 Human Molecular Genetics, 2001, Vol. 10, No. 4 339–352

ARTICLE

Sequence, structure and pathology of the fully annotated terminal 2 Mb of the short arm of human chromosome 16

Rachael J. Daniels[1], John F. Peden[1], Christine Lloyd[2,+], Sharon W. Horsley[1], Kevin Clark[1], Cristina Tufarelli[1], Lyndal Kearney[1], Veronica J. Buckle[1], Norman A. Doggett[3], Jonathan Flint[1] and Douglas R. Higgs[1,§]

[1]MRC Molecular Haematology Unit, Weatherall Institute for Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK, [2]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK and [3]Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received 22 November 2000 ; Revised and Accepted 16 December 2000

We have sequenced 1949 kb from the terminal Giemsa light band of human chromosome 16p, enabling us to fully annotate the region extending from the telomeric repeats to the previously published tuberous sclerosis disease 2 (*TSC2*) and polycystic kidney disease 1 (*PKD1*) genes. This region can be subdivided into two GC-rich, Alu-rich domains and one GC-rich, Alu-poor domain. The entire region is extremely gene rich, containing 100 confirmed genes and 20 predicted genes. Many of the genes encode widely expressed proteins orchestrating basic cellular processes (e.g. DNA recombination, repair, transcription, RNA processing, signal transduction, intracellular signalling and mRNA translation). Others, such as the α globin genes (*HBA1* and *HBA2*), *PDIP* and *BAIAP3*, are specialized tissue-restricted genes. Some of the genes have been previously implicated in the pathophysiology of important human genetic diseases (e.g. asthma, cataracts and the ATR-16 syndrome). Others are known disease genes for α thalassaemia, adult polycystic kidney disease and tuberous sclerosis. There is also linkage evidence for bipolar affective disorder, epilepsy and autism in this region. Sixty-three chromosomal deletions reported here and elsewhere allow us to interpret the results of removing progressively larger numbers of genes from this well defined human telomeric region.

INTRODUCTION

Although the Human Genome Project is nearing completion, the extent to which the current sequence is accurately assembled and annotated varies considerably from one region to another. In addition to identifying genes, fully annotated sequence will allow us to address global relationships between chromosome structure and function. In particular, we will be able to relate long-range, primary DNA sequence to the key processes of nuclear metabolism including transcription, replication, recombination, repair, methylation, chromatin assembly and nuclear positioning. Extensive preliminary data already suggest that correlations exist between chromosome banding, DNA sequence composition and these processes (1,2).

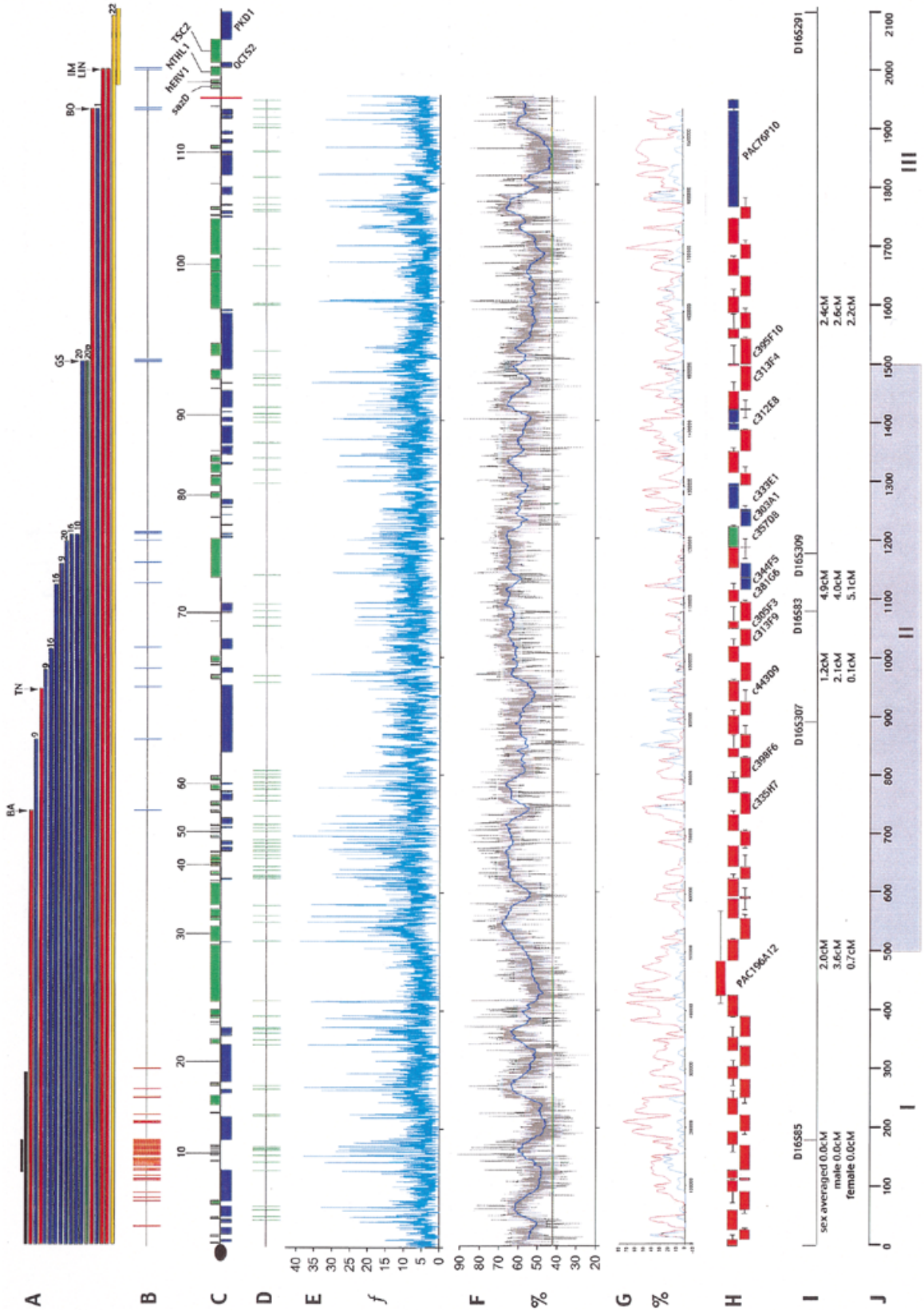
On a smaller scale, it is known that *cis*-acting sequences which control expression of specific genes may be located tens or hundreds of kilobases from the gene they regulate. It will therefore be important to establish whether regions of the genome that encode proteins are organized at a level above the unit of the gene and address the question of whether sequence analysis can help identify structurally discrete chromosomal domains that contribute to or reflect function.

The terminal 285 kb of 16p13.3, which includes the α globin genes, has been previously characterized using a variety of functional assays, allowing us to relate primary sequence to known biological function (3). Here, we have sequenced the terminal ~2 Mb of human chromosome 16p, enabling us to fully annotate a contig extending from the telomeric repeats to the previously published tuberous sclerosis disease 2 (*TSC2*) and polycystic kidney disease 1 (*PKD1*) genes (4,5). This segment of the Giemsa light band 16p13.3 is GC-rich and Alu dense containing many putative CpG islands and genes. In addition, 63 deletions from this 2 Mb region and their corresponding phenotypes, including the ATR-16 syndrome, are reported here and elsewhere (6–11) allowing us to interpret the effects of deleting progressively larger numbers of genes from this well defined chromosomal region.

Given its very high gene density and proximity to a human telomere, it is not surprising that, in addition to α thalassaemia (12), the ATR-16 syndrome (6), tuberous sclerosis (4) and the

[+]Christine Lloyd headed the production of the sequence at the Sanger Centre. A full list of past and present members of staff who contributed to generating this sequence is given in the Acknowledgements.

[§]To whom correspondence should be addressed. Tel: +44 1865 222393; Fax: +44 1865 222500; Email: drhiggs@molbiol.ox.ac.uk


A
B
C
D
E
f
F
%
G
%
H
I
J
BA
TN
GS
BO
IM
LIN
TSC2
NTRL1
hERV1
sacD
PKD1
OCTS2
D16585
D165307
D16583
D165309
D165291
sex averaged 0.0cM
male 0.0cM
female 0.0cM
2.0cM
3.6cM
0.7cM
1.2cM
2.1cM
0.1cM
4.9cM
4.0cM
5.1cM
2.4cM
2.6cM
2.2cM
PAC196A12
C325H7
C398F6
c44309
C305F3
C313F9
C344F5
C381G6
C35708
C303A1
C333E1
C312E8
C313F4
C395F10
PAC76F10
I
II
III

adult form of polycystic kidney disease (5), several previously characterized human genetic disease genes may also lie in this gene-rich region. These include asthma (13), cataracts with micro-ophthalmia (14), susceptibility to bipolar affective disorder (15), epilepsy (16) and various forms of autism (17–19). This highly annotated sequence extending 2 Mb from the 16p telomere should facilitate rapid identification of disease genes falling in this region. These data therefore provide an ideal opportunity to evaluate the extent to which DNA sequence analysis of the human genome will contribute to our understanding of chromosome structure, function and pathology.

RESULTS

Construction and sequencing of the 16p13.3 contig

A physical map of overlapping cosmids spanning the terminal 2 Mb of chromosome 16p was constructed (Fig. 1H) using three different restriction enzymes and multiple hybridizations with internal and end-clone fragments (see Materials and Methods). In most cases, several (3–14) cosmids were identified at each screening stage from a chromosome 16-specific cosmid library (20), providing significant depth (average, six cosmids), and hence confidence, in the resulting map. Clones chosen to represent the minimal tiling path were analysed using FISH to ensure that they mapped uniquely to the terminal region of 16p. In three instances, the genomic DNA was represented by a single cosmid [c398F6 (AL023882), c313F9/c305F3 (AL031707/AL031706) and c381G6 (AL031598) (Fig. 1H)]. Restriction maps of these regions were later confirmed using sequence data from the clones themselves but their structure has not yet been confirmed in genomic DNA. Two regions (around co-ordinates 426–485 and 1775–1949 kb) are not represented in the chromosome 16 cosmid library. In these cases, a human PAC library (RPCI) (21) was screened with flanking probes and recombinants spanning each gap [PAC196A12 (AL049542) and PAC76P10 (AL132867) (Fig. 1H)] were identified. The provenance of these clones was confirmed from sequence data (see further results on PAC76P10 below), restriction mapping and fluorescence *in situ* hybridization (FISH) analysis.

Sixty-five cosmids and PACs representing the minimal tiling path were sequenced. Of these, 59 are finished with an error rate of <1 in 10 000 bp. As described for other areas of the human genome (22,23), some small segments were consistently difficult to sequence but, in most cases, this could be overcome by applying methods for analysing sequence with a high GC content (see Materials and Methods). Six clones remain unfinished due to difficulties in obtaining sequence.

Clone c357D8 (green box in Fig. 1H) is contiguous, but includes regions of single-stranded sequence and poor quality data, invariably flanked by long tracts of Alu repeats. The missing strands have so far proved impossible to sequence using a variety of technologies.

Five clones (blue boxes in Fig. 1H) each contain a single gap (850, <100, 600 and 1250 bp and ~8 kb) not represented in the M13 shotgun libraries of the individual clones. These gaps are flanked by repeat sequences with tracts of very high GC content and it appears that polymerase consistently stalls at specific sequences. Despite considerable effort, completing these gaps, one of which contains an exon belonging to *CACNA1H* (gene no. 72), is beyond the scope of this current study, but efforts are continuing to complete these clones to a similar standard. It is interesting that 5 of the 16 ATR-16 centromeric chromosomal breakpoints fall within the 182 kb region spanned by four of these same clones, c344F5, c357D8, c303A1 and c333E1 (Fig. 1H), suggesting that there may be some link between chromosome breakage and segments that are difficult to clone, PCR and sequence. There is also evidence for a high degree of genetic recombination occurring in and around these clones (see Relationship between structure, gene expression and recombination).

Overview of the telomeric region of 16p13.3

The overall structure of this area is consistent with previous observations on GC-rich telomeric regions of the human genome (1,3) but provides further detail and resolution. The average GC content of the entire 1949 kb sequence is 57.5%, ranging from 47.2 to 65.3% when subdivided into 100 kb fragments (Fig. 2A), which is higher than the average for the human genome (~42%) (22). Superficial observation of the GC content suggested that this 2 Mb segment may be divided into three: Region I (1–500 kb) with an average GC content of 54.1%, Region II (501–1500 kb) with an average GC content of 60.9%

Figure 1. (Opposite) Summary of the key features of the 2 Mb region, which can also be accessed with further details at http://www.molbiol.ox.ac.uk/~haem/HMG/16p.html. (**A**) The bars represent the material deleted from a series of individuals. The smaller black bar denotes the range of the most common α globin deletions (12) and the larger black bar represents an individual with the largest known deletion with no phenotype other than α thalassaemia (11). Blue bars show the extent of 16p deleted material from individuals with ATR-16 who also have additional aneuplodies, and the chromosomal origin of the translocated material is shown at the end of the deletion. The deletion in patient GS, who has an unbalanced translocation involving an acrocentric p arm, is represented in green. Red bars show the extent of deleted material from ATR-16 individuals (BA, TN, BO, IM, LIN) currently presumed to be purely monosomic for this region of the genome. The large yellow bar represents the 16p deleted material from a patient with an unbalanced translocation who suffers from both tuberous sclerosis and polycystic kidney disease and the small yellow bar represents the extent of interstitial deletions causing tuberous sclerosis and polycystic kidney disease (79). (In most cases, the breakpoint is given as the midpoint of the cosmid in which the FISH signal changes; thus, the actual breakpoint could lie 30 kb in either direction.) (**B**) Breakpoints (both telomeric and centromeric where relevant) in this region from both α thalassaemia (red lines) and ATR-16 individuals (blue lines). (**C**) Genes identified throughout the region. The black oval denotes the telomeric repeat $(TTAGGG)_n$ region. Green boxes above the line show genes transcribed towards the centromere. Blue boxes below the line show genes transcribed towards the telomere. The red bar indicates the end of our analysis and annotation, but the physical map is contiguous with the PKD1 region, as shown (4,5,80,81). (**D**) Putative CpG islands. (**E**) Frequency of CpG dinucleotides per 200 bp. (**F**) Percentage GC content over a 200 bp window in grey. The blue line shows the percentage GC content of a moving window of 20 kb, stepping by 200 bp. The green line shows the average for the whole genome, currently estimated to be 42%. (**G**) Percentage of Alu (in red) and LINE (in blue) repeats as a proportion of total bases across a 1000 bp window, smoothed over 10 kb. (**H**) Minimal tiling path of clones. The filled red boxes represent the fully sequenced clones. Clones that contain sequencing gaps are shown in blue; the contiguous but unfinished clone is shown in green. (**I**) Polymorphic markers from the CEPH consortium chromosome 16 linkage map with the sex-averaged, male and female genetic distances between them shown in centiMorgans (47). (**J**) Scale bar in kilobases. Regions I, II and III are indicated.

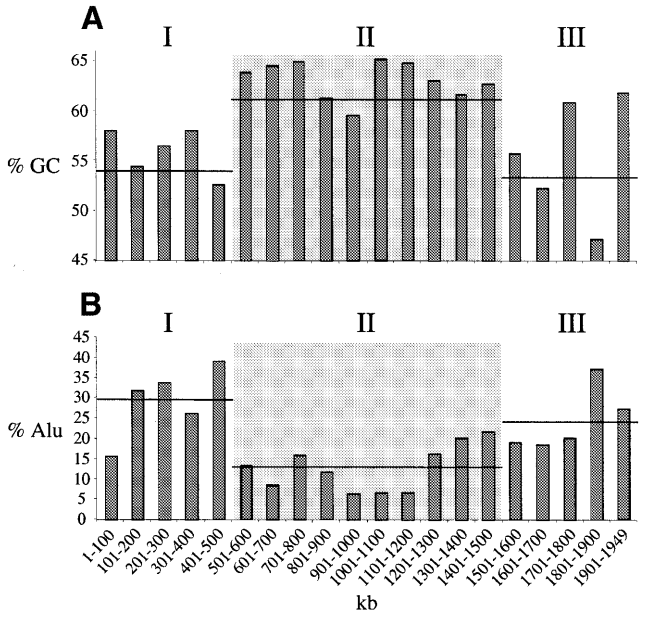


Figure 2. (**A** and **B**) Variation in GC and Alu content. These graphs show (A) the percentage of GC nucleotides and (B) the percentage of Alu repeats in 100 kb segments (non-overlapping). The shaded box divides the three regions, I, II and III. The horizontal lines represent the average for each region.

and Region III (1501–1949 kb) with an average GC content of 53.7% (Fig. 2A). Assessing the GC content using a 20 kb moving average plotted at the midpoint showed that the regular ~90 kb wavelength in GC content previously noted in the terminal 285 kb (3) extends throughout Region I (Fig. 1F). Genes that are transcribed towards the centromere appear to lie within the peaks of this wave pattern and those transcribed towards the telomere appear to lie within the troughs. The GC content remains relatively constant across Region II with a slight dip between co-ordinates 800 and 1000 kb (Fig. 2A), then, in Region III, further modulation in GC content occurs (Fig. 1F).

The repetitive elements are summarized in Table 1. As for other GC-rich isochores, the average Alu density is high (19.6%) whereas the density of LINE repeats is low (5.1%). The highest Alu density (29.5%) occurs in Region I (Figs 1G and 2B). In Region II, the Alu density is lower (12.7%) and this segment contains many tandem and simple repeats, a relative increase in the frequency of LINEs and fewer low complexity repeats than adjacent regions. Alu density increases again (24.2%) in Region III.

When the entire masked sequence was compared with itself, a 34 kb region was identified (between co-ordinates 1213 and 1247 kb) composed of one direct and two inverted repeats, the largest of which is 10 kb (data not shown). This is the same region that contains the clones which proved most difficult to clone and sequence; also located within this region are four members of the mast cell tryptase gene family: human transmembrane tryptase, tryptase beta III, tryptase beta I and tryptase beta II (13,24–26). Directly centromeric to these are three more tryptase-like genes that appear to be pseudogenes. Although the tryptase genes beta I–III have previously been localized to 16p13.3 (13), there is some disagreement in the order and number of tryptase genes presented here (based on genomic sequence data) with the previous report [based on restriction map data (13)].

Table 1. Summary of repetitive elements

Repeat	No. of elements	Length occupied (bp)	% of sequence
SINEs	1531	392 256	20.12
ALUs	1442	382 124	19.60
MIRs	89	10 132	0.52
LINEs	279	99 319	5.09
LINE1	208	87 122	4.47
LINE2	66	11 553	0.59
L3/CR1	5	644	0.03
LTR elements	126	48 848	2.51
MaLRs	70	20 490	1.05
ERVL	10	5025	0.26
ERV_classI	40	20 462	1.05
ERV_classII	2	2003	0.10
DNA elements	83	18 134	0.93
MER1_type	51	9809	0.50
MER2_type	19	6296	0.32
Unclassified	1	2302	0.12
Total interspersed repeats		560 859	28.77
Small RNA	4	304	0.02
Satellites	1	1084	0.06
Simple repeats	322	35 625	1.83
Low complexity	222	16 442	0.84

Synteny between human and mouse sequences

As previously shown, the telomeric 172 kb segment containing the α globin cluster and five of the genes located telomerically are syntenic to an interstitial fragment of mouse chromosome 11 (27). The remainder of the 2 Mb region between *LUC7L* (gene no. 16) up to and including at least the mouse orthologue of the *PKD1* gene, which lies beyond the region sequenced here, is syntenic to mouse chromosome 17 (28–30).

Identification of genes and estimates of gene density

The entire sequence was masked for repeats and initially annotated by sequence homology using the BLAST suite of programs (31,32) to search nucleotide [dbEST (33) and EMBL (34)] and protein [SWISS-PROT and TrEMBL (35)] sequence databases. The sequence was also analysed with the exon prediction programs GRAIL1.3, MZEF and XPOUND and the gene prediction programs GENSCAN, FGENES and FGENESH (36–41; V.V. Solovyev, unpublished data, see http://genomic.sanger.ac.uk). All sequences and analyses were processed using an automated system and stored in ACEDB (http://www.acedb.org). After extensive review and editing of these data, we classified (Table 2) and characterized (Table 3) 120 genes in the 1949 kb telomeric region. We found corroborative evidence (spliced ESTs, peptide homology or CpG

Table 2. Summary of genes types identified within this region

Category	Genes	Description
A	52	Known human gene
B	2	Like gene (i.e. 80% peptide sequence homology over whole length)
C	18	Similar to (i.e. significant peptide sequence homology over one or more exons)
D	11	Gene and exon predictions supported by spliced ESTs
D2	1	Supported only by spliced ESTs
D3	1	Conservation of genomic sequence between species
E	15	Pseudogenes
F	5	GENSCAN predictions supported by a CpG island
F2	15	GENSCAN predictions only
Total	120	

island) for 105; the remainder were supported by GENSCAN predictions alone.

Of the *ab initio* gene prediction programs used, GENSCAN was found to be the most accurate at predicting known genes (category A and B, Table 2). The accuracy increased when genes and exons were predicted by more than one program. In some regions, GENSCAN either over- or under-predicted. For example, around 844–960 kb, GENSCAN predicted six genes which, from further analysis, appear to be one gene and in another case, (630–669 kb), seven closely spaced genes (including two category A) were predicted to be a single transcript. Close inspection of this region revealed at least five clusters of ESTs and five CpG islands. We were able to amplify each putative gene from HeLa mRNA using internal primers but failed to amplify between these genes (data not shown), supporting the interpretation that they are separate genes.

Within this 2 Mb region, gene density is not uniform (Fig. 1C). On average, there is approximately one gene every 16 kb in this region, consistent with previous observations that telomeric, Giemsa light bands are gene dense (1,3,42). There seems to be no bias towards small or large genes. The smallest genomic coverage is *TRG4* (gene no. 36), which has a single exon spanning 75 bp; the largest is *C16orf26* (gene no. 62) spanning 116 kb. *CACNA1H* (gene no. 72) has the greatest number of exons (35) and there are several genes with only one exon, of which the largest is *IGFALS* (gene no. 106). *CRAMP1L* (gene no. 99) has the largest exon at 4029 bp. The smallest exon (9 bp) lies in *NUBP2* (gene no. 105).

The association of CpG islands and genes

We have shown previously that, in the terminal 285 kb, most of the genes are associated with CpG islands (3), with the island lying at the 5′ end of their associated genes, spanning the promoter. Putative CpG islands were initially identified from the frequency of CpG dinucleotides (15 CpGs in at least two adjacent 200 bp windows) but discarding sequences containing

Table 3. Genes identified within the terminal 1949 kb of 16p

Clone	Gene	HGNC name	Start	End	Associated nos	Protein acc. no.	Cat.	Trans.	CpG	CpG start	CpG end	Further details	OMIM	PubMed ID
HSPTEL	1	*DDX11P*	1685	4086	HS.166048	TR:Q92498	E	Pos.				Pseudogene	(601150)	9013641/ 9054936
HSPTEL	2	*(16pHQG2)*	4037	9529	HS.21346		E	Neg.				Pseudogene		9054936
HSLAW2	2.1	*(GENSCAN1)*	12932	16397			F2	Pos.				GENSCAN1		
HSLAW2	3	*IL9RP3*	17109	29489	HS.247991	SP:Q01113	E	Neg.				Interleukin 9 receptor IL9R pseudogene		8666384
HSNFG9	3.1	*POLR3K*	36979	43634	HS.110857	GP:AF051316_1	A	Neg.	A	43013	44230	RNA polIII CII subunit		9869639
HSNFG9	4	*C16orf33*	43881	47669	HS.15277		A	Pos.	A	43013	44230	Unknown function		1591777/ 9054936
HSNFG9	5	*C16orf8*	48058	66371	HS.57988		A	Neg.	B/C	50996	53162	Possible role in signal transduction		8838797
HSRA36		CpG island							B/C	62018	62873	CpG island		
HSRA36	6	*MPG*	68247	75845	HS.79396	SP:P29372	A	Pos.	D	67449	68631	*N*-methylpurine DNA glycosylase, DNA repair enzyme	156565	8475094
HSRA36	7	*C16orf35*	74275	128837	HS.19699	TR:Q12980	A	Neg.	E	128068	129511	Conserved gene telomeric to α-globin cluster, unknown function	600928	8575760
HSGG1	8	*HBZ*	142854	144504	HS.77253	SP:P02008	A	Pos.	F	143754	144866	ζ-globin, embryonic oxygen transport	142310	6963223
HSGG1	9	*HBZP*	152946	155254	HS.77253	SP:P02008	E	Pos.	G	154344	156720	ζ-globin pseudogene	142300	6963223
HSGG1	10	*HBAP2*	155997	156768	HS.146994	TR:E963826	E	Pos.		154344	156720	Globin pseudogene		3952001
HSGG1	11	*HBAP1*	158655	159453	HS.251577	SP:P01922	E	Pos.				Globin pseudogene		7407925
HSGG1	12	*HBA2*	162875	163709	HS.251577	TR:E974440	A	Pos.	H	162370	163447	Haemoglobin, α2	141850	345245/ 6446404
HSGG1	13	*HBA1*	166674	167521	HS.75792	TR:E963826	A	Pos.	I	166174	167254	Haemoglobin, α1	141800	345245/ 6446404
HSGG1	14	*(ROP)*	168553	168652	HY3		E	Pos.				Pseudogene		2448755
HSGG1	15	*HBQ1*	170335	171177	HS.247921	SP:P09105	A	Pos.	J	170162	171761	Haemoglobin, θ1	142240	9054936

Continued overleaf

Table 3. *Continued.*

Clone	Gene	HGNC name	Start	End	Associated nos	Protein acc. no.	Cat.	Trans.	CpG	CpG start	CpG end	Further details	OMIM	PubMed ID
HSCOS12	16	*LUC7L*	178971	219373	HS.16803	TR:BAA91500	A	Neg.	K	218840	219496	Possibly an RNA binding protein		9054936
HS310H5	16.1	*(PREDICTION1)*	224735	226818			F	Pos.	L			PREDICTION1		
HS310H5	17	*C16orf9*	238465	256658	HS.4768		D	Pos.	L			Unknown function		9054936
HS314G4	17.1	*RGS11*	258849	266449	HS.65756	SP:O94810	A	Neg.	M	265336	266715	Regulator of G protein signalling 11	603895	9789084
HS314G4	18	*ARHGDIG*	271219	274238	HS.121516	SP:Q99819	A	Pos.	N	270599	271532	Rho GDP dissociation inhibitor (GDI) γ	602844	9113980
HS314G4	19	*PDIP*	273743	277753	HS.66581	SP:Q13087	A	Pos.				Protein disulphide isomerase PDIP precursor		96152236
HS314G4	20	*AXIN1*	277979	343000	HS.184434	TR:O15169	A	Neg.	15	342509	343998	Axin, inhibits axin formation in embryo	603816	97373830
HS415C1	21	*C16orf43*	342969	353067			D3	Pos.	16	350520	352026	Conserved homology to chicken		
HS367G8	22	*MAAT1*	357931	361065	HS.110587	SP:Q13084	A	Neg.	17	360396	361271	Melanoma associated antigen recognised by T lymphocytes	604853	7751637
HS367G8	23	*TMEM6*	361326	372442	HS.25426	TR:BAB16376	A	Neg.	18/19	367951	368367	60% similar to NAG5 over second half		11006113
HS367G8		CpG island							18/19	371725	373345	CpG island		
HS367G8	24	*RPL23AP5*	377283	377809	HS.184776	SP:P29316	E	Pos.				60S ribosomal protein L23A-like	(602326)	1538749/ 9582194
HS367G8	25	*(GENSCAN2)*	379934	382348			F2	Pos.				GENSCAN2		
HS367G8	26	*NME4*	387758	391289	HS.9235	SP:O00746	A	Pos.		387445	388470	Nucleoside diphosphate kinase : NDKM	601818	9099850
HS359F1	27	*DECR2*	392498	403027	MM.35760	TR:Q9WV68	A	Pos.	20	391370	392667	2,4-dienoyl Coenzyme A reductase 2, peroxisomal		
HS359F1	28	*KIAA0665*	416152	513404	HS.119004	TR:O75154	A	Pos.	21	415304	417350	Unknown function		9734811
HS356B8	29	*C16orf10*	517377	517873	HS.252860		D	Neg.	22	517261	519046	Unknown function		
HS366D1	30	*SOLH*	518397	545167	HS.55836	TR:O75808	A	Pos.	22	517261	519046	Small optic lobes homologue, possible candidate gene for CATM	603267	9722942
HS366D1	31	*C16orf11*	555665	560025	HS.121190/ HS.201275	SP:Q13049	C	Pos.	23	550308	551713	Unknown: possibly a gene, query zinc finger-like		
HS407A10	32	*PIGQ*	560512	574667	HS.18079	TR:O14927	A	Pos.	24	560341	560881	*N*-acetylglucosamyl transferase component GPI1		9463366
HS398G5	33	*RASL8C*	580823	619798	HS.24655	TR:Q12829	B	Pos.	25	578976	581240	RAR (RAS-like GTPase)-like	(602672)	
HS398G5	34	*C16orf12*	623117	624592			D	Pos.	26	623170	624634	Pseudogene		
HS398G5	35	*C16orf13*	624966	626845	HS.239500		C	Neg.	27	626243	628016	Unknown function		
HS349E10	36	*TRG4*	627271	627341			A	Neg.		626243	628016	tRNA gene		
HS349E10	37	*C16orf14*	632448	638786	HS.41514		D	Pos.	28	632133	633235	Similar to non-muscle α actinin 1		
HS349E10		CpG island							29			CpG island		
HS349E10	38	*C16orf15*	646537	648838	FBan0003909		D	Pos.	30	645429	647041	Unknown function		
HS313D11	39	*C16orf16*	649088	651604	HS.222312		D	Pos.		651125	651946	Weakly similar to proline rich protein MP3		
HS313D11	40	*C16orf17*	651597	652526	HS.25890	O43942	C	Pos.		651125	651946	Similar to transducin		
HS313D11	41	*C16orf18*	652748	653496			D	Pos.				Unknown function		
HS313D11	42	*C16orf19*	653489	658358			D	Pos.				Unknown function		
HS313D11	43	*C16orf39*	658727	664702	HS.21497	TR:Q94263	C	Pos.	31	658464	659596	Similar to AK001902 (hs), similar to K08F11.5 (ce) which is weakly similar to the RAS gene family		7906398
HS313D11	44	*RHBDL*	666320	668795	HS.137572	TR:O75783	A	Pos.	32	666043	676347	Rhomboid-related protein	603264	9662444
HS313D11	45	*STUB1*	670611	673574	HS.25197	TR:Q9WUD1	A	Pos.	33	666043	676347	STIP1 homology and U box containing protein 1		10330192
HS313D11	46	*C16orf20*	673072	674855	MM.29000	TR:Q18258	C	Neg.	34			Similar to Ce F17C8.5		
HS313D11	47	*C16orf21*	675150	680902	HS.111520	TR:Q9XZ25	C	Neg.	35	680540	681293	Similar to FBan0007609		
HS380A1	48	*C16orf22*	684779	687665	HS.241432	TR:Q9VTL7	B	Neg.	36	684457	686475	Possible G protein receptor, cg14134		
HS380A1	49	*(PREDICTION2)*	696111	701766			F	Pos.	37	690931	691983	PREDICTION2		
HS380A1	50	*C16orf23*	705773	708008	HS.124915	TR:CAB58552	C	Pos.	38	705281	706526	Unknown		
HS444G9	51	*C16orf24*	711790	713122	HS.201282	TR:Q9VDC8	C	Pos.	39	711153	712690	Similar to FBan0003337		
HS444G9	52	*C16orf25*	713118	716943	HS.201282		D	Neg.	40	717169	719949	Unknown function		
HS444G9	53	*HAGHL*	717873	720243	AI761206	SP:Q16775	C	Pos.	40	717169	719949	Similar to HAGH	(138760)	8550579

Continued opposite

Table 3. *Continued.*

Clone	Gene	HGNC name	Start	End	Associated nos	Protein acc. no.	Cat.	Trans.	CpG	CpG start	CpG end	Further details	OMIM	PubMed ID
HS444G9	54	*NARFL*	720249	731519	HS.22158	TR:Q9UHQ1	C	Neg.	41	730744	731932	Weakly similar to ORF YNL240c (*S.cerevisiae*), similar to Let1 (*K.lactis*)		2227358
HS335H7	55	*(GENSCAN3)*	738928	745592			F2	Pos.				GENSCAN3		
HS335H7	56	*MSLN*	751354	759393	HS.155981	TR:Q14859	A	Pos.	42	755447	756341	Pre/pro-megakaryocyte potentiating factor precursor	(601051)	8552591/ 8150545/ 7665620
HS335H7	57	*C16orf37*	759960	773394			C	Neg.	43	765283	766347	Similar to pre/pro-megakaryocyte potentiating factor precursor	(601051)	
HS335H7	58	*C16orf40*	775502	778888	HS.101742	SP:P39219	C	Neg.	44	777909	779873	Ribosomal large subunit pseudouridine synthase C-like (ec)		9278503
HS321D2	59	*C16orf41*	779575	788602	HS.153850	TR:Q9VQU2	C	Pos.	45	777909	779873	Some homology with holliday junction DNA helicase RUVB-like, some homology with ch112 protein sc		3279394/ 2842314/ 9278503
HS321D2	60	*GNG13*	788569	791227	AB030207	GP:AB030207_1	A	Neg.	46	790760	791338	Gγ subunit		
HS321D2	61	*(PREDICTION3)*	795975	804393			F	Pos.	47			PREDICTION3		
HS398F6		CpG island							48			CpG island		
HS398F6		CpG island							49			CpG island		
HS360B4	62	*C16orf26*	844147	960496	HS.58362/ AL133278	TR:BAB14218	A	Neg.				Possible integral membrane, similar to P96418, MTCY08D5.31C		8610181
HS443D9		CpG island							50			CpG island		
HS394H11	63	*SOX8*	969801	976481	AF164104	TR:AAF35886	A	Pos.	51/52	969390	974834	SRY (sex determining region Y) box 8		10662550/ 10684944
HS394H11		CpG island							51/52			CpG island		
HS394H11	64	*(GENSCAN4)*	976542	978662			F2	Pos.				GENSCAN4		
HS394H11	65	*(GENSCAN5)*	980643	990026			F2	Neg.				GENSCAN5		
HS394H11	66	*(GENSCAN6)*	993145	995972			F2	Pos.				GENSCAN6		
HS422E10	67	*(GENSCAN7)*	999069	1009354			F2	Pos.				GENSCAN7		
HS422E10	68	*(GENSCAN8)*	1019949	1038792			F2	Neg.				GENSCAN8		
HS313F9		CpG island							53			CpG island		
HS349E11	69	*SSTR5*	1067560	1068807	HS.241375	SP:P35346	A	Pos.	54	1066950	1069120	Somatostatin receptor type 5	182455	7908405/ 8373420/ 8078491
HS349E11	70	*(ZSIG37P)*	1082223	1083468	AF192499	GP:AF192499_1	E	Neg.	55	1081042	1082813	Putative secreted protein pseudogene		
HS349E11	71	*(PREDICTION4)*	1085537	1099421			F	Neg.	56			PREDICTION4		
HS344F5	72	*CACNA1H*	1142602	1210751	HS.122359	TR:O95802	A	Pos.	57			Voltage-dependent t type calcium channel α 1H subunit		9670923/ 9930755
HS357D8	73	*TPSG1*	1210463	1214066	HS.268558	TR:Q9UBB2	A	Neg.				HS transmembrane tryptase		10521469
HS357D8	74	*TPSB2*	1217147	1219022	AF099143	GP:AF099143_1	A	Neg.		1220924	1221344	Mast cell tryptase βIII		2187193/ 9920877
HS303A1	75	*TPSB1*	1232518	1234393	HS.250700	TR:Q15661	A	Neg.				Tryptase I	191080	2187193/ 9920877
HS303A1	76	*TPSD1*	1244964	1247185	HS.241387	TR:O95824	A	Pos.		1242409	1242832	Putative mast cell MMCP7-like II tryptase		2203827/ 2187193/ 9920877
HS333E1	77	*TPSP1*	1249901	1251918	HS.170971	TR:Q9UQI1	E	Neg.				Most similar to βII		9920877
HS333E1	78	*TPSP2*	1262439	1265129		GP:AB038652_1	E	Neg.				Most similar to *Sus scrofa* MCT7		
HS333E1	79	*TPSP3*	1268876	1277617		sp:Q02844	E	Neg.				Most similar to mouse MCT7		1454796
HS333E1	80	*(GENSCAN9)*	1278694	1289817			F2	Pos.				GENSCAN9		
HS358B7	81	*UBE2I*	1300513	1315659	HS.84285	SP:P50550	A	Pos.	58	1297219	1297968	Ubiquitin conjugating enzyme E2 aka UBC9	601661	8668529/ 9067428/ 8565643/ 8798754
HS358B7	82	*RPS20P2*	1317301	1318464	HS.8102	SP:P17075	E	Pos.				40S ribosomal protein S20-like	(603682)	8479924/ 2357470
HS316G12	83	*BAIAP3*	1322297	1338080	HS.101516	TR:O94812	A	Pos.	59	1321577	1323375	BAI-associated protein 3, aka BAIAP3/KIAA0734	604009	9872452/ 9790924
HS316G12	84	*C16orf42*	1337208	1340445	HS.134846	GP:AE003708_24	C	Neg.	60	1337687	1338802	Similar to UND313 (sc)		8896276
HS316G12	85	*C16orf27*	1340590	1352267	HS.241575		D	Pos.	60	1337687	1338802	Similar to protein kinase C substrate		
HS399E4		CpG island							61			CpG island		

Continued overleaf

Table 3. *Continued.*

Clone	Gene	HGNC name	Start	End	Associated nos	Protein acc. no.	Cat.	Trans.	CpG	CpG start	CpG end	Further details	OMIM	PubMed ID
HS399E4	86	*C16orf28*	1353928	1368325	EMBL:AK027013		A	Neg.				Unknown function		
HS312E8	87	*TJP1P*	1359754	1360524	HS.74614		E	Pos.				Human tight junction (zonula occludens) protein ZO 1 pseudogene	(601009)	8395056
HS399E4	88	*UNKL*	1372573	1403332	HS.118261/ HS.161279	TR:Q24580	C	Neg.	62	1402403	1403660	Ce protein similar to Dm Cys3His finger protein, similarity to UNK (unkempt)		1339381
HS312E8	89	*(PREDICTION5)*	1409653	1418350			F	Neg.	63			PREDICTION5		
HS312E8		CpG island							64			CpG island		
HS312E8	90	*(GENSCAN10)*	1421507	1422043			F2	Pos.				GENSCAN10		
HS312E8	91	*C16orf29*	1426833	1427878	EMBL:AL036619		D	Neg.	65	1426023	1427043	Unknown function		
HS390E6	92	*CLCN7*	1434342	1464013	HS.80768	SP:P51798	A	Neg.	66	1463526	1464539	Putative chloride channel protein 7, aka CLC7	602727	8543009/ 9565675
HS305C8	93	*RPS3AP2*	1466836	1467610	HS.77039	SP:P49241	E	Pos.		1463526	1464539	40S ribosomal protein S3A-like pseudogene		63013004!!/ 1549582
HS305C8	94	*C16orf38*	1475053	1475313		SP:Q15818	C	Neg.	67	1476705	1477135	Similar to portion of neuronal pentraxin i precursor NPX1 or NP1	(602367)	8884281/ 7695898
HS305C8	95	*KIAA0683*	1482372	1499460	HS.226275	TR:O75168	A	Pos.	68	1479924	1481251	Unknown function		9734811
HS305C8	96	*KIAA0590*	1499433	1599050	HS.111862	TR:O60332	A	Neg.	69	1599060	1604099	Unknown function		9628581
HS380F5	97	*C16orf30*	1523243	1543878	HS.164158	TR:BAB14926	A	Pos.	70			Unknown function		
HS395F10	98	*(GENSCAN11)*	1599168	1601813			F2	Neg.				GENSCAN11		
HS395F10	99	*CRAMP1L*	1603928	1666913	HS.128494/ HS.117900	GP:Y13674_1	A	Pos.	71	1599060	1604099	Similar to EG:95B7.2, similar to Y13674 DMCRAMP_1		9310333
HS431H6	100	*C16orf34*	1667309	1691674	HS.172035	GP:BAB14434_1	A	Pos.	72	1666944	1667862	HN1-like		
HS329F2	101	*MAPK8IP3*	1695345	1757897	HS.88500	GP:AF178637_1	A	Pos.	73	1694829	1695937	Similar to sperm-specific protein, Mm JNK/SAPK-associated protein 1 (?MAPK81P3)		9480848/ 7906398
HS371H6	102	*NME3*	1759262	1762287	HS.81687	SP:Q13232	A	Neg.	74	1759387	1763033	Nucleoside diphosphate kinase 3	601817	7638209/ 9067290
HS371H6		CpG island										CpG island		
HS371H6	103	*(GENSCAN12)*	1762308	1765318			F2	Pos.				GENSCAN12		
HS371H6	104	*C16orf31*	1765787	1770784	HS.7247	TR:Q9V3U3	C	Neg.	75			Unknown function		
HSAC76P10	105	*NUBP2*	1772032	1778277	HS.256549	TR:Q9Y5Y2	A	Pos.	75			Putative nucleotide binding protein		10486206
HSAC76P10	106	*IGFALS*	1779530	1782842	HS.839	SP:P35858	A	Neg.	76	1779732	1781733	Insulin-like growth factor binding protein complex acid labile chain precursor, aka ALS	601489	2473065/ 1379671
HSAC76P10	107	*HAGH*	1798283	1812169	HS.155482	SP:q16775	A	Neg.	77	1815265	1816783	Hydroxyacylglutathione hydrolase	138760	10508780/ 8550579
HSAC76P10	108	*C16orf36*	1816387	1817145	HS.246859	gp:AB041600_1	A	Pos.	77	1815265	1816783	Similar to homoprotocatechuate catabolism bifunctional isomerase/decarboxylase		8384293/ 8223600
HSAC76P10	109	*(GENSCAN13)*	1830980	1867010			F2	Neg.				GENSCAN13		
HSAC76P10	110	*C16orf32*	1868091	1873801	HS.99512		D2	Neg.	78	1861564	1862093	Unknown function		
HSAC76P10	111	*(GENSCAN14)*	1876501	1882921			F2	Pos.				GENSCAN14		
HSAC76P10	112	*(GENSCAN15)*	1886578	1893384			F2	Neg.				GENSCAN15		
HSAC76P10	113	*HS3ST5*	1901134	1907723	HS.48384	TR:Q9Y662	C	Neg.	79	1907337	1908483	Heparin-sulphate-D-glucosaminyl-3-*O*-sulfotransferase 3B-like	(604058)	9988767
HSAC76P10		CpG island							80			CpG island		
HSAC76P10		CpG island							81			CpG island		
HSAC76P10	114	*SEPX1*	1927758	1932766	EMBL:AF166124	TR:AAF21429	A	Neg.	82			Selenoprotein X		10608886
HSAC76P10	115	*RPL3L*	1934253	1943572	HS.159191	SP:RL3L_HUMAN	A	Neg.	83			60S ribosomal protein L3-like	(604163)	8921388
HSAC76P10	116	*NDUFB10*	1948451	no end	HS.198274	SP:O96000	A	Pos.	84			NADH ubiquinone oxidoreductase PDSW subunit	603843	9878551

Clone, clone-containing telomeric end of the gene; Gene, sequential gene number; HGNC name, approved Human Genome Nomenclature Committee symbol (gene names in brackets have not been approved); Start/End, telomeric and centromeric extent of gene in base pairs; Associated nos, Unigene cluster ID or representative EST accession number; Protein acc. no., protein accession number for the protein with the highest similarity; Cat., category (see Table 2); Trans., orientation of transcription (Pos., towards centromere; Neg., towards telomere); CpG, CpG island number; CpG start/end, telomeric and centromeric extent of CpG island in base pairs; Further details, limited description; OMIM, OMIM reference number, brackets indicate a reference to the gene that is similar to the gene identified here; PubMed ID, most relevant reference. Genes now excluded from the region between the telomere and 1949 kb include *ZNF174*, *DNL1*, *CCNF*, *DCI*, *HMOX2*, *TNP2*, *testisin*, *PRM1*, *PRM2*, *ABC*, *HAUSP*, *E4F1* and *PGP*.

GC-rich repetitive DNA. Thus, we identified 84 CpG islands within the entire 1949 kb region equally distributed throughout Regions I–III. In contrast, computational methods, such as CPGREPORT [EMBOSS (43)], when searching with conventional criteria for CpG islands (CpG observed/expected frequency > 0.6, %GC > 50, over 200 bp), overestimated and identified 234 putative islands.

The presence of a CpG island is invariably thought to indicate the presence of a nearby gene (44). In the region studied here, 79 genes are associated with CpG islands and 41 are not. This is consistent with previous observations showing that most housekeeping genes and half of all tissue-restricted genes are associated with CpG islands (45); our observations [66% (79/120)], are somewhat higher than the genome average of 56% reported previously (44). Seven putative 'bi-directional' CpG islands are each associated with two genes, one in either transcriptional orientation. Five of these appear to contain two CpG peaks very close together, possibly incorporating two separate CpG islands. Three genes [*C16orf8* (gene no. 5), *TMEM6* (gene no. 23) and *SOX8* (gene no. 63)] each have two associated CpG islands.

Five CpG islands are associated with a predicted gene for which there is no other corroborative evidence (EST or protein homology matches, category F in Table 2). For nine CpG islands, we found no convincing evidence for any gene close by. These putative CpG islands may not be biologically active (unmethylated) or may be associated with genes which do not conform to the current prediction criteria or may be expressed in a tissue or developmental stage-specific manner so that they are not represented in any of the current sequence databases. Unless these 'orphan CpG islands' mark some other chromosomal element, it seems likely that additional genes near these CpG islands will be identified in the future, in which case the gene density in this region may be somewhat higher than currently estimated.

Relationship between structure, gene expression and recombination

Many of the 79 genes associated with CpG islands are widely expressed, although some (e.g. the α globin genes and *PDIP*) are expressed in a highly tissue-specific manner. We detected no pattern to the orientation of genes in this region: 57.5% (69/120) of the genes are transcribed towards the centromere and 42.5% (51/120) towards the telomere.

We have previously shown, using an *in situ* hybridization assay (46), that the terminal 50 kb of 16p replicates (or separates) later in S phase than the adjacent 250 kb which replicates early in the cell cycle (46). Provisional data suggest that most of the 2 Mb region also replicates early in the cell cycle although there is a remarkable dip in replication, or separation of chromatids, in the central portion of Region II which is currently under investigation (V. Buckle, unpublished data). Although this occurs in a relatively gene-poor region of this contig, at present there appears to be no clear correlation between replication timing and GC content.

Microsatellite markers allow us to relate the physical map to recombination events recorded in the CEPH consortium linkage map of chromosome 16 (Fig. 1I) (47). Data from the CEPH map indicate that this 2 Mb region has a higher recombination rate (male 12.3 cM, female 8.0 cM, sex-averaged 10.5 cM) than the genome average (1.1 cM/Mb). Recombination events occur most frequently between co-ordinates 900 and 1200 kb (male 6.1 cM and female 5.2 cM for a region of just 0.3 Mb).

Many chromosomal rearrangements have been reported from this segment of chromosome 16 including truncations (6,7,48), interstitial deletions (49) and translocations (10). All known breakpoints, both telomeric and centromeric, are plotted in Figure 1B. While initial inspection suggests that some breakpoints cluster around the α globin complex, this can be explained by the fact that many of these (red bars in Fig. 1B) are highly selected deletions that cause α thalassaemia. Five breakpoints, associated with ATR-16 syndrome, cluster close to the inverted and tandem repeats (at 1121–1304 kb) which encompass the tryptase gene cluster. Further observations are required to establish if this represents a preferred site of chromosomal breakage.

The effects of monosomy for 16p13.3

Deletions that remove the α globin genes (co-ordinates 162–168 kb) give rise to the well-defined haematological phenotype of α thalassaemia. Small deletions (within co-ordinates 129.5–178.2 kb) from the α globin cluster are very common (1–90% carrier frequency) in individuals from tropical and subtropical regions of the world. These deletions are confined to the α globin gene cluster and the surrounding genes remain intact (12; see Conclusions). We recently reported a series of 21 rare, interstitial deletions that remove the α globin genes and a variable number of genes flanking the α globin cluster (11). The largest of these extends for 268 kb and removes 15 functional genes (larger black bar in Fig. 1A). Two heterozygotes for this deletion have α thalassaemia but otherwise appear phenotypically normal (11), demonstrating that, apart from the α globin genes, none of the genes in the terminal 268 kb region of 16p is haploinsufficient.

Here, we have extended this analysis by investigating 16 individuals with still larger deletions (up to 2 Mb) from chromosome 16p (Fig. 1A). All were initially brought to our attention because they have α thalassaemia. Fourteen also have a variety of developmental abnormalities and all have some degree of learning difficulty and therefore have alpha thalassaemia with mental retardation syndrome [ATR-16, OMIM 141750 (6,50)]. Most of these patients have unbalanced translocations making it impossible to distinguish phenotypic features due to monosomy for 16p or trisomy for the other unbalanced chromosome (10,51). However, five patients appear to have pure monosomy of 16p based on cytogenetic studies, multiprobe FISH analysis (52) and, in some cases, analysis of the chromosomal breakpoint (Fig. 1A, red bars).

Two such patients (BA: deletion ~757 kb and TN: deletion ~951 kb) have no reported physical abnormalities but their cognitive abilities fall in the low-average range in contrast to their close relatives. Three patients with substantially larger deletions (BO: deletion ~1900 kb; IM: deletion ~2000 kb; LIN: deletion ~2000 kb) were previously shown to have facial dysmorphism with a variety of physical abnormalities and significant learning difficulties (6,8,9). The patient GS (deletion ~1595 kb), who is also dysmorphic with learning difficulties, has an unbalanced translocation involving satellite DNA from chromosome 21p. Since there may be no contribution to

the phenotype from this additional material, this patient's phenotype is predominantly due to monosomy for 16p. It is already known that removal of the gene *TSC2* causes the clearly defined phenotype of tuberous sclerosis (4,53,54) and thus *TSC2* at ~2050 kb effectively delimits the ATR-16 phenotype to this terminal 16p region.

The simplest conclusion is that the larger the region of monosomy, the more genes are deleted and the more severe the phenotype. Although it is clear that deletion of some genes may contribute more than others and, in some cases, direct disruption of a gene at a breakpoint may have a bearing on phenotype, there may not be a critical gene that explains all features of ATR-16 syndrome. However, the interpretation of these data is complex and is discussed further below.

Relationship to other known diseases

Previous reports have implicated the terminal region of 16p13.3 in several important human genetic diseases in addition to α thalassaemia, ATR-16 syndrome, tuberous sclerosis and the adult polycystic kidney disease.

The pathophysiology of asthma may involve members of the tryptase gene family (55–59). Here, we have shown that four mast cell tryptase genes, and three putative tryptase pseudogenes, lie in a 60 kb region, 1240 kb from the telomere of chromosome 16, as reported previously (13). Although these genes are not exclusively responsible for this polygenic disorder, they appear to play an important role in its pathophysiology (55–59).

One of the markers (*D16S521*), previously linked to a bipolar affective disorder (15), lies 34 kb from the telomere although other markers linked to this disease lie at least 2.5 Mb from the telomere (60–62). Autosomal recessive, idiopathic myoclonic epilepsy of infancy has been mapped to a broad region between *D16S3024* (at 1594 kb from the telomere) and *D16S423* (16). While this includes some of the 2 Mb contig, the highest LOD score (q = 0) corresponds to *D16S3027* located at least 2.5 Mb from the telomere.

Significant linkage exists between autism and markers in 16p13.3 (18,19) although the peak probability of linkage lies beyond this 2 Mb region. Autism with Tourette's syndrome has been reported in patients trisomic for 16p13.1-pter (17).

Cataracts with micro-ophthalmia (CATM) maps to 16p13.3 in a single family with a translocation involving chromosomes 2 and 16 (14). Both balanced and unbalanced translocations are associated with CATM indicating that a gene on one of these chromosomes is disrupted by the translocation. The breakpoint in chromosome 16 has been localized to band p13.3 by cytogenetic studies. Although the breakpoint in this family has not yet been refined, *SOLH* (gene no. 30) is a candidate because of its role in eye formation (63,64).

CONCLUSIONS

We have completed and fully annotated the sequence of a human telomere extending 2 Mb from the most terminal $(TTAGGG)_n$ repeats. This work highlights some deficiencies in the current public databases (such as: http://www.ncbi.nlm.nih.gov/genome/guide/HsChr16.shtml and http://www.ensembl.org) in which some of the released sequence generated here appears to be misassembled and only sparsely annotated.

The entire region is rich in CpG islands and genes, consistent with previous predictions that the greatest density of genes will occur in GC-rich, telomeric regions of the genome. It is interesting that this relatively small segment of the human genome (0.07%) contains 120 confirmed genes, predicted genes and pseudogenes; this is approximately half as many as identified from the whole of chromosome 21 (284 genes in 33.5 Mb, 1.12% of the genome) (23). It is also approximately three times as gene dense as the equivalent subtelomeric region of human chromosome 22q (22). This extreme variation in gene density emphasizes the difficulty in accurately predicting the number of genes in the human genome using isolated segments of the genome.

This telomeric sequence appears to be divisible into three segments (Regions I–III) on the basis of GC content and Alu density consistent with previous observations on isochores and chromosome 'flavors' (65,66). Throughout these segments there is marked variation in GC content [particularly in Region I (Figs 1F and 2A)] which was not seen when the same sequence was randomized, suggesting a biological basis for this phenomenon. At present the mechanism underlying this variation cannot be clearly related to transcription, replication or recombination. It remains possible that these variations reflect or contribute to some aspect of the higher order folding or organization of the chromosome.

At a higher level of resolution, we examined the distribution of genes along the chromosome. Again, no clear patterns emerge with respect to size, type or orientation but it is clear that tissue-restricted genes are intermingled with widely expressed genes. The question arises of how such genes are independently regulated and it has been frequently proposed that each gene may be sequestered in an independent structural and functional domain. Despite the popularity of such models (67,68), to date, no DNA sequence basis for subdivisions of the chromosome has emerged. Given the considerable overlap between genes and regulatory elements in the well-characterized terminal 285 kb region of 16p it seems unlikely that this region could be simply subdivided into independently-regulated chromosomal domains as described by Prioleau *et al.* (69). The identification and characterization of putative chromatin 'boundary elements' in other segments of the genome (70,71) suggest that if such chromosomal subdivisions exist in this telomeric region of the chromosome they may be difficult to predict from primary sequence.

The distribution of all known chromosomal breakpoints in this area is quite uneven and presumably reflects complex interactions between ascertainment bias, natural selection and the locations of preferred sites of recombination. One group (red bars in Fig. 1B), selected because they cause α thalassaemia, cluster around the α globin genes. It is interesting that none of the common, highly selected forms of α thalassaemia (12) removes the flanking genes even though rare individuals with larger deletions appear phenotypically normal. Presumably, although deletions extending into these highly conserved, widely expressed genes can be tolerated in rare heterozygotes (11), as a group, such individuals may be at some selective disadvantage. These deletions would almost certainly be lethal in homozygotes. The second group (blue bars in Fig. 1B) was identified because these individuals have α thalassaemia with

learning difficulties and, in most cases, additional developmental abnormalities. None of these breakpoints falls in the region 268–757 kb. Presumably although such telomeric deletions would cause α thalassaemia, one might predict that they do not produce any easily discernible phenotype and therefore do not commonly come to medical attention. The breakpoints clustered ~1200 kb from the 16p telomere occur near a repetitive region that contains a block of tryptase genes and pseudogenes, has a high rate of recombination and contains cosmids that have proven difficult to clone and sequence. It remains to be determined whether this is a preferred site of chromosome breakage.

The acquisition of fully annotated sequence has enabled us to begin to relate long-range DNA sequence to chromosome structure function and pathology. Clearly this sequence resource will now enable us to extend these studies and construct microarrays to specifically analyse this region in a systematic, unbiased manner.

MATERIALS AND METHODS

Physical mapping

The clones to complete the physical map were obtained from either the Los Alamos chromosome 16-specific cosmid library (20) or from the HGMP PAC library RPCI1 (21) by direct radioactive hybridization of filters using precisely known markers or probes derived from the clones themselves. Probes were labelled using the MegaPrime DNA Labelling Kit and [α-^{32}P]dCTP (Amersham Pharmacia Biotech). Clones were grown up using standard conditions [using ampicillin for SuperCos1 based cosmids (Stratagene) or kanamycin for the PACs]. DNA was prepared using standard techniques (Hybaid and Qiagen). Clones were digested separately with *Eco*RI, *Not*I and *Hin*dIII and electrophoresed on 0.8% agarose gels. The gels were stained with Vistra Green (Amersham Pharmacia Biotech) and scanned on a Storm PhosphorImager (Molecular Dynamics, Amersham Pharmacia Biotech) to identify exact fragment sizes. Southern blots of the gels were hybridized using end fragments generated from the clones themselves, using the Gene Images DNA Labelling kit (Amersham Pharmacia Biotech) to identify positive clones and combined with the restriction enzyme data to allow an accurate map of the growing contig to be established. Based on these data, a minimum tiling path of clones was chosen to represent the physical map (Fig. 1H). All of these clones were analysed using standard FISH techniques to confirm their location on human chromosome 16p13.3 (72). EMBL IDs with accession numbers in brackets for the minimum tiling path are as follows, in order, telomeric to centromeric: HSPTEL (Z84812), HSLAW2 (Z84723), HSNFG9 (Z69719), HSRA36 (Z69720), HSGG4 (Z84722), HSX94 (Z84813), HS24F8 (Z69666), HSGG1 (Z84721), HScos12 (Z69706), HSRJ14 (Z69890), HS310H5 (Z69705), HS314G4 (Z69667), HS419C1 (Z99754), HS333B10 (Z81450), HS415C1 (Z98272), HS367G8 (Z97634), HS359F1 (AL023881), HSC196A12 (AL049542), HS356B8 (Z98882), HS366D1 (Z97986), HS407A10 (Z98883), HS338H10 (Z98881), HS398G5 (Z84479), HS349E10 (AL022341), HS313D11 (Z92544), HS380A1 (Z97653), HS444G9 (Z98258), HS335H7 (AL031258), HS321D2 (AL031033), HS398F6 (AL023882), HS360A4 (Al031008), HS360B4 (AL031716), HS306A4 (AL008727), HS366D3 (Z93041), HS443D9 (Z92845), HS394H11 (Z99757), HS422E10 (AL024496), HS313F9 (AL031707), HS305F3 (AL031706), HS349E11 (AL031713), HS381G6 (AL031598), HS344F5 (AL031712), HS302G6 (AL031703), HS357D8 (AL031715), HS303A1 (AL031704), HS333E1 (AL031711), HS358B7 (AL031714), HS316G12 (AL031709), HS399E4 (AL031721), HS312E8 (AL032819), LA16–438F12 (AL137252), HS390E6 (AL031600), HS305C8 (AL031705), HS385E7 (AL031720), HS380F5 (AL031719), HS313F4 (Z97633), HS425C2 (AL133297), HS395F10 (Z97652), HS315G5 (AL031708), HS431H6 (AL031009), HS329F2 (AL031710), HS361A3 (AL031717), HS371H6 (AL031718) and HSAC76P10 (AL132867). The GenBank accession number for the complete 1949 kb is AE005175.

End-clone production

The terminal fragments for chosen clones were obtained in the following manner. The DNA was cleaved with *Sac*I (or *Xho*I or *Apa*I) for SuperCos1 cosmids or *Xho*I (or *Apa*I) for the CyPAC2n clones (these enzymes were chosen because they did not cut within the vector and could be heat-inactivated). The digests were heat-inactivated and ligase and ligase buffer (Promega) were added according to the manufacturers instructions. The ligations and transformations were performed using standard protocols. The DNA was extracted using Hybaid's Miniprep kit and a test quantity of DNA digested with the original enzyme described above to confirm that the new clone produced a single linear fragment. The correct subclones were then digested with the original enzyme and *Not*I to release the vector from the two terminal fragments. This digest was electrophoresed on a 0.8% low melting point agarose preparative gel and the terminal fragments excised. No further purification was required before labelling either radioactively or non-radioactively as described above, except for incubation at 65°C for 5 min to melt the agarose slice.

Sequencing

The cosmids and PACs were sequenced using a standard shotgun approach (73). In brief, the clone DNA was sonicated and 1.4–2.0 kb sized fragments were ligated into M13 or pUC vectors and transformed. Restriction digest data were used to estimate the size of each clone and around 200 sequence reads per 10 kb were generated using fluorescent dye-labelled terminators and primers on ABI 373A and ABI 377 sequencing machines (PE Applied Biosystems). The M13 subclones were sequenced using forward primers, while both forward and reverse primers were used to sequence the pUC subclones.

The sequence reads were base-called using phred (74) and assembled using phrap (http://www.phrap.org) into a GAP database (75) for editing. Standard finishing methods were employed to bring about gap closure and resolve sequence ambiguities. Various software tools were used to check the quality of the sequence and restriction digests were used to confirm the assembly of each clone.

Sequencing gaps that failed to be resolved by standard shotgun and finishing approaches were tackled by a number of techniques. (i) Using an oligo-screening strategy to identify further M13 clones that may extend the gap sequence or close

the gap altogether (76). (ii) Sequencing subclones from a short insert library generated either from a pUC subclone that spanned the gap or from a subcloned restriction fragment (77). (iii) PCR across the gap and direct sequencing of the PCR product using the original and internal primers. (iv) Direct sequencing of the cosmid DNA using primers that flank the gap [using 3 µg template DNA, 16 µl of standard ABI BigDye (PE Applied Biosystems) sequencing mix and 45 cycles]. (v) Application of the previous methods (iii and iv) but substituting ABI BigDye dGTP, increasing the PCR and sequencing denaturing temperature to 98°C and/or adding 1 M betaine to the PCR and sequencing reactions. (vi) Using standard manual sequencing techniques (Amersham).

Phenotypes of patients with 16p monosomy

The clinical features of these patients are briefly described here but have been or will be presented in detail elsewhere.

Patient BA. A preliminary report of this patient (78) described her as a phenotypically normal 14 year-old girl with a marked discrepancy between verbal and performance IQs, measured at 89 and 75, respectively. The chromosomal breakpoint in this patient lies in c335H7, ~757 kb from the 16p telomere.

Patients TN. Two brothers have delayed development for speech and walking; one also has a left iris coloboma. Their mother also has this deletion and is clearly intellectually different from her siblings (unpublished data). The TN breakpoint lies in c443D9, ~951 kb from the 16p telomere.

Patient GS. A boy aged 3 years. He had moderate delay in receptive language abilities and severe delay in expressive language abilities (unpublished data). The breakpoint lies between c313F4 and c395F10, ~1595 kb from the 16p telomere.

Patient BO. This patient was described in detail by Wilkie *et al.* (6) and references therein. At 15 years of age, he was moderately to severely retarded (IQ 53) with mild facial dysmorphism and minor congenital abnormalities. The breakpoint lies in PAC76P10, ~1900 kb from the 16p telomere.

Patient IM. This patient was described as having developmental delay and at 8 years of age had the mental ability of a 5-year-old (8). The breakpoint lies telomeric to the PKD1 region, ~2000 kb from the 16p telomere.

Patient LIN. This patient was described as having developmental delay with sign language developing at 2 years-of-age and walking by 2 years (9). She also has a variety of mild dysmorphic features. The breakpoint lies telomeric to the PKD1 region, ~2000 kb from the 16p telomere.

Breakpoint analysis

Chromosome 16p deletion patients were analysed using standard FISH techniques as described previously (72), with selected clones from the minimum tiling path to localize the chromosome 16 breakpoint to one or two clones. Multiprobe FISH analyses were performed as described previously (52).

ACKNOWLEDGEMENTS

Members of the Sanger Centre team are: Rachael Ainscough, Claire Bagguley, Karen Barlow, Caroline Baynes, Lisa Beard, Victoria Cobley, Gerard Coville, Sancha Donnelly, Andrew Ellington, Kerry Fleming, Debbie Frame, John Frankland, Audrey Fraser, Lisa Gilby, Rebekah Hall, Gretta Hall-Tamlyn, Sarah Holmes, Bijay Jassal, Matthew Jones, Jo Kershaw, Andrew Kimberley, Andrew King, Julia Lightning, Madeleine Moore, Chantal Percy, Adelaide Pettett, Ratna Shownkeen, Matthew Sims, Charlie Steward, Daniel Thomas, Karen Thomas, Justine Wallis, David Willey, Laurens Wilming and John Woodward. The authors would also like to thank: Richard Gibbons, Jane Rogers (Sanger Centre), M. Gardner, M. Descartes, Helen Brown, Ahmed Daghir, Hadley Wood, Christopher Ward, Peter Harris, the HUGO Nomenclature Committee (H. Wain, M. Lush, M. Wright, R. Lovering, E. Bruford and S. Povey), the Medical Research Council, the Wellcome Trust (J.F.) and the UK HGMP Resource Centre.

REFERENCES

1. Craig, J.M. and Bickmore, W.A. (1993) Chromosome bands—flavours to savour. *Bioessays*, **15**, 349–354.
2. Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**, 3–17.
3. Flint, J., Thomas, K., Micklem, G., Raynham, H., Clark, K., Doggett, N.A., King, A. and Higgs, D.R. (1997) The relationship between chromosome structure and function at a human telomeric region. *Nature Genet.*, **15**, 252–257.
4. European Chromosome 16 Tuberous Sclerosis Consortium (1993) Identification and characterization of the tuberous sclerosis gene on chromosome 16. *Cell*, **75**, 1305–1315.
5. European Polycystic Kidney Disease Consortium (1994) The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. *Cell*, **77**, 881–894.
6. Wilkie, A.O.M., Buckle, V.J., Harris, P.C., Lamb, J., Barton, N.J., Reeders, S.T., Lindenbaum, R.H., Nicholls, R.D., Barrow, M., Bethlenfalvay, N.C. *et al.* (1990) Clinical features and molecular analysis of the α thalassaemia/mental retardation syndromes. I. Cases due to deletions involving chromosome band 16p13.3. *Am. J. Hum. Genet.*, **46**, 1112–1126.
7. Lamb, J., Harris, P.C., Wilkie, A.O.M., Wood, W.G., Dauwerse, J.G. and Higgs, D.R. (1993) De novo truncation of chromosome 16p and healing with $(TTAGGG)_n$ in the α-thalassemia/mental retardation syndrome (ATR-16). *Am. J. Hum. Genet.*, **52**, 668–676.
8. Fei, Y.J., Liu, J.C., McKie, V.C. and Huisman, T.H. (1992) Hb H disease and mild mental retardation in a black girl with a Hb S heterozygosity. *Hemoglobin*, **16**, 431–434.
9. Lindor, N.M., Valdes, M.G., Wick, M., Thibodeau, S.N. and Jalal, S. (1997) De novo 16p deletion: ATR-16 syndrome. *Am. J. Med. Genet.*, **72**, 451–454.
10. Lamb, J., Wilkie, A.O.M., Harris, P.C., Buckle, V.J., Lindenbaum, R.H., Barton, N.J., Reeders, S.T., Weatherall, D.J. and Higgs, D.R. (1989) Detection of breakpoints in submicroscopic chromosomal translocation, illustrating an important mechanism for genetic disease. *Lancet*, **2**, 819–824.
11. Horsley, S.W., Daniels, R.J., Anguita, E., Raynham, H.A., Peden, J.F., Villegas, A., Vickers, M.A., Green, S., Chui, D.H.K., Ayyub, H., *et al.* (2001) Monosomy for the most telomeric, gene-rich region of human chromosome 16p causes minimal phenotypic effects. *Eur. J. Hum. Genet.*, in press.
12. Higgs, D.R., Vickers, M.A., Wilkie, A.O.M., Pretorius, I.-M., Jarman, A.P. and Weatherall, D.J. (1989) A review of the molecular genetics of the human α-globin gene cluster. *Blood*, **73**, 1081–1104.
13. Pallaoro, M., Fejzo, M.S., Shayesteh, L., Blount, J.L. and Caughey, G.H. (1999) Characterization of genes encoding known and novel human mast cell tryptases on chromosome 16p13.3. *J. Biol. Chem.*, **274**, 3355–3362.
14. Yokoyama, Y., Narahara, K., Tsuji, K., Ninomiya, S. and Seino, Y. (1992) Autosomal dominant congenital cataract and microphthalmia associated with a familial t(2;16) translocation. *Hum. Genet.*, **90**, 177–178.

15. Detera-Wadleigh, S.D., Barden, N., Craddock, N., Ewald, H., Foroud, T., Kelsoe, J. and McQuillin, A. (1999) Chromosomes 12 and 16 Workshop. *Am. J. Med. Genet.*, **88**, 255–259.
16. Zara, F., Gennaro, E., Stabile, M., Carbone, I., Malacarne, M., Majello, L., Santangelo, R., Antonio de Falco, F. and Bricarelli, F.D. (2000) Mapping of a locus for a familial autosomal recessive idiopathic myoclonic epilepsy of infancy to chromosome 16p13. *Am. J. Hum. Genet.*, **66**, 1552–1557.
17. Hebebrand, J., Martin, M., Körner, J., Roitzheim, B., de Braganca, K., Werner, W. and Remschmidt, H. (1994) Partial trisomy 16p in an adolescent with autistic disorder and Tourette's syndrome. *Am. J. Med. Genet.*, **54**, 268–270.
18. International Molecular Genetic Study of Autism Consortium (1998) A full genome screen for autism with evidence for linkage to a region on chromosome 7q. *Hum. Mol. Genet.*, **7**, 571–578.
19. Philippe, A., Martinez, M., Guilloud-Bataille, M., Gillberg, C., Råstam, M., Sponheim, E., Coleman, M., Zappella, M., Aschauer, H., van Malldergerme, L. *et al.* (1999) Genome-wide scan for autism susceptibility genes. *Hum. Mol. Genet.*, **8**, 805–812.
20. Stallings, R.L., Torney, D.C., Hildebrand, C.E., Longmire, J.L., Deaven, L.L., Jett, J.H., Doggett, N.A. and Moyzis, R.K. (1990) Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl Acad. Sci. USA*, **87**, 6218–6222.
21. Ioannou, P.A., Amemiya, C.T., Garnes, J., Kroisel, P.M., Shizuya, H., Chen, C., Batzer, M.A. and de Jong, P.J. (1994) A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genet.*, **6**, 84–89.
22. Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
23. The Chromosome 21 Mapping and Sequencing Consortium (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, 311–319.
24. Wong, G.W., Tang, Y., Feyfant, E., Sali, A., Li, L., Li, Y., Huang, C., Friend, D.S., Krilis, S.A. and Stevens, R.L. (1999) Identification of a new member of the tryptase family of mouse and human mast cell proteases which possesses a novel COOH-terminal hydrophobic extension. *J. Biol. Chem.*, **274**, 30784–30793.
25. Miller, J.S., Moxley, G. and Schwartz, L.B. (1990) Cloning and characterization of a second complementary DNA for human tryptase. *J. Clin. Invest.*, **86**, 864–870.
26. Vanderslice, P., Ballinger, S.M., Tam, E.K., Goldstein, S.M., Craik, C.S. and Caughey, G.H. (1990) Human mast cell tryptase: multiple cDNAs and genes reveal a multigene serine protease family. *Proc. Natl Acad. Sci. USA*, **87**, 3811–3815.
27. Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R.J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K.C., McMorrow, T. *et al.* (2001) Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α globin cluster. *Hum. Mol. Genet.*, **10**, 371–382.
28. Tufarelli, C., Frischauf, A.-M., Hardison, R., Flint, J. and Higgs, D.R. (2001) Characterisation of a widely expressed gene (*LUC7-LIKE*) defining the centromeric boundary of the human α globin domain. *Genomics*, **71**, in press.
29. Olsson, P.G., Sutherland, H.F., Nowicka, U., Korn, B., Poutska, A. and Frischauf, A.M. (1995) The mouse homologue of the tuberin gene (TSC2) maps to a conserved synteny group between mouse chromosome 17 and human 16p13.3. *Genomics*, **25**, 339–340.
30. Olsson, P.G., Lohning, C., Horsley, S., Kearney, L., Harris, P.C. and Frischauf, A. (1996) The mouse homologue of the polycystic kidney disease gene (Pkd1) is a single-copy gene. *Genomics*, **34**, 233–235.
31. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
32. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database searcch programs. *Nucleic Acids Res.*, **25**, 3389–3402.
33. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
34. Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G. and Tuli, M.A. (2000) The EMBO nucleotide sequence database. *Nucleic Acids Res.*, **28**, 19–23.
35. Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
36. Xu, Y., Mural, R., Shah, M. and Uberbacher, E. (1994) Recognizing exons in genomic sequence using GRAIL II. *Genet. Eng.*, **16**, 241–253.
37. Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. *Ismb*, **3**, 367–375.
38. Zhang, M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA*, **94**, 565–568.
39. Thomas, A. and Skolnick, M.H. (1994) A probabilistic model for detecting coding regions in DNA sequences. *IMA J. Math. Appl. Med. Biol.*, **11**, 149–160.
40. Uberbacher, E.C. and Mural, R.J. (1991) Locating protein coding regions in human DNA sequences using a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA*, **88**, 11261–11265.
41. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
42. Saccone, S., De Sario, A., Della Valle, G. and Bernardi, G. (1992) The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. *Proc. Natl Acad. Sci. USA*, **89**, 4913–4917.
43. Rice, P., Longden, O. and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
44. Cross, S.H. and Bird, A.P. (1995) CpG islands and genes. *Curr. Opin. Genet. Dev.*, **5**, 309–314.
45. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
46. Smith, Z.E. and Higgs, D.R. (1999) The pattern of replication at a human telomeric region (16p13.3): its relationship to chromosome structure and gene expression. *Hum. Mol. Genet.*, **8**, 1373–1386.
47. Kozman, H.M., Keith, T.P., Donis-Keller, H., White, R.L., Weissenbach, J., Dean, M., Vergnaud, G., Kidd, K., Gussella, J., Royle, N.J. *et al.* (1995) The CEPH Consortium linkage map of human chromosome 16. *Genomics*, **25**, 44–58.
48. Flint, J., Craddock, C.F., Villegas, A., Bentley, D.P., Williams, H.J., Galanello, R., Cao, A., Wood, W.G., Ayyub, H. and Higgs, D.R. (1994) Healing of broken human chromosomes by the addition of telomeric repeats. *Am. J. Hum. Genet.*, **55**, 505–512.
49. Hatton, C., Wilkie, A.O.M., Drysdale, H.C., Wood, W.G., Vickers, M.A., Sharpe, J., Ayyub, H., Pretorius, I.-M., Buckle, V.J. and Higgs, D.R. (1990) Alpha thalassemia caused by a large (62 kb) deletion upstream of the human α globin gene cluster. *Blood*, **76**, 221–227.
50. Weatherall, D.J., Higgs, D.R., Bunch, C., Old, J.M., Hunt, D.M., Pressley, L., Clegg, J.B., Bethlenfalvay, N.C., Sjolin, S., Koler, R.D. *et al.* (1981) Hemoglobin H disease and mental retardation. A new syndrome or a remarkable coincidence? *N. Engl. J. Med.*, **305**, 607–612.
51. Gibbons, R.J. and Higgs, D.R. (2001) The alpha thalassemia/mental retardation syndromes. In Steinberg, M.H., Forget, B.G., Higgs, D.R. and Nagel, R.L. (eds), *Disorders of Hemoglobin*. Cambridge University Press, Cambridge, UK.
52. Knight, S.J., Horsley, S.W., Regan, R., Lawrie, N.M., Maher, E.J., Cardy, D.L., Flint, J. and Kearney, L. (1997) Development and clinical application of an innovative fluorescence *in situ* hybridization technique which detects submicroscopic rearrangements involving telomeres. *Eur. J. Hum. Genet.*, **5**, 1–8.
53. Harris, P.C. (1997) The TSC2/PKD1 contiguous gene syndrome. *Contrib. Nephrol.*, **122**, 76–82.
54. Cheadle, J.P., Reeve, M.P., Sampson, J.R. and Kwiatkowski, D.J. (2000) Molecular genetic advances in tuberous sclerosis. *Hum. Genet.*, **107**, 97–114.
55. De Sanctis, G.T., Merchant, M., Beier, D.R., Dredge, R.G., Grobholz, J.K., Martin, T.R., Lander, E.S. and Drazen, J.M. (1995) Quantitative locus analysis of airway hyperresponsiveness in A/J and C57BL/6J mice. *Nature Genet.*, **11**, 150–154.
56. Caughey, G.H. (1997) Of mites and men: trypsin-like proteases in the lungs. *Am. J. Respir. Cell Mol. Biol.*, **16**, 621–628.
57. Hunt, J.E., Friend, D.S., Gurish, M.F., Feyfant, E., Sali, A., Huang, C., Ghildyal, N., Stechschulte, S., Austen, K.F. and Stevens, R.L. (1997) Mouse mast cell protease **9**, a novel member of the chromosome 14 family of serine proteases that is selectively expressed in uterine mast cells. *J. Biol. Chem.*, **272**, 29158–29166.
58. Johnson, P.R., Ammit, A.J., Carlin, S.M., Armour, C.L., Caughey, G.H. and Black, J.L. (1997) Mast cell tryptase potentiates histamine-induced contraction in human sensitized bronchus. *Eur. Respir. J.*, **10**, 38–43.

59. Rice, K.D., Tanaka, R.D., Katz, B.A., Numerof, R.P. and Moore, W.R. (1998) Inhibitors of tryptase for the treatment of mast cell-mediated diseases. *Curr. Pharm. Des.*, **4**, 381–396.
60. McInnes, L.A., Escamilla, M.A., Service, S.K., Reus, V.I., Leon, P., Silva, S., Rojas, E., Spesny, M., Baharloo, S., Blakenship, K. *et al.* (1996) A complete genome screen for genes predisposing to severe bipolar disorder in two Costa Rican pedigrees. *Proc. Natl Acad. Sci. USA*, **93**, 13060–13065.
61. Ewald, H., Mors, P., Flint, T., Koed, K., Eiberg, H. and Kruse, T.A. (1995) A possible locus for manic depressive illness on chromosome 16p13. *Psychiatr. Genet.*, **5**, 71–81.
62. Edenberg, H.J., Foroud, T., Conneally, P.M., Sorbel, J.J., Carr, K., Crose, C., Willig, C., Zhao, J., Miller, M., Bowman, E. *et al.* (1997) Initial genomic scan of the NIMH genetics initiative bipolar pedigrees: chromosomes 3, 5, 15, 16, 17 and 22. *Am. J. Med. Genet.*, **74**, 238–246.
63. Kamei, M., Webb, G.C., Young, I.G. and Campbell, H.D. (1998) SOLH, a human homologue of the Drosophila melanogaster small optic lobes gene is a member of the calpain and zin-finger gene families and maps to human chromosome 16p13.3 near CATM (cataract with microphthalmia). *Genomics*, **51**, 197–206.
64. Kamei, M., Webb, G.C., Heydon, K., Hendry, I.A., Young, I.G. and Campbell, H.D. (2000) Solh, the mouse homologue of the Drosophila melanogaster small optic lobes gene: organization, chromosomal mapping and localization of gene product to the olfactory bulb. *Genomics*, **64**, 82–89.
65. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–957.
66. Holmquist, G.P. (1992) Review article: Chromosomal bands, their chromatin flavors and their functional features. *Am. J. Hum. Genet.*, **51**, 17–37.
67. Kitzberg, D., Selig, S. and Cedar, H. (1991) Chromosome structure and eukaryotic gene organization. *Curr. Opin. Genet. Dev.*, **1**, 534–537.
68. Kellum, R. and Elgin, S.C. (1998) Chromatin boundaries: punctuating the genome. *Curr. Biol.*, **8**, R521–R524.
69. Prioleau, M.N., Nony, P., Simpson, M. and Felsenfeld, G. (1999) An insulator element and condensed chromatin region separate the chicken beta-globin locus from an independently regulated erythroid-specific folate receptor gene. *EMBO J.*, **18**, 4035–4048.
70. Bell, A.C. and Felsenfeld, G. (1999) Stopped at the border: boundaries and insulators. *Curr. Opin. Genet. Dev.*, **9**, 191–198.
71. Sun, F.L. and Elgin, S.C. (1999) Putting boundaries on silence. *Cell*, **99**, 459–462.
72. Buckle, V.J. and Rack, K. (1993) Fluorescent *in situ* hybridisation. In Davies, K.E. (ed.), *Human Genetic Diseases*. IRL Press, Oxford, UK, pp. 59–80.
73. Bankier, A.T., Weston, K.M. and Barrell, B.G. (1987) Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. *Methods Enzymol.*, **155**, 51–93.
74. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
75. Bonfield, J.K., Smith, K.f. and Staden, R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res.*, **23**, 4992–4999.
76. Flint, J., Sims, M., Clark, K., Staden, R. and Thomas, K. (1998) An oligo-screening strategy to fill gaps found during shotgun sequencing projects. *DNA Seq.*, **8**, 241–245.
77. McMurray, A.A., Sulston, J.E. and Quail, M.A. (1998) Short-insert libraries as a method of problem solving in genome sequencing. *Genome Res.*, **8**, 562–566.
78. Waye, J.S., Chui, D.H.K., Higgs, D.R., Hetherington, R. and Olivieri, N.F. (1995) *De novo* deletion of the entire ζ–α globin gene cluster in a girl with Hb H disease (Abstract). *Blood*, **86**, 8a.
79. Brook-Carter, P.T., Peral, B., Ward, C.J., Thompson, P., Hughes, J., Maheshwar, M.M., Nellist, M., Gamble, V., Harris, P.C. and Sampson, J.R. (1994) Deletion of the *TSC2* and *PKD1* genes associated with severe infantile polycystic kidney disease—a contiguous gene syndrome. *Nature Genet.*, **8**, 328–332.
80. Burn, T.C., Connors, T.D., Van Raay, T.J., Dackowski, W.R., Millholland, J.M., Klinger, K.W. and Landes, G.M. (1996) Generation of a transcriptional map for a 700-kb region surrounding the polycystic kidney disease type 1 (PKD1) and tuberous sclerosis type 2 (TSC2) disease genes on human chromosome 16p13.3. *Genome Res.*, **6**, 525–537.
81. Aspinwall, R., Rothwell, D.G., Roldan-Arjona, T., Anselmino, C., Ward, C.J., Cheadle, J.P., Sampson, J.R., Lindahl, T., Harris, P.C. and Hickson, I.D. (1997) Cloning and characterization of a functional human homolog of *E.coli* endonuclease III. *Proc. Natl Acad. Sci. USA*, **194**, 109–114.