpn*I *Hind*III	371bp+4770bp	*Eco*RV* *Hind*III	339bp+4802bp
	E7				pGL3-Enhancer		371bp+5016bp		339bp+5048bp
8	**B8**	RA-1(F) RA-1(R)	3663-4776	1113bp	pGL3-Basic	*Kpn*I *Nar*I	1188bp+4702bp	*Eco*RV* *Bgl*II	654bp+5236bp
	P8				pGL3-Promoter		1380bp+4702bp		654bp+5428bp
9	**B9**	RA-1(F) REP20(R)	3663-4142	479bp	pGL3-Basic	*Kpn*I *Hind*III	486bp+4770bp	*Eco*RV* *Kpn*I	449bp+4807bp
10	**B10**	RA-2(F) RA-2(R)	4733-5772	1039bp	pGL3-Basic	*Kpn*I *Hind*III	1046bp+4770bp	*Acc*I* *Kpn*I	990bp+2014bp+2812bp
	P10				pGL3-Promoter		1239bp+4770bp		990bp+2206bp+2812bp

Table 6.2.3 continued

11	**B11**	RA-2(F) REP22(R)	4733-5226	493bp	pGL3-Basic	*Kpn*I *Hind*III	500bp+4770bp	*Mlu*I* *Bgl*II	473bp+4797bp
12	**B12**	RA-3(F) RA-3(R)	5728-6880	1152bp	pGL3-Basic	*Kpn*I *Hind*III	1160bp+4770bp	*Bgl*II*	179bp+5751bp
	P12				pGL3-Promoter		1352bp+4770bp		179bp+5943bp
13	**B13**	RA-3(F) REP24(R)	5728-6247	519bp	pGL3-Basic	*Kpn*I *Hind*III	526bp+4770bp	*Xba*I* *Kpn*I	306bp+4990bp
14	**B14**	RA-4(F) RA-4(R)	6843-8009	1166bp	pGL3-Basic	*Kpn*I *Hind*III	1173bp+4770bp	*Mlu*I* *Bgl*II	1146bp+4797bp
	P14				pGL3-Promoter		1365bp+4770bp		1146bp+4989bp
15	**B15**	RA-4(F) REP26(R)	6843-7319	476bp	pGL3-Basic	*Kpn*I *Hind*III	481bp+4770bp	*Nsi*I* *Bgl*II	432bp+4819bp
16	**B16**	RA-5(F) RA-5(R)	7830-8939	1109bp	pGL3-Basic	*Kpn*I *Hind*III	1116bp+4770bp	*Bgl*II*	177bp+5709bp
	P16				pGL3-Promoter		1308bp+4770bp		177bp+5901bp
17	**B17**	RA-5(F) REP28(R)	7830-8443	613bp	pGL3-Basic	*Kpn*I *Hind*III	620bp+4770bp	*Mlu*I* *Bgl*II	593bp+4797bp
18	**B18**	RA-6(F) INTRON-1(R)	8838-9850	1012bp	pGL3-Basic	*Sac*I *Hind*III	1018bp+4776bp	*Pst*I* *Bgl*II	703bp+5091bp
	P18				pGL3-Promoter		1210bp+4776bp		703bp+5283bp
19	**B19**	RA-7(F) RA-7(R)	11177-12183	1006bp	pGL3-Basic	*Kpn*I *Hind*III	1013bp+4770bp	*Eco*RV* *Bgl*II	590bp+5193bp
	P19				pGL3-Promoter		1205bp+4770bp		590bp+5385bp
20	**B20**	RA-8(F) RA-8(R)	12132-13280	1148bp	pGL3-Basic	*Kpn*I *Nco*I	1188bp+4737bp	*Acc*I* *Kpn*I	429bp+2684bp+2812bp
	P20				pGL3-Promoter		1380bp+4737bp		429bp+2876bp+2812bp
21	**B21**	RA-9(F) RA-9(R)	13194-14286	1092bp	pGL3-Basic	*Kpn*I *Nco*I	1132bp+4737bp	*Bgl*II*	485bp+5384bp
	P21				pGL3-Promoter		1324bp+4737bp		485bp+5576bp

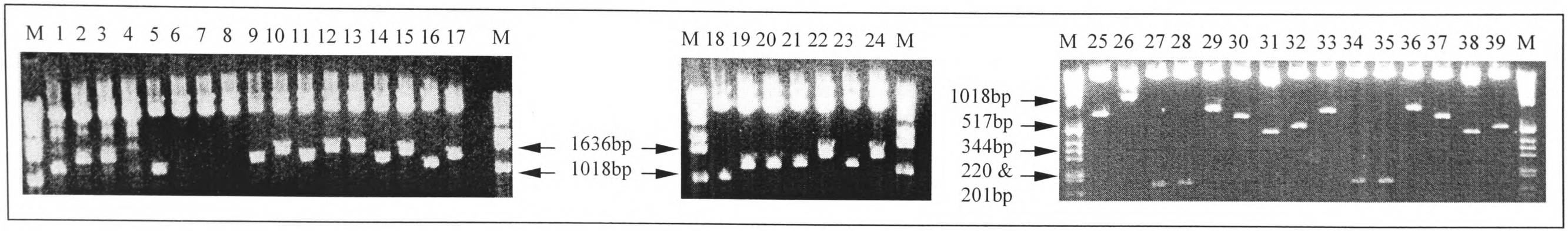


Figure 6.2.9: Examples of clones screened for incorporation of the correct insert into pGL3 luciferase reporter vectors. Restriction digestion of approximately 1µg of plasmid DNA was carried out as outlined in table 6.2.3. Lanes 1-24 are examples of clones digested with vector-specific enzymes, to excise the entire insert. Lanes 25-39 are examples of clones digested with enzymes cutting both within the vector and the insert, to confirm identity and orientation of the insert. Lane 1 is *Kpn*I+*Nar*I-digested B8 clone, while lanes 2-4 are similarly digested P8 clones. Lane 5 is *Kpn*I+*Hind*III-digested B10 clone, while lanes 6-8 are similarly digested P10 clones (all three negative). Lanes 9 and 10 are *Kpn*I+*Hind*III-digested B12 and P12 clones, respectively. Lane 11 is *Kpn*I+*Hind*III-digested B14 clone, while lanes 12 and 13 are similarly digested P14 clones. Lanes 14 and 15 are *Kpn*I+*Hind*III-digested B16 and P16 clones, respectively. Lanes 16 and 17 are *Sac*I+*Hind*III-digested clones B18 and P18, respectively. Lane 18 is *Kpn*I+*Hind*III-digested B19 clone, while lanes 19 and 20 are similarly digested P19 clones. Lanes 21 and 22 are *Kpn*I+*Nco*I-digested B20 and P20 clones, respectively, while lanes 23 and 24 are similarly digested B21 and P21 clones, respectively. Lanes 25 and 33 are *Eco*RV+*Bgl*II-digested B8 and P8 clones, respectively. Lane 26 is *Acc*I+*Kpn*I-digested B10 clone. Lanes 27 and 34 are *Bgl*II-digested B12 and P12 clones, respectively, while lanes 28 and 35 are similarly digested B16 and P16 clones, respectively. Lanes 29 and 36 are *Pst*I+*Bgl*II-digested B18 and P18 clones, respectively. Lanes 30 and 37 are *Eco*RV+*Bgl*II-digested B19 and P19 clones, respectively. Lanes 31 and 38 are *Acc*I+*Kpn*I-digested B20 and P20 clones, respectively. Lanes 32 and 39 are *Bgl*II-digested B21 and P21 clones, respectively. Digests 25-39 were analysed on a 2% (w/v) agarose gel. Lanes M are 1µg of 1kb ladder. Positive clones were obtained for all constructs, apart from P10 (lanes 6-8). A positive P10 clone was acquired after repeating the entire cloning procedure.

6.2.5 Search for regions bearing core promoter elements of the mouse *Nat2* gene

The 21 constructs in pGL3-Basic vector (table 6.2.3) were used in dual-luciferase reporter assays (section 2.3.5.1), following transient transfection (section 2.2.4.2) of BNL.CL2 cells (section 6.2.1). Co-tranfections were carried out with 10μg of purified plasmid DNA from each construct, plus 10μg of purified pRL-TK *Renilla* luciferase reporter vector (Appendix 1.2e). Control co-transfections were performed using 10μg of pRL-TK vector and an equal amount of pGL3-Basic or pGL3-Control vector (Appendix 1.2), lacking insert. Lysates of transfected BNL.CL2 cells were used to measure the luminescence generated by the firefly and *Renilla* luciferases (section 2.3.5.1), which are products of the pGL3 and pRL-TK vectors, respectively. *Renilla* luciferase activity was used as an internal standard, to account for variation in transfection efficiency with the BNL.CL2 cells.

Only construct B7 provided high luciferase activity, comparable to that of the pGL3-Control vector used as positive control (figure 6.2.10). Construct B7 covers the genomic region from 133bp upstream to 232bp within the NCE of *Nat2* (nucleotide position 3255-3620 of 129/Ola clone A; Appendix 2), therefore, it contains the transcription initiation site of the gene, mapped by RNase protection assay (section 6.2.3). The same genomic region was predicted by NIX analysis to contain a putative promoter (section 4.2.2.2).

The luminescence produced by constructs B11 and B17 was higher than the background luminescence generated by the pGL3-Basic vector alone, but much lower (15- and 12-fold, respectively) than the luminescence produced by the pGL3-Control vector (figure 6.2.10). NIX analysis (section 4.2.2.2) did not predict any promoter elements in the genomic regions covered by these two constructs. It is unlikely that either construct B11 or B17 contains a functional promoter. The luminescence generated by the rest of the constructs was either undetectable or similar to the background luminescence of the pGL3-Basic vector. No luminescence was detected in lysates of mock-transfected cells (figure 6.2.10).

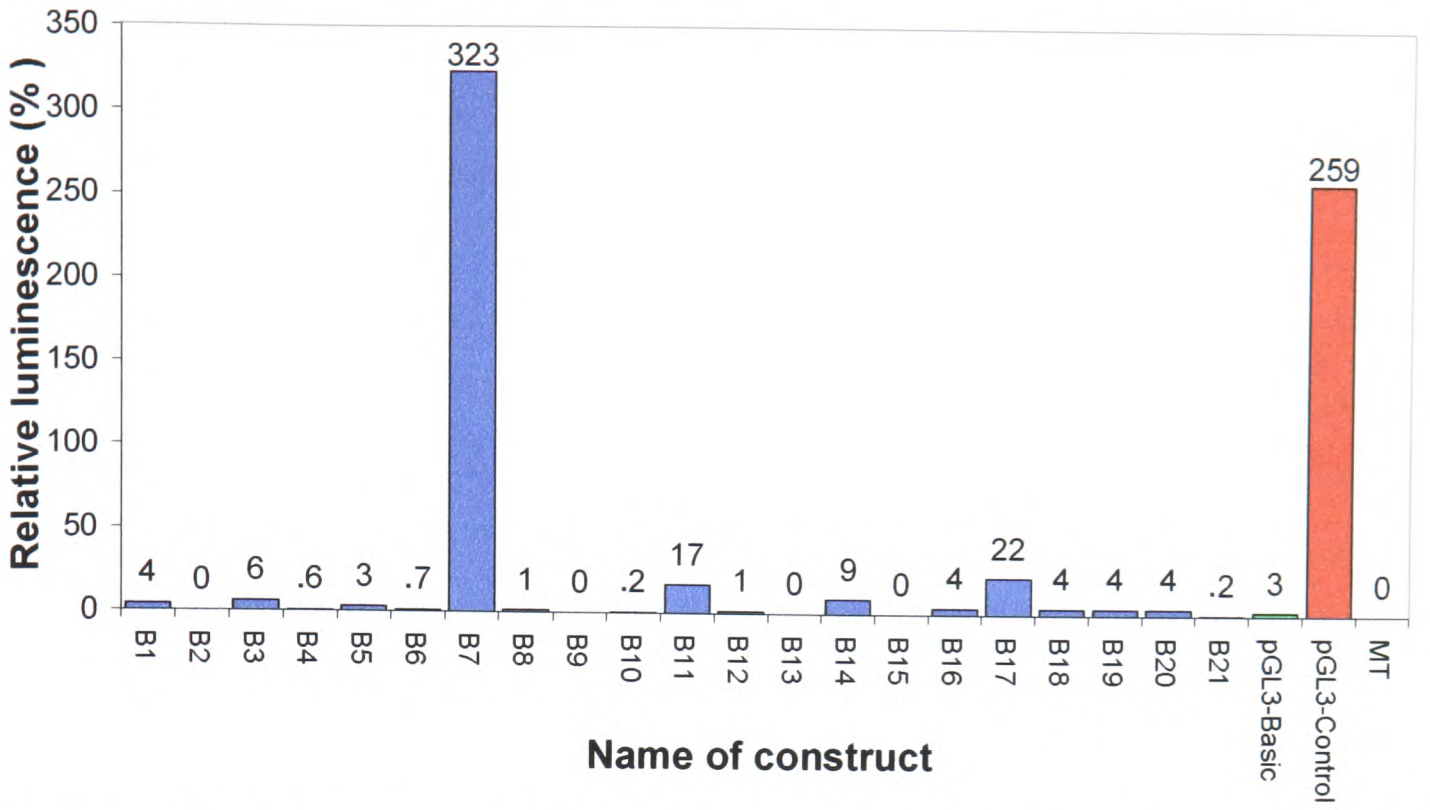


Figure 6.2.10: Luciferase activities generated by pGL3-Basic constructs in BNL.CL2 cells. Plasmid DNA from constructs B1-B21 (table 6.2.3) was co-transfected with pRL-TK vector into BNL.CL2 cells. The generated cell lysates were subjected to dual-luciferase reporter assay, measuring both firefly and *Renilla* luciferase activity. The luminescence produced by firefly luciferase was expressed as percentage relative to the luminescence produced by *Renilla* luciferase (relative luminescence). Vector pGL3-Control (without insert) was co-transfected with pRL-TK, to serve as a positive control for firefly luciferase activity. Vector pGL3-Basic (without insert) was co-transfected with pRL-TK, to provide a measure of background firefly luminescence. Mock-transfected cells (MT) were used as a negative control for luciferase activity. The relative luminescence for each cell lysate is shown at the top of the corresponding bar.

6.2.6 Search for regions bearing putative transcription regulatory elements for the mouse *Nat2* gene

The 15 constructs in pGL3-Promoter vector, plus construct E7 in pGL3-Enhancer vector (table 6.2.3), were used in dual-luciferase reporter assays, as described in the previous section (6.2.5). An additional control co-transfection was performed, using 10μg of pRL-TK vector with an equal amount of pGL3-Promoter vector, lacking insert. Luminescence was measured in cell lysates prepared from more than one plate of BNL.CL2 cells, independently transfected with each construct. Table 6.2.4 shows the mean relative luminescence calculated for each set of measurements and the corresponding standard deviations.

The produced luciferase activity varied between constructs (table 6.2.4 and figure 6.2.11). Most of the constructs generated luciferase activity higher than that of the pGL3-Promoter (Appendix 1.2b), but comparable to that of the pGL3-Control (Appendix 1.2d) vector without insert. On average, constructs covering the region upstream of the NCE and downstream of the coding region of *Nat2* provided higher levels of luciferase activity, compared with the constructs spanning the 6.1kb intron of the gene. Construct P4 provided significantly higher ($p<0.05$) and construct P10 significantly lower ($p<0.01$) luciferase activity than either the pGL3-Promoter or the pGL3-Control vector alone (table 6.2.4 and figure 6.2.11). Therefore, the inserts of constructs P4 and P10 might contain enhancer and repressor elements, respectively.

Table 6.2.4: Results of the search for transcription regulatory elements around the *Nat2* gene. Dual-luciferase reporter assays were performed with lysates of BNL.CL2 cells transfected with pGL3-Promoter constructs (table 6.2.3). More than one independent measurement (n) were taken, using the products of multiple transfections with the same construct. The relative luminescence is provided for each set of measurements. Genomic segment 7 (figure 6.2.8), identified to contain a core promoter (construct B7 in figure 6.2.10), was cloned into the pGL3-Enhancer vector (construct E7 below), which lacks a promoter, but has a SV40 enhancer (Appendix 1.2c).

Name of construct	Relative luminescence (%) (mean ± standard deviation)
P1	408±77 (n=2)
P2	clone not available
P3	486±18 (n=2)
P4	817±226 (n=2)
P5	592±73 (n=2)
P6	448±29 (n=2)
E7	281±32 (n=3)
P8	241±30 (n=6)
P10	110±10 (n=6)
P12	239±48 (n=6)
P14	265±87 (n=6)
P16	252±67 (n=6)
P18	197±32 (n=6)
P19	469±144 (n=6)
P20	350±171 (n=6)
P21	321±93 (n=6)
pGL3-Promoter	174±48 (n=11)
pGL3-Control	417±183 (n=6)
pGL3-Basic	4±3.3 (n=9)
MT*	0 (n=6)

**Mock transfected cells*

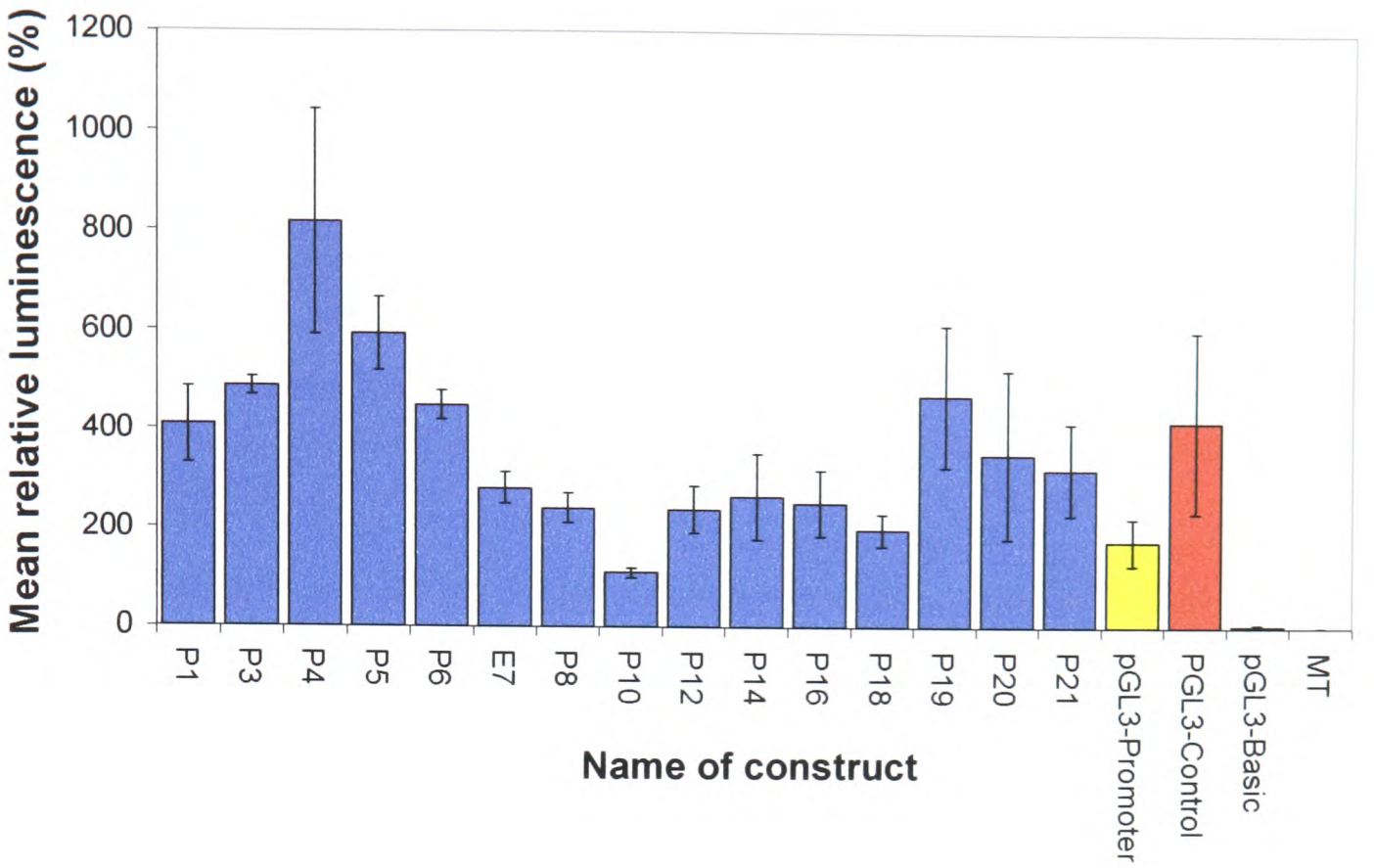


Figure 6.2.11: Luciferase activities generated by pGL3-Promoter constructs in BNL.CL2 cells. Plasmid DNA from constructs P1-P21 (table 6.2.3) was co-transfected with pRL-TK vector into BNL.CL2 cells. The generated cell lysates were subjected to dual-luciferase reporter assay, measuring both firefly and *Renilla* luciferase activity. The luminescence produced by firefly luciferase was expressed as percentage relative to that produced by *Renilla* luciferase. Vectors pGL3-Control and pGL3-Promoter, lacking insert, were used to provide a measure of the firefly luminescence produced in the presence or the absence of a SV40 enhancer, respectively. Vector pGL3-Basic, also lacking insert, was used to provide a measure of the background firefly luminescence. Mock-transfected cells (MT) were used as a negative control for luciferase activity.

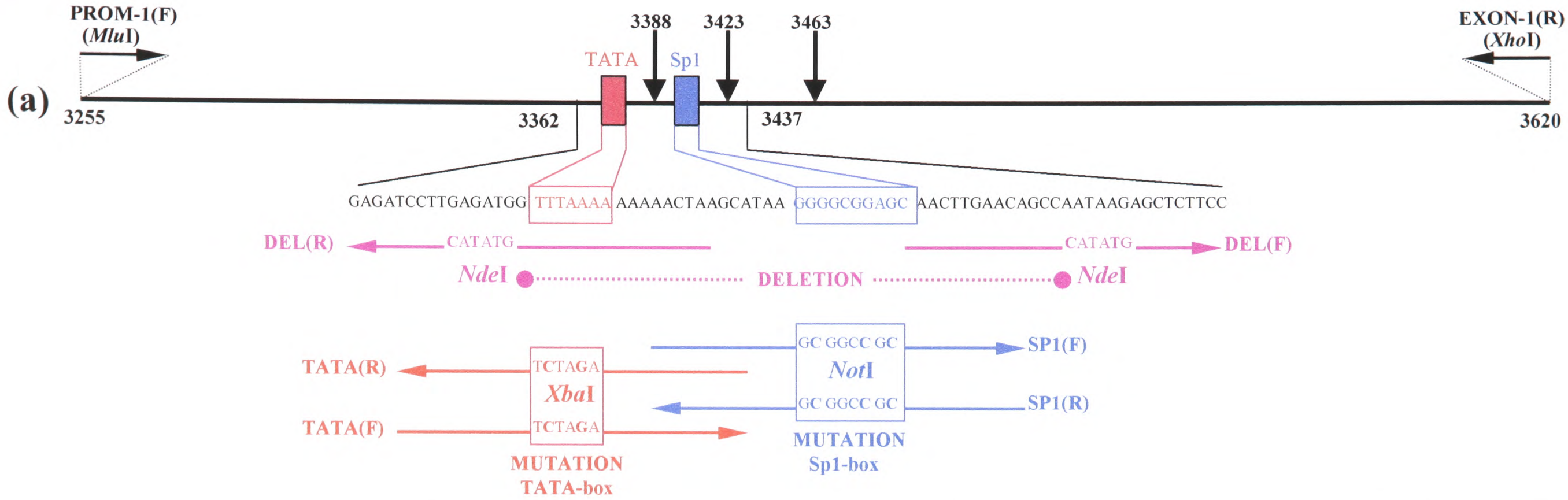
6.2.7 Inactivation of the *Nat2* promoter by deletion or mutation

The 365bp insert of construct B7, providing high levels of reporter gene expression in BNL.CL2 cells (figure 6.2.10), was analysed for the presence of consensus regulatory sequences, deposited in the TRANSFAC database (http://transfac.gbf.de). Two of the identified elements were likely to compose a core promoter for the *Nat2* gene: a) A TATA-box (TTTAAAA) at position 3378 of 129/Ola clone A (Appendix 2), which has a core similarity of 0.928 to the consensus

sequence TATAAAA (Beebee and Burke, 1990) and was also predicted by NIX analysis (section 4.2.2.2); b) an Sp1-box (GGGGCGGAGC) at position 3400 of 129/Ola clone A (Appendix 2), which has a core similarity of 1.000 to the consensus sequence GGGCGG (Latchman, 1998b) and a perfect match to the promoter sequence of the murine genes for hypoxanthine phosphoribosyl transferase (*Hprt*) and dihydrofolate reductase (*Dhfr*) (Dynan, 1986).

To investigate whether the promoter activity of construct B7 is due to any of these particular elements, three strategies were adopted (figure 6.2.12): 1) deletion of region 3371-3428 of 129/Ola clone A (Appendix 2), containing both candidate promoter elements (construct DEL); 2) Introduction of point mutations only to the putative TATA-box [construct TATA(mt)]; 3) Introduction of point mutations only to the putative Sp1-box [construct SP1(mt)]. The three constructs were otherwise identical to construct B7, containing the intact promoter (figure 6.2.12). The cloning steps leading to the generation of each construct are described in figure 6.2.13. Constructs B7, DEL, TATA(mt) and SP1(mt) were separately co-transfected with pRL-TK vector into BNL.CL2 cells, and the generated cell lysates were analysed by dual-luciferase reporter assay (figure 6.2.14).

Deletion of both TATA and Sp1 elements in construct DEL caused a 97% decrease in relative luminescence (figure 6.2.14). The effect of mutations in the TATA-box of construct TATA(mt) and the Sp1-box of construct SP1(mt) was also substantial, leading to a reduction in relative luminescence by 90% and 92%, respectively (figure 6.2.14). This demonstrates that both of the predicted elements are crucial components of the promoter driving high levels of luciferase expression in BNL.CL2 cells transfected with the B7 construct. These elements are adjacent to the transcription initiation site of murine *Nat2* (section 6.2.3) and would be expected to regulate expression of the gene *in vivo*.



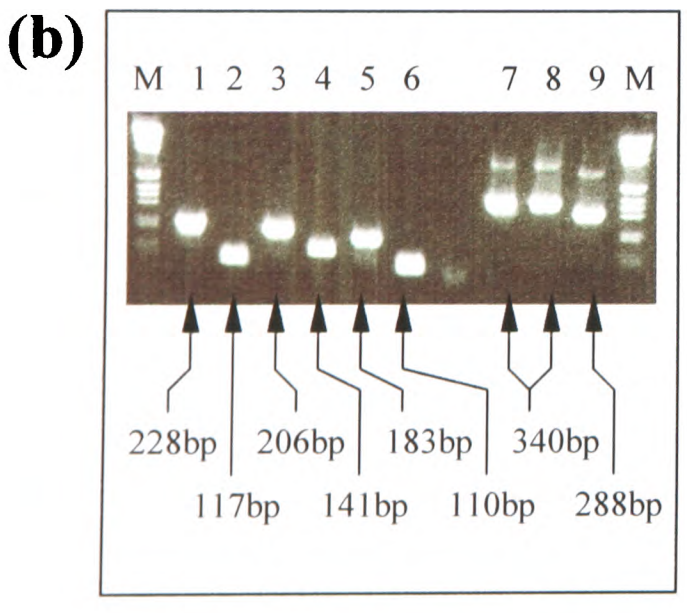


Figure 6.2.12: Strategies for deletion or mutation of candidate promoter elements upstream of the mouse *Nat2* gene. a) Diagram showing the insert of the B7 construct (black horizontal line), relative to the beginning of 129/Ola clone A (Appendix 2). The sequence of the 75bp region containing the putative TATA and Sp1 boxes is provided. Positions 3388, 3423 and 3463 (thick vertical arrows) correspond to the *Nat2* transcription initiation sites determined by RNase protection assay (section 6.2.3). Primer pairs PROM-1(F)/DEL(R) and DEL(F)/EXON-1(R) (table 6.2.2) were used to amplify the portions of the insert upstream (DEL-5′) and downstream (DEL-3′) of the candidate promoter elements, respectively. The two portions were ligated at the *Nde*I site of the DEL-primers (coloured purple), introducing a 57bp deletion (purple dotted line) spanning both elements (construct DEL). Primers TATA(F) and TATA(R) (table 6.2.2) span the TATA-box and contain a *Xba*I site. Amplification of the 5′ portion (TATA-5′) of the insert with primer pair PROM-1(F)/TATA(R) and of the 3′ portion (TATA-3′) with primer pair TATA(F)/EXON-1(R), followed by ligation at the *Xba*I site of the TATA-primers (coloured red), introduced two point mutations (in bold) to the TATA-box [construct TATA(mt)]. Primers SP1(F) and SP1(R) (table 6.2.2) span the Sp1-box and contain a *Not*I site. Amplification of the 5′ portion (SP1-5′) of the insert with primer pair PROM-1(F)/SP1(R) and of the 3′ portion (SP1-3′) with primer pair SP1(F)/EXON-1(R), followed by ligation at the *Not*I site of the SP1-primers (coloured blue), introduced three point mutations (in bold) to the Sp1-box [construct SP1(mt)]. b) Amplification of the 5′ and 3′ portions of each insert, as described in (a). Following ligation of the two portions, amplification of the entire insert was performed with primers PROM-1(F) and EXON-1(R). The products are shown on a 3% (w/v) agarose gel. Lanes 1, 2 and 7 are the amplified TATA-3′ and TATA-5′ portions, and their ligation product TATA(mt), respectively. Lanes 3, 4 and 8 are the amplified SP1-3′ and SP1-5′ portions, and their ligation product SP1(mt), respectively. Lanes 5, 6 and 9 are the amplified DEL-3′ and DEL-5′ portions, and their ligation product DEL, respectively. Lanes M are 1μg of 1kb ladder.

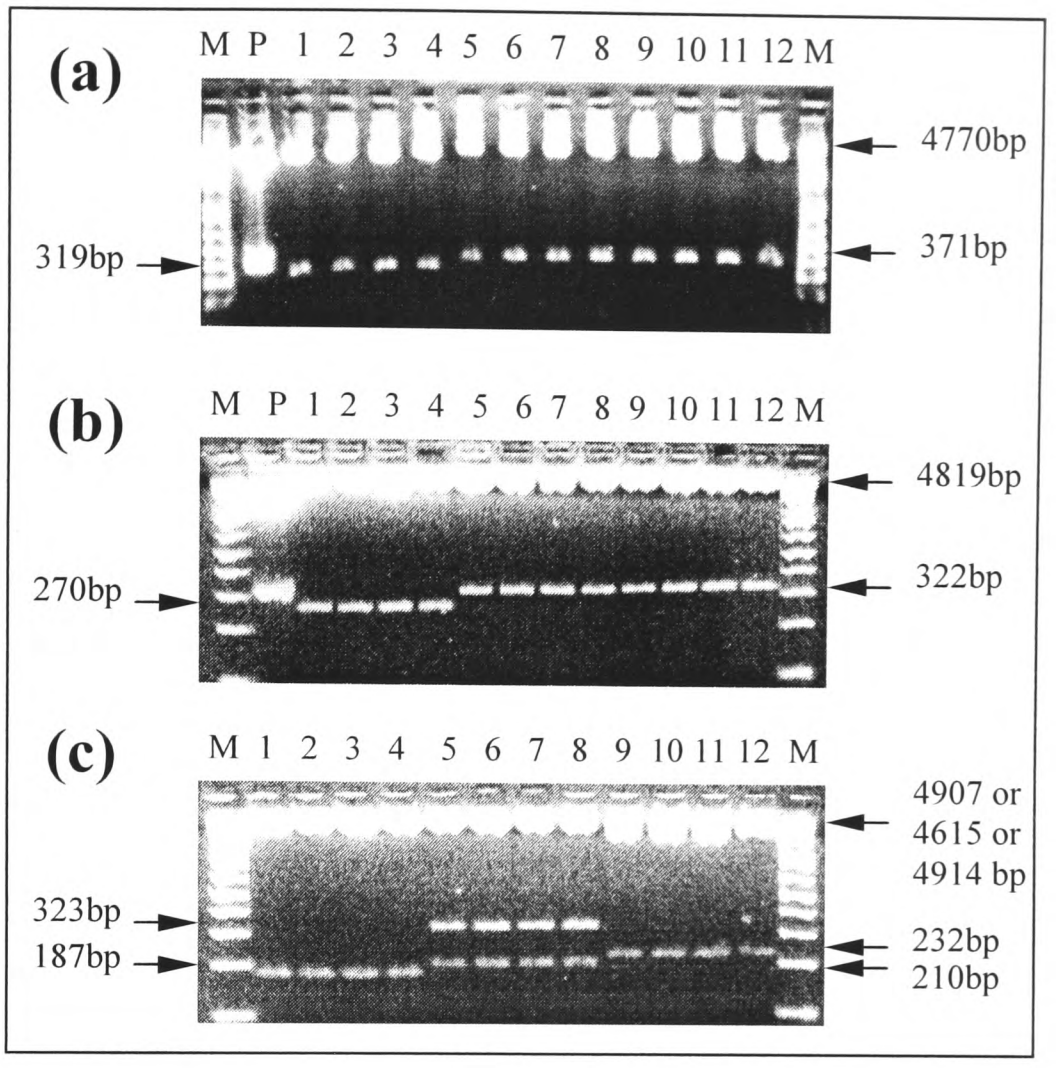


Figure 6.2.13: Generation of constructs DEL, TATA(mt) and SP1(mt). Ligation of the mutated inserts (figure 6.2.12) to the *Mlu*I and *Xho*I sites of the pGL3-Basic vector was followed by transformation of JM109 *E. coli* cells (section 2.2.3.10). Several colonies from each plate were screened by PCR with primers PROM-1(F)/ EXON-1(R) (table 6.2.2). The figure shows digestion of plasmid DNA (1μg) from positive colonies, expected to carry the DEL (lanes 1-4), SP1(mt) (lanes 5-8) and TATA(mt) (lanes 9-12) constructs, in order to check for a) the presence and b) the orientation of each insert in the pGL3-Basic vector, as well as c) to confirm successful mutation or deletion of the candidate promoter elements. Digestion of plasmid DNA from the intact B7 construct is shown in lanes P, for comparison. Lanes M are 1μg of 1kb Plus DNA ladder. Products were analysed on 2.5% (w/v) agarose gels. a) Digestion with *Kpn*I and *Hind*III, cutting at positions 5 and 53 of the vector, respectively (Appendix 1.2a). All clones contain the correct size insert. b) Digestion with *Eco*RV, cutting at position 33 of each insert (position 3288 of 129/Ola clone A; Appendix 2), and *Bgl*II, cutting at position 36 of the vector (Appendix 1.2a). All inserts have been cloned in the right orientation into the pGL3-Basic vector. c) Lanes 1-4: Digestion of DEL clones with *Nde*I (site of deletion) and *Bgl*II (vector). Lanes 5-8: Digestion of SP1(mt) clones with *Not*I (site of Sp1-box mutation and position 4651 of the vector) and *Bgl*II (vector). Lanes 9-12: Digestion of TATA(mt) clones with *Xba*I (site of TATA-box mutation) and *Bgl*II (vector). Incorporation of the mutations was successful. The inserts of selected DEL, TATA(mt) and SP1(mt) clones were also sequenced with primer PROM-1(F).

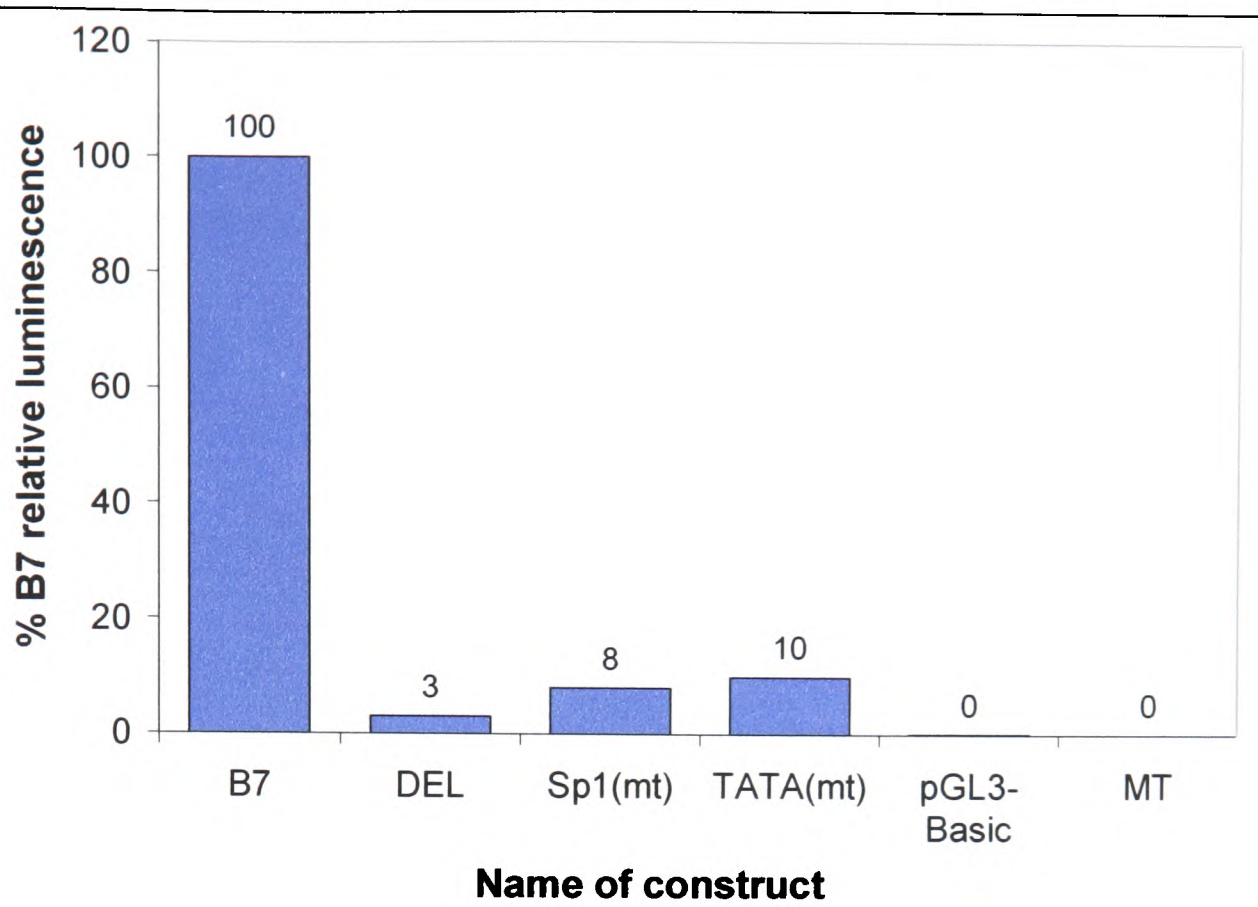


Figure 6.2.14: Luciferase activity following deletion or mutation of the putative TATA and Sp1 elements of the *Nat2* promoter. Plasmid DNA from constructs B7, DEL, SP1(mt) and TATA(mt) (figures 6.2.12 and 6.2.13) was co-transfected with pRL-TK vector into BNL.CL2 cells. The generated cell lysates were subjected to dual-luciferase reporter assay. The luminescence produced by firefly luciferase was expressed relative to that produced by *Renilla* luciferase. Vector pGL3-Basic (without insert) was co-transfected with pRL-TK, to provide a measure of the background firefly luminescence. Mock-transfected cells (MT) were used as a negative control for luciferase activity. The relative luminescence calculated for DEL, SP1(mt) and TATA(mt) constructs is presented as percentage of the relative luminescence calculated for the intact B7 construct (100%). The corresponding values are provided at the top of each bar.

6.2.8 Promoter activity of 5′-nested deletion constructs spanning the region upstream of the *Nat2* non-coding exon

To investigate whether the genomic region upstream of the *Nat2* NCE contains elements that might regulate activity of the core promoter (section 6.2.7), the 2682bp region between the *Nat1* gene and the *Nat2* NCE (nucleotide position 938-3620 of 129/Ola clone A; Appendix 2) was progressively deleted from the 5′ end and cloned into the pGL3-Basic vector (figures 6.2.15 and 6.2.16). The generated constructs (D1-D8), as well as construct B7 (section 6.2.5), were co-transfected with pRL-TK vector into BNL.CL2 cells, the lysates of which were then assayed for firefly and *Renilla* luciferase activities (figure 6.2.17).

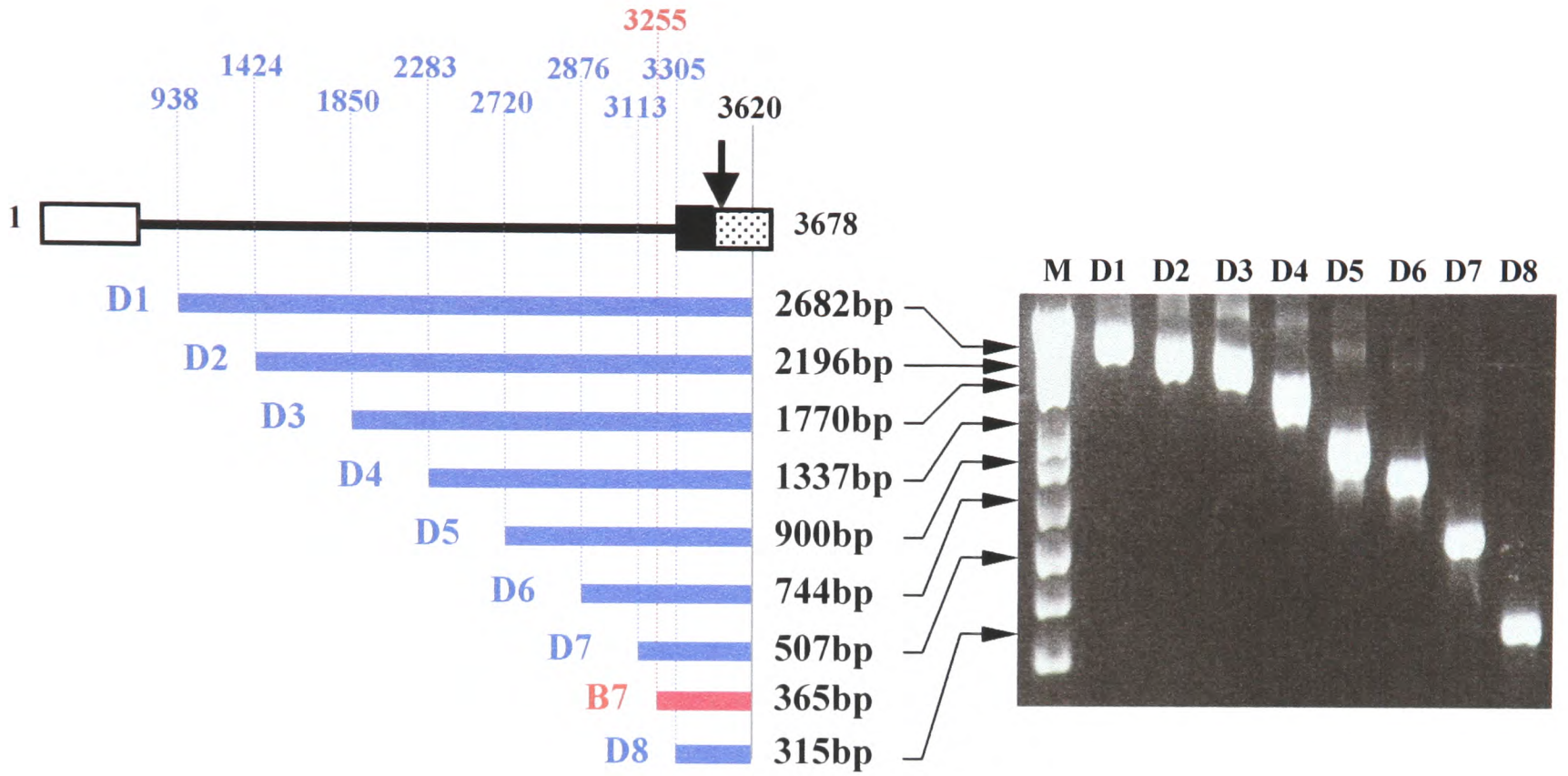


Figure 6.2.15: Deletion constructs spanning the region upstream of the *Nat2* NCE. Fragments D1-D8 were amplified with *Pfu*-DNA polymerase, using the forward primers REP3(F), REP5(F), REP7(F), REP9(F), REP11(F), REP13(F), REP15(F) and REP17(F), respectively, and the same reverse primer EXON-1(R) (table 6.2.2). All fragments have the same 3′ end at position 3620 of 129/Ola clone A (Appendix 2) and nested 5′ ends, the position of which is indicated in reference to the same clone. The relative position of the B7 insert (section 6.2.4) is also shown (thick red line). The longest fragment (D1) covers the region 487bp downstream of the *Nat*1 gene (white box) to 232bp within the *Nat2* NCE (dotted box). The region of the core *Nat2* promoter (black box) is also shown. The thick vertical arrow indicates the relative position of the *Sac*I site used for screening of plasmid clones, after ligation of each fragment into the pGL3-Basic vector (figure 6.2.16).

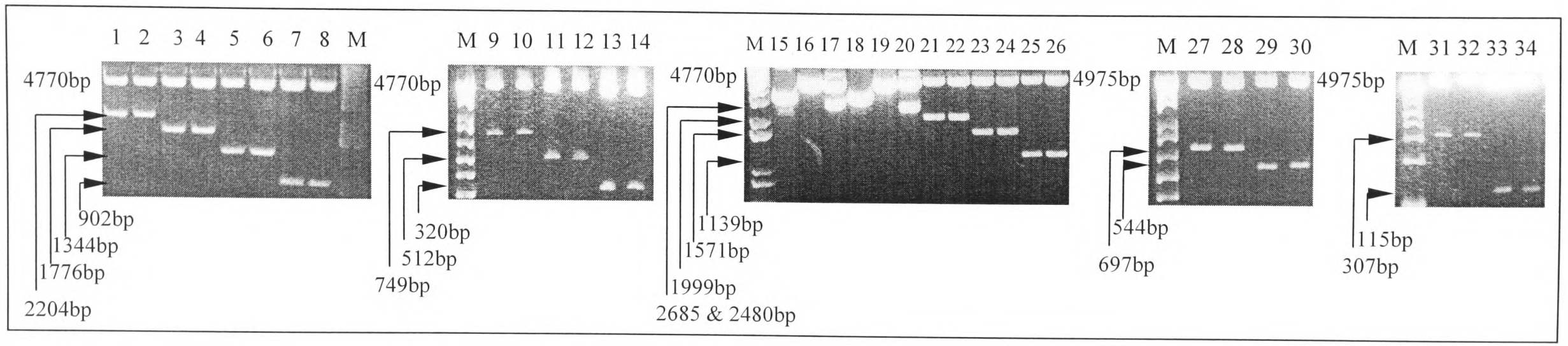


Figure 6.2.16: Generation of deletion constructs D1-D8. Ligation of inserts D1-D8 (figure 6.2.15) to the *Mlu*I and *Xho*I sites of the pGL3-Basic vector was followed by transformation of JM109 *E. coli* cells (section 2.2.3.10). Several colonies from each plate were screened by PCR with primers REP17(F)/EXON-1(R) (table 6.2.2). Two positive colonies from each plate were then checked for the presence of the correct insert by digestion of 1μg plasmid DNA. Lanes 1-17: Digestion with *Kpn*I and *Hind*III, cutting at positions 5 and 53 of the vector, respectively (Appendix 1.2a), in order to excise the entire insert. Digests 9-14 were analysed on a 2% (w/v) agarose gel. The order of the samples is: D2 clones (lanes 1 and 2), D3 clones (lanes 3 and 4), D4 clones (lanes 5 and 6), D5 clones (lanes 7 and 8), D6 clones (lanes 9 and 10), D7 clones (lanes 11 and 12), D8 clones (lanes 13 and 14) and D1 clones (lanes 15-17). Lanes 18-34: Digestion with *Sac*I, cutting at positions 3428 of 129/Ola clone A (Appendix 2) and 11 of the vector (Appendix 1.2a), to check the insert orientation. Digests, 27-30 and 31-34 were analysed on 2% and 3% (w/v) agarose gels, respectively. The order of the samples is: D1 clones (lanes 18-20), D2 clones (lanes 21 and 22), D3 clones (lanes 23 and 24), D4 clones (lanes 25 and 26), D5 clones (lanes 27 and 28), D6 clones (lanes 29 and 30), D7 clones (lanes 31 and 32) and D8 clones (lanes 33 and 34). Lanes M are 1μg of 1kb Plus ladder. All clones provided the expected restriction fragments, apart from those in lanes 15, 16, 18 and 19.

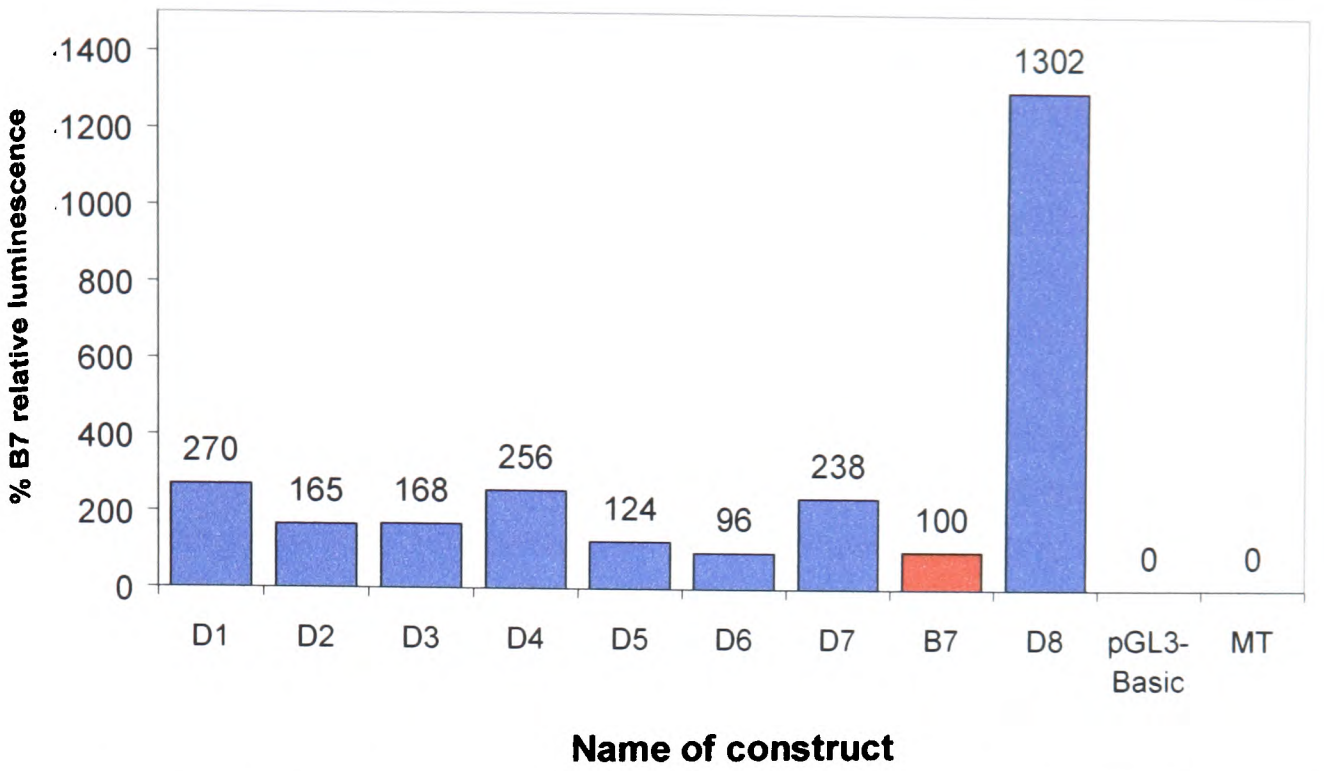


Figure 6.2.17: Luciferase activities generated by the 5′-nested deletion constructs. Plasmid DNA from constructs D1-D8 (figures 6.2.15 and 6.2.16) was co-transfected with pRL-TK vector into BNL.CL2 cells. A co-transfection with pRL-TK vector and construct B7 (figure 6.2.10) was also performed for comparison. The generated cell lysates were subjected to dual-luciferase reporter assay. The luminescence produced by firefly luciferase was expressed relative to that produced by *Renilla* luciferase. Vector pGL3-Basic (without insert) was co-transfected with pRL-TK, to provide a measure of background firefly luminescence. Mock-transfected cells (MT) were used as a negative control for luciferase activity. The relative luminescence calculated for each construct is presented as percentage of the relative luminescence calculated for the B7 construct (red bar). The corresponding values are provided on top of the bars.

Luciferase activity was low in lysates of cells transfected with constructs D1-D7, as well as B7, but increased sharply (by up to 13-fold) in cells transfected with construct D8 (figure 6.2.17). The latter is the shortest one of the deletion constructs, extending only 73bp upstream of the *Nat2* core promoter (figure 6.2.15). It appears that the activity of the core promoter is affected by upstream negative elements, probably located within region 3255-3305 of 129/Ola clone A (Appendix 2), i.e. in the 50bp non-overlapping portion between constructs B7 and D8 (figure 6.2.15). TRANSFAC analysis predicted several transcription factor binding sites in this region, including one C/EBPβ and five GATA sites with core similarity of 1.0 to the corresponding consensus sequences. Introduction of short deletions or point mutations to this candidate region may allow precise localisation of the postulated negative element(s). The observed variation in luciferase activity between constructs D1-D7

(figure 6.2.17) is probably due to sub-optimal experimental conditions, although the presence of a repressor could mask the effect of other upstream regulatory elements to a variable degree.

6.2.9 Identification of transcription factor binding sites near the mouse *Nat2* promoter by electrophoretic mobility shift assay (EMSA)

As a first step towards investigation for transcription factors regulating expression of the mouse *Nat2* gene, the region upstream of the NCE was examined for the presence of protein binding sites by EMSA. PCR was used to amplify five probes (GS1-GS5), spanning the 735bp region from position 3437 (beginning of NCE) to 2702 (686bp upstream of the NCE) of 129/Ola clone A (Appendix 2). The primers used for amplification are shown in table 6.2.5 and the corresponding products in figure 6.2.18. Primers mNAT2-516 and mNAT2-691 (table 6.2.5) were used to amplify part of the *Nat2* coding region (figure 6.2.18), to serve as a control probe (NSC) for non-specific binding in EMSAs.

Optimisation of EMSA was initially carried out using various oligonucleotide probes (figure 6.2.19). Incubation (section 2.2.7.3) of a ^{32}P-labelled (section 2.2.7.1) Oct1-specific oligonucleotide (table 6.2.6) with nuclear protein extract from HeLa cells or with whole protein extract from BNL.CL2 cells (section 2.2.7.2) shifted bands on the autoradiogram, following electrophoresis of the reaction product on a non-denaturing 4% (w/v) polyacrylamide gel (section 2.2.2.2). The observed binding was specific, as the shifted bands were removed by competition with 50-fold molar excess of unlabelled specific probe (Oct1 oligo; table 6.2.6), but remained unaffected by competition with the same molar excess of unlabelled non-specific probe (AP2 oligo; table 6.2.6). The same experimental conditions also allowed specific binding of AP2 protein to an AP2-specific oligonucleotide (table 6.2.6 and figure 6.2.19).

Further optimisation with the longer (144bp) GS1 probe (figure 6.2.18) was carried out, using serial dilutions (1:1 to 1:50) of BNL.CL2 cell extract. Optimal binding was observed with a low dilution (1:2) of extract. The effect of salt strength was also investigated. NaCl concentrations, ranging from 40 to 120mM, did not affect

protein binding to the GS1 probe. A final concentration of 70mM NaCl was, therefore, used throughout.

Table 6.2.5: Primers used for the amplification of EMSA probes. The nucleotide sites of primer annealing are indicated relative to the beginning of 129/Ola clone A (Appendix 2). Primers marked with a (F) are forward, while those marked with a (R) are reverse. Forward primer mNAT2-516 was used with reverse primer mNAT2-691 to amplify part of the coding region of the *Nat2* gene (NSC probe). The annealing sites of these primers are, therefore, indicated relative to the beginning of the *Nat2* coding region (adenosine of the ATG translation initiation codon as position 1).

Primer name	Sequence (5′→3′)	Tm (°C)	Specificity
GS-1(F)	3293-CTCACGCAAAGAACGGAC-3310	62.1	GS1 probe
GS-1(R)	3437-GGAAGAGCTCTTATTGGC-3420	56.4	
GS-2(F)	3137-GGAATGTGGGTGACATG-3153	58.0	GS2 probe
GS-2(R)	3293-GATATCATGATATTTTTGC-3275	49.1	
GS-3(F)	2979-ACATAGGACAGTTGTCG-2995	50.7	GS3 probe
GS-3(R)	3138-CCTAATGTATGCCTTTCC-3121	54.4	
GS-4(F)	2818-ACTACATGCCACCAAGC-2834	56.9	GS4 probe
GS-4(R)	2996-ACGACAACTGTCCTATG-2980	50.7	
GS-5(F)	2702-CCCTCAGAACCATGCAG-2718	60.4	GS5 probe
GS-5(R)	2817-CGGGTGTGGCTCAGGTG-2801	67.0	
mNAT2-516	508-TTCCAAACCAAGAATTTA-525	46.0	NSC probe (mouse *Nat2*)
mNAT2-691	691- ACTCCTTCTGGGGTCTGCA -673	61.4	

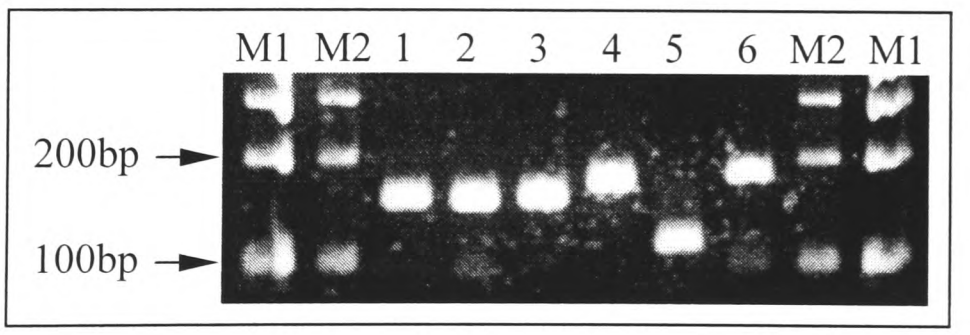


Figure 6.2.18: Amplification of the probes used for EMSA. Amplifications were carried out with *Pfu*-DNA polymerase, using 50ng of plasmid DNA from 129/Ola clone A as template. The primers used are shown in table 6.2.5. The products were gel-purified (section 2.2.1.3) and analysed on a 3% (w/v) agarose gel. Probes GS1-GS5 (lanes 1-5) are 144, 156, 159, 178 and 115bp in size, respectively, and span region 3437-2702 on 129/Ola clone A (Appendix 2). Probe NSC (lane 6) is 183bp in size and spans region 10358-10541 of 129/Ola clone A, corresponding to region 508-691 of the *Nat2* coding region. Lanes M1 and M2 are 1µg of 1kb Plus ladder and 5µl of BioMarker EXT ladder, respectively.

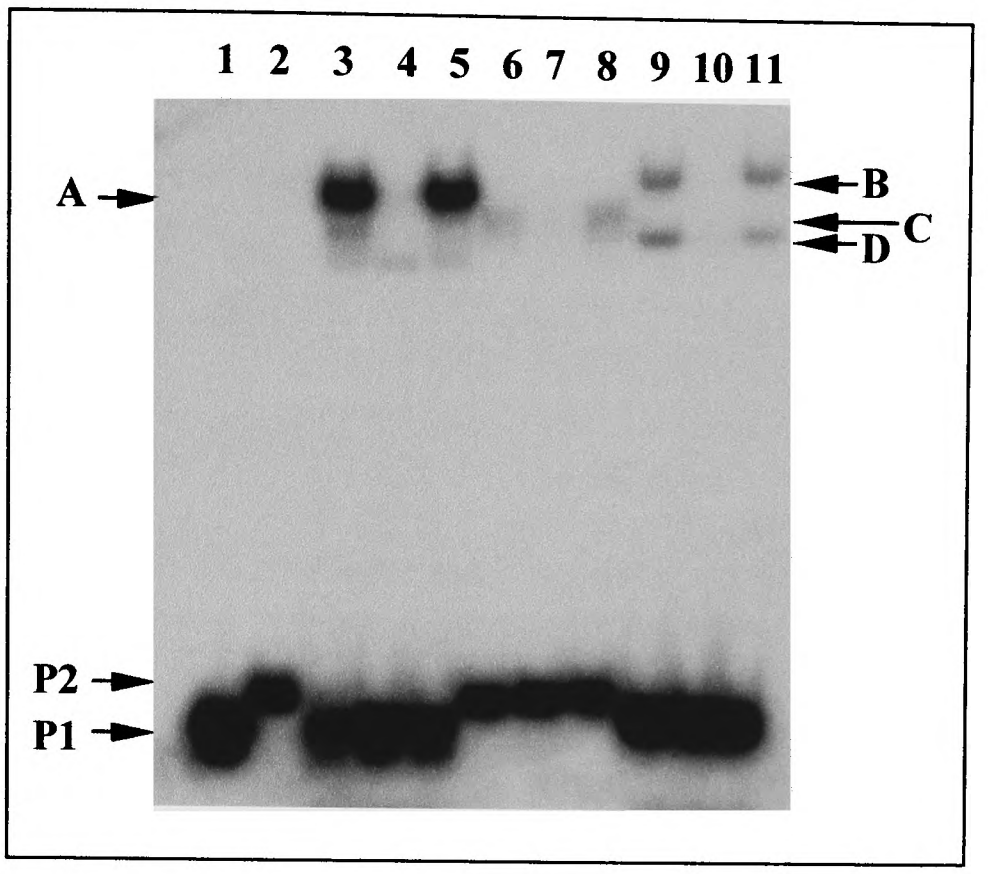


Figure 6.2.19: Optimisation of EMSA. Oct1-specific oligonucleotide probe P1 (lane 1) and AP2-specific oligonucleotide probe P2 (lane 2) were end-labelled with [γ-^{32}P]-dATP, using T4 polynucleotide kinase (section 2.2.7.1). The sequence of each probe is provided in table 6.2.6. The radiolabelled probes (35fmol) were incubated (section 2.2.7.3) with protein extract, in the presence or absence of a 50-fold molar excess (1.75pmol) of unlabelled specific or non-specific competitor. The reaction products were subjected to electrophoresis on a 4% (w/v) non-denaturing polyacrylamide gel (section 2.2.2.2). Lanes 3-5 are the products of incubation of radiolabelled Oct1-specific probe with nuclear protein extract from HeLa cells (Promega), in the absence of competitor (lane 3) or in the presence of specific (Oct1) (lane 4) or non-specific (AP2) (lane 5) competitor. Lanes 6-8 are the products of incubation of radiolabelled AP2-specific probe with AP2 protein (Promega), in the absence of competitor (lane 6) or in the presence of specific (AP2) (lane 7) or non-specific (Oct1) (lane 8) competitor. Lanes 9-11 are the products of incubation of radiolabelled Oct1-specific probe with BNL.CL2 whole cell protein extract, in the absence of competitor (lane 9) or in the presence of specific (Oct1) (lane 10) or non-specific (AP2) (lane 11) competitor. Bands A-D correspond to the protein-bound radiolabelled probes, retarded on the gel due to the increase in their molecular weight. Incubation of the Oct1-specific oligonucleotide probe with extract from murine BNL.CL2 cells provided two retarded bands (B and D), indicating that these cells must express two isoforms of the Oct1 transcription factor. These are distinct from the single Oct1 isoform detected in human HeLa cells (band A).

Table 6.2.6: Oligonucleotide probes used for EMSA. The specificity and sequence of each double-stranded oligonucleotide probe is presented (from Promega's Technical Bulletin #110 on Gel Shift Assay Systems). The probes were used for optimisation of EMSA (figure 6.2.19), as well as in competition assays with probes GS1-GS5 (figure 6.2.20).

Transcription factor	Sequence of specific oligonucleotide probe
AP2	5′-GAT CGA ACT GAC CGC CCG CGG CCC GT-3′ 3′-CTA GCT TGA CTG GCG GGC GCC GGG CA-5′
Sp1	5′-ATT CGA TCG GGG CGG GGC GAG C-3′ 3′-TAA GCT AGC CCC GCC CCG CTC G-5′
AP1 (c-Jun)	5′-CGC TTG ATG AGT CAG CCG GAA-3′ 3′-GCG AAC TAC TCA GTC GGC CTT-5′
Oct1	5′-TGT CGA ATG CAA ATC ACT AGA A-3′ 3′-ACA GCT TAC GTT TAG TGA TCT T-5′
CREB	5′-AGA GAT TGC CTG ACG TCA GAG AGC TAG-3′ 3′- TCT CTA ACG GAC TGC AGT CTC TCG ATC-5′
NF-κB	5′-AGT TGA GGG GAC TTT CCC AGG C-3′ 3′-TCA ACT CCC CTG AAA GGG TCC G-5′

Incubation of the GS1 probe (figure 6.2.18), containing the *Nat2* promoter, with BNL.CL2 cell extract produced five shifted bands (figure 6.2.20a). Protein binding was specific, as these bands were not affected by non-specific competition with unlabelled NSC probe, but were completely removed by specific competition with unlabelled GS1 probe. It appears that the GS1 probe contains binding sites for more than one transcription factor expressed endogenously in BNL.CL2 cells. Competition with unlabelled Sp1-specific oligonucleotide (table 6.2.6) changed the relative intensity of the shifted bands (lane 6 in figure 6.2.20a), indicating the likely presence of an Sp1 site within the promoter region of *Nat2*, consistent with previous observations (section 6.2.7). The rest of the oligonucleotide competitors listed in table 6.2.6 did not interfere with protein binding to the GS1 probe (figure 6.2.20a), in agreement with the TRANSFAC analysis presented in table 6.2.7.

Of the remaining four probes (figure 6.2.18), covering the region upstream of the *Nat2* core promoter, only probes GS3 and GS5 provided a potentially specific binding pattern (figures 6.2.20b and c, respectively). Incubation of probe GS3 with BNL.CL2 cell extract produced five shifted bands, four of which (A-D) were removed by competition with unlabelled GS3 probe and were, therefore, likely to represent specific protein binding (figure 6.2.20b). However, the intensity of these bands was

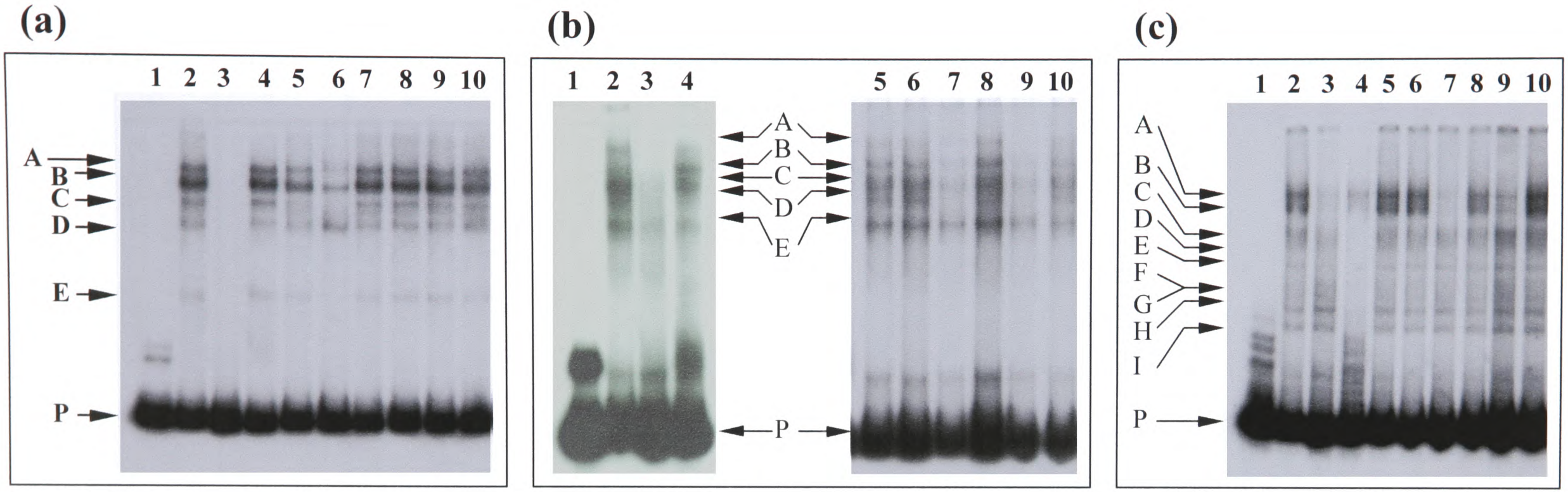


Figure 6.2.20: EMSAs using BNL.CL2 whole cell protein extract and probes GS1, GS3 and GS5. Probes (35fmol), end-labelled with [γ-^{32}P]-dATP, were incubated with freshly thawed BNL.CL2 whole cell protein extract, diluted 1:2 in extraction buffer (section 2.2.7). The final concentration of salt in the reaction mixture was 70mM. The binding reaction products were analysed on 4% (w/v) non-denaturing polyacrylamide gels (section 2.2.2.2). EMSAs were performed a) with GS1 probe, b) with GS3 probe and c) with GS5 probe. Lane 1 is the "probe only" control (section 2.2.7.3), lacking BNL.CL2 extract from the binding reaction mixture. Lane 2 is the product of incubation of the radiolabelled probe with BNL.CL2 extract. Lane 3 is the product of a "specific competition" reaction (section 2.2.7.3) with 1.75pmol of unlabelled specific probe. Lane 4 is the product of a "non-specific competition" reaction (section 2.2.7.3) with 1.75pmol of unlabelled NSC probe. Lanes 5-10 are the products of competition reactions with 1.75pmol of AP2, Sp1, AP1, Oct1, CREB and NF-κB oligonucleotide probes (table 6.2.6), respectively. The band of unbound probe is indicated with P. Probe bands running higher than P (e.g. lane 1 in b) are artefacts due to the large size (~150bp) of the free probe. The shifted bands of protein-bound probe are indicated with capital letters (A-I). A decrease in the relative intensity of the shifted bands in lanes 3-10 indicates interference of the unlabelled competitor with protein binding to the labelled probe.

reduced, following competition with unlabelled NSC probe, indicating that part of the observed binding may be non-specific (figure 6.2.20b). Probe GS5 provided nine shifted bands, only two of which (A and B) could be specific, although their intensity was significantly reduced by competition with unlabelled NSC probe (figure 6.2.20c). Only competition with unlabelled AP1- and CREB-specific oligonucleotides (table 6.2.6) interfered with protein binding to probes GS3 and GS5 (lanes 7 and 9 in figures 6.2.20b and c, respectively). The GS3 and GS5 probes are, therefore, likely to contain binding sites for AP1 and CREB transcription factors, present in BNL.CL2 protein extract, in agreement with the results of the TRANSFAC analysis (table 6.2.7).

Table 6.2.7: Screening of probes for putative transcription factor binding sites by TRANSFAC analysis. Summary indicating the type of transcription factors, predicted to bind to probes GS1-GS5, as well as the orientation [forward (+) or reverse (-)] of the DNA strand bearing the relevant recognition sites. Only good predictions, with a core similarity of ≥0.9 and a matrix similarity of ≥0.85 to the consensus sequence, were included. When binding to a specific probe is supported by EMSA (figure 6.2.20), the transcription factor is indicated in bold. The position of each probe is indicated relative to the beginning of 129/Ola clone A (Appendix 2).

Probe name (position)	Transcription factor (strand orientation)
GS5 (2702-2817)	ATF(-), **CREB(-)**, **AP1(-)**, **AP1FJ(-)**, ER(+), **CREB(-)**, TCF11(-), TCF11(+), **AP1(+)**, **AP1FJ(+)**, RORA2(-), CMYB(-), GATA1(+), GATA2(+), GATA3(+), LMO2COM(+), IK2(+), MZF1(+), GATA1(+), LMO2COM(+), NKX25(-), E47(-), LMO2COM(-), MyoD(-), SREBP1(+), δEF1(+), NF1(-), Oct1(-)
GS4 (2818-2996)	GFI1(+), TCF11(-), AP1FJ(+), GRE(-), NF1(-), S8(+), TST1(+), IK2(+), NFAT(+), HFH8(-), HFH3(-), GATA(-), GATA1(-), GATA2(-), GATA1(-), GATA3(-), LMO2COM(-), GFI1(+), S8(+), GFI1(+), TATA(+), IRF1(+), IRF1(+), IRF2(-), HNF3B(-), CMYB(+), C/EBP(+), GATA1(-), SRF(+)
GS3 (2979-3138)	C/EBP(+), GATA1(-), SRF(+), LMO2COM(-), CETS1P54(-), CETS1P54(+), XBP1(-), ARNT(-), ATF(-), **CREB(-)**, **AP1FJ(-)**, **CREB(-)**, **ATF2(-)**, TCF11(+), Oct1(-), CAAT(-), GATA1(+), LMO2COM(+), GATA3(+), BRN2(1), NFAT(+), GKLF(+), HNF3B(-), IK1(+), MZF1(+), IK2(+), NFAT(+), SRY(+), STAT(-), NFAT(+), BRN2(-), TCF11(-), **AP1FJ(+)**, **AP1(+)**
GS2 (3137-3293)	C/EBP(-), MEF2(-), AP1FJ(+), RORA1(-), IK2(-), LYF1(-), GATA3(+), Oct1(+), GATA1(-), LMO2COM(-), MYCMAX(+), NMYC(-), USF(+), GATA1(-), LMO2COM(-), GATA1(+), GATA1(+), GATA1(-), LMO2COM(+), LMO2COM(-), GATA3(-), GATA2(-), C/EBP(+), VMYB(+)
GS1 (3293-3437)	NKX25(+), CMYB(-), NFAT(-), IK1(-), IK2(-), STAF(+), LYF1(-), TATA(+), GC(+), RFX1(+), **Sp1(+)**, NF1(-), NKX25(-), CAAT(+), NFY(+), NRF2(-), IK2(+), IK1(+), SRE(-)

6.3 Discussion

Despite the wealth of information regarding the enzymatic properties and tissue-specific distribution of human NAT1 and its murine homologue NAT2, very little is known about the molecular mechanisms regulating expression of their genes *in vivo*. The expression profile of human *NAT1* and murine *Nat2* resembles that of housekeeping genes, such as those for hypoxanthine phosphoribosyl transferase and dihydrofolate reductase (Dynan, 1986). Both human *NAT1* and murine *Nat2* are expressed in the preimplantation embryo (section 5.2.2; Payton *et al.*, 1999b; Smelt *et al.*, 2000), during pre- and early post-natal development (Pacifici *et al.*, 1986; Stanley *et al.*, 1998; Mitchell *at al.* 1999; Estrada *et al.*, 2000), and throughout adulthood (section 5.1). Expression appears to be ubiquitous both in the foetus (Pacifici *et al.*, 1986; Stanley *et al.*, 1998) and the mature organism (sections 5.1 and 5.2.1), suggesting involvement of NAT in basic metabolic processes of the cell, e.g. in folate catabolism, as previously postulated (Minchin, 1995; Ward *et al.*, 1995; Payton *et al.*, 1999b).

Ubiquitous expression may still be subject to regulation by various intra- or extra-cellular stimuli (Dynan, 1986), such as steroid hormones, cell cycle regulators, cAMP, phorbol esters, bacterial lipopolysaccharide (LPS), growth factors, heavy metals and cytokines (La Thangue and Rigby, 1988; Thanos and Maniatis, 1995; Chao and Young, 1996; Latchman, 1998a). Transcriptional regulation is a complex cellular process, involving binding of multiple transcription factors to *cis*-acting sequences linked to the transcribed gene. These sequences can either be constituents of the core promoter or binding sites for various constitutive or inducible regulatory factors (section 1.5.1). Study of the *cis*-acting sequences associated with a gene is an essential first step towards understanding of the *trans*-acting factors mediating transcription initiation and regulation, but requires extensive knowledge of the exon-intron organisation of the gene and the sequence of its surrounding region.

Previous studies on the transcriptional control of the genes for NAT have focused on the region immediately upstream of the coding exon. As mentioned earlier (section 6.1), Estrada-Rodgers *et al.* (1998a) identified several transcription initiation sites, located 112-151bp upstream of the coding exon of the mouse *Nat2* gene, but did not search for promoter elements in that region. Other investigators (Minchin *et al.*,

1998) have mapped the transcription initiation site and core promoter of the human *NAT1* gene 184-213bp and 246-264bp upstream of the coding region, respectively.

Following the work described in sections 5.2.4 and 5.2.6, it has become evident that both human *NAT1* and murine *Nat2* possess NCEs, implying that transcription can start from sites additional to those described above (Estrada-Rodgers *et al.*, 1998a; Minchin *et al.*, 1998). The presence of NCEs also suggests the differential use of more than one promoter. The work presented in this Chapter has involved investigation for transcriptional regulatory elements upstream of the NCE of murine *Nat2* (section 5.2.4). Sequencing of 129/Ola clone A (section 4.2.2.1) has been fundamental to the search for such elements, while of key inıportance has also been the characterisation of the mouse embryonic liver BNL.CL2 cell line, which expresses *Nat2* transcripts initiated from the NCE (section 6.2.1). The exact position of the *Nat2* transcription initiation site was identified by RNase mapping with a ribo-probe specific for the 5′ flanking region of the NCE, after transient transfection of BNL.CL2 cells with 129/Ola clone A DNA (section 6.2.3).

Although it was postulated that transcription of the mouse *Nat2* gene could be regulated by at least two alternative promoters, screening of the entire 129/Ola clone A identified only one region driving high levels of reporter gene expression. This region encompasses the beginning of the *Nat2* NCE (section 6.2.5) and contains a TATA- and a Sp1-box (section 1.5.1). Deletion of both of these elements practically abolished reporter gene expression. This was not simply due to the loss of the *Nat2* transcription initiation site, also contained in the deleted region, since mutation of either the TATA- or the Sp1-box also led to a substantial decline in core promoter activity (section 6.2.7). The identified TATA-box has the less typical TTTAAAA sequence, also found in the promoters of the genes for somatostatin and tyrosine hydroxylase (Tavianini *et al.*, 1984; Montminy, 1997). Although substitution of the adenosine at position 2 of the consensus sequence (TATAAAA) is rare (1% frequency) (Beebee and Burke, 1990), the TATA-box described in the present study is functional, but requires activation by Sp1 for maximum efficiency. Sp1 is known to initiate transcription of many genes lacking a TATA-box (Dynan, 1986; Klein *et al.*, 1998; Zhang *et al.*, 1998; Barski *et al.*, 1999; Geiger *et al.*, 2000), via interaction with initiator or downstream promoter elements (section 1.5.1). However, the Sp1-box

described in the present study has been demonstrated to be capable of driving high levels of reporter gene expression only in the presence of an intact TATA-box (section 6.2.7). Interestingly, the identified Sp1-box has a sequence identical to that found in the promoter of the mouse *Hprt* and *Dhfr* genes (Dynan *et al.*, 1986; Melton *et al.*, 1986), and could serve to maintain constitutive expression of the *Nat2* gene.

Overall, the identified promoter has an unusually compact structure, with the Sp1-box located downstream of the TATA-box and the major transcription initiation site flanked by the two elements (figure 6.2.12). Transcription usually starts from sites located about 30bp downstream of the TATA-box (Müller *et al.*, 1988; Weis and Reinberg, 1992; Pugh, 1996), but this distance can be significantly shortened when the TATA-box has an atypical sequence (Le Van Thai *et al.*, 1993; Kiang *et al.*, 1997; Madiai *et al.*, 1999). Moreover, although Sp1 is generally believed to bind upstream of the TATA-box (Dynan and Tjian, 1985; Latchman, 1998a; 1998b), there are examples of genes with Sp1-boxes downstream of the TATA-box (Mezquita *et al.*, 1993; Duprez *et al.*, 1994; Dong *et al.*, 2000; Ito *et al.*, 2000). It appears that the Sp1-box closest to the transcription start site has the strongest stimulatory effect (La Thangue and Rigby, 1988).

The lack of alternative promoter elements within the 6.1kb intron of murine *Nat2* is remarkable. Based on computational analysis, Estrada-Rodgers *et al.* (1998a) suggested a number of TATA-boxes as putative promoter elements in the vicinity of the *Nat2* coding region, but did not verify their predictions experimentally. These TATA-boxes were also identified by TRANSFAC analysis in the present study. Most of them show a weak sequence homology to the consensus or have an orientation opposite to that of the transcribed *Nat2* gene. Moreover, the more sophisticated NIX analysis (section 4.2.2.2) did not predict a core promoter in the proposed region. Contrary to the TRANSFAC analysis, which looks for individual matches between the inspected sequence and various consensus sequences deposited in the database, NIX analysis uses complex algorithms to evaluate the features of the entire analysed region (www.hgmp.mrc.ac.uk/NIX/). This may involve screening for *cis*-acting sequences (e.g. CCAAT-, Sp1- or TATA-box), clustered close to a putative Inr or the beginning of a predicted exon. For example, although NIX analysis provided only one good prediction of a promoter acting in the orientation of the *Nat2* gene (section

4.2.2.2), TRANSFAC analysis of 129/Ola clone A identified 116 TATA-boxes. However, only one of these boxes is associated with a Sp1-box, to form the functional promoter identified both experimentally (section 6.2.7) and by NIX analysis (section 4.2.2.2). The entire 129/Ola clone A contains three additional Sp1-boxes, only one of which shows a perfect match to the consensus. Interestingly, this box is located within the NCE, only 72bp downstream of the Sp1-box of the core promoter.

Although it is unlikely that any of the elements predicted by Estrada-Rodgers *et al.* (1998a) could drive expression of the mouse *Nat2* gene, the presence of a functional promoter proximal to the coding region has been experimentally confirmed for the human *NAT1* gene (Minchin *et al.*, 1998). This consists of a typical AP1-box, flanked by two ATCATTT motifs. The AP1-box has been demonstrated to bind a protein complex containing c-Fos, but not c-Jun, and is essential for core promoter activity together with the 3′ flanking repeat. The 5′ flanking repeat is not required for activity, but its deletion rendered the promoter responsive to treatment with phorbol esters. Although not perfectly conserved, these elements are present in the mouse and are conveniently positioned about 100bp (position 9580-9598 of 129/Ola clone A; Appendix 2) upstream of the transcription initiation sites identified by Estrada-Rodgers *et al.* (1998a). However, they are not active in BNL.CL2 cells, unless a specific stimulus is required for induction. It is also possible that the BNL.CL2 cells lack the necessary activating factors or express a repressor that blocks transcription from this putative promoter. Using different cell lines may reveal alternative promoter elements driving expression of the mouse *Nat2* gene in a cell type-specific manner.

The pattern of reporter gene expression generated by the 5′-nested deletion constructs (D1-D8), spanning the region upstream of the mouse *Nat2* NCE, suggested the presence of negative regulatory elements in the vicinity of the *Nat2* promoter (section 6.2.8). The low constitutive expression of many housekeeping genes is directed by TATA-less promoters possessing one or more Sp1-boxes (Bienz-Tadmor *et al.*, 1985; Dynan, 1986; Müller *et al.*, 1988; Weis and Reinberg, 1992; Klein *et al.*, 1998; Barski *et al.*, 1999). Murine *Nat2* has been postulated to play a role in the endogenous metabolism of folate (Payton *et al.*, 1999b) and may have the characteristics of a housekeeping gene. Any negative upstream regulatory elements might, therefore, maintain a low level of *Nat2* expression by reducing the strength of

its TATA-containing core promoter. This negative regulation could be cancelled following an appropriate stimulus. Surprisingly, portions of the same region cloned in pGL3-Promoter vector (constructs P1-P6) appeared to positively regulate reporter gene expression (section 6.2.6). In this case, however, the effect of the analysed fragments was assessed in relation with the heterologous SV40 promoter of the reporter vector, rather than the native *Nat2* promoter, as in the case of constructs D1-D8. Moreover, the inserts of constructs P1-P6 had limited overlap, contrary to those of constructs D1-D8 which shared the same 3′ end. Fragmentation of the region upstream of the *Nat2* NCE in constructs P1-P6 may have a) interrupted important interactions between the promoter and upstream repressor elements, masking their postulated negative effect, or b) revealed the positive effect of putative upstream enhancer elements by removing the repressor and/or placing them close to the SV40 promoter of the vector (Hatzopoulos *et al.*, 1988). It is also possible that, in their natural genomic position, the putative enhancer elements of constructs P1-P6 direct their positive action towards the upstream *Nat1*, rather than the downstream *Nat2* promoter, since c*is*-acting sequences can function in a position- and orientation-independent fashion (Hatzopoulos *et al.*, 1988; Müller *et al.*, 1988; Latchman, 1998a; 1998b).

The dual-luciferase reporter assay is a very powerful technique for detecting promoter activity, as in sections 6.2.5 and 6.2.7. However, when using the technique to evaluate the effect of different regulatory elements (enhancers or repressors) on the activity of a specific promoter, one has to bear in mind a number of technical problems that may affect accuracy and reproducibility of the results. One of the major drawbacks is variability in the transfection efficiency. Cells may grow at different rates on similarly processed plates, leading to variable levels of reporter gene expression. Measurement of the *Renilla* luciferase activity, used as an internal control, should help overcome this problem. However, depending on the strength of the regulatory elements carried by the pGL3 vector, the promoter of the pRL-TK vector may become subject to *trans*-activation in co-transfection experiments (Promega's technical manual #TM040 on dual-luciferase reporter assay). This prevents neutral constitutive expression of the *Renilla* luciferase gene, diminishing its value as an internal control. When performing the experiments described in sections 6.2.6 and 6.2.8, all possible care was taken to avoid such problems. For example, cell cultures

were monitored, to ensure uniform growth rates before and after transfection. Large variations in pRL-TK activity between co-transfections were also avoided, as much as possible. However, it may be of use to perform additional quantitative assays, in order to pin-point the regions harbouring transcription regulatory elements for *Nat2*.

Study of the *cis*-acting sequences associated with a gene is the first step towards identification of the *trans*-acting factors regulating its transcription. While certain transcription factors are constitutively active in all cell types, others can only function in response to a specific stimulus (Latchman, 1998a). The use of EMSA (section 6.2.9) demonstrated specific protein binding to the promoter region of the mouse *Nat2* gene (probe GS1; figure 6.2.20a). Although further work is required towards characterisation of these proteins, it is reasonable to assume that they must target specific elements of the core promoter. In support of this hypothesis, competition by a Sp1-specific oligonucleotide interfered with protein binding to the GS1 probe, confirming the presence of an active Sp1 binding site in the region. This is consistent with the results of the reporter gene assays, described above.

Probes GS3 and GS5, extending about 600bp upstream of the promoter region, were also capable of binding proteins, but the strength and specificity of interaction was significantly lower (figure 6.2.20b and c). The partial competition observed with the non-specific NSC probe may imply either non-specific binding of cellular proteins to the GS3 and GS5 probes or presence of specific binding sites on the NSC probe. The latter was designed to match a coding sequence (part of the *Nat2* coding region), so it would not be expected to bind transcription factors. However, the majority of transcription factors have short target sequences, which often diverge from the consensus without obvious consequences (Johnson and McKnight, 1989). TRANSFAC analysis identified a number of putative binding sites shared between NSC, GS3 and GS5 probes, including sites for AP1, GATA2, GATA3 and C/EBP factors. Competition between the specific and non-specific probe for binding of these factors could explain the observed decrease in the intensity of the shifted bands.

The use of various oligonucleotides as competitors of probes GS3 and GS5 in EMSAs suggested binding of AP1 and CREB to more than one site in the region upstream of the *Nat2* promoter. In support of these observations, TRANSFAC analysis identified 7 putative AP1 and 4 putative CREB sites in the region covered by

the two probes. One of the CREB sites (CRE) is also specific for factor ATF2, which, though a member of the AP1 family, binds preferentially to CREs (Karin *et al.*, 1997). The AP1 family includes various isoforms of Jun, Fos and ATF proteins. They are all bZIP proteins, capable of forming homodimers or heterodimers which bind to the AP1 site. Fos cannot directly bind to DNA, but forms heterodimers with Jun. These are more stable and have higher affinity for the AP1 site than the Jun-Jun homodimers. Phosphorylation of c-Jun at serines 63 and 73 further increases stability of the complex and enhances its ability to activate transcription. Transient induction of c-Fos and c-Jun occurs very rapidly in response to a variety of stimuli, including growth factors, cytokines and phorbol esters. The AP1 complexes have been implicated in many biological processes, including cell proliferation, differentiation and apoptosis. Despite their various abnormalities, knockout mice deficient in c-Fos can survive, while c-Jun knockouts die *in utero*. On the other hand, simultaneous overexpression of c-Fos and c-Jun leads to tumourigenesis (Lee *et al.*, 1987a; 1987b; Rauscher *et al.*, 1988; Li *et al.*, 1992; Kerppola and Curran, 1995; Karin *et al.*, 1997).

CREB is a bZIP transcription factor binding to the palindromic cAMP response element (CRE). Upon stimulation by hormones or growth factors, the levels of second messanger cAMP rise, allowing activation of protein kinase A, which in turn phosphorylates CREB at serine 133. The activated protein binds to CREs as a dimer, stimulating transcription of numerous target genes. Phosphorylated CREB is known to interact with the CREB-binding protein (CBP), a zinc-finger protein with a C-terminal glutamine-rich domain. CBP is believed to act as a bridge between the CREB complex and the TFIIB factor of the core transcriptional apparatus. Interestingly, CREB is competing with AP1-complexes for binding to both the CBP and the CRE. Other transcription factors (e.g. nuclear receptors, NF-κB, STATs and SREBPs) also interact with CBP, presumably leading to an antagonism between different regulatory pathways. Cross-binding of AP1 factors to CRE may provide an alternative explanation for the observed competition of GS3 and GS5 probes by both the AP1- and CREB-specific oligonucleotide probes used in EMSAs (section 6.2.9) (Rauscher *et al.*, 1988; Roesler *et al.*, 1988; Karin *et al.*, 1997; Montminy, 1997; Shikama *et al.*, 1997; Latchman, 1998a; Sassone-Corsi, 1998).

organs	C/EBPβ[1]	C/EBPβ[2]	c-Jun[3]	CREB[4]	CREB[5]	ATF2[6]	E47[7]	GATA1[8]	GATA2[9]	GATA3[10]	MEF2[11]	MEF2[12]	MZF1[13]	Oct1[14]	Oct6[15]	Sp1[16]	TBP[17]	XBP1[18]	STAT4[19]	STAT5[20]	δCREB[21]	Arnt[22]	SREBP1[23]	Th1[24]	RFX1[25]
B-cells															+										
T-cells										+															
adrenal gland						+																	+		
autonomic ganglia																								+	
bone marrow																			+	+					
brain				+	+	+					+	+											+		
cardiac muscle												+													
erythroid cells								+																	
fetal brain										+													+		
fetal heart																							+		
fetal kidney																							+		
fetal liver																							+		
fetal lung																							+		
gut				+																				+	
heart				+																+			+	+	
intestine																								+	
kidney				+								+								+			+		
kidney after LPS induction		+																							
mammary gland																				+					
liver		+		+																		+	+		
lung						+						+							+	+			+	+	
mast cells								+																	
megacaryo-cytic cells								+																	
meninges						+																			
muscle						+																	+		
myeloid cells													+												
myogenic cells											+														
pancreas																							+		
pharyngeal arches																								+	
placenta										+		+											+	+	
skeletal muscle												+							+						
smooth muscle												+													
spinal cord						+																			
spleen		+		+		+													+	+					
stomach						+																			
testis				+															+						+
thalamus						+																			
thymus				+		+														+					
ubiquitous	+		+		+		+							+		+	+	+			+				
widespread		+							+																

[1-25]TRANSFAC accession nos.:[1]T00017; [2]T00581; [3]T00131 & T00133; [4]T00163; [5]T00989; [6]T00167; [7]T00207; [8]T00305 & T00306; [9]T00308; [10]T00310 & T00311; [11]T00505; [12]T01005; [13]T00529; [14]T00641 & T00644; [15]T00655; [16]T00759; [17]T00796; [18]T00902; [19]T01576; [20]T00944 & T01579; [21]T01311; [22]T01346; [23]T01556, T01557 & T01558; [24]T01569; [25]T01666 & T01673.

The data generated by TRANSFAC analysis of the genomic region around the *Nat2* promoter can serve as the basis for prediction of the tissue- or developmental-specific expression of the *Nat2* gene. Table 6.3.1 provides information about the expression pattern of transcription factors likely to control activity of the *Nat2* core promoter. The combination of predicted factors suggests ubiquitous expression of the *Nat2* gene, consistent with the experimental results of several studies (sections 5.1, 5.2.1 and 5.2.2). Ubiquitously expressed factors, such as TBP, Sp1 and C/EBPβ (sections 1.5.1) are specific for most eukaryotic promoters. The CCAAT-box binding factor C/EBPβ is of special interest, because it is most abundant in the liver and is known to activate liver-specific expression of several genes, including a group of LPS-inducible acute-phase response genes (Akira *et al.*, 1990). Two C/EBPβ binding sites within 68 and 85bp from the TATA-box of the *Nat2* promoter could, therefore, play a role in maintaining high levels of hepatic *Nat2* expression. Of particular importance is also Oct1, a POU protein involved in cellular transcription and DNA replication (O'Neil *et al.*, 1988). Oct1 binds to the characteristic "octamer motif" and is constitutively expressed in many cell types, unlike other family members (e.g. Oct2 and Oct6), which are B-cell-specific (Bendall *et al.*, 1997; TRANSFAC accession no. T00655).

Table 6.3.1: Expression pattern of transcription factors likely to regulate activity of the mouse *Nat2* core promoter. The results of TRANSFAC analysis, performed for the 735bp region covered by EMSA probes GS1-GS5 (section 6.2.9), were further subjected to CYTOMER analysis (http://transfac.gbf.de/CYTOMER/). The CYTOMER database contains information about the tissue- and developmental-specific expression of mammalian transcription factors, allowing deduction of the expression pattern of a gene, based on the combination of its putative regulators. The "+" mark denotes expression of a transcription factor in a particular tissue. Isoforms of a transcription factor with distinct expression patterns are presented in different columns, while those with the same expression pattern are presented in a single column. The accession numbers for all predicted factors are provided in the footnote.

Of special interest are the properties of some of the tissue-specific regulatory factors predicted to bind to elements close to the *Nat2* promoter (table 6.3.1). The STAT proteins comprise a large family of transcription factors regulating gene expression in response to various cytokines. Upon stimulation, STATs become phosphorylated by tyrosine kinases, then dimerise and enter the nucleus, where they bind to DNA (Ihle and Kerr, 1995; Darnell, 1997; Horvath and Darnell, 1997). The SREBP proteins mediate sterol-regulated gene expression. In unstimulated cells, they stay bound to the membranes of the endocytoplasmic reticulum. Upon stimulation, the N-terminal part of the molecule is released by proteolytic cleavage and moves to the nucleus, where it acts as a bHLH transcription factor (Pahl and Baeuerle, 1996; Brown and Goldstein, 1997). Members of the GATA family were predicted to bind to at least 10 sites in the analysed region. GATA1 is a zinc finger protein essential for erythropoiesis. It is also present in megakaryocytes and other types of blood cells. Knockout mice deficient in GATA1 do not survive, due to failure in erythroid differentiation, but targeted inactivation in megakaryocytes is partly compensated for by the GATA2 isoform, which is widely expressed. GATA3 is T-cell specific (Shivdasani *et al.*, 1997; TRANSFAC accession nos. T00310 and T00311). Some of the above-mentioned transcription factors, namely AP1, Oct1 and GATA1, are known to act synergistically with Sp1 (Lee *et al.*, 1987a; TRANSFAC accession nos. T00305 and T00641) and could, therefore, directly interact with the *Nat2* promoter.

NAT2 activity has been demonstrated to vary among tissues of the adult mouse, with levels being highest in the liver, thymus, spleen and white blood cells and lowest in the brain, heart, muscle, bone marrow and oesophagus (Chung *et al.*, 1993; Payton *et al.*, 1999b). Similar variations have been observed in the amount of immunohistochemically detectable mouse NAT2 protein, among cell types of the adult liver, spleen and kidney (Stanley *et al.*, 1997). These variations could be the result of transcriptional regulation of the *Nat2* gene in a tissue-specific manner. Interestingly, the brain and heart, which provided the lowest NAT2 activity in the adult mouse (Chung *et al.*, 1993; Payton *et al.*, 1999b) exhibited a most pronounced immunohistochemical staining in the developing embryo (Stanley *et al.*, 1998). This is in consistence with the proposed role of murine NAT2 in folate metabolism (Payton *et al.*, 1999b). The quick folate turnover observed during early gestation is believed to be crucial for DNA synthesis in the rapidly proliferating cells of the developing

embryonic tissues (Scott *et al.*, 1993; McNulty *et al.*, 1993; 1995). It has been postulated that retarded cell proliferation, most likely due to an imbalance in folate metabolism, is the major cause for neural tube and heart malformations (Lucock, 2000; Barber *et al.*, 1999). In these developing tissues, high levels of NAT2 expression might be important for maintaining the balance between folate intake and excretion. On the other hand, there is no need for high level of NAT2 expression in the non-proliferating neural and cardiac tissues of the mature organism, where the role of folate is less important. Understanding the mechanisms regulating expression of the genes for mammalian NATs will help unravel the biological role of each isoenzyme.

CHAPTER 7

Conclusions and future work

Epidemiological research has long been considered as an effective way of identifying risks associated with exposure to synthetic chemicals (Vainio and McGregor, 1992). NAT has been in the focus of such research for almost 50 years, since it was first realised that individuals differ in their capacity to N-acetylate xenobiotics, including commonly used drugs and established carcinogens (Grant *et al.*, 2000). Epidemiological evidence has pointed to an association between the slow NAT2 phenotype and increased risk for occupational or smoking-related bladder cancer (Marcus *et al.*, 2000a). However, other studies investigating the possible link between NAT and different types of cancer have provided inconclusive and often conflicting results (Hein *et al.*, 2000a). The statistical power of these studies may be limited by the small size of the analysed group of affected individuals (Taioli, 1999). Moreover, predisposition to complex diseases can be influenced by factors, such as age, sex and lifestyle (Levy *et al.*, 1996). When involvement of a specific xenobiotic metabolising enzyme is investigated, the power of population studies is further limited by differences in individual levels of exposure and the concurrent action of other polymorphic enzymes of xenobiotic metabolism (Vainio and McGregor, 1992; Levy *et al.*, 1996). Large-scale case-control studies (Taioli, 1999; Wikman *et al.*, 2001) and meta-analyses of published results (Marcus *et al*, 2000a; 2000b) may provide more conclusive information about the possible links between NAT and disease.

Animal models can be used to overcome many of the problems associated with epidemiological research (Vainio and McGregor, 1992; Levy *et al.*, 1996). The laboratory mouse, an already popular model in pharmacological and toxicological studies, is becoming increasingly valuable with the advancement of the Mouse Genome Project (Moore, 1999; West *et al.*, 2000). Mouse inbred and congenic strains can be used to evaluate the effect of xenobiotics on living organisms of defined genetic background and under controlled experimental conditions, without the ethical

concerns associated with the use of human subjects (Levy *et al.*, 1996). Mouse embryos can also be accessed at any stage of gestation, allowing investigation of gene involvement during development (Copp, 1995).

The enzymatic properties and substrate specificities of murine NAT1 and NAT2 isoenzymes have been thoroughly studied (Martell *et al.*, 1991; 1992; Kelly and Sim, 1994). Murine NAT1 is similar to human NAT2 and is likely to N-acetylate only xenobiotic substrates, while murine NAT2 is functionally equivalent to human NAT1 and may be involved in endogenous metabolism (Sim *et al.*, 2000). Based on these similarities, the mouse has been established as a suitable model for studying human NAT, but its use has been limited by the scarcity of genetic polymorphism and the lack of sufficient knowledge about expression of the murine *Nat* genes *in vivo*.

The aim of the work presented in this thesis has been to develop improved mouse models for NAT, either in the form of newly characterised strains with polymorphisms in the *Nat* genes or in the form of genetically engineered strains lacking or overexpressing NAT. To allow comparison with their human homologues, the mouse *Nat* genes were analysed for their chromosomal localisation and tissue-specific expression profile. The genomic structure and control of expression of murine *Nat2* was also studied. The following sections will summarise the major findings of this work and provide the framework of related current and future research.

7.1 Mouse strains polymorphic for NAT

Previously characterised fast and slow acetylating inbred strains of the mouse, rat and Syrian hamster carry polymorphisms only at the *Nat2* locus (Hein *et al.*, 1997). Therefore, existing rodent models for NAT are useful only for investigating the effects of variation in human NAT1 activity. The identification of novel functional polymorphism in the *Nat1* gene of the wild-derived inbred strain *M. spretus* (Chapter 3) establishes the mouse as the sole, currently available, animal model mimicking the human NAT2 polymorphism. Human NAT2 and murine NAT1 isoenzymes specialise in the metabolism of arylamine and hydrazine drugs (Martell *et al.*, 1992; Vatsis and Weber, 1994; Ware and Svensson, 1996). Therefore, the slow acetylating *M. spretus* strain will be useful for assessing the efficacy and/or toxicity of drugs undergoing N-acetylation *in vivo*.

M. spretus is a slow acetylating strain with both NAT1- and NAT2-specific substrates (section 3.2.7), as it carries mutations in both the *Nat1* and *Nat2* gene (section 3.2.2). Established arylamine carcinogens, such as 2-AF, are N-acetylated by both NAT isoenzymes in mice and humans (Vatsis and Weber, 1994; Martell *et al.*, 1992; Kelly and Sim, 1994). However, most studies to date (sections 1.2.4 and 1.3.3) have examined only the human/mouse NAT2 polymorphism as a susceptibility factor for chemical-induced cancer. In humans, it has been difficult to assess the combined effect of polymorphisms at both *NAT* loci, mainly due to the lack of a clear *NAT1* genotype-phenotype correlation (section 1.2.3.1). In mice, such studies require strains with mutations in both *Nat1* and *Nat2* genes. *M. spretus* could, therefore, serve as a model for studying the links between reduced acetylation capacity and cancer risk.

Genetic mapping has linked the slow NAT2 phenotype to teratogen-induced orofacial clefting in A/J mice (Karolyi *et al.*, 1987; 1988; 1990). It might be of interest to investigate whether *M. spretus*, also a slow acetylator for NAT2, shows increased susceptibility to the same developmental defects. This would implicate NAT2 in early developmental processes, possibly via its postulated role in folate catabolism. On the other hand, a lack of association would imply the presence of another nearby locus responsible for the high predisposition of A/J mice to chemical teratogenesis.

7.2 Mouse knock-out and transgenic strains for NAT

Although slow acetylating mouse strains, such as the A/J and *M. spretus*, can be useful for evaluating the toxic effects of xenobiotics on the early embryo, the most effective way of studying the role of NAT in development is the production of NAT2-deficient mouse strains. A collaborative effort, part of which was described in sections 4.2.1 and 4.2.2, has led to the production of *Nat2* knock-out mice, currently being used to produce congenic lines for the "*null*" allele (Cornish *et al.*, 2001). It appears that *Nat2* deficiency *per se* does not compromise survival or reproductive fitness of the offspring, as no major phenotypic anomalies were observed. However, a detailed histological analysis may reveal subtle differences between the knock-out and wild-type mice, that will help elucidate the endogenous role of murine NAT2.

The *Nat2*-null mice will also be useful in pharmacological and toxicological studies, as they are likely to show different sensitivities to drugs, carcinogens or teratogens, compared with the wild-type controls. In order to avoid gene complementarity (Gonzalez, 1998), it may be necessary to knock out all three *Nat* genes on a single construct. A previously identified 130kb PAC clone, containing all three murine *Nat* genes (Fakis *et al.*, 2000), could be useful for this purpose.

The targeting construct carrying the inactivated *Nat2* gene (section 4.2.1.2) was also used to generate mice with highly active copies of human *NAT1* (section 4.3). These mice have severe abnormalities and rarely survive throughout gestation (Pinter *et al.*, 2001), suggesting that NAT overexpression has deleterious effects on the developing embryo. Detailed phenotypic analysis of the transgenic mice at various stages of gestation is required in order to unravel the role of NAT in development. The generation of inducible transgenic mice, overexpressing NAT in a spatially- or temporally-controlled manner, will further allow evaluation of the impact that excessive acetylation may have on carcinogen bioactivation.

7.3 Localisation and genomic structure of the mouse *Nat* genes

Previous studies (Mattano *et al.*, 1988; Fakis *et al.*, 2000) have determined the cytogenetic and approximate genetic position of the mouse *Nat* genes. However, their physical map location on chromosome 8 remains unknown. The present study has identified three 1.1cM genomic intervals likely to contain the entire *Nat* gene cluster (section 4.2.3). These intervals cover part of the 28.2-33.8cM region, where *Nat2* has previously been localised by linkage analysis (Mattano *et al.*, 1988), and are located between established WI/MIT YAC contigs. Accurate mapping of the *Nat* gene cluster will facilitate comparison with the syntenic 8p22 chromosomal region, bearing the *NAT* genes in humans (Hickman *et al.*, 1994; Matas *et al.*, 1997). It is also likely that the *Nat* genes are part of the same syntenic group not only in mice and humans, but in other mammalian species too. Data generated from the recently launched Rat Genome Project (Twigger *et al.*, 2002) will allow interesting comparisons. A comparative approach may help pinpoint the position of other loci in the neighbourhood of *Nat*, including genes previously implicated in chemical teratogenesis and carcinogenesis (Farrington *et al.*, 1996; Diehl and Erickson, 1997; Van Alewijk *et al.*, 1999). Genetic

markers, such as the SNPs (section 3.2.2) and SSLPs (section 4.2.4) described in the present study, may facilitate such investigations.

The input of the Mouse Genome Project towards sequencing of the genomic region carrying the *Nat* genes is currently very limited (section 4.2.2.2). This is not surprising, since the *Nat* region is not included in the YAC-based physical map of murine chromosome 8 (section 4.2.3.1). Sequencing analysis of 129/Ola clone A (section 4.2.2.1) has uncovered interesting elements in the vicinity of the *Nat2* gene (section 4.2.2.2). A region about 6kb upstream of the single coding exon of *Nat2* was demonstrated by RT-PCR to form an upstream non-coding exon (section 5.2.4). Human *NAT2* also has an upstream NCE (Ohsako and Deguchi, 1990; section 5.2.6), while comparison of ESTs with the draft sequence of human chromosome 8 has further revealed three NCEs for the human *NAT1* gene (section 5.2.6). Based on their exon-intron structure similarity, murine *Nat2* is likely to be the genetic orthologue of human *NAT1*, an assumption supported by the functional identity of the corresponding protein products. The structure of the genes for mammalian NAT, where an intronless coding region is separated from one or more upstream NCE by large introns, is unusual and may imply operation of complex regulatory mechanisms. NAT-encoding transcripts are likely to be initiated from more than one transcription initiation site or be subject to alternative splicing and/or polyadenylation. Future work will involve characterisation and quantification of these transcripts in normal and diseased tissue, using standard and "real-time" RT-PCR techniques. The use of mouse cDNA libraries and Rapid Amplification of cDNA Ends (RACE) methodologies will further allow isolation of full-length transcripts encoding for all three murine *Nat* genes.

7.4 Expression profile and transcriptional regulation of the mouse *Nat* genes

Expression of the murine *Nat* genes was investigated at the level of transcription by RT-PCR. *Nat1* transcript was detected only in the liver (section 5.2.1), implying a role for its protein product restricted to xenobiotic metabolism. In contrast, *Nat2* expression was evident in all of the analysed tissues (section 5.2.1), in agreement with previous studies employing protein detection methods (Chung *et al.*, 1993; Stanley *et al.*, 1997). *Nat2* was confirmed to be the only murine *Nat* gene

expressed in ES cells, suggesting that it could be important for early embryogenesis. Similarly, only *NAT1* transcript was detected in human blastocysts (section 5.2.2).

Genes, such as murine *Nat2* and human *NAT1*, activated at very early stages of development and maintaining their ubiquitous expression throughout life, are likely to serve fundamental functions (Latchman, 1998b). To better understand transcription of murine *Nat2*, the core promoter and transcription initiation site of the gene were accurately mapped (Chapter 6). The *Nat2* core promoter has a functional TATA-box and a Sp1-box, which is often associated with housekeeping genes and is known to drive constitutive expression (Dynan, 1986; Latchman, 1998a). As demonstrated by EMSA, the promoter region of the *Nat2* gene is capable of specific protein binding (section 6.2.9). Incubation with pure TBP and Sp1 proteins should confirm interaction of these two factors with the *Nat2* core promoter. Further characterisation of *Nat2* regulatory elements will require EMSAs with factors (e.g. AP1 and CREB) predicted to bind to the region upstream of the core promoter (section 6.2.9), as well as detailed search for protein-binding sites by DNaseI footprinting. The nature of the factors binding to the identified sequences can be determined by TRANSFAC analysis and confirmed by EMSA (Latchman, 1995). Reporter gene assays, DNaseI footprinting and EMSA can then be used to investigate the effect of mutations introduced to the identified elements. It may also be worthwhile to examine whether any of the repetitive sequences around *Nat2* (section 4.2.2.2) might serve regulatory functions. Understanding of the *in vivo* regulation of the *Nat2* promoter will further require analysis of the interactions between candidate *cis*-elements and specific *trans*-acting factors in the context of chromosome structure and nuclear organisation, using *in vivo* DNA footprinting by ligation-mediated PCR, or chromatin immunoprecipitation followed by PCR or Southern blotting (Latchman, 1995; Orlando 2000).

Although the present study has identified only one active promoter for murine *Nat2*, located upstream of the NCE (section 6.2.5), an alternative promoter could drive transcription from a site adjacent to the coding region (Estrada-Rodgers *et al.*, 1998a). Applying the methodologies described in Chapter 6 to different cell types may provide additional information about the mechanisms underlying transcriptional regulation of murine *Nat2*. Similar studies could also be performed for murine *Nat1* and *Nat3*, as well as for NAT-encoding genes of other organisms.

7.5 The role of murine NAT3

Although murine *Nat3* was cloned shortly after the *Nat1* and *Nat2* genes (Kelly and Sim, 1992; 1994), the properties and biological role of its protein product remain poorly understood. Recombinant NAT3 protein, expressed from the *Nat3*1* allele of common inbred mouse strains (e.g. Balb/c, A/J, C57Bl/6J), provides only marginal activity towards common NAT substrates (section 3.2.7; Kelly and Sim, 1994; Fretland *et al.*, 1997; Estrada-Rodgers *et al.*, 1998b). Similar results were also obtained after recombinant expression of the newly identified *Nat3*2* and *Nat3*3* alleles from the MCA and MSP strains, respectively (section 3.2.7). None of the mouse strains analysed so far appears to carry frameshift or nonsense mutations in the coding region of the *Nat3* gene (section 3.2.2). Recombinant NAT3 protein has been detected in the soluble fraction of bacterial cell lysates by SDS-PAGE (Kelly and Sim, 1994) and Western blot analysis (Ian Mills, PartII Thesis, Oxford 1996), indicating that the *Nat3* gene is functional in recombinant expression systems. In contrast, results here showed *Nat3* expression to be very restricted *in vivo*; *Nat3* transcript was found in the spleen by RT-PCR, although no NAT3 protein was detected by Western blotting (section 5.2.1). It is possible that NAT3 may be produced in very low amounts in the spleen, where it could serve some unknown function. Assaying for recombinant NAT3 activity against a library of compounds may open the way towards identification of putative endogenous or xenobiotic substrates for the isoenzyme. Future work may also involve detailed analysis of the expression profile of the *Nat3* gene, using suitable cDNA libraries. Sequencing of the region flanking the *Nat3* gene will further allow search for transcriptional regulatory elements. Murine NAT3 may have evolved to perform enzymatic functions distinct to those of other NAT isoenzymes and this is a possibility worth investigating.

Xenobiotic metabolising enzymes, including NATs, are anticipated to become an area of intensive research in the growing field of pharmacogenomics. Pharmacological studies will continue to employ animal experimentation and the mouse is poised to become the premier model of choice following completion of the Mouse Genome Project. Understanding of the molecular mechanisms regulating tissue- and developmental-specific expression of the murine *Nat* genes will help resolve their biological role and identify potential links with disease.

REFERENCES

- Abe M, Deguchi T, Suzuki T (1993). The structure and characterisation of a fourth allele of polymorphic N-acetyltransferase gene found in the Japanese population. *Biochemical and Biophysical Research Communications* **191:** 811-816.
- Abu-Zeid M, Nagata K, Miyata M, Ozawa S, Fukuhara M, Yamazoe Y, Kato R (1991). An arylamine acetyltransferase (AT-1) from Syrian golden hamster liver: Cloning, complete nucleotide sequence and expression in mammalian cells. *Molecular Carcinogenesis* **4:** 81-88.
- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE *et al.* (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185-2195.
- Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H *et al.* (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651-1656.
- Adjaye J, Bolton V, Monk M (1999). Developmental expression of specific genes detected in high-quality cDNA libraries from single human preimplantation embryos. *Gene* **237:** 373-383.
- Adjaye J, Daniels R, Bolton V, Monk M (1997). cDNA libraries from single human preimplantation embryos. *Genomics* **46:** 337-344.
- Adjaye J, Daniels R, Monk M (1998). The construction of cDNA libraries from human single preimplantation embryos and their use in the study of gene expression during development. *Journal of Assisted Reproduction and Genetics* **15:** 344-348.
- Agundez JAG, Jimenez-Jimenez FJ, Luengo A, Molina JA, Orti-Pareja M, Vazquez A, Ramos F *et al.* (1998a). Slow allotypic variants of the *NAT2* gene and susceptibility to early-onset Parkinson's disease. *Neurology* **51**: 1587-1592.
- Agundez JAG, Ladero JM, Olivera M, Abildua R, Roman JM, Benitez J (1995). Genetic analysis of the arylamine N-acetyltransferase polymorphism in breast cancer patients. *Oncology* **52:** 7-11.
- Agundez JAG, Martinez C, Olivera M, Gallardo L, Ladero JM, Rosado C, Prados J *et al.* (1998b). Expression in human prostate of drug- and carcinogen-metabolising enzymes: Association with prostate cancer risk. *British Journal of Cancer* **78:** 1361-1367.
- Agundez JAG, Martinez C, Olivera M, Ledesma MC, Ladero JM, Benitez J (1994). Molecular analysis of the arylamine N-acetyltransferase polymorphism in a Spanish population. *Clinical Pharmacology and Therapeutics* **56:** 202-209.
- Agundez JAG, Menaya JG, Tejeda R, Lago F, Chavez M, Benitez J (1996a). Genetic analysis of the *NAT2* and *CYP2D6* polymorphisms in white patients with non-insulin-dependent diabetes mellitus. *Pharmacogenetics* **6:** 465-472.

- Agundez JAG, Olivera M, Ladero JM, Rodriguez-Lescure A, Ledesma MC, Diaz-Rubio M, Meyer UA *et al.* (1996b). Increased risk for hepatocellular carcinoma in *NAT2*-slow acetylators and *CYP2D6*-rapid metabolisers. *Pharmacogenetics* **6:** 501-512.
- Agundez JAG, Olivera M, Martinez C, Ladero JM, Benitez J (1996c). Identification and prevalence study of 17 allelic variants of the human *NAT2* gene in a white population. *Pharmacogenetics* **6:** 423-428.
- Akira S, Isshiki H, Sugita T, Tanabe O, Kinoshita S, Nishio Y, Nakajima T, Hirano T, Kishimoto T (1990). A nuclear factor for IL-6 expression (NF-IL6) is a member of the C/EBP family. *The EMBO Journal* **9:** 1897-1906.
- Akusjärvi G, Mathews MB, Andersson P, Vennström B, Pettersson U (1980). Structure of genes for virus-associated RNAI and RNAII of adenovirus typse 2. *Proceedings of the National Academy of Sciences of the USA* **77:** 2424-2428.
- Ambrosone CB, Freudenheim JL, Graham S, Marshall JR, Vena JE, Brasure JR, Michalek AM *et al.* (1996). Cigarette smoking, N-acetyltransferase 2 genetic polymorphisms and breast cancer risk. *Journal of the American Medical Association* **276:** 1494-1501.
- Ambrosone CB, Freudenheim JL, Sinha R, Graham S, Marshal JR, Vena JE, Laughlin R *et al.* (1998). Breast cancer risk, meat consumption and N-acetyltransferase (*NAT2*) genetic polymorphisms. *International Journal of Cancer* **75:** 825-830.
- Andreadis A, Gallego ME, Nadal-Ginard B (1987). Generation of protein isoform diversity by alternative splicing; Mechanisms and biological implications. *Annual Reviews of Cell Biology* **3:** 207-242.
- Andres HH, Klem AJ, Szabo SM, Weber WW (1985). New spectrophotometric and radiochemical assays for acetyl-CoA:arylamine N-acetyltransferase, applicable to a variety of arylamines. *Analytical Biochemistry* **145:** 367-375.
- Andres HH, Kolb HJ, Schreiber RJ, Weiss L (1983a). Characterisation of the active site, substrate specificity and kinetic properties of acetyl-CoA:arylamine N-acetyltransferase from pigeon liver. *Biochimica et Biophysica Acta* **746:** 193-201.
- Andres HH, Kolb HJ, Weiss L (1983b). Purification and physical-chemical properties of acetyl-CoA:arylamine N-acetyltransferase from pigeon liver. *Biochemica and Biophysica Acta* **746:** 182-192.
- Andres HH, Vogel RS, Tarr GE, Johnson L, Weber WW (1987). Purification, physicochemical and kinetic properties of liver acetyl-CoA:arylamine N-acetyltransferase from rapid acetylator rabbits. *Molecular Pharmacology* **31:** 446-456.
- Arbieva ZH, Banerjee K, Kim SY, Edassery SL, Maniatis VS, Horrigan SK, Westbrook CA (2000). High-resolution physical map and transcript identification of a prostate cancer deletion interval on 8p22. *Genome Research* **10:** 244-257.
- Avner P (1998). Complex traits and polygenic inheritance in the mouse. *Methods: A Companion to Methods in Enzymology* **14:** 191-198.
- Baeuerle PA, Baltimore D (1996). NF-κB: Ten years after. *Cell* **87:** 13-20.

- Bailey LB, Gregory JF (1999a). Folate metabolism and requirements. *Journal of Nutrition* **129:** 779-782.
- Bailey LB, Gregory JF (1999b). Polymorphisms of methylenetetrahydrofolate reductase and other enzymes: Metabolic significance, risks and impact on folate requirement. *Journal of Nutrition* **129:** 919-922.
- Barber RC, Lammer EJ, Shaw GM, Greer KA, Finnell RH (1999). The role of folate transport and metabolism in neural tube defect risk. *Molecular Genetics and Metabolism* **66:** 1-9.
- Bardonaro M, Nordstrom JL (1994). Different mechanisms are responsible for the low accumulation of transcripts from intronless and 3′ splice site deleted genes. *Biochemical and Biophysical Research Communications* **203:** 128-132.
- Barnes WM (1994). PCR amplification of up to 35kb DNA with high fidelity and high yield from λ bacteriophage templates. *Proceedings of the National Academy of Sciences of the USA* **91:** 2216-2220.
- Barski OA, Gabbay KH, Bohren KM (1999). Characterisation of the human aldehyde reductase gene and promoter. *Genomics* **60:** 188-198.
- Battey J, Jordan E, Cox D, Dove W (1999). An action plan for mouse genomics. *Nature Genetics* **21:** 73-75.
- Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MFW *et al.* (2000). Genealogies of mouse inbred strains. *Nature Genetics* **24:** 23-25.
- Beebee T, Burke J (1990). Gene structure and transcription. IRL Press, Oxford, UK.
- Bell DA, Badawi AF, Lang NP, Ilett KF, Hirvonen A (1995a). Polymorphism in the N-acetyltransferase (NAT1) polyadenylation signal: Association of the *NAT1*10* allele with higher N-acetylation activity in bladder and colon tissue. *Cancer Research* **55:** 5226-5229.
- Bell DA, Stephens EA, Castranio T, Umbach DM, Watson M, Deakin M *et al.* (1995b). Polyadenylation polymorphism in the acetyltransferase 1 gene (*NAT1*) increases risk of colorectal cancer. *Cancer Research* **55:** 3537-3542.
- Bell DA, Taylor JA, Butler MA, Stephens EA, Wiest J, Brubaker LH, Kadlubar FF *et al.* (1993). Genotype/phenotype discordance for human arylamine N-acetyltransferase (NAT2) reveals a new slow-acetylator allele common in African-Americans. *Carcinogenesis* **14:** 1689-1692.
- Bendall HH, Scherer DC, Edson CR, Ballard DW, Oltz EM (1997). Transcription factor NF-κB regulates inducible Oct-2 gene expression in precursor B lymphocytes. *Journal of Biological Chemistry* **272:** 28826-28828.
- Bengel D, Heils A, Petri S, Seemann M, Glatz K, Andrews A *et al.* (1997). Gene structure and 5′-flanking regulatory region of the murine serotonin transporter. *Molecular Brain Research* **44:** 286-292.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2000). GenBank. *Nucleic Acids Research* **28:** 15-18.

- Berk AJ (1999). Activation of RNA polymerase II transcription. *Current Opinion in Cell Biology* **11:** 330-335.
- Bertilsson L, Dahl M-L, Sjöqvist F, Aberg-Wistedt A, Humble M, Johansson I, Lundqvist E, Ingelman-Sundberg M (1993). Molecular basis for rational megaprescribing in ultrarapid hydroxylators of debrisoquine. *The Lancet* **341:** 63.
- Bienz-Tadmor O, Zakut-Houri R, Nesco S, Givol D, Oren M (1985). The 5′ region of the p53 gene: evolutionary conservation and evidence for a negative regulatory element. *The EMBO Journal* **4:** 3209-3213.
- Bigler J, Chen C, Potter JD (1997). Determination of human *NAT2* acetylator genotype by oligonucleotide ligation assay. *BioTechniques* **22:** 682-690.
- Birnstiel ML, Busslinger M, Strub K (1985). Transcription termination and 3′ processing: The end is in site! *Cell* **41:** 349-359.
- Black DL (2000). Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell* **103:** 367-370.
- Blake JA, Richardson JE, Davisson MT, Eppig JT and the Mouse Genome Informatics Group (1997). The Mouse Genome Database (MGD). A comprehensive public resource of genetic, phenotypic and genomic data. *Nucleic Acids Research* **25:** 85-91.
- Blum M, Demierre A, Grant DM, Heim M, Meyer U (1991). Molecular mechanism of slow acetylation of drugs and carcinogens in humans. *Proceedings of the National Academy of Sciences of the USA* **88**: 5237-5241.
- Blum M, Grant D, Demierre A, Meyer UA (1989a). Nucleotide sequence of a full-length cDNA for arylamine N-acetyltransferase from rabbit liver. *Nucleic Acids Research* **17:** 3589.
- Blum M, Grant DM, Demierre A, Meyer UA (1989b). N-acetylation pharmacogenetics: A gene deletion causes absence of arylamine N-acetyltransferase in liver of slow acetylator rabbits. *Proceedings of the National Academy of Sciences of the USA* **86:** 9554-9557.
- Blum M, Grant DM, McBride W, Heim M, Meyer UA (1990a). Human N-acetyltransferase genes: Isolation, chromosomal localisation and functional expression. *DNA and Cell Biology* **9:** 193-203.
- Blum M, Heim M, Meyer UA (1990b). Nucleotide sequence of rabbit *NAT2* encoding polymorphic liver arylamine N-acetyltransferase (NAT). *Nucleic Acids Research* **18:** 5295.
- Blum M, Heim M, Meyer UA (1990c). Nucleotide sequence of rabbit *NAT1* encoding monomorphic arylamine N-acetyltransferase. *Nucleic Acids Research* **18:** 5287.
- Boissy RJ, Watson MA, Umbach DM, Deakin M, Elder J, Strange RC, Bell DA (2000). A pilot study investigating the role of NAT1 and NAT2 polymorphisms in gastric adenocarcinoma. *International Journal of Cancer* **87:** 507-511.
- Bookstein R, Levy A, MacGrogan D, Lewis TB, Weissenbach J, O'Connell P, Leach AJ (1994). Yeast artificial chromosome and radiation hybrid map of loci in

chromosome band 8p22, a common region of allelic loss in multiple human cancers. *Genomics* **24:** 317-323.

- Borjigin J, Li X, Snyder SH (1999). The pineal gland and melatonin: Molecular and pharmacologic regulation. *Annual Reviews in Pharmacology and Toxicology* **39:** 53-65.
- Bork P, Copley R (2001). Filling in the gaps. *Nature* **409:** 818-820.
- Boussif O, Zanta MA and Behr J-P (1996). Optimised galenics improve *in vitro* gene transfer with cationic molecules up to 1000-fold. *Gene Therapy* **3:** 1074-1080.
- Brady KP, Rowe LB, Her H, Stevens TJ, Eppig J, Sussman DJ *et al.* (1997). Genetic mapping of 262 loci derived from expressed sequences in the murine interspecific cross using single-strand conformational polymorphism analysis. *Genome Research* **7:** 1085-1093.
- Brewer G (1971). Annotation: Human ecology, an expanding role for the human geneticist. *American Journal of Human Genetics* **23:** 92.
- Briggs MR, Kadonaga JT, Bell SP, Tjian R (1986). Purification and biochemical characterisation of the promoter-specific transcription factor, Sp1. *Science* **234:** 47-52.
- Brockmöller J, Cascorbi I, Kerb R, Roots I (1996). Combined analysis of inherited polymorphisms in arylamine N-acetyltransferase 2, glutathione S-transferases M1 and T1, microsomal epoxide hydrolase and cytochrome P450 enzymes as modulators of bladder cancer risk. *Cancer Research* **56:** 3915-3925.
- Brown MS, Goldstein JL (1997). The SREBP pathway: Regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor. *Cell* **89:** 331-340.
- Brown PO, Botstein D (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21:** 33-37.
- Bruce C, Whitelaw A, Archibald AL, Harris S, McClenaghan M, Simons JP, Clark AJ (1991). Targeting expression to the mammary gland: intronic sequences can enhance the efficiency of gene expression in transgenic mice. *Transgenic Research* **1:** 3-13.
- Bruhn C, Brockmöller J, Cascorbi I, Roots I, Borchert H-H (1999). Correlation between genotype and phenotype of the human arylamine N-acetyltransferase type 1 (NAT1). *Biochemical Pharmacology* **58:** 1759-1764.
- Bui TD, Woe H-Y, Khan I, Burton PBJ, Moniz CF (1996). Two nuclear proteins bind to the promoter P3 region of the human parathyroid hormone-related peptide gene. *Biochemical and Biophysical Research Communications* **226:** 222-225.
- Bunschoten A, Tiemersma E, Schouls L, Kampman E (2000). Simultaneous determination of polymorphism in N-acetyltransferase 1 and 2 genes by reverse line blot hybridisation. *Analytical Biochemistry* **285:** 156-162.
- Buratowski S (1997). Multiple TATA-binding factors come back into style. *Cell* **91:** 13-15.

- Buratowski S (2000). Snapshots of RNA polymerase II transcription initiation. *Current Opinion in Cell Biology* **12:** 320-325.
- Burley SK (1996). Picking up the TAB. *Nature* **381:** 112-113.
- Burley SK, Roeder RG (1996). Biochemistry and structural biology of transcription factor IID (TFIID). *Annual Reviews of Biochemistry* **65:** 769-799.
- Butcher NJ, Ilett KF, Minchin RF (1998). Functional polymorphism of the arylamine N-acetyltransferase type 1 gene caused by $C^{190}T$ and $G^{560}A$ mutations. *Pharmacogenetics* **8:** 67-72.
- Buters JTM, Sakai S, Richter T, Pineau T, Alexander DL, Savas U, Doehmer J *et al.* (1999). Cytochrome P450 CYP1B1 determines susceptibility to 7,12-dimethylbenz[α]anthracene-induced lymphomas. *Proceedings of the National Academy of Sciences of the USA* **96:** 1977-1982.
- Butler JEF, Kadonaga JT (2001). Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes and Development* **15:** 2515-2519.
- Caldwell J, Gardner I, Swales N (1995). An introduction to drug disposition: The basic principles of absorption, distribution, metabolism and excretion. *Toxicologic Pathology* **23:** 102-114.
- Camilo E, Zimmerman J, Mason JB, Golner B, Russel R, Selhub J, Rosenberg IH (1996). Folate synthesised by bacteria in the human upper small intestine is assimilated by the host. *Gastroenterology* **110:** 991-998.
- Cantor CR (1990). Orchestrating the Human Genome Project. *Science* **248:** 49-51.
- Capecchi MR (1989). The new mouse genetics: Altering the genome by gene targeting. *Trends in Genetics* **5:** 70-76.
- Caron H, Van Schaik B, Van der Mee M, Baas F, Riggins G, Van Sluis P, Hermus M-C *et al.* (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289-1292.
- Carson PE, Flanagan CL, Ickes CE, Alving AS (1956). Enzymatic deficiency in primaquine sensitive erythrocytes. *Science* **124:** 484-485.
- Carter M, Urlich S, Oofuji Y, Williams DA, Ross ME (1999). *Crooked tail* (*Cd*) models of human folate-responsive neural tube defects. *Human Molecular Genetics* **8:** 2199-2204.
- Cartwright RA, Glashan RW, Rogers HJ, Ahmad RA, Barham-Hall D, Higgins E, Kahn MA (1982). The role of N-acetyltransferase phenotype in bladder carcinogenesis: A pharmacogenetic epidemiological approach to bladder cancer. *Lancet* **2:** 842-846.
- Carver EA, Stubbs L (1997). Zooming in on the human-mouse comparative map: Genome conservation re-examined on a high-resolution scale. *Genome Research* **7**: 1123-1137.
- Cascorbi I, Brockmöller J, Bauer S, Reum T, Roots I (1996a). *NAT2*12A* (803A→G) codes for rapid arylamine N-acetylation in humans. *Pharmacogenetics* **6:** 257-259.

- Cascorbi I, Brockmöller J, Mrozikiewicz PM, Bauer S, Loddenkemper R, Roots I (1996b). Homozygous rapid arylamine N-acetyltransferase (NAT2) genotype as a susceptibility factor for lung cancer. *Cancer Research* **56:** 3961-3966.
- Cascorbi I, Drakoulis N, Brockmöller J, Maurer A, Sperling K, Roots I (1995). Arylamine N-acetyltransferase (*NAT2*) mutations and their allelic linkage in unrelated caucasian individuals: Correlation with phenotypic activity. *American Journal of Human Genetics* **57:** 581-592.
- Chabot B (1996). Directing alternative splicing: Cast and scenarios. *Trends in Genetics* **12:** 472-478.
- Chao DM, Young RA (1996). Activation without a vital ingredient. *Nature* **383:** 119-120.
- Chen J, Stampfer MJ, Hough HL, Garcia-Closas M, Willett WC, Hennekens CH, Kelsey KT *et al.* (1998). A prospective study of N-acetyltransferase genotype, red meat intake, and risk of colorectal cancer. *Cancer Research* **58:** 3307-3311.
- Cheng S (1995). Longer PCR amplifications. In *PCR Strategies*, edited by Innis MA, Gelfand DH and Sninsky JJ. Academic Press, London, UK; pp.313-324.
- Cheng S, Fockler C, Barnes WM, Higuchi R (1994). Effective amplification of long targets from cloned inserts and human genomic DNA. *Proceedings of the National Academy of Sciences of the USA* **91:** 5695-5699.
- Chiang P-W, Zhang R, Stubbs L, Zhang L, Zhu L, Kurnit DM (1998). Comparison of murine *Supt4h* and a nearly identical expressed, processed gene: evidence of sequence conservation through gene conversion extending into the untranslated regions. *Nucleic Acids Research* **26:** 4960-4964.
- Chumakov IM, Rigault P, Le Gall I, Bellanné-Chantelot C, Billault A, Guillou S, Soularue P *et al.* (1995). A YAC contig map of the human genome. *Nature* **377:** 175-183.
- Chung JG, Lee JH, Ho CC, Lai JM, Chou YC, Teng HH, Hung CF *et al.* (1997). A survey of arylamine N-acetyltransferase activity in common fruits and vegetables. *Journal of Food Biochemistry* **20:** 481-490.
- Chung JG, Levy GN, Weber WW (1993). Distribution of 2-aminofluorene and p-aminobenzoic acid N-acetyltransferase activity in the tissues of C57BL/6J rapid and B6.A-*Nat*^{S} slow acetylator congenic mice. *Drug Metabolism and Disposition* **21:** 1057-1063.
- Clark AJ, Archibald AL, McClenaghan M, Simons JP, Wallace R, Whitelaw CBA (1993). Enhancing the efficiency of transgene expression. *Philosophical Transactions of the Royal Society of London B* **339:** 225-232.
- Clarke R, Smith AD, Jobst KA, Refsum H, Sutton L, Ueland PM (1998). Folate, vitamin B12 and serum total homocysteine levels in confirmed Alzheimer's disease. *Archives of Neurology* **55:** 1449-1455.
- Claverie J-M (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics* **6:** 1735-1744.

- Colgan J, Manley JL (1995). Cooperation between core promoter elements influences transcriptional activity *in vivo*. *Proceedings of the National Academy of Sciences of the USA* **92:** 1955-1959.
- Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L (1998). New goals for the US Human Genome Project: 1998-2003. *Science* **282:** 682-689.
- Cooper TA, Mattox W (1997). The regulation of splice-site selection, and its role in human disease. *American Journal of Human Genetics* **61:** 259-266.
- Copp AJ (1995). Death before birth: Clues from gene knockouts and mutations. *Trends in Genetics* **11:** 87-93.
- Corcoran J (1998). What are the molecular mechanisms of neural tube defects? *BioEssays* **20:** 6-8.
- Cornish V, Pinter K, Smith A, Boukouvala S, Mo M, Payton M, Sim E (2001). Murine arylamine N-acetyltransferase type 2; A knock-out mouse progress report. In *Proceedings of the Second International NAT Workshop*, 5-6 October 2001, Oxford, UK; p. T6.
- Coroneos E, Gordon JW, Kelly SL, Wang PD, Sim E (1991). Drug metabolising N-acetyltransferase activity in human cell lines. *Biochimica et Biophysica Acta* **1073:** 593-599.
- Coroneos E, Sim E (1991). N-acetyltransferase in human urothelium and bladder cancer cell lines. *Biochemical Society Transactions* **19:** 129S.
- Coroneos E, Sim E (1993). Arylamine N-acetyltransferase activity in human cultured cell lines. *Biochemical Journal* **294:** 481-486.
- Coughtrie MWH, Sharp S, Maxwell K, Innes NP (1998). Biology and function of the reversible sulfation pathway catalysed by human sulfotransferases and sulfatases. *Chemico-Biological Interactions* **109:** 3-27.
- Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, Thompson NE, Burgess RR *et al.* (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* **288:** 640-649.
- Cramer P, Bushnell DA, Kornberg RD (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292:** 1863-1876.
- Cribb AE, Grant DM, Miller MA, Spielberg SP (1991). Expression of monomorphic arylamine N-acetyltransferase (NAT1) in human leucocytes. *Journal of Pharmacology and Experimental Therapeutics* **259:** 1241-1246.
- Curran JL (1997). Human linkage mapping. In *Genome Mapping: A Practical Approach*, edited by Dear PH, Oxford University Press, New York, USA.
- Czeizel AE and Dudás I (1992). Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. *New England Journal of Medicine* **327:** 1832-1835.
- Dabora SL, Sigalas I, Hall F, Eng C, Vijg J, Kwiatkowski DJ (1998). Comprehensive mutation analysis of *TSC1* using two-dimensional DNA electrophoresis with DGGE. *Annals of Human Genetics* **62:** 491-504.

- Daniels R, Adjaye J, Bolton V, Monk M (1998). Detection of a novel splice variant of the hypoxanthine-guanine phosphoribosyl transferase gene in human oocytes and preimplantation embryos: implications for a RT-PCR-based preimplantation diagnosis of Lesch-Nyhan syndrome. *Molecular Human Reproduction* **4:** 785-789.
- Darnell JE (1997). STATs and gene regulation. *Science* **277:** 1630-1635.
- De Leon JH, Martell KJ, Vatsis KP, Weber WW (1995). Slow acetylation in mice is caused by a labile and catalytically impaired mutant N-acetyltransferase (NAT2 9). *Drug Metabolism and Disposition* **23:** 1354-1361.
- De Leon JH, Vatsis KP, Weber WW (2000). Characterisation of naturally occuring and recombinant human N-actyltransferase variants encoded by *NAT1**. *Molecular Pharmacology* **58:** 288-299.
- Debiec-Rychter M, Land SJ, King CM (1999). Histological localisation of acetyltransferases in human tissue. *Cancer Letters* **143:** 99-102.
- Deguchi T (1992). Sequences and expression of alleles of polymorphic arylamine N-acetyltransferase of human liver. *Journal of Biological Chemistry* **267:** 18140-18147.
- Deguchi T, Mashimo M, Suzuki T (1990). Correlations between acetylator phenotypes and genotypes of polymorphic arylamine N-acetyltransferase in human liver. *Journal of Biological Chemistry* **265:** 12757-12760.
- Deitz AC, Zheng W, Leff MA, Gross M, Wen WQ, Doll MA, Xiao GH *et al.* (2000). N-acetyltransferase-2 genetic polymorphism, well-done meat intake, and breast cancer risk among postmenopausal women. *Cancer Epidemiology, Biomarkers and Prevention* **9:** 905-910.
- Delaforge M (1998). Importance of metabolism in pharmacological studies: Possible *in vitro* predictability. *Nuclear Medicine and Biology* **25:** 705-709.
- Delfino RJ, Sinha R, Smith C, West J, White E, Lin HJ, Liao SY *et al.* (2000a). Breast cancer, heterocyclic aromatic amines from meat and acetyltransferase 2 genotype. *Carcinogenesis* **21:** 607-615.
- Delfino RJ, Smith C, West JG, Lin HJ, White E, Liao SY, Gim JS *et al.* (2000b). Breast cancer, passive and active cigarette smoking and N-acetyltransferase 2 genotype. *Pharmacogenetics* **10:** 461-469.
- Deloménie C, Fouix S, Longuemaux S, Brahimi N, Bizet C, Picard B *et al.* (2001). Identification and functional characterisation of arylamine N-acetyltransferases in eubacteria: Evidence for highly selective acetylation of 5-aminosalicylic acid. *Journal of Bacteriology* **183:** 3417-3427.
- Deloménie C, Goodfellow GH, Krishnamoorthy R, Grant DM, Dupret J-M (1997). Study of the role of the highly conserved Arg^{9} and Arg^{64} in the catalytic function of human N-acetyltransferases NAT1 and NAT2 by site-directed mutagenesis. *Biochemical Journal* **323:** 207-215.
- Deloménie C, Sica L, Grant DM, Krishnamoorthy R, Dupret J-M (1996). Genotyping of the polymorphic N-acetyltransferase (*NAT2**) gene locus in two native African populations. *Pharmacogenetics* **6:** 177-185.

- Deloukas P, Matthews LH, Ashurst J, Burton J, Gilbert JGR, Jones M, Stavrides G *et al.* (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865-871.
- Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L *et al.* (1998). A physical map of 30,000 human genes. *Science* **282:**744-746.
- Denoyelle F, Marlin S, Weil D, Moatti L, Chauvin P, Garabédian É-N *et al.* (1999). Clinical features of the prevalent form of childhood deafness, *DFNB1*, due to a connexin-26 gene defect: implications for genetic counselling. *The Lancet* **353:** 1298-1303.
- Derewlany LO, Knie B, Koren G (1994). Arylamine N-acetyltransferase activity of the human placenta. *Journal of Pharmacology and Experimental Therapeutics* **269:** 756-760.
- Derewlany LO, Koren G (1994). Biotransformation of carcinogenic arylamines and arylamides by human placenta. *Journal of Laboratory and Clinical Medicine* **124:** 134-141.
- Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P *et al.* (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152-154.
- Diehl SR, Erickson RP (1997). Genome scan for teratogen-induced clefting susceptibility loci in the mouse: Evidence of both allelic and locus heterogeneity distinguishing cleft lip and cleft palate. *Proceedings of the National Academy of Sciences of the USA* **94:** 5231-5236.
- Dietrich WF, Copeland NG, Gilbert DJ, Miller JC, Jenkins NA, Lander ES (1995). Mapping the mouse genome: Current status and future prospects. *Proceedings of the National Academy of Sciences of the USA* **92:** 10849-10853.
- Doll MA, Fretland AJ, Deitz AC, Hein DW (1995). Determination of human NAT2 acetylator genotype by restriction fragment length polymorphism and allele specific amplification. *Analytical Biochemistry* **231:** 412-420.
- Doll MA, Hein DW (1995). Cloning, sequencing and expression of *NAT1* and *NAT2* encoding genes from rapid and slow acetylator inbred rats. *Pharmacogenetics* **5:** 247-251.
- Doll MA, Jiang W, Deutz AC, Rustan TD, Hein DW (1997). Identification of a novel allele at the human *NAT1* acetyltransferase locus. *Biochemical and Biophysical Research Communications* **233:** 584-591.
- Dong S, Lester L, Johnson LF (2000). Transcriptional control elements and complex initiation pattern of the TATA-less bidirectional human thymidylate synthase promoter. *Journal of Cellular Biochemistry* **77:** 50-64.
- Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM (1999). Expression profiling using cDNA microarrays. *Nature Genetics* **21:** 10-14.
- Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402:** 489-495.

- Dupret J-M, Goodfellow GH, Janezic SA, Grant DM (1994). Structure-function studies of human arylamine N-acetyltransferases NAT1 and NAT2; Functional analysis of recombinant NAT1/NAT2 chimaeras expressed in *Escherichia coli*. *Journal of Biological Chemistry* **269:** 26830-26835.
- Dupret J-M, Grant DM (1992). Site-derected mutagenesis of recombinant human arylamine N-acetyltransferase expressed in *Escherichia coli*; Evidence for direct involvement of Cys^{68} in the catalytic mechanism of polymorphic human NAT2. *Journal of Biological Chemistry* **267:** 7381-7385.
- Duprez D, Treagger J, Pecqueur C, Vigny M (1994). Organisation and promoter activity of the retinoic-acid-induced-heparin-binding (RIHB) gene. *European Journal of Biochemistry* **224:** 931-941.
- Duthie SJ (1999). Folic acid deficiency and cancer: mechanisms of DNA instability. *British Medical Bulletin* **55:** 578-592.
- Dynan WS (1986). Promoters for housekeeping genes. *Trends in Genetics* **2:** 196-197.
- Dynan WS, Sazer S, Tjian R, Schimke RT (1986). Transcription factor Sp1 recognises a DNA sequence in the mouse dihydrofolate reductase gene. *Nature* **319:** 246-248.
- Dynan WS, Tjian R (1985). Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. *Nature* **316:** 774-778.
- Ebisawa T, Deguchi T (1991). Structure and restriction fragment length polymorphism of genes for human liver arylamine N-acetyltransferases. *Biochemical and Biophysical Research Communications* **177:** 1252-1257.
- Ebisawa T, Sasaki Y, Deguchi T (1995). Complementary DNAs for two arylamine N-acetyltransferases with identical 5′ non-coding regions from rat pineal gland.
- Edwalds-Gilbert G, Veraldi KL, Milcarek C (1997). Althernative poly(A) site selection in complex transcription units: Means to an end? *Nucleic Acids Research* **25:** 2547-2561.
- Elliott DJ (2000). Splicing and the single cell. *Histology and Histopathology* **15:** 239-249.
- Elliott DJ, Stutz F, Lescure A, Rosbash M (1994). mRNA nuclear export. *Current Opinions in Genetics and Development* **4:** 305-309.
- Eppig JT, Blake JA, Davisson MT, Richardson JE (1998). Informatics for mouse genetics and genome mapping. *Methods: A Companion to Methods in Enzymology* **14:** 179-190.
- Estrada L, Kanelakis KC, Levy GN, Weber WW (2000). Tissue- and gender-specific expression of N-acetyltransferase 2 (*NAT2**) during development of the outbred mouse strain CD-1. *Drug Metabolism and Disposition* **28:** 139-146.
- Estrada-Rodgers L, Levy GN, Weber WW (1998a). Characterisation of a hormone response element in the mouse N-acetyltransferase 2 (*Nat2**) promoter. *Gene Expression* **7:** 13-24.

- Estrada-Rodgers L, Levy GN, Weber WW (1998b). Substrate selectivity of mouse N-acetyltransferase 1, 2, and 3 expressed in COS-1 cells. *Drug Metabolism and Disposition* **26:** 502-505.
- Evans DAP (1964). Acetylation polymorphisms. *Proceedings of the Royal Society of Medicine* **57:** 508-518.
- Evans DAP (1992). N-acetyltransferase. In *Pharmacogenetics of Drug Metabolism,* edited by Kalow W. Pergamon Press, New York, USA; pp.95-178.
- Evans DAP, Manley KA, McKusick VA (1960). Genetic control of isoniazid metabolism in man. *British Medical Journal* **2:** 485-461.
- Evans DAP, White TA (1964). Human acetylation polymorphism. *Journal of Laboratory and Clinical Medicine* **63:** 394-404.
- Evans WE, Relling MV (1999). Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science* **286:** 487-491.
- Fakis G, Boukouvala S, Buckle V, Payton M, Denning C, Sim E (2000). Chromosomal localisation and mapping of the genes for murine arylamine N-acetyltransferases (NATs), enzymes involved in the metabolism of carcinogens: Identification of a novel upstream non-coding exon for murine *Nat2*. *Cytogenetics and Cell Genetics* **90:** 134-138.
- Fang SH, Chung JG, Chang WC, Chang SS (1997). Evidence for arylamine N-acetyltransferase activity in the fungi *Candida albicans*. *Toxicology Letters* **92:** 109-116.
- Farrington SM, Cunningham C, Boyle SM, Wyllie AH, Dunlop MG (1996). Detailed physical and deletion mapping of 8p with isolation of YAC clones from tumour suppressor loci involved in colorectal cancer. *Oncogene* **12:** 1803-1808.
- Ferguson RJ, Doll MA, Rustan TD, Baumstark BR, Hein DW (1994a). Syrian hamster monomorphic N-acetyltransferase (*NAT1*) alleles: Amplification, cloning, sequencing, and expression in *E. coli*. *Pharmacogenetics* **4:** 82-90.
- Ferguson RJ, Doll MA, Rustan TD, Gray K, Hein DW (1994b). Cloning, expression and functional characterisation of two mutant ($NAT2^{191}$ and $NAT2^{341/803}$) and wild-type human polymorphic N-acetyltransferase (*NAT2*) alleles. *Drug Metabolism and Disposition* **22:** 371-376.
- Ferguson RJ, Doll MA, Rustan TD, Hein DW (1996). Cloning, expression and functional characterisation of rapid and slow acetylator polymorphic N-acetyltransferase encoding genes of the Syrian hamster. *Pharmacogenetics* **6:** 55-66.
- Filiadis IF, Georgiou I, Alamanos Y, Kranas V, Giannakopoulos X, Lolis D (1999). Genotypes of N-acetyltransferase-2 and risk of bladder cancer: A case-control study. *Journal of Urology* **161:** 1672-1675.
- Flammang TJ, Couch LH, Levy GN, Weber WW, Wise CK (1992). DNA adduct levels in congenic rapid and slow acetylator mouse strains following chronic administration of 4-aminobiphenyl. *Carcinogenesis* **13:** 1887-1891.

- Fletcher BH, Cassady AI, Summers KM, Cavanagh AC (2001). The murine chaperonin 10 gene family contains an intronless, putative gene for early pregnancy factor, *Cpn10-rs1*. *Mammalian Genome* **12:** 133-140.
- Franke S, Klowitz I, Schnakenberg E, Rommel B, Van de Ven W, Bullerdiek J, Schloot W (1994). Isolation and mapping of a cosmid clone containing the human *NAT2* gene. *Biochemical and Biophysical Research Communications* **199:** 52-55.
- Fretland AJ, Doll MA, Gray K, Feng Y, Hein D (1997). Cloning, sequencing and recombinant expression of *NAT1*, *NAT2* and *NAT3* derived from the C3H/HeJ (rapid) and A/HeJ (slow) acetylator inbred mouse: Functional characterisation of the activation and deactivation of aromatic amine carcinogens. *Toxicology and Applied Pharmacology* **142:** 360-366.
- Fretland AJ, Leff MA, Doll MA, Hein DW (2001). Functional characterisation of human N-acetyltransferase 2 (NAT2) single nucleotide polymorphisms. *Pharmacogenetics* **11:** 207-215.
- Frymoyer JW, Jacox RF (1963). Studies of genetically controlled sulfadiazine acetylation in rabbit livers: Possible identification of the heterozygous trait. *Journal of Laboratory and Clinical Medicine* **62:** 905-909.
- Fu L, Suen CKM, Waseem A, White KN (1997). Variable requirement for splicing signals for nucleocytoplasmic export of mRNAs. *Biochemistry and Molecular Biology International* **42:** 329-337.
- Fujiwara Y, Emi M, Ohata H, Kato Y, Nakajima T, Mori T, Nakamura Y (1993). Evidence for the presence of two tumour suppressor genes on chromosome 8p for colorectal carcinoma. *Cancer Research* **53:** 1172-1174.
- Fujiwara Y, Ohata H, Emi M, Okui K, Koyama K, Tsuchiya E, Nakajima T (1994). A 3Mb physical map of the chromosome region 8p21.3-p22, including a 600kb region commonly deleted in human hepatocellular carcinoma, colorectal cancer and non-small cell lung cancer. *Genes Chromosomes and Cancer* **10:** 7-14.
- Fukutome K, Watanabe M, Shiraishi T, Murata M, Uemura H, Kubota Y, Kawanamura J *et al.* (1999). N-acetyltransferase 1 genetic polymorphism influences the risk of prostate cancer development. *Cancer Letters* **136:** 83-87.
- Galli-Taliadoros LA, Sedgwick JD, Wood SA, Körner H (1995). Gene knock-out technology: A methodological overview for the interested novice. *Journal of Immunological Methods* **181:** 1-15.
- Gawronska-Szklarz B, Luszawska-Kutrzeba T, Czaja-Bulsa G, Kurzawski G (1999). Relationship between acetylation polymorphism and risk of atopic diseases. *Clinical Pharmacology and Therapeutics* **65:** 562-569.
- Gearing KL, Cairns W, Okret S, Gustafsson J-A (1993). Heterogeneity in the 5′ untranslated region of the rat glucocorticoid receptor mRNA. *Journal of Steroid Biochemistry and Molecular Biology* **46:** 635-639.
- Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M *et al.* (1994). Homeodomain-DNA recognition. *Cell* **78:** 211-223.

- Geiger A, Decaux JF, Burcelin R, Le Cam A, Salazar G, Charron MJ *et al.* (2000). Structural and functional characterisations of the 5′-flanking region of the mouse glucagon receptor gene: Comparison with the rat gene. *Biochemical and Biophysical Research Communications* **272:** 912-921.
- Gelbart WM (1998). Databases in genomic research. *Science* **282:** 659-661.
- Geoghegan FL, McPartlin JM, Weir DG, Scott JM (1995). Para-acetamidobenzoylglutamate is a suitable indicator of folate catabolism in rats. *Journal of Nutrition* **125:** 2563-2570.
- Gerber H-P, Seipel K, Georgiev O, Höfferer M, Hug M, Rusconi S, Schaffner W (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263:** 808-811.
- Gertig DM, Hankinson SE, Hough H, Spiegelman D, Colditz GA, Willett W, Kelsey KT *et al.* (1999). N-acetyltransferase 2 genotypes, meat intake and breast cancer risk. *International Journal of Cancer* **80:** 13-17.
- Glatt H, Engelke CEH, Pabel U, Teubner W, Jones AL, Coughtrie MWH, Andrae U *et al.* (2000). Sulfotransferases: Genetics and role in toxicology. *Toxicology Letters* **112-113:** 341-348.
- Glowinski IB, Weber WW (1982a). Genetic regulation of aromatic amine N-acetylation in inbred mice. *Journal of Biological Chemistry* **257:** 1424-1430.
- Glowinski IB, Weber WW (1982b). Biochemical characterisation of genetically variant aromatic amine N-acetyltransferases in A/J and C57BL/6J mice. *Journal of Biological Chemistry* **257:** 1431-1437.
- Gnatt AL, Cramer P, Fu J, Bushnell DA, Kornberg RD (2001). Structural basis of transcription: An RNA polymerase II elongation complex at 3.3Å resolution. *Science* **292:** 1876-1882.
- Goebel C, Vogel C, Wulferink M, Mittmann S, Sachs B, Schraa S *et al.* (1999). Procainamide, a drug causing lupus, induces prostaglandin H sunthase-2 and formation of T cell-sensitising drug metabolites in mouse macrophages. *Chemical Research and Toxicology* **12:** 488-500.
- Gonzalez FJ (1998). The study of xenobiotic-metabolising enzymes and their role in toxicity *in vivo* using targeted gene disruption. *Toxicology Letters* **102-103:** 161-166.
- Gonzalez FJ (2001). The use of gene knockout mice to unravel the mechanisms of toxicity and chemical carcinogenesis. *Toxicology Letters* **120:** 199-208.
- Gonzalez FJ, Kimura S (1999). Role of gene knockout mice in understanding the mechanisms of chemical toxicity and carcinogenesis. *Cancer Letters* **143:** 199-204.
- Gonzalez FJ, Kimura S (2001). Understanding the role of xenobiotic-metabolism in chemical carcinogenesis using gene knockout mice. *Mutation Research* **477:** 79-87.
- Gonzalez FJ, Nebert DW (1990). Evolution of the P450 gene superfamily: Animal-plant "warfare", molecular drive and human genetic differences in drug oxidation. *Trends in Genetics* **6:** 182-186.

- Gonzalez MV, Alvarez V, Pello MF, Menendez MJ, Suarez C, Coto E (1998). Genetic polymorphism of N-acetyltransferase-2, glutathione S-transferase-M1, and cytochromes P450IIE1 and P450IID6 in the susceptibility to head and neck cancer. *Journal of Clinical Pathology* **51:** 294-298.
- Goodfellow GH, Dupret J-M, Grant DM (2000). Identification of amino acids imparting acceptor substrate selectivity to human arylamine acetyltransferases NAT1 and NAT2. *Biochemical Journal* **348:** 159-166.
- Grant DM (1993). Molecular genetics of N-acetyltransferases. *Pharmacogenetics* **3:** 45-50.
- Grant DM, Blum M, Beer M, Meyer UA (1991). Monomorphic and polymorphic human arylamine N-acetyltransferases: a comparison of liver isozymes and expressed products of two cloned genes. *Molecular Pharmacology* **39:**184-191.
- Grant DM, Blum M, Demierre A, Meyer UA (1989a). Nucleotide sequence of an intronless gene for a human arylamine N-acetyltransferase related to polymorphic drug acetylation. *Nucleic Acids Research* **17:** 3978.
- Grant DM, Goodfellow GH, Sugamori KS, Durette K (2000). Pharmacogenetics of the human arylamine N-acetyltransferases. *Pharmacology* **61:** 204-211.
- Grant DM, Lottspeich F, Meyer UA (1989b). Evidence for two closely related isozymes of arylamine N-acetyltransferase in human liver. *FEBS Letters* **244:** 203-207.
- Graveley BR (2001). Alternative splicing: Increasing diversity in the proteomic world. *Trends in Genetics* **17:** 100-107.
- Gray TA, Hernandez L, Carey AH, Schaldach MA, Smithwick MJ, Rus K *et al.* (2000). The ancient source of a distinct gene family encoding proteins featuring RING and C_3H zinc-finger motifs with abundant expression in developing brain and nervous system. *Genomics* **66:** 76-86.
- Green J, Banks E, Berrington A, Darby S, Deo H, Newton R (2000). N-acetyltransferase 2 and bladder cancer: an overview and consideration of the evidence for gene-environment interaction. *British Journal of Cancer* **83:** 412-417.
- Green MD, Tephly TR (1998). Glucuronidation of amine substrates by purified and expressed UDP-glucuronosyltransferase proteins. *Drug Metabolism and Disposition* **26:** 860-867.
- Greenblatt J (1997). RNA polymerase II holoenzyme and transcriptional regulation. *Current Opinions in Cell Biology* **9:** 310-319.
- Greene NDE, Gerrelli D, Van Straaten HWM, Copp AJ (1998). Abnormalities of floor plate, notochord and somite differentiation in the *loop-tail* (*Lp*) mouse: a model of severe neural tube defects. *Mechanisms of Development* **73:** 59-72.
- Gregory JF (2001). Case study: Folate bioavailability. *Journal of Nutrition* **131:** 1376S-1382S.
- Gross M, Kruisselbrink T, Anderson K, Lang N, McGovern P, Delongchamp R, Kadlubar F (1999). Distribution and concordance of N-acetyltransferase genotype

and phenotype in an American population. *Cancer Epidemiology, Biomarkers and Prevention* **8:** 683-692.

- Gruss P, Efstratiadis A, Karathanasis S, König M, Khoury G (1981). Synthesis of stable unspliced mRNA from an intronless simian virus 40-rat preproinsulin gene recombinant. *Proceedings of the National Academy of Sciences of the USA* **78:** 6091-6095.
- Gruss P, Khoury G (1980). Rescue of a splicing defective mutant by insertion of an heterologous intron. *Nature* **286:** 634-637.
- Gu H, Marth JD, Orban PC, Mossmann H, Rajewsky K (1994). Deletion of a DNA polymerase β gene segment in T cells using cell type-specific gene targeting. *Science* **265:** 103-106.
- Güray T, Güvenç (1997a). Characterisation of sheep liver N-acetyltransferase. *Biochemical Archives* **13:** 107-111.
- Güray T, Güvenç (1997b). Sheep tissue acetyl coenzyme A-dependent arylamine N-acetyltransferase. *Comparative Biochemistry and Physiology* **118C:** 305-310.
- Guyer MS, Collins FS (1995). How is the Human Genome Project doing and what have we learned so far? *Proceedings of the National Academy of Sciences of the USA* **92:** 10841-10848.
- Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S *et al.* (1994). The 1993-94 Genethon human genetic linkage map. *Nature Genetics* **7:** 246-339.
- Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, Muselet D *et al.* (1996). A radiation hybrid map of the human genome. *Human Molecular Genetics* **5:** 339-346.
- Haldi ML, Strickland C, Lim P, VanBerkel V, Chen X-N, Noya D, Korenberg JR *et al.* (1996). A comprehensive large-insert yeast artificial chromosome library for physical mapping of the mouse genome. *Mammalian Genome* **7:** 767-769.
- Halford S, Freedman MS, Bellingham J, Inglis SL, Poopalasundaram S, Soni BG *et al.* (2001). Characterisation of a novel human opsin gene with wide tissue expression and identification of embedded and flanking genes on chromosome 1q43. *Genomics* **72:** 203-208.
- Hanauer A, Mandel JL (1984). The glyceraldehyde 3 phosphate dehydrogenase gene family: structure of a human cDNA and of an X chromosome linked pseudogene; amazing complexity of the gene family in mouse. *The EMBO Journal* **3:** 2627-2633.
- Hanauer SB, Dassopoulos T (2001). Evolving treatment strategies for inflammatory bowel disease. *Annual Reviews of Medicine* **52:** 299-318.
- Hanna PE (1996). Metabolic activation and detoxification of arylamines. *Current Medicinal Chemistry* **3:** 195-210.
- Hardouin SN, Nagy A (2000). Mouse models for human disease. *Clinical Genetics* **57:** 237-244.
- Harhangi BS, Oostra BA, Heutink P, van Duijn CM, Hofman A, Breteler MMB (1999). N-acetyltransferase-2 polymorphism in Parkinson's disease: The

Rotterdam study. *Journal of Neurology, Neurosurgery and Psychiatry* **67:** 518-520.

- Harrison SM, Dunwoodie SL, Arkell RM, Lehrach H, Beddington RSP (1995). Isolation of novel tissue-specific genes from cDNA libraries representing the individual tissue constituents of the gastrulating mouse embryo. *Development* **121:** 2479-2489.
- Haselbeck R, Greer CL (1993). Minimum intron requirements for tRNA splicing and nuclear transport in *Xenopus* oocytes. *Biochemistry* **32:** 8575-8581.
- Hasty P, Bradley A (1992). Gene targeting vectors for mammalian cells. In *Gene Targeting: A Practical Approach*, edited by Joyner AL. Oxford University Press, New York, USA; pp.1-32.
- Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park H-S, Yoyoda A *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* **405:** 311-319.
- Hatzopoulos AK, Schlokat U, Gruss P (1988). Enhancers and other *cis*-acting regulatory sequences. In *Transcription and Splicing*, edited by Hames BD and Glover DM. IRL Press, New York, USA; pp.97-129.
- Hawkins JD (1996). Post-transcriptional processing of RNA. In *Gene Structure and Expression*. Cambridge University Press, Cambridge, UK; pp. 110-123.
- Hayashi T, Suzuki K (1993). Gene organisation of human protein C inhibitor, a member of SERPIN family proteins encoded in five exons. *International Journal of Hematology* **58:** 213-224.
- Hayes RB, Bi W, Rothman N, Broly F, Caporaso N, Feng P, You X *et al.* (1993). N-acetylation phenotype and genotype and risk of bladder cancer in benzidine-exposed workers. *Carcinogenesis* **14:** 675-678.
- Hearse DJ, Weber WW (1973). Multiple N-acetyltransferases and drug metabolism: Tissue distribution, characterisation and significance of mammalian N-acetyltransferase. *Biochemical Journal* **132:** 519-526.
- Hein DW, Doll MA, Fretland AJ, Gray K, Deitz AC, Feng Y *et al.* (1997). Rodent models of human acetylation polymorphism: Comparisons of recombinant acetyltransferases. *Mutation Researh* **376:** 101-106.
- Hein DW, Doll MA, Fretland AJ, Leff MA, Webb SJ, Xiao GH, Devanaboyina U-S *et al.* (2000a). Molecular genetics and epidemiology of the *NAT1* and *NAT2* acetylation polymorphisms. *Cancer Epidemiology, Biomarkers and Prevention* **9:** 29-42.
- Hein DW, Ferguson RJ, Doll MA, Rustan TD, Gray K (1994). Molecular genetics of human polymorphic N-acetyltransferase: Enzymatic analysis of 15 recombinant wild-type, mutant and chimeric NAT2 allozymes. *Human Molecular Genetics* **3:** 729-734.
- Hein DW, Grant DM, Sim E (2000b). Update on consensus arylamine N-acetyltransferase gene nomenclature. *Pharmacogenetics* **10:** 291-292.
- Hein DW, Omichinski JG, Brewer JA, Weber WW (1982). A unique pharmacogenetic expression of the N-acetylation polymorphism in the inbred hamster. *The Journal of Pharmacology and Experimental Therapeutics* **220:** 8-15.

- Hein DW, Rustan TD, Bucher KD, Martin WJ, Furman EJ (1991). Acetylator phenotype-dependent and –independent expression of arylamine N-acetyltransferase isoenzymes in rapid and slow acetylator inbred rat liver. *Drug Metabolism and Disposition* **19:** 933-937.
- Hein DW, Trinidad A, Yerokun T, Ferguson RJ, Kirlin WG, Weber WW (1988). Genetic control of acetyl coenzyme A-dependent arylamine N-acetyltransferase, hydrazine N-acetyltransferase, and N-hydroxy-arylamine O-acetyltransferase enzymes in C57BL/6J, A/J, AC57F_1, and the rapid and slow acetylator A.B6 and B6.A congenic inbred mouse. *Drug Metabolism and Disposition* **16:** 341-347.
- Henderson CJ, Smith AG, Ure J, Brown K, Bacon EJ, Wolf CR (1998). Increased skin tumorigenesis in mice lacking pi class glutathione S-transferases. *Proceedings of the National Academy of Sciences of the USA* **95:** 5275-5280.
- Hengen PN (1994). Long and accurate PCR. *Trends in Biochemical Sciences* **19:** 341.
- Henning S, Cascorbi I, Münchow B, Jahnke V, Roots I (1999). Association of arylamine N-acetyltransferases *NAT1* and *NAT2* genotypes to laryngeal cancer risk. *Pharmacogenetics* **9:** 103-111.
- Herbert A, Rich A (1999). RNA processing and the evolution of eukaryotes. *Nature Genetics* **21:** 265-269.
- Hernandez N (1993). TBP, a universal eukaryotic transcription factor? *Genes and Development* **7:** 1291-1308.
- Herrup K (1994). Transgenic and ES cell chimeric mice as tools for the study of the nervous system. *Discussion in Neuroscience* **X (3+4):** 25-63.
- Hickman D, Palamanda JR, Unadkat JD, Sim E (1995). Enzyme kinetic properties of human recombinant arylamine N-acetyltransferase 2 allotypic variants expressed in *Escherichia coli*. *Biochemical Pharmacology* **50:** 697-703.
- Hickman D, Pope J, Patil SD, Fakis G, Smelt V, Stanley LA, *et al.* (1998). Expression of arylamine N-acetyltransferase in human intestine. *Gut* **42:** 402-409.
- Hickman D, Risch A, Buckle V, Spurr NK, Jeremiah SJ, McCarthy A and Sim E (1994). Chromosomal localisation of human genes for arylamine N-acetyltransferase. *Biochemical Journal* **297:** 441-445.
- Hickman D, Sim E (1991). N-acetyltransferase polymorphism: Comparison of phenotype and genotype in humans. *Biochemical Pharmacology* **42:** 1007-1014.
- Hirata S, Koh T, Yamada-Mouri N, Hoshi K, Kato J (1996). The untranslated first exon "exon 0S" of the rat estrogen receptor (ER) gene. *FEBS Letters* **394:** 371-373.
- Hirvonen A (1999). Polymorphisms of xenobiotic-metabolising enzymes and susceptibility to cancer. *Environmental Health Perspectives* **107S:** 37-47.
- Ho CC, Lin TH, Lai YS, Chung JG, Levy GN, Weber WW (1996). Kinetics of acetyl coenzyme A: Arylamine N-acetyltransferase from rapid and slow acetylator frog tissues. *Drug Metabolism and Disposition* **24:** 137-141.
- Hodges D, Bernstein SI (1994). Genetic and biochemical analysis of alternative RNA splicing. *Advances in Genetics* **31:** 207-280.

- Hooper M, Hardy K, Handyside A, Hunter S, Monk M (1987). HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326:** 292-295.
- Horowitz DS and Krainer AR (1994). Mechanisms for selecting 5′ splice sites in mammalian pre-mRNA splicing. *Trends in Genetics* **10:** 100-106.
- Horvath CM, Darnell JE (1997). The state of the STATs: Recent developments in the study of signal transduction to the nucleus. *Current Opinions in Cell Biology* **9:** 233-239.
- Hou S-M, Ryberg D, Fält S, Deverill A, Tefre T, Børresen A-L, Haugen A *et al.* (2000). GSTM1 and NAT2 polymorphisms in operable and non-operable lung cancer patients. *Carcinogenesis* **21:** 49-54.
- Hou VC, Conboy JG (2001). Regulation of alternative pre-mRNA splicing during erythroid differentiation. *Current Opinions in Hematology* **8:** 74-79.
- Hsieh F-I, Pu Y-S, Chern H-D, Hsu L-I, Chiou H-Y, Chen C-J (1999). Genetic polymorphisms of N-acetyltransferase 1 and 2 and risk of cigarette smoking-related bladder cancer. *British Journal of Cancer* **81:** 537-541.
- Hu Z-Z, Zhuang L, Meng J, Dufau ML (1998). Transcriptional regulation of the generic promoter III of the rat prolactin receptor gene by C/EBPβ and Sp1. *Journal of Biological Chemistry* **273:** 26225-26235.
- Hu Z-Z, Zhuang L, Meng J, Leondires M, Dufau ML (1999). The human prolactin receptor gene structure and alternative promoter utilisation: The generic promoter hPIII and a novel human promoter hP_N. *Journal of Clinical Endocrinology and Metabolism* **84:** 1153-1156.
- Huang C-S, Chern H-D, Shen C-Y, Hsu S-M, Chang K-J (1999a). Association between N-acetyltransferase 2 (*NAT2*) genetic polymorphism and development of breast cancer in post-menopausal Chinese women in Taiwan, an area of great increase in breast cancer incidence. *International Journal of Cancer* **82:** 175-177.
- Huang Y, Carmichael GG (1997). The mouse histone H2a gene contains a small element that facilitates cytoplasmic accumulation of intronless gene transcripts and of unspliced HIV-1-related mRNAs. *Proceedings of the National Academy of Sciences of the USA* **94:** 10104-10109.
- Huang Y, Wimler KM, Carmichael GG (1999b). Intronless mRNA transport elements may affect multiple steps of pre-mRNA processing. *The EMBO Journal* **18:** 1642-1652.
- Hubbard A, Moyes C, Wyllie AH, Smith CAD, Harisson DJ (1998). N-acetyltransferase 1: Two polymorphisms in coding sequence identified in colorectal cancer patients. *British Journal of Cancer* **77:** 913-916.
- Hubbard AL, Harrison DJ, Moyes C, Wyllie AH, Cunningham C, Manion E, Smith CAD (1997). N-acetyltransferase 2 genotype in colorectal cancer and selective gene retention in cancers with chromosome 8p deletions. *Gut* **41:** 229-234.

- Hudson TJ, Church DM, Greenaway S, Nguyen H, Cook A, Steen RG, Van Etten WJ *et al.* (2001). A radiation hybrid map of mouse genes. *Nature Genetics* **29:** 201-205.
- Hudson TJ, Stein LD, Gerety SS, Ma J, Castle AB, Silva J, Slonim DK *et al.* (1995). An STS-based map of the human genome. *Science* **270:** 1945-1954.
- Hughes NC, Janezic SA, McQueen KL, Jewett MAS, Castranio T, Bell DA *et al.* (1998a). Identification and characterisation of variant alleles of human acetyltransferase NAT1 with defective function using p-aminosalicylate as an in-vivo and in-vitro probe. *Pharmacogenetics* **8:** 55-66.
- Hughes SJ, Glover TW, Zhu XX, Kuick R, Thoraval D, Orringer MB, Beer DG, Hanash S (1998b). A novel amplicon at 8p22-23 results in overexpression of cathepsin B in esophageal adenocarcinoma. *Proceeding of the National Academy of Sciences USA* **95:** 12410-12415.
- Humphrey T, Proudfoot NJ (1988). A beginning to the biochemistry of polyadenylation. *Trends in Genetics* **4:** 243-245.
- Hunter DJ, Hankinson SE, Hough H, Gertig DM, Garcia-Closas M, Spiegelmn D, Manson JAE *et al.* (1997). A prospective study of NAT2 acetylation genotype, cigarette smoking and risk of breast cancer. *Carcinogenesis* **18:** 2127-2132.
- Huo L, Scarpulla RC (1999). Multiple 5′-untranslated exons in the nuclear respiratory factor 1 gene span 47kb and contribute to transcript heterogeneity and translational efficiency. *Gene* **233:** 213-224.
- Idle JR, Armstrong M, Boddy AV, Boustead C, Cholerton S, Cooper J, Daly AK, Ellis J, Gregory W, Hadidi H, et al (1992). The pharmacogenetics of chemical carcinogenesis. *Pharmacogenetics* **2:** 246-258.
- IHGSC: The International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.
- Ihle JN, Kerr IM (1995). Jaks and Stats in signalling by the cytokine receptor superfamily. *Trends in Genetics* **11:** 69-74.
- Ilett KF, Ingram DM, Carpenter DS, Teitel CH, Lang NP, Kadlubar FF *et al.* (1994). Expression of monomorphic and polymorphic N-acetyltransferases in human colon. *Biochemical Pharmacology* **47:** 914-917.
- Ilett KF, Kadlubar FF, Minchin RF (1999). 1998 International meeting on the arylamine N-acetyltransferases: Synopsis of the workshop on nomenclature, biochemistry, molecular biology, interspecies comparisons and role in human disease risk. *Drug Metabolism and Disposition* **27:** 957-959.
- Inatomi H, Katoh T, Kawamoto T, Matsumoto T (1999). *NAT2* gene polymorphism as a possible marker for susceptibility to bladder cancer in Japanese. *International Journal of Urology* **6:** 446-454.
- Ingelman-Sundberg M (1998). Functional consequences of polymorphism of xenobiotic metabolising enzymes. *Toxicology Letters* **102-103:** 155-160.
- Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H, Chen C, Baltzer MA *et al.* (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genetics* **6:** 84-89.

- Ishii H, Baffa R, Numata S-I, Murakumo Y, Rattan S, Inoue H, Mori M *et al.* (1999). The *FEZ1* gene at chromosome 8p22 encodes a leucine-zipper protein, and its expression is altered in multiple human tumours. *Proceedings of the National Academy of Sciences of the USA* **96:** 3928-3933.
- Ito E, Xie G-X, Maruyama K, Pierce Palmer P (2000). A core-promoter region functions bi-directionally for human opioid-receptor-like gene ORL1 and its 5′-adjacent gene GAIP. *Journal of Molecular Biology* **304:** 259-270.
- Jackson IJ (2001). Mouse mutagenesis on target. *Nature Genetics* **28:** 198-200.
- Jencks WP, Gresser M, Valenzuela MS, Huneeus FC (1972). Acetyl coenzyme A: arylamine acetyltransferase. Measurement of the steady state concentration of the acetyl-enzyme intermediate. *Journal of Biological Chemistry* **247:** 3756-3760.
- Jenne JW (1965). Partial purification and properties of the isoniazid trans-acetylase in human liver. Its relationship to the acetylation of p-aminosalicylic acid. *Journal of Clinical Investigation* **44:** 1992-2002.
- Jenne JW, McDonald FM, Mendoza E (1961). A study of the renal clearances, metabolic inactivation rates, and serum fall-off interaction of isoniazid and para-aminosalicylic acid in man. *American Reviews of Respiratory Disease* **84:** 371-378.
- Jeong Y, Won J, Yim J (1998). Cloning and structure of a rabbit protein inhibitor of neuronal nitric oxide synthase (PIN) gene and its pseudogene. *Gene* **214:** 67-75.
- Jitrapakdee S, Booker GW, Cassady AI, Wallace JC (1997). The rat pyruvate carboxylase gene structure: Alternate promoters generate multiple transcripts with the 5′-end heterogeneity. *Journal of Biological Chemistry* **272:** 20522-20530.
- Johns LE, Houlston RS (2000). N-acetyltransferase-2 and bladder cancer risk: A meta-analysis. *Environmental and Molecular Mutagenesis* **36:** 221-227.
- Johnson PF, McKnight (1989). Eukaryotic transcriptional regulatory proteins. *Annual Reviews in Biochemistry* **58:** 799-839.
- Jolly C, Vourc'h C, Robert-Nicoud M, Morimoto RI (1999). Intron-independent association of splicing factors with active genes. *Journal of Cell Biology* **145:** 1133-1143.
- Jolly DJ, Okayama H, Berg P, Esty AC, Filpula D, Bohlen P, Johnson GG *et al.* (1983). Isolation and characterisation of a full-length expressible cDNA for human hypoxanthine phosphoribosyltransferase. *Proceedings of the National Academy of Sciences of the USA* **80:** 477-481.
- Jourenkova-Mironova N, Wikman H, Bouchardy C, Mitrunen K, Dayer P, Benhamou S, Hirvonen A (1999). Role of N-acetyltransferase 1 and 2 (*NAT1* and *NAT2*) genoypes in susceptibility to oral/pharyngeal and laryngeal cancers. *Pharmacogenetics* **9:** 533-537.
- Juberg DR, Bond JT, Weber WW (1991). N-acetylation of aromatic amines: genetic polymorphism in inbred rat strains. *Pharmacogenetics* **1:** 50-57.
- Juriloff DM, Harris MJ (2000). Mouse models for neural tube closure defects. *Human Molecular Genetics* **9:** 993-1000.

- Justice MJ, Zheng B, Woychik RP, Bradley A (1997). Using targeted large deletions and high-efficiency N-ethyl-N-nitrosourea mutagenesis for functional analyses of the mammalian genome. *Methods: A Companion to Methods in Enzymology* **13:** 423-436.
- Kalow W, Genest K (1957). A method for the detection of atypical forms of human serum cholinesterase. Determination of dibucaine numbers. *Canadian Journal of Biochemistry and Physiology* **35:** 339-346.
- Karin M, Liu ZL, Zandi E (1997). AP-1 function and regulation. *Current Opinions in Cell Biology* **9:** 240-246.
- Karolyi J, Erickson RP, Liu S (1988). Genetics of susceptibility to 6-aminonicotinamide-induced cleft palate in the mouse: Studies in congenic and recombinant inbred strains. *Teratology* **37:** 283-287.
- Karolyi J, Erickson RP, Liu S, Killewald L (1990). Major effects on teratogen-induced facial clefting in mice determined by a single genetic region. *Genetics* **126:** 201-205.
- Karolyi J, Liu S, Erickson RP (1987). Susceptibility to phenytoin-induced cleft lip with or without cleft palate: Many genes are involved. *Genetical Research* **49:** 43-49.
- Katoh T, Boissy R, Nagata N, Kitagawa K, Kuroda Y, Itoh H, Kawamoto T, Bell DA (2000). Inherited polymorphism in the N-acetyltransferase 1 (*NAT1*) and 2 (*NAT2*) genes and susceptibility to gastric and colorectal adenocarcinoma. *International Journal of Cancer* **85:** 46-49.
- Katoh T, Inatomi H, Yang M, Kawamoto T, Matsumoto T, Bell DA (1999). Arylamine N-acetyltransferase 1 (NAT1) and 2 (NAT2) genes and risk of urothelial transitional cell carcinoma among Japanese. *Pharmacogenetics* **9:** 401-404.
- Katoh T, Kaneko S, Boissy R, Watson M, Ikemura K, Bell DA (1998). A pilot study testing the association between N-acetyltransferase 1 and 2 and risk of oral squamous cell cacinoma in Japanese people. *Carcinogenesis* **19:** 1803-1807.
- Kelly SL, Sim E (1991). Expression of N-acetyltransferase in a human monocytic cell line, U937. *Human and Experimental Toxicology* **10:** 33-38.
- Kelly SL, Sim E (1992). Arylamine N–acetyltransferase in mice. *Biochemical Society Transactions* **20:** 106S.
- Kelly SL, Sim E (1994). Arylamine N-acetyltransferase in Balb/c mice: Identification of a novel mouse isoenzyme by cloning and expression *in vitro*. *Biochemical Journal* **302:** 347-353.
- Kerppola T, Curran T (1995). Zen and the art of Fos and Jun. *Nature* **373:** 199-200.
- Kiang DT, Jin N, Tu Z-J, Lin HH (1997). Upstream genomic sequence of the human connexin 26 gene. *Gene* **199:** 165-171.
- Kim SH, Moores JC, David D, Respess JG, Jolly DJ, Friedmann T (1986). The organisation of the human HPRT gene. *Nucleic Acids Research* **14:** 3103-3118.

- Kimura S, Gonzalez FJ (2000). Applications of genetically manipulated mice in pharmacogenetics and pharmacogenomics. *Pharmacology* **61:** 147-153.
- Kimura S, Kawabe M, Ward JM, Morishima H, Kadlubar FF, Hammons GJ, Fernandez-Salguero P *et al.* (1999). CYP1A2 is not the primary enzyme responsible for 4-aminobiphenyl-induced hepatocarcinogenesis in mice. *Carcinogenesis* **20:** 1825-1830.
- King CM, Land SJ, Jones RF, Debiec-Rychter M, Lee M-S, Wang CY (1997). Role of acetyltransferases in the metabolism and carcinogenicity of aromatic amines. *Mutation Research* **376:** 123-128.
- Kirkwood BR (1988). Essentials of medical statistics. Blackwell Science, Oxford, UK.
- Kiss I, Sándor J, Ember I (2000). Allelic polymorphism of *GSTM1* and *NAT2* genes modifies dietary-induced DNA damage in colorectal mucosa. *European Journal for Cancer Prevention* **9:** 429-432.
- Kiyokawa T, Umemoto T, Watanabe Y, Matsushita S, Shida H (1997). Two distinct pathways for intronless mRNA expression: One related, the other unrelated to human immunodeficiency virus Rev and human T-cell leukemia virus type I Rex functions. *Biological Signals* **6:** 134-142.
- Klein M, Pieri I, Uhlmann F, Pfizenmaier K, Eisel U (1998). Cloning and characterisation of promoter and 5′-UTR of the NMDA receptor subunit ε_2: Evidence for alternative splicing of 5′-non-coding exon. *Gene* **208:** 259-269.
- Klemme LM, Roberts KP, Hoffman LB, Ensrud KM, Siiteri JE, Hamilton DW (1999). Cloning and characterisation of the rat *Crisp-1* gene. *Gene* **240:** 279-288.
- Kloth MT, Gee RL, Messing EM, Swaminathan S (1994). Expression of N-acetyltransferase (NAT) in cultured human uroepithelial cells. *Carcinogenesis* **15:** 2781-2787.
- Klug A, Schwabe JWR (1995). Zinc fingers. *The FASEB Journal* **9:** 597-604.
- Knowles MA (1999). Identification of novel bladder tumour suppressor genes. *Electrophoresis* **20:** 269-279.
- Knowles MA, Shaw ME, Proctor AJ (1993). Deletion mapping of chromosome 8 in cancers of the urinary bladder using restriction fragment length polymorphism and microsatelite polymorphisms. *Oncogene* **8:** 1357-1364.
- Korb M, Ke Y, Johnson LF (1993). Stimulation of gene expression by introns: conversion of an inhibitory intron to a stimulatory intron by alteration of the splice donor sequence. *Nucleic Acids Research* **21:** 5901-5908.
- Kornberg TB (1993). Understanding the homeodomain. *Journal of Biological Chemistry* **268:** 26813-26816.
- Krajinovic M, Ghadirian P, Richer C, Sinnett H, Gandini S, Perret C, Lacroix A *et al.* (2001). Genetic susceptibility to breast cancer in French-Canadians: Role of carcinogen-metabolising enzymes and gene-environment interactions. *International Journal of Cancer* **92:** 220-225.

- Kuehn MR, Bradley A, Robertson EJ, Evans MJ (1987). A potential animal model for Lesch-Nyhan syndrome through introduction of HPRT mutations into mice. *Nature* **326:** 295-298.
- Kühn R, Schwenk F, Aguet M, Rajewsky K (1995). Inducible gene targeting in mice. *Science* **269:** 1427-1429.
- Kundu TK, Rao MRS (1999). CpG islands in chromatin organisation and gene expression. *Journal of Biochemistry* **125:** 217-222.
- La Du BN (1992). Overview of pharmacogenetics. In *Pharmacogenetics of Drug Metabolism*, edited by Kalow W. Pergamon Press, New York, USA; pp. 1-12.
- La Thangue NB, Rigby PWJ (1988). *Trans*-acting protein factors and the regulation of eukaryotic transcription. In *Transcription and Splicing*, edited by Hames BD and Glover DM. IRL Press, New York, USA; pp.97-129.
- Labuda D, Krajinovic M, Richer C, Skoll A, Sinnett H, Yotova V, Sinnett D (1999). Rapid detection of CYP1A1, CYP2D6 and NAT variants by multiplex polymerase chain reaction and allele-specific oligonucleotide assay. *Analytical Biochemistry* **275:** 84-92.
- Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH (1998). New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB. *Genes and Development* **12:** 34-44.
- Lamb P, McKnight SL (1991). Diversity and specificity in transcriptional regulation: The benefits of heterotypic dimerisation. *Trends in Biochemical Sciences* **16:** 417-422.
- Lan MS, Li Q, Lu J, Modi WS, Notkins AL (1994). Genomic organisation, 5′-upstream sequence, and chromosomal localisation of an insulinoma-associated intronless gene, IA-1. *Journal of Biological Chemistry* **269:** 14170-14174.
- Land SJ, Jones RE, King CM (1994). Biochemical and genetic analysis of two acetyltransferases from hamster tissues that can metabolise aromatic amine derivatives. *Carcinogenesis* **15:** 1585-1595.
- Lander ES (1999). Array of hope. *Nature Genetics* **21:** 3-4.
- Lang NP, Butler MA, Massengill J, Lawson M, Stotts RC, Hauer-Jensen M, Kadlubar FF (1994). Rapid metabolic phenotypes for acetyltransferase and cytochrome P4501A2 and putative exposure to food-borne heterocyclic amines increase the risk for colorectal cancer or polyps. *Cancer Epidemiology, Biomarkers and Prevention* **3:** 675-682.
- Lania L, Majello B, De Luca P (1997). Transcriptional regulation by the Sp family proteins. *International Journal of Biochemistry and Cell Biology* **29:** 1313-1323.
- Latchman D (1998a). Eukaryotic transcription factors. Academic Press, London, UK.
- Latchman D (1998b). Gene regulation: A eukaryotic perspective. Stanley Thornes Publishers, Cheltenham, UK.
- Latchman D, editor (1995). Transcription factors: A Practical Approach. Oxford University Press, Oxford, UK; pp. 1-47.

- Le Van Thai A, Coste E, Allen JM, Palmiter RD, Weber MJ (1993). Identification of a neuron-specific promoter of human aromatic L-amino acid decarboxylase gene. *Molecular Brain Research* **17:** 227-238.
- Lee EJD, Zhao B, Seow-Choen F (1998). Relationship between polymorphism of N-acetyltransferase gene and susceptibility to colorectal carcinoma in a Chinese population. *Pharmacogenetics* **8:** 513-517.
- Lee JH, Chung JG, Lai JM, Levy GN, Weber WW (1997). Kinetics of arylamine N-acetyltransferase in tissues from human breast cancer. *Cancer Letters* **111:** 39-50.
- Lee SS-T, Buters JTM, Pineau T, Fernandez-Salguero P, Gonzalez FJ (1996). Role of CYP2E1 in the hepatotoxicity of acetaminophen. *Journal of Biological Chemistry* **271:** 12063-12067.
- Lee TI, Young RA (2000). Transcription of eukaryotic protein-coding genes. *Annual Reviews of Genetics* **34:** 77-137.
- Lee TX, Johnson LF (1998). Pre-mRNA processing enhancer (PPE) element increases the expression of an intronless thymidylate synthase gene but does not affect intron-dependent S phase regulation. *Journal of Cellular Biochemistry* **69:** 104-116.
- Lee W, Haslinger A, Karin M, Tjian R (1987a). Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40. *Nature* **325:** 368-372.
- Lee W, Mitchell P, Tjian R (1987b). Purified transcription factor AP-1 interacts with TPA-inducible enhancer elements. *Cell* **49** 741-752.
- Leff MA, Epstein PN, Doll MA, Fretland AJ, Devanboyina U-S, Rustan TD *et al.* (1999a). Prostate-specific human N-acetyltransferase 2 (NAT2) expression in the mouse. *The Journal of Pharmacology and Experimental Therapeutics* **290:** 182-187.
- Leff MA, Fretland AJ, Doll MA, Hein DW (1999b). Novel human N-acetyltransferase 2 alleles that differ in mechanism for slow acetylator phenotype. *Journal of Biological Chemistry* **274:** 34519-34522.
- Lemon B, Tjian R (2000). Orchestrated response: A symphony of transcription factors for gene control. *Genes and Development* **14:** 2551-2569.
- Lemos MC, Regateiro FJ (1998). N-acetyltransferase genotypes in the Portugese population. *Pharmacogenetics* **8:** 561-564.
- Lerebours F, Olschwang S, Thuille B, Schmitz A, Fouchet P, Laurent-Puig P, Boman F *et al.* (1999). Deletion mapping of the tumour suppressor locus involved in colorectal cancer on chromosome band 8p21. *Genes Chromosomes and Cancer* **25:** 147-153.
- Levy GN, Chung J-G, Weber WW (1994). 2-Aminofluorene metabolism and DNA adduct formation by mononuclear leukocytes from rapid and slow acetylator mouse strains. *Carcinogenesis* **15:** 353-357.

- Levy GN, Martell KJ, De Leon JH, Weber WW (1992). Metabolic, molecular genetic and toxicological aspects of the acetylation polymorphism in inbred mice. *Pharmacogenetics* **2:** 197-206.
- Levy GN, Martell KJ, Weber WW (1993). Polymorphic N-acetylation of 2-aminofluorene by cell-free colon extracts from inbred mice. *Pharmacogenetics* **3:** 71-76.
- Levy GN, Rodgers L, Weber WW (1996). Effects of heredity on response to drugs and environmental chemicals: Construction of rodent models. *Chemical Research in Toxicology* **9:** 1215-1224.
- Levy GN, Weber WW (1989). 2-Aminofluorene-DNA adduct formation in acetylator congenic mouse lines. *Carcinogenesis* **10:** 705-709.
- Levy GN, Weber WW (1992). 2-Aminofluorene-DNA adducts in mouse urinary bladder: Effect of age, sex and acetylator phenotype. *Carcinogenesis* **13:** 159-164.
- Lewis BA, Kim T-K, Orkin SH (2000). A downstream element in the human β-globin promoter: Evidence of extended sequence-specific transcription factor IID contacts. *Proceedings of the National Academy of Sciences of the USA* **97:** 7172-7177.
- Li L, Chambard J-C, Karin M, Olson EN (1992). Fos and Jun repress transcriptional activation by myogenin and MyoD: the amino terminus of Jun can mediate repression. *Genes and Development* **5:** 676-689.
- Li YC, Hung CF, Yeh FT, Lin JP, Chung JG (2001). Luteolin-inhibited arylamine N-acetyltransferase activity and DNA-2-aminofluorene adduct in human and mouse leukemia cells. *Food Chemistry and Toxicology* **39:** 641-647.
- Lijnzaad P, Helgesen C, Rodriguez-Tomé P (1998). The Radiation Hybrid Database. *Nucleic Acids Research* **26:** 102-105.
- Lin HJ, Han CY, Lin BK, Hardy S (1993). Slow acetylator mutations in the human polymorphic N-acetyltransferase gene in 786 Asians, blacks, Hispanics and whites: application to metabolic epidemiology. *American Journal of Human Genetics* **52:** 827-834.
- Lin HJ, Han C-Y, Lin BK, Hardy S (1994). Ethnic distribution of slow acetylator mutations in the polymorphic N-acetyltransferase (*NAT2*) gene. *Pharmacogenetics* **4:** 125-134.
- Lin HJ, Probst-Hensch NM, Hughes NC, Sakamoto GT, Louie AD, Kau IH *et al.* (1998). Variants of N-acetyltransferase NAT1 and a case-control study of colorectal adenomas. *Pharmacogenetics* **8:** 269-281.
- Littlewood T, Evan G (1995). Helix-loop-helix. *Protein Profile* **2:** 621-653.
- Lo N-W, Lau JTY (1996). Novel heterogeneity exists in the 5'-untranslated region of the β-galactoside α2,6-sialyltransferase mRNAs in the human B-lymphoblastoid cell line, Louckes. *Biochemical and Biophysical Research Communications* **228:** 380-385.
- Lobe CG, Nagy A (1998). Conditional genome alteration in mice. *BioEssays* **20:** 200-208.

- Lo-Guidice J-M, Allorge D, Chevalier D, Debuysère H, Fazio F, Lafitte J-J, Broly F (2000). Molecular analysis of the N-acetyltransferase 1 gene (*NAT1**) using polymerase chain reaction-restriction fragment-single strand conformation polymorphism assay. *Pharmacogenetics* **10:** 293-300.
- Lopez AJ (1998). Alternative splicing of pre-mRNA: Developmental consequences and mechanisms of regulation. *Annual Reviews of Genetics* **32:** 279-305.
- Lorber B, Giegé R (1999). Biochemical aspects and handling of macromolrcular solutions and crystals. In *Crystallization of Nucleic Acids and Proteins: A Practical Approach*, edited by Ducruix A and Giegé R. Oxford University Press, Oxford UK; pp. 17-43.
- Lu AYH (1998). Drug-metabolism research challenges in the new millenium: Individual variability in drug therapy and drug safety. *Drug Metabolism and Disposition* **26:** 1217-1222.
- Lucock M (2000). Folic acid: Nutritional Biochemistry, Molecular Biology and role in disease processes. *Molecular Genetics and Metabolism* **71:** 121-138.
- Lugones LG, Scholtmeijer K, Klootwijk R, Wessels JGH (1999). Introns are necessary for mRNA accumulation in *Schizophyllum commune*. *Molecular Microbiology* **32:** 681-689.
- Lutchman M, Pack S, Kim AC, Azim A, Emmert-Buck M, Van Huffel C, Zhuang Z *et al.* (1999). Loss of heterozygosity on 8p in prostate cancer implicates a role for dematin in tumour progression. *Cancer Genetics and Cytogenetics* **115:** 65-69.
- Macé K, Bowman ED, Vautravers P, Shields PG, Harris CC, Pfeifer AMA (1998). Characterisation of xenobiotic-metabolising enzyme expression in human bronchial mucosa and peripheral lung tissues. *European Journal of Cancer* **34:** 914-920.
- Madiai F, Hackshaw KV, Chiu I-M (1999). Characterisation of the entire transcription unit of the mouse fibroblast growth factor 1 (FGF-1) gene: Tissue-specific expression of the FGF-1.A mRNA. *Journal of Biological Chemistry* **274:** 11937-11944.
- Maglott DR, Katz KS, Sicotte H, Pruitt KD (2000). NCBI's LocusLink and RefSeq. *Nucleic Acids Research* **28:** 126-128.
- Mahgoub A, Dring LG, Idle JR, Lancaster R, Smith RL (1977). Polymorphic hydroxylation of debrisoquine in man. *The Lancet* **i:** 584-586.
- Makeyev AV, Chkheidze AN, Liebhaber SA (1999). A set of highly conserved RNA-binding proteins, αCP-1 and αCP-2, implicated in the mRNA stabilisation, are coexpressed from an intronless gene and its intron-containing paralog. *Journal of Biological Chemistry* **274:** 24849-24857.
- Mallon A-M, Strivens M (1998). DNA sequence analysis and comparative sequencing. *Methods: A Companion to Methods in Enzymology* **14:** 160-178.
- Marcus PM, Hayes RB, Vineis P, Garcia-Closas M, Caporaso NE, Autrup H, Branch RA *et al.* (2000a). Cigarette smoking, N-acetyltransferase 2 acetylation

status, and bladder cancer risk: A case-series meta-analysis of a gene-environment interaction. *Cancer Epidemiology, Biomarkers and Prevention* **9:** 461-467.

- Marcus PM, Vineis P, Rothman N (2000b). NAT2 slow acetylation and bladder cancer risk: A meta-analysis of 22 case-control studies conducted in the general population. *Pharmacogenetics* **10:** 115-122.
- Marin M, Karis A, Visser P, Grosveld F, Philipsen S (1997). Transcription factor Sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. *Cell* **89:** 619-628.
- Marra M, Hillier L, Kucaba T, Allen M, Barstead R, Beck C, Blistain A *et al.* (1999). An encyclopedia of mouse genes. *Nature Genetics* **21:** 191-194.
- Marshall E (1999). A high-stakes gamble on genome sequencing. *Science* **284:** 1906-1909.
- Marshall E (2001). Celera assembles mouse genome; Public labs plan new strategy. *Science* **292:** 822-823.
- Martell KJ, Levy GN, Weber WW (1992). Cloned mouse N-acetyltransferases: Enzymatic Properties of Expressed *Nat-1* and *Nat-2* gene products. *Molecular Pharmacology* **42:** 265-272.
- Martell KJ, Vatsis KP, Weber WW (1991). Molecular genetic basis of rapid and slow acetylation in mice. *Molecular Pharmacology* **40:** 218-227.
- Mas C, Bourgeois F, Bulfone A, Levancher B, Mugnier C, Simonneau M (2000). Cloning and expression analysis of a novel gene, RP42, mapping to an autism susceptibility locus on 6q16. *Genomics* **65:** 70-74.
- Matas N, Thygesen P, Stacey M, Risch A, Sim E (1997). Mapping AAC1, AAC2 and AACP, the genes for arylamine N-acetyltransferases, carcinogen metabolising enzymes on human chromosome 8p22, a region frequently deleted in tumours. *Cytogenetics and Cell Genetics* **77:** 290-295.
- Matsuda Y, Chapman VM (1995). Application of fluorescence *in situ* hybridisation in genome analysis of the mouse. *Electrophoresis* **16:** 261-272.
- Mattano SS, Erickson RP, Nesbitt MN, Weber WW (1988). Linkage of *Nat* and *Es-1* in the mouse and development of strains congenic for N-acetyltransferase. *Journal of Heredity* **79:** 430-433.
- Mattano SS, Land S, King CM, Weber WW (1989). Purification and biochemical characterisation of hepatic arylamine N-acetyltransferase from rapid and slow acetylator mice: Identity with arylhydroxamic acid N,O-acyltransferase and N-hydroxyarylamine O-acetyltransferase. *Molecular Pharmacology* **35:** 599-609.
- Mattano SS, Weber WW (1987). Kinetics of arylamine N-acetyltransferase in tissues from rapid and slow acetylator mice. *Carcinogenesis* **8:** 133-137.
- May DG (1994). Genetic differences in drug disposition. *Journal of Clinical Pharmacology* **34:** 881-897.
- McCarthy L, Hunter K, Schalkwyk L, Riba L, Anson S, Mott R, Newell W *et al.* (1995). Efficient high-resolution genetic mapping of mouse interspersed repetitive sequence PCR products, toward integrated genetic and physical mapping of the

mouse genome. *Proceedings of the National Academy of Sciences of the USA* **92:** 5302-5306.

- McCracken S, Fong N, Yankulov K, Ballantyne S, Pan G, Greenblatt J, Patterson SD *et al.* (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription.
- McGurk KA, Brierley CH, Burchell B (1998). Drug glucuronidation by human renal UDP-glucuronosyltransferases. *Biochemical Pharmacology* **55:** 1005-1012.
- McLeod HL, Evans WE (2001). Pharmacogenomics: Unlocking the human genome for better drug therapy. *Annual Reviews in Pharmacology and Toxicology* **41:** 101-121.
- McNulty H, McPartlin JM, Weir DG, Scott JM (1993). Folate catabolism is increased during pregnancy in rats. *Journal of Nutrition* **123:** 1089-1093.
- McNulty H, McPartlin JM, Weir DG, Scott JM (1995). Folate catabolism is related to growth rate in weaning rats. *Journal of Nutrition* **125:** 99-103.
- McPartlin J, Courtney G, McNulty H, Weir D, Scott J (1992). The quantitative analysis of exogenous folate catabolites in human urine. *Analytical Biochemistry* **206:** 256-261.
- McPherson JD (1997). Sequence ready - or not? *Genome Research* **7:** 1111-1113.
- Meisler MH (1996). The role of the laboratory mouse in the Human Genome Project. *American Journal of Human Genetics* **59:** 764-771.
- Melton DW, McEwan C, McKie AB, Reid AM (1986). Expression of the mouse HPRT gene: Deletional analysis of the promoter region of an X-chromosome linked housekeeping gene. *Cell* **44:** 319-328.
- Mewes HW, Albermann K, Bähr M, Frishman D, Gleissner A, Hani J, Heumann K *et al.* (1997). Overview of the yeast genome. *Nature* **387:** 7-8.
- Meyer UA (2000). Pharmacogenetics and adverse drug reactions. *The Lancet* **356:** 1667-1671.
- Mezquita J, López-Ibor B, Pau M, Mezquita C (1993). Intron and intronless transcription of the chicken polybiquitin gene UbII. *FEBS Letters* **319:** 244-248.
- Millikan RC (2000). *NAT1*10* and *NAT1*11* polymorphisms and breast cancer risk. *Cancer Epidemiology, Biomarkers and Prevention* **9:** 217-219.
- Millikan RC, Pittman GS, Newman B, Tse C-KJ, Selmin O, Rockhill B, Savitz D *et al.* (1998). Cigarette smoking, N-acetyltransferases 1 and 2, and breast cancer risk. *Cancer Epidemiology, Biomarkers and Prevention* **7:** 371-378.
- Minchin RF (1995). Acetylation of p-amino benzoylglutamate, a folic acid catabolite, by recombinant arylamine N-acetyltransferase and U937 cells. *Biochemical Journal* **307:** 1-3.
- Minchin RF, Butcher NJ, Pope C (1998). Characterisation of the Human N-acetyltransferase type I gene promoter: Identification of ATCATTT repeats flanking an AP-1 like motif in the minimum promoter sequence. *Proceedings of the First International Workshop on the Arylamine N-Acetyltransferases*, October 22-24, Queensland, Australia; p.2.

- Minchin RF, Kadlubar FF, Ilett KF (1993). Role of acetylation in colorectal cancer. *Mutation Research* **290:** 35-42.
- Mitchell MK, Futscher BW, McQueen CA (1999). Developmental expression of N-acetyltransferases in C57Bl/6 mice. *Drug Metabolism and Disposition* **27:** 261-264.
- Mitchell PJ, Tjian R (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245:** 371-378.
- Mitchell RS, Bell JC (1957). Clinical implications of isoniazid blood levels in pulmonary tuberculosis. *New England Journal of Medicine* **257:** 1066-1070.
- Miyata M, Kudo G, Lee Y-H, Yang TJ, Gelboin HV, Fernandez-Salguero P, Kimura S *et al.* (1999). Targeted disruption of the microsomal epoxide hydrolase gene. *Journal of Biological Chemistry* **274:** 23963-23968.
- Monaco AP, Larin Z (1994). YACs, BACs, PACs and MACs: Artificial chromosomes as research tools. *Trends in Biotechnology* **12:** 280-286.
- Montminy M (1997). Transcriptional regulation by cyclic AMP. *Annual Reviews of Biochemistry* **66:** 807-822.
- Moore KJ (1999). Utilisation of mouse models in the discovery of human disease genes. *Drug Discovery Today* **4:** 123-128.
- Morabia A, Bernstein MS, Bouchardy I, Kurtz J, Morris MA (2000). Breast cancer and active and passive smoking: The role of the N-acetyltransferase 2 genotype. *American Journal of Epidemiology* **152:** 226-232.
- Moreadith RW, Radforrd NB (1997). Gene targeting in embryonic stem cells: The new physiology and metabolism. *Journal of Molecular Medicine* **75:** 208-216.
- MRC Vitamin Study Research Group (1991). Prevention of neural tube defects: Results of the Medical Research Council vitamin study. *The Lancet* **338:** 131-137.
- Mrozikiewicz PM, Drakoulis N, Roots I (1994). Polymorphic arylamine N-acetyltransferase (*NAT2*) genes in children with insulin-dependent diabetes mellitus. *Clinical Pharmacology and Therapeutics* **56:** 626-634.
- Müller MM, Gerster T, Schaffner W (1988). Enhancer sequences and the regulation of gene transcription. *European Journal of Biochemistry* **176:** 485-495.
- Müller U (1999). Ten years of gene targeting: Targeted mouse mutants, from vector design to phenotype analysis. *Mechanisms of Development* **82:** 3-21.
- Mummidi S, Ahuja SS, McDaniel BL, Ahuja SK (1997). The human CC chemokine receptor 5 (CCR5) gene: Multiple transcripts with 5′-end heterogeneity, dual promoter usage and evidence for polymorphisms within the regulatory regions and noncoding exons. *Journal of Biological Chemistry* **272:** 30662-30671.
- Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA *et al.* (2000). A whole-genome assembly of Drosophila. *Science* **287:** 2196-2204.
- Nagao M, Sugimura T (1993). Carcinogenic factors in food with relevance to colon cancer development. *Mutation Research* **290:** 43-51.

- Nagata K, Ozawa S, Miyata M, Shimada M, Yamazoe Y, Kato R (1994). Primary structure and molecular basis of polymorphic appearance of an acetyltransferase (AT-II) in hamsters. *Pharmacogenetics* **4:** 91-100.
- Nakada Y, Taniura H, Uetsuki T, Inazawa J, Yoshikawa K (1998). The human chromosomal gene for necdin, a neuronal growth suppressor, in the Prader-Willi syndrome deletion region. *Gene* **213:** 65-72.
- Nebert DW (1997). Polymorphisms in drug metabolising enzymes: What is their clinical relevance and why do they exist? *American Journal of Human Genetics* **60:** 265-271.
- Nebert DW, Bingham E (2001). Pharmacogenomics: Out of the lab and into the community. *Trends in Biotechnology* **19:** 519-523.
- Nebert DW, Duffy JJ (1997). How knockout mouse lines will be used to study the role of drug-metabolising enzymes and their receptors during reproduction and development, and in environmental toxicity, cancer and oxidative stress. *Biochemical Pharmacology* **53:** 249-254.
- Nebert DW, McKinnon RA, Puga A (1996). Human drug-metabolising enzyme polymorphisms: Effects on risk of toxicity and cancer. *DNA and Cell Biology* **15:** 273-280.
- Nehls M, Kyewski B, Messerle M, Waldschütz R, Schüddekopf K, Smith AJH, Boehm T (1996). Two genetically separable steps in the differentiation of thymic epithelium. *Science* **272:** 886-889.
- Nehls M, Messerle M, Sirulnik A, Smith AJH, Boehm T (1994). Two large insert vectors, λPS and λKO, facilitate rapid mapping and targeted disruption of mammalian genes. *BioTechniques* **17:** 770-775.
- Nelson DR, Koymans L, Kamataki T, Stegeman JJ, Feyereisen R, Waxman DJ, Waterman MR *et al.* (1996). P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* **6:** 1-42.
- Nerlov C, Ziff EB (1994). Three levels of functional interaction determine the activity of CCAAT/enhancer binding protein-α on the serum albumin promoter. *Genes and Development* **8:** 350-362.
- Nevins JR (1983). The pathway of eukaryotic mRNA formation. *Annual Reviews of Biochemistry* **52:** 441-466.
- Nilsson NE, Tryselius Y, Owman C (2000). Genomic organisation of the leukotriene B_4 receptor locus of human chromosome 14. *Biochemical and Biophysical Research Communications* **274:** 383-388.
- Nishioka Y (1995). The origin of common laboratory mice. *Genome* **38:** 1-7.
- Norton PA (1994). Alternative pre-mRNA splicing: factors involved in splice site selection. *Journal of Cell Science* **107:**1-7.
- Nusbaum C, Slonim DK, Harris KL, Birren BW, Steen RG, Stein LD, Miller J *et al.* (1999). A YAC-based physical map of the mouse genome. *Nature Genetics* **22:** 388-393.

- Nyberg F, Hou S-M, Hemminki K, Lambert B, Pershagen G (1998). Glutathione S-transferase μ1 and N-acetyltransferase 2 genetic polymorphisms and exposure to tobacco smoke in nonsmoking and smoking lung cancer patients and population controls. *Cancer Epidemiology, Biomarkers and Prevention* **7:** 875-883.
- O'Neil EA, Fletcher C, Burrow CR, Heintz N, Roeder RG, Kelly TJ (1988). Transcription factor OTF-1 is functionally identical to the DNA replication factor NF-III. *Science* **241:** 1210-1213.
- O'Neil WM, Gilfix BM, DiGirolamo A, Tsoukas CM, Wainer I (1997). N-acetylation among HIV-positive patients and patients with AIDS: When is fast, fast and slow, slow? *Clinical Pharmacology and Therapeutics* **62:** 261-271.
- Ohbayashi T, Schmidt EE, Makino Y, Kishimoto T, Nebeshima Y, Muramatsu M *et al.* (1996). Promoter structure of the mouse TATA-binding protein (TBP) gene. *Biochemical and Biophysical Research Communications* **225:** 275-280.
- Ohsako S, Deguchi T (1990). Cloning and expression of cDNAs for polymorphic and monomorphic arylamine N-acetyltransferases from human liver. *Journal of Biological Chemistry* **265:** 4630-4634.
- Ohsako S, Ohtomi M, Sakamoto Y, Uyemura K, Deguchi T (1988). Arylamine N-acetyltransferase from chicken liver: II. Cloning of cDNA and expression in Chinese hamster ovary cells. *Journal of Biological Chemistry* **263:** 7534-7538.
- Okkels H, Sigsgaard T, Wolf H, Autrup H (1997). Arylamine N-acetyltransferase 1 (NAT1) and 2 (NAT2) polymorphisms in susceptibility to bladder cancer: The influence of smoking. *Cancer Epidemiology, Biomarkers and Prevention* **6:** 225-231.
- Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, Beasley EM, Brady SD *et al.* (2001). A high-resolution radiation hybrid map of the human genome draft sequence. *Science* **291:** 1298-1302.
- Olson MV (2001). Clone by clone by clone. *Nature* **409:** 816-818.
- Omenn GS (2001). Prospects for pharmacogenetics and ecogenetics in the new millenium. *Drug Metabolism and Disposition* **29:** 611-614.
- Orlando V (2000). Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitations. *Trends in Biochemical Sciences* **25:** 99-104.
- Orphanides G, Lagrange T, Reinberg D (1996). The general transcription factors of RNA polymerase II. *Genes and Development* **10:** 2657-2683.
- Ouellette BFF, Boguski MS (1997). Database divisions and homology search files: A guide for the perplexed. *Genome Research* **7:** 952-955.
- Oyama T, Miyoshi Y, Koyama K, Nakagawa H, Yamori T, Ito T, Matsuda H *et al.* (2000). Isolation of a novel gene on 8p21.3-22 whose expression is reduced significantly in human colorectal cancers with liver metastasis. *Genes Chromosomes and Cancer* **29:** 9-15.
- Ozawa S, Abu-Zeid M, Kawakubo Y, Toyama S, Yamazoe Y, Kato R (1990). Monomorphic and polymorphic isozymes of arylamine N-acetyltransferases in

hamster liver: Purification of the isozymes and genetic basis of N-acetylation polymorphism. *Carcinogenesis* **11:** 2137-2144.

- Pacifici GM, Bencini C, Rane A (1986). Acetyltransferase in humans: Development and tissue distribution. *Pharmacology* **32:** 283-291.
- Pahl HL, Baeuerle PA (1996). Control of gene expression by proteolysis. *Current Opinions in Cell Biology* **8:** 340-347.
- Palmiter RD, Sandgren EP, Avarbock MR, Allen DD, Brinster RL (1991). Heterologous introns can enhance expression of transgenes in mice. *Proceedings of the National Academy of Sciences of the USA* **88:** 478-482.
- Palotie A, Heiskanen M, Laan M, Horelli-Kuitunen N (1996). High-resolution fluoresence *in situ* hybridisation: A new approach in genome mapping. *Annals of Medicine* **28:** 101-106.
- Patek PQ, Collins JL, Cohn M (1978). Transformed cell lines susceptible or resistant to *in vivo* surveillance against tumorigenesis. *Nature* **276:** 510-511.
- Payton M, Auty R, Delgoda R, Everett M, Sim E (1999a). Cloning and characterisation of arylamine N-aetyltransferase genes from *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*: Increased expression results in isoniazid resistance. *Journal of Bacteriology* **181:** 1343-1347.
- Payton M, Mushtaq A, Yu T-W, Wu L-J, Sinclair J, Sim E (2001). Eubacterial arylamine N-acetyltransferases-identification and comparison of 18 members of the protein family with conserved active site cysteine, histidine and aspartate residues. *Microbiology* **147:** 1137-1147.
- Payton M, Smelt V, Upton A, Sim E (1999b). A method for genotyping murine arylamine N-acetyltransferase type 2 (*NAT2*): A gene expressed in preimplantation embryonic stem cells encoding an enzyme acetylating the folate catabolite p-aminobenzoylglutamate. *Biochemical Pharmacology* **58:** 779-785.
- Payton MA, Sim E (1998). Genotyping human arylamine N-acetyltransferase type 1 (NAT1): The identification of two novel allelic variants. *Biochemical Pharmacology* **55:** 361-366.
- Peltonen L, McKusick VA (2001). Dissecting human disease in the postgenomic era. *Science* **291:** 1224-1229.
- Persson K, Heby O, Berger F (1999). The functional intronless S-adenosylmethionine decarboxylase gene of the mouse (*Amd-2*) is linked to the ornithine decarboxylase gene (*Odc*) on chromosome 12 and is present in distantly related species of the genus *Mus*. *Mammalian Genome* **10:** 784-788.
- Pfost DR, Boyce-Jacino MT, Grant DM (2000). A SNPshot: pharmacogenetics and the future of drug therapy. *Trends in Biotechnology* **18:** 334-338.
- Philips AV, Cooper TA (2000). RNA processing and human disease. *Cellular and Molecular Life Sciences* **57:** 235-249.
- Pinter K, Brook F, Johnson N, Payton M, Sim E (2001). Over-expression of human *NAT1* in the mouse. In *Proceedings of the Second International NAT Workshop*, 5-6 October 2001, Oxford, UK; p. T2.

- Plagge A, Kelsey G, Allen ND (2000). Directed mutagenesis in embryonic stem cells. In *Mouse Genetics and Transgenics: A Practical Approach*, edited by Jackson IJ and Abbott CM. Oxford University Press, New York, USA; pp. 247-284.
- Pompeo F, Brooke E, Kawamura A, Mushtaq A, Sim E (2002). The pharmacogenetics of NAT: Structural aspects. *Pharmacogenomics* **in press.**
- Probst-Hensch NM, Haile RW, Ingles SA, Longnecker MP, Han CY, Lin BK, Lee DB *et al.* (1995). Acetylation polymorphism and prevalence of colorectal adenomas. *Cancer Research* **55:** 2017-2020.
- Probst-Hensch NM, Haile RW, Li DS, Sakamoto GT, Louie AD, Lin BK, Frankl HD *et al.* (1996). Lack of association between the polyadenylation polymorphism in the NAT1 (acetyltransferase 1) gene and colorectal adenomas. *Carcinogenesis* **17:** 2125-2129.
- Proudfoot NJ (2000). Connecting transcription to messenger RNA processing. *Trends in Biochemical Science* **25:** 290-293.
- Proudfoot NJ, Whitelaw E (1988). Termination and 3′ end processing of eukaryotic RNA. In *Transcription and Splicing*, edited by Hames BD and Glover DM. IRL Press, New York, USA; pp.97-129.
- Pugh BF (1996). Mechanisms of transcription complex assembly. *Current Opinions in Cell Biology* **8:** 303-311.
- Radjendirane V, JosephP, Lee Y-H, Kimura S, Klein-Szanto AJP, Gonzalez FJ, Jaiswal AK (1998). Disruption of the DT diaphorase (NQO1) gene in mice leads to increased menadione toxicity. *Journal of Biological Chemistry* **273:** 7382-7389.
- Rafiq M, Suen CKM, Choudhury N, Joannou CL, White KN, Evans RW (1997). Expression of recombinant human ceruloplasmin-An absolute requirement for splicing signals in the expression cassette. *FEBS Letters* **407:** 132-136.
- Rampazzo A, Pivotto F, Occhi G, Tiso N, Bortoluzzi S, Rowen L *et al.* (2000). Characterisation of C14orf4, a novel intronless human gene containing a polyglutamine repeat, mapped to the ARVD1 critical region. *Biochemical and Biophysical Research Communications* **278:** 766-774.
- Rauscher FJ, Voulalas PJ, Franza BR, Curran T (1988). Fos and Jun bind cooperatively to the AP-1 site: Reconstitution *in vitro*. *Genes and Development* **2:** 1687-1699.
- Reeves PT, Minchin RF, Ilett KF (1988). Induction of sulfamethazine acetylation by hydrocortisone in the rabbit. *Drug Metabolism and Disposition* **16:** 110-115.
- Reeves PT, Minchin RF, Ilett KF (1989). *In vivo* mechanisms for the enhanced acetylation of sulfamethazine in the rabbit after hydrocortisone treatment. *Journal of Pharmacology and Experimental Therapeutics* **248:** 348-352.
- Rhodes M, Straw R, Fernando S, Evans A, Lacey T, Dearlove A, *et al.* (1998). A high-resolution microsatellite map of the mouse genome. *Genome Research* **8:** 531-542.

- Riddle B, Jencks WP (1971). Acetyl-coenzyme A: Arylamine N-acetyltransferase; Role of the acetyl-enzyme intermediate and the effects of substituents on the rate. *Journal of Biological Chemistry* **246:** 3250-3258.
- Riggins GJ, Strausberg RL (2001). Genome and genetic resources from the Cancer Genome Anatomy Project. *Human Molecular Genetics* **10:** 663-667.
- Rio DC (1993). Splicing of pre-mRNA: Mechanism, regulation and role in development. *Current Opinions in Genetics and Development* **3:** 574-584.
- Risch A, Smelt V, Lane D, Stanley L, van der Slot W, Ward A, Sim E (1996). Arylamine N-acetyltransferase in erythrocytes of cystic fibrosis patients. *Pharmacology and Toxicology* **4:** 235-240.
- Risch A, Wallace DMA, Bathers S, Sim E (1995). Slow acetylation genotype is a susceptibility factor in occupational and smoking related bladder cancer. *Human Molecular Genetics* **4:** 231-236.
- Roberts RM, Liu L, Guo Q, Leaman D, Bixby J (1998). The evolution of type I interferons. *Journal of Interferon and Cytokine Research* **18:** 805-816.
- Rocha L, Garcia C, de Mendoca A, Gil JP, Bishop DT, Lechner MC (1999). N-acetyltransferase (*NAT2*) genotype and susceptibility to sporadic Alzheimer's disease. *Pharmacogenetics* **9:** 9-15.
- Rodrigues-Lima F, Deloménie C, Goodfellow GH, Grant DM, Dupret J-M (2001). Homology modelling and structural analysis of human arylamine N-acetyltransferase NAT1. Evidence for the conservation of a cysteine protease catalytic domain and an active-site loop. *Biochemical Journal* **356:** 327-334.
- Rodriguez JW, Kirlin WG, Ferguson RJ, Doll MA, Gray K, Rustan TD, Kemp K, Urso P, Hein DW (1993). Human acetylator genotype: Relationship to colorectal cancer incidence and arylamine N-acetyltransferase expression in colon cytosol. *Archives of Toxicology* **67:** 445-452.
- Roeder RG (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences* **21:** 327-335.
- Roesler WJ, Vandenbark GR, Hanson RW (1988). Cyclic AMP and the induction of eukaryotic gene transcription. *Journal of Biological Chemistry* **263:** 9063-9066.
- Rogalla P, Kazmierczak B, Flohr AM, Hauke S, Bullerdiek J (2000). Back to the roots of a new exon-The molecular archaeology of a *SP100* splice variant. *Genomics* **63:** 117-122.
- Rosenberg MP (1997). Gene knockout and transgenic technologies in risk assessment: The next generation. *Molecular Carcinogenesis* **20:** 262-274.
- Rothman N, Hayes RB, Bi W, Caporaso N, Broly F, Woosley RL, Yin S *et al.* (1993). Correlation between N-acetyltransferase activity and *NAT2* genotype in Chinese males. *Pharmacogenetics* **3:** 250-255.
- Rothstein MA, Epps PG (2001). Ethical and legal implications of pharmacogenomics. *Nature Reviews in Genetics* **2:** 228-231.
- Rougraff PM, Paxton R (1987). Purification and partial characterisation of chicken liver acetyl coenzyme A:arylamine N-acetyltransferase. *Comparative Biochemistry and Physiology* **86B:** 601-606.

- Rubin GM (2001). Comparing species. *Nature* **409:** 820-821.
- Ryan AK, Rosenfeld MG (1997). POU domain family values: Flexibility, partnerships and developmental codes. *Genes and Development* **11:** 1207-1225.
- Sabath DE, Broome HE, Prystowsky MB (1990). Glyceraldehyde-3-phosphate dehydrogenase mRNA is a major interleukin 2-induced transcript in a cloned T-helper lymphocyte. *Gene* **91:** 185-191.
- Sadrieh N, Davis CD, Snyderwine EG (1996). N-acetyltansferase expression and metabolic activation of the food-derived heterocyclic amines in the human mammary gland. *Cancer Research* **56:** 2683-2687.
- Sambrook J, Fritch EF, Maniatis T (1989). Molecular Cloning: A Laboratory Manual. 2nd edition. Cold Spring Harbor Press, New York.
- Sandri-Goldin RM (1998). ICP27 mediates HSV RNA export by shuttling through a leucine-rich nuclear export signal and binding viral intronless RNAs through an RGG motif. *Genes and Development* **12:** 868-879.
- Sandy J, Mushtaq A, Kawamura A, Sim E, Noble M (2002). The structure of arylamine N-acetyltransferase from *Mycobacterium smegmatis*, an enzyme which inactivates the anti-tubercular drug isoniazid. *Journal of Molecular Biology* **in press.**
- Sasaki Y, Ohsako S, Deguchi T (1991). Molecular and genetic analysis of arylamine N-acetyltransferase of rabbit liver. *Journal of Biological Chemistry* **266:** 13243-13250.
- Sassone-Corsi P (1998). Coupling gene expression to cAMP signalling: Role of CREB and CREM. *International Journal of Biochemistry and Cell Biology* **30:** 27-38.
- Sauer B (1998). Inducible gene targeting in mice using the *Cre/lox* system. *Methods: A Companion to Methods in Enzymology* **14:** 381-392.
- Sawzdargo M, Nguyen T, Lee DK, Lynch KR, Cheng R, Heng HHQ *et al.* (1999). Identification and cloning of three novel human G protein-coupled receptor genes GPR52, ΨGPR53 and GPR55: GPR55 is extensively expressed in human brain. *Molecular Brain Research* **64:** 193-198.
- Schenborn ET, Mierendorf RC (1985). A novel transcription property of SP6 and T7 RNA polymerases: Dependence on template structure. *Nucleic Acids Research* **13:** 6223-6236.
- Schnakenberg E, Ehlers C, Feyerabend W, Werdin R, Hubotter R, Dreikorn K *et al.* (1998). Genotyping of the polymorphic N-acetyltransferase (NAT2) and loss of heterozygosity in bladder cancer patients. *Clinical Genetics* **53:** 396-402.
- Schöneberg T, Schulz A, Grosse R, Achade R, Henklein P, Schultz G *et al.* (1999). A novel subgroup of class I G-protein-coupled receptors. *Biochimica et Biophysica Acta* **1446:** 57-70.
- Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, *et al.* (1996). A gene map of the human genome. *Science* **274:** 540-546.

- Schwabe JWR, Rhodes D (1991). Beyond zinc fingers: Steroid hormone receptors have a novel structural motif for DNA recognition. *Trends in Biochemical Sciences* **16:** 291-296.
- Scott JM, McPartlin J, Molloy A, McNulty H, Halligan A, Darling M, Weir DG (1993). Folate metabolism in pregnancy. In *Chemistry and Biology of Pteridines and Folates*, edited by Ayling JF. Plenum Press, New York, USA; 727-732.
- Sedlacek Z, Münstermann E, Dhorne-Pollet S, Otto C, Bock D, Schütz G *et al.* (1999). Human and mouse *XAP-5* and *XAP-5-like* (*X5L*) genes: Identification of an ancient functional retrosposon differentially expressed in testis. *Genomics* **61:** 125-132.
- Sekine A, Saito S, Iida A, Mitsunobu Y, Higuchi S, Harigae S, Nakamura Y (2001). Identification of single-nucleotide polymorphisms (SNPs) of human N-acetyltransferase genes NAT1, NAT2, AANAT, ARD1 and L1CAM in the Japanese population. *Journal of Human Genetics* **46:** 314-319.
- Shibata K, Itoh M, Aizawa K, Nagaoka S, Sasaki N, Carninci P, Konno H *et al.* (2000). RIKEN Integrated Sequence Analysis System (RISA)—384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Research* **10:** 1757-1771.
- Shibuta K, Nakashima T, Abe M, Mashimo M, Mori M, Ueo H, Akiyoshi T *et al.* (1994). Molecular genotyping for N-acetylation polymorphism in Japanese patients with colorectal cancer. *Cancer* **74:** 3108-3112.
- Shikama N, Lyon J, La Thangue NB (1997). The p300/CBP family: integrating signals with transcription factors and chromatin. *Trends in Cell Biology* **7:** 230-236.
- Shishikura K, Hohjoh H, Tokunaga K (2000). Novel allele containing a 190C>T nonsynonymous substitution in the N-acetyltransferase (NAT2) gene. *Human Mutation* **5:** 581.
- Shivdasani RA, Fujiwara Y, McDevitt MA, Orkin SH (1997). A lineage-selective knockout establishes the critical role of transcription factor GATA-1 in megakaryocyte growth and platelet development. *The EMBO Journal* **16:** 3965-3973.
- Shizuya H, Birren B, Kim U-J, Mancino V, Slepak T, Tachiiri Y, Simon M (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the USA* **89:** 8794-8797.
- Sim E, Hickman D (1991). Polymorphism in human N-acetyltransferase - the case of the missing allele. *Trends in Pharmacological Sciences* **12:** 211-213.
- Sim E, Payton M, Noble M, Minchin R (2000). An update on genetic, structural and functional studies of arylamine N-acetyltransferases in eukaryotes and prokaryotes. *Human Molecular Genetics* **9:** 2435-2441.
- Sim E, Stanley LA, Risch A, Thygesen P (1995). Xenogenetics in multifactorial disease susceptibility. *Trends in Genetics* **11:** 509-512.
- Sinclair JC, Delgoda R, Noble MEM, Jarmin S, Goh NK, Sim E (1998). Purification, characterisation and crystallisation of an N-hydryarylamine O-

acetyltransferase from *Salmonella typhimurium*. *Protein Expression and Purification* **12:** 371-380.

- Sinclair JC, Sandy J, Delgoda R, Sim E, Noble MEM (2000). Structure of arylamine N-acetyltransferase reveals a catalytic triad. *Nature Structural Biology* **7:** 560-564.
- Sinclair JC, Sim E (1997). A fragment consisting of the first 204 amino-terminal amino acids of human arylamine N-acetyltransferase one (NAT1) and the first transacetylation step of catalysis. *Biochemical Pharmacology* **53:** 11-16.
- Skoda RC, Gonzalez FJ, Demierre A, Meyer UA (1988). Two mutant alleles of the human cytochrome P450dbl gene (P450C2D1) associated with genetically deficient metabolism of debrisoquine and other drugs. *Proceedings of the National Academy of Sciences of the USA* **85:** 5240-5243.
- Smaglik P, Abbott A (2000). Project offers free mouse sequence. *Science* **407:** 663-664.
- Smelt VA, Mardon HJ, Redman CWG, Sim E (1997). Acetylation of arylamines by the placenta. *European Journal of Drug Metabolism and Pharmacokinetics* **22:** 403-408.
- Smelt VA, Mardon HJ, Sim E (1998). Placental expression of arylamine N-acetyltransferases: Evidence for linkage disequilibrium between *NAT1*10* and *NAT2*4* alleles of the two human arylamine N-acetyltransferase loci *NAT1* and *NAT2*. *Pharmacology and Toxicology* **83:**149-157.
- Smelt VA, Upton A, Adjaye J, Payton MA, Boukouvala S, Johnson N, Mardon HJ, Sim E (2000). Expression of arylamine N-acetyltransferases in pre-term placentas and in human pre-implantation embryos. *Human Molecular Genetics* **9:** 1101-1107.
- Smigielski EM, Sirotkin K, Ward M, Sherry ST (2000). dbSNP: A database of single nucleotide polymorphisms. *Nucleic Acids Research* **28:** 352-355.
- Smith CAD, Smith G, Wolf CR (1994). Genetic polymorphisms in xenobiotic metabolism. *European Journal of Cancer* **30A:** 1921-1935.
- Smith CAD, Wadelius M, Gough AC, Harrison DJ, Wolf CR, Rane A (1997). A simplified assay for the arylamine N-acetyltransferase 2 polymorphism validated by phenotyping with isoniazid. *Journal of Medical Genetics* **34:** 758-760.
- Smith CWJ, Patton JG, Nadal-Ginard B (1989). Alternative splicing in the control of gene expression. *Annual Reviews in Genetics* **23:** 527-577.
- Smith CWJ, Valcárcel J (2000). Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends in Biochemical Sciences* **25:** 381-388.
- Smolen TN, Brewer JA, Weber WW (1993). Testosterone modulation of N-acetylation in mouse kidney. *The Journal of Pharmacology and Experimental Therapeutics* **264:** 854-858.
- Sonenberg N (1994). mRNA translation: Influence of the 5′ and 3′ untranslated regions. *Current Opinions in Genetics and Development* **4:** 310-315.
- Sosinsky A, Glusman G, Lancet D (2000). The genomic structure of human olfactory receptor genes. *Genomics* **70:** 49-61.

- Southern E, Mir K, Shchepinov M (1999). Molecular interactions on microarrays. *Nature Genetics* **21:** 5-9.
- Spurr NK, Gough AC, Chinegwundoh FI, Smith CA (1995). Polymorphisms in drug-metabolising enzymes as modifiers of cancer risk. *Clinical Chemistry* **41:** 1864-1869.
- Stacey M, Matas N, Drake M, Payton M, Fakis G, Greenland J, Sim E (1999). Arylamine N-acetyltransferase type 2 (NAT2), chromosome 8 aneuploidy, and identification of a novel *NAT1* cosmid clone: An investigation in bladder cancer by interphase FISH. *Genes, Chromosomes and Cancer* **25:** 376-383.
- Stacey M, Payton M, Fakis G, Pope J, Sim E (1998). Characterisation of arylamine N-acetyltransferase in the human urothelial cell line RT112. *Pathogenesis* **1:** 99-105.
- Stacey M, Thygesen P, Stanley LA, Matas N, Risch A, Sim E (1996). Arylamine N-acetyltransferase as a potential biomarker in bladder cancer: Fluorescent *in situ* hybridisation and immunohistochemistry studies. *Biomarkers* **1:** 55-61.
- Stanley LA, Copp AJ, Pope J, Rolls S, Smelt V, Perry VH *et al.* (1998). Immunochemical detection of arylamine N-acetyltransferase during mouse embryonic development and in adult mouse brain. *Teratology* **58:** 174-182.
- Stanley LA, Coroneos E, Cuff R, Hickman D, Ward A, Sim E (1996). Immunochemical detection of arylamine N-acetyltransferase in normal and neoplastic bladder. *The Journal of Histochemistry and Cytochemistry* **44:** 1059-1067.
- Stanley LA, Mills IG, Sim E (1997). Localisation of polymorphic N-acetyltransferase (NAT2) in tissues of inbred mice. *Pharmacogenetics* **7:** 121-130.
- Stojdl DF, Bell JC (1999). SR protein kinases: The splice of life. *Biochemistry and Cell Biology* **77:** 293-298.
- Strachan T, Abitbol M, Davidson D, Beckmann JS (1997). A new dimension for the human genome project: Towards comprehensive expression maps. *Nature Genetics* **16:** 126-132.
- Strahl BD, Allis CD (2000). The language of covalent histone modifications. *Nature* **403:** 41-45.
- Struhl K (1989). Helix-turn-helix, zinc-finger and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. *Trends in Biochemical Sciences* **14:** 137-140.
- Struhl K (1994). Duality of TBP, the universal transcription factor. *Science* **263:** 1103-1104.
- Sugiyama A, Noguchi K, Kitanaka C, Katou N, Tashiro F, Ono T *et al.* (1999). Molecular cloning and chromosomal mapping of mouse intronless *myc* gene acting as a potent apoptosis inducer. *Gene* **226:** 273-283.
- Sun PC, Uppaluri R, Schmidt AP, Pashia ME, Quant EC, Sunwoo JB, Collin SM *et al.* (2001). Transcript map of the 8p23 putative tumour suppressor region. *Genomics* **75:** 17-25.
- Surridge C (1996). The core curriculum. *Nature* **380:** 287-288.

- Svejstrup JQ, Vichi P, Egly J-M (1996). The multiple roles of transcription/repair factor TFIIH. *Trends in Biochemical Sciences* **21:** 346-350.
- Taioli E (1999). International collaborative study on genetic susceptibility to environmental carcinogens. *Cancer Epidemiology, Biomarkers and Prevention* **8:** 727-728.
- Tanaka A, Mita S, Ohta S, Kyozuka J, Shimamoto K, Nakamura K (1990). Enhancement of foreign gene expression by a didcot intron in rice but not in tobacco is correlated with an increased level of mRNA and an efficient splicing of the intron. *Nucleic Acids Research* **18:** 6767-6770.
- Tavianini MA, Hayes TE, Magazin MD, Minth CD, Dixon JE (1984). Isolation, characterisation and DNA sequence of the rat somatostatin gene. *Journal of Biological Chemistry* **259:** 11798-11803.
- Taylor BA (2000). Mapping phenotypic trait loci. In *Mouse Genetics and Transgenics: A Practical Approach*, edited by Jackson IJ and Abbott CM. Oxford University Press, New York, USA; pp. 87-120.
- Taylor JA, Umbach DM, Stephens E, Castranio T, Paulson D, Robertson C, Mohler JL *et al.* (1998). The role of N-acetylation polymorphisms in smoking-associated bladder cancer: Evidence of a gene-gene-exposure three-way interaction. *Cancer Research* **58:** 3603-3610.
- Thanos D, Maniatis T (1995). NF-κB: A lesson in family values. *Cell* **80:** 529-532.
- The *Arabidopsis* Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796-815.
- The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012-2018.
- The International Human Genome Mapping Consortium (2001). A physical map of the human genome. *Nature* **409:** 934-941.
- The International Mouse Mutagenesis Consortium (2001). Functional annotation of mouse genome sequences. *Science* **291:** 1251-1255.
- The International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928-933.
- The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium (2001). Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685-690.
- Thomas KR, Capecchi MR (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell* **51:** 503-512.
- Thygesen P, Risch A, Stacey M, Fakis G, Takle L, Knowles M, Sim E (1999). Genes for arylamine N-acetyltransferases in relation to loss of the short arm of chromosome 8 in bladder cancer. *Pharmacogenetics* **9:** 1-8.

- Tomarev SI, Piatigorsky J (1996). Lens crystallins of invertebrates: Diversity and recruitment from detoxification enzymes and novel proteins. *European Journal of Biochemistry* **235:** 449-465.
- Trepanier LA, Cribb A, Spielberg SP, Ray K (1998). Deficiency of cytosolic arylamine N-acetylation in the domestic cat and wild felids caused by the presence of a single *NAT1*-like gene. *Pharmacogenetics* **8:** 169-179.
- Trepanier LA, Ray K, Winand NJ, Spielberg SP, Cribb AE (1997). Cytosolic arylamine N-acetyltransferase (NAT) deficiency in the dog and other canids due to an absence of *NAT* genes. *Biochemical Pharmacology* **54:** 73-80.
- Trubetskoy AM, Okenquist SA, Lenz J (1999). R region sequences in the long terminal repeat of a murine retrovirus specifically increase expression of unspliced RNAs. *Journal of Virology* **73:** 3477-3483.
- Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J *et al.* (2002). Rat Genome Database (RGD): Mapping disease onto the genome. *Nucleic Acids Research* **30:** 125-128.
- Ueland PM, Hustad S, Schneede J, Refsum H, Vollset SE (2001). Biological and clinical implications of the MTHFR C677T polymorphism. *Trends in Pharmacological Sciences* **22:** 195-201.
- Upton A, Sim E (1999). Presence of an arylamine N-acetyltransferase gene homologue and N-acetylation activity in domestic cattle. *Biochemical Society Meeting*, Cork, Ireland.
- Upton A, Smelt V, Mushtaq A, Aplin R, Johnson N, Mardon H *et al.* (2000). Placental arylamine N-acetyltransferase type 1: potential contributory source of urinary folate catabolite p-acetamidobenzoylglutamate during pregnancy. *Biochimica et Biophysica Acta* **1524:** 143-148.
- Upton AM, Mushtaq A, Victor TC, Sampson SL, Sandy J, Smith D-M, van Helden PV *et al.* (2001). Arylamine N-acetyltransferase of *Mycobacterium tuberculosis* is a polymorphic enzyme and a site of isoniazid metabolism. *Molecular Microbiology* **42:** 309-317.
- Vainio H and McGregor D (1992). Approaches to the prediction of human cancer risk. *Pharmacogenetics* **2:** 337-343.
- Valentine JL, Lee SS-T, Seaton MJ, Asgharian B, Farris G, Corton JC, Gonzalez FJ *et al.* (1996). Reduction of benzene metabolism and toxicity in mice that lack CYP2E1 expression. *Toxicology and Applied Pharmacology* **141:** 205-213.
- Van Alewijk DC, Van der Weiden MM, Eussen BJ, Van Den Andel-Thijssen LD, Ehren-van Eskelen CC, Konig JJ, Van Steenbrugge GJ *et al.* (1999). Identification of a homozygous deletion at 8p12-21 in a human prostate cancer xenograft. *Genes Chromosomes and Cancer* **24:** 119-126.
- Van Etten WJ, Steen RG, Nguyen H, Castle AB, Slonim DK, Ge B, Nusbaum C, *et al.* (1999). Radiation hybrid map of the mouse genome. *Nature Genetics* **22:** 384-387.
- Vandromme M, Gauthier-Rouvière C, Lamb N, Fernandez A (1996). Regulation of transcription factor localisation: fine-tuning of gene expression. *Trends in Biochemical Sciences* **21:** 59-64.

- Vatsis KP, Martell KJ, Weber WW (1991). Diverse point mutations in the human gene for polymorphic N-acetyltransferase. *Proceedings of the National Academy of Sciences of the USA* **88:** 6333-6337.
- Vatsis KP, Weber WW (1993). Structural heterogeneity of Caucasian N-acetyltranferase at the NAT1 gene locus. *Archives of Biochemistry and Biophysics* **301:** 71-76.
- Vatsis KP, Weber WW (1994). Human N-acetyltransferases. In *Conjugation-Deconjugation Reactions in Drug Metabolism and Toxicity: Handbook of Experimental Pharmacology,* edited by Kaufmann FC. Springer-Verlag, Berlin, Germany; vol. 112, pp. 109-130.
- Vatsis KP, Weber WW, Bell DA, Dupret J-M, Price-Evans DA, Grant DM *et al.* (1995). Nomenclature for N-acetyltransferases. *Pharmacogenetics* **5:** 1-17.
- Vaziri SJ, Hughes NC, Sampson H, Darlington G, Jewett MAS, Grant DM (2001). Variation in enzymes of arylamine procarcinogen biotransformation among bladder cancer patients and control subjects. *Pharmacogenetics* **11:** 7-20.
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995). Serial analysis of gene expression. *Science* **270:** 484-487.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO *et al.* (2001). The sequence of the human genome. *Science* **291:** 1304-1351.
- Verma IM, Stevenson JK, Schwarz EM, Van Antwerp D, Miyamoto S (1995). Rel/NF-κB/IκB family: Intimate tales of association and dissociation. *Genes and Development* **9:** 2723-2735.
- Verrijzer CP, Tjian R (1996). TAFs mediate transcriptional activation and promoter selectivity. *Trends in Biochemical Sciences* **21:** 338-342.
- Verrijzer CP, Van der Vliet C (1993). POU domain transcription factors. *Biochimica et Biophysica Acta* **1173:** 1-21.
- Vineis P, Bartsch H, Caporaso N, Harrington AM, Kadlubar FF, Landi MT, Malaveille C *et al.* (1994). Genetically based N-acetyltransferase metabolic polymorphism and low-level environmental exposure to carcinogens. *Nature* **369:** 154-156.
- Wade PA, Pruss D, Wolffe AP (1997). Histone acetylation: Chromatin in action. *Trends in Biochemical Sciences* **22:** 128-132.
- Wadelius M, Autrup JL, Stubbins MJ, Andersson S-O, Johansson J-E, Wadelius C, Wolf CR *et al.* (1999). Polymorphisms in *NAT2, CYP2D6, CYP2C19* and *GSTP1* and their association with prostate cancer. *Pharmacogenetics* **9:** 333-340.
- Wang JC, Radford DM, Holt MS, Helms C, Goate A, Brandt W , Parik M, Phillips NJ, DeSchyver K, Schuh ME, Fair KL, Ritter JH, Marshall P, Donnis-Keller H (1999). Sequence-ready contig for the 1.4-cM ductal carcinoma *in situ* loss of heterozygocity region on chromosome 8p22-p23. *Genomics* **60:** 1-11.
- Ward A, Hickman D, Gordon JW, Sim E (1992). Arylamine N-acetyltransferase in human red blood cells. *Biochemical Pharmacology* **44:** 1099-1104.

- Ward A, Summers MJ, Sim E (1995). Purification of recombinant human N-acetyltransferase type 1 (NAT1) expressed in *E. coli* and characterisation of its potential role in folate metabolism. *Biochemical Pharmacology* **49:** 1759-1767.
- Ware JA, Svensson CK (1996). Longitudinal distribution of arylamine N-acetyltransferases in the intestine of the hamster, mouse, and rat. *Biochemical Pharmacology* **52:** 1613-1620.
- Watanabe M, Sofuni T, Nohmi T (1992). Involvement of Cys69 residue in the catalytic mechanism of N-hydroxyarylamine O-acetyltransferase of *Salmonella typhimurium*. Sequence similarity at the amino acid level suggests a common catalytic mechanism of acetyltransferase for *S. typhimurium* and higher organisms. *Journal of Biological Chemistry* **267:** 8429-8436.
- Watson AP, Wang PD, Sim E (1990). Arylamine N-acetyltransferase from fast (C57BL6) and slow (A/J) N-acetylating strains of mice. *Biochemical Pharmacology* **39:** 647-654.
- Watson JD (1990). The Human Genome Project: Past, present and future. *Science* **248:** 44-49.
- Weber WW, Cohen SN (1967). N-acetylation of drugs: Isolation and properties of an N-acetyltransferase from rabbit liver. *Molecular Pharmacology* **3:** 266-273.
- Weber WW, Hein DW (1985). N-acetylation pharmacogenetics. *Pharmacological Reviews* **37**: 25-78.
- Weber WW, Vatsis KP (1993). Individual variability in p-aminobenzoic acid N-acetylation by human N-acetyltransferase (NAT1) of peripheral blood. *Pharmacogenetics* **3:** 209-212.
- Webster NJG, Huang Z (1999). Hormonal regulation of alternative splicing. *Frontiers in Hormone Research* **25:** 1-17.
- Weis L, Reinberg D (1992). Transcription by RNA polymerase II: Initiator-directed formation of transcription-competent complexes. *The FASEB Journal* **6:** 3300-3309.
- Weisburger JH (1997). A perspective on the history and significance of carcinogenic and mutagenic N-substituted aryl compounds in human health. *Mutation Research* **376:** 261-266.
- Welfare MR, Cooper J, Bassendine MF, Daly AK (1997). Relationship between acetylator status, smoking and diet and colorectal cancer risk in the north-east of England. *Carcinogenesis* **18:** 1351-1354.
- West DB, Iakougova O, Olsson C, Ross D, Ohmen J, Chatterjee A (2000). Mouse genetics/genomics: An effective approach for drug target discovery and validation. *Medical Research Reviews* **20:** 216-230.
- Westphal GA, Reich K, Schultz TG, Neumann C, Hallier E, Schnuch A (2000). N-acetyltransferase 1 and 2 polymorphisms in para-substituted arylamine-induced contact allergy. *British Journal of Dermatology* **142:** 1121-1127.
- Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA *et al.* (2000). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **28:** 10-14.

- Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD *et al.* (2001). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **29:** 11-16.
- Wikman H, Thiel S, Jäger B, Schmezer P, Spiegelhalder B, Edler L, Dienemann H *et al.* (2001). Relevance of N-acetyltranferase 1 and 2 (NAT1, NAT2) genetic polymorphisms in non-small cell lung cancer susceptibility. *Pharmacogenetics* **11:** 157-168.
- Williams JA, Phillips DH (2000). Mammary Expression of xenobiotic metabolising enzymes and their potential role in breast cancer. *Cancer Research* **60:** 4667-4677.
- Williams JA, Stone AM, Fakis G, Johnson N, Cordell JA, Meinl W *et al.* (2001). N-acetyltransferases, sulfotransferases and heterocyclic amine activation in the breast. *Pharmacogenetics* **11:**371-388.
- Williams RS, Wagner PD (2000). Transgenic animals in integrative biology: Approaches and interpretations outcome. *Journal of Applied Physiology* **88:** 1119-1126.
- Williams RT (1959). The metabolism and detoxication of drugs, toxic substances and other organic compounds. In *Detoxication Mechanisms*. Champman & Hall, London, UK; pp. 734-740.
- Windle B, Silvas E, Parra I (1995). High resolution microscopic mapping of DNA using multicolour fluorescent hybridisation. *Electrophoresis* **16:** 273-278.
- Windmill KF, McKinnon RA, Zhu X, Gaedigk A, Grant DM, McManus ME (1997). The role of xenobiotic metabolising enzymes in arylamine toxicity and carcinogenesis: Functional and localisation studies. *Mutation Research* **376:** 153-160.
- Windridge S, Ball MT, Copp AJ, Hobart MJ, Payton M, Sim E, Stanley LA (2001). Expression of NAT1, NAT2 and NAT3 in the *Splotch* mouse embryo: Association with neural tube defects. In *Proceedings of the Second International NAT Workshop*, 5-6 October 2001, Oxford, UK; p. T3.
- Wong FW-Y, Chan W-Y, Lee SS-T (1998). Resistance to carbon tetrachloride-induced hepatotoxicity in mice which lack CYP2E1 expression. *Toxicology and Applied Pharmacology* **153:** 109-118.
- Wong L-L (1998). Cytochrome P450 monooxygenases. *Current Opinions in Chemical Biology* **2:** 263-268.
- Woolhouse NM, Qureshi MM, Bastaki SMA, Patel M, Abdulrazzaq Y, Bayoumi AL (1997). Polymorphic N-acetyltransferase (NAT2) genotyping of Emiratis. *Pharmacogenetics* **7:** 73-82.
- Yang M, Katoh T, Delongchamp R, Ozawa S, Kohshi K, Kawamoto T (2000). Relationship between *NAT1* genotype and phenotype in a Japanese population. *Pharmacogenetics* **10:** 225-232.
- You Y, Browning VL, Schimenti JC (1997). Generation of radiation-induced deletion complexes in the mouse genome, using embryonic stem cells. *Methods: A Companion to Methods in Enzymology* **13:** 409-421.

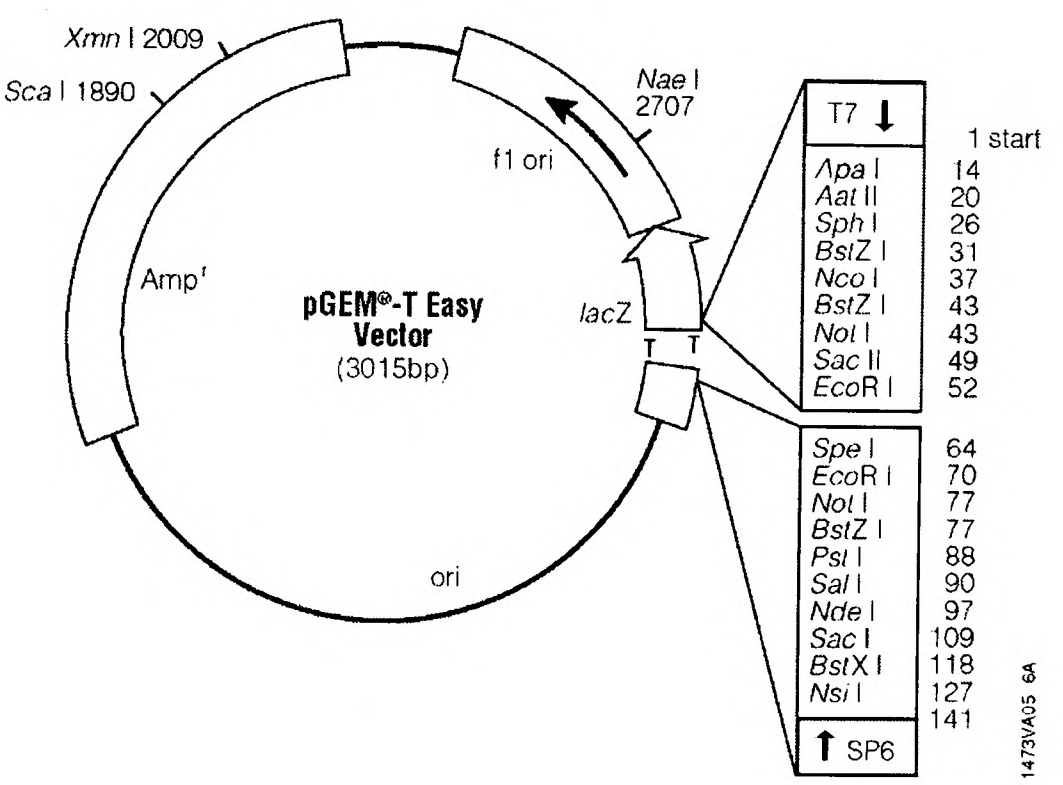
- Zaher H, Svensson CK (1994). Glucocorticoid induction of hepatic acetyl CoA:arylamine N-acetyltransferase activity in the rat. *Research Communications in Chemical Pathology and Pharmacology* **83:** 195-208.
- Zaphiropoulos PG (1998). Mechanisms of pre-mRNA splicing: Classical versus non-classical pathways. *Histology and Histopathology* **13:** 585-589.
- Zhang L, Wong SC, Matherly LH (1998). Transcript heterogeneity of the human reduced folate carrier results from the use of multiple promoters and variable splicing of alternative upstream exons. *Biochemical Journal* **332:** 733-780.
- Zhao B, Lee E, Yeoh PN, Gong NH (1998). Detection of mutations and polymorphism of N-acetyltransferase 1 gene in Indian, Malay and Chinese populations. *Pharmacogenetics* **8:** 299-304.
- Zhao J, Hyman L, Moore C (1999). Formation of mRNA 3′ ends in eukaryotes: Mechanism, regulation and interrelationships with other steps in mRNA synthesis. *Microbiology and Molecular Biology Reviews* **63:** 405-445.
- Zheng W, Deitz AC, Campbell DR, Wen W-Q, Cerhan JR, Sellers TA, Folsom AR *et al.* (1999). N-acetyltransferase 1 genetic polymorphism, cigarette smoking, well-done meat intake and breast cancer risk. *Cancer Epidemiology, Biomarkers and Prevention* **8:** 233-239.
- Zhou T, Chiang C-M (2001). The intronless and TATA-less human *TAF$_{II}$55* gene contains a functional initiator and a downstream promoter element. *Journal of Biological Chemistry* **276:** 25503-25511.
- Zimmermann N, Daugherty BL, Kavanaugh JL, El-Awar FY, Moulton EA, Rothenberg ME (2000). Analysis of the CC chemokine receptor 3 gene reveals a complex 5′ exon organisation, a functional role for untranslated exon 1, and a broadly active promoter with eosinophil-selective elements. *Blood* **96:** 2346-2354.

APPENDIX 1

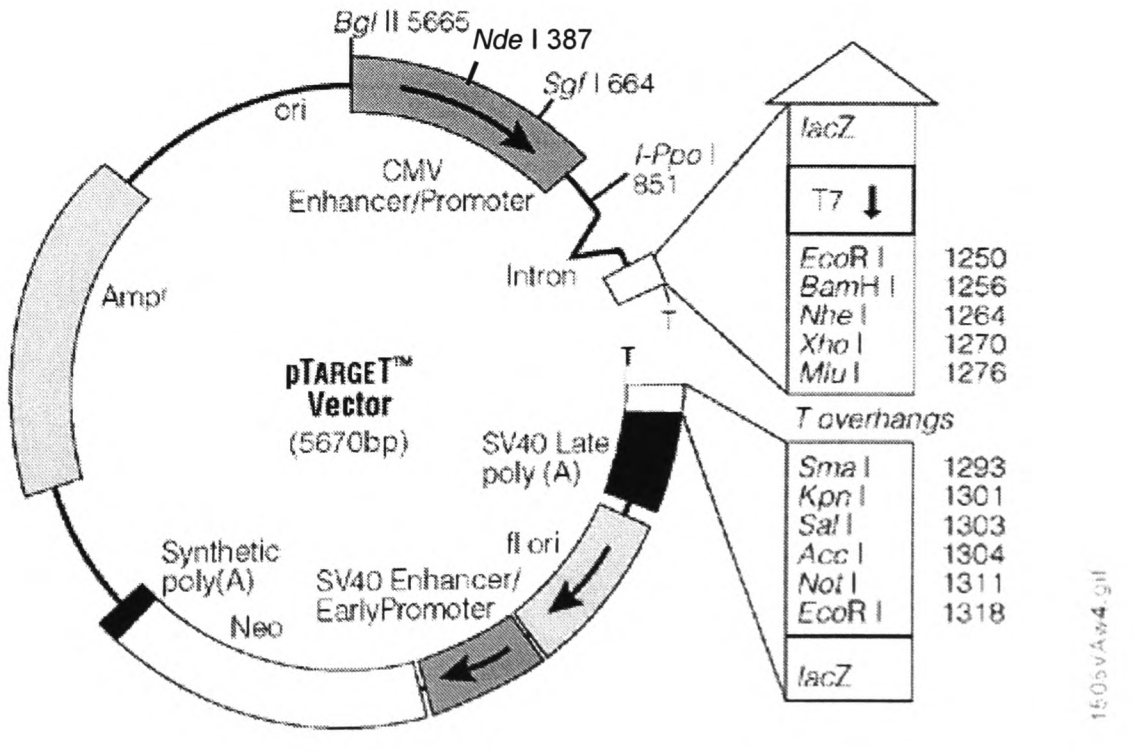
Cloning vectors

1.1 T-tailed vectors

(a)



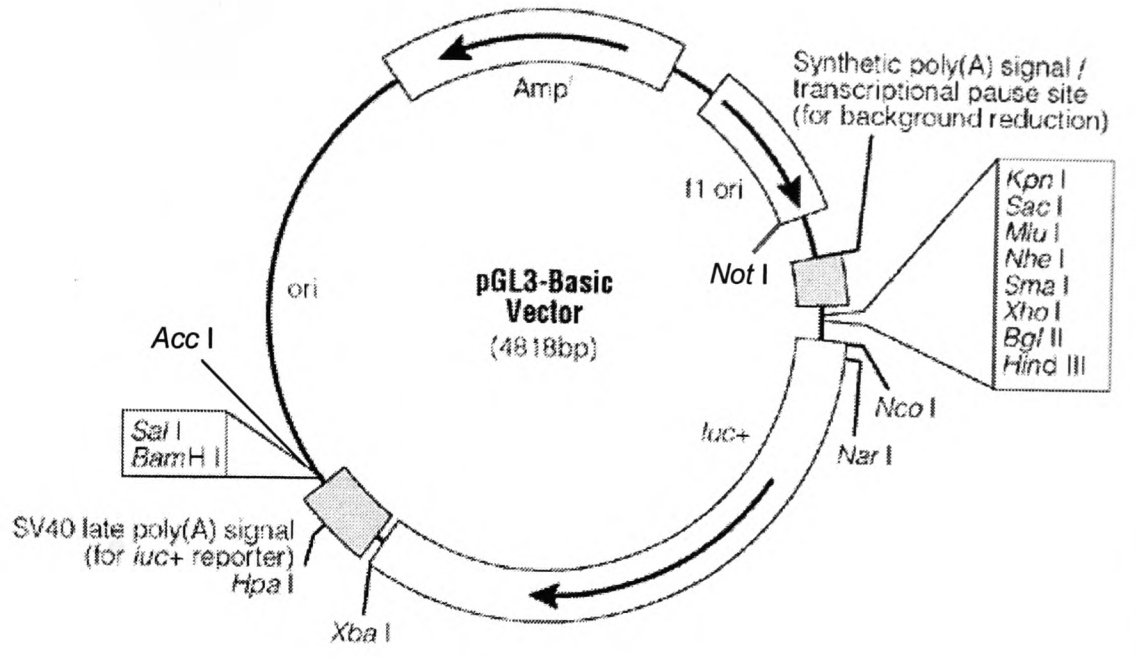
(b)



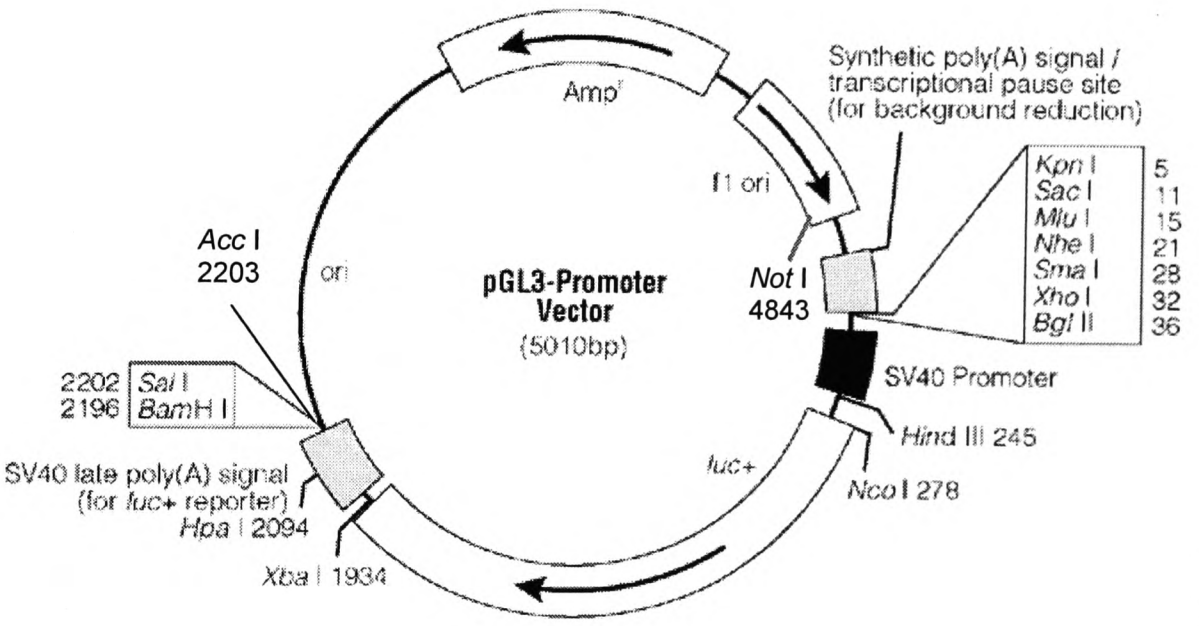
From Promega's product catalogue

1.2 Reporter vectors

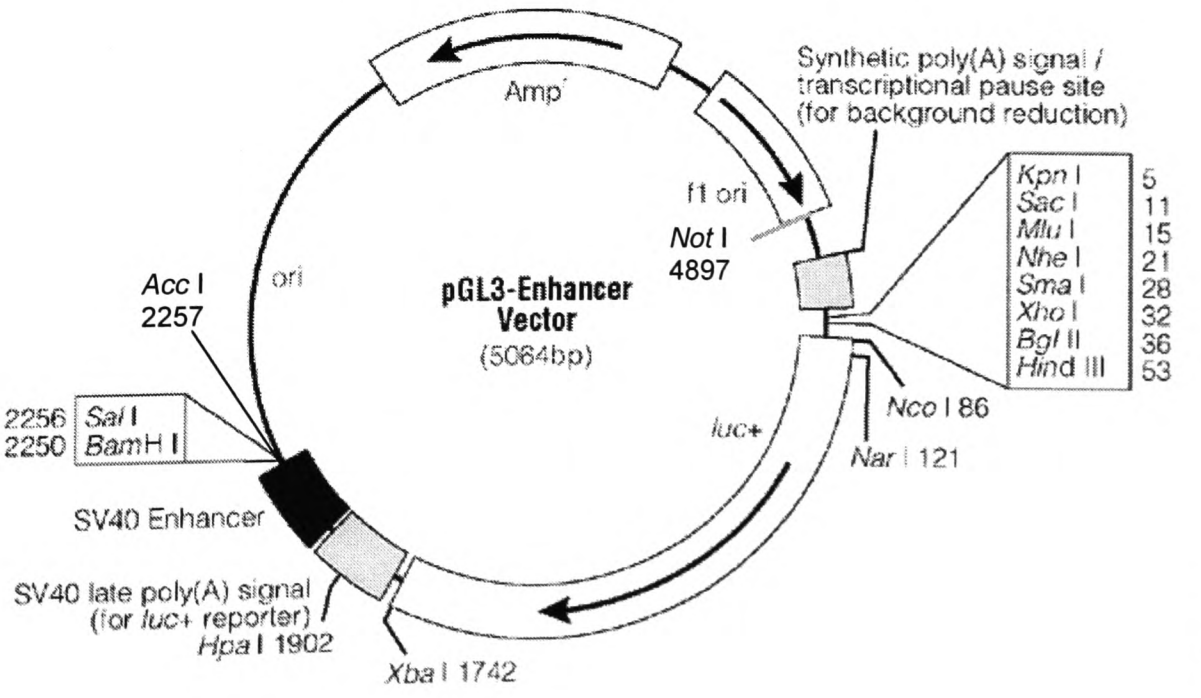
(a)



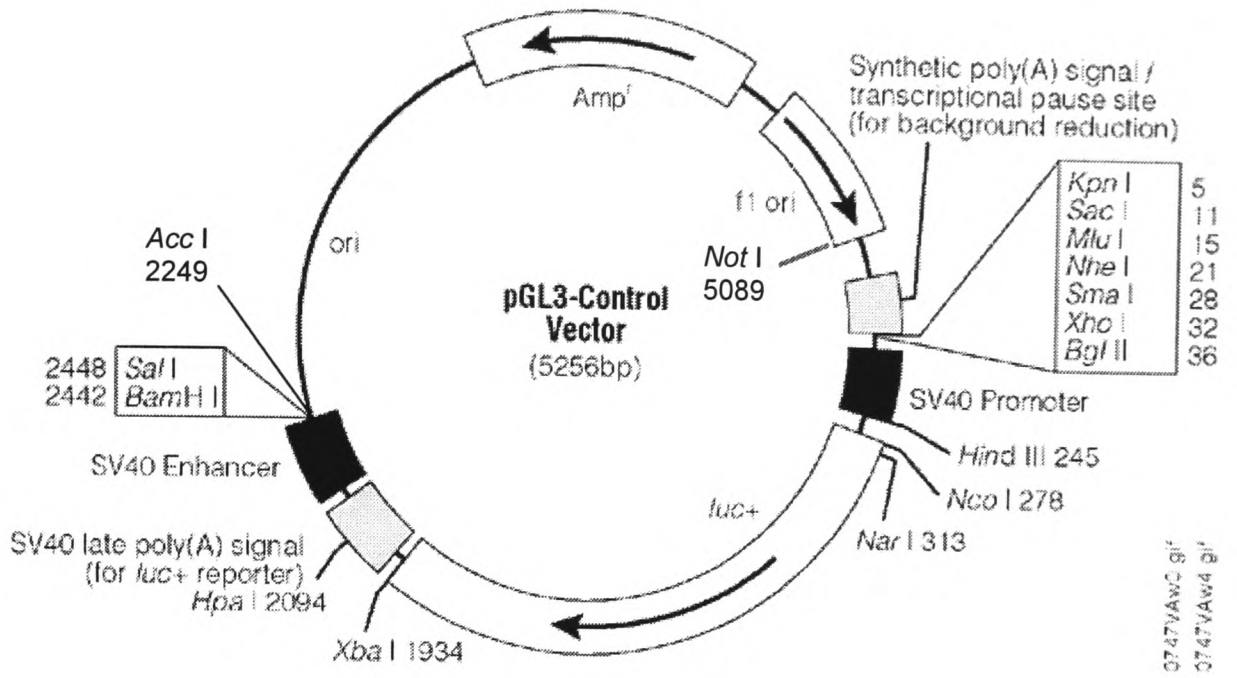
(b)



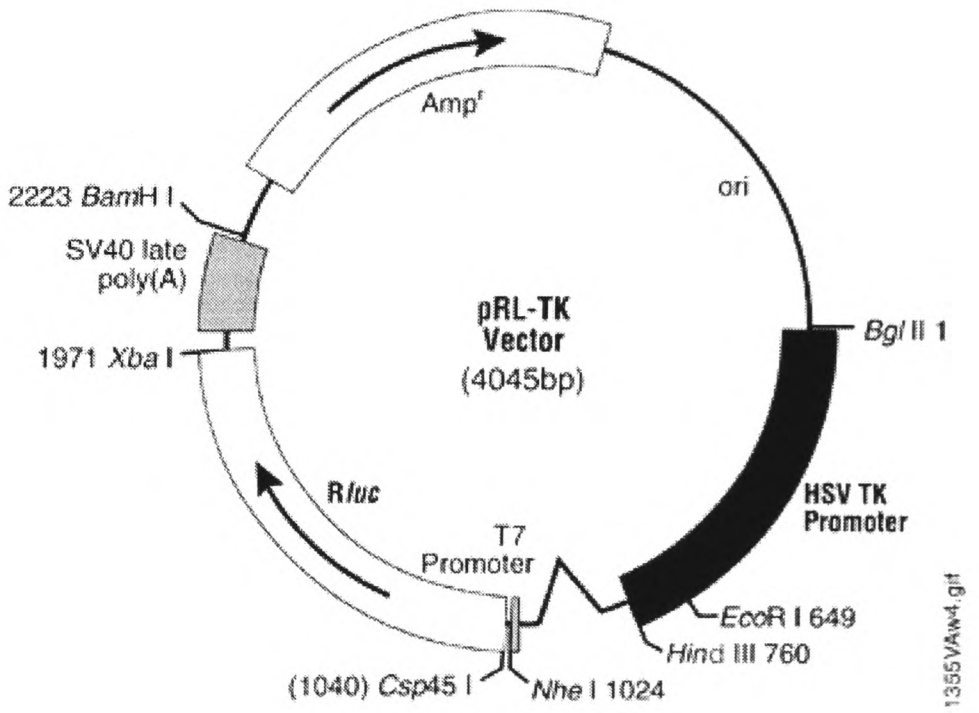
(c)



(d)

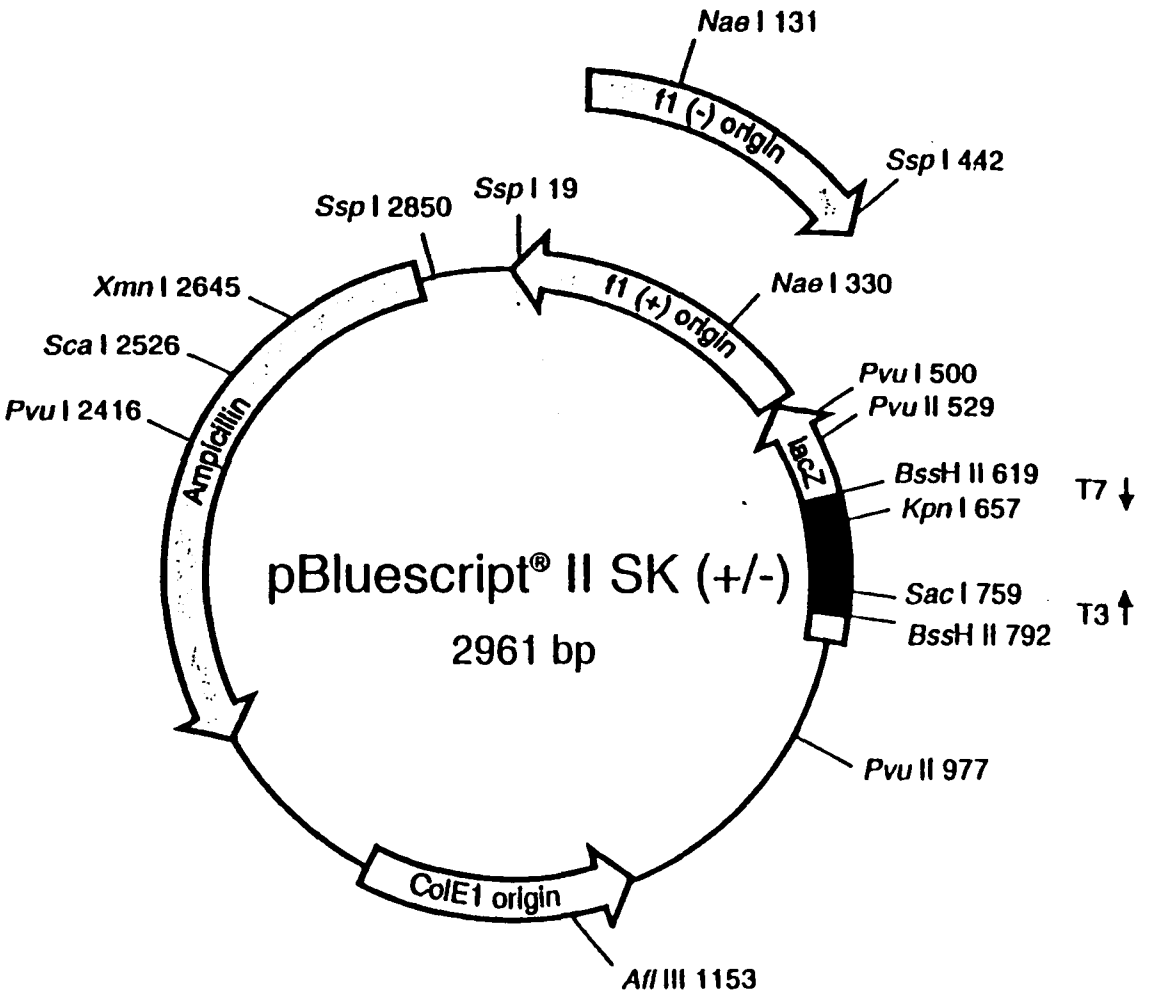


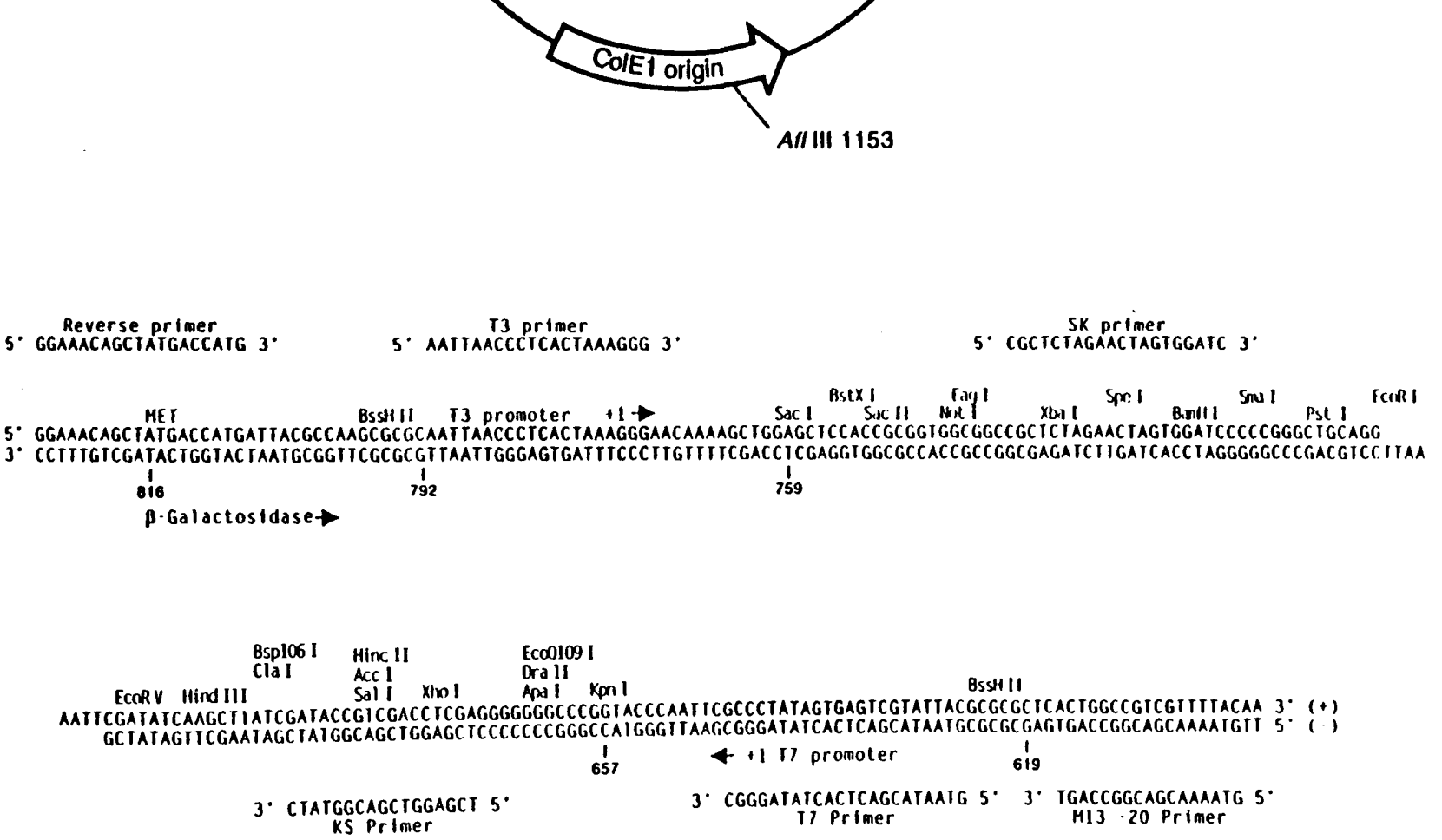
(e)



From Promega's product catalogue

1.3 The pBluescript (pBS) SK(-) vector





From Stratagene's product catalogue

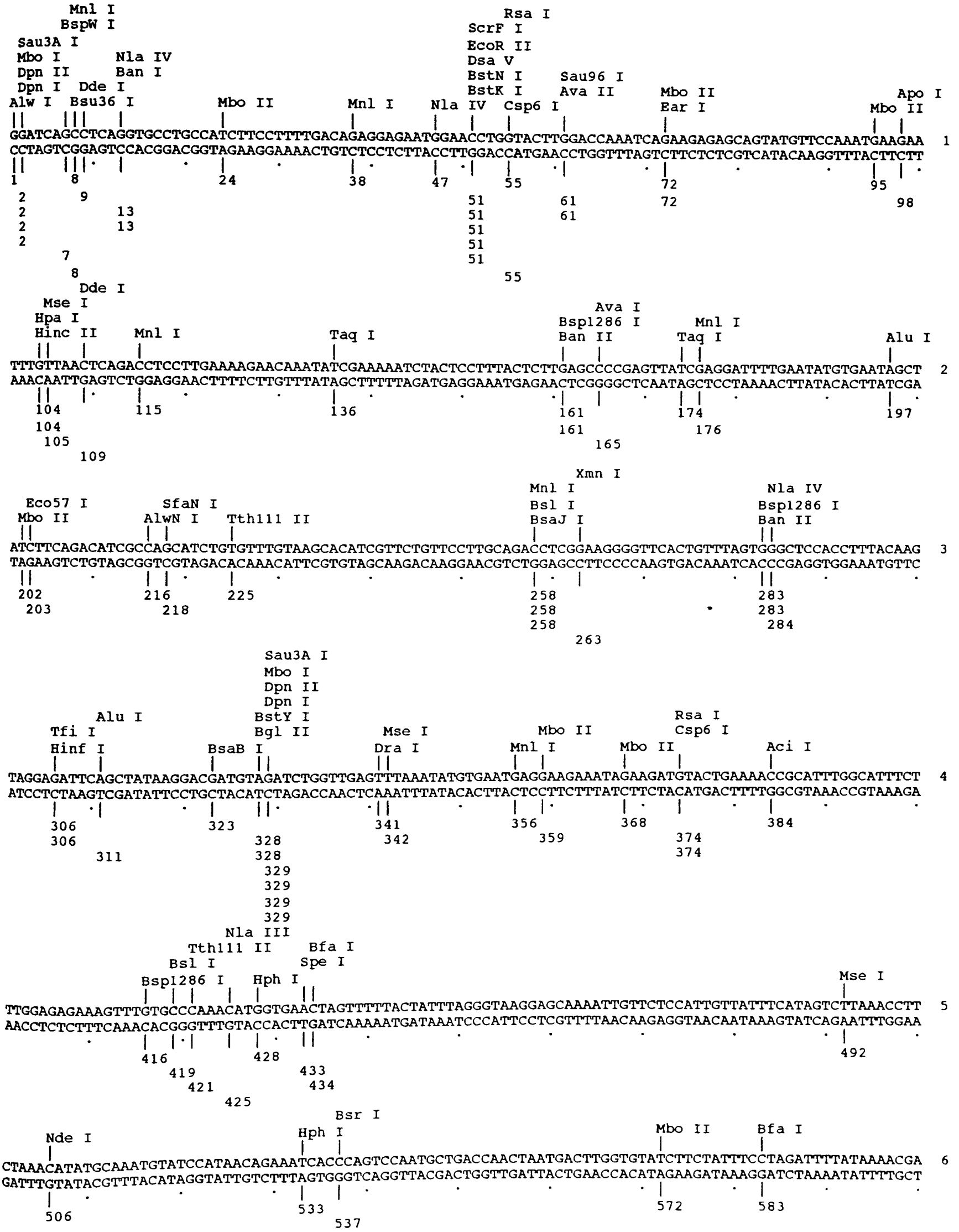
APPENDIX 2

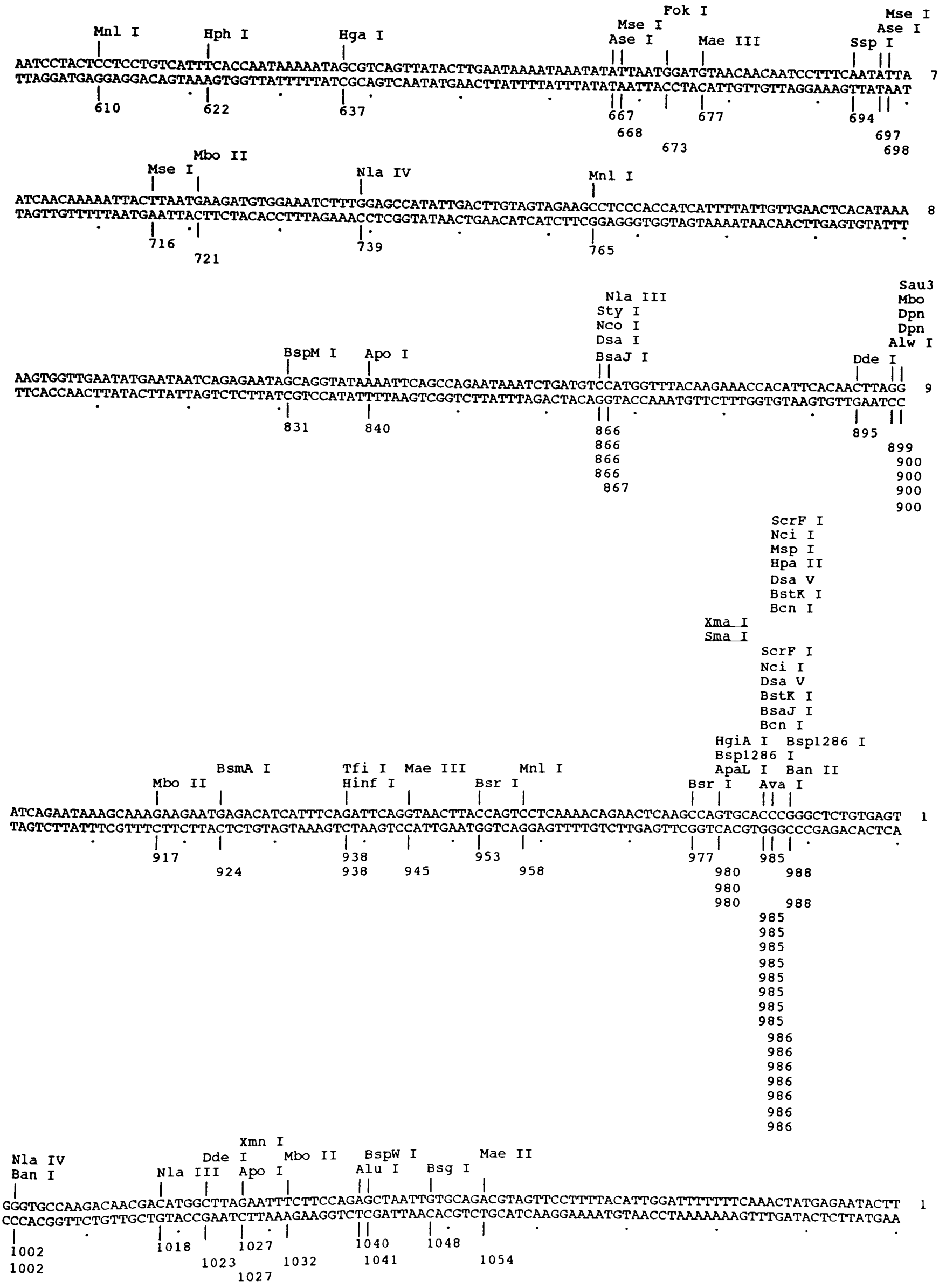
The nucleotide sequence of 129/Ola clone A

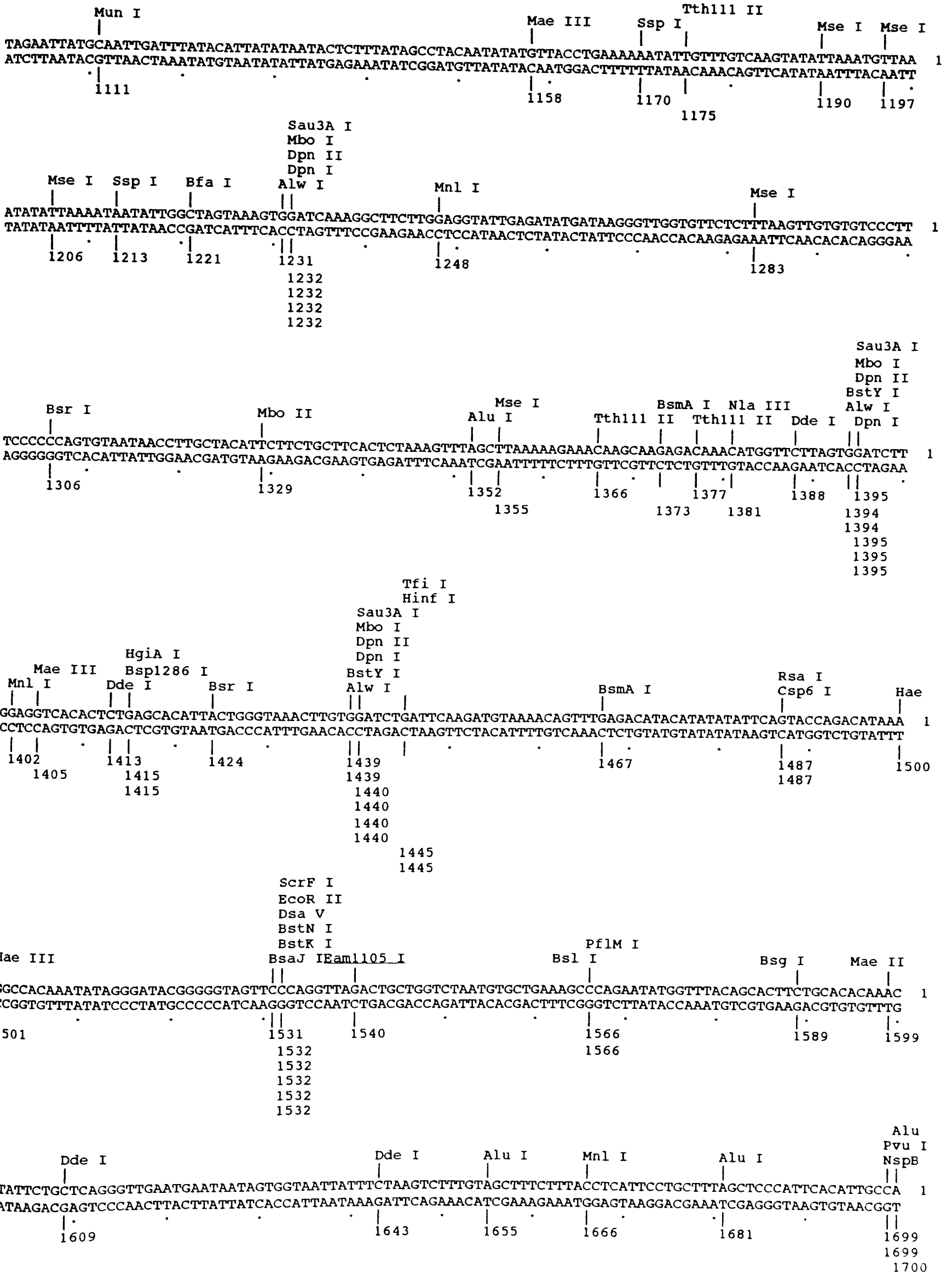
DNA Strider™ 1.2.1 ### Sunday, March 10, 2002 7:56:57 PM

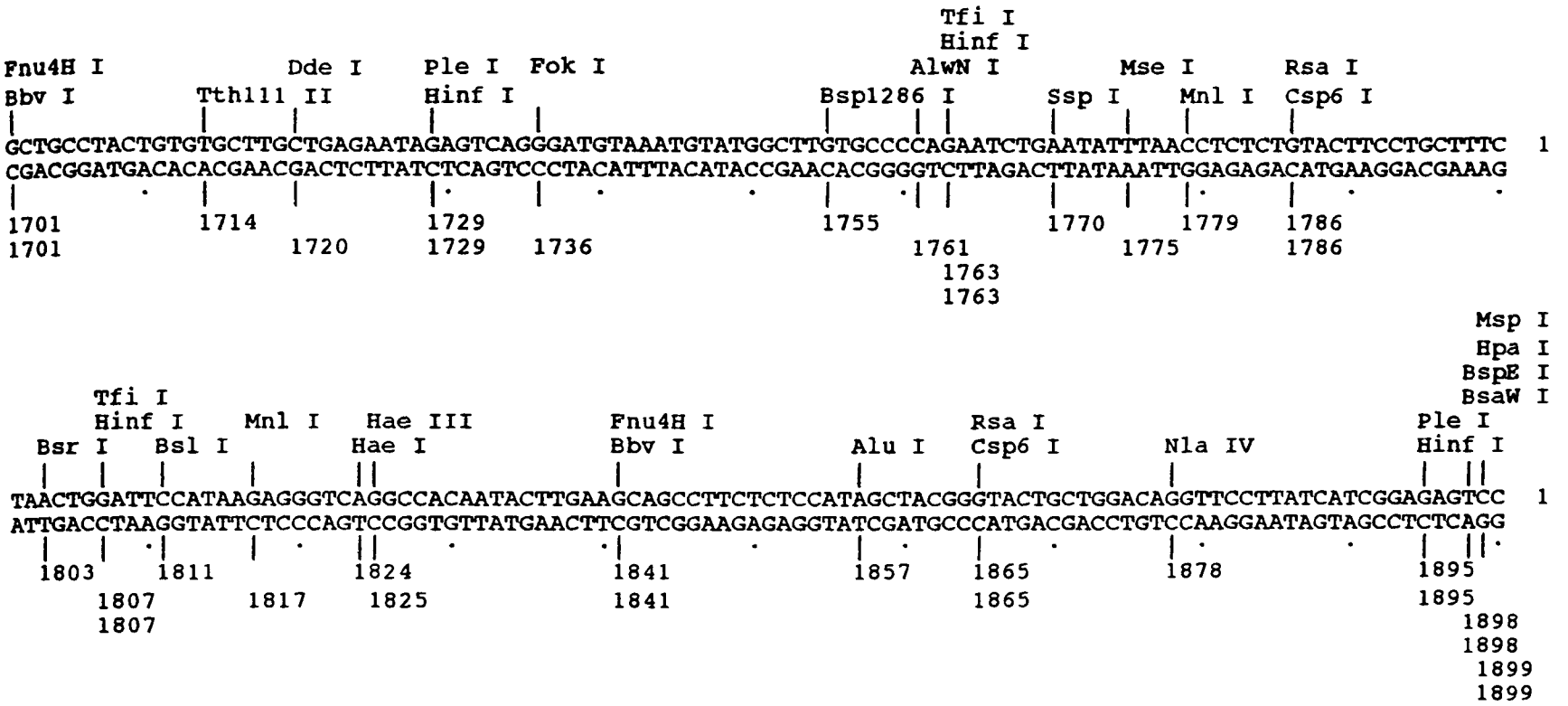
Complete cloneA [1 to 14357] -> Restriction Map

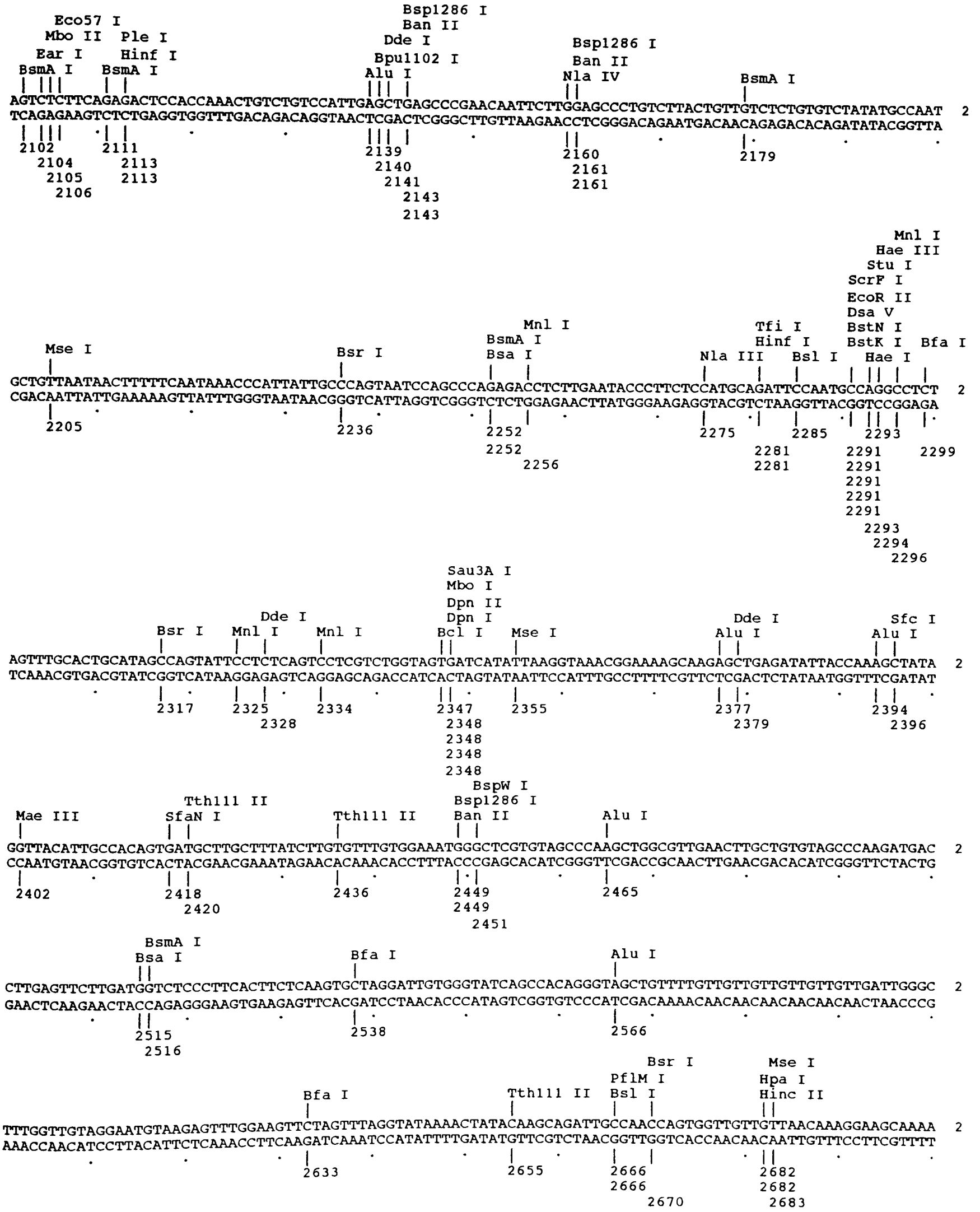
DNA sequence 14357 b.p. GGATCAGCCTCA ... TTATAAAGAGCT linear

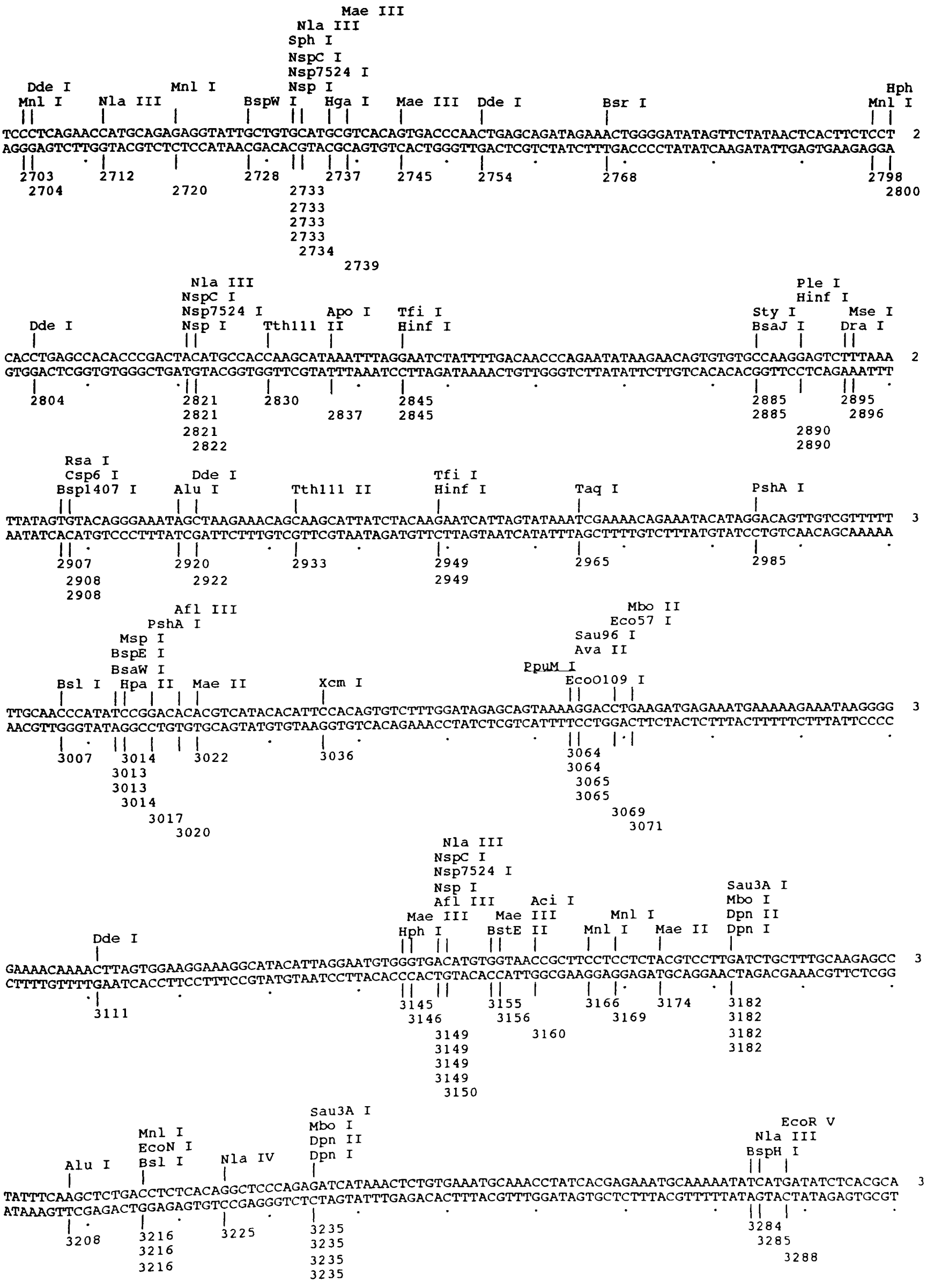


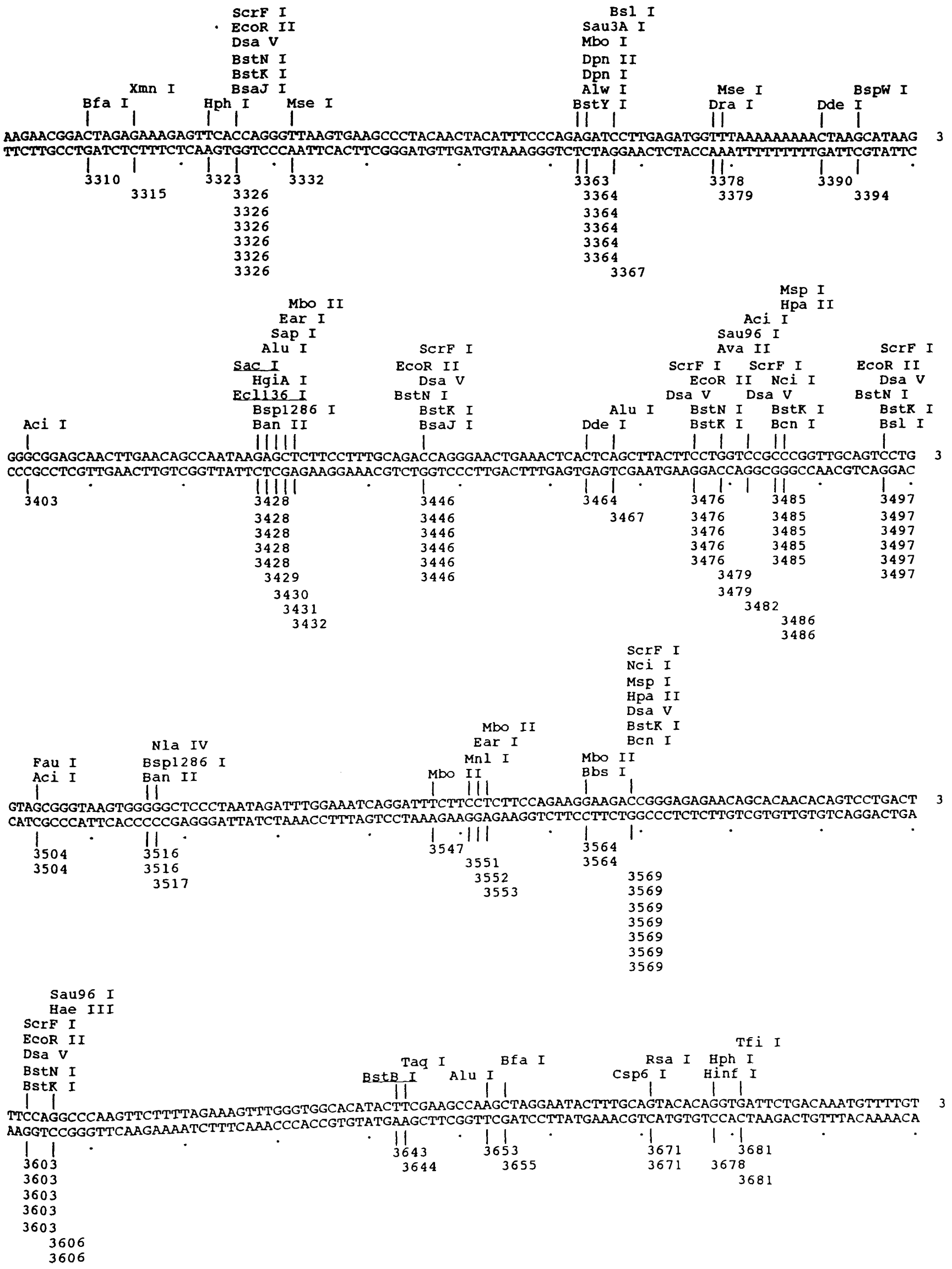

Mnl I
Hph I
Hga I
Fok I
Mse I
Ase I
Ase I
Mae III
Mse I
Ase I
Ssp I
AATCCTACTCCTCCTGTCATTTCACCAATAAAAATAGCGTCAGTTATACTTGAATAAAATAAATATATTAATGGATGTAACAACAATCCTTTCAATATTA 7
TTAGGATGAGGAGGACAGTAAAGTGGTTATTTTTATCGCAGTCAATATGAACTTATTTTATTTATATAATTACCTACATTGTTGTTAGGAAAGTTATAAT
610
622
637
667
668
673
677
694
697
698
Mbo II
Mse I
Nla IV
Mnl I
ATCAACAAAAATTACTTAATGAAGATGTGGAAATCTTTGGAGCCATATTGACTTGTAGTAGAAGCCTCCCACCATCATTTTATTGTTGAACTCACATAAA 8
TAGTTGTTTTTAATGAATTACTTCTACACCTTTAGAAACCTCGGTATAACTGAACATCATCTTCGGAGGGTGGTAGTAAAATAACAACTTGAGTGTATTT
716
721
739
765
Nla III
Sty I
Nco I
Dsa I
Sau3
Mbo
Dpn
Dpn
Alw I
BspM I
Apo I
BsaJ I
Dde I
AAGTGGTTGAATATGAATAATCAGAGAATAGCAGGTATAAAATTCAGCCAGAATAAATCTGATGTCCATGGTTTACAAGAAACCACATTCACAACTTAGG 9
TTCACCAACTTATACTTATTAGTCTCTTATCGTCCATATTTTAAGTCGGTCTTATTTAGACTACAGGTACCAAATGTTCTTTGGTGTAAGTGTTGAATCC
831
840
866
866
866
866
867
895
899
900
900
900
900
ScrF I
Nci I
Msp I
Hpa II
Dsa V
BstK I
Bcn I
Xma I
Sma I
ScrF I
Nci I
Dsa V
BstK I
BsaJ I
Bcn I
HgiA I
Bspl286 I
Bspl286 I
ApaL I
Ban II
BsmA I
Tfi I
Mae III
Mnl I
Mbo II
Hinf I
Bsr I
Bsr I
Ava I
ATCAGAATAAAGCAAAGAAGAATGAGACATCATTTCAGATTCAGGTAACTTACCAGTCCTCAAAACAGAACTCAAGCCAGTGCACCCGGGCTCTGTGAGT 1
TAGTCTTATTTCGTTTCTTCTTACTCTGTAGTAAAGTCTAAGTCCATTGAATGGTCAGGAGTTTTGTCTTGAGTTCGGTCACGTGGGCCCGAGACACTCA
917
924
938
938
945
953
958
977
980
980
980
985
985
985
985
985
985
985
985
985
986
986
986
986
986
986
986
Nla IV
Ban I
Xmn I
Dde I
Mbo II
BspW I
Mae II
Nla III
Apo I
Alu I
Bsg I
GGGTGCCAAGACAACGACATGGCTTAGAATTTCTTCCAGAGCTAATTGTGCAGACGTAGTTCCTTTTACATTGGATTTTTTTCAAACTATGAGAATACTT 1
CCCACGGTTCTGTTGCTGTACCGAATCTTAAAGAAGGTCTCGATTAACACGTCTGCATCAAGGAAAATGTAACCTAAAAAAAGTTTGATACTCTTATGAA
1002
1002
1018
1023
1027
1027
1032
1040
1041
1048
1054

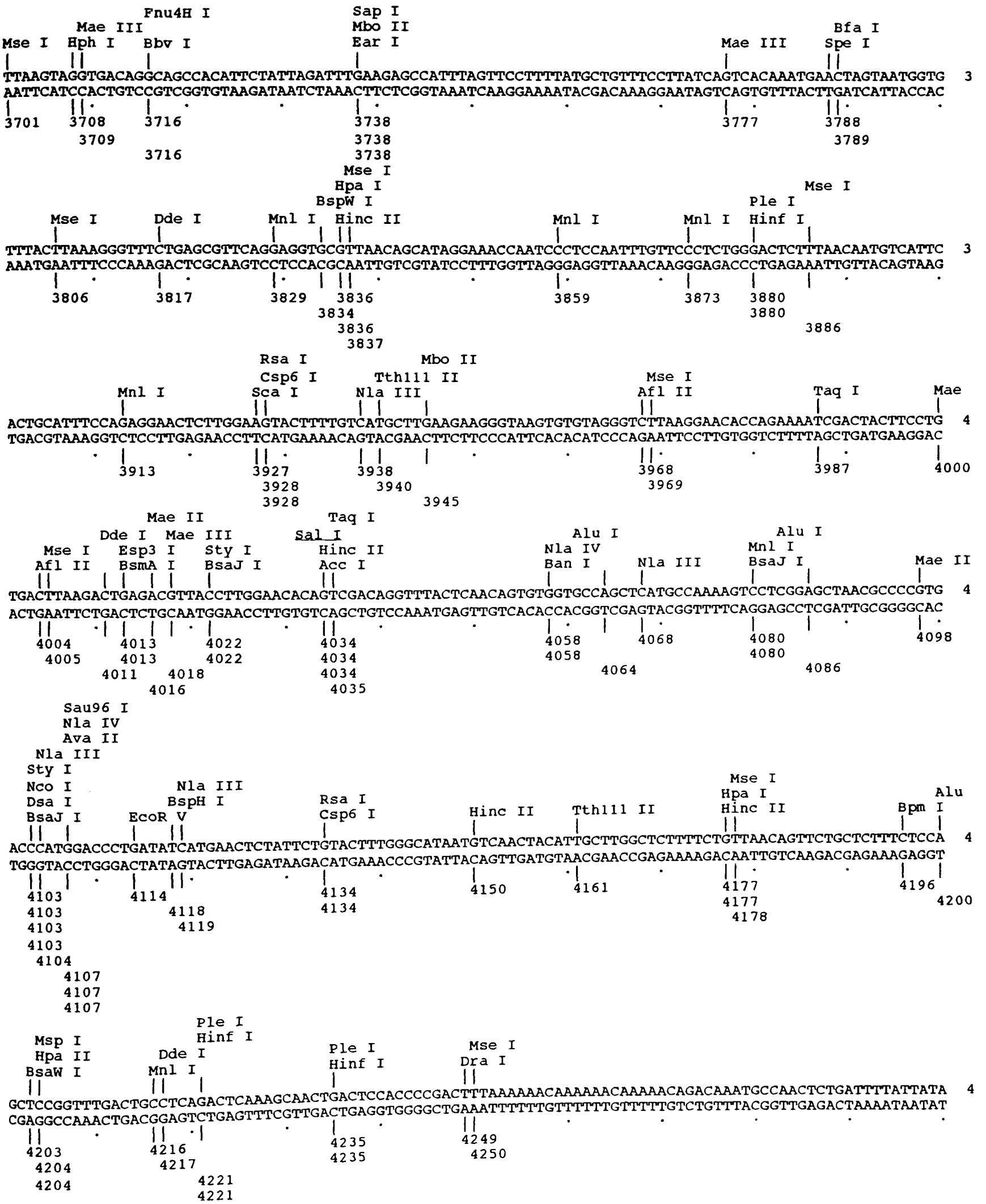

Mun I
Tth111 II
Mae III
Ssp I
Mse I
Mse I
TAGAATTATGCAATTGATTTATACATTATATAATACTCTTTATAGCCTACAATATATGTTACCTGAAAAAATATTGTTTGTCAAGTATATTAAATGTTAA 1
ATCTTAATACGTTAACTAAATATGTAATATATTATGAGAAATATCGGATGTTATATACAATGGACTTTTTTATAACAAACAGTTCATATAATTTACAATT
1111
1158
1170
1175
1190
1197
Sau3A I
Mbo I
Dpn II
Dpn I
Mse I
Ssp I
Bfa I
Alw I
Mnl I
Mse I
ATATATTAAAATAATATTGGCTAGTAAAGTGGATCAAAGGCTTCTTGGAGGTATTGAGATATGATAAGGGTTGGTGTTCTCTTTAAGTTGTGTGTCCCTT 1
TATATAATTTTATTATAACCGATCATTTCACCTAGTTTCCGAAGAACCTCCATAACTCTATACTATTCCCAACCACAAGAGAAATTCAACACACAGGGAA
1206
1213
1221
1231
1232
1232
1232
1232
1248
1283
Sau3A I
Mbo I
Dpn II
BstY I
Alw I
Dpn I
Bsr I
Mbo II
Mse I
Alu I
Tth111 II
BsmA I
Nla III
Tth111 II
Dde I
TCCCCCCAGTGTAATAACCTTGCTACATTCTTCTGCTTCACTCTAAAGTTTAGCTTAAAAAGAAACAAGCAAGAGACAAACATGGTTCTTAGTGGATCTT 1
AGGGGGGTCACATTATTGGAACGATGTAAGAAGACGAAGTGAGATTTCAAATCGAATTTTTCTTTGTTCGTTCTCTGTTTGTACCAAGAATCACCTAGAA
1306
1329
1352
1355
1366
1373
1377
1381
1388
1394
1394
1395
1395
1395
Tfi I
Hinf I
Sau3A I
Mbo I
Dpn II
Dpn I
HgiA I
Bsp1286 I
Mae III
Mnl I
Dde I
Bsr I
BstY I
Alw I
BsmA I
Rsa I
Csp6 I
Hae
GGAGGTCACACTCTGAGCACATTACTGGGTAAACTTGTGGATCTGATTCAAGATGTAAAACAGTTTGAGACATACATATATATTCAGTACCAGACATAAA 1
CCTCCAGTGTGAGACTCGTGTAATGACCCATTTGAACACCTAGACTAAGTTCTACATTTTGTCAAACTCTGTATGTATATATAAGTCATGGTCTGTATTT
1402
1405
1413
1415
1415
1424
1439
1439
1440
1440
1440
1440
1445
1445
1467
1487
1487
1500
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Hae III
BsaJ I
Eam1105 I
PflM I
Bsl I
Bsg I
Mae II
GGCCACAAATATAGGGATACGGGGGTAGTTCCCAGGTTAGACTGCTGGTCTAATGTGCTGAAAGCCCAGAATATGGTTTACAGCACTTCTGCACACAAAC 1
CCGGTGTTTATATCCCTATGCCCCCATCAAGGGTCCAATCTGACGACCAGATTACACGACTTTCGGGTCTTATACCAAATGTCGTGAAGACGTGTGTTTG
1501
1531
1532
1532
1532
1532
1532
1540
1566
1566
1589
1599
Alu
Pvu I
NspB
Dde I
Dde I
Alu I
Mnl I
Alu I
GTATTCTGCTCAGGGTTGAATGAATAATAGTGGTAATTATTTCTAAGTCTTTGTAGCTTTCTTTACCTCATTCCTGCTTTAGCTCCCATTCACATTGCCA 1
CATAAGACGAGTCCCAACTTACTTATTATCACCATTAATAAAGATTCAGAAACATCGAAAGAAATGGAGTAAGGACGAAATCGAGGGTAAGTGTAACGGT
1609
1643
1655
1666
1681
1699
1699
1700

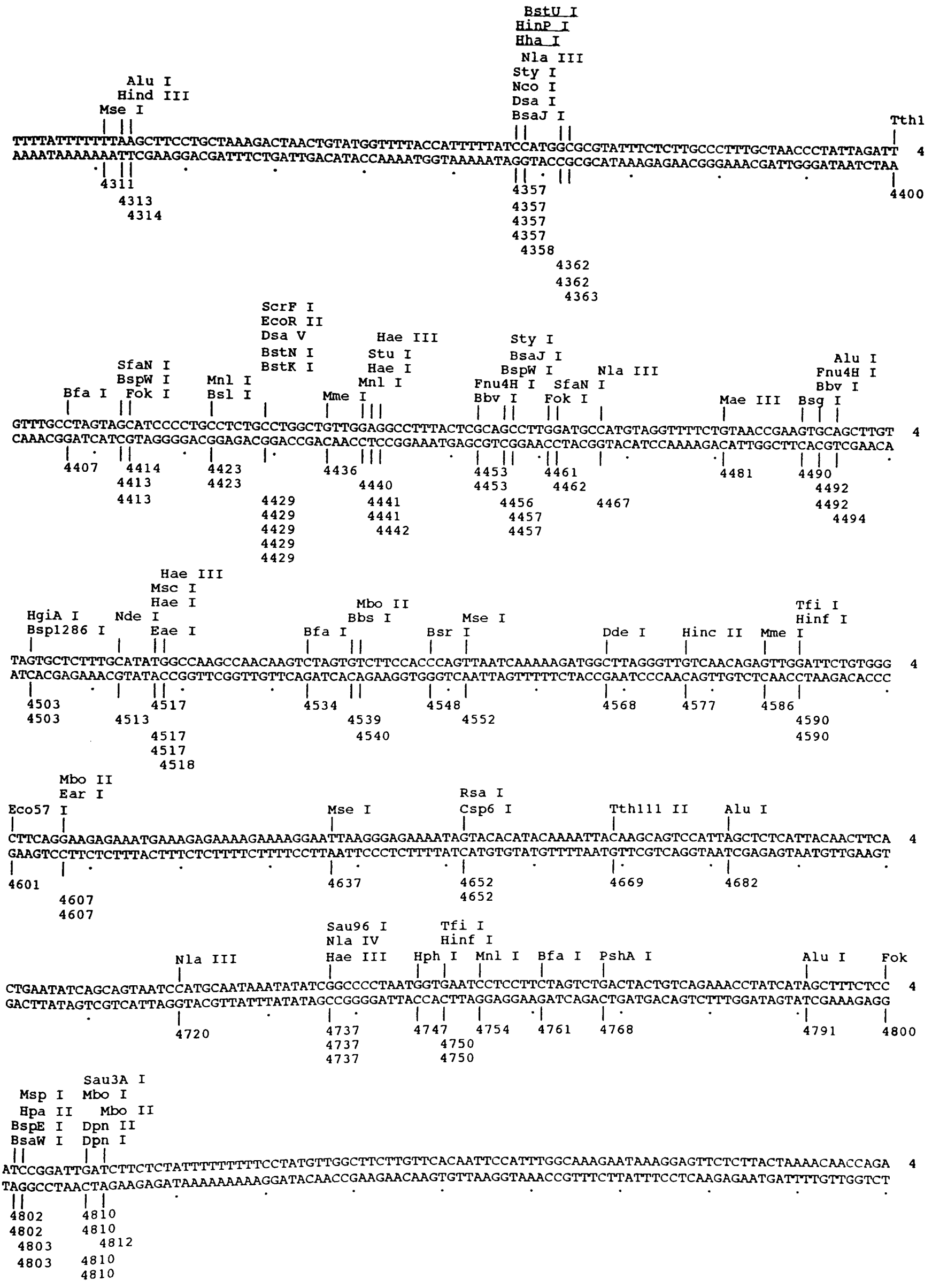

Tfi I
Hinf I
Fnu4H I
Bbv I
Tth111 II
Dde I
Ple I
Hinf I
Fok I
AlwN I
Bsp1286 I
Mse I
Ssp I
Mnl I
Rsa I
Csp6 I
GCTGCCTACTGTGTGCTTGCTGAGAATAGAGTCAGGGATGTAAATGTATGGCTTGTGCCCCAGAATCTGAATATTTAACCTCTCTGTACTTCCTGCTTTC 1
CGACGGATGACACACGAACGACTCTTATCTCAGTCCCTACATTTACATACCGAACACGGGGTCTTAGACTTATAAATTGGAGAGACATGAAGGACGAAAG
1701 1714 1729 1755 1770 1779 1786
1701 1720 1729 1736 1761 1775 1786
1763
1763
Msp I
Hpa I
BspE I
BsaW I
Tfi I
Hinf I
Bsr I
Bsl I
Mnl I
Hae III
Hae I
Fnu4H I
Bbv I
Alu I
Rsa I
Csp6 I
Nla IV
Ple I
Hinf I
TAACTGGATTCCATAAGAGGGTCAGGCCACAATACTTGAAGCAGCCTTCTCTCCATAGCTACGGGTACTGCTGGACAGGTTCCTTATCATCGGAGAGTCC 1
ATTGACCTAAGGTATTCTCCCAGTCCGGTGTTATGAACTTCGTCGGAAGAGAGGTATCGATGCCCATGACGACCTGTCCAAGGAATAGTAGCCTCTCAGG
1803 1811 1824 1841 1857 1865 1878 1895
1807 1817 1825 1841 1865 1895
1807
1898
1898
1899
1899
Bsr I
ScrF I
EcoR II
Dsa V
BstN I
BstK I
BsaJ I
BsaJ I
Mae III
Tfi I
Hae III
Sau3A I
Mbo I
Dpn II
Dpn I
Alw I
BstY I
Sfc I
Pst I
Fnu4H I
Bbv I
Mnl I
Bsl I
Fok I
Mbo II
Ear I
Bsl I
Nla IV
Mnl I
Mnl I
Fok I
Sau96 I
Hae III
Dde I
Alu I
Hinf I
Sau96 I
AlwN I
Mae III
Alw I
GGAGCTATCAGGATTCCCAGGGCCAGTTACCATCAAGGGAAGTGTTGGCTTTTGCACAGTTGCTGTCACACAACCCGATCCCCAGCCCCTACCTACATTC 2
CCTCGATAGTCCTAAGGGTCCCGGTCAATGGTAGTTCCCTTCACAACCGAAAACGTGTCAACGACAGTGTGTTGGGCTAGGGGTCGGGGATGGATGTAAG
1903 1912 1920 1957 1965 1977
1912 1921 1977
1916 1926 1977
1917 1977
1917 1977
1917
1917
1917
1923
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Sau3A I
Mbo I
Dpn II
Dpn I
Alw I
BstY I
Sfc I
Pst I
Fnu4H I
Bbv I
Ple I
Hinf I
Tth111 II
CTGCTTGGGAGTCAGAGATCCTGGCTGCAGGCTCTTGCTATGAACTTCCCCCATCCTCACAAGGGATAAAAGGAGGCTCCCTCTTCTGGCCCATCCTCTT 2
GACGAACCCTCAGTCTCTAGGACCGACGTCCGAGAACGATACTTGAAGGGGGTAGGAGTGTTCCCTATTTTCCTCCGAGGGAGAAGACCGGGTAGGAGAA
2002 2009 2016 2024 2052 2055 2073 2075 2079 2080 2081 2082 2088 2088 2092 2095 2098
2009 2017 2024 2055 2055
2017 2025
2017 2025
2017
2017
2020
2020
2020
2020
2020

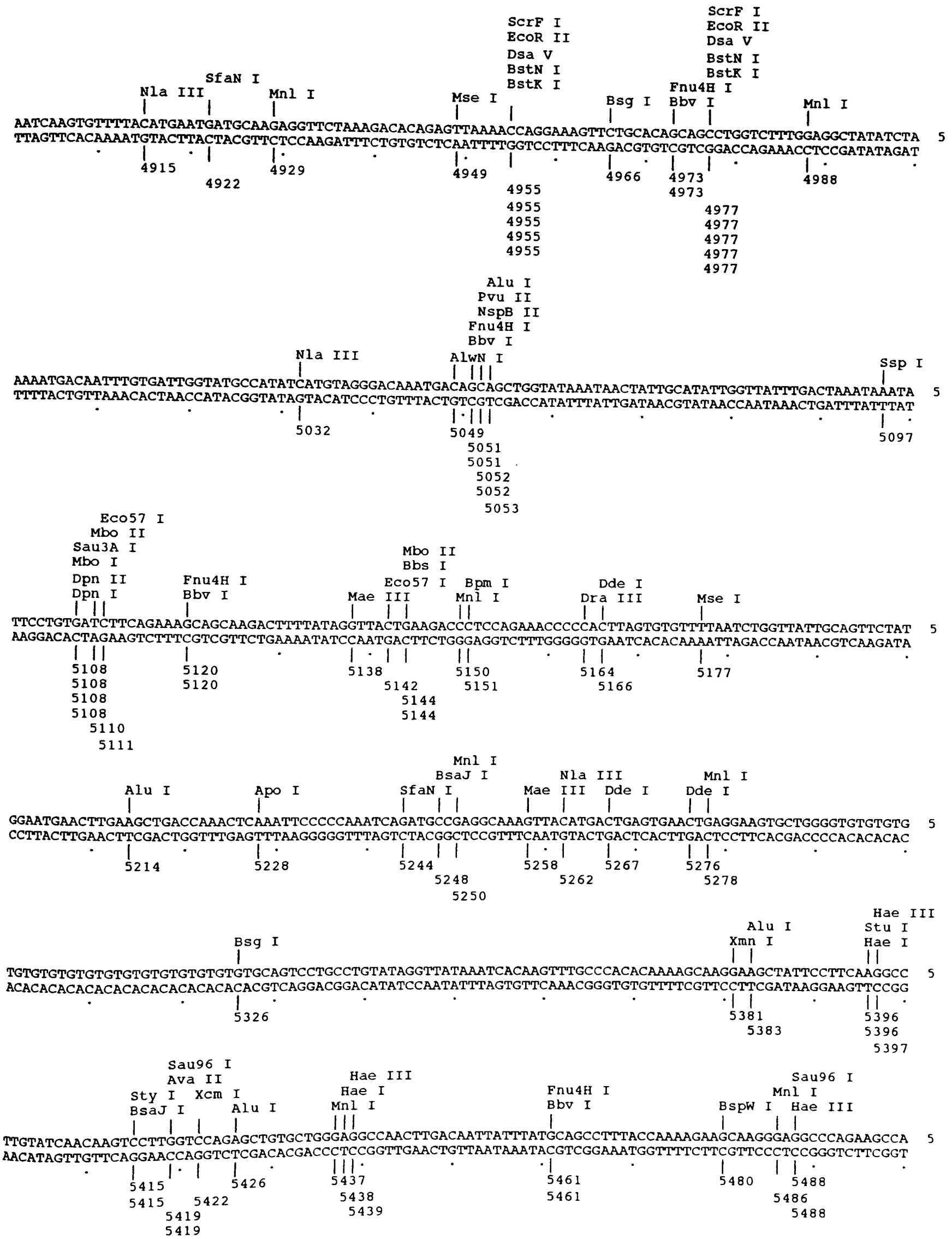

Eco57 I
Mbo II
Ple I
Ear I
Hinf I
BsmA I
BsmA I
Bsp1286 I
Ban II
Dde I
Bpu1102 I
Alu I
Bsp1286 I
Ban II
Nla IV
BsmA I
AGTCTCTTCAGAGACTCCACCAAACTGTCTGTCCATTGAGCTGAGCCCGAACAATTCTTGGAGCCCTGTCTTACTGTTGTCTCTGTGTCTATATGCCAAT 2
TCAGAGAAGTCTCTGAGGTGGTTTGACAGACAGGTAACTCGACTCGGGCTTGTTAAGAACCTCGGGACAGAATGACAACAGAGACACAGATATACGGTTA
2102
2104
2105
2106
2111
2113
2113
2139
2140
2141
2143
2143
2160
2161
2161
2179
Mnl I
Hae III
Stu I
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Bfa I
Mse I
Mnl I
BsmA I
Bsa I
Bsr I
Tfi I
Hinf I
Nla III
Bsl I
Hae I
GCTGTTAATAACTTTTTCAATAAACCCATTATTGCCCAGTAATCCAGCCCAGAGACCTCTTGAATACCCTTCTCCATGCAGATTCCAATGCCAGGCCTCT 2
CGACAATTATTGAAAAAGTTATTTGGGTAATAACGGGTCATTAGGTCGGGTCTCTGGAGAACTTATGGGAAGAGGTACGTCTAAGGTTACGGTCCGGAGA
2205
2236
2252
2252
2256
2275
2281
2281
2285
2291
2291
2291
2291
2291
2293
2293
2294
2296
2299
Sau3A I
Mbo I
Dpn II
Dpn I
Dde I
Dde I
Sfc I
Bsr I
Mnl I
Mnl I
Bcl I
Mse I
Alu I
Alu I
AGTTTGCACTGCATAGCCAGTATTCCTCTCAGTCCTCGTCTGGTAGTGATCATATTAAGGTAAACGGAAAAGCAAGAGCTGAGATATTACCAAAGCTATA 2
TCAAACGTGACGTATCGGTCATAAGGAGAGTCAGGAGCAGACCATCACTAGTATAATTCCATTTGCCTTTTCGTTCTCGACTCTATAATGGTTTCGATAT
2317
2325
2328
2334
2347
2348
2348
2348
2348
2355
2377
2379
2394
2396
BspW I
Tth111 II
Bsp1286 I
Mae III
SfaN I
Tth111 II
Ban II
Alu I
GGTTACATTGCCACAGTGATGCTTGCTTTATCTTGTGTTTGTGGAAATGGGCTCGTGTAGCCCAAGCTGGCGTTGAACTTGCTGTGTAGCCCAAGATGAC 2
CCAATGTAACGGTGTCACTACGAACGAAATAGAACACAAACACCTTTACCCGAGCACATCGGGTTCGACCGCAACTTGAACGACACATCGGGTTCTACTG
2402
2418
2420
2436
2449
2449
2451
2465
BsmA I
Bsa I
Bfa I
Alu I
CTTGAGTTCTTGATGGTCTCCCTTCACTTCTCAAGTGCTAGGATTGTGGGTATCAGCCACAGGGTAGCTGTTTTGTTGTTGTTGTTGTTGATTGGGC 2
GAACTCAAGAACTACCAGAGGGAAGTGAAGAGTTCACGATCCTAACACCCATAGTCGGTGTCCCATCGACAAAACAACAACAACAACAACTAACCCG
2515
2516
2538
2566
Bsr I
Mse I
PflM I
Hpa I
Bfa I
Tth111 II
Bsl I
Hinc II
TTTGGTTGTAGGAATGTAAGAGTTTGGAAGTTCTAGTTTAGGTATAAAACTATACAAGCAGATTGCCAACCAGTGGTTGTTGTTAACAAAGGAAGCAAAA 2
AAACCAACATCCTTACATTCTCAAACCTTCAAGATCAAATCCATATTTTGATATGTTCGTCTAACGGTTGGTCACCAACAACAATTGTTTCCTTCGTTTT
2633
2655
2666
2666
2670
2682
2682
2683

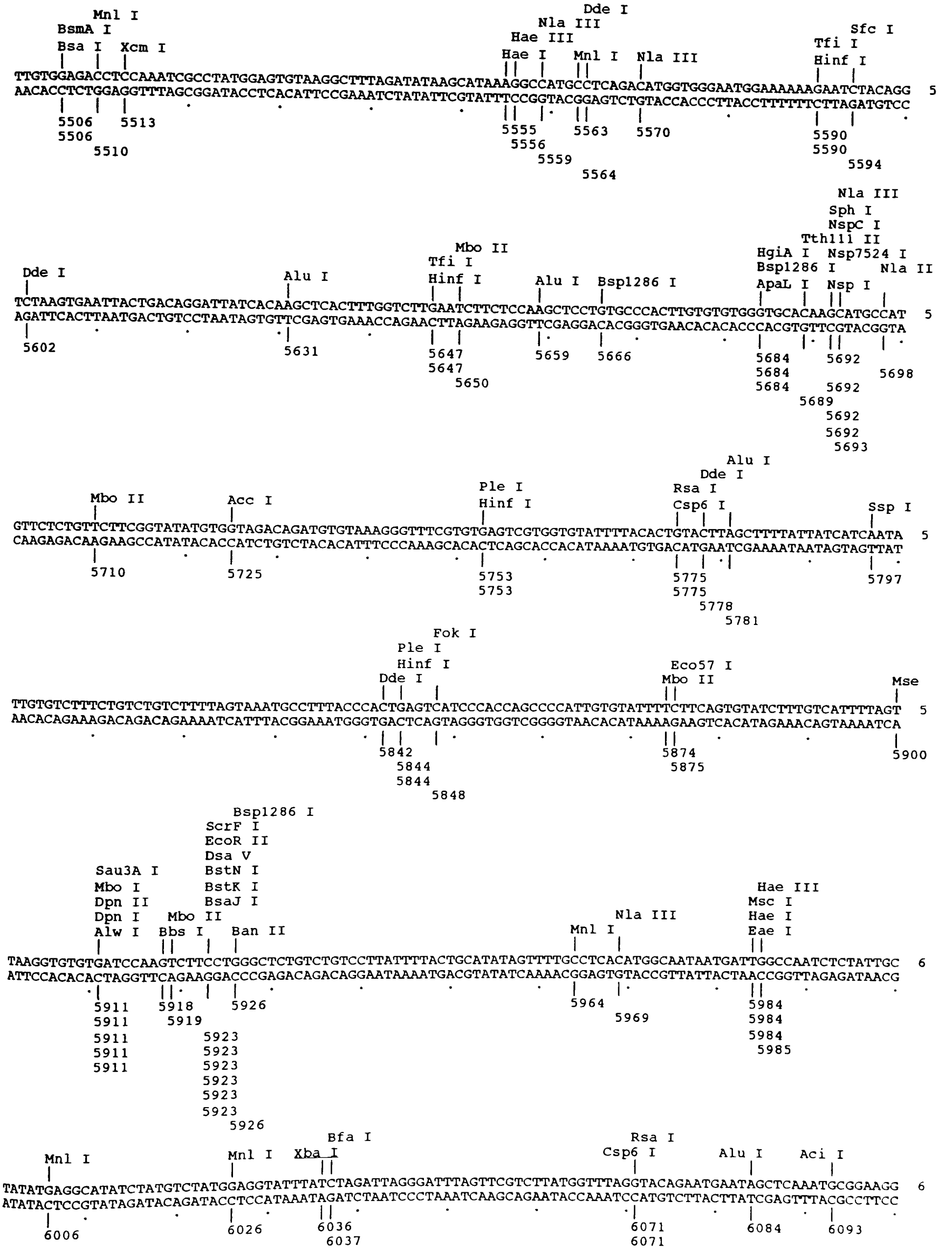

Mae III
Nla III
Sph I
NspC I
Nsp7524 I
Dde I
Mnl I
Nsp I
Hph
Mnl I
Nla III
BspW I
Hga I
Mae III
Dde I
Bsr I
Mnl I
TCCCTCAGAACCATGCAGAGAGGTATTGCTGTGCATGCGTCACAGTGACCCAACTGAGCAGATAGAAACTGGGGATATAGTTCTATAACTCACTTCTCCT 2
AGGGAGTCTTGGTACGTCTCTCCATAACGACACGTACGCAGTGTCACTGGGTTGACTCGTCTATCTTTGACCCCTATATCAAGATATTGAGTGAAGAGGA
2703 2712 2720 2728 2737 2745 2754 2768 2798
2704 2733 2800
2733
2733
2733
2734
2739
Nla III
NspC I
Nsp7524 I
Nsp I
Apo I
Tfi I
Ple I
Hinf I
Dde I
Tth111 II
Hinf I
Sty I
BsaJ I
Mse I
Dra I
CACCTGAGCCACACCCGACTACATGCCACCAAGCATAAATTTAGGAATCTATTTTGACAACCCAGAATATAAGAACAGTGTGTGCCAAGGAGTCTTTAAA 2
GTGGACTCGGTGTGGGCTGATGTACGGTGGTTCGTATTTAAATCCTTAGATAAAACTGTTGGGTCTTATATTCTTGTCACACACGGTTCCTCAGAAATTT
2804 2821 2830 2837 2845 2885 2895
2821 2845 2885 2896
2821 2890
2822 2890
Rsa I
Csp6 I
Dde I
Tfi I
Bspl407 I
Alu I
Tth111 II
Hinf I
Taq I
PshA I
TTATAGTGTACAGGGAAATAGCTAAGAAACAGCAAGCATTATCTACAAGAATCATTAGTATAAATCGAAAACAGAAATACATAGGACAGTTGTCGTTTTT 3
AATATCACATGTCCCTTTATCGATTCTTTGTCGTTCGTAATAGATGTTCTTAGTAATCATATTTAGCTTTTGTCTTTATGTATCCTGTCAACAGCAAAAA
2907 2920 2933 2949 2965 2985
2908 2922 2949
2908
Afl III
PshA I
Msp I
BspE I
BsaW I
Mbo II
Eco57 I
Sau96 I
Ava II
PpuM I
Bsl I
Hpa II
Mae II
Xcm I
EcoO109 I
TTGCAACCCATATCCGGACACACGTCATACACATTCCACAGTGTCTTTGGATAGAGCAGTAAAAGGACCTGAAGATGAGAAATGAAAAAGAAATAAGGGG 3
AACGTTGGGTATAGGCCTGTGTGCAGTATGTGTAAGGTGTCACAGAAACCTATCTCGTCATTTTCCTGGACTTCTACTCTTTACTTTTTCTTTATTCCCC
3007 3014 3022 3036 3064 3069 3071
3013 3064
3013 3065
3014 3065
3017
3020
Nla III
NspC I
Nsp7524 I
Nsp I
Afl III
Aci I
Sau3A I
Mbo I
Dpn II
Dpn I
Mae III
Hph I
Mae III
BstE II
Mnl I
Mnl I
Mae II
Dde I
GAAAACAAAACTTAGTGGAAGGAAAGGCATACATTAGGAATGTGGGTGACATGTGGTAACCGCTTCCTCCTCTACGTCCTTGATCTGCTTTGCAAGAGCC 3
CTTTTGTTTTGAATCACCTTCCTTTCCGTATGTAATCCTTACACCCACTGTACACCATTGGCGAAGGAGGAGATGCAGGAACTAGACGAAACGTTCTCGG
3111 3145 3155 3166 3174 3182
3146 3156 3169 3182
3149 3160 3182
3149 3182
3149
3149
3150
Sau3A I
Mbo I
Dpn II
Dpn I
EcoR V
Nla III
BspH I
Mnl I
EcoN I
Bsl I
Nla IV
Alu I
TATTTCAAGCTCTGACCTCTCACAGGCTCCCAGAGATCATAAACTCTGTGAAATGCAAACCTATCACGAGAAATGCAAAAATATCATGATATCTCACGCA 3
ATAAAGTTCGAGACTGGAGAGTGTCCGAGGGTCTCTAGTATTTGAGACACTTTACGTTTGGATAGTGCTCTTTACGTTTTTATAGTACTATAGAGTGCGT
3208 3216 3225 3235 3284 3288
3216 3235 3285
3216 3235
3235

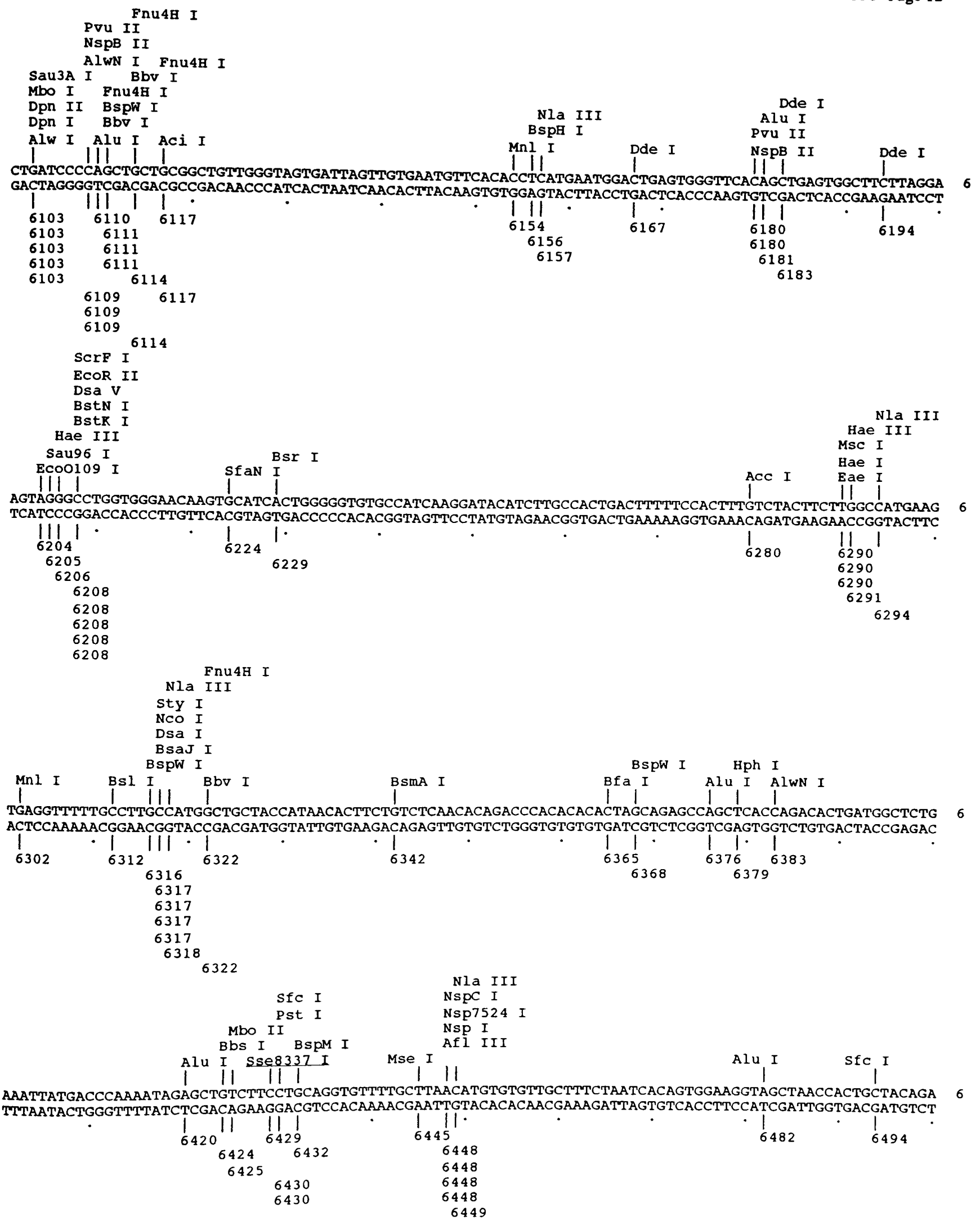

ScrF I
EcoR II
Dsa V
BstN I
BstK I
BsaJ I
Xmn I
Bfa I
Hph I
Mse I
Bsl I
Sau3A I
Mbo I
Dpn II
Dpn I
Alw I
BstY I
Mse I
Dra I
BspW I
Dde I
AAGAACGGACTAGAGAAAGAGTTCACCAGGGTTAAGTGAAGCCCTACAACTACATTTCCCAGAGATCCTTGAGATGGTTTAAAAAAAAACTAAGCATAAG 3
TTCTTGCCTGATCTCTTTCTCAAGTGGTCCCAATTCACTTCGGGATGTTGATGTAAAGGGTCTCTAGGAACTCTACCAAATTTTTTTTGATTCGTATTC
3310 3323 3332 3363 3378 3390
3315 3326 3364 3379 3394
3326 3364
3326 3364
3326 3364
3326 3364
3326 3367
Mbo II
Ear I
Sap I
Alu I
Sac I
HgiA I
Ecl136 I
Bsp1286 I
Ban II
ScrF I
EcoR II
Dsa V
BstN I
BstK I
BsaJ I
Msp I
Hpa II
Aci I
Sau96 I
Ava II
ScrF I
EcoR II
Nci I
Dsa V
Alu I
Dde I
BstN I
BstK I
Bcn I
Aci I
GGGCGGAGCAACTTGAACAGCCAATAAGAGCTCTTCCTTTGCAGACCAGGGAACTGAAACTCACTCAGCTTACTTCCTGGTCCGCCCGGTTGCAGTCCTG 3
CCCGCCTCGTTGAACTTGTCGGTTATTCTCGAGAAGGAAACGTCTGGTCCCTTGACTTTGAGTGAGTCGAATGAAGGACCAGGCGGGCCAACGTCAGGAC
3403 3428 3446 3464 3476 3485 3497
3428 3446 3467 3476 3485 3497
3428 3446 3476 3485 3497
3429 3446 3476 3479 3497
3430 3479
3431 3482
3432 3486
3486
ScrF I
Nci I
Msp I
Hpa II
Dsa V
BstK I
Bcn I
Mbo II
Ear I
Mnl I
Mbo II
Nla IV
Fau I
Aci I
Bsp1286 I
Ban II
Mbo II
Bbs I
GTAGCGGGTAAGTGGGGGCTCCCTAATAGATTTGGAAATCAGGATTTCTTCCTCTTCCAGAAGGAAGACCGGGAGAGAACAGCACAACACAGTCCTGACT 3
CATCGCCCATTCACCCCCGAGGGATTATCTAAACCTTTAGTCCTAAAGAAGGAGAAGGTCTTCCTTCTGGCCCTCTCTTGTCGTGTTGTGTCAGGACTGA
3504 3516 3547 3551 3564 3569
3504 3516 3552 3564 3569
3517 3553 3569
3569
3569
3569
3569
Sau96 I
Hae III
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Taq I
BstB I
Alu I
Bfa I
Rsa I
Csp6 I
Hph I
Hinf I
Tfi I
TTCCAGGCCCAAGTTCTTTTAGAAAGTTTGGGTGGCACATACTTCGAAGCCAAGCTAGGAATACTTTGCAGTACACAGGTGATTCTGACAAATGTTTTGT 3
AAGGTCCGGGTTCAAGAAAATCTTTCAAACCCACCGTGTATGAAGCTTCGGTTCGATCCTTATGAAACGTCATGTGTCCACTAAGACTGTTTACAAAACA
3603 3643 3653 3671 3681
3603 3644 3655 3671 3678
3603 3681
3603
3603
3606
3606

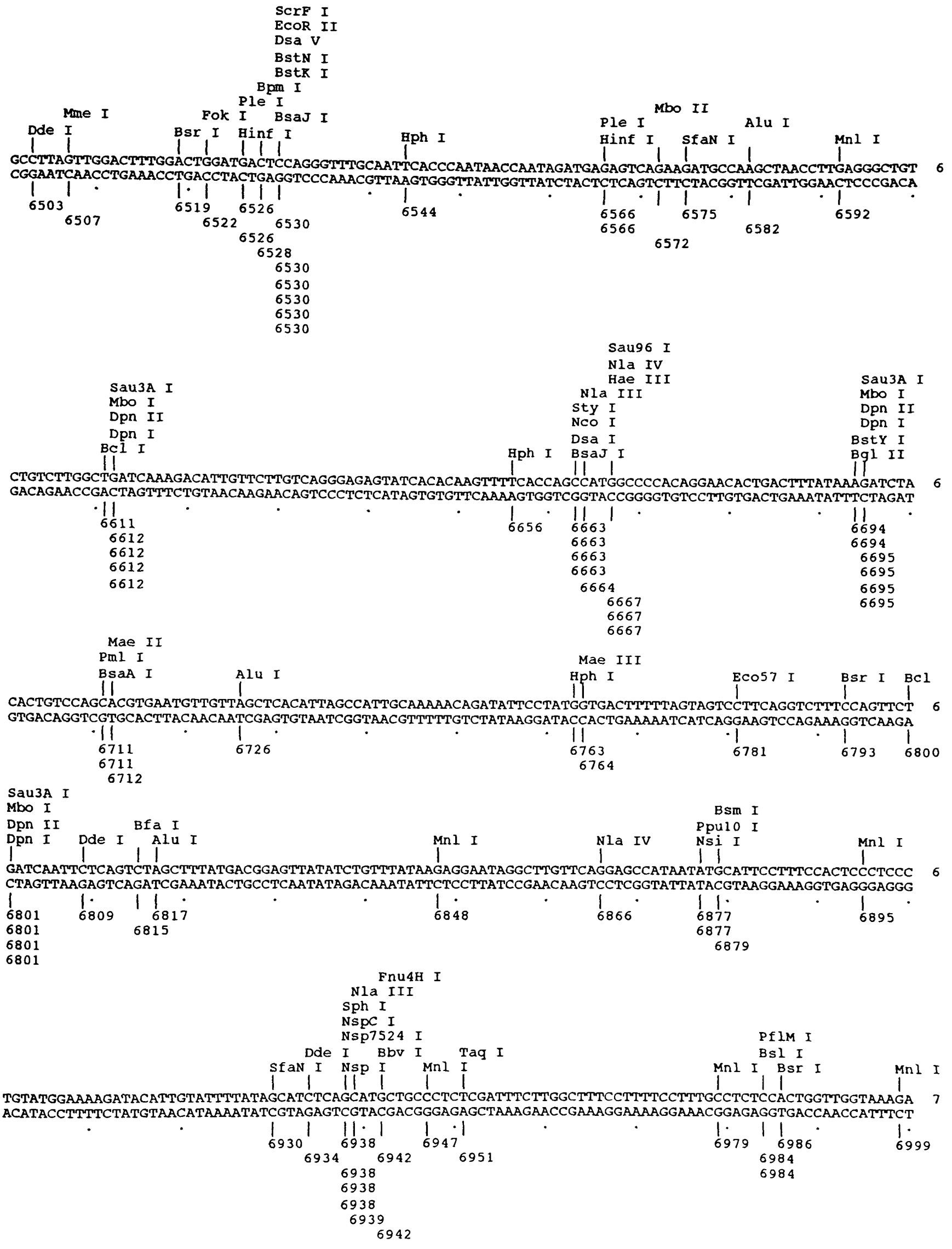

Fnu4H I
Sap I
Mae III
Mbo II
Bfa I
Mse I Hph I Bbv I
Ear I
Mae III Spe I
TTAAGTAGGTGACAGGCAGCCACATTCTATTAGATTTGAAGAGCCATTTAGTTCCTTTTATGCTGTTTCCTTATCAGTCACAAATGAACTAGTAATGGTG 3
AATTCATCCACTGTCCGTCGGTGTAAGATAATCTAAACTTCTCGGTAAATCAAGGAAAATACGACAAAGGAATAGTCAGTGTTTACTTGATCATTACCAC
3701 3708 3716 3738 3777 3788
3709 3738 3789
3716 3738
Mse I
Hpa I
Mse I
BspW I
Ple I
Mse I Dde I Mnl I Hinc II Mnl I Mnl I Hinf I
TTTACTTAAAGGGTTTCTGAGCGTTCAGGAGGTGCGTTAACAGCATAGGAAACCAATCCCTCCAATTTGTTCCCTCTGGGACTCTTTAACAATGTCATTC 3
AAATGAATTTCCCAAAGACTCGCAAGTCCTCCACGCAATTGTCGTATCCTTTGGTTAGGGAGGTTAAACAAGGGAGACCCTGAGAAATTGTTACAGTAAG
3806 3817 3829 3836 3859 3873 3880
3834 3880
3836 3886
3837
Rsa I
Mbo II
Csp6 I
Tth111 II
Mse I
Mnl I
Sca I
Nla III
Afl II
Taq I
Mae
ACTGCATTTCCAGAGGAACTCTTGGAAGTACTTTTGTCATGCTTGAAGAAGGGTAAGTGTGTAGGGTCTTAAGGAACACCAGAAAATCGACTACTTCCTG 4
TGACGTAAAGGTCTCCTTGAGAACCTTCATGAAAACAGTACGAACTTCTTCCCATTCACACATCCCAGAATTCCTTGTGGTCTTTTAGCTGATGAAGGAC
3913 3927 3938 3968 3987 4000
3928 3940 3969
3928 3945
Mae II
Taq I
Dde I Mae III
Sal I
Alu I
Alu I
Mse I Esp3 I Sty I
Hinc II
Nla IV
Mnl I
Afl II BsmA I BsaJ I
Acc I
Ban I Nla III BsaJ I
Mae II
TGACTTAAGACTGAGACGTTACCTTGGAACACAGTCGACAGGTTTACTCAACAGTGTGGTGCCAGCTCATGCCAAAAGTCCTCGGAGCTAACGCCCCGTG 4
ACTGAATTCTGACTCTGCAATGGAACCTTGTGTCAGCTGTCCAAATGAGTTGTCACACCACGGTCGAGTACGGTTTTCAGGAGCCTCGATTGCGGGGCAC
4004 4013 4022 4034 4058 4068 4080 4098
4005 4013 4022 4034 4058 4080
4011 4018 4034 4064 4086
4016 4035
Sau96 I
Nla IV
Ava II
Nla III
Sty I
Nco I
Nla III
Dsa I
BspH I
Rsa I
Mse I
Hpa I
Alu
BsaJ I
EcoR V
Csp6 I
Hinc II
Tth111 II
Hinc II
Bpm I
ACCCATGGACCCTGATATCATGAACTCTATTCTGTACTTTGGGCATAATGTCAACTACATTGCTTGGCTCTTTTCTGTTAACAGTTCTGCTCTTTCTCCA 4
TGGGTACCTGGGACTATAGTACTTGAGATAAGACATGAAACCCGTATTACAGTTGATGTAACGAACCGAGAAAAGACAATTGTCAAGACGAGAAAGAGGT
4103 4114 4134 4150 4161 4177 4196
4103 4118 4134 4177 4200
4103 4119 4178
4103
4104
4107
4107
4107
Ple I
Hinf I
Msp I
Hpa II
Dde I
Ple I
Mse I
BsaW I
Mnl I
Hinf I
Dra I
GCTCCGGTTTGACTGCCTCAGACTCAAAGCAACTGACTCCACCCCGACTTTAAAAAAACAAAAAAACAAAAACAGACAAATGCCAACTCTGATTTTATTATA 4
CGAGGCCAAACTGACGGAGTCTGAGTTTCGTTGACTGAGGTGGGGCTGAAATTTTTTGTTTTTTGTTTTTGTCTGTTTACGGTTGAGACTAAAATAATAT
4203 4216 4235 4249
4204 4217 4235 4250
4204 4221
4221

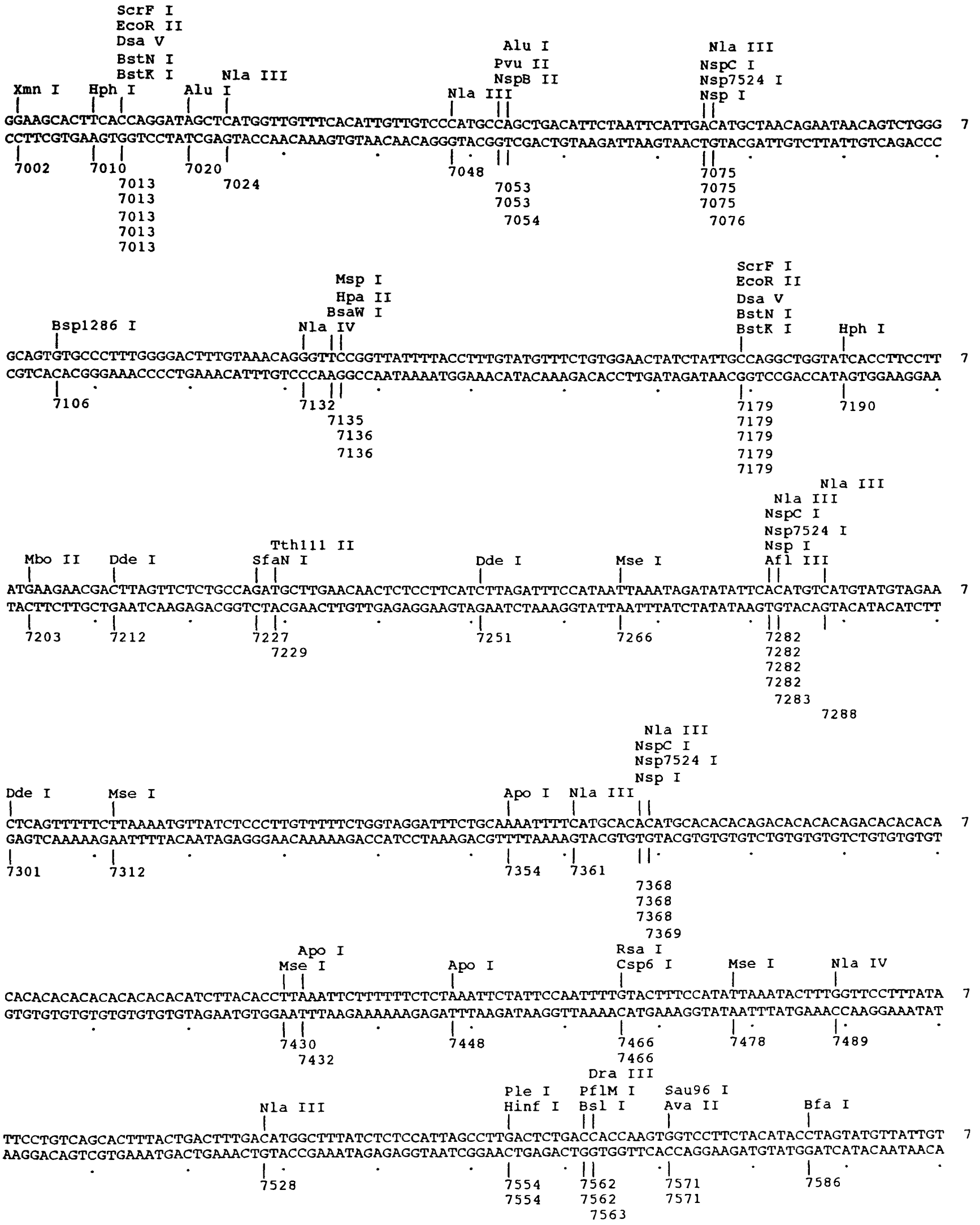

BstU I
HinP I
Hha I
Nla III
Sty I
Nco I
Dsa I
BsaJ I
Alu I
Hind III
Mse I
Tth1
TTTTATTTTTTTAAGCTTCCTGCTAAAGACTAACTGTATGGTTTTACCATTTTTATCCATGGCGCGTATTTCTCTTGCCCTTTGCTAACCCTATTAGATT 4
AAAATAAAAAAATTCGAAGGACGATTTCTGATTGACATACCAAAATGGTAAAAATAGGTACCGCGCATAAAGAGAACGGGAAACGATTGGGATAATCTAA
4311
4313
4314
4357
4357
4357
4357
4358
4362
4362
4363
4400
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Hae III
Stu I
Hae I
Sty I
BsaJ I
BspW I
Nla III
Fnu4H I
SfaN I
Bbv I
Fok I
Alu I
Fnu4H I
Bbv I
SfaN I
BspW I
Fok I
Bfa I
Mnl I
Bsl I
Mme I
Mnl I
Mae III
Bsg I
GTTTGCCTAGTAGCATCCCCTGCCTCTGCCTGGCTGTTGGAGGCCTTTACTCGCAGCCTTGGATGCCATGTAGGTTTTCTGTAACCGAAGTGCAGCTTGT 4
CAAACGGATCATCGTAGGGGACGGAGACGGACCGACAACCTCCGGAAATGAGCGTCGGAACCTACGGTACATCCAAAAGACATTGGCTTCACGTCGAACA
4407
4414
4413
4413
4423
4423
4429
4429
4429
4429
4429
4436
4440
4441
4441
4442
4453
4453
4456
4457
4457
4461
4462
4467
4481
4490
4492
4492
4494
Hae III
Msc I
Hae I
HgiA I
Nde I
Bsp1286 I
Eae I
Mbo II
Bbs I
Bfa I
Bsr I
Mse I
Dde I
Hinc II
Mme I
Tfi I
Hinf I
TAGTGCTCTTTGCATATGGCCAAGCCAACAAGTCTAGTGTCTTCCACCCAGTTAATCAAAAAGATGGCTTAGGGTTGTCAACAGAGTTGGATTCTGTGGG 4
ATCACGAGAAACGTATACCGGTTCGGTTGTTCAGATCACAGAAGGTGGGTCAATTAGTTTTTCTACCGAATCCCAACAGTTGTCTCAACCTAAGACACCC
4503
4503
4513
4517
4517
4517
4518
4534
4539
4540
4548
4552
4568
4577
4586
4590
4590
Mbo II
Ear I
Eco57 I
Mse I
Rsa I
Csp6 I
Tth111 II
Alu I
CTTCAGGAAGAGAAATGAAAGAGAAAAGAAAAGGAATTAAGGGAGAAAATAGTACACATACAAAATTACAAGCAGTCCATTAGCTCTCATTACAACTTCA 4
GAAGTCCTTCTCTTTACTTTCTCTTTTCTTTTCCTTAATTCCCTCTTTTATCATGTGTATGTTTTAATGTTCGTCAGGTAATCGAGAGTAATGTTGAAGT
4601
4607
4607
4637
4652
4652
4669
4682
Sau96 I
Nla IV
Tfi I
Hinf I
Nla III
Hae III
Hph I
Mnl I
Bfa I
PshA I
Alu I
Fok
CTGAATATCAGCAGTAATCCATGCAATAAATATATCGGCCCCTAATGGTGAATCCTCCTTCTAGTCTGACTACTGTCAGAAACCTATCATAGCTTTCTCC 4
GACTTATAGTCGTCATTAGGTACGTTATTTATATAGCCGGGGATTACCACTTAGGAGGAAGATCAGACTGATGACAGTCTTTGGATAGTATCGAAAGAGG
4720
4737
4737
4737
4747
4750
4750
4754
4761
4768
4791
4800
Sau3A I
Msp I
Mbo I
Hpa II
Mbo II
BspE I
Dpn II
BsaW I
Dpn I
ATCCGGATTGATCTTCTCTATTTTTTTTTCCTATGTTGGCTTCTTGTTCACAATTCCATTTGGCAAAGAATAAAGGAGTTCTCTTACTAAAACAACCAGA 4
TAGGCCTAACTAGAAGAGATAAAAAAAAAGGATACAACCGAAGAACAAGTGTTAAGGTAAACCGTTTCTTATTTCCTCAAGAGAATGATTTTGTTGGTCT
4802
4802
4803
4803
4810
4810
4812
4810
4810

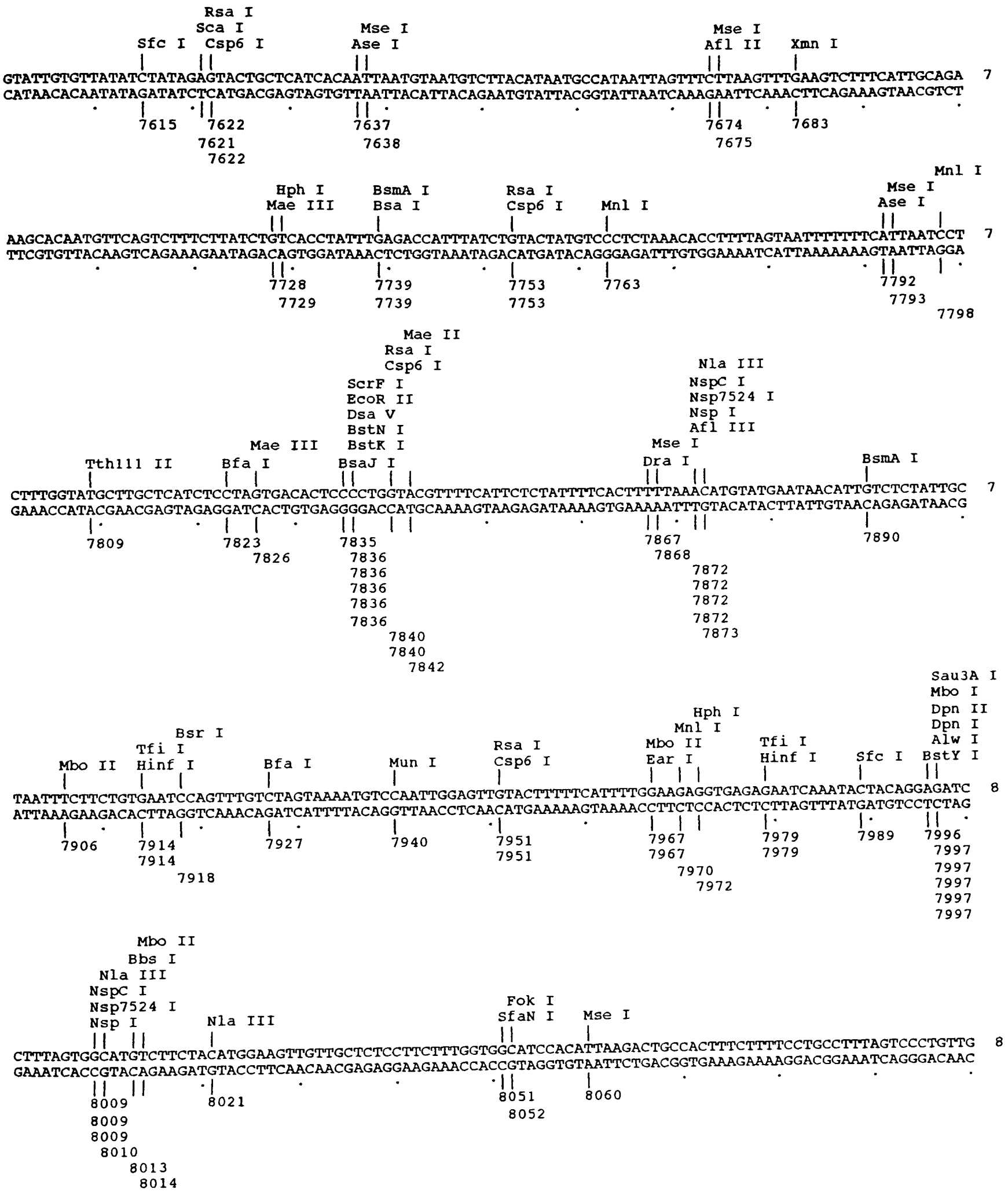

ScrF I
EcoR II
Dsa V
BstN I
BstK I
Fnu4H I
Bbv I
SfaN I
Nla III
Mnl I
Mse I
Bsg I
AATCAAGTGTTTTACATGAATGATGCAAGAGGTTCTAAAGACACAGAGTTAAAACCAGGAAAGTTCTGCACAGCAGCCTGGTCTTTGGAGGCTATATCTA 5
TTAGTTCACAAAATGTACTTACTACGTTCTCCAAGATTTCTGTGTCTCAATTTTGGTCCTTTCAAGACGTGTCGTCGGACCAGAAACCTCCGATATAGAT
4915
4922
4929
4949
4955
4966
4973
4977
4988
Alu I
Pvu II
NspB II
Fnu4H I
Bbv I
Nla III
AlwN I
Ssp I
AAAATGACAATTTGTGATTGGTATGCCATATCATGTAGGGACAAATGACAGCAGCTGGTATAAATAACTATTGCATATTGGTTATTTGACTAAATAAATA 5
TTTTACTGTTAAACACTAACCATACGGTATAGTACATCCCTGTTTACTGTCGTCGACCATATTTATTGATAACGTATAACCAATAAACTGATTTATTTAT
5032
5049
5051
5052
5053
5097
Eco57 I
Mbo II
Sau3A I
Mbo I
Dpn II
Dpn I
Fnu4H I
Bbv I
Mae III
Bpm I
Mnl I
Dde I
Dra III
Mse I
TTCCTGTGATCTTCAGAAAGCAGCAAGACTTTTATAGGTTACTGAAGACCCTCCAGAAACCCCCACTTAGTGTGTTTTAATCTGGTTATTGCAGTTCTAT 5
AAGGACACTAGAAGTCTTTCGTCGTTCTGAAAATATCCAATGACTTCTGGGAGGTCTTTGGGGGTGAATCACACAAAATTAGACCAATAACGTCAAGATA
5108
5110
5111
5120
5138
5142
5144
5150
5151
5164
5166
5177
Mbo II
Bbs I
Alu I
Apo I
Mnl I
BsaJ I
SfaN I
Nla III
Mae III
Dde I
Mnl I
Dde I
GGAATGAACTTGAAGCTGACCAAACTCAAATTCCCCCAAATCAGATGCCGAGGCAAAGTTACATGACTGAGTGAACTGAGGAAGTGCTGGGGTGTGTGTG 5
CCTTACTTGAACTTCGACTGGTTTGAGTTTAAGGGGGTTTAGTCTACGGCTCCGTTTCAATGTACTGACTCACTTGACTCCTTCACGACCCCACACACAC
5214
5228
5244
5248
5250
5258
5262
5267
5276
5278
Hae III
Alu I
Stu I
Bsg I
Xmn I
Hae I
TGTGTGTGTGTGTGTGTGTGTGTGTGCAGTCCTGCCTGTATAGGTTATAAATCACAAGTTTGCCCACACAAAAGCAAGGAAGCTATTCCTTCAAGGCC 5
ACACACACACACACACACACACACACACGTCAGGACGGACATATCCAATATTTAGTGTTCAAACGGGTGTGTTTTCGTTCCTTCGATAAGGAAGTTCCGG
5326
5381
5383
5396
5397
Sau96 I
Ava II
Sty I
BsaJ I
Xcm I
Alu I
Hae III
Hae I
Mnl I
Fnu4H I
Bbv I
Sau96 I
Mnl I
BspW I
Hae III
TTGTATCAACAAGTCCTTGGTCCAGAGCTGTGCTGGGAGGCCAACTTGACAATTATTTATGCAGCCTTTACCAAAAGAAGCAAGGGAGGCCCAGAAGCCA 5
AACATAGTTGTTCAGGAACCAGGTCTCGACACGACCCTCCGGTTGAACTGTTAATAAATACGTCGGAAATGGTTTTCTTCGTTCCCTCCGGGTCTTCGGT
5415
5419
5422
5426
5437
5438
5439
5461
5480
5486
5488

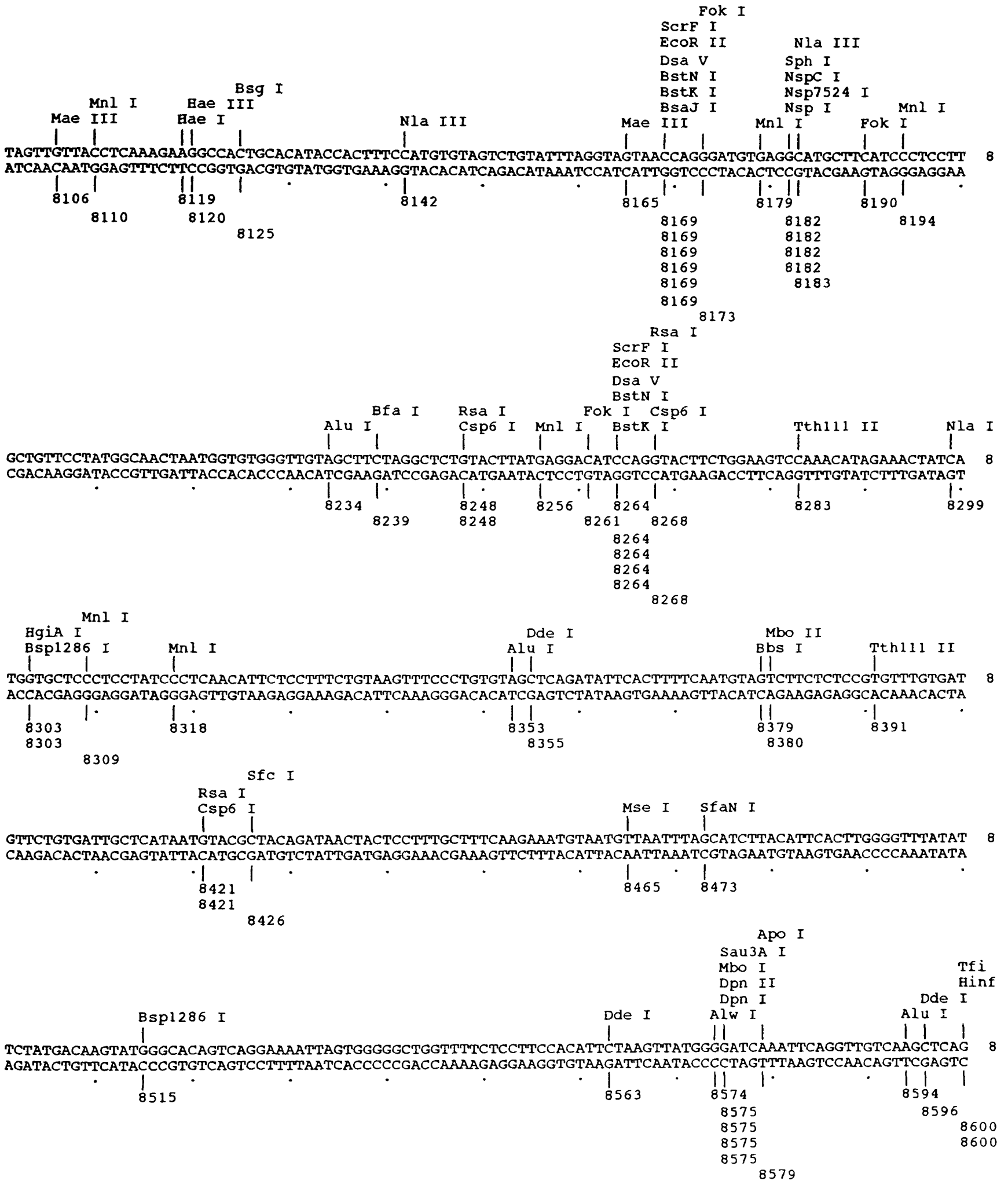

Mnl I
BsmA I
Bsa I
Xcm I
Dde I
Nla III
Hae III
Hae I
Mnl I
Nla III
Sfc I
Tfi I
Hinf I
TTGTGGAGACCTCCAAATCGCCTATGGAGTGTAAGGCTTTAGATATAAGCATAAAGGCCATGCCTCAGACATGGTGGGAATGGAAAAAAGAATCTACAGG 5
AACACCTCTGGAGGTTTAGCGGATACCTCACATTCCGAAATCTATATTCGTATTTCCGGTACGGAGTCTGTACCACCCTTACCTTTTTTCTTAGATGTCC
5506
5513
5506
5510
5555
5556
5559
5563
5564
5570
5590
5590
5594
Nla III
Sph I
NspC I
Tth111 II
Mbo II
Tfi I
Hinf I
HgiA I
Nsp7524 I
Bsp1286 I
Nla II
Dde I
Alu I
Alu I
Bsp1286 I
ApaL I
Nsp I
TCTAAGTGAATTACTGACAGGATTATCACAAGCTCACTTTGGTCTTGAATCTTCTCCAAGCTCCTGTGCCCACTTGTGTGTGGGTGCACAAGCATGCCAT 5
AGATTCACTTAATGACTGTCCTAATAGTGTTCGAGTGAAACCAGAACTTAGAAGAGGTTCGAGGACACGGGTGAACACACACCCACGTGTTCGTACGGTA
5602
5631
5647
5647
5650
5659
5666
5684
5684
5684
5689
5692
5692
5692
5693
5692
5698
Alu I
Dde I
Mbo II
Acc I
Ple I
Hinf I
Rsa I
Csp6 I
Ssp I
GTTCTCTGTTCTTCGGTATATGTGGTAGACAGATGTGTAAAGGGTTTCGTGTGAGTCGTGGTGTATTTTACACTGTACTTAGCTTTTATTATCATCAATA 5
CAAGAGACAAGAAGCCATATACACCATCTGTCTACACATTTCCCAAAGCACACTCAGCACCACATAAAATGTGACATGAATCGAAAATAATAGTAGTTAT
5710
5725
5753
5775
5778
5781
5797
Fok I
Ple I
Hinf I
Dde I
Eco57 I
Mbo II
Mse
TTGTGTCTTTCTGTCTGTCTTTTAGTAAATGCCTTTACCCACTGAGTCATCCCACCAGCCCCATTGTGTATTTTCTTCAGTGTATCTTTGTCATTTTAGT 5
AACACAGAAAGACAGACAGAAAATCATTTACGGAAATGGGTGACTCAGTAGGGTGGTCGGGGTAACACATAAAAGAAGTCACATAGAAACAGTAAAATCA
5842
5844
5844
5848
5874
5875
5900
Bsp1286 I
ScrF I
EcoR II
Dsa V
Sau3A I
BstN I
Mbo I
BstK I
Dpn II
BsaJ I
Dpn I
Mbo II
Alw I
Bbs I
Ban II
Nla III
Mnl I
Hae III
Msc I
Hae I
Eae I
TAAGGTGTGTGATCCAAGTCTTCCTGGGCTCTGTCTGTCCTTATTTTACTGCATATAGTTTTGCCTCACATGGCAATAATGATTGGCCAATCTCTATTGC 6
ATTCCACACACTAGGTTCAGAAGGACCCGAGACAGACAGGAATAAAATGACGTATATCAAAACGGAGTGTACCGTTATTACTAACCGGTTAGAGATAACG
5911
5911
5911
5911
5911
5911
5918
5919
5923
5923
5923
5923
5923
5923
5926
5926
5964
5969
5984
5984
5984
5985
Bfa I
Rsa I
Mnl I
Mnl I
Xba I
Csp6 I
Alu I
Aci I
TATATGAGGCATATCTATGTCTATGGAGGTATTTATCTAGATTAGGGATTTAGTTCGTCTTATGGTTTAGGTACAGAATGAATAGCTCAAATGCGGAAGG 6
ATATACTCCGTATAGATACAGATACCTCCATAAATAGATCTAATCCCTAAATCAAGCAGAATACCAAATCCATGTCTTACTTATCGAGTTTACGCCTTCC
6006
6026
6036
6037
6071
6071
6084
6093

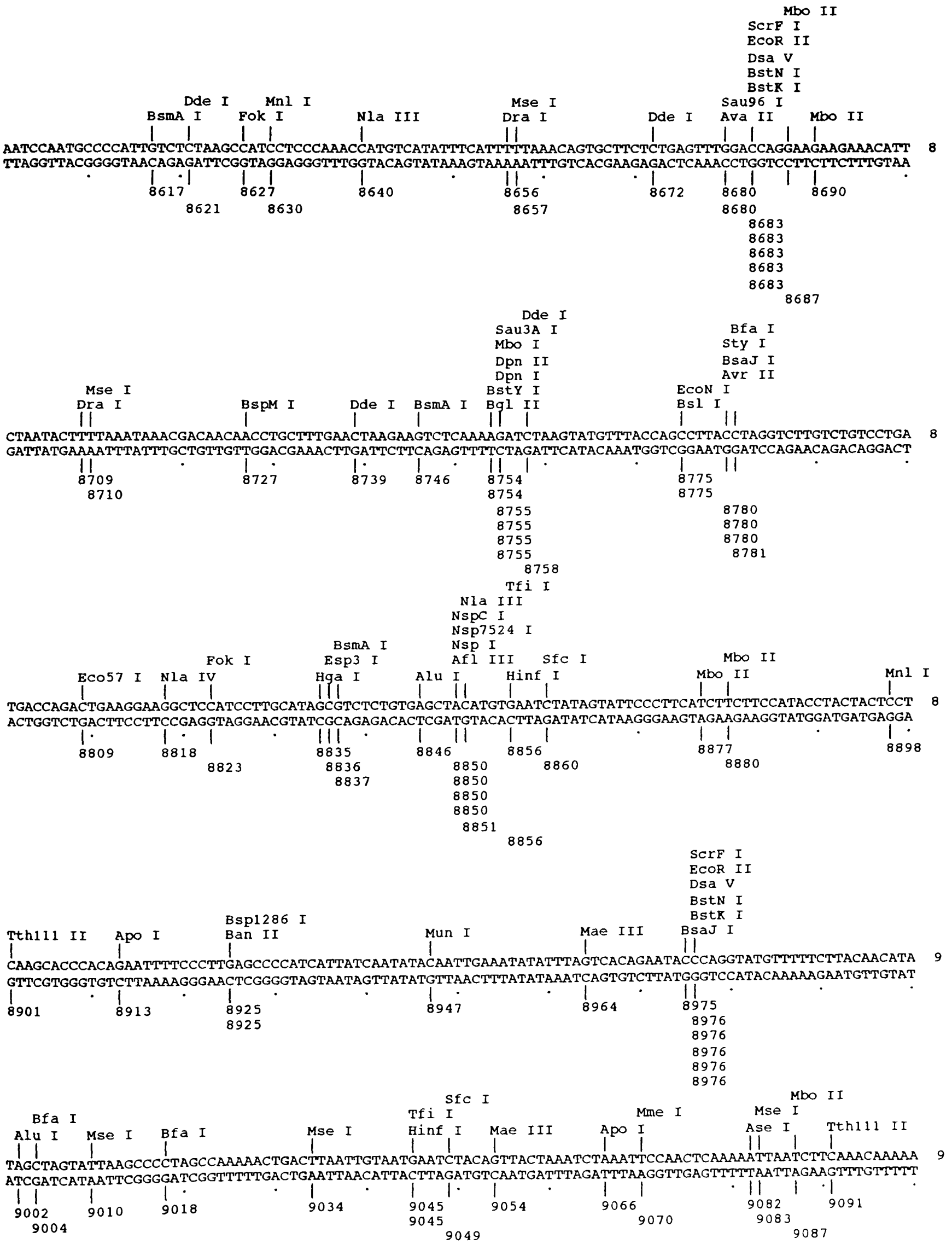

Fnu4H I
Pvu II
NspB II
AlwN I Fnu4H I
Sau3A I Bbv I
Mbo I Fnu4H I
Dpn II BspW I
Dpn I Bbv I
Alw I Alu I Aci I
Nla III
BspH I
Mnl I
Dde I
Dde I
Alu I
Pvu II
NspB II
Dde I
CTGATCCCCAGCTGCTGCGGCTGTTGGGTAGTGATTAGTTGTGAATGTTCACACCTCATGAATGGACTGAGTGGGTTCACAGCTGAGTGGCTTCTTAGGA 6
GACTAGGGGTCGACGACGCCGACAACCCATCACTAATCAACACTTACAAGTGTGGAGTACTTACCTGACTCACCCAAGTGTCGACTCACCGAAGAATCCT
6103 6110 6117
6103 6111
6103 6111
6103 6111
6103 6114
6109 6117
6109
6109
6114
6154
6156
6157
6167
6180
6180
6181
6183
6194
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Hae III
Nla III
Hae III
Msc I
Hae I
Eae I
Sau96 I
EcoO109 I
Bsr I
SfaN I
Acc I
AGTAGGGCCTGGTGGGAACAAGTGCATCACTGGGGGTGTGCCATCAAGGATACATCTTGCCACTGACTTTTTCCACTTTGTCTACTTCTTGGCCATGAAG 6
TCATCCCGGACCACCCTTGTTCACGTAGTGACCCCCACACGGTAGTTCCTATGTAGAACGGTGACTGAAAAAGGTGAAACAGATGAAGAACCGGTACTTC
6204 6224
6205 6229
6206
6208
6208
6208
6208
6280
6290
6290
6290
6291
6294
Fnu4H I
Nla III
Sty I
Nco I
Dsa I
BsaJ I
BspW I
BspW I
Hph I
Mnl I
Bsl I
Bbv I
BsmA I
Bfa I
Alu I
AlwN I
TGAGGTTTTTGCCTTGCCATGGCTGCTACCATAACACTTCTGTCTCAACACAGACCCACACACACTAGCAGAGCCAGCTCACCAGACACTGATGGCTCTG 6
ACTCCAAAAACGGAACGGTACCGACGATGGTATTGTGAAGACAGAGTTGTGTCTGGGTGTGTGTGATCGTCTCGGTCGAGTGGTCTGTGACTACCGAGAC
6302 6312 6316 6317 6317 6317 6317 6318 6322
6322
6342
6365 6368
6376 6379
6383
Nla III
NspC I
Nsp7524 I
Nsp I
Afl III
Sfc I
Pst I
Mbo II
Bbs I BspM I
Alu I Sse8337 I
Mse I
Alu I
Sfc I
AAATTATGACCCAAAATAGAGCTGTCTTCCTGCAGGTGTTTTGCTTAACATGTGTGTTGCTTTCTAATCACAGTGGAAGGTAGCTAACCACTGCTACAGA 6
TTTAATACTGGGTTTTATCTCGACAGAAGGACGTCCACAAAACGAATTGTACACACAACGAAAGATTAGTGTCACCTTCCATCGATTGGTGACGATGTCT
6420 6424 6425 6429 6430 6430 6432
6445 6448 6448 6448 6448 6449
6482
6494

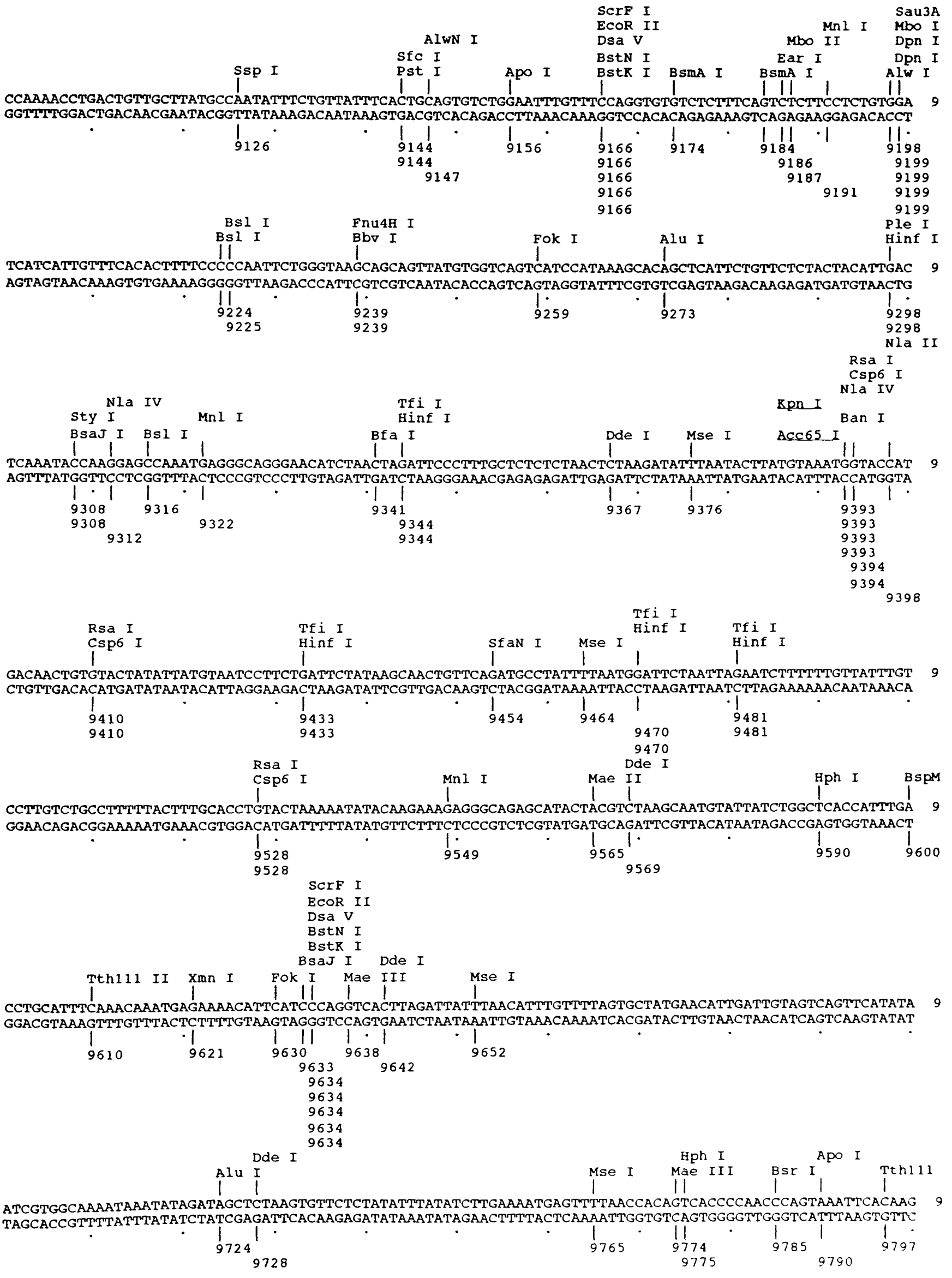

ScrF I
EcoR II
Dsa V
BstN I
BstK I
Bpm I
Ple I
Mme I
Fok I
BsaJ I
Mbo II
Dde I
Bsr I
Hinf I
Hph I
Ple I
Alu I
Hinf I
SfaN I
Mnl I
GCCTTAGTTGGACTTTGGACTGGATGACTCCAGGGTTTGCAATTCACCCAATAACCAATAGATGAGAGTCAGAAGATGCCAAGCTAACCTTGAGGGCTGT 6
CGGAATCAACCTGAAACCTGACCTACTGAGGTCCCAAACGTTAAGTGGGTTATTGGTTATCTACTCTCAGTCTTCTACGGTTCGATTGGAACTCCCGACA
6503
6507
6519
6522
6526
6526
6528
6530
6530
6530
6530
6530
6530
6544
6566
6566
6572
6575
6582
6592
Sau96 I
Nla IV
Hae III
Sau3A I
Mbo I
Dpn II
Dpn I
Nla III
Sty I
Nco I
Dsa I
BstY I
Bgl II
Bcl I
Hph I
BsaJ I
CTGTCTTGGCTGATCAAAGACATTGTTCTTGTCAGGGAGAGTATCACACAAGTTTTCACCAGCCATGGCCCCACAGGAACACTGACTTTATAAAGATCTA 6
GACAGAACCGACTAGTTTCTGTAACAAGAACAGTCCCTCTCATAGTGTGTTCAAAAGTGGTCGGTACCGGGGGTGTCCTTGTGACTGAAATATTTCTAGAT
6611
6612
6612
6612
6612
6656
6663
6663
6663
6663
6664
6667
6667
6694
6694
6695
6695
6695
6695
Mae II
Pml I
BsaA I
Alu I
Mae III
Hph I
Eco57 I
Bsr I
Bcl
CACTGTCCAGCACGTGAATGTTGTTAGCTCACATTAGCCATTGCAAAAACAGATATTCCTATGGTGACTTTTTAGTAGTCCTTCAGGTCTTTCCAGTTCT 6
GTGACAGGTCGTGCACTTACAACAATCGAGTGTAATCGGTAACGTTTTTGTCTATAAGGATACCACTGAAAAATCATCAGGAAGTCCAGAAAGGTCAAGA
6711
6711
6712
6726
6763
6764
6781
6793
6800
Sau3A I
Mbo I
Dpn II
Dpn I
Bfa I
Dde I
Alu I
Mnl I
Nla IV
Bsm I
Ppu10 I
Nsi I
Mnl I
GATCAATTCTCAGTCTAGCTTTATGACGGAGTTATATCTGTTTATAAGAGGAATAGGCTTGTTCAGGAGCCATAATATGCATTCCTTTCCACTCCCTCCC 6
CTAGTTAAGAGTCAGATCGAAATACTGCCTCAATATAGACAAATATTCTCCTTATCCGAACAAGTCCTCGGTATTATACGTAAGGAAAGGTGAGGGAGGG
6801
6801
6801
6801
6809
6815
6817
6848
6866
6877
6877
6879
6895
Fnu4H I
Nla III
Sph I
NspC I
Nsp7524 I
Dde I
Bbv I
Taq I
PflM I
Bsl I
SfaN I
Nsp I
Mnl I
Mnl I
Bsr I
Mnl I
TGTATGGAAAAGATACATTGTATTTTATAGCATCTCAGCATGCTGCCCTCTCGATTTCTTGGCTTTCCTTTTCCTTTGCCTCTCCACTGGTTGGTAAAGA 7
ACATACCTTTTCTATGTAACATAAAATATCGTAGAGTCGTACGACGGGAGAGCTAAAGAACCGAAAGGAAAAGGAAACGGAGAGGTGACCAACCATTTCT
6930
6934
6938
6938
6938
6938
6939
6942
6942
6947
6951
6979
6984
6984
6986
6999

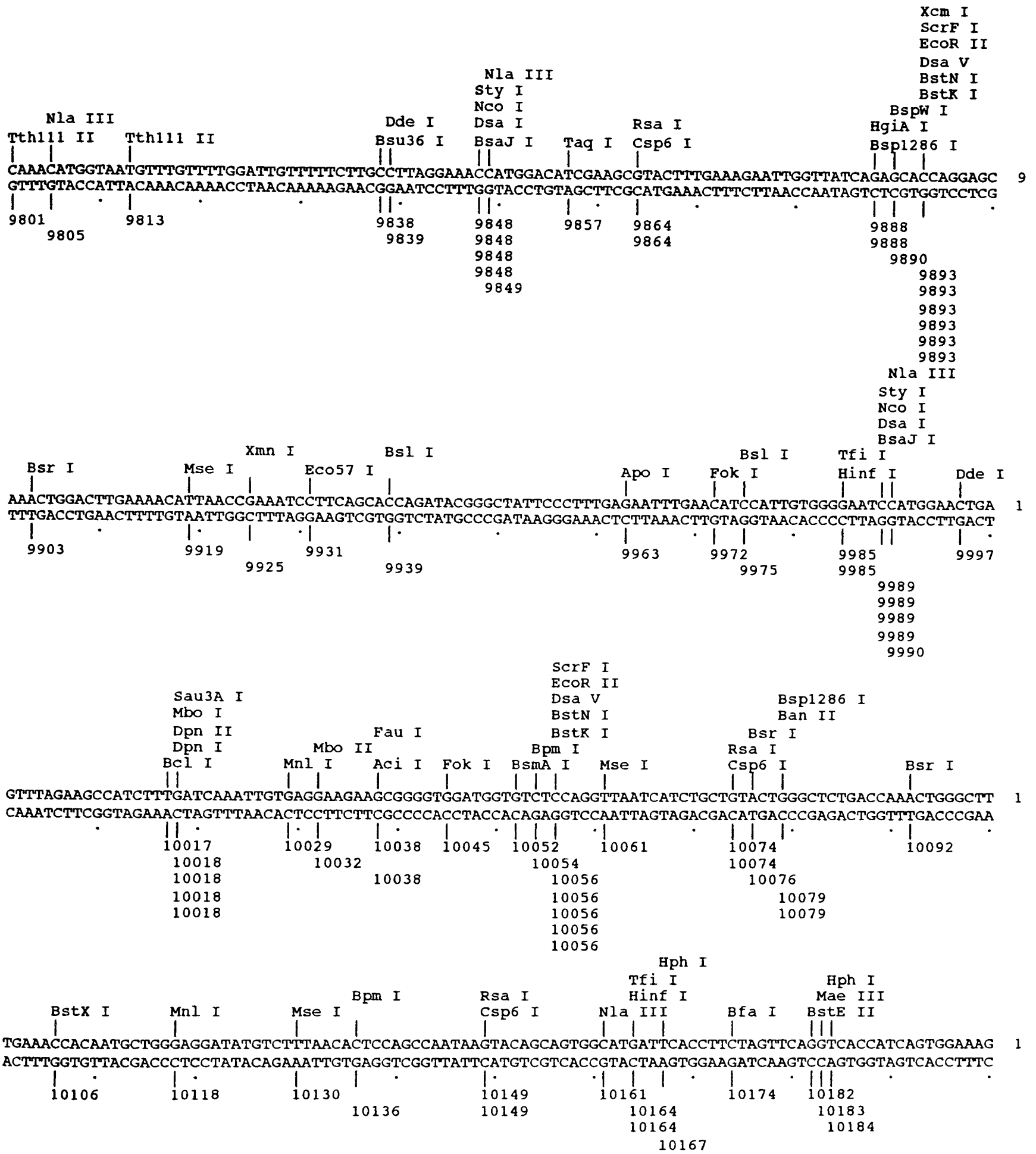

ScrF I
EcoR II
Dsa V
BstN I
BstK I
Nla III
Xmn I
Hph I
Alu I
Alu I
Pvu II
NspB II
Nla III
Nla III
NspC I
Nsp7524 I
Nsp I
GGAAGCACTTCACCAGGATAGCTCATGGTTGTTTCACATTGTTGTCCCATGCCAGCTGACATTCTAATTCATTGACATGCTAACAGAATAACAGTCTGGG 7
CCTTCGTGAAGTGGTCCTATCGAGTACCAACAAAGTGTAACAACAGGGTACGGTCGACTGTAAGATTAAGTAACTGTACGATTGTCTTATTGTCAGACCC
7002 7010 7013 7013 7013 7013 7013 7020 7024 7048 7053 7053 7054 7075 7075 7075 7076
Msp I
Hpa II
BsaW I
ScrF I
EcoR II
Dsa V
BstN I
Bsp1286 I
Nla IV
BstK I
Hph I
GCAGTGTGCCCTTTGGGGACTTTGTAAACAGGGTTCCGGTTATTTTACCTTTGTATGTTTCTGTGGAACTATCTATTGCCAGGCTGGTATCACCTTCCTT 7
CGTCACACGGGAAACCCCTGAAACATTTGTCCCAAGGCCAATAAAATGGAAACATACAAAGACACCTTGATAGATAACGGTCCGACCATAGTGGAAGGAA
7106 7132 7135 7136 7136 7179 7179 7179 7179 7179 7190
Nla III
Nla III
NspC I
Nsp7524 I
Nsp I
Tth111 II
Mbo II
Dde I
SfaN I
Dde I
Mse I
Afl III
ATGAAGAACGACTTAGTTCTCTGCCAGATGCTTGAACAACTCTCCTTCATCTTAGATTTCCATAATTAAATAGATATATTCACATGTCATGTATGTAGAA 7
TACTTCTTGCTGAATCAAGAGACGGTCTACGAACTTGTTGAGAGGAAGTAGAATCTAAAGGTATTAATTTATCTATATAAGTGTACAGTACATACATCTT
7203 7212 7227 7229 7251 7266 7282 7282 7282 7283 7288
Nla III
NspC I
Nsp7524 I
Nsp I
Dde I
Mse I
Apo I
Nla III
CTCAGTTTTTCTTAAAATGTTATCTCCCTTGTTTTTCTGGTAGGATTTCTGCAAAATTTTCATGCACACATGCACACACAGACACACACAGACACACACA 7
GAGTCAAAAAGAATTTTACAATAGAGGGAACAAAAAGACCATCCTAAAGACGTTTTAAAAGTACGTGTGTACGTGTGTGTCTGTGTGTGTCTGTGTGTGT
7301 7312 7354 7361 7368 7368 7368 7369
Apo I
Mse I
Apo I
Rsa I
Csp6 I
Mse I
Nla IV
CACACACACACACACACACATCTTACACCTTAAATTCTTTTTTTCTCTAAATTCTATTCCAATTTTGTACTTTCCATATTAAATACTTTGGTTCCTTTATA 7
GTGTGTGTGTGTGTGTGTGTAGAATGTGGAATTTAAGAAAAAAGAGATTTAAGATAAGGTTAAAACATGAAAGGTATAATTTATGAAACCAAGGAAATAT
7430 7432 7448 7466 7466 7478 7489
Dra III
Ple I
Hinf I
PflM I
Bsl I
Sau96 I
Ava II
Nla III
Bfa I
TTCCTGTCAGCACTTTACTGACTTTGACATGGCTTTATCTCTCCATTAGCCTTGACTCTGACCACCAAGTGGTCCTTCTACATACCTAGTATGTTATTGT 7
AAGGACAGTCGTGAAATGACTGAAACTGTACCGAAATAGAGAGGTAATCGGAACTGAGACTGGTGGTTCACCAGGAAGATGTATGGATCATACAATAACA
7528 7554 7554 7562 7562 7563 7571 7571 7586

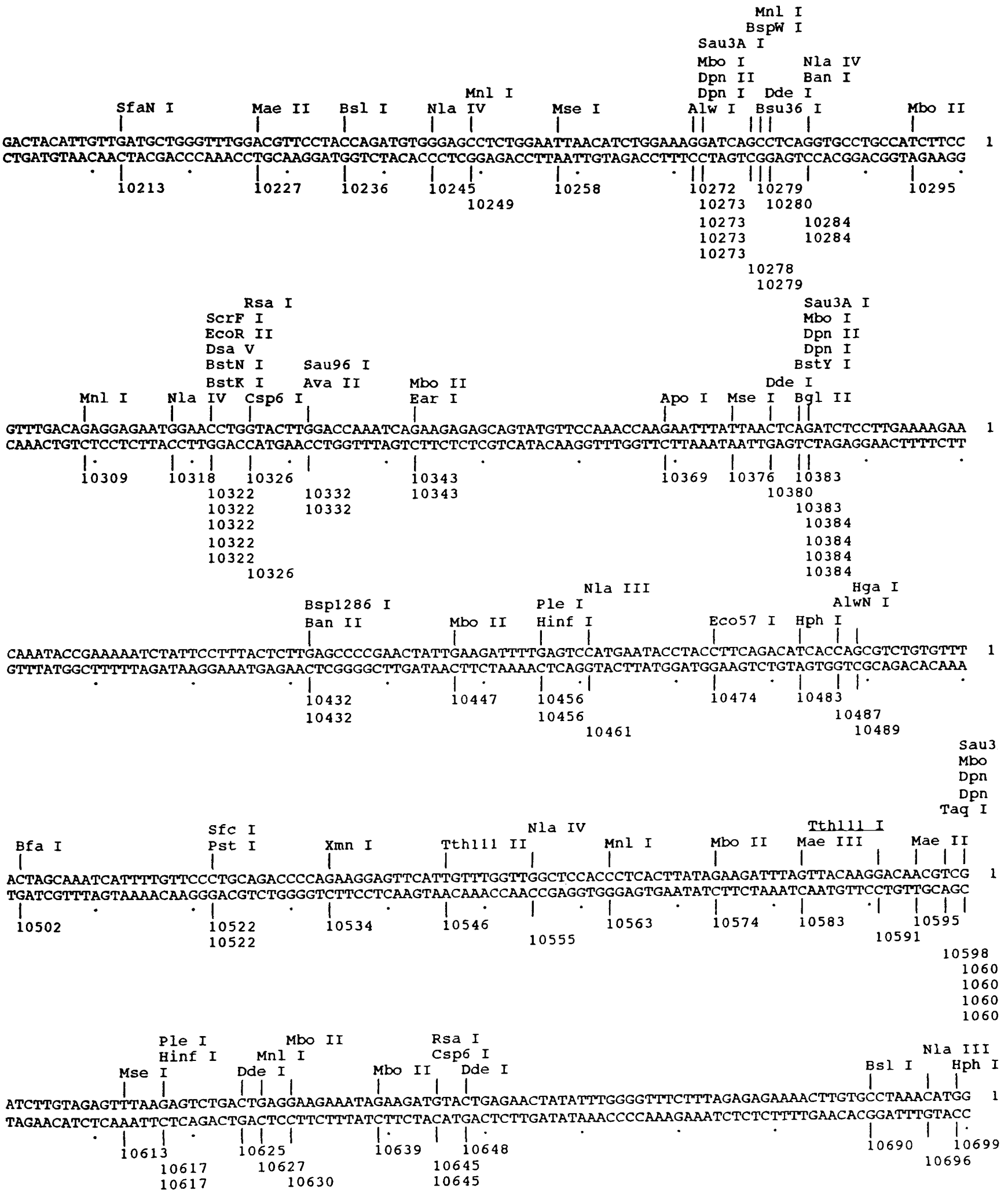

Rsa I
Sca I
Sfc I Csp6 I
Mse I
Ase I
Mse I
Afl II
Xmn I
GTATTGTGTTATATCTATAGAGTACTGCTCATCACAATTAATGTAATGTCTTACATAATGCCATAATTAGTTTCTTAAGTTTGAAGTCTTTCATTGCAGA 7
CATAACACAATATAGATATCTCATGACGAGTAGTGTTAATTACATTACAGAATGTATTACGGTATTAATCAAAGAATTCAAACTTCAGAAAGTAACGTCT
7615 7622 7637 7674 7683
7621 7638 7675
7622
Mnl I
Hph I
Mae III
BsmA I
Bsa I
Rsa I
Csp6 I
Mnl I
Mse I
Ase I
AAGCACAATGTTCAGTCTTTCTTATCTGTCACCTATTTGAGACCATTTATCTGTACTATGTCCCTCTAAACACCTTTTAGTAATTTTTTTCATTAATCCT 7
TTCGTGTTACAAGTCAGAAAGAATAGACAGTGGATAAACTCTGGTAAATAGACATGATACAGGGAGATTTGTGGAAAATCATTAAAAAAAGTAATTAGGA
7728 7739 7753 7763 7792
7729 7739 7753 7793
7798
Mae II
Rsa I
Csp6 I
Nla III
ScrF I
EcoR II
Dsa V
BstN I
BstK I
NspC I
Nsp7524 I
Nsp I
Afl III
Mae III
Mse I
Tth111 II
Bfa I
BsaJ I
Dra I
BsmA I
CTTTGGTATGCTTGCTCATCTCCTAGTGACACTCCCCTGGTACGTTTTCATTCTCTATTTTCACTTTTTAAACATGTATGAATAACATTGTCTCTATTGC 7
GAAACCATACGAACGAGTAGAGGATCACTGTGAGGGGACCATGCAAAAGTAAGAGATAAAAGTGAAAAATTTGTACATACTTATTGTAACAGAGATAACG
7809 7823 7835 7867 7890
7826 7836 7868
7836 7872
7836 7872
7840 7872 7873
7840
7842
Sau3A I
Mbo I
Dpn II
Dpn I
Alw I
Hph I
Mnl I
Bsr I
Tfi I
Mbo II
Hinf I
Bfa I
Mun I
Rsa I
Csp6 I
Mbo II
Ear I
Tfi I
Hinf I
Sfc I
BstY I
TAATTTCTTCTGTGAATCCAGTTTGTCTAGTAAAATGTCCAATTGGAGTTGTACTTTTTCATTTTGGAAGAGGTGAGAGAATCAAATACTACAGGAGATC 8
ATTAAAGAAGACACTTAGGTCAAACAGATCATTTTACAGGTTAACCTCAACATGAAAAAGTAAAACCTTCTCCACTCTCTTAGTTTATGATGTCCTCTAG
7906 7914 7927 7940 7951 7967 7979 7989 7996
7914 7951 7967 7979 7997
7918 7970 7997
7972 7997
7997
7997
Mbo II
Bbs I
Nla III
NspC I
Nsp7524 I
Nsp I
Nla III
Fok I
SfaN I
Mse I
CTTTAGTGGCATGTCTTCTACATGGAAGTTGTTGCTCTCCTTCTTTGGTGGCATCCACATTAAGACTGCCACTTTCTTTTCCTGCCTTTAGTCCCTGTTG 8
GAAATCACCGTACAGAAGATGTACCTTCAACAACGAGAGGAAGAAACCACCGTAGGTGTAATTCTGACGGTGAAAGAAAAGGACGGAAATCAGGGACAAC
8009 8021 8051 8060
8009 8052
8009
8010
8013
8014

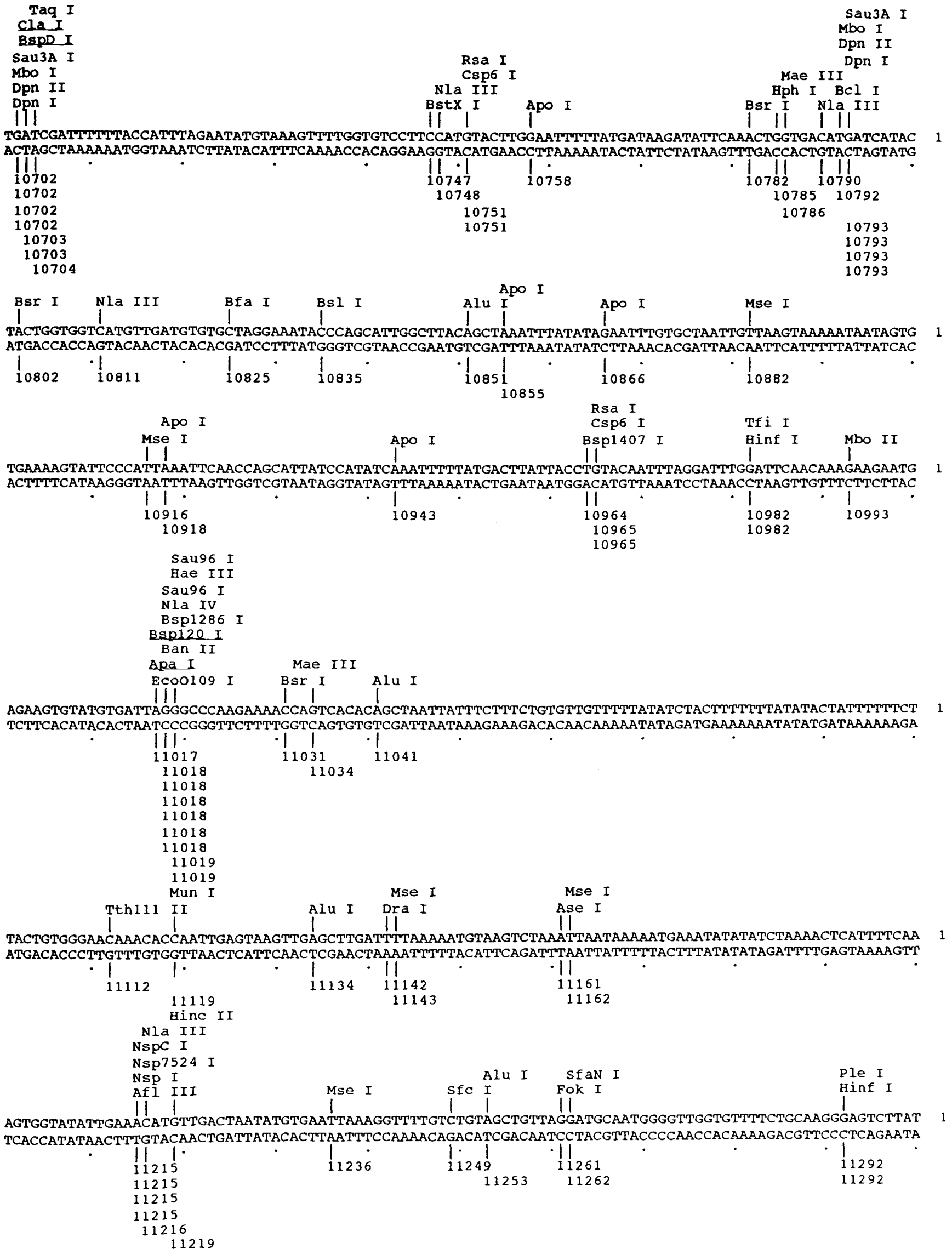

Fok I
ScrF I
EcoR II
Nla III
Dsa V
Sph I
BstN I
NspC I
BstK I
Nsp7524 I
BsaJ I
Nsp I
Mnl I
Bsg I
Mnl I
Hae III
Mae III
Hae I
Nla III
Mae III
Mnl I
Fok I
TAGTTGTTACCTCAAAGAAGGCCACTGCACATACCACTTTCCATGTGTAGTCTGTATTTAGGTAGTAACCAGGGATGTGAGGCATGCTTCATCCCTCCTT 8
ATCAACAATGGAGTTTCTTCCGGTGACGTGTATGGTGAAAGGTACACATCAGACATAAATCCATCATTGGTCCCTACACTCCGTACGAAGTAGGGAGGAA
8106
8119
8142
8165
8179
8190
8110
8120
8125
8169
8169
8169
8169
8169
8169
8173
8182
8182
8182
8182
8183
8194
Rsa I
ScrF I
EcoR II
Dsa V
BstN I
Bfa I
Rsa I
Fok I
Csp6 I
Alu I
Csp6 I
Mnl I
BstK I
Tth111 II
Nla I
GCTGTTCCTATGGCAACTAATGGTGTGGGTTGTAGCTTCTAGGCTCTGTACTTATGAGGACATCCAGGTACTTCTGGAAGTCCAAACATAGAAACTATCA 8
CGACAAGGATACCGTTGATTACCACACCCAACATCGAAGATCCGAGACATGAATACTCCTGTAGGTCCATGAAGACCTTCAGGTTTGTATCTTTGATAGT
8234
8239
8248
8248
8256
8261
8264
8264
8264
8264
8264
8268
8268
8283
8299
Mnl I
HgiA I
Bsp1286 I
Mnl I
Dde I
Alu I
Mbo II
Bbs I
Tth111 II
TGGTGCTCCCTCCTATCCCTCAACATTCTCCTTTCTGTAAGTTTCCCTGTGTAGCTCAGATATTCACTTTTCAATGTAGTCTTCTCTCCGTGTTTGTGAT 8
ACCACGAGGGAGGATAGGGAGTTGTAAGAGGAAAGACATTCAAAGGGACACATCGAGTCTATAAGTGAAAAGTTACATCAGAAGAGAGGCACAAACACTA
8303
8303
8309
8318
8353
8355
8379
8380
8391
Sfc I
Rsa I
Csp6 I
Mse I
SfaN I
GTTCTGTGATTGCTCATAATGTACGCTACAGATAACTACTCCTTTGCTTTCAAGAAATGTAATGTTAATTTAGCATCTTACATTCACTTGGGGTTTATAT 8
CAAGACACTAACGAGTATTACATGCGATGTCTATTGATGAGGAAACGAAAGTTCTTTACATTACAATTAAATCGTAGAATGTAAGTGAACCCCAAATATA
8421
8421
8426
8465
8473
Apo I
Sau3A I
Mbo I
Dpn II
Dpn I
Tfi
Hinf
Dde I
Bsp1286 I
Dde I
Alw I
Alu I
TCTATGACAAGTATGGGCACAGTCAGGAAAATTAGTGGGGGCTGGTTTTCTCCTTCCACATTCTAAGTTATGGGGATCAAATTCAGGTTGTCAAGCTCAG 8
AGATACTGTTCATACCCGTGTCAGTCCTTTTAATCACCCCCGACCAAAAGAGGAAGGTGTAAGATTCAATACCCCTAGTTTAAGTCCAACAGTTCGAGTC
8515
8563
8574
8575
8575
8575
8575
8579
8594
8596
8600
8600

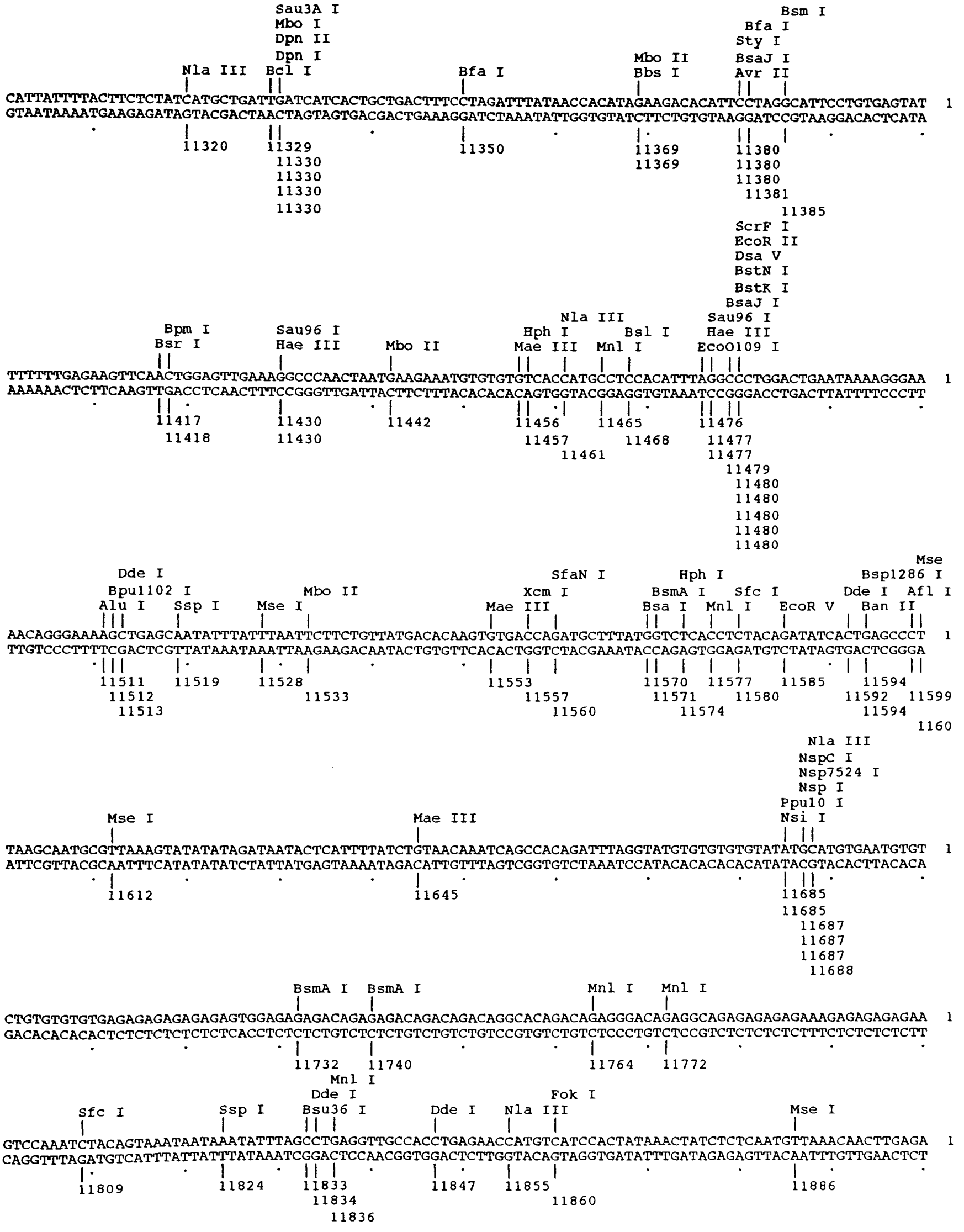

Mbo II
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Dde I
Mnl I
Mse I
Sau96 I
BsmA I
Fok I
Nla III
Dra I
Dde I
Ava II
Mbo II
AATCCAATGCCCCATTGTCTCTAAGCCATCCTCCCAAACCATGTCATATTTCATTTTTAAACAGTGCTTCTCTGAGTTTGGACCAGGAAGAAGAAACATT 8
TTAGGTTACGGGGTAACAGAGATTCGGTAGGAGGGTTTGGTACAGTATAAAGTAAAAAATTTGTCACGAAGAGACTCAAACCTGGTCCTTCTTCTTTGTAA
8617 8627 8640 8656 8672 8680 8690
8621 8630 8657 8680
8683
8683
8683
8683
8683
8687
Dde I
Sau3A I
Mbo I
Dpn II
Dpn I
Bfa I
Sty I
BsaJ I
Avr II
Mse I
Dra I
BstY I
Bgl II
BspM I
Dde I
BsmA I
EcoN I
Bsl I
CTAATACTTTTAAATAAACGACAACAACCTGCTTTGAACTAAGAAGTCTCAAAAGATCTAAGTATGTTTACCAGCCTTACCTAGGTCTTGTCTGTCCTGA 8
GATTATGAAAATTTATTTGCTGTTGTTGGACGAAACTTGATTCTTCAGAGTTTTCTAGATTCATACAAATGGTCGGAATGGATCCAGAACAGACAGGACT
8709 8727 8739 8746 8754 8775
8710 8754 8775
8755
8755
8755
8755
8758
8780
8780
8780
8781
Tfi I
Nla III
NspC I
Nsp7524 I
BsmA I
Nsp I
Fok I
Esp3 I
Afl III
Sfc I
Mbo II
Eco57 I
Nla IV
Hga I
Alu I
Hinf I
Mbo II
Mnl I
TGACCAGACTGAAGGAAGGCTCCATCCTTGCATAGCGTCTCTGTGAGCTACATGTGAATCTATAGTATTCCCTTCATCTTCTTCCATACCTACTACTCCT 8
ACTGGTCTGACTTCCTTCCGAGGTAGGAACGTATCGCAGAGACACTCGATGTACACTTAGATATCATAAGGGAAGTAGAAGAAGGTATGGATGATGAGGA
8809 8818 8835 8846 8856 8877 8898
8823 8836 8837 8850 8850 8850 8850 8851 8856 8860 8880
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Bsp1286 I
Tth111 II
Apo I
Ban II
Mun I
Mae III
BsaJ I
CAAGCACCCACAGAATTTTCCCTTGAGCCCCATCATTATCAATATACAATTGAAATATATTTAGTCACAGAATACCCAGGTATGTTTTTCTTACAACATA 9
GTTCGTGGGTGTCTTAAAAGGGAACTCGGGGTAGTAATAGTTATATGTTAACTTTATATAAATCAGTGTCTTATGGGTCCATACAAAAAGAATGTTGTAT
8901 8913 8925 8947 8964 8975
8925 8976
8976
8976
8976
8976
Sfc I
Mbo II
Bfa I
Tfi I
Mme I
Mse I
Alu I
Mse I
Bfa I
Mse I
Hinf I
Mae III
Apo I
Ase I
Tth111 II
TAGCTAGTATTAAGCCCCTAGCCAAAAACTGACTTAATTGTAATGAATCTACAGTTACTAAATCTAAATTCCAACTCAAAAATTAATCTTCAAACAAAAA 9
ATCGATCATAATTCGGGGATCGGTTTTTGACTGAATTAACATTACTTAGATGTCAATGATTTAGATTTAAGGTTGAGTTTTTAATTAGAAGTTTGTTTTT
9002 9010 9018 9034 9045 9054 9066 9082 9091
9004 9045 9070 9083
9049 9087

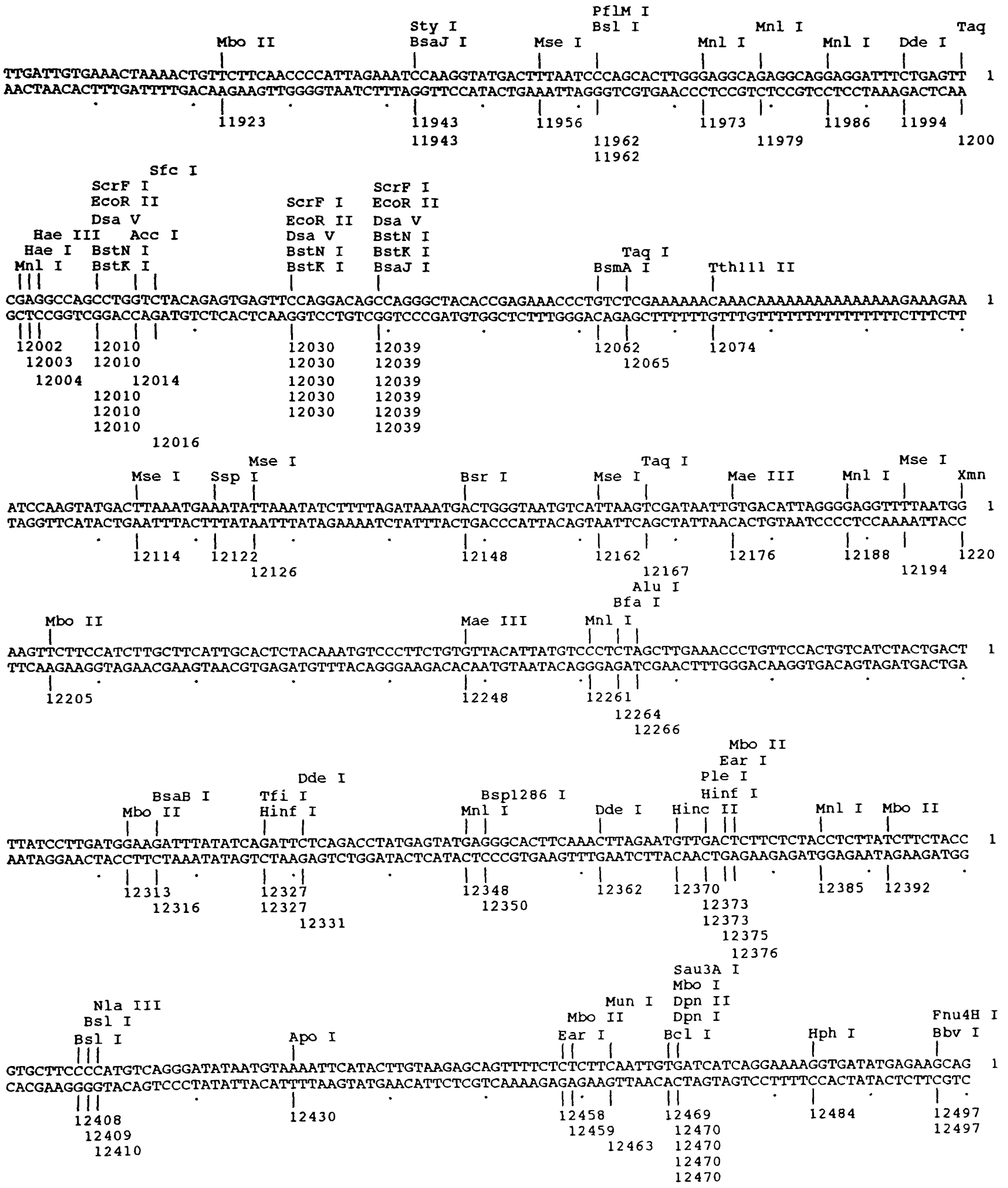

ScrF I
EcoR II
Dsa V
BstN I
BstK I
Sau3A
Mbo I
Dpn I
Dpn I
Alw I
AlwN I
Sfc I
Pst I
Ssp I
Apo I
BsmA I
Mnl I
Mbo II
Ear I
BsmA I
CCAAAACCTGACTGTTGCTTATGCCAATATTTCTGTTATTTCACTGCAGTGTCTGGAATTTGTTTCCAGGTGTGTCTCTTTCAGTCTCTTCCTCTGTGGA 9
GGTTTTGGACTGACAACGAATACGGTTATAAAGACAATAAAGTGACGTCACAGACCTTAAACAAAGGTCCACACAGAGAAAGTCAGAGAAGGAGACACCT
9126
9144
9144
9147
9156
9166
9166
9166
9166
9166
9174
9184
9186
9187
9191
9198
9199
9199
9199
9199
Ple I
Hinf I
Bsl I
Bsl I
Fnu4H I
Bbv I
Fok I
Alu I
TCATCATTGTTTCACACTTTTCCCCCAATTCTGGGTAAGCAGCAGTTATGTGGTCAGTCATCCATAAAGCACAGCTCATTCTGTTCTCTACTACATTGAC 9
AGTAGTAACAAAGTGTGAAAAGGGGGTTAAGACCCATTCGTCGTCAATACACCAGTCAGTAGGTATTTCGTGTCGAGTAAGACAAGAGATGATGTAACTG
9224
9225
9239
9239
9259
9273
9298
9298
Nla II
Rsa I
Csp6 I
Nla IV
Nla IV
Sty I
Mnl I
Tfi I
Hinf I
Kpn I
Ban I
BsaJ I
Bsl I
Bfa I
Dde I
Mse I
Acc65 I
TCAAATACCAAGGAGCCAAATGAGGGCAGGGAACATCTAACTAGATTCCCTTTGCTCTCTCTAACTCTAAGATATTTAATACTTATGTAAATGGTACCAT 9
AGTTTATGGTTCCTCGGTTTACTCCCGTCCCTTGTAGATTGATCTAAGGGAAACGAGAGAGATTGAGATTCTATAAATTATGAATACATTTACCATGGTA
9308
9308
9312
9316
9322
9341
9344
9344
9367
9376
9393
9393
9393
9393
9394
9394
9398
Rsa I
Csp6 I
Tfi I
Hinf I
SfaN I
Mse I
Tfi I
Hinf I
GACAACTGTGTACTATATTATGTAATCCTTCTGATTCTATAAGCAACTGTTCAGATGCCTATTTTAATGGATTCTAATTAGAATCTTTTTTGTTATTTGT 9
CTGTTGACACATGATATAATACATTAGGAAGACTAAGATATTCGTTGACAAGTCTACGGATAAAATTACCTAAGATTAATCTTAGAAAAAACAATAAACA
9410
9410
9433
9433
9454
9464
9470
9470
9481
9481
Rsa I
Csp6 I
Mnl I
Mae II
Dde I
Hph I
BspM
CCTTGTCTGCCTTTTTACTTTGCACCTGTACTAAAAATATACAAGAAAGAGGGCAGAGCATACTACGTCTAAGCAATGTATTATCTGGCTCACCATTTGA 9
GGAACAGACGGAAAAATGAAACGTGGACATGATTTTTATATGTTCTTTCTCCCGTCTCGTATGATGCAGATTCGTTACATAATAGACCGAGTGGTAAACT
9528
9528
9549
9565
9569
9590
9600
ScrF I
EcoR II
Dsa V
BstN I
BstK I
BsaJ I
Dde I
Tth111 II
Xmn I
Fok I
Mae III
Mse I
CCTGCATTTCAAACAAATGAGAAAACATTCATCCCAGGTCACTTAGATTATTTAACATTTGTTTTAGTGCTATGAACATTGATTGTAGTCAGTTCATATA 9
GGACGTAAAGTTTGTTTACTCTTTTGTAAGTAGGGTCCAGTGAATCTAATAAATTGTAAACAAAATCACGATACTTGTAACTAACATCAGTCAAGTATAT
9610
9621
9630
9633
9634
9634
9634
9634
9634
9638
9642
9652
Dde I
Alu I
Hph I
Apo I
Mse I
Mae III
Bsr I
Tth111
ATCGTGGCAAAATAAATATAGATAGCTCTAAGTGTTCTCTATATTTATATCTTGAAAATGAGTTTTAACCACAGTCACCCCAACCCAGTAAATTCACAAG 9
TAGCACCGTTTTATTTATATCTATCGAGATTCACAAGAGATATAAATATAGAACTTTTACTCAAAATTGGTGTCAGTGGGGTTGGGTCATTTAAGTGTTC
9724
9728
9765
9774
9775
9785
9790
9797

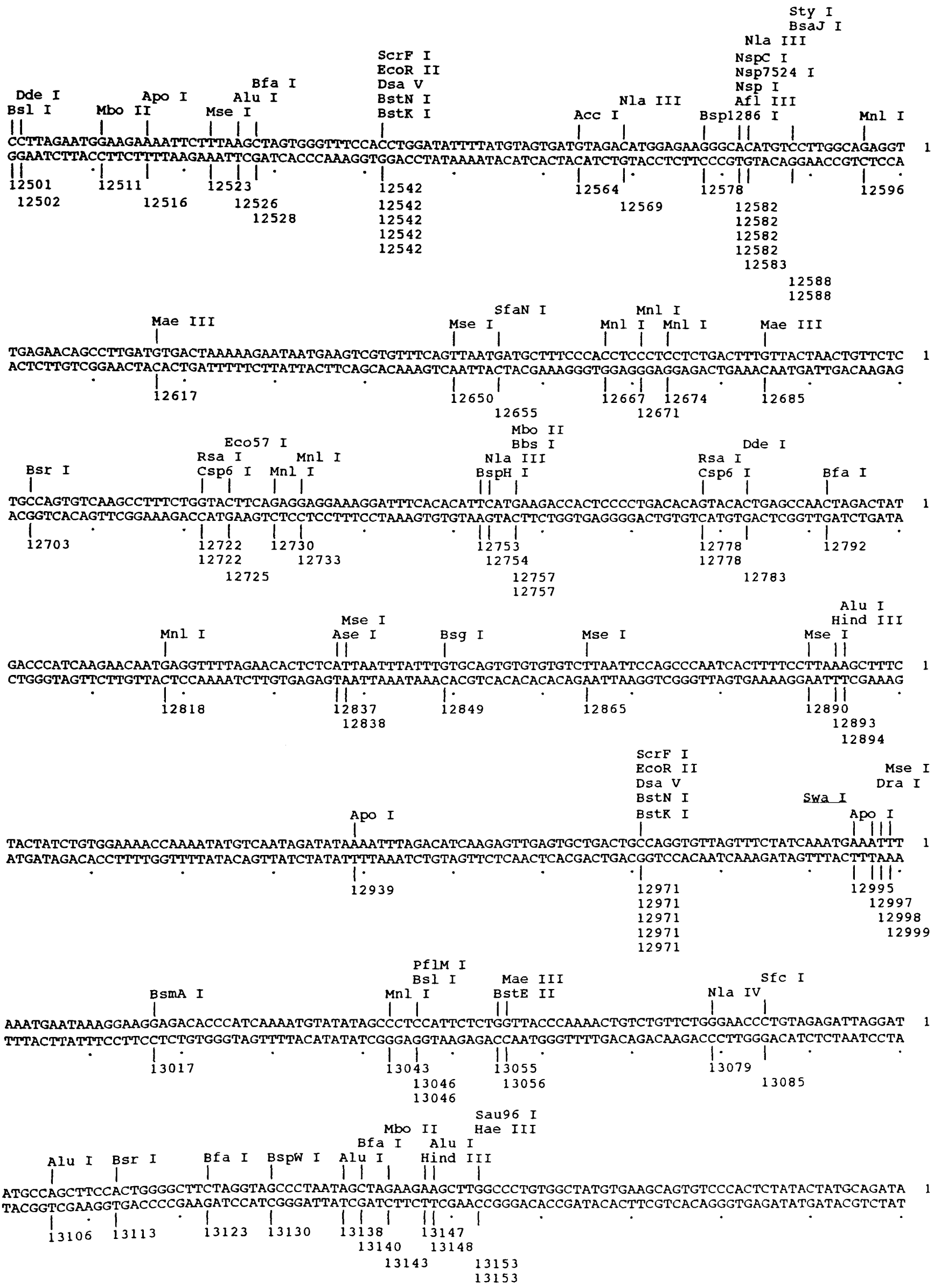

Xcm I
ScrF I
EcoR II
Dsa V
BstN I
BstK I
BspW I
HgiA I
Bsp1286 I
Nla III
Sty I
Nco I
Dsa I
Dde I
Bsu36 I
BsaJ I
Taq I
Rsa I
Csp6 I
Nla III
Tth111 II
Tth111 II
CAAACATGGTAATGTTTGTTTTGGATTGTTTTTCTTGCCTTAGGAAACCATGGACATCGAAGCGTACTTTGAAAGAATTGGTTATCAGAGCACCAGGAGC 9
GTTTGTACCATTACAAACAAAACCTAACAAAAAGAACGGAATCCTTTGGTACCTGTAGCTTCGCATGAAACTTTCTTAACCAATAGTCTCGTGGTCCTCG
9801 9805 9813 9838 9839 9848 9848 9848 9848 9849 9857 9864 9864 9888 9888 9890 9893 9893 9893 9893 9893 9893
Nla III
Sty I
Nco I
Dsa I
BsaJ I
Xmn I
Bsl I
Bsl I
Tfi I
Bsr I
Mse I
Eco57 I
Apo I
Fok I
Hinf I
Dde I
AAACTGGACTTGAAAACATTAACCGAAATCCTTCAGCACCAGATACGGGCTATTCCCTTTGAGAATTTGAACATCCATTGTGGGGAATCCATGGAACTGA 1
TTTGACCTGAACTTTTGTAATTGGCTTTAGGAAGTCGTGGTCTATGCCCGATAAGGGAAACTCTTAAACTTGTAGGTAACACCCCTTAGGTACCTTGACT
9903 9919 9925 9931 9939 9963 9972 9975 9985 9985 9989 9989 9989 9989 9990 9997
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Sau3A I
Mbo I
Dpn II
Dpn I
Bsp1286 I
Ban II
Bsr I
Rsa I
Csp6 I
Mbo II
Fau I
Bpm I
BsmA I
Mse I
Bcl I
Mnl I
Aci I
Fok I
Bsr I
GTTTAGAAGCCATCTTTGATCAAATTGTGAGGAAGAAGCGGGGTGGATGGTGTCTCCAGGTTAATCATCTGCTGTACTGGGCTCTGACCAAACTGGGCTT 1
CAAATCTTCGGTAGAAACTAGTTTAACACTCCTTCTTCGCCCCACCTACCACAGAGGTCCAATTAGTAGACGACATGACCCGAGACTGGTTTGACCCGAA
10017 10018 10018 10018 10018 10029 10032 10038 10038 10045 10052 10054 10056 10056 10056 10056 10056 10061 10074 10074 10076 10079 10079 10092
Hph I
Tfi I
Hph I
Mae III
Bpm I
Rsa I
Hinf I
BstX I
Mnl I
Mse I
Csp6 I
Nla III
Bfa I
BstE II
TGAAACCACAATGCTGGGAGGATATGTCTTTAACACTCCAGCCAATAAGTACAGCAGTGGCATGATTCACCTTCTAGTTCAGGTCACCATCAGTGGAAAG 1
ACTTTGGTGTTACGACCCTCCTATACAGAAATTGTGAGGTCGGTTATTCATGTCGTCACCGTACTAAGTGGAAGATCAAGTCCAGTGGTAGTCACCTTTC
10106 10118 10130 10136 10149 10149 10161 10164 10164 10167 10174 10182 10183 10184

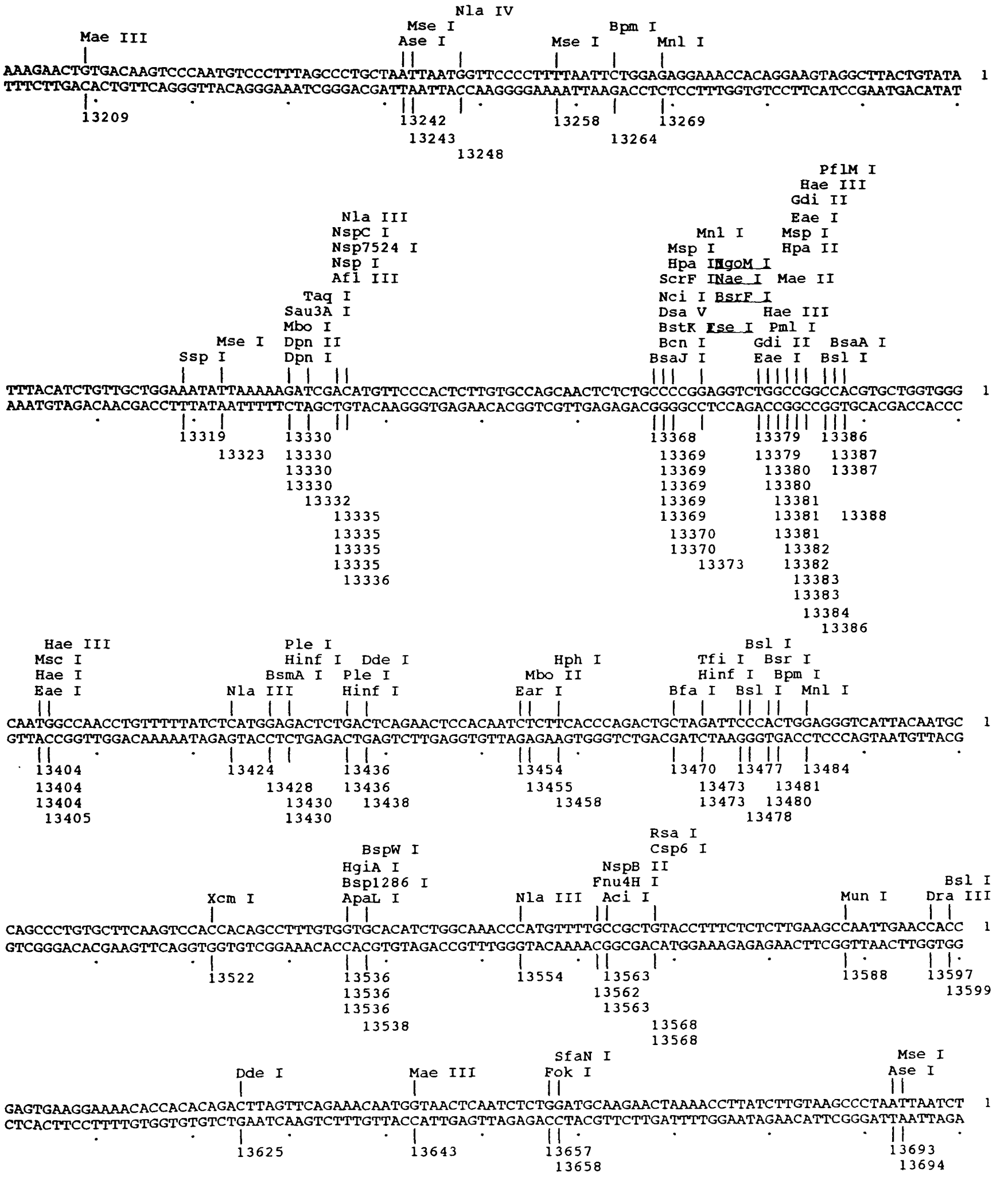

Mnl I
BspW I
Sau3A I
Mbo I
Dpn II
Dpn I
Nla IV
Ban I
Dde I
Mnl I
SfaN I
Mae II
Bsl I
Nla IV
Mse I
Alw I
Bsu36 I
Mbo II
GACTACATTGTTGATGCTGGGTTTGGACGTTCCTACCAGATGTGGGAGCCTCTGGAATTAACATCTGGAAAGGATCAGCCTCAGGTGCCTGCCATCTTCC 1
CTGATGTAACAACTACGACCCAAACCTGCAAGGATGGTCTACACCCTCGGAGACCTTAATTGTAGACCTTTCCTAGTCGGAGTCCACGGACGGTAGAAGG
10213 10227 10236 10245 10249 10258 10272 10273 10273 10273 10273 10278 10279 10279 10280 10284 10284 10295
Rsa I
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Sau96 I
Ava II
Mbo II
Ear I
Sau3A I
Mbo I
Dpn II
Dpn I
BstY I
Dde I
Mnl I
Nla IV
Csp6 I
Apo I
Mse I
Bgl II
GTTTGACAGAGGAGAATGGAACCTGGTACTTGGACCAAATCAGAAGAGAGCAGTATGTTCCAAACCAAGAATTTATTAACTCAGATCTCCTTGAAAAGAA 1
CAAACTGTCTCCTCTTACCTTGGACCATGAACCTGGTTTAGTCTTCTCTCGTCATACAAGGTTTGGTTCTTAAATAATTGAGTCTAGAGGAACTTTTCTT
10309 10318 10322 10322 10322 10322 10322 10326 10326 10332 10332 10343 10343 10369 10376 10380 10383 10383 10384 10384 10384 10384
Nla III
Hga I
AlwN I
Bsp1286 I
Ban II
Mbo II
Ple I
Hinf I
Eco57 I
Hph I
CAAATACCGAAAAATCTATTCCTTTACTCTTGAGCCCCGAACTATTGAAGATTTTGAGTCCATGAATACCTACCTTCAGACATCACCAGCGTCTGTGTTT 1
GTTTATGGCTTTTTAGATAAGGAAATGAGAACTCGGGGCTTGATAACTTCTAAAACTCAGGTACTTATGGATGGAAGTCTGTAGTGGTCGCAGACACAAA
10432 10432 10447 10456 10456 10461 10474 10483 10487 10489
Sau3
Mbo
Dpn
Dpn
Taq I
Sfc I
Pst I
Nla IV
Tth111 I
Tth111 II
Mnl I
Mbo II
Mae III
Mae II
Bfa I
Xmn I
ACTAGCAAATCATTTTGTTCCCTGCAGACCCCAGAAGGAGTTCATTGTTTGGTTGGCTCCACCCTCACTTATAGAAGATTTAGTTACAAGGACAACGTCG 1
TGATCGTTTAGTAAAACAAGGGACGTCTGGGGTCTTCCTCAAGTAACAAACCAACCGAGGTGGGAGTGAATATCTTCTAAATCAATGTTCCTGTTGCAGC
10502 10522 10522 10534 10546 10555 10563 10574 10583 10591 10595 10598 1060 1060 1060 1060
Ple I
Hinf I
Mbo II
Mnl I
Rsa I
Csp6 I
Nla III
Mse I
Dde I
Mbo II
Dde I
Bsl I
Hph I
ATCTTGTAGAGTTTAAGAGTCTGACTGAGGAAGAAATAGAAGATGTACTGAGAACTATATTTGGGGTTTCTTTAGAGAGAAAACTTGTGCCTAAACATGG 1
TAGAACATCTCAAATTCTCAGACTGACTCCTTCTTTATCTTCTACATGACTCTTGATATAAACCCCAAAGAAATCTCTCTTTTGAACACGGATTTGTACC
10613 10617 10617 10625 10627 10630 10639 10645 10645 10648 10690 10696 10699

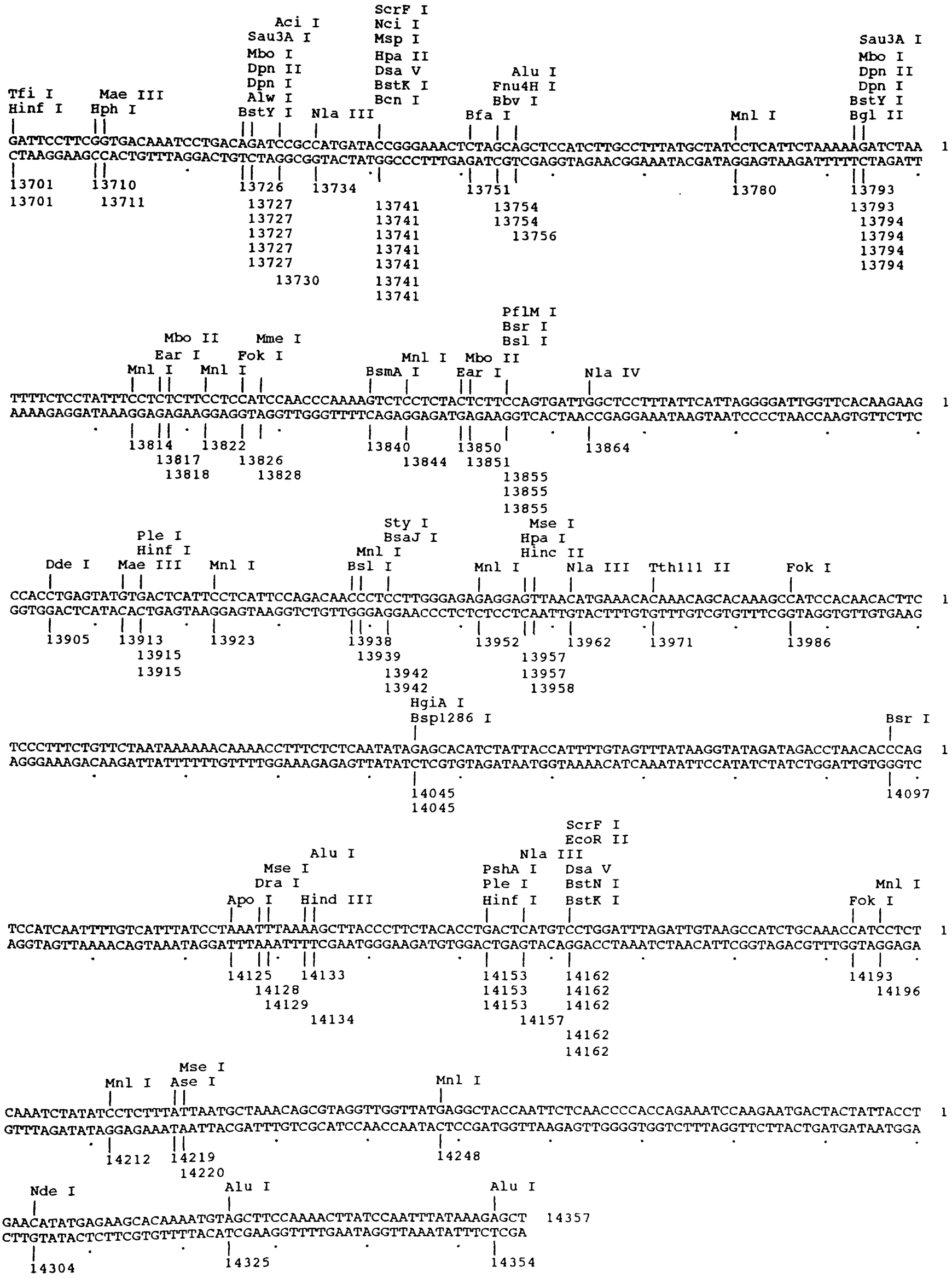

Taq I
Cla I
BspD I
Sau3A I
Mbo I
Dpn II
Dpn I
Rsa I
Csp6 I
Nla III
BstX I
Apo I
Mae III
Hph I
Bcl I
Bsr I
Nla III
Sau3A I
Mbo I
Dpn II
Dpn I
TGATCGATTTTTTACCATTTAGAATATGTAAAGTTTTGGTGTCCTTCCATGTACTTGGAATTTTTATGATAAGATATTCAAACTGGTGACATGATCATAC 1
ACTAGCTAAAAAATGGTAAATCTTATACATTTCAAAACCACAGGAAGGTACATGAACCTTAAAAATACTATTCTATAAGTTTGACCACTGTACTAGTATG
10702
10702
10702
10702
10703
10703
10704
10747
10748
10751
10751
10758
10782
10785
10786
10790
10792
10793
10793
10793
10793
Bsr I
Nla III
Bfa I
Bsl I
Apo I
Alu I
Apo I
Mse I
TACTGGTGGTCATGTTGATGTGTGCTAGGAAATACCCAGCATTGGCTTACAGCTAAATTTATATAGAATTTGTGCTAATTGTTAAGTAAAAATAATAGTG 1
ATGACCACCAGTACAACTACACACGATCCTTTATGGGTCGTAACCGAATGTCGATTTAAATATATCTTAAACACGATTAACAATTCATTTTTATTATCAC
10802
10811
10825
10835
10851
10855
10866
10882
Apo I
Mse I
Apo I
Rsa I
Csp6 I
Bsp1407 I
Tfi I
Hinf I
Mbo II
TGAAAAGTATTCCCATTAAATTCAACCAGCATTATCCATATCAAATTTTTATGACTTATTACCTGTACAATTTAGGATTTGGATTCAACAAAGAAGAATG 1
ACTTTTCATAAGGGTAATTTAAGTTGGTCGTAATAGGTATAGTTTAAAAATACTGAATAATGGACATGTTAAATCCTAAACCTAAGTTGTTTCTTCTTAC
10916
10918
10943
10964
10965
10965
10982
10982
10993
Sau96 I
Hae III
Sau96 I
Nla IV
Bsp1286 I
Bsp120 I
Ban II
Apa I
Mae III
EcoO109 I
Bsr I
Alu I
AGAAGTGTATGTGATTAGGGCCCAAGAAAACCAGTCACACAGCTAATTATTTCTTTCTGTGTTGTTTTTATATCTACTTTTTTTATATACTATTTTTTCT 1
TCTTCACATACACTAATCCCGGGTTCTTTTGGTCAGTGTGTCGATTAATAAAGAAAGACACAACAAAAATATAGATGAAAAAAATATATGATAAAAAAGA
11017
11018
11018
11018
11018
11018
11018
11019
11019
11031
11034
11041
Mun I
Tth111 II
Alu I
Mse I
Dra I
Mse I
Ase I
TACTGTGGGAACAAACACCAATTGAGTAAGTTGAGCTTGATTTTAAAAATGTAAGTCTAAATTAATAAAAATGAAATATATATCTAAAACTCATTTTCAA 1
ATGACACCCTTGTTTGTGGTTAACTCATTCAACTCGAACTAAAATTTTTACATTCAGATTTAATTATTTTTACTTTATATATAGATTTTGAGTAAAAGTT
11112
11119
11134
11142
11143
11161
11162
Hinc II
Nla III
NspC I
Nsp7524 I
Nsp I
Afl III
Mse I
Sfc I
Alu I
SfaN I
Fok I
Ple I
Hinf I
AGTGGTATATTGAAACATGTTGACTAATATGTGAATTAAAGGTTTTGTCTGTAGCTGTTAGGATGCAATGGGGTTGGTGTTTTCTGCAAGGGAGTCTTAT 1
TCACCATATAACTTTGTACAACTGATTATACACTTAATTTCCAAAACAGACATCGACAATCCTACGTTACCCCAACCACAAAAGACGTTCCCTCAGAATA
11215
11215
11215
11215
11215
11216
11219
11236
11249
11253
11261
11262
11292
11292


Sau3A I
Mbo I
Dpn II
Dpn I
Nla III
Bcl I
Bfa I
Mbo II
Bbs I
Bsm I
Bfa I
Sty I
BsaJ I
Avr II
CATTATTTTACTTCTCTATCATGCTGATTGATCATCACTGCTGACTTTCCTAGATTTATAACCACATAGAAGACACATTCCTAGGCATTCCTGTGAGTAT 1
GTAATAAAATGAAGAGATAGTACGACTAACTAGTAGTGACGACTGAAAGGATCTAAATATTGGTGTATCTTCTGTGTAAGGATCCGTAAGGACACTCATA
11320
11329
11330
11330
11330
11330
11350
11369
11369
11380
11380
11380
11381
11385
ScrF I
EcoR II
Dsa V
BstN I
BstK I
BsaJ I
Bpm I
Bsr I
Sau96 I
Hae III
Mbo II
Nla III
Hph I
Mae III
Bsl I
Mnl I
Sau96 I
Hae III
EcoO109 I
TTTTTTGAGAAGTTCAACTGGAGTTGAAAGGCCCAACTAATGAAGAAATGTGTGTGTCACCATGCCTCCACATTTAGGCCCTGGACTGAATAAAAGGGAA 1
AAAAAACTCTTCAAGTTGACCTCAACTTTCCGGGTTGATTACTTCTTTACACACACAGTGGTACGGAGGTGTAAATCCGGGACCTGACTTATTTTCCCTT
11417
11418
11430
11430
11442
11456
11457
11461
11465
11468
11476
11477
11477
11479
11480
11480
11480
11480
11480
Dde I
Bpu1102 I
Alu I
Ssp I
Mse I
Mbo II
SfaN I
Xcm I
Mae III
Hph I
BsmA I
Bsa I
Mnl I
Sfc I
EcoR V
Mse
Bsp1286 I
Dde I
Ban II
Afl I
AACAGGGAAAAGCTGAGCAATATTTATTTAATTCTTCTGTTATGACACAAGTGTGACCAGATGCTTTATGGTCTCACCTCTACAGATATCACTGAGCCCT 1
TTGTCCCTTTTCGACTCGTTATAAATAAATTAAGAAGACAATACTGTGTTCACACTGGTCTACGAAATACCAGAGTGGAGATGTCTATAGTGACTCGGGA
11511
11512
11513
11519
11528
11533
11553
11557
11560
11570
11571
11574
11577
11580
11585
11592
11594
11594
11599
1160
Nla III
NspC I
Nsp7524 I
Nsp I
Ppu10 I
Nsi I
Mse I
Mae III
TAAGCAATGCGTTAAAGTATATATAGATAATACTCATTTTATCTGTAACAAATCAGCCACAGATTTAGGTATGTGTGTGTGTATATGCATGTGAATGTGT 1
ATTCGTTACGCAATTTCATATATATCTATTATGAGTAAAATAGACATTGTTTAGTCGGTGTCTAAATCCATACACACACACATATACGTACACTTACACA
11612
11645
11685
11685
11687
11687
11687
11688
BsmA I
BsmA I
Mnl I
Mnl I
CTGTGTGTGTGAGAGAGAGAGAGAGTGGAGAGAGACAGAGAGACAGACAGACAGGCACAGACAGAGGGACAGAGGCAGAGAGAGAGAAAGAGAGAGAGAA 1
GACACACACACTCTCTCTCTCTCTCACCTCTCTCTGTCTCTCTGTCTGTCTGTCCGTGTCTGTCTCCCTGTCTCCGTCTCTCTCTCTTTCTCTCTCTCTT
11732
11740
11764
11772
Mnl I
Dde I
Fok I
Sfc I
Ssp I
Bsu36 I
Dde I
Nla III
Mse I
GTCCAAATCTACAGTAAATAATAAATATTTAGCCTGAGGTTGCCACCTGAGAACCATGTCATCCACTATAAACTATCTCTCAATGTTAAACAACTTGAGA 1
CAGGTTTAGATGTCATTTATTATTTATAAATCGGACTCCAACGGTGGACTCTTGGTACAGTAGGTGATATTTGATAGAGAGTTACAATTTGTTGAACTCT
11809
11824
11833
11834
11836
11847
11855
11860
11886


PflM I
Sty I
Bsl I
Mnl I
Taq
Mbo II
BsaJ I
Mse I
Mnl I
Mnl I
Dde I
TTGATTGTGAAACTAAAACTGTTCTTCAACCCCATTAGAAATCCAAGGTATGACTTTAATCCCAGCACTTGGGAGGCAGAGGCAGGAGGATTTCTGAGTT 1
AACTAACACTTTGATTTTGACAAGAAGTTGGGGTAATCTTTAGGTTCCATACTGAAATTAGGGTCGTGAACCCTCCGTCTCCGTCCTCCTAAAGACTCAA
11923
11943
11956
11973
11986
11994
11943
11962
11979
1200
11962
Sfc I
ScrF I
EcoR II
Dsa V
ScrF I
EcoR II
ScrF I
EcoR II
Dsa V
Hae III
Acc I
Hae I
BstN I
Mnl I
BstK I
Taq I
BsmA I
Tth111 II
CGAGGCCAGCCTGGTCTACAGAGTGAGTTCCAGGACAGCCAGGGCTACACCGAGAAACCCTGTCTCGAAAAAACAAACAAAAAAAAAAAAAAAGAAAGAA 1
GCTCCGGTCGGACCAGATGTCTCACTCAAGGTCCTGTCGGTCCCGATGTGGCTCTTTGGGACAGAGCTTTTTTGTTTGTTTTTTTTTTTTTTCTTTCTT
12002
12003
12004
12010
12010
12014
12010
12010
12010
12030
12030
12030
12030
12039
12039
12039
12039
12039
12062
12065
12074
12016
Mse I
Taq I
Mse I
Mse I
Ssp I
Bsr I
Mse I
Mae III
Mnl I
Xmn
ATCCAAGTATGACTTAAATGAAATATTAAATATCTTTTAGATAAATGACTGGGTAATGTCATTAAGTCGATAATTGTGACATTAGGGGAGGTTTTAATGG 1
TAGGTTCATACTGAATTTACTTTATAATTTATAGAAAATCTATTTACTGACCCATTACAGTAATTCAGCTATTAACACTGTAATCCCCTCCAAAATTACC
12114
12122
12126
12148
12162
12167
Alu I
Bfa I
12176
12188
12194
1220
Mbo II
Mae III
Mnl I
AAGTTCTTCCATCTTGCTTCATTGCACTCTACAAATGTCCCTTCTGTGTTACATTATGTCCCTCTAGCTTGAAACCCTGTTCCACTGTCATCTACTGACT 1
TTCAAGAAGGTAGAACGAAGTAACGTGAGATGTTTACAGGGAAGACACAATGTAATACAGGGAGATCGAACTTTGGGACAAGGTGACAGTAGATGACTGA
12205
12248
12261
12264
12266
Mbo II
Ear I
Ple I
Hinf I
Dde I
BsaB I
Mbo II
Tfi I
Hinf I
Bsp1286 I
Mnl I
Dde I
Hinc II
Mnl I
Mbo II
TTATCCTTGATGGAAGATTTATATCAGATTCTCAGACCTATGAGTATGAGGGCACTTCAAACTTAGAATGTTGACTCTTCTCTACCTCTTATCTTCTACC 1
AATAGGAACTACCTTCTAAATATAGTCTAAGAGTCTGGATACTCATACTCCCGTGAAGTTTGAATCTTACAACTGAGAAGAGATGGAGAATAGAAGATGG
12313
12316
12327
12327
12331
12348
12350
12362
12370
12373
12373
12375
12376
12385
12392
Sau3A I
Mbo I
Dpn II
Dpn I
Nla III
Bsl I
Bsl I
Apo I
Mun I
Mbo II
Ear I
Bcl I
Hph I
Fnu4H I
Bbv I
GTGCTTCCCCATGTCAGGGATATAATGTAAAATTCATACTTGTAAGAGCAGTTTTCTCTCTTCAATTGTGATCATCAGGAAAAGGTGATATGAGAAGCAG 1
CACGAAGGGGTACAGTCCCTATATTACATTTTAAGTATGAACATTCTCGTCAAAAGAGAGAAGTTAACACTAGTAGTCCTTTTCCACTATACTCTTCGTC
12408
12409
12410
12430
12458
12459
12463
12469
12470
12470
12470
12470
12484
12497
12497


Sty I
BsaJ I
Nla III
NspC I
Nsp7524 I
Nsp I
Afl III
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Bfa I
Dde I
Apo I
Alu I
Bsl I
Mbo II
Mse I
Nla III
Acc I
Bsp1286 I
Mnl I
CCTTAGAATGGAAGAAAATTCTTTAAGCTAGTGGGTTTCCACCTGGATATTTTATGTAGTGATGTAGACATGGAGAAGGGCACATGTCCTTGGCAGAGGT 1
GGAATCTTACCTTCTTTTAAGAAATTCGATCACCCAAAGGTGGACCTATAAAATACATCACTACATCTGTACCTCTTCCCGTGTACAGGAACCGTCTCCA
12501 12511 12523 12542 12564 12578 12596
12502 12516 12526 12542 12569 12582
12528 12542 12582
12542 12582
12542 12582
12583
12588
12588
SfaN I
Mnl I
Mae III
Mse I
Mnl I
Mnl I
Mae III
TGAGAACAGCCTTGATGTGACTAAAAAGAATAATGAAGTCGTGTTTCAGTTAATGATGCTTTCCCACCTCCCTCCTCTGACTTTGTTACTAACTGTTCTC 1
ACTCTTGTCGGAACTACACTGATTTTTCTTATTACTTCAGCACAAAGTCAATTACTACGAAAGGGTGGAGGGAGGAGACTGAAACAATGATTGACAAGAG
12617 12650 12667 12674 12685
12655 12671
Mbo II
Bbs I
Eco57 I
Rsa I
Mnl I
Nla III
Dde I
Rsa I
Bsr I
Csp6 I
Mnl I
BspH I
Rsa I
Csp6 I
Bfa I
TGCCAGTGTCAAGCCTTTCTGGTACTTCAGAGGAGGAAAGGATTTCACACATTCATGAAGACCACTCCCCTGACACAGTACACTGAGCCAACTAGACTAT 1
ACGGTCACAGTTCGGAAAGACCATGAAGTCTCCTCCTTTCCTAAAGTGTGTAAGTACTTCTGGTGAGGGGACTGTGTCATGTGACTCGGTTGATCTGATA
12703 12722 12730 12753 12778 12792
12722 12733 12754 12778
12725 12757 12783
12757
Alu I
Hind III
Mse I
Mnl I
Ase I
Bsg I
Mse I
Mse I
GACCCATCAAGAACAATGAGGTTTTAGAACACTCTCATTAATTTATTTGTGCAGTGTGTGTGTCTTAATTCCAGCCCAATCACTTTTCCTTAAAGCTTTC 1
CTGGGTAGTTCTTGTTACTCCAAAATCTTGTGAGAGTAATTAAATAAACACGTCACACACACAGAATTAAGGTCGGGTTAGTGAAAAGGAATTTCGAAAG
12818 12837 12849 12865 12890
12838 12893
12894
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Mse I
Dra I
Swa I
Apo I
Apo I
TACTATCTGTGGAAAACCAAAATATGTCAATAGATATAAAATTTAGACATCAAGAGTTGAGTGCTGACTGCCAGGTGTTAGTTTCTATCAAATGAAATTT 1
ATGATAGACACCTTTTGGTTTTATACAGTTATCTATATTTTAAATCTGTAGTTCTCAACTCACGACTGACGGTCCACAATCAAAGATAGTTTACTTTAAA
12939 12971 12995
12971 12997
12971 12998
12971 12999
12971
PflM I
Bsl I
Mae III
Sfc I
BsmA I
Mnl I
BstE II
Nla IV
AAATGAATAAAGGAAGGAGACACCCATCAAAATGTATATAGCCCTCCATTCTCTGGTTACCCAAAACTGTCTGTTCTGGGAACCCTGTAGAGATTAGGAT 1
TTTACTTATTTCCTTCCTCTGTGGGTAGTTTTACATATATCGGGAGGTAAGAGACCAATGGGTTTTGACAGACAAGACCCTTGGGACATCTCTAATCCTA
13017 13043 13055 13079
13046 13056 13085
13046
Sau96 I
Mbo II
Hae III
Bfa I
Alu I
Alu I
Bsr I
Bfa I
BspW I
Alu I
Hind III
ATGCCAGCTTCCACTGGGGCTTCTAGGTAGCCCTAATAGCTAGAAGAAGCTTGGCCCTGTGGCTATGTGAAGCAGTGTCCCACTCTATACTATGCAGATA 1
TACGGTCGAAGGTGACCCCGAAGATCCATCGGGATTATCGATCTTCTTCGAACCGGGACACCGATACACTTCGTCACAGGGTGAGATATGATACGTCTAT
13106 13113 13123 13130 13138 13147
13140 13148
13143 13153
13153


Nla IV
Mse I
Bpm I
Mae III
Ase I
Mse I
Mnl I
AAAGAACTGTGACAAGTCCCAATGTCCCTTTAGCCCTGCTAATTAATGGTTCCCCTTTTAATTCTGGAGAGGAAACCACAGGAAGTAGGCTTACTGTATA 1
TTTCTTGACACTGTTCAGGGTTACAGGGAAATCGGGACGATTAATTACCAAGGGGAAAATTAAGACCTCTCCTTTGGTGTCCTTCATCCGAATGACATAT
13209
13242
13258
13269
13243
13264
13248
PflM I
Hae III
Gdi II
Eae I
Nla III
NspC I
Nsp7524 I
Nsp I
Afl III
Mnl I
Msp I
Msp I
Hpa II
Hpa II
NgoM I
ScrF I
Nae I
Mae II
Taq I
Nci I
BsrF I
Sau3A I
Dsa V
Hae III
Mbo I
BstK I
Tse I
Pml I
Mse I
Dpn II
Bcn I
Gdi II
BsaA I
Ssp I
Dpn I
BsaJ I
Eae I
Bsl I
TTTACATCTGTTGCTGGAAATATTAAAAAGATCGACATGTTCCCACTCTTGTGCCAGCAACTCTCTGCCCCGGAGGTCTGGCCGGCCACGTGCTGGTGGG 1
AAATGTAGACAACGACCTTTATAATTTTTCTAGCTGTACAAGGGTGAGAACACGGTCGTTGAGAGACGGGGCCTCCAGACCGGCCGGTGCACGACCACCC
13319
13323
13330
13330
13330
13330
13332
13335
13335
13335
13335
13336
13368
13369
13369
13369
13369
13369
13370
13370
13373
13379
13379
13380
13380
13381
13381
13382
13382
13383
13383
13384
13386
13386
13387
13388
Hae III
Ple I
Bsl I
Msc I
Hae I
Eae I
Hinf I
BsmA I
Nla III
Dde I
Ple I
Hinf I
Hph I
Mbo II
Ear I
Tfi I
Hinf I
Bsr I
Bpm I
Bfa I
Bsl I
Mnl I
CAATGGCCAACCTGTTTTTATCTCATGGAGACTCTGACTCAGAACTCCACAATCTCTTCACCCAGACTGCTAGATTCCCACTGGAGGGTCATTACAATGC 1
GTTACCGGTTGGACAAAAATAGAGTACCTCTGAGACTGAGTCTTGAGGTGTTAGAGAAGTGGGTCTGACGATCTAAGGGTGACCTCCCAGTAATGTTACG
13404
13404
13404
13405
13424
13428
13430
13430
13436
13436
13438
13454
13455
13458
13470
13473
13473
13478
13477
13481
13480
13484
Rsa I
Csp6 I
BspW I
HgiA I
Bsp1286 I
ApaL I
NspB II
Fnu4H I
Aci I
Bsl I
Dra III
Xcm I
Nla III
Mun I
CAGCCCTGTGCTTCAAGTCCACCACAGCCTTTGTGGTGCACATCTGGCAAACCCATGTTTTGCCGCTGTACCTTTCTCTCTTGAAGCCAATTGAACCACC 1
GTCGGGACACGAAGTTCAGGTGGTGTCGGAAACACCACGTGTAGACCGTTTGGGTACAAAACGGCGACATGGAAAGAGAGAACTTCGGTTAACTTGGTGG
13522
13536
13536
13536
13538
13554
13562
13563
13563
13568
13568
13588
13597
13599
SfaN I
Mse I
Dde I
Mae III
Fok I
Ase I
GAGTGAAGGAAAACACCACACAGACTTAGTTCAGAAACAATGGTAACTCAATCTCTGGATGCAAGAACTAAAACCTTATCTTGTAAGCCCTAATTAATCT 1
CTCACTTCCTTTTGTGGTGTGTCTGAATCAAGTCTTTGTTACCATTGAGTTAGAGACCTACGTTCTTGATTTTGGAATAGAACATTCGGGATTAATTAGA
13625
13643
13657
13658
13693
13694


Tfi I
Hinf I
Mae III
Hph I
Aci I
Sau3A I
Mbo I
Dpn II
Dpn I
Alw I
BstY I
Nla III
ScrF I
Nci I
Msp I
Hpa II
Dsa V
BstK I
Bcn I
Bfa I
Alu I
Fnu4H I
Bbv I
Mnl I
Sau3A I
Mbo I
Dpn II
Dpn I
BstY I
Bgl II
GATTCCTTCGGTGACAAATCCTGACAGATCCGCCATGATACCGGGAAACTCTAGCAGCTCCATCTTGCCTTTATGCTATCCTCATTCTAAAAAGATCTAA 1
CTAAGGAAGCCACTGTTTAGGACTGTCTAGGCGGTACTATGGCCCTTTGAGATCGTCGAGGTAGAACGGAAATACGATAGGAGTAAGATTTTTCTAGATT
13701 13710 13726 13734 13751 13780 13793
13701 13711 13727 13741 13754 13793
13727 13741 13754 13794
13727 13741 13756 13794
13727 13741 13794
13727 13741 13794
13730 13741
13741
Mbo II
Ear I
Mme I
Fok I
Mnl I
Mnl I
PflM I
Bsr I
Bsl I
Mnl I
Mbo II
BsmA I
Ear I
Nla IV
TTTTCTCCTATTTCCTCTCTTCCTCCATCCAACCCAAAAGTCTCCTCTACTCTTCCAGTGATTGGCTCCTTTATTCATTAGGGGATTGGTTCACAAGAAG 1
AAAAGAGGATAAAGGAGAGAAGGAGGTAGGTTGGGTTTTCAGAGGAGATGAGAAGGTCACTAACCGAGGAAATAAGTAATCCCCTAACCAAGTGTTCTTC
13814 13822 13840 13850 13864
13817 13826 13844 13851
13818 13828 13855
13855
13855
Ple I
Hinf I
Sty I
BsaJ I
Mse I
Hpa I
Hinc II
Dde I
Mae III
Mnl I
Mnl I
Bsl I
Mnl I
Nla III
Tth111 II
Fok I
CCACCTGAGTATGTGACTCATTCCTCATTCCAGACAACCCTCCTTGGGAGAGAGGAGTTAACATGAAACACAAACAGCACAAAGCCATCCACAACACTTC 1
GGTGGACTCATACACTGAGTAAGGAGTAAGGTCTGTTGGGAGGAACCCTCTCTCCTCAATTGTACTTTGTGTTTGTCGTGTTTCGGTAGGTGTTGTGAAG
13905 13913 13923 13938 13952 13962 13971 13986
13915 13939 13957
13942 13958
13942
HgiA I
Bsp1286 I
Bsr I
TCCCTTTCTGTTCTAATAAAAAACAAAACCTTTCTCTCAATATAGAGCACATCTATTACCATTTTGTAGTTTATAAGGTATAGATAGACCTAACACCCAG 1
AGGGAAAGACAAGATTATTTTTTGTTTTGGAAAGAGAGTTATATCTCGTGTAGATAATGGTAAAACATCAAATATTCCATATCTATCTGGATTGTGGGTC
14045 14097
14045
ScrF I
EcoR II
Alu I
Nla III
Mse I
Dra I
PshA I
Dsa V
Ple I
BstN I
Mnl I
Apo I
Hind III
Hinf I
BstK I
Fok I
TCCATCAATTTTGTCATTTATCCTAAATTTAAAAGCTTACCCTTCTACACCTGACTCATGTCCTGGATTTAGATTGTAAGCCATCTGCAAACCATCCTCT 1
AGGTAGTTAAAACAGTAAATAGGATTTAAATTTTCGAATGGGAAGATGTGGACTGAGTACAGGACCTAAATCTAACATTCGGTAGACGTTTGGTAGGAGA
14125 14133 14153 14162 14193
14128 14153 14162 14196
14129 14153 14162
14134 14157
14162
14162
Mse I
Mnl I
Ase I
Mnl I
CAAATCTATATCCTCTTTATTAATGCTAAACAGCGTAGGTTGGTTATGAGGCTACCAATTCTCAACCCCACCAGAAATCCAAGAATGACTACTATTACCT 1
GTTTAGATATAGGAGAAATAATTACGATTTGTCGCATCCAACCAATACTCCGATGGTTAAGAGTTGGGGTGGTCTTTAGGTTCTTACTGATGATAATGGA
14212 14219 14248
14220
Nde I
Alu I
Alu I
GAACATATGAGAAGCACAAAATGTAGCTTCCAAAACTTATCCAATTTATAAAGAGCT 14357
CTTGTATACTCTTCGTGTTTTACATCGAAGGTTTTGAATAGGTTAAATATTTCTCGA
14304 14325 14354