Genome-wide DNaseI Hypersensitive Sites Profiles in Laboratory Mouse Strains by DNase-seq



Mona Hosseini

The Wellcome Trust Centre for Human Genetics

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2013

Acknowledgments

I wish to thank my supervisor Professor Jonathan Flint for his unfailing support, motivating scientific discussions, discipline, and constructive criticism and Professor Douglas Higgs for co-supervising my project. I wish to extend my gratitude to many people who helped and inspired me during this 4 year DPhil programme: Dr Marco De Gobbi for his assistance with developing DNase-seq protocol. Dr Andrew Edwards and Dr James Cleak for their help with animal experiments. Dr Jacqueline Sloane Stanley for her assistance with measuring Ter119+ cell fragments. Dr Ben Davies for his kind help of preparing MEFs from C57BL/6J. Polinka Hernandez-Pliego and Stuart Davidson for their assistance with MEF cell culture. Dr Stefania Segditsas for her useful discussions on qPCR design. Professor Richard Mott, Dr Leo Goodstadt, Dr Jim Hughes, and Georg Otto, Dr Richard Copley for their assistance and comments on the analysis of the data. Dr Monika Kowalczyk for preparing RNASeq libraries. Dr Jérôme Nicod, Dr Binnaz Yalchin, Lorna Gregory, Amarjit Bhomra for technical support. Mrs Rosie Butler for her friendship and encouragements. IT support team, Ruth Porter and Mark Gibbons for their constant help especially when my thesis draft threatened by Mac book failure! People who were part of dynamically changing Flint group and colleagues at the Wellcome Trust Centre for Human Genetics for their welcoming attitude anytime I needed their assistance and my college, Christ Church, which offered me accommodation and wonderful support for the whole duration of my DPhil, particularly I would like to thank my graduate Academic Assistant Miss Melanie Radburn and the college House Manager Mr Philip Tootill. Beside those who helped me technically, I wish to thank my parents, Mr Mohammad Reza Hosseini and Mrs Fatemeh Rezaei Nazari, my siblings (Nona, Fedia and Seyed Mahdi), my nieces and nephew (Hana, Elena, Mohhamad Parsa), and my dear family and friends (Marta, Areej, Sophie, Aisha, Narges, Zahra1, Amelie, Shima, Carme, Elham, Mahdieh, Afsaneh, Zahra2, Sara, Leila, …) who helped me to live during these years far from my dear homeland Iran with happiness, joy and positive attitude! And last but not the least I wish to thank Wellcome Trust and Clarendon Fund for funding my DPhil programme in Genomic Medicine and Statistics at the University of Oxford.

Declarations

I declare the results presented here are all my own work with the exception of the following:

Dr Marco De Gobbi and Dr Jim Hughes assisted me to develop DNase-seq library preparation protocol from existing DNase-chip protocol. Dr Jacqueline Sloane Stanley measured Ter119+ cell fragments. Dr Ben Davies prepared Mouse Embryonic Fibroblasts from C57BL/6J (MEFs). Polinka Hernandez-Pliego and Stuart Davidson provided assistance with MEF cell culture. Dr Monika Kowalczyk prepared RNASeq libraries from eight inbred mouse strains. Professor Jonathan Flint, Professor Richard Mott, Dr Leo Goodstadt, Dr Jim Hughes, and Georg Otto, Dr Richard Copley assisted with, commented on, and developed the automated peak calling pipeline development and genome-wide analysis of the DNase-seq and RNASeq data.

Abstract

Variation at regulatory elements, identified through hypersensitivity to digestion by Deoxyribonuclease I (DNase I), is believed to contribute to variation in complex traits, but the extent and consequences of this variation are poorly characterized. To investigate the relationship between sequence variation, and the functional consequences of variation in chromatin accessibility, genome-wide DNase I hypersensitive sites (DHS) of terminally differentiated erythroblasts were studied in eight inbred strains of mice studied (A/J, AKR/J, BALBc/J, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J, and LP/J). These strains were selected because of the availability of their genome sequence and quantitative trait loci (QTL) data.

After confirming that next generation sequencing could identify DNase I hypersensitive sites with high sensitivity and specificity, and that differential peaks could be found, an automated peak calling pipeline was developed and optimized. 36,693 DHS peaks were identified covering 9.1 Mb (0.29%) of mouse genome. There was no indication of within strain variation. Between strains reproducible variation was observed at approximately 5% of DNase hypersensitive sites (1,397 DHSs). Variable DHSs were more likely to be enhancers than promoters and less likely to occur at conserved regions of the genome. Only 36% of such variable DHSs contain a sequence variant predictive of site variation and 12% contain at least one variant that disrupts transcription factor binding sites. The majority (86%) of variable DHSs differ in size/shape and the remaining 14% demonstrate discrete variation in single peak or cluster of peaks. Sequence variants within variable DHS are more likely to be associated

with complex traits than those in non-variant DHS, and variants associated with complex traits preferentially occur in enhancer-like elements. Changes at a small proportion (7%) of discretely variable DHS are associated with changes in nearby transcriptional activity. Our results show that whilst DNA sequence variation is not the major determinant of variation in open chromatin, where such variants exist they are likely to be causal for complex traits.

Table of Contents

List of abbreviations **8**

Chapter 1 : Introduction **10**
1.1 Overview **10**
1.2 Objectives **15**
1.3 Background **15**
1.3.1 Transcriptional regulatory genetic variation and human phenotype variation and diseases 15
1.4 Transcriptional regulatory elements classifications **18**
1.4.1 Promoter 19
1.4.2 Distal regulatory elements 22
1.5 Genomic Regulatory Elements identifiable features **25**
1.5.1 Regulatory elements and conservation 25
1.5.2 DHS and gene regulation 26
1.5.3 Chromatin accessibility and gene regulation 28
1.5.4 Epigenetics and gene gegulation 29
1.5.5 Genomic analysis of TFBSs and gene regulation 32
1.6 Non-coding functional elements in the human genome **33**
1.7 Summary **35**

Chapter 2 : Methods **36**
2.1 Primary cells **36**
2.1.1 Terminally differentiated erythroblasts 36
2.1.2 C57BL/6J Mouse Embryonic Fibroblasts (MEFs) 36
2.2 DNase-seq **36**
2.2.1 DNase I assay 36
2.2.2 DNase-seq library preparation steps 38
2.2.3 Quantification of DNase I digestion by qPCR 41
2.2.4 DNaseseq analysis 44
2.3 RNASeq **46**
2.3.1 RNA sample preparation 46
2.3.2 RNASeq analysis 47

Chapter 3 : DNaseseq library preparation **48**
3.1 Introduction **48**
3.2 Developing DNaseI-seq protocol By modifying DNase-chip assay in K562 cells 48
3.2.1 Sorted vs. unsorted cells 52
3.2.2 DNase-chip on primary mouse erythroid cells 54
3.3 Quantification of DNase I digestion by qPCR **55**
3.3.1 qPCR fold change calculation 57
3.3.2 Assessing the efficiency of controls (C1, C2, and C3) and tests (R2, R3, and -14) primer sets 61
3.4 DNase I assays and selection of mid-phase of digestions by qPCR **71**
3.5 Preparing DNase-seq library from C57BL/6J MEFs **78**
3.5.1 Designing primers for SYBR-green qPCR DHS sensitive and control primer pairs for C57BL/6J MEFs 80
3.6 Summary **90**

Chapter 4 : DNase-seq data analysis 91
4.1 Introduction 91
4.2 DNase I sequencing Data sets 92
4.3 Sequencing results from C57BL/6J-1 and C3H/HeJ-1 95
4.4 DNase-seq data on GAII platform 103
4.4.1 Differential DHS peaks with GAII platforms 106
4.5 Within strain variation 110
4.6 Variable sites 115
4.7 Tissue specific DHS: erythroid vs MEFs 119
4.8 Conclusion 122

Chapter 5 : Genome-Wide analysis of sites of DNase hypersensitivy 123
5.1 Introduction 123
5.2 Genome-wide peak caller 124
5.3 Comparing DHS peaks with alpha and beta globin known regulatory elements 136
5.4 DHS characteristics 139
5.5 Detection of variable peaks 142
5.6 Classification of Variable Peaks 144
5.7 Comparison of variable and non variable peaks 152
5.8 The relationship between sequence variation and DHS peak variation . 155
5.9 Correlation of DHS variation and Transcription Factor Binding Sites (TFBSs) variation 165
5.10 Conclusion 166

Chapter 6 : Functional consequences of DHS variation 167
6.1 RNA-Seq library preparation and data analysis 168
6.2 Relationship between DHS and RNA expression variation in erythroblasts 171
6.3 Relationship between DHS and phenotypic variation 182
6.3.1 Testing the enrichment of candidate variants in variable/non-variable DHS sites within QTLs: discriminating functional variants 182
6.4 Conclusion 184

Chapter 7 : Discussion 185
7.1 Variable DHSs prevalence 186
7.2 Molecular basis of DHS variation 186
7.3 DHS variation and gene expression 188
7.4 Variable DHS and QTL analysis 189
7.5 Conclusion 191

Supplementary Material 193

References 205

LIST OF ABBREVIATIONS

ACS	Active chromatin sequence
Actb	Actin, beta
BRE	TFIIB-recognition element
ChIP	Chromatin immunoprecipitation
ChIP-Seq	Chromatin immunoprecipitation with massively parallel sequencing
CRE	*Cis*-acting regulatory elements
CTCF	CCCTC-binding factor
DCE	Downstream core element
DHSs	DNase I hypersensitive sites
DNase I	Deoxyribonuclease I
eDHSs	Erythroid-specific DNase I hypersensitive sites
EB	Elution buffer
ENCODE	Encyclopedia of DNA elements
FAIRE	Formaldehyde-assisted isolation of regulatory elements
FDR	False discovery rate
Gapdh	Glyceraldehyde-3-phosphate dehydrogenase
GTFs	General transcription factors
GWAS	Genome-wide association studies
HS	Heterogeneous stock
Inr	Initiator element
LCR	Locus control regions
LM-PCR	Ligation mediated PCR
LMRs	Low-methylated regions
MAD	Median absolute deviation
MAPQ	Mapping quality
MARs	Matrix scaffold-attachment regions
MEF	Mouse embryonic fibroblasts
MNase	Micrococcal nuclease
MPSS	Massively parallel signature sequencing
MSC	Multispecies conserved sequences
NHGRI	National Human Genome Research Institute
PCR	Polymerase chain reaction
PIC	Pre-initiation complexes
PREs	Polycomb group response elements
Q20	*Phred* quality score 20
qPCR	Quantitative PCR
QTL	Quantitative trait locus
RIN	RNA integrity number
rSNP	Regulatory single nucleotide polymorphism
SARs	Scaffold-attachment regions
SDP	Strain distribution pattern
SNPs	Single nucleotide polymorphisms
TAS	Trait/disease-associated SNPs

TF	Transcription factor
TFBSs	Transcription factor binding sites
TSS	Transcriptional start sites

CHAPTER I : INTRODUCTION

1.1 OVERVIEW

It is increasingly apparent that the differences outside coding regions are a major, if not the major, contributor to common disease and to quantitative variation in many phenotypes of biomedical importance. We know the consequence of changes that affect coding regions of genes, through the alteration of protein structure, the truncation or elimination of transcription, the production or deletion of exons; we are much less clear about the consequences of changes that lie outside coding regions (Chorley et al., 2008).

Genome-wide association studies (GWAS) in human indicate many functional sequence variants lie within non-coding (intronic/intergenic) regions and only 12% of trait/disease-associated SNPs (TASs) map to or in tight linkage disequilibrium with protein-coding loci (Manolio, 2010, Hardy and Singleton, 2009). Hindorff and Sethupathyb et al. have developed an online database from published TASs i.e. National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwa studies) (Hindorff et al.). Among 531 TASs (465 unique TASs) they reviewed in 2009, 88% were intronic (45%, n=208) or intergenic (43%, n=199) and only 9% (n=41) were nonsynonymous (Hindorff et al., 2009, Kent et al., 2002).

There is evidence that regulatory sequence variants may contribute to the genetic component of phenotypic variation and to disease. One well-documented example is a single base pair change that creates a novel promoter element, driving the expression of a novel transcript that, in turn, shuts down the transcription of the alpha globin genes (De Gobbi et al., 2006). This regulatory

single nucleotide polymorphism (rSNP) was discovered by examining a family with a hereditary anemia (alpha thalassaemia). Importantly, the rSNP lies outside all known globin regulatory regions, does not coincide with a region of sequence conservation and is not predicted to be functional from existing analysis of non-coding DNA (Encyclopedia of DNA elements (ENCODE)) project (Feingold et al., 2004)). Since the human genome contains upwards of 10 million SNPs, the vast majority of which lie outside conserved regions and have no known function, the discovery of this functional rSNP raises the possibility that many other sequence variants in the genome also have regulatory effects. This example demonstrates a molecular mechanism through which a sequence variant alters a regulatory element, which in turn has functional consequences. It raises the question of the prevalence of such a mechanism in the genome. My thesis addresses the question of the extent to which underlying sequence variation relates to variation in gene regulation, and how frequently that variation has functional consequences.

In order to investigate the relationship between sequence and regulatory element variation I took advantage of a recent high-throughput sequencing project that has generated complete sequence from 17 inbred mouse strains. These data, generated through collaboration between the Sanger Institute and Oxford University, provide a critical and unique resource for determining the function of sequence variants. The sequence was acquired by >20 fold genome coverage with next generation reads (Illumina) (Keane et al., 2011, Yalcin et al., 2011).

I was able to assess regulatory regions across the genome because of their known sensitivity to Deoxyribonuclease I (DNase I). In order for regulatory

proteins to bind DNA they must either displace nucleosome packaging (which consists of chromatin) or the DNA must be maintained in an open configuration, free from nucleosomes. This naked DNA is also accessible to nucleases, so a simple assay of protein accessibility is its ease of digestion. Regions of DNA that are especially sensitive to nuclease digestion (DNase I hypersensitive sites, DHSs) are known to identify a number of different types of regulatory elements, including promoters, enhancers, silencers, insulators, and locus control regions (LCR) (Gross and Garrard, 1988, Wu, 1980, Elgin, 1984, Felsenfeld et al., 1996, Felsenfeld and Groudine, 2003).

My project required a method for screening the genome for DHSs. Traditionally DHSs have been identified by Southern blotting, a highly accurate method (Wu, 1980) but one that is technically demanding and not suitable to high throughput applications. Over the last years a number of methods have been developed that can deliver whole genome level analysis of DHSs. Chromatin immunoprecipitation (ChIP) coupled to DNA microarrays (Ren et al., 2000) or high-throughput DNA sequencing (Wei et al., 2006, Johnson et al., 2007) are able to deliver high throughput analyses, but do not deliver nucleotide-level resolution of DHS. However dense massively parallel sequencing around DNase I cleavage sites, assayed by DNase I digestion of intact nuclei, has been shown to discover in vivo occupancy sites of DNA-binding proteins genome-wide, at single nucleotide resolution (Hesselberth et al., 2009, Boyle and Furey, 2009). I used Solexa high-throughput sequencing for genome-wide DHSs identification.

A major source of variation in assays of DHS is the purity and state of differentiation of the tissue. DHS profiles differ in various tissues reflecting tissue-specific variation of regulatory elements (Bernstein et al., 2012, Neph et

al., 2012a, Thurman et al., 2012). As I intended to study the extent and cause of variation between individuals, using a tissue with multiple different cell types and at different stages of differentiation (such as the dentate gyrus of the hippocampus), would introduce large amounts of variation in the assay, compromising power to detect effects. Therefore I used terminally differentiated mouse erythroblasts. These cells can be obtained from the spleens of mice in which haemolytic anaemia has been induced. This is a standard method, used routinely in Professor Higgs' laboratory where I worked (Spivak et al., 1973).

My project proceeded from the construction and validation of Illumina libraries on three inbred mouse strains (A/J, C57BL/6J, and C3H/HeJ), to the analysis of eight strains. Selection of these specific inbred strains was due to the fact that these inbred strains have been completely sequenced as part of the Mouse Genomes Sequencing Consortium (A/J, AKR/J, C3H/HeJ, C57BL/6J, BALBc/J, CBA/J, DBA/2J and LP/J)(Keane et al., 2011). In addition, these eight inbred mouse strains are the founders of heterogeneous stock (HS) mice from which more than 840 quantitative trait loci (QTLs) have been identified (Valdar et al., 2006). I anticipated that sufficient single/paired end reads would be generated from a single flow cell to allow unambiguous determination of the position of all DHSs in the tissue (DHSs account for about 1% of the genome (Gross and Garrard, 1988)). I compared the position of DHSs in each strain to the sequence variants identified from whole genome sequencing. The final part of my project involved testing whether the sequence variants cause any observed variation in the transcription or gene regulation (as assayed by variation in DHSs positions).

In order to investigate the functional consequences of variation in DHSs, I assessed the consequences of DHS variation on gene regulation and phenotypic variation. To address phenotypic variation, I took advantage of genome-wide mapping data in the mouse for multiple complex traits. Over 840 QTLs for 97 phenotypes were mapped in over 2,000 HS mice descended from eight inbred progenitor strains (A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J and LP/J). The large number of recombinants that have accumulated since the founding of the stock means that genetic mapping resolution is an order of magnitude higher than that obtained by mapping in backcrosses or intercrosses (Valdar et al., 2006). I would be able to relate sequence variation to DHSs and to relevant phenotypic QTLs. The HS data include mapping of QTLs for disease models (asthma, anxiety and type 2 diabetes), as well as haematological (including full blood cell count, and phenotypes relevant to the tissue I analyzed), immunological, biochemical and anatomical assays (Yalcin et al., 2005). Therefore we could ask if sequence variants contributing to DHS variation are also likely to influence quantitative traits.

1.2 OBJECTIVES

To study the relationship between sequence variation, DHS variation and phenotypic variation between individuals, I aimed to identify DNase I hypersensitive sites genome-wide in 8 inbred mouse strains by DNase-seq. With these data I aimed to estimate the proportion of heritable variation in DHS and determine first whether local sequence variants contribute to the mouse strain variation in DNase I hypersensitivity sites; second, I aimed to obtain evidence for the functional consequences of a number of such variants by testing the effect of variable DHS on gene expression profiles and phenotypic variation. I set out to find candidate variants with an effect on complex phenotypes, whose involvement could then later be confirmed by engineering mutations in transgenic animals. This would be a major step forward towards understanding the relationship between sequence and phenotypic variation.

1.3 BACKGROUND

In this section I briefly review the relationship between variation in transcriptional regulatory elements and phenotypic variation and human disease. I review evidence for the involvement of regulatory sites in monogenic (Mendelian and non-Mendelian) human diseases as well as their involvement in common complex human diseases.

1.3.1 TRANSCRIPTIONAL REGULATORY GENETIC VARIATION AND HUMAN PHENOTYPE VARIATION AND DISEASES

Competent spatial, temporal and tissue specific expression of genes is essential in the precise execution of different steps of biological processes including

proliferation, development, differentiation, apoptosis, and aging (Maston et al., 2006). Regulatory sequences have a critical role in shaping the diversity of structure and function of different cells from fixed genetic material in metazoans. Deregulation of gene expression on the other hand can cause pathological conditions (Knight, 2005).

Following the completion of the human genome project and the cataloguing of ~20,000-25,000 protein-coding genes (Collins et al., 2004), understanding the molecular mechanisms of expression patterns and identifying the regulatory elements associated with each gene and how gene expression alters in pathological condition became important aims for explaining the contribution of non-coding functional elements to phenotype diversity.

1.3.1.1 Monogenic Mendelian and non-Mendelian human diseases with regulatory involvement

Structural variation in coding sequence can explain the pathophysiology of many Mendelian /monogenic disorders. Online Mendelian Inheritance in Man (OMIM) lists 14,213 gene descriptions among which 3,748 phenotypic descriptions and/or the molecular basis are known (2012b). Regulatory element involvement in human conditions has not been studied that extensively; however, there are now some classical examples of regulatory mutations leading to disease phenotype.

β-globin promoter mutations in β thalassemia were among the first *cis*-regulatory elements characterized (Antonarakis et al., 1984). For instance, mutations in ATATAA (TATA box) compromise TF2D binding and mutations in GCCAGACCC compromise the binding of erythroid cell specific transcription

factor Erythroid Kruppel-like Factor (EKLF) (Feng et al., 1994).

Mutation in *PAX6* coding sequence leads to aniridia (absence of iris); however absence of this mutation in number of aniridia patients whom their cytogenetic studies identified breakpoints far downstream of *PAX6* gene suggested the role of distal *cis*-acting regulatory elements. Detailed DNase I hypersensitivity mapping of the *Pax6* critical genomic region (i.e. 20 kb upstream to P0 promoter to ~170 kb downstream of the gene), together with multi-species sequence conservation studies, led to the identification of multiple novel eye-specific enhancers in the *Pax6* downstream regulatory region in the last intron of the *Elp4* gene (McBride et al., 2011).

Other studies also predicted significant roles for regulatory elements (including core promoters, proximal promoters, enhancers, silencers, insulators and LCRs) in phenotypic variation and disease mechanisms (Pastinen and Hudson, 2004, Grice et al., 2005, Manolio et al., 2008, Emison et al., 2005, Maston et al., 2006, Noonan and McCallion, 2010). Preaxial polydactyly due to point mutations of a ZRS enhancer 1 Mb upstream of the *SHH* gene (Lettice et al., 2003) and X-linked deafness due to an enhancer microdeletion 900 kb upstream of *POU3F4* gene (deKok et al., 1996) are two examples of involvement of long-range regulatory elements involved in human diseases.

Non-coding mutations account for 1%-2% of disease mutations; the majority of these are mapped to proximal and distal promoters within 1 kb of transcriptional start sites (TSS) (Epstein, 2009). 1,459 regulatory mutations have been identified over 700 genes causing inherited disorders in human (April 2012 statistics complied by the Human Gene Mutation Database). To date, this figure has been expanded to 1,711 out of 97,797 human disease mutation (1.74% of

disease mutations are regulatory mutations) (2012a).

1.3.1.2 Common complex human diseases association with non-coding variants

Genome-wide association studies (GWAS) are designed to study the association of hundred of thousands of single nucleotide polymorphisms (SNPs) with a trait/disease in thousands of individuals (Hardy and Singleton, 2009). This method has been widely used in recent years to study the genetic basis of complex traits (Manolio, 2010). To date, published GWASs (1,617 studies with significant associations $p \leq 5 \times 10^{-8}$) have identified 249 loci involved in complex traits in humans (Hindorff et al.). 88% of GWAS significant signals lie in non-coding genomic loci (45% intronic and 43% intergenic) (Hindorff et al., 2009). These human common disease and quantitative trait-associated variants are enriched in regulatory genomic regions marked by DNase I hypersensitivity (Maurano et al., 2012). The ENCODE consortium examined 5,654 noncoding GWAS variants (including 5,134 unique SNPs) associated with 207 common disease and 447 quantitative traits. ENCODE reported that 76.6% of non-coding GWAS SNPs lie within DHS (2,931 SNPs) or in complete linkage disequilibrium (LD) with SNPs in adjacent DHS (999 SNPs). Thus, disease associated variants are significantly more likely to lie within DHS than in other regions of the genome.

1.4 TRANSCRIPTIONAL REGULATORY ELEMENTS CLASSIFICATIONS

The regulation of eukaryotic protein-coding gene expression can occur in various steps, including transcription initiation and elongation, as well as mRNA

processing, transport, translation and stability among which regulation mostly occurs at transcription initiation step. RNA polymerase II is responsible for protein-coding gene transcription. Transcriptional regulation of these genes usually occurs through two types of *cis*-acting regulatory elements (CRE): a promoter (core promoter and proximal promoter elements) and distal regulatory elements (enhancers, silencers, insulators, and LCR) (Maston et al., 2006) (Figure 1-1). CREs are usually a non-coding DNA containing a binding site for transcription factor (TF) that activate or maintain transcription and *trans*-acting regulatory sequence encode TFs. *Cis*-acting regulatory elements interact with *trans*-acting transcription factors through their recognition sites to enhance/repress transcription (Wittkopp and Kalay, 2012). Eukaryotic transcriptional machinery consists of general transcription factors (GTFs) which form pre-initiation complexes (PIC) when bound to core promoters and direct RNA polymerase II to transcriptional start sites (Maston et al., 2006). Eukaryotic regulatory elements function synergistically (McKnight and Kingsbury, 1982). Detailed characteristics of each type are described below.

1.4.1 PROMOTER

1.4.1.1 Core promoter

A core promoter is a region at the beginning of a gene providing a scaffold for basic transcriptional machinery. It determines the transcription start site (TSS) position and the direction of transcription (Smale and Kadonaga, 2003). Numerous core promoters have been described so far, including TATA box (25 bp upstream to TSS), Initiator element (Inr), Downstream Promoter Element (DPE), Downstream Core Element (DCE), TFIIB-Recognition Element (BRE).

These core promoter motifs are not universally present in all ~10,000 estimated human promoters. Inr are the most common form (present in 1/2 core promoters) followed by DPE and BRE found in almost 1/4 of core promoters; a TATA box is present in one eighth (Gershenzon and Ioshikhes, 2005, Bucher, 1990).

1.4.1.2 Proximal promoter elements

A proximal promoter element is a region up to few hundred base pairs immediately upstream to core promoter, usually containing multiple activator binding sites. Many mammalian promoters are characterized by the presence of unmethylated CpG dinucleotides. In mammalian genomes CpG dinucleotides are typically methylated on the cytosine residue. However, in the vicinity of housekeeping genes and regions implicated in gene regulation (many tissue-specific genes) CpG dinucleotides are not methylated (Gardinergarden and Frommer, 1987, Takai and Jones, 2002). These regions, called CpG islands, are short DNA segments (200 bp to 2 kb) with a high C+G nucleotide content (>50%) and frequency of CpG dinucleotides (observed/expected frequency ≥ 0.6) (Venter et al., 2001). Almost 60% of human promoters are associated with CpG islands but they are less frequent in the mouse genome (15,500 CpG islands in mouse vs 27,000 in human) (Waterston et al., 2002). Their strong association with promoters makes them a reliable tool for promoter detection and gene prediction (Ioshikhes and Zhang, 2000, Jones, 2006). CpG island methylation is involved in gene inactivation and plays a role in imprinting and tissue-specific gene expression (Venter et al., 2001) and also leads to repressive chromatin structure (Jones et al., 1998).

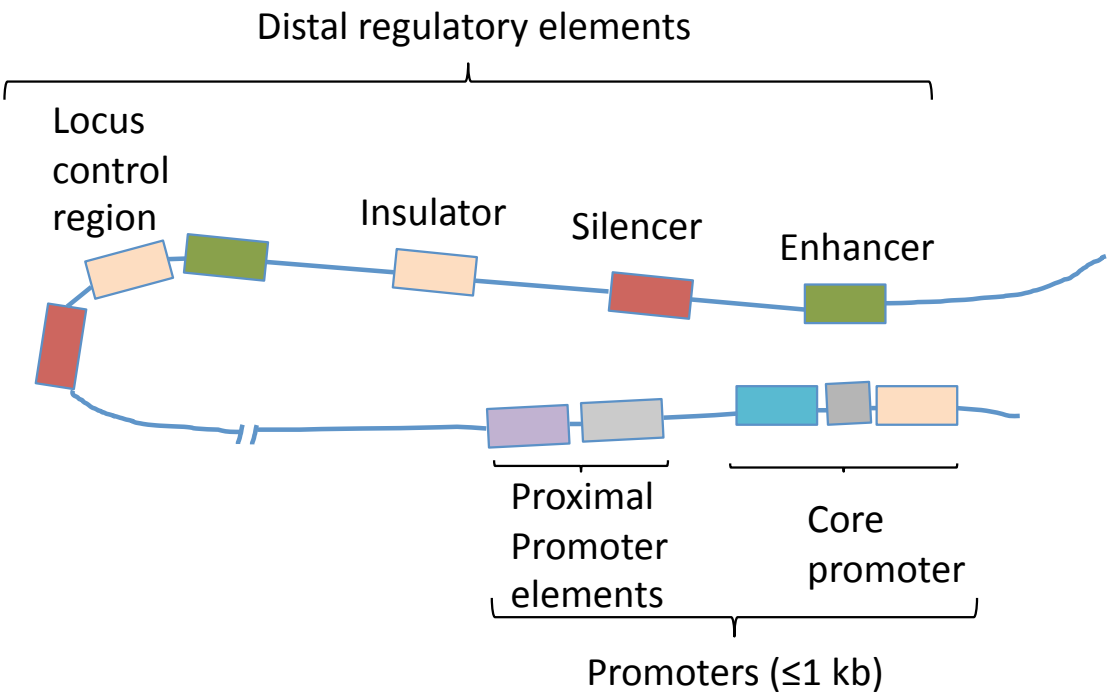


Figure 1-1 - Proximal and distal gene regulatory elements. Promoters consist of core promoter and proximal promoter elements within a span of 1 kb and upstream distal regulatory elements could be located within 1 Mb from the promoter (*Figure adapted from Maston et al., 2006*).

1.4.2 DISTAL REGULATORY ELEMENTS

1.4.2.1 Enhancers

Enhancers are typically long-distance regulatory elements that markedly increase transcription of a gene in a spatial- and temporal-specific manner. Their function is independent of distance and orientation to promoters. They typically contain clusters of closely grouped transcription factor binding sites (TFBSs) (Maston et al., 2006). Their function has been shown to be modular and additive (Visel et al., 2009a). In extreme cases, enhancers function >1 Mb of their target gene promoter (Lettice et al., 2003). Different models (looping, tracking, linking) have been suggested to explain the mechanism of action of distal enhancers (Bulger and Groudine, 2002). Cohesin is involved in higher order chromatin structure organization and it is thought to facilitate and stabilize promoter-distant enhancer (long-range) interactions by forming chromatin loops that introduces enhancer promoter proximity by interacting with transcriptional repressor CCCTC-binding factor (CTCF) binding sites (Ong and Corces, 2011).

1.4.2.2 Silencers

Silencers are sequence specific regulatory elements repressing transcription of their target gene through binding to negative TFs (called repressors). Similar to enhancers, their function is independent of orientation and distance from the promoter. They may be located independently far from their target genes, or in the intronic or 3' UTR region of their target gene. They can act synergistically (Maston et al., 2006).

Two classes of silencers (short-range and long-range) have been described in Dorsophila: short-range silencers affect their target gene within ~

100 bp whereas long-range silencers affect multiple promoters and enhancers up to few kilo basepairs (Kulkarni and Arnosti, 2005). Polycomb group Response Elements (PREs) are a class of silencers that have dual silencers/anti-silencer functions based on the availability of non-coding transcription across PRE and the type of protein that they bound to. The study of PREs might shed a light on understanding on how a repressive transcriptional state switches to an active state (Schmitt et al., 2005).

1.4.2.3 Insulators (boundary elements)

Insulators (~0.5-3 kb in length) protect genes from the influence of transcriptional regulation of neighboring genes. As a result, they limit the activity domain of regulatory elements and subsequently divide the genome into discrete expression parts. Insulators function in an orientation-independent and position-dependent manner. Their two main functions are to block the spread of heterochromatin (heterochromatin-barrier activity) and avoid enhancer-promoter interaction (enhancer-blocking activity).

CTCF is the only known protein that mediates insulator activity in vertebrates (Maston et al., 2006). CTCF binding sites are associated with precise nucleosome positioning (up to 20 nucleosomes in either side) (Ohlsson et al., 2010). Cohesin also mediates CTCF-dependent enhancer activity blockade by recruiting insulator sites (Wendt et al., 2008). CTCF also functions in imprinting and X-chromosomal inactivation. Allele-specific binding of CTCF to pivotal regions in regulation of several imprinted genes has been reported (Ohlsson et al., 2010).

1.4.2.4 Locus Control Regions (LCRs)

LCRs are groups of regulatory elements implicated in the regulation of an entire locus or gene cluster and they strongly enhance physiological expression of their linked transgenes in a copy number-dependent, position-independent and tissue-specific manner. LCRs are marked by clusters of DNase I hypersensitive sites which contain both tissue-specific and ubiquitous TFBSs (Li et al., 2002). They provide open chromatin for the genes they regulate. LCRs are composed of numerous *cis*-acting elements (enhancers, insulators, silencers, nuclear-matrix or chromosomal scaffold-attachment regions (MARs or SARs)). They are commonly located upstream of their linked gene(s) but they can also be found in the intron of their target gene, downstream of the gene or in the introns of their adjacent gene (Maston et al., 2006). Like enhancers and silencers, LCRs have long-range activity through loop formation and formation of clustering of sequences into "active chromatin hubs" which sustains open chromatin domains (Tolhuis et al., 2002).

1.5 GENOMIC REGULATORY ELEMENTS IDENTIFIABLE FEATURES

1.5.1 REGULATORY ELEMENTS AND CONSERVATION

One way to identify *cis*-regulatory elements is to determine evolutionary conserved motif sequences (Lenhard et al., 2003, Ureta-Vidal et al., 2003) through multiple-species comparison, a method termed phylogenetic footprinting (Tagle et al., 1988). Under selective pressure in evolution, the accumulation of mutations in functional genomic loci occurs more slowly than in neutral DNA. The use of phylogenetic footprinting, combined with TFBS data, provides better predictive accuracy of the location of potential regulatory elements than TFBS data alone (Wasserman and Sandelin, 2004).

Analysis of multispecies conserved sequences (MCS) in non-coding regions can improve the annotation of regulatory elements, but the method has the limitation that regulatory elements are not fully conserved (Hughes et al., 2005). Although regulatory elements are more conserved in comparison to nonfunctional sequences they can endure various sequence alterations, including inversion, without apparent functional consequences (Jones, 2006, Ludwig et al., 2000).

Dermitzakis and Clark studied the efficacy of phylogenetic footprinting to predict functional regulatory elements. Based on their functional assays of 20 regulatory sites in human and rodents, they estimated 32-40% of functional regulatory elements in human are not functional in rodents, which indicates a high turnover of TFBSs (2002). Measures of evolutionary constraint detected only 29-61% of functional regulatory elements in the study of the zebrafish *Phox2b* gene regulatory region. Reliance on standard evolutionary constraint

methods overlooked 42-51% of significant functional information *in vivo* (McGaughey et al., 2008). These findings are consistent with ENCODE project pilot data. Since 5% of the human genome is highly conserved (MCS), and of this protein-coding genes account for 1.5%, then the remaining 3.5% is likely to be functional non-coding sequences (Siepel et al., 2005). However, unlike protein-coding sequences, many regulatory elements are not identified as MCSs (King et al., 2007, Birney et al., 2007) suggesting a more comprehensive approach is required to study regulatory elements. Other studies showed non-conserved functional regulatory elements. For example some regulatory elements of the Beta globin locus are not conserved between human and mouse (Hughes et al., 2005, King et al., 2005).

1.5.2 DHS AND GENE REGULATION

The identification and characterization of non-coding functional elements is critical for an understanding of the dynamics of gene regulation and gene expression. Identifying regions of the genome, which are hypersensitive to cleavage of DNase I is a conventional and robust technique to map active functional elements (Keene et al., 1981, McGhee et al., 1981).

1.5.2.1 Eukaryotic dynamic chromosomal structural organization and DHS

Eukaryotic genomic nucleus DNA is packaged in regularly spaced protein complexes called chromatin. The basic repeating unit of chromatin is the nucleosome, which consists of DNA wound almost twice around a core of histone octamers (two copies of each canonical histone: H2A, H2B, H3 and H4) and

connected by linker DNA to form higher order structures. Nucleosome DNA in all eukaryotes contains 147 bp, whereas linker DNA length varies in different species (15-55 bp) (Wu, 1980, Hewish and Burgoyne, 1973, Weintraub et al., 1975). A chromatin fiber is a flexibly jointed chain of nucleosomes. Accurate positioning of nucleosomes leads to appropriate genomic function. Packaged DNA is less accessible than naked DNA to other proteins, including the proteins involved in transcriptional regulation, so that chromatin packing renders the DNA transcriptionally inactive (closed chromatin). In contrast regions of the genome that are free of chromatin (open chromatin) are available to transcription initiation complexes and are transcriptionally active. Consequently, one way to find where genes are being transcribed is to look for regions of open chromatin.

Nucleosome depleted DNA is accessible also to proteins that digest DNA, (DNases). The DNA component of nucleosome is less sensitive to DNase digestion compared to linker DNA. Therefore, it is possible to study transcriptional competency by looking for sequence that is hypersensitive to cleavage by the non-specific endonuclease DNase I. These are DNase I hypersensitive regions (Lu and Richardson, 2004, Felsenfeld and Groudine, 2003, Lodish et al., 2004, Jansen et al., 2012, Sproul et al., 2005). Gene regulatory elements are associated with nucleosome-free regions, which are hypersensitive to DNase I cleavage (DNase I hypersensitive sites (DHS)) (Stalder et al., 1980, Varshavsky et al., 1978, Wu, 1980, Saragosti et al., 1980). Binding of trans-acting factors in nucleosome depleted regions and consequent local chromatin conformational alterations lead to increase in chromatin accessibility and DHSs formation (Felsenfeld, 1992, Felsenfeld and Groudine, 2003, Gross and Garrard,

1988, Cockerill, 2011).

1.5.3 CHROMATIN ACCESSIBILITY AND GENE REGULATION

Chromatin organization mediates gene activity at various interconnected levels (DNA sequence, nucleosome, chromatin folding and chromosome arrangement). Efficient and accurate chromatin packaging is essential for temporal, spatial and cell specific gene regulation. The three main levels of chromatin organization are nucleosome packaging, the 30 nm fiber (estimated linear compaction of 100-200 kb μm-1) and higher folding levels which form metaphase chromosomes (Fransz and de Jong, 2011).

1.5.3.1 Nucleosome positioning and gene regulation

It has been known for many years that nucleosomal positioning with respect to DNA affects gene regulation (Straka and Horz, 1991, Lohr and Lopez, 1995). Nucleosomal DNA arrangement controls gene expression by preventing the interaction between DNA and DNA-binding proteins, such as activators and repressors (Richmond and Davey, 2003). DNA sequence, histone modifications, DNA methylation, TFs, and chromatin remodeling modify chromatin activity through their effect on nucleosome positioning and chromatin accessibility (Tsankov et al., 2010). For example more accessible chromatin conformation indicates more active genes (Weintraub and Groudine, 1976, Axel and Garel, 1977, Stalder et al., 1980).

Recently, whole-genome nucleosome positioning has been studied in yeast (Yuan et al., 2005, Lee et al., 2007), worm (Johnson et al., 2006) and human (Ozsolak et al., 2007). Nucleosomes positioned 200bp either side of polymerase II promoters were first described by high-throughput nucleosome mapping in

the yeast genome (Yuan et al., 2005). A study of 3,692 promoters in 7 human cell lines, also showed that actively transcribed genes associate with nucleosome free regions at their TSSs and that nucleosomes are a major determinant of DNA accessibility for TFs (Ozsolak et al., 2007). An association between nucleosome patterning and gene expression levels was demonstrated in yeast and human cells (Schones and Zhao, 2008). These genome-wide studies used micrococcal nuclease (MNase) digestion (to preferentially digest nucleosome-free (linker) DNA over nucleosomal DNA) and tiling microarrays (Yuan et al., 2005, Lee et al., 2007, Ozsolak et al., 2007) or high-throughput sequencing (Johnson et al., 2006, Albert et al., 2007) to map nucleosome occupancy and positioning. Nucleosome position also can be detected by specific histone modification ChIP-seq data in certain genomic regions (Barski et al., 2007b, Schmid and Bucher, 2007).

1.5.3.2 Mediator, Cohesin and gene regulation

Mediator, a transcriptional coactivator, and Cohesin provide another evidence for connection between chromatin structure and cell-type specific gene expression. Cell-type specific Mediator Cohesin complex occupancy mediates enhancer-promoter chromatin loop formation of active genes in mouse embryonic stem cells (Kagey et al., 2010).

1.5.4 EPIGENETICS AND GENE GEGULATION

Cell type specific gene expression patterns in metazoans are maintained through epigenetic mechanisms, heritable changes excluding DNA sequence variants. These chromatin modifications include combinations of DNA methylation and post-translational histone modification and are cell-type specific (Schones and Zhao, 2008).

1.5.4.1 DNA methylation

DNA methylation, the addition of methyl groups to cytosine in CpG dinucleotides, is known to have a gene regulatory function in mammalian genomes (Bird, 2002). CpG islands (Gardinergarden and Frommer, 1987, Takai and Jones, 2002) are associated with many human genes and are resistant to this covalent modification (Ioshikhes and Zhang, 2000). DNA methylation leads to inactivation of CpG islands (Bird, 2002) as well as alteration in the binding characteristics of TFBSs (Bell et al., 1999, Hark et al., 2000, Tate and Bird, 1993) Whole-genome base-pair resolution analysis of methylation in mouse showed CpG poor regions have reduced levels of methylation (~30%). These low-methylated regions (LMRs) contain DNA-binding sites and have distal active regulatory elements features including DNase I hypersensitivity (Stadler et al., 2011).

1.5.4.2 Post-translational histone modifications

Various models have been suggested to describe the role of histone modifications in transcription. These include charge neutralization (Roth and Allis, 1992), a histone code (Strahl and Allis, 2000, Turner, 2000) and signaling pathway (Schreiber and Bernstein, 2002) models). Histone covalent changes generally occur at the N-terminal tails and take the form of methylation, phosphorilation, acetylation and ubiquitylation.

Whole-genome approaches have been used to study histone modification profiles in different model organisms (ChIP-chip, ChIP-seq, and ChIP-SAGE) (Schones and Zhao, 2008). These studies indicate that transcription positively correlates with histone H3 lysine 4 methylation (H3K4me) and histone acetylation. Further evidence of H3K4 trimethylation (H3K4me3) signals at

silent promoters suggest this histone mark might be involved in priming an active chromatin state, indicating previous transcriptional activity. Many histone marks of this class cluster at regulatory elements and introduce significant redundancy in the system. For instance, H3K9me3, H3K27me3, and H4K20me3 are mainly observed in repressed chromatin and heterochromatin formation. Also it has been shown that H3K4me3 and H3K27me3 with opposite functions coexist at regulatory elements. The relative level of these modifications is thought to influence gene activity (Roh et al., 2006).

Distinct histone modifications have been described at specific regulatory sites (Heintzman et al., 2007, Birney et al., 2007, Barski et al., 2007a). For example, high levels of H3K4me1 and very low levels of H3K4me3 are a signature of some enhancers (Heintzman et al., 2007), although there is evidence of high H3K4me3 levels in some other enhancers (Barski et al., 2007a). Active promoters show high levels of H2BK5me1, H3K9me1, H3K27me1, H4K20me1 and H3K36me3 (Barski et al., 2007a). High levels of H3K4me3 and low levels of H3K4me1 have also been described as a signature of active promoters. H3K4me2 and H3K27Ac also indicate presence of promoters or enhancers. These modifications might be used to improve genomic annotations as they reflect district functional elements. The deposition of H2A.Z and H3.3, histone variants, in nucleosomes indicates active chromatin implicating lower nucleosomal stability and higher nucleosomal mobility (Schones and Zhao, 2008, Suto et al., 2000). Several histone modifications are associated with DHSs; for instance, H3K4me1-3 and H3K9me1 (Xi et al., 2007, Barski et al., 2007a).

1.5.5 GENOMIC ANALYSIS OF TFBSS AND GENE REGULATION

One approach to study gene regulation is based on transcription factors, which preferentially bind specific DNA motifs (Lemon and Tjian, 2000). ChIP-seq (Chromatin immunoprecipitation with massively parallel sequencing) is a gold-standard method to study individual TFs genome-wide. The limitation is that one can only study a single TF in a specific condition within a specific tissue in a single experiment (Johnson et al., 2007).

Sequence-specific motifs of TFs can be accessed from databases (e.g. TRANSFAC or JASPAR (Wingender et al., 1996, Sandelin et al., 2004)) but among genomic sequences there are multiple matches to binding motifs, not all of which actually bind TFs. CENTIPEDE is a recently developed algorithm, which incorporates cell-specific experimental data and sequence information to map bound TFBSs (Pique-Regi et al., 2011).

Various genome-wide approaches have been used to study TF binding, including methods which assess chromatin accessibility such as DNase I sensitivity or FAIRE (formaldehyde-assisted isolation of regulatory elements) assays (Boyle et al., 2008, Gaulton et al., 2010), protection of bound TFBS from DNase I digestion (DNase footprinting) (Fu et al., 2008, Hesselberth et al., 2009, Chen et al., 2010, Boyle et al., 2011), ChIP-seq to study histone modifications (Heintzman et al., 2007) or p300 (a coactivator protein) (Visel et al., 2009b, Pique-Regi et al., 2011).

1.5.5.1 DHS identification: conventional vs genome-wide approaches

The conventional method for identifying DHS is an indirect end-labeling Southern blotting technique, described in the 1980s (Wu, 1980). This method is

very accurate but laborious and not scalable to genome-wide investigations. To date, several genome-wide approaches have been introduced to study DHSs including DNase-chip (array), which combines DNase assay with array technology (Crawford et al., 2006a, Sabo et al., 2006), DNase-seq/massively parallel signature sequencing (MPSS), which combines DNase assay with high throughput sequencing (Crawford et al., 2006b, Hesselberth et al., 2009), DNase-qPCR (quantitative chromatin profiling) (Crawford et al., 2004), using active chromatin sequence (ACS) libraries (Sabo et al., 2004), or combining experimental data and computational methods to predict bound TFBSs which coincide with DHSs (Pique-Regi et al., 2011).

1.5.5.2 FAIRE

FAIRE is an alternative genome-wide approach to study chromatin accessibility and nucleosome occupancy. This technique uses formaldehyde to crosslink chromatin in vivo, followed by sonication and phenol-chloroform extraction. Recovered DNA from the aqueous phase is then labeled and hybridized to microarrays (Nagy et al., 2003). This method was successfully tested in yeast (Hogan et al., 2006) and human cells to study nucleosome-depleted regions which encompass active regulatory sites (Giresi et al., 2007, Paul et al., 2011).

1.6 NON-CODING FUNCTIONAL ELEMENTS IN THE HUMAN GENOME

The ENCODE project showed recently that 80.4% of the human genome conveys biochemical functions through systematic mapping of functional elements in the human genome and estimated that 99% of the human genome lies within 1.7 kb of these sites. Also, TFBS analysis (ChIP-seq and DNase I footprint studies)

showed that 95% of the human genome lies within 8 kb of TFBSs. Non-coding variants were at least as frequent as coding variants among ENCODE functional annotated sites. Overall, 8.5% of bases in the genome are involved either in TFBSs (4.6%) or a DHS footprint (5.7%), which is much more than protein-coding sequences and sequences subject to evolutionary constraint (Bernstein et al., 2012).

Considering DNase I sensitivity as a hallmark of genomic regulatory sites, a DHS study of 125 cell types mapped 2.89 million unique DHSs in the human genome, many of which coincided with 4.8 million sites mapped by FAIRE assay in 25 cell types. ~ 2 million DHSs per cell type (FDR 1%) have been identified, which represent 1% of the genome in each cell type (3.9% in total). Analysing TFBSs in K562 cells showed overall a 98.5% coincidence of ChIP-seq peaks and DHSs (Thurman et al., 2012). 8.4 million distinct DNase I footprints, cleavage protected genomic regions secondary to binding to regulatory factors, have been identified across 41 cell types (FDR 1%) as well as a ~50 bp promoter DNase I footprint which localizes transcription initiation (Neph et al., 2012b). About 5% (n= 135,735) of DHS were within 2.5 kb of TSSs. The other 95%, so-called distal DHSs, were evenly distributed in intronic and intergenic regions. Distal DHSs are mostly cell-type specific (Thurman et al., 2012).

1.7 SUMMARY

DHS marks chromatin accessibility at active regulatory sites and ENCODE data show that DHSs are enriched for disease/trait associated non-coding GWAS variants. The number and location of DHS varies between tissues, reflecting tissue-specific gene regulation. However knowledge regarding the extent and cause of DHS variation between individuals is scarce.

In order to study the relationship between local sequence variation, DHS variation and functional consequences, I set out to profile the DHS of eight inbred mouse strains from which complete sequence and QTL data were available. My aim was to use DNase-seq in terminally differentiated erythroblasts to quantify genome-wide the proportion of variation in DHS. I would then be able to estimate heritable DHS variation and examined the association between local sequence variants and variable DHSs. To assess the functional outcome of DHS variation, the relationship between DHS variation, gene regulation and quantitative traits phenotypic variation would then be assessed.

CHAPTER 2 : METHODS

In this chapter, I describe the materials and methods of my project.

2.1 PRIMARY CELLS

I performed DNase I assays on harvested primary cells (terminally differentiated erythroblasts) and fibroblasts primary cultured cells (mouse embryonic fibroblasts (MEFs)).

2.1.1 TERMINALLY DIFFERENTIATED ERYTHROBLASTS

Mature primary erythroid cells were obtained from phenylhydrazine-treated adult (6-9 weeks old) female mice as described (Spivak et al., 1973). Mice for this experiment were the progenitors of the heterogeneous stock (HS) (A/J [JAX #646], AKR/J [JAX #648], BALB/cJ [JAX #651], CBA/J [JAX #656], C3H/HeJ [JAX #659], C57BL/6J [JAX#664], DBA/2J [JAX #671] and LP/J [JAX #676]). TER119-positive cells were magnetically sorted, after indirect labeling from the aliquot of the samples.

2.1.2 C57BL/6J MOUSE EMBRYONIC FIBROBLASTS (MEFS)

Dr Jacqueline Sloane-Stanley from Prof Higgs group prepared 1st and 2nd batches of C57BL6/J MEFs. 3rd batch was prepared from C57BL/6J MEFs and stored at -80°C freezer by Dr Ben Davies group. Standard protocols used for revitalizing and culture of frozen MEFs to generate adequate number of cells for DNaseI assay (min 5 x 10^7 Cells) in defined number of passages. Triple 175 ml flasks were used for passage IV.

2.2 DNASE-SEQ

2.2.1 DNASE I ASSAY

DNase I assay was performed on ~ 5 x10^7- 1 x 10^8 cells. Cells from mainly terminally differentiated erythroblast (~60-80% cells) were pellet down and

washed by ice-cold mouse PBS (0.9 M). A pre-lysis aliquot was taken as a baseline control for the assessment of nuclei release. Then, nuclei were released by resuspending pellet in 1ml lysis buffer in strictly one minute (for solution recipes see Appendix 1). Nuclei were pellet down and washed by ice-cold RSB at 4°C centrifuge. A post-lysis aliquot was taken. Pre-lysis and post-lysis aliquots were observed under the microscope for assessment of nuclei release. Figure 2-1 shows examples of what observed at pre-lysis and post-lysis samples. If < 80% of nuclei are released the lysis step needs to be repeated.

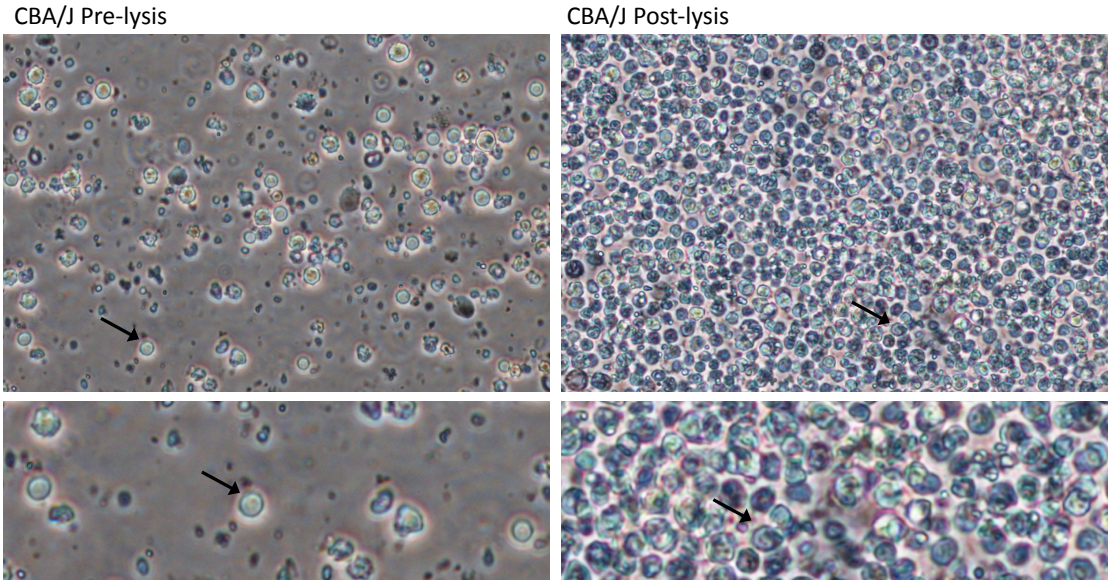


Figure 2-1 – Pre and post-lysis aliquots of cells harvested from phenylhydrazine treated CBA/J. Right panel shows the image of pre-lysis cells harvested from spleen of a phenylhydrazine treated CBA/J under the microscope. The arrow points a terminally differentiated erythroblast (the dominant population), which has bright cytoplasm around central nucleus. Left panel the same cells after lysis buffer exposure. Nuclei are released in majority of cells. Bright cytoplasm is not visible and nuclei look enlarged and bizarre. Panels at the bottom are magnified images from the top panels for better visibility and arrows point same locations in top and bottom panels.

Released nuclei were digested with eight concentrations of DNase I (0_{ice}, 0_{37}, 0.5, 1, 2, 4, 8, 16, 32, 64 µl of 10 U/µl DNase I in 1 ml reaction volume) at 37°C (except 0_{ice}) until the sample containing 16 µl reached water-like viscosity and

the digestion stopped instantly by transferring the tubes to ice box following by transferring the reactions to lysis buffer-containing tubes. DNAs were extracted by phenol/chloroform-extraction, and ethanol-precipitation. Phases of DNase I digestion were selected for library preparation by qPCR. Solexa compatible libraries were constructed and sequenced by Illumina Genome Analyzer (GA-IIx) and HiSeq2000.

Figure 2-2 illustrates DNase-seq library preparation steps, which is slightly modified version of the standard protocol (Quail et al., 2008).

2.2.2 DNASE-SEQ LIBRARY PREPARATION STEPS

- Blunt end DNaseI digested DNA

1.5-2 μg of each DNase I digested DNA was diluted from different digestions (usually the mid phase of DNase I (total 5-6 μg)) in 20 μl of Sigma water. Then, 20 μl of NEB2 buffer, 5 μl of 10 mM dNTPs, 2.4 μl of T4 DNA Pol, 2 μl of 100x BSA, and 150.6 μl of Sigma water added to the reaction and it was incubated 12°C for 1 hour. This step was followed by Phenol/Chloroform purification and Ethanol precipitation and DNA pellets were resuspended in 50 μl of Sigma water.

- Klenow reaction adding "A" bases to the 3' end of the DNA fragments

1mM dATP 10 μl, 5 μl of 10x Klenow Buffer (Promega M195A), 1 μl Promega DNA Polymerase I Large (Klenow) Fragment Exonuclease Minus (M218A), 9 μl of Sigma water were added to 25 μl of blunted DNA and it was incubated 30 min at 37°C. This step was followed by QIAGEN PCR purification and DNA resuspension in 30 μl elution buffer (EB) provided by the kit.

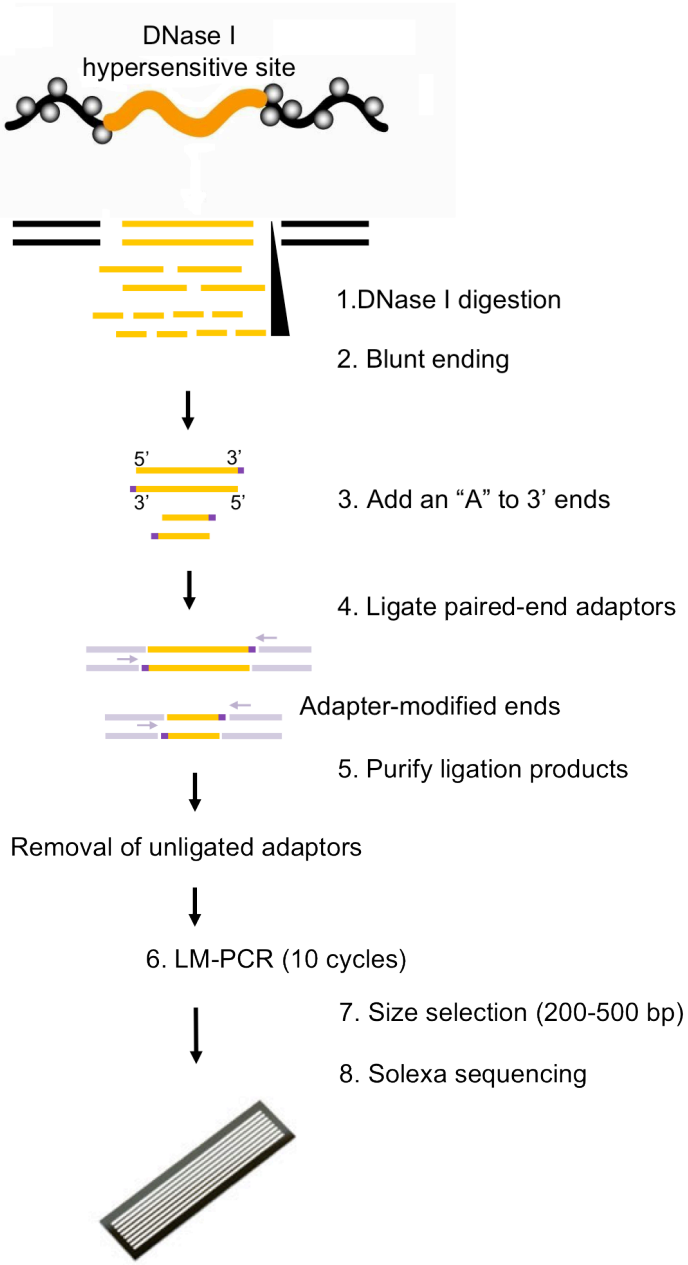


Figure 2-2 – DNase-seq library preparation steps.

- Ligation Solexa adapters

22.5 μl of Sigma water, 10 μl of 5x Ligase Buffer Invitrogen, 2 μl of 100 μM PE adapter oligo mix, 0.5 μl of T4 Ligase were added to 15 μl of blunted/"A" added DNA and was incubated overnight at 16°C. This step was followed by QIAGEN PCR purification and DNA elution in 30 μl EB.

- PCR

26.5 μl of Sigma water, 10 μl of 5x Phusion HF Buffer (Finnzymes NEB), 1 μl of 25 μM Primer "PE1", 1 μl of 25 μM Primer "PE2", 1 μl of 10 mM dNTPs, 0.5 μl of Phusion Hot Start High Fidelity Taq (Finnzymes NEB F-540S) were added to 10 μl of blunted/"A" added/adapters ligated DNA to prepare PCR reactions.

- PCR protocol: A. 98^C for 30 sec

  B. 10 cycles of:

  98°C for 10 sec

  65°C for 30 sec

  72°C for 15 sec

  C. 72°C for 5 min

  D. Hold at 4°C

- Electrophoresed on 2% Low Range Ultra agarose gel (#161-3107) in TAE buffer and size select (200-500 bp) to avoid primers and adapters (Figure 2-3). This step was followed by the instructions of QIAquick Gel Extraction Kit to purify on QIAquick columns and DNA elution in 30 μl EB.

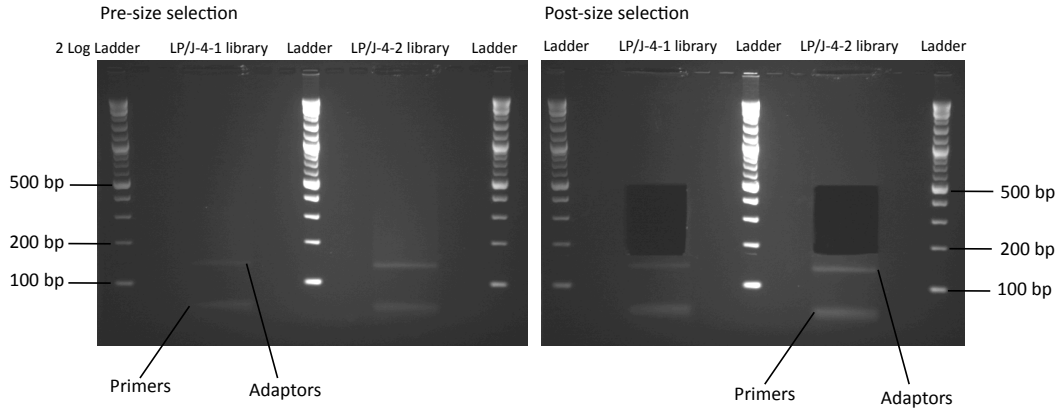


Figure 2-3 – Size selection step of DNase-seq library preparation. Two PCRs were performed per sample. It is critically important to avoid contaminations. Electrophoresis tank should be carefully cleaned by 10% bleach and rinse carefully. There should be an empty lane between library and ladder lane. A sterile scalpel used per cut to avoid contamination. A control picture obtained to control avoidance of primers and adaptors post-size selection.

2.2.3 QUANTIFICATION OF DNASE I DIGESTION BY QPCR

Real-time PCR was used to choose similarly digested material in different samples. Three pairs of forward and reverse primers were designed to amplify ~ 140-160bp products by using the program Primer3 (Rozen and Skaletsky, 2000) for murine positive (MCS-R2, MCS-R3, minus14) and negative (Control-1, Control-2, Control-3) DHSs and their efficiency, uniqueness and whether or not they contain any SNP in different 8 strains were interrogated. MCS-R2 and MCS-R3 are two of the regulatory elements of the alpha globin genes (Hughes et al., 2005). Minus14 amplifies a regulatory element that controls alpha globin gene expression. Controls one and two are regions in the alpha globin locus which appeared inactive using H3K4me2 ChIP on chip. Control-3 measures the *Pkdrej* gene, which is sperm specific and is very condensed in all tissues. Among these, the best positive and negative set of primers were chosen for qPCR based on similar efficiency for test and control amplification and specificity of primers due

to melting curve analysis and running the end products on the gel. The standard curves were used to calculate the percentage of copies of the MCS-R3 amplicon remaining in 5 ng of DNase I-treated genomic DNA in different phases of digestion (different DNase I concentrations).

The amount of template DNA was standardized by correcting for amplification of the DNase I-insensitive Control-2 or Control-3 sequence (McArthur et al., 2001). Among 3 replicates of DNase I digestion per strains, mid-phase of digestion of the sample, which presented more digestion with increasing amount of DNase I, was selected for GAII library preparation. PCR was performed by using SYBR green PCR iQTM SYBR Green Supermix (BioRad), and real-time PCR was performed on iCycler iQTM real-time PCR detection system (BioRad). Genomic DNA was quantified three-times by using the NanoDrop 2000 Spectrophotometer (Thermoscientific) and a PicoGreen dsDNA Assay kit (Invitrogen). Five ng of genomic DNA digested with eight different concentrations of DNase I were loaded onto 96-well optical PCR plates by using a iCycler machine. PCRs performed in triplicate generated highly reproducible results (SD 0.25) and outliers were excluded from the analysis.

2.2.3.1 Primer design for MEF SYBRgreen qPCR

Primer3 free software and SYBRgreen real time PCR (qPCR) primers design guidelines recently published (Thornton and Basu, 2011) used for designing 3 primer pairs for controls 1-3 as C57BL/6J MEFs DHS negative sites and 2 primer pairs for two housekeeping genes (beta-actin (*Actb*) and Glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*)) as C57BL/6J MEFs DHS positive sites. I aligned the initial C57BL/6J DNase-seq result from GAII platform with the sequence of these sites to optimize our primer design by downsizing the source

sequence of the DHS positive sites to the area under DHS peak at *Actb* and *Gapdh* loci (indicating the existence of regulatory elements of these sites) and confirming the absence of peaks in choosing the source sequence for designing DHS negative controls.

First the following criteria used for primer design by primer3:

- Primer length: 20-28 bp (optimal = 25 bp)
- 60 < primer Tm < 70 ° C (optimal = 64 ° C), maximum Tm difference between forward and reverse primer = 2 ° C and it should be higher than Tm of any estimated template secondary structures
- Annealing temperature = 59 – 60 ° C
- Product size = 80-150 bp (optimal = 120 bp)
- 50 < CG% < 60
- Avoid repeats (maximum dinucleotide repeats = 4)
- Avoid >4 single base runs
- Avoid >2 GC clamps at 5' end
- Avoid 3' complementary to prevent primer-dimer formation
- Concentration of divalent cations set at 3.5 mM (adjusted for MgCl in SYBR green supermix)
- Concentration of dNTPs = 0.2 mM

Then, the existence of primer secondary structure (self dimer, cross dimer and hairpins) and amplicons' secondary structures were verified by Beacon Designer TM free edition (http://free.premierbiosoft.com/) and free online Integrated DNA Technologies (IDT) mFold software (http://eu.idtdna.com/UNAFold?) respectively. And finally specificity of primer hybridization was assessed by alignment tool (BLAST), and LookSeq used for SNP detection.

2.2.4 DNASESEQ ANALYSIS

2.2.4.1 Sequence Alignment

In order to mitigate the impact of inter-strain sequence variation on mapping quality we first created pseudo genomes for each of the eight strains by introducing (Keane et al., 2011) single nucleotide polymorphisms from the relevant strain into the MGSC37 mouse reference. We then aligned reads from each lane to the appropriate mouse strain genome using STAMPY (Lunter and Goodson, 2011). Purpose-written software was used to calculate the genomic coverage for DNA fragments represented by each read pair. In particular, for short fragments where the two reads of a pair overlap, sequencing primers were removed and bases of the overlapping region were only counted once. Conversely, the coverage included the entire aligned DNA insert and not merely two ends sequenced in the read pair. Read map alignments were filtered for quality, removing reads with a mapping quality values < 2. Alignments to repetitive parts of the genome were filtered by excluding genomic regions with a CRG 50-mer Alignability score of < 0.5 (Derrien et al., 2012). These are duplicate regions that are present in more than two (near) identical copies in the genome.

2.2.4.2 Peak detection and analysis

We called DNase hypersensitive sites using purpose-written software. Let X_j be the count of aligned reads covering position *j* in the genome. To allow for local variation in the coverage of aligned reads, we computed the deviation *dj* of the overall mean coverage from the local mean coverage (in a window of width 2L+1) and adjusted *Xj* to *Wj = Xj-dj*. This maintains the same average read coverage, μ, where μ is the overall mean coverage. A small number of sites where the adjusted values were negative were truncated to 0. Then we called peaks on

the transformed scores $Yj = log_2(Wj/p\mu)$ where the parameter p controls the stringency of the peak detection. We called peaks by searching for high-scoring segments (Karlin and Altschul, 1990), defined as intervals [m,n] such that $S_{mn} = \sum_{m<j<n} Yj$, satisfies the conditions $S_{mn} > 0$ and $S_{mn} > S_{uv}$ where [u,v] is any interval containing [m,n]. Yj is negative whenever $Xj\text{-}dj < p\mu$, so we only call peaks in regions where adjusted counts are at least p times the mean coverage. The motivation for the definition of Y is first that Y is effectively independent of the average coverage (so that we can compare data with different read coverage, e.g. from an Illumina GA-II and a HiSeq2000) and second that peaks with extremely high values (e.g. to due to unresolved repeats being mapped to the same locus) carry relatively little additional weight.

Variation between peaks was estimated under the assumption that the distribution of reads at a peak can be modeled as a negative binomial distribution. We used the DESeq package to estimate the differences between peaks in different strains (Anders and Huber, 2010). Correlation between peaks and sequence variation was performed in linear models using the statistical software package R. The strain distribution pattern of a variant was expressed as a vector, so that for example a sequence variant present only in strains A/J and LP/J would be the vector ABBBBBBA (where strains are in the order A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J, LP/J). We asked if the same vector is present in the pattern of strain variation of DHS peaks.

The degree of variation was estimated by the heritability of the site, defined as half the between strain component of variance divided by the sum of the within strains variance and half the between strain variance (Hegmann and Possidente, 1981). Although the distribution of peak heights violates some

assumptions behind this calculation, the heritability provides an intuitive measure of the extent of variation.

Peak coordinates were associated with the following genomic features: sequence variants identified in the eight inbred progenitor strains of the HS (Keane et al., 2011, Yalcin et al., 2011); vertebrate sequence alignments taken from the multiz30way track of the UCSC genome browser (Blanchette et al., 2004); regions enriched for H3K4Me1 (enhancer feature) and H3K4me3 enrichment (promoter feature) (Kowalczyk et al., 2012); transcription factor binding motifs from TRANSFAC database identified by running FIMO (Bailey et al., 2009) with a detection threshold of $P < 10^{-5}$ (FIMO was run on the MGSC37 mouse reference and on sequences for each strain by introducing single nucleotide polymorphisms into the reference sequence (as described above). Statistical analyses were carried out using the R software package (R-Development-Core-Team, 2004).

2.3 RNASEQ

2.3.1 RNA SAMPLE PREPARATION

Total RNAs extracted from phenylhydrazine treated mouse spleens of given strains. RNA Integrity Number (RIN) measured by Agilent 2100 Bioanalyzer. Minimum of RIN=9 considered as a cut-off for proceeding samples. Ter119+ cell measured from an aliquot obtained from the spleens. TURBO DNA-free™ Kit used for decontaminating RNA preparations from DNA. Since globin in highly expressed in erythroid cells, globin depletion performed to avoid masking effect of globin on gene expression study by GLOBINclear™ Kit (#AM1981). Total RNA was split into poly(A)+ and poly(A)- RNA using mRNA isolation performed using PolyATtract® mRNA Isolation Systems (#Z5300). Poly(A)+ RNA libraries were

generated using the Illumina mRNA-Seq paired-end kit after globin depletion using GlobinClear (Ambion). Poly(A)- RNA libraries were generated using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre), after depletion of ribosomal transcripts with RiboMinus Eukaryote Kit for RNA sequencing (Invitrogen). Poly(A)+ and poly(A)- libraries were sequenced on Illumina HiSeq2000 and GAIIx respectively. Minimum amount of 10 μg total RNA or 100-200 ng globin-depleted mRNA is required for HiSeq Solexa sequencing. Dr Monika Kowalczyk did RNA samples preparations.

2.3.2 RNASEQ ANALYSIS

Read map alignments were filtered for quality using samtools version 0.1.18 (Li et al., 2009) and reads with a MAPQ value < 15 were removed. Sequence reads that overlap with transcript intervals were counted using the R package IRanges and a table of transcript and exon coordinates provided by UCSC, based on the mouse reference genome assembly mm9. Reads were mapped to n = 26347 (out of 37681) transcripts. Normalization of read numbers was carried out using DESeq version 1.8.3 (Anders and Huber, 2010). Transfrags were assembled de-novo using Cufflinks version 1.3.0 (Trapnell et al., 2010) (n = 74390 transfrags) and these transfrags were linked to the reference genome using cuffcompare. Differential expression of transcripts in the strains used was analyzed using cuffdiff.

CHAPTER 3 : DNASESEQ LIBRARY PREPARATION

3.1 INTRODUCTION

In this chapter I describe the work performed to establish a reliable high throughput sequence based analysis of DNase I hypersensitive sites. Conventional methods to identify DHSs (e.g. Southern blotting) are not scalable to genome-wide studies. DNase-chip assays have been developed as a potential method to identify DHSs in high resolution across the genome (Crawford et al., 2006a). I tested whether it was possible to generate DNase I digested DNA suitable for Solexa sequencing. A DNase-chip assay has been already shown to effectively identify the DHSs previously mapped at the alpha and beta globin locus by Southern blot strategies (Figure 3-1). I aimed to modify this method to produce Solexa sequencing compatible DNase I digested material.

3.2 DEVELOPING DNASEI-SEQ PROTOCOL BY MODIFYING DNASE-CHIP ASSAY IN K562 CELLS

The DNase-seq protocol builds on traditional Southern blot technique to identify regulatory elements in intact chromatin. I modified a DNase-chip assay and compared the results obtained from K562 human erythroleukemic cell line (Andersson et al., 1979) to the DHSs already identified in this cell type in two well characterized human loci (alpha and beta globin). The nuclei from K562 cells were exposed to a range of different concentrations of DNase I. The samples from the mid-phase of digestion were blunted, ligated to linkers and amplified by 30 cycles ligation mediated PCR (LM-PCR). This DNA library was labelled and co-hybridized with a control genomic DNA to a custom alpha and beta globin tiled microarray (De Gobbi et al., 2007) (Figure3-1A). The procedure was then

modified to generate DNase I-digested DNA library compatible with Solexa sequencing. In order to obtain this it was required to use Solexa specific adaptors and primers, and to decrease the number of PCR cycles from 30 to 10 (Figure3-1B). Finally, the DNA library was labeled and hybridized on the same custom microarray.

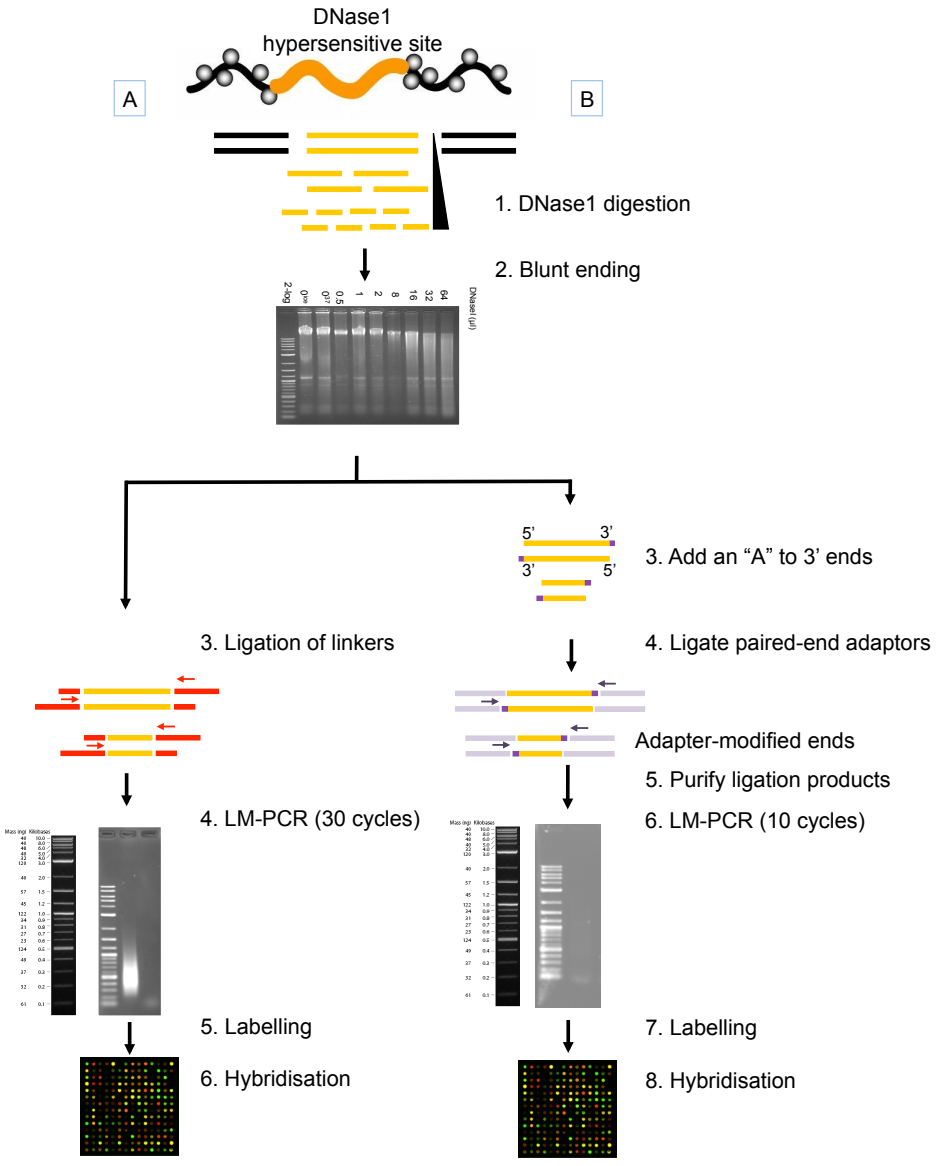


Figure 3-1 - Outline of DNase-chip method and its modified version. (A) K562 cells were digested with different concentrations of DNase1 and the mid-phase of digestion was used for blunting, ligation, LM-PCR amplification and hybridization on microarray. (B) Solexa specific linkers and primers were used and the number of PCR cycles reduced to 10.

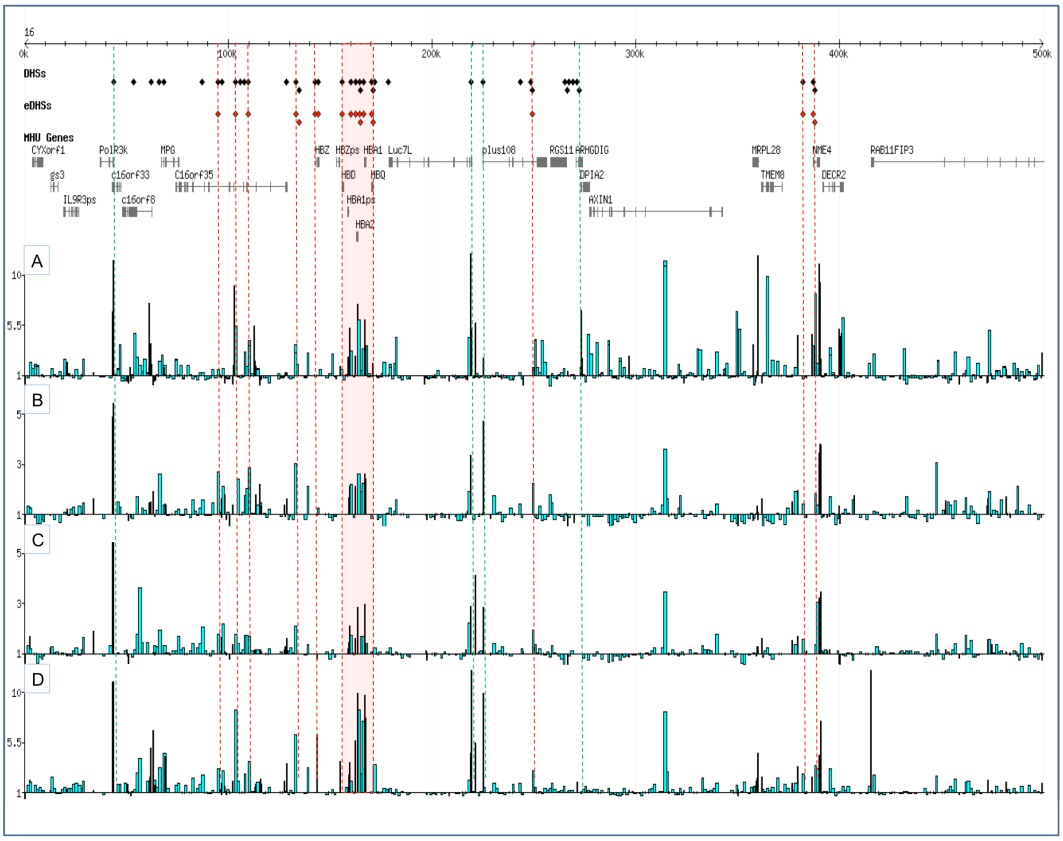


Figure 3-2 - Comparison of DHSs mapping at the telomeric region of human chromosome 16p in K562 cells. The position of the constitutive (DHSs, black diamonds) and erythroid-specific DNase1 hypersensitive sites (eDHSs, red diamonds), and the genes located in the telomeric region of chromosome 16p are shown at the top. Below this, the tracks from the different DNase-chip experiments are shown. The Y axis represents the enrichment of DNase I digested DNA over input DNA from one experiment i.e. the signal from the DNase assay is normalized against the signal from input DNA (genomic DNA from K562 cell line) by array express software. The DNase material was labelled in cy3 and the input in cy5 to test the efficiency of each probe to interpret enrichment. The signal below the line indicates that region is depleted in signal in DNase track relative to genomic input i.e. more signal was observed in input than the DNase enrich sample. The X axis represents the alpha globin locus. Green and red dashed vertical lines run from DHSs and eDHSs through the array data respectively. Panel A: old K562 DNase-chip (LM-PCR, 30 cycles). Panel B: new K562 DNase-chip (LM-PCR, 30 cycles). Panel C: new K562 DNase-chip (LM-PCR, 10 cycles). Panel D: new K562 DNase-chip (Solexa PCR, 10 cycles).

The results from the standard DNase-chip assay reproduced data previously obtained by the same technique and by Southern blotting (Figure 3-2 A and B).

In the standard protocol, the reduction of the number of cycles from 30 to 10 did not affect the DHSs signals on the array (Figure3-2 C) and the use of the Solexa

modified protocol, as shown in Figure 3-2, did not interfere with the detection of DHSs on the microarray (Figure 3-2 D).

In conclusion, using Solexa-sequencing specific adaptors and reducing the number of LM-PCR cycles, we obtained results on DNase-chip similar to the standard protocol. This suggests that this new DNase I digested DNA library is compatible with high throughput sequencing for genome-wide DHSs detection.

Next, I performed DNase I assay on primary mouse cells of the same tissue at similar developmental stage. To perform DNase I assays, I digested nuclei of terminally differentiated primary erythroblasts (majorly TER119-positive cells) harvested from the spleens of the phenylhydrazine-treated mice with 8 different concentrations of DNaseI (Figure 3-3).

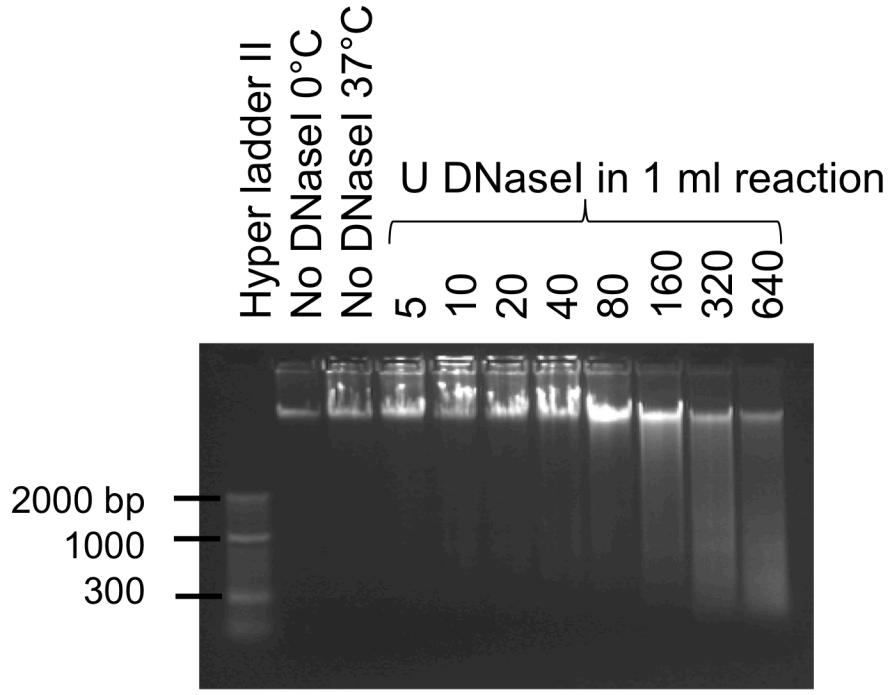


Figure 3-3 - Gel electrophoresis of DNaseI-digested DNA. 8 different concentrations of DNase I used are labelled by the volume of 10 U/µl DNaseI used per 1 ml of reaction volume (0.5, 1, 2, 4, 8, 16, 32, 64 µl).

I performed all DNase I digestions on at least three biological replicates. Since valid DNase I hypersensitive sites should be detected in more than one concentration of DNaseI by Southern blot and DNase-chip, and not all the DNase I hypersensitive sites are equally sensitive (Crawford et al., 2006a), I selected the combination of four different phases of digestion based on the qPCR results to generate the DNase I-seq Solexa compatible libraries.

3.2.1 SORTED VS. UNSORTED CELLS

Mature primary erythroid cells were obtained from phenylhydrazine-treated adult mice as described (see Methods). I performed a DNase I assay on sorted (TER119+) vs. unsorted cells harvested from the anemic spleens of phenylhydrazine-treated C57BL/6J mice. Figure 3-4 shows that the assay did not

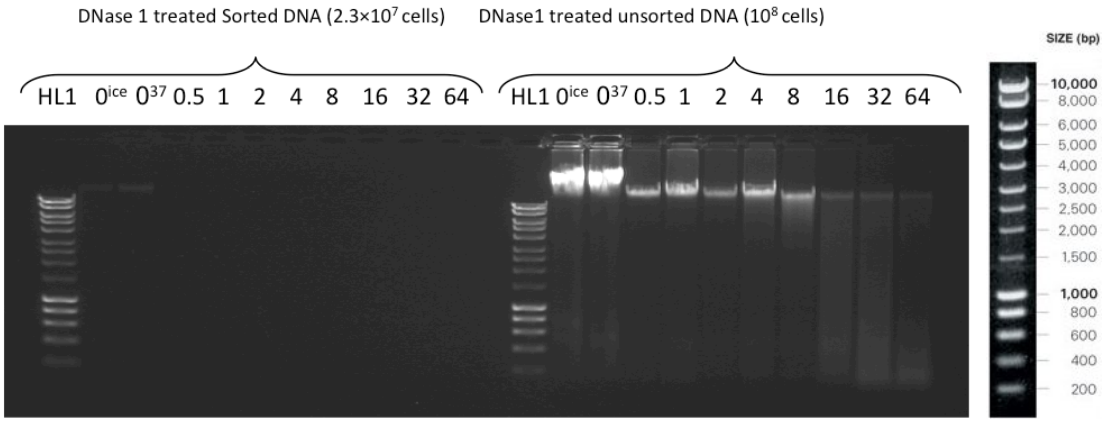


Figure 3-4 - DNase I assay on TER119-positive cells vs. non-selected cells from spleens of phenylhydrazine treated C57BL/6J mice.

work on selected TER119-positive cells. Based on this finding I decided to use unsorted cells. However, TER119-positive cells were counted to give an idea how

the spleen responded to drug-induced hemolytic anemia. Table 3-1 summarizes the frequencies of TER119-positive cells in replicates of given strains.

Table 3-1 - Descriptive data of the replicates of the eight inbred strains used for DNase I assay.

Strain replicates		Age (weeks)		No. of cells for DNase assay	Frequency of TER119+ cells %
		Injection	Harvesting		
A/J	1	7	8	1×10^8	66
	2	7	8	1×10^8	57
	3	7	8	7×10^7	60
AKR/J	1	8	9	9×10^7	70
	2	8	9	9×10^7	73
	3	8	9	9×10^7	74
BALBc/J	1	6	7	8×10^7	72
	2	6	7	7×10^7	74
	3	6	7	7.5×10^7	76
C3H/HeJ	1	6	7	1×10^8	78
	2	6	7	1×10^8	67
	3	6	7	1.2×10^8	77
C57BL/6J	1	8	9	9×10^7	63
	2	8	9	1.2×10^8	63
	3	8	9	1.3×10^8	65
CBA/J	1	7	8	1×10^8	62
	2	7	8	9×10^7	62
	3	7	8	9×10^7	57
	4	6	7	1×10^8	82
	5	6	7	9×10^7	79
	6	6	7	9.5×10^7	82
DBA/2J	1	6	7	1×10^8	71
	2	6	7	1.3×10^8	74
	3	6	7	1.4×10^8	76
LP/J	1	7	8	8×10^7	63
	2	9	10	6.5×10^7	55
	3	9	10	8×10^7	71
	4	6	7	9.5×10^7	80
	5	6	7	8.5×10^7	82
	6	6	7	6×10^7	82
Mean ± SD		7 ± 1	8 ± 1	$(9.4 \pm 1.9) \times 10^7$	70 ± 8

3.2.2 DNASE-CHIP ON PRIMARY MOUSE ERYTHROID CELLS

Solexa compatible libraries were prepared from different pooled phases of DNase I assay on unsorted primary mouse erythroid cells (C57BL/6J)(Figure 3-5). The quality of these libraries initially tested by a previously described hybridization array, which designed by Professor Higgs group to interrogate the mouse alpha globin locus. DNase I digested DNA hybridized to this array and I expected to detect a specific enrichment of signal at known hypersensitive sites.

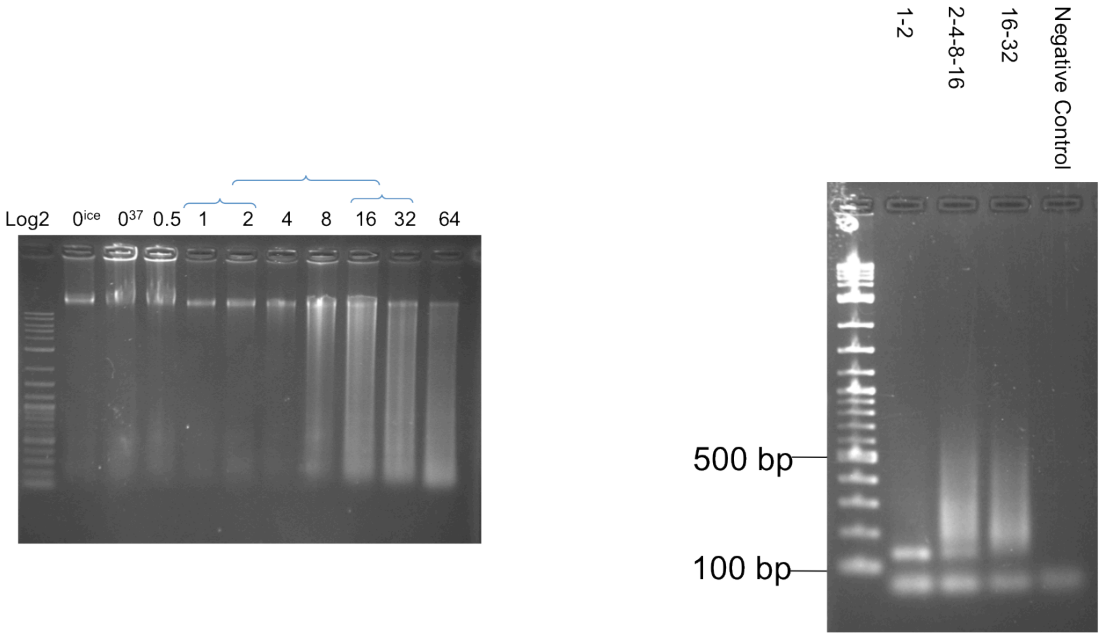


Figure 3-5 - Library preparation from mouse primary erythroid cells. We pooled the reaction products of DNase I assay from 1 and 2, 2 to 16 and 16 and 32. The DNA has been blunt ended and amplified in a PCR. For the next steps of the experiment we need DNA in the 250-550bp range to avoid contamination with Solexa primers and oligos.

Figure 3-6 shows a snapshot of DNase-chip result in alpha globin locus. These peaks coincide where the known DNase hypersensitivity sites reside. However, this custom tilled microarray only enabled me to interrogate a small part of the genome in our analysis. By DNase-seq I can theoretically perform a genome-wide analysis.

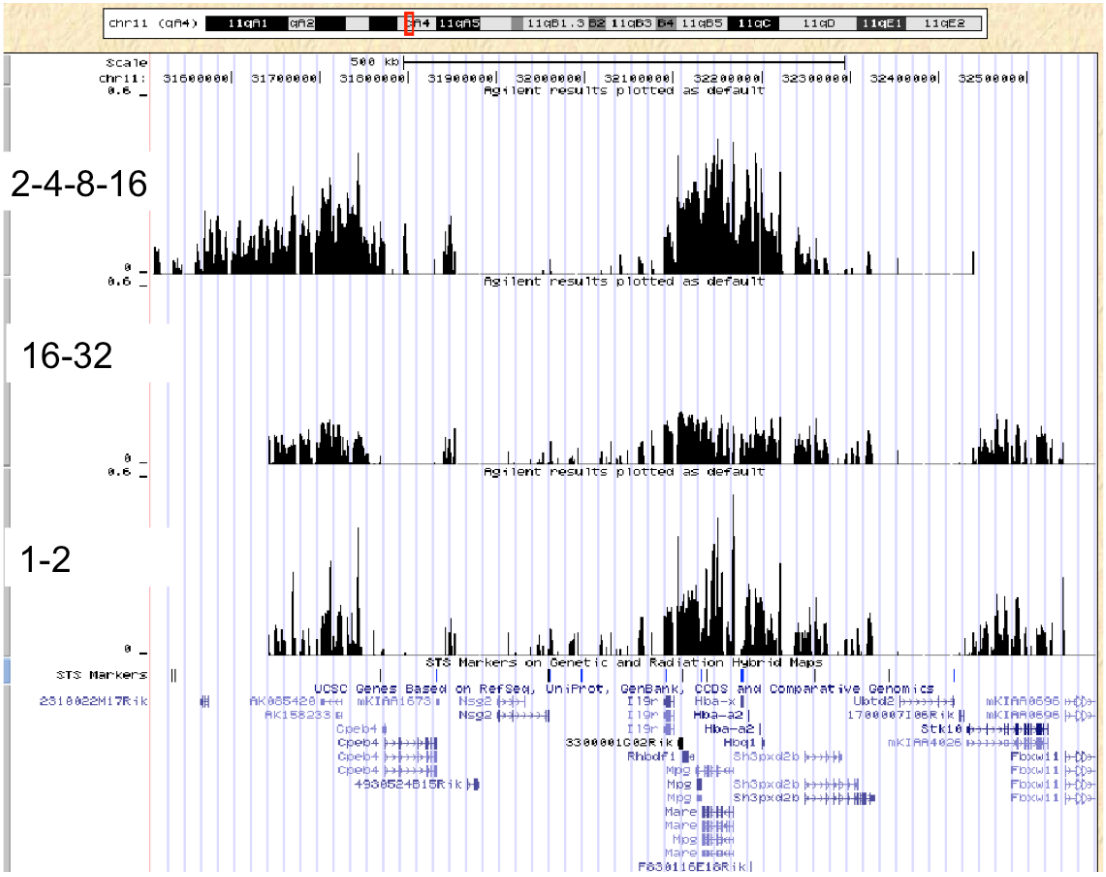


Figure 3-6 - a snapshot of DNase-chip results at mouse alpha globin locus on chromosome 11. Libraries from DNase assay pooled reaction phases (low-digestion: 1,2, mid phase of digestion: 2,3,8,16 and high-digestion: 16,32) hybridized on a custom Agilent microarray. Coinciding enrichment signals and known DHS at alpha globin locus suggest good quality of library preparation. It also highlights the importance of choosing a comparable phase of DNase I assay for library preparation as different phases of DNase assay might capture various DHSs with various sensitivity and specificity.

3.3 QUANTIFICATION OF DNASE I DIGESTION BY QPCR

My goal was to select comparable samples in which adequate digestion had occurred to reveal DNase I hypersensitive sites, but which were not over-digested. Unfortunately it proved difficult to calibrate digestion either by the amount of enzyme or the time of digestion. Instead, I used real-time quantitative PCR (qPCR) based methods to interrogate DNase I sensitivity at specific genomic

positions. To monitor DNase I digestion quantitatively, and to select a sample suitable for construction of Solexa libraries, I analyzed several aliquots from the same nuclear preparation by qPCR, I quantified the degree of digestion by comparison with known erythroblast DNase I hypersensitive sites on mouse chromosome 11, around the alpha globin locus (De Gobbi et al., 2006).

First I designed three sets of forward and three sets of reverse primers for positive (R2, R3 and minus14) and negative (C1- C3) erythroid specific DHSs (see methods). Similar melting temperature (tm) (~60 ° C) and GC content (~ 50%) of forward and reverse primers were considered in primer designing with Primer3.

Appendix 2 summarizes the primers details and the result of in silico primer specificity testing and SNP detection. UCSC BLAT (http://genome.ucsc .edu/cgi-bin/hgBlat?command=start) was used to check the specific aligning of primers to the mouse genome. LookSeq (http://mus.well.ox.ac.uk/mus/) was used to check if the forward and reverse primers include any SNP in eight given strains that might affect annealing of primers to the template and efficiency of primers. C1 reverse primer does not uniquely align control 1 site at chromosome 11 it also anneals to plus strand of chromosome 8. C2 forward primer contains a SNP in AKR/J (A base replaces C in other strains). Therefore, designed C1 primer set was predicted not to be an ideal primer set. C2 primer set in AKR/J also might not be an ideal DHS negative primer set and an alternative preferably should be chosen.

The efficiency of primers initially tested with a modified Touchdown PCR using HotStarTaq DNA polymerase (Qiagen) on 6 out of eight genomic DNA from given strains (see Appendix 3). PCR products were run on 3% electrophoresis gel to test specificity of primers (Figure 3-7). Single bands at expected size (~ 150 bp) indicated the appropriate function of primers.

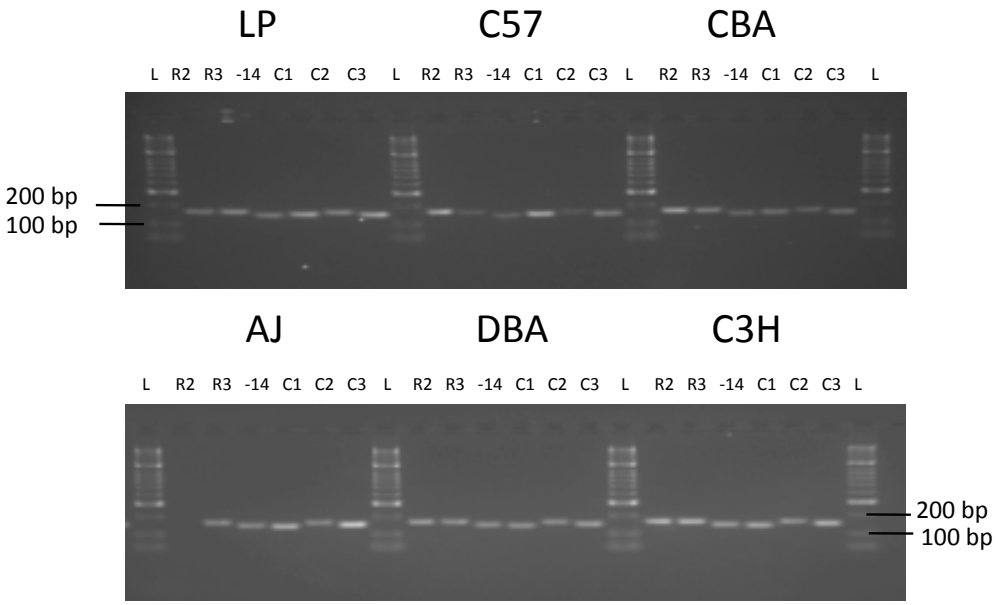


Figure 3-7 - Primers were tested on genomic DNA of six inbred mouse strains. Single bands at expected size (~ 150 bp) were observed. An empty lane at A/J R2 is due to faulty gel loading.

3.3.1 QPCR FOLD CHANGE CALCULATION

The amplification kinetic of DNase I sensitive vs. control templates was studied by comparing amplification (Ct values) in three DNaseI assays (C57BL/6J-1, C3H/HeJ-1, and A/J-1). dCt, ddCt, and fold change measured based on the following definitions and formula:

A1= endogenous control (i.e. insensitive site to DNase I digestion) in our experiments includes: C1, C2, and C3

A2= test PCR (i.e. DNase I sensitive site) in our experiments includes: R2, R3, and minus14

S0= untreated samples for both A1 and A2 (here 0ice samples of the first replicates of the following strains: C57BL/6J-1-0ice, C3H/HeJ-1-0ice, and A/J-1-0ice)

S1= DNase I treated samples (here No. 1 of DNase I assay of first replicates of the following strains: C57BL/6J-1-1, C3H/HeJ-1-1, and A/J-1-1)

$dCt_{S0} = Ct_{A2} - Ct_{A1}$ of untreated sample

$dCt_{S1} = Ct_{A2} - Ct_{A1}$ of DNase I treated sample

$ddCt = dCt_{S1} - dCt_{S0}$

If 2^ -ddCt > 1 then fold change = 2^-ddCt

If 2^ -ddCt < 1 then fold change = -1/2^-ddCt

Every sample measured in triplicates. Outliers (which drove up Ct SD > 0.25) were excluded from the analysis. Since C1 was not a reliable control (failed unique alignment to the genome) was excluded from fold change calculations. Fold changes were measured twice (considering C2 and C3 as an endogenous control (A1)). Figure 3-8 summarizes the dCt, ddCt, 2^-ddCt and fold change values. Fold change > 1.5 or < -1.5 are indicative of real change. 5 ng/μl of all templates were loaded on the PCR plate. As the abundance of DNase I sensitive templates (A2) relative to insensitive sites (A1) is expected to be lower in treated sample (S1) in compare with untreated samples (S0), I expected to observe fold change < - 1.5. Fold change results were overall in the expected direction but only in few cases $\sim \leq 1.5$. The explanation could be that the efficiency of primer sets for tests and controls are not comparable or the efficiencies of the primers are not within desirable range. Therefore, for selecting best control and test

primer sets to assess DNase I digestion, I performed detailed experiments to test primers efficiencies. Figures 3-9 - 3-15 and Appendix 2 show the analysis of the specificity and efficiency of primers of real-time PCR with SYBR Green.

To assess DNase I digestion in different samples, I quantified abundance of DNA templates with different primers sets for controls (insensitive sites to DNase I digestion) and tests (sensitive sites to DNase I digestion). It is critical to fulfill the assumption of qPCR to quantify template abundance, which is the similarity between control and test primer efficiencies and to be in acceptable range (90-110 %). Efficiency of 100% means the amount of template doubles per cycle.

A1	C1	Ct Mean	SD										
	C3H-1-0ice	26.44	0.21										
	C3H-1-1	26.26	0.11										
	AJ-1-0ice	25.27	0.22										
	AJ-1-1	25.47	0.15										
	C57-1-0ice	25.39	0.21										
	C57-1-1	25.45	0.16										
	C2												
	C3H-1-0ice	26.53	0.15										
	C3H-1-1	26.27	0.25*										
	AJ-1-0ice	25.34	0.12										
	AJ-1-1	25.15	0.12*										
	C57-1-0ice	25.4	0.21										
	C57-1-1	25.57	0.20										
	C3												
	C3H-1-0ice	25.83	0.12										
	C3H-1-1	26.01	0.22										
	AJ-1-0ice	25.17	0.24										
	AJ-1-1	24.86	0.04*										
	C57-1-0ice	25.08	0.07*										
	C57-1-1	25.03	0.14										
				C2 as A1					C3 as A1				
A2	R2	Ct Mean	SD	dCt	ddCt	2^-ddCt	Fold change		dCt	ddCt	2^-ddCt	Fold change	
	C3H-1-0ice	26.32	0.23	-0.21					0.49				
	C3H-1-1	26.48	0.21	0.21	0.42	0.75	-1.34		0.47	-0.02	1.01	1.01	
	AJ-1-0ice	24.87	0.05	-0.47					-0.3				
	AJ-1-1	25.09	0.04	-0.06	0.41	0.75	-1.33		0.23	0.53	0.69	-1.44	
	C57-1-0ice	24.79	0.15	-0.61					-0.29				
	C57-1-1	25.26	0.06	-0.31	0.3	0.81	-1.23		0.23	0.52	0.70	-1.43	
	R3												
	C3H-1-0ice	25.99	0.04	-0.54					0.16				
	C3H-1-1	26.26	0.00*	-0.01	0.53	0.69	-1.44		0.25	0.09	0.94	-1.06	
	AJ-1-0ice	24.73	0.11	-0.61					-0.44				
	AJ-1-1	24.95	0.19	-0.2	0.41	0.75	-1.33		0.09	0.53	0.69	-1.44	
	C57-1-0ice	24.87	0.15	-0.53					-0.21				
	C57-1-1	25.12	0.18	-0.45	0.08	0.95	-1.06		0.09	0.3	0.81	-1.23	
	minus14												
	C3H-1-0ice	27.96	0.06*	1.43					2.13				
	C3H-1-1	28.44	0.23	2.17	0.74	0.60	-1.67		2.43	0.3	0.81	-1.23	
	AJ-1-0ice	26.59	0.15*	1.25					1.42				
	AJ-1-1	27.29	0.21*	2.14	0.89	0.54	-1.85		2.43	1.01	0.50	-2.01	
	C57-1-0ice	27.69	0.22*	2.29					2.61				
	C57-1-1	27.46	0.19	1.89	-0.4	1.32	1.32		2.43	-0.18	1.13	1.13	

Figure 3-8 - dCt, ddCt and fold change calculations of C3H/HeJ, A/J and C57BL/6J DNase I treated and untreated samples SYBR green qPCR. Controls (A1: C1, C2, and C3) were run on a separate plate from tests (A2: R2, R3, and minus14). Thresholds were set on similar values (threshold C1 = 102.10 vs. threshold C2 = 102.41), and to meet 95% precision (Ct mean SD ≤ 0.25), outliers were excluded from the analysis. The SD values with * are adjusted values after excluding outliers. 12.5 µl of SYBR green supermix, 0.5 µl of each forward and reverse primers (10 pM/µl), 1 µl of DNA template (5 ng/µl), Sigma water up to 25 µl were used per reaction and all the templates were tested in triplicates.

3.3.2 ASSESSING THE EFFICIENCY OF CONTROLS (C1, C2, AND C3) AND TESTS (R2, R3, AND -14) PRIMER SETS

First, efficiency of primers in C57-1-0ice was measured. It is important to choose same threshold in different plates to compare the qPCR results. Standard curves of 6 dilution series (dilution factor=2) were generated (20, 10, 5, 2.5, 1.25, 0.625 ng/µl) and primer efficiencies and R^2s were calculated by Bio-Rad iQ5 Optical System Software version 2.1.97.1001. Efficiencies were calculated by a standard formula (E= [10^(-1/slope)-1]*100). Efficiencies within the range of 90-110% and $R^2 \geq 0.95$ are desirable. Samples loaded in triplicate. SD of the replicates was measured. SD≤0.167 indicates 99.7% precision and SD≤0.25 represents 95% precision. Since designed C1 proved not to be a good candidate for qPCR by in silico BLAT search for unique alignment to the mouse genome and wide melting curve peak characteristics it was excluded from further evaluations.

Figures 3-10 and 3-11 show the efficiencies of given primers on C57-1-0ice template in two experiments. Global comparisons between efficiency of positive DHS primer sets (R2, R3, and minus 14) and DNase I insensitive control primer sets (C1, C2, and C3), R^2 and SD of Ct values in both experiments indicates C3 and C2 primers can be considered both as acceptable primer sets; however C3 have a better efficiency to amplify DNase I untreated template. R2 and R3 also are both more efficient primers than minus14. Further assessments were performed on C57 genomic DNA, C57-1-64 (as the most extreme phase of digestion providing the chance to test efficiency of primers in DNase I treated samples) (Figures 3-12 and 3-13). Efficiency of primers on genomic C57BL/6J

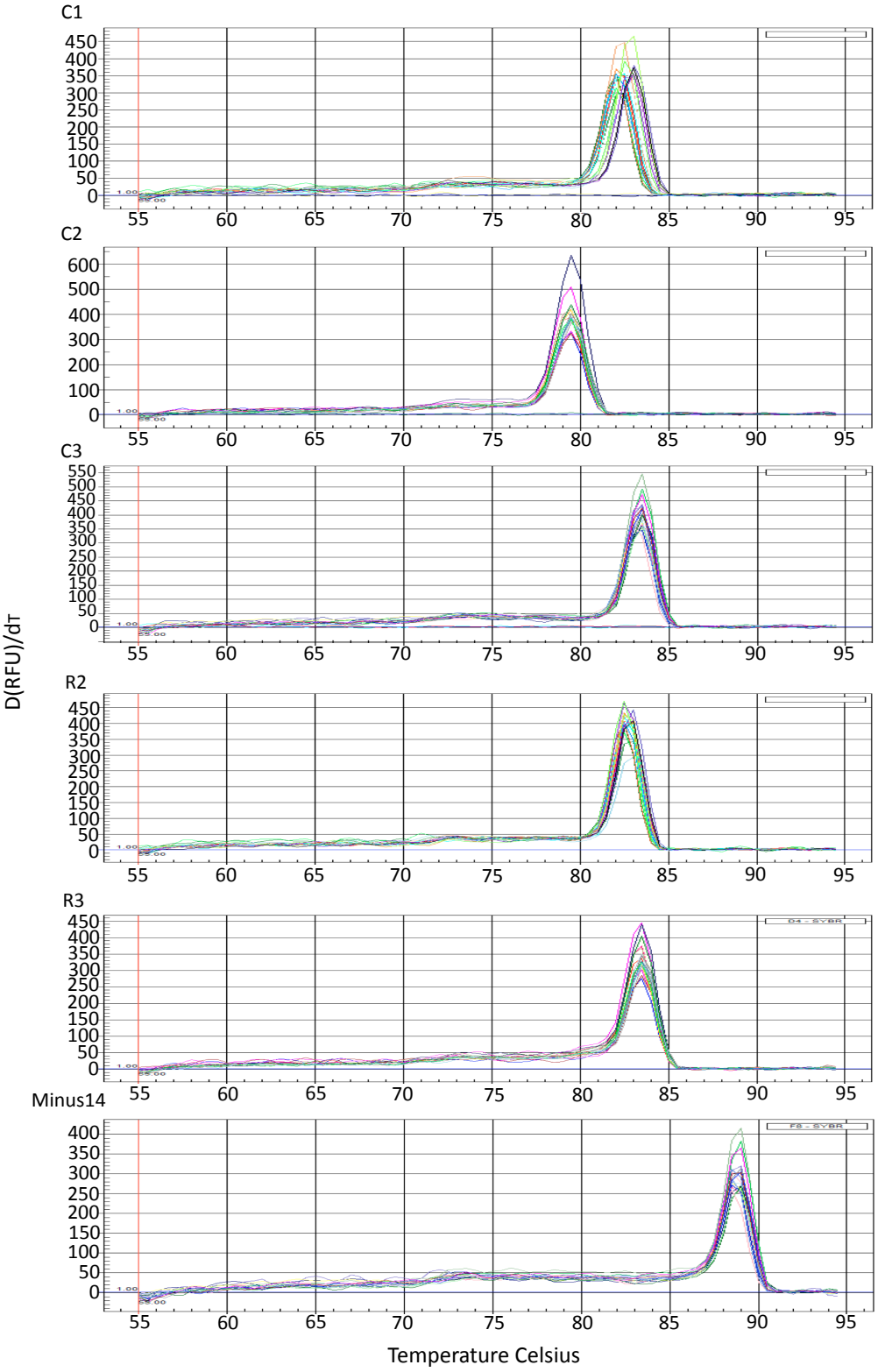


Figure 3-9 - Melting curve analysis of the qPCR products. Following the final step of amplification, the samples were melted by increasing the temperature slowly to 94.5 °C and monitoring the fluorescent signal continually. The rate of change of fluorescence is shown as a function of temperature. Major peaks occur, centred for C1 product (147 bp), C3 (147 bp), R2 (161 bp), R3 (159 bp) at 83, at 79 for C2 (150 bp), and at 87 for minus14 (142 bp). The bigger the PCR product size the sooner the peak appears. C1 broad peak indicates the unspecificity of C1 primer pair so I omitted it from the following experiments.

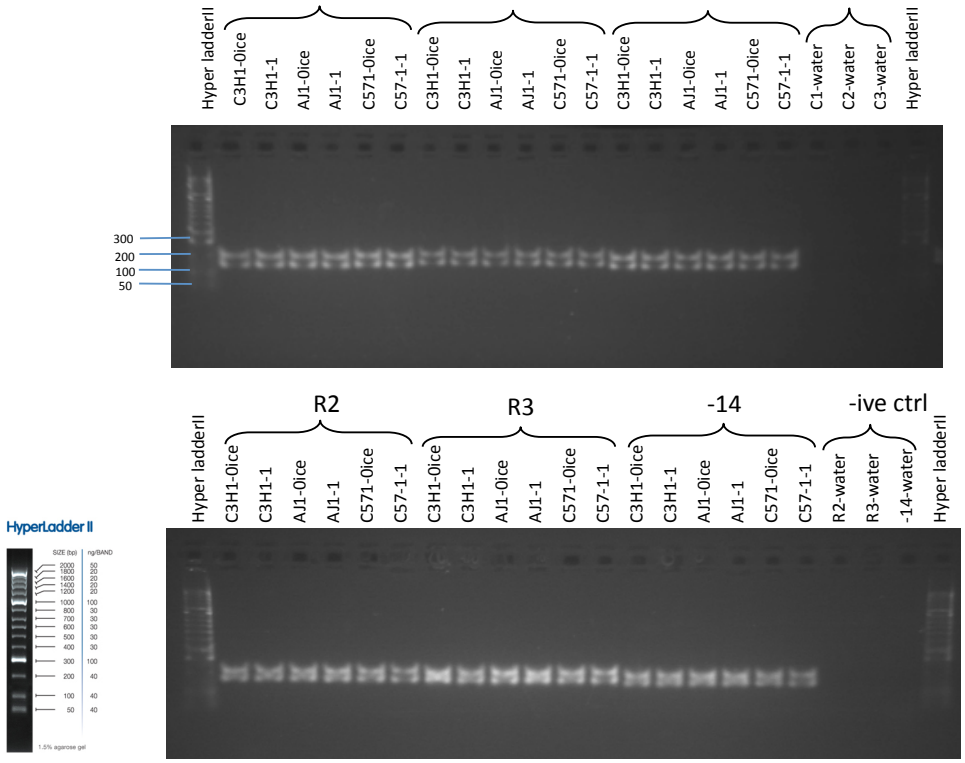


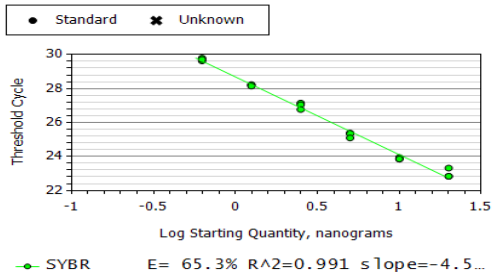
Figure 3-10 - Agarose gel electrophoresis of the PCR products of C1, C2, C3, R2, R3, minus14 and the negative control. Single bands of the expected sizes are evident for the template-dependent PCRs (C1 product (147 bp), C3 (147 bp), R2 (161 bp), R3 (159 bp), C2 (150 bp), and minus14 (142 bp)), whereas for the negative control no single band is seen.

DNA was similar to C57-0ice-1 as expected. C2 showed better efficiency than C3. R2 and R3 had almost similar efficiencies and amplified DNase I extremely treated template significantly better than minus14. Since primer efficiencies at C57-1-64 were not within accepted range (except C2), I chose to further assess primer efficiencies in DNase I treated samples by testing their efficiencies on C57-1-16 template to avoid probable extreme condition of C57-1-64, which contains maximum amount of DNase I among all treated samples (Figure 3-14). Similar to other DNase I treated template (i.e. C57-1-64), C57-1-16 more

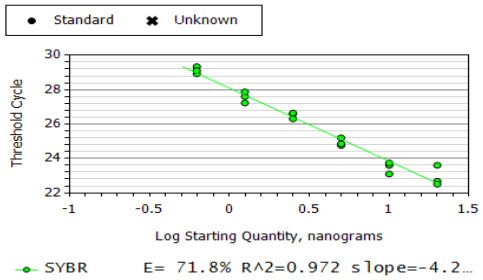
efficiently amplified by C2 and R3 (both satisfy 90-110 efficiency assumption) and have comparable efficiencies. C3 works better in untreated templates but its efficiency is not acceptable in treated samples. Since R2 and R3 had almost similar efficiencies in both untreated and treated templates, R2 could be used as a back up primer set for DHS positive amplicon.

Table 3-2 summarizes the qPCR primer efficiencies, R^2, precision of the results by SD of Ct means in dilution series of standard curve. All the dilution series was initially generated by triplicate samples for 6 dilution series of templates (20, 10, 5, 2.5, 1.25, 0.625 ng/µl). In some cases, the samples which deviated SD > 2.5 or decreased efficiencies and R^2 significantly were excluded from the analysis here. Based on the qPCR results (including standard curves, primer efficiencies and melting curves), primers C2-C3 and R3 were selected as they introduce comparable primer efficiency for test and control in treated and non-treated samples.

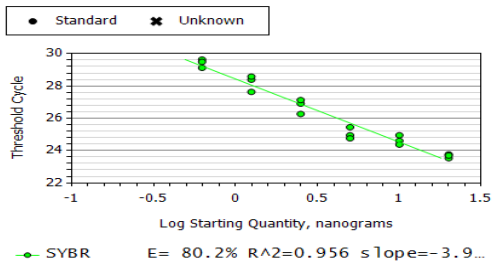
C2



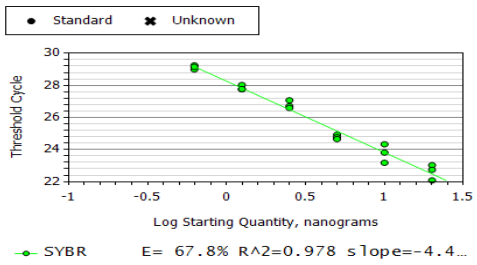
C3



R2



R3



-14

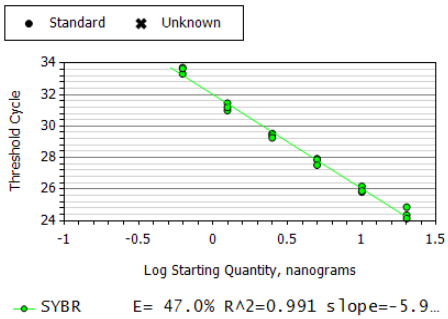


Figure 3-11 - Assessment of primer sets efficiencies on C57-0ice template: first experiment. Threshold was set at 101.32 for the analysis. R2 and C3 showed better efficiencies. Minus 14 showed the lowest efficiency.

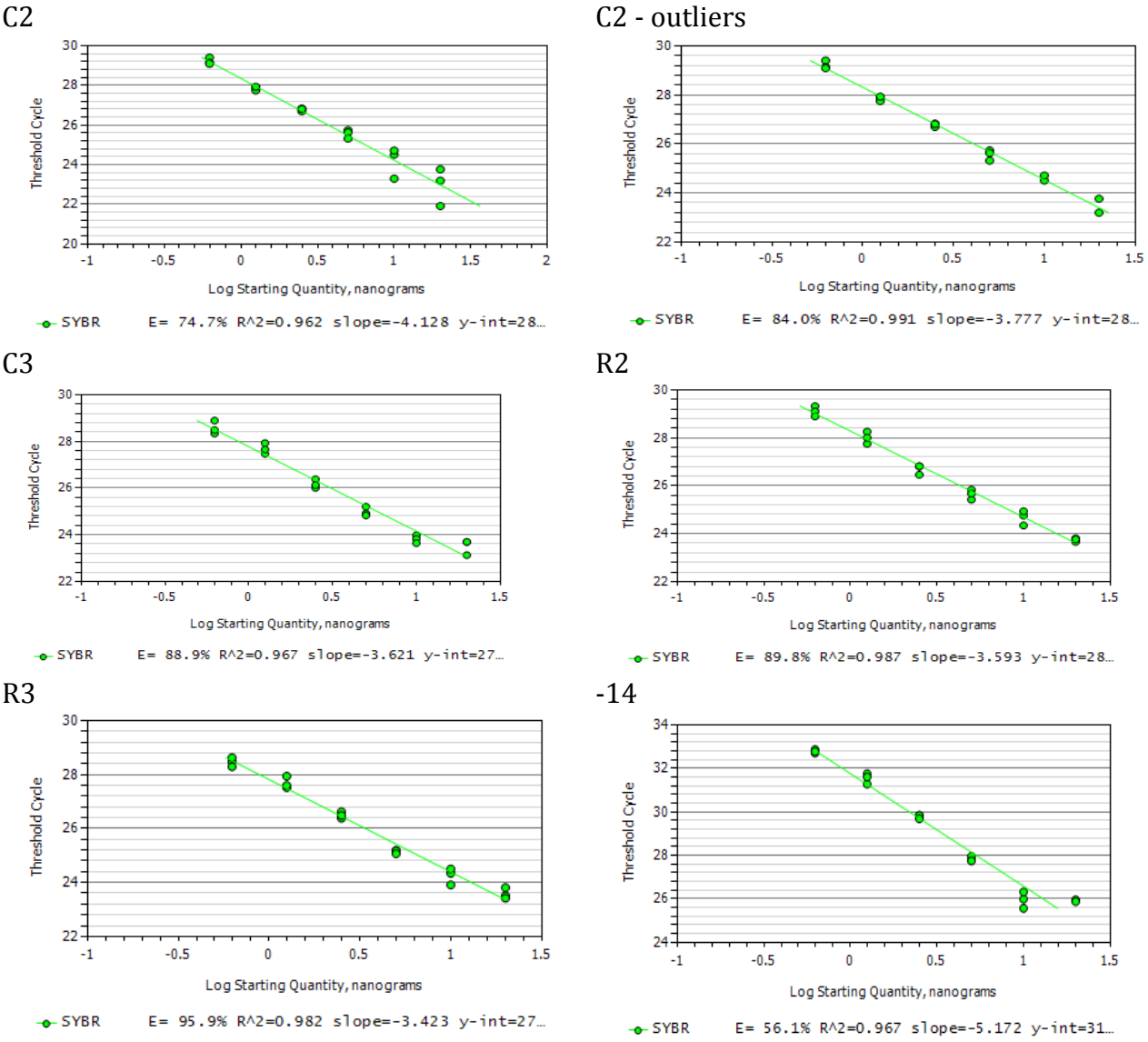
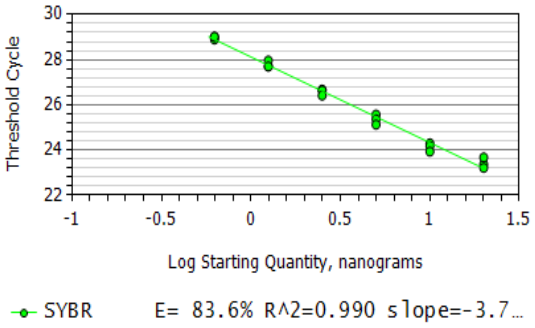
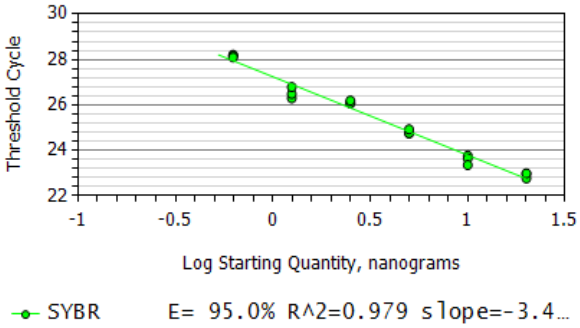


Figure 3-12 - Assessment of primer sets efficiencies on C57-0ice template: second experiment. Threshold was set at 101.87 for the analysis. R2 and C3 with ~ 90% efficiency and R3 with ~ 96% efficiency are within acceptable range (90-110%). Minus 14 showed the lowest efficiency. Excluding outliers from standard curve calculations improved C2 efficiency to 84% efficiency in C57-0ice template (DNase I untreated sample).

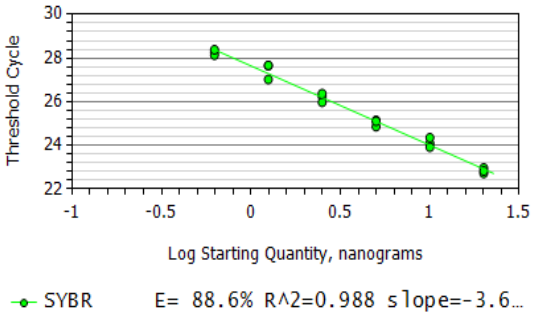
C2



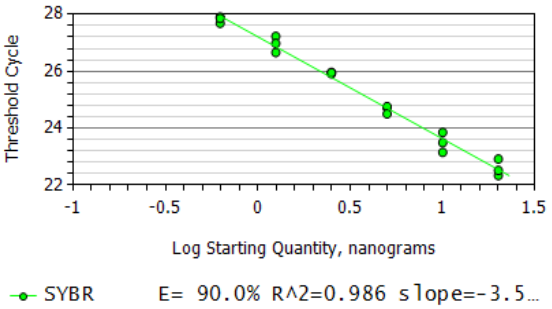
C3



R2



R3



-14

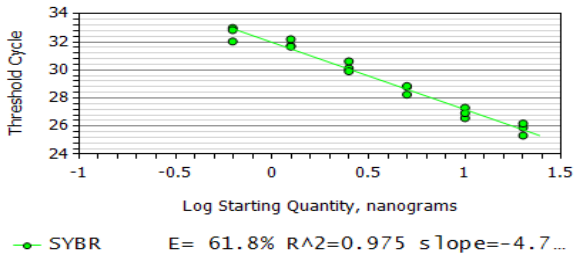
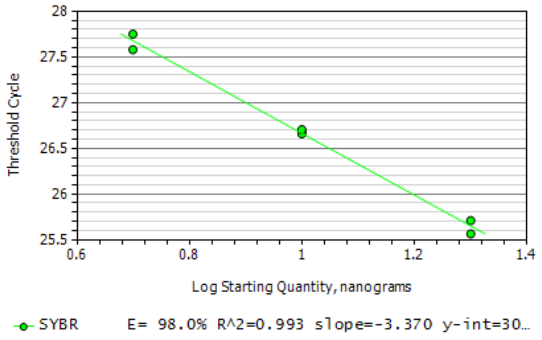
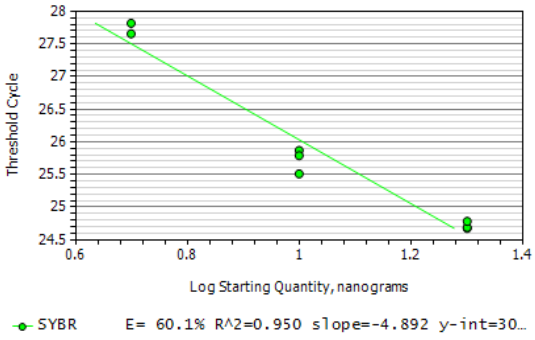


Figure 3-13 - Assessment of primer sets efficiencies on C57BL/6J genomic DNA template. Minus 14 was the weakest primer set. Threshold was set at 100.45.

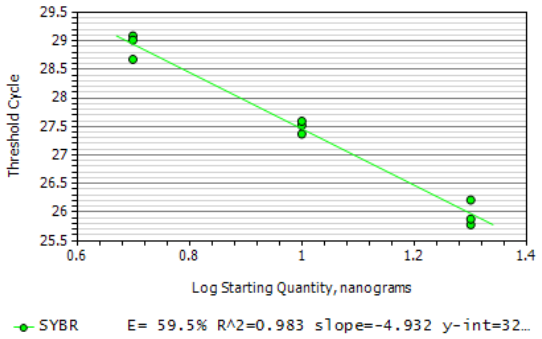
C2



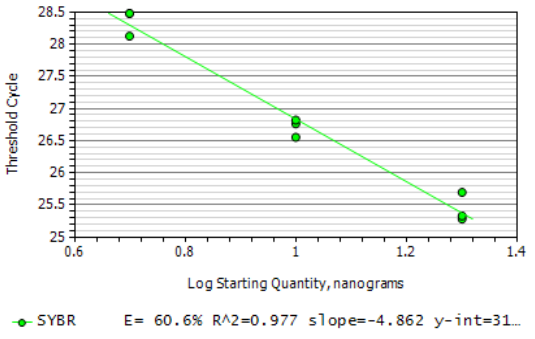
C3



R2



R3



-14

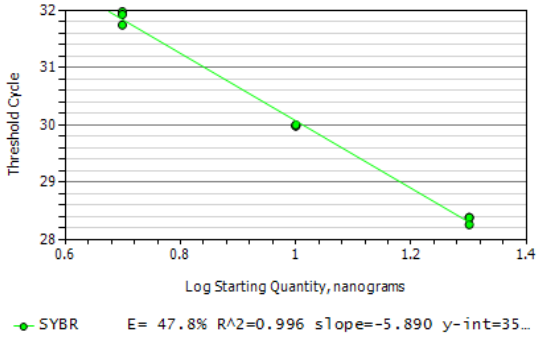
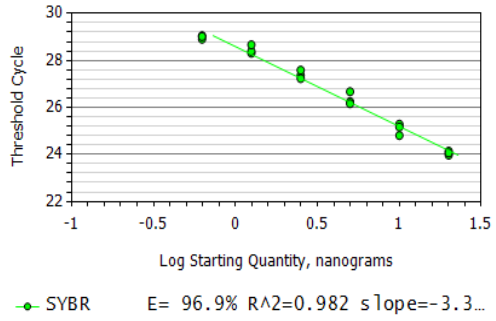
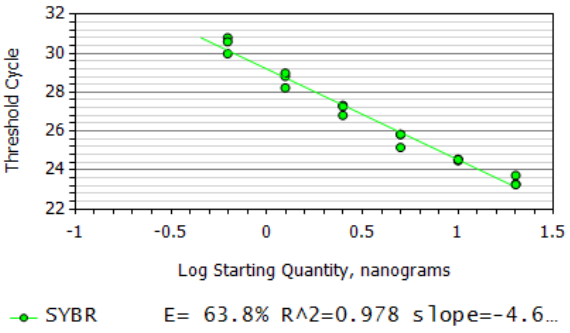


Figure 3-14 - Assessment of primer sets efficiencies on C57-1-64 template. Threshold was set at 101.98. Since standard 5-6 series of 2 times dilution did not generate acceptable efficiencies, I included 3 higher concentrations (20, 10, 5 ng/µl) to roughly estimate the efficiency of primers at the most DNase I treated template (No. 64). Outliers, which introduced Ct SD>2.5, were excluded from this analysis. C2 had better efficiency in treated templates. Minus 14 had the lowest efficiency among all the primer sets.

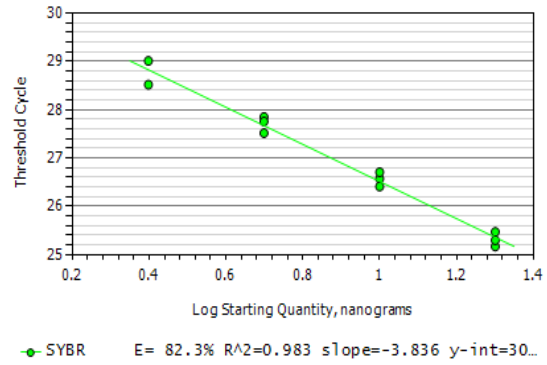
C2



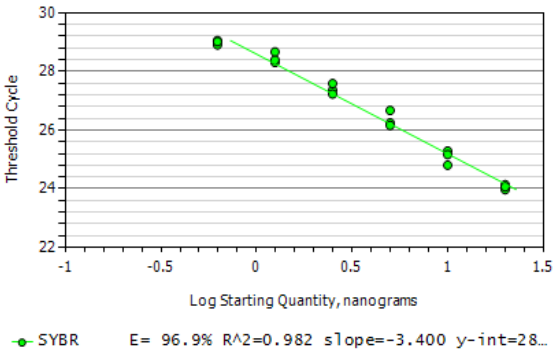
C3



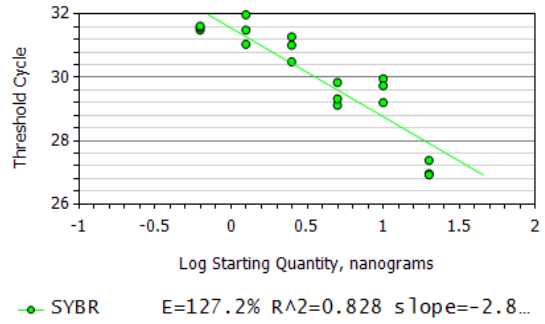
R2



R3



-14



-14 excluding outliers and dilution 10 ng/µl

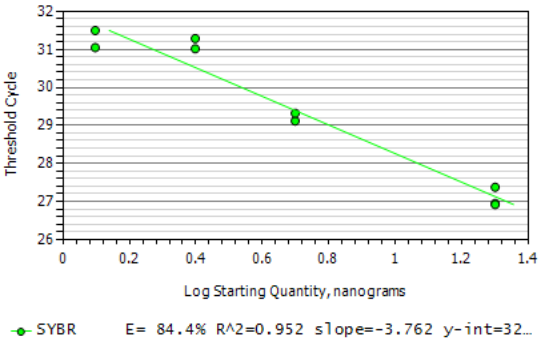


Figure 3-15 - Assessment of primer sets efficiencies on C57-1-16 template. Threshold was set at 102.67. C2 and R3 had better efficiencies to amplify DNase I treated template.

Table 3-2 - qPCR primer set efficiencies of DNase I untreated and DNase I treated templates

Template	Threshold	Primer set	Efficiency (%)	Slope	R^2	Ct mean (SD)					
						20	10	5	2.5	1.25	0.625
C57-1-0ice-1st	101.32	C2	65.3 (63.3*)	-4.5 (-4.7*)	0.991 (0.995*)	0.28 (0.00*)	0.05	0.16	0.20	0.05	0.08
		C3	71.8 (68.1*)	-4.2 (-4.4*)	0.972 (0.993)	0.59 (0.12*)	0.34 (0.10*)	0.23	0.18	0.32 (0.19*)	0.21
		R2	80.2 (76.2*)	-3.9 (-4.0*)	0.956 (0.963*)	0.12	0.29 (0.17*)	0.36 (0.14*)	0.45 (0.16*)	0.50 (0.14*)	0.27 (0.10*)
		R3	67.8 (71.4*)	-4.4 (-4.3*)	0.978 (0.981)	0.48 (0.20*)	0.58 (0.37*)	0.14	0.25	0.15	0.13
		-14	47.0(45.9*)	-5.9 (-6.1*)	0.991 (0.994)	0.37 (0.15*)	0.19	0.22	0.14	0.23	0.21
C57-1-0ice-2nd	101.87	C2	74.7 (84.0*)	-4.1 (-3.8*)	0.962 (0.991*)	0.90 (0.40*)	0.76 (0.15*)	0.22	0.07	0.09	0.17
		C3	88.9 (88.5*)	-3.6 (-3.6*)	0.991 (0.963*)	0.33 (0.04*)	0.17	0.20	0.19	0.23	0.29
		R2	89.8 (91.1*)	-3.6 (-3.6*)	0.967 (0.989*)	0.08	0.30 (0.13*)	0.21	0.20	0.25	0.21
		R3	95.9 (97.9*)	-3.4 (-3.4*)	0.987 (0.985*)	0.31 (0.20*)	0.11	0.07	0.12	0.22	0.17
		-14	56.1 (58.0*)	-5.2 (-5.0*)	0.982 (0.974*)	1.64 (0.06*)	0.38 (0.23*)	0.11	0.09	0.24	0.08
C57 genomic DNA	100.45	C2	83.6	-3.8	0.990	0.24	0.19	0.22	0.14	0.15	0.07
		C3	95.0	-3.4	0.979	0.12	0.21	0.10	0.07	0.25	0.05
		R2	88.6 (87.6*)	-3.6 (-3.7*)	0.988 (0.989)	0.13	0.22	0.16	0.20	0.37 (0.01*)	0.12
		R3	90.0 (85.8*)	-3.5 (-3.7*)	0.986 (0.990*)	0.30 (0.13*)	0.35 (0.25*)	0.14	0.03	0.29 (0.18*)	0.12
		-14	61.8 (63.6*)	-4.7 (-4.7*)	0.975 (0.995*)	0.44 (0.21*)	0.37 (0.03*)	0.34 (0.04*)	0.36 (0.27*)	0.2 (0.15*)	0.52 (0.11*)
C57-1-64†	101.98	C2	98.0	-3.4	0.993	0.10	0.02	0.12			
		C3	60.1	-4.8	0.950	0.06	0.19	0.11			
		R2	59.5	-4.9	0.983	0.22	0.11	0.23			
		R3	60.6	-4.8	0.977	0.23	0.14	0.20			
		-14	47.8	-5.9	0.996	0.07	0.12	0.01			
C57-1-16	102.57	C2	96.9 (96.3*)	-3.3 (-3.4*)	0.982 (0.986*)	0.08	0.25	0.06*	0.18	0.19	0.08
		C3	63.8 (59.8*)	-4.6 (-4.9*)	0.978 (0.995*)	0.02*	0.04	0.00*	0.04*	0.12*	0.14*
		R2	124.3 (82.3*)	-2.8 (-3.8*)	0.934 (0.983*)	0.15	0.15	0.17	0.35*		
		R3	96.2 (98.1*)	-3.4 (-3.4*)	0.949 (0.956*)	0.27	0.12	0.03	0.25*	0.25*	0.10
		-14	127.2 (84.4*)	-2.8 (-3.7*)	0.828 (0.952*)	0.26		0.14*	0.32*	0.19*	

*** Values obtained after excluding outliers from the calculations.**

† Standard curves analyzed including 3 dilution series 20, 10, 5 ng/µl.

3.4 DNASE I ASSAYS AND SELECTION OF MID-PHASE OF DIGESTIONS BY QPCR

I carried out DNase I digestion on genomic DNA of three biological replicates of phenylhydrazine treated mice of each of A/J, AKR/J, BALBc/J, C57BL/6J, C3H/HeJ, DBA/2J, and 6 replicates of CBA/J and LP/J strains (Figure 3-16). TER119-positive cells were measured (Figure 3-17) and DNA quantities were calculated using NanoDrop (triplicate) and PicoGreen DNA quantification by a standard protocol (duplicate). qPCRs were done with C2 - R3 in all strains except AKR/J (C3 - R3 in this case) as C2 contains a SNP in the AKR/J strain and the phases for library preparation were selected (Figure 3-18).

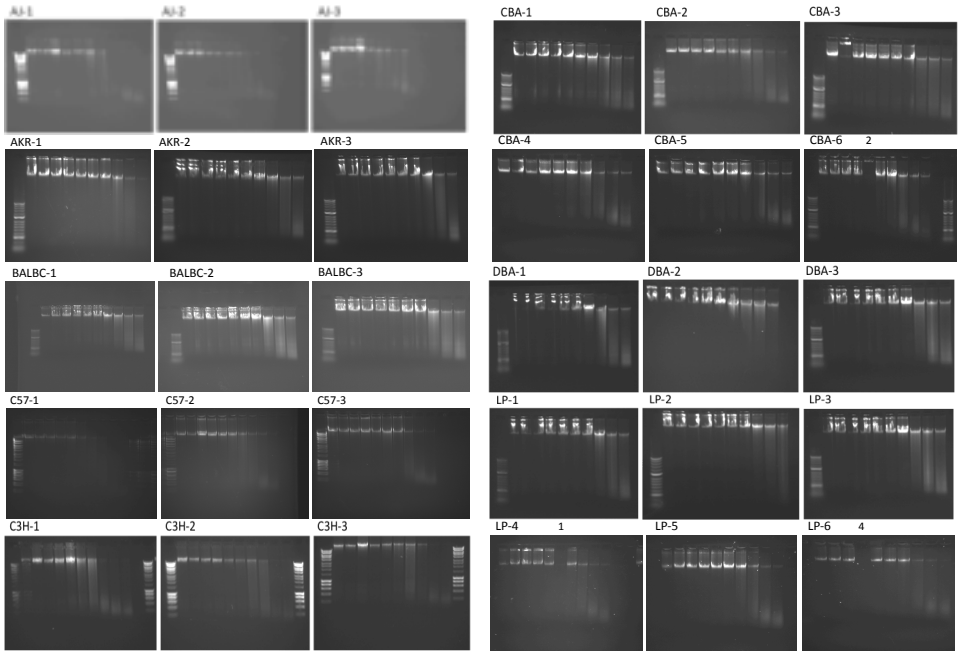


Figure 3-16 - DNase I digestion of genomic DNA of eight given inbred mouse strains. 1/10 diluted DNAs eluted from 0ice, 0_{37}, 0.5, 1, 2, 4, 8, 16, 32, and 64 phases of DNase I digestion are loaded on the 1% agarose gel (TAE buffer) at 110 v for 25 minutes. Absence of obvious band in a lane indicates loss of DNA pellet during phenol/chloroform extraction process (e.g. LP-4-1, LP-6-4, and CBA-6-2).

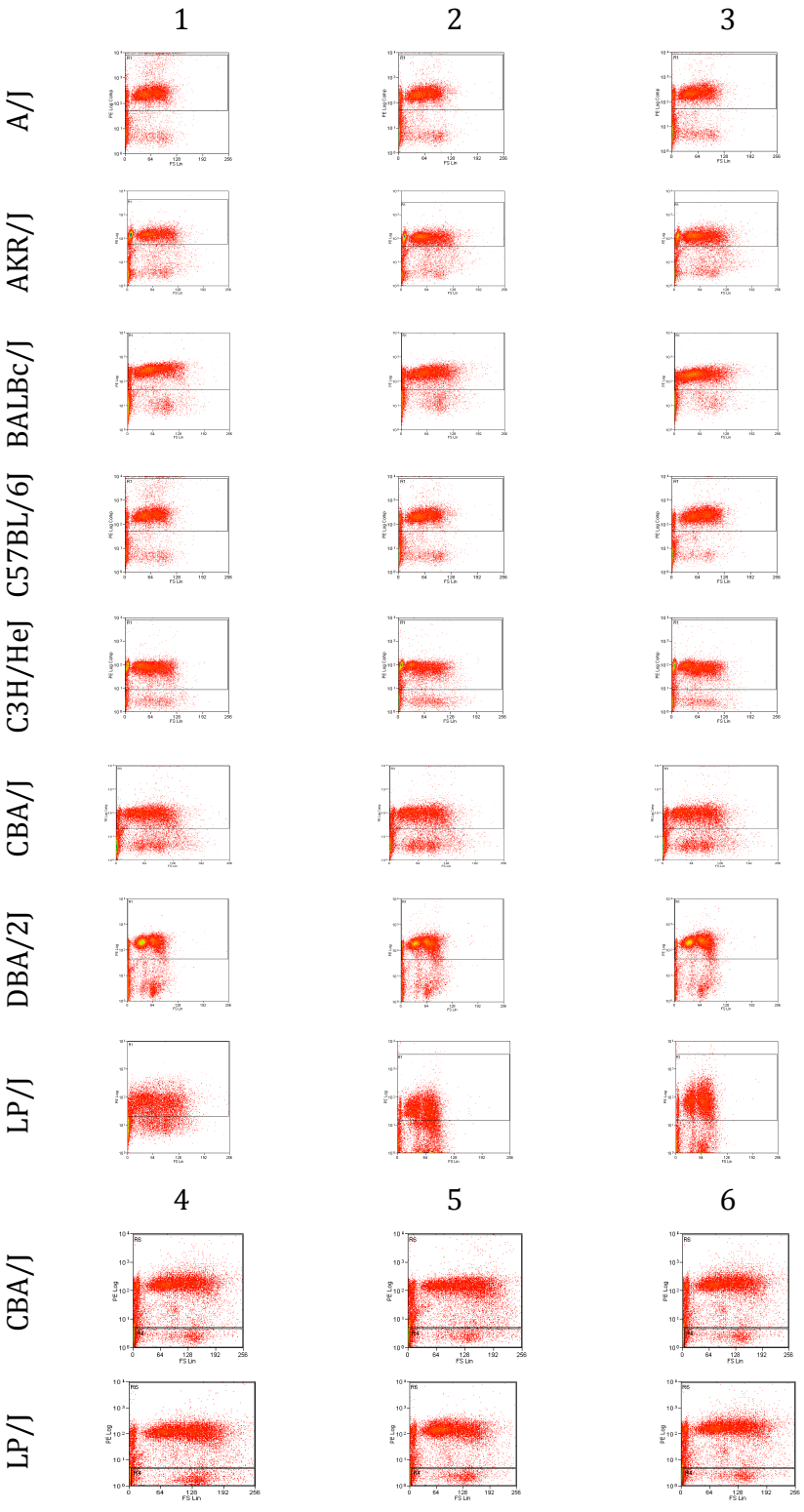


Figure 3-17 - Flow cytometric analyses of mouse primary erythroblasts. TER119 positive erythroid cells were auto–magnetic- activated cell sorting–purified from the spleen of phenylhydrazine- treated adult mice and restained with anti-CD71 to show the equivalent maturation stages in all mice examined. Dr Jackie Sloan Stanley measured TER119 positive cells. Table 3-1 contains the percentage of TER119 positive cells in each experiment.

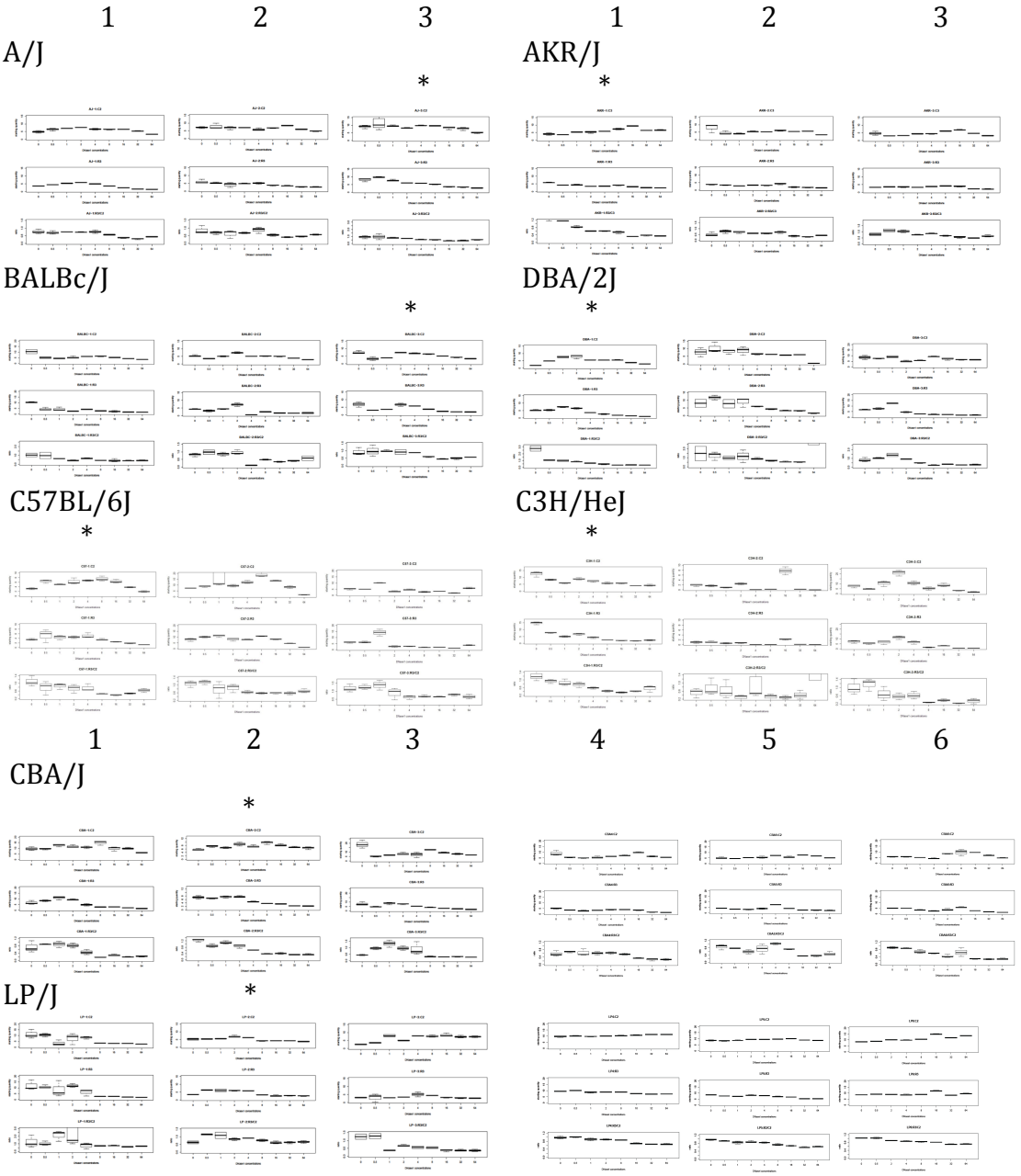


Figure 3-18 – Quantification of DNaseI digestion by qPCR in the eight mice strains. Each column represents a biological replicate (3 per each of A/J, AKR/J, C57BL/6J, C3H/HeJ, BALBc/J, and DBA/2J strains and 6 per each of CBA/J, and LP/J). The initially selected digestions for library preparations by GAII are starred. The first and second rows in each panel illustrate estimated starting quantities of the template (Y axis) in samples with increasing titration of DNaseI (X axis) for DHS negative and positive set of primers respectively (i.e. C2 and R3 except AKR/J: C3 and R3). 8 different concentrations of DNaseI used are labelled by the volume of 10 U/µl DNaseI used per 1 ml of reaction volume (0.5, 1, 2, 4, 8, 16, 32, 64 µl). The third row shows the R3/C2 ratio (except AKR/J: R3/C3, please see methods for further details). Samples No. 1, 2, 4 and 8 were chosen for GAII library preparation for all the eight selected replicates. For better visualization of this figure please see appendices 4-11.

Theoretically, templates are expected to be equally amplified pre and post DNase I digestion at DNase I resistant sites (controls such as C2, or C3); however, amplification curve of DHS primers was expected to shift to the right as less template was expected to be available to be amplified post DNase I digestion in compare with untreated templates. Figure 3-19 is an example of what I observed in my samples. Standard curves for both DHS and control primers sets were included in each qPCR plate and the starting quantity of templates were estimated by iQ5 software. R3/C2 or R3/C3 for AKR/J ratios are expected to be close to 1 at 0ice samples and the higher the DNase I concentration the lower the ratio at mid-phase of digestion. To capture similarly digested material for the library preparation, 4 phases at mid-phase of digestion which showed declining trend in R3/C2 or R3/C3 ratio were selected (see Table 3-3).

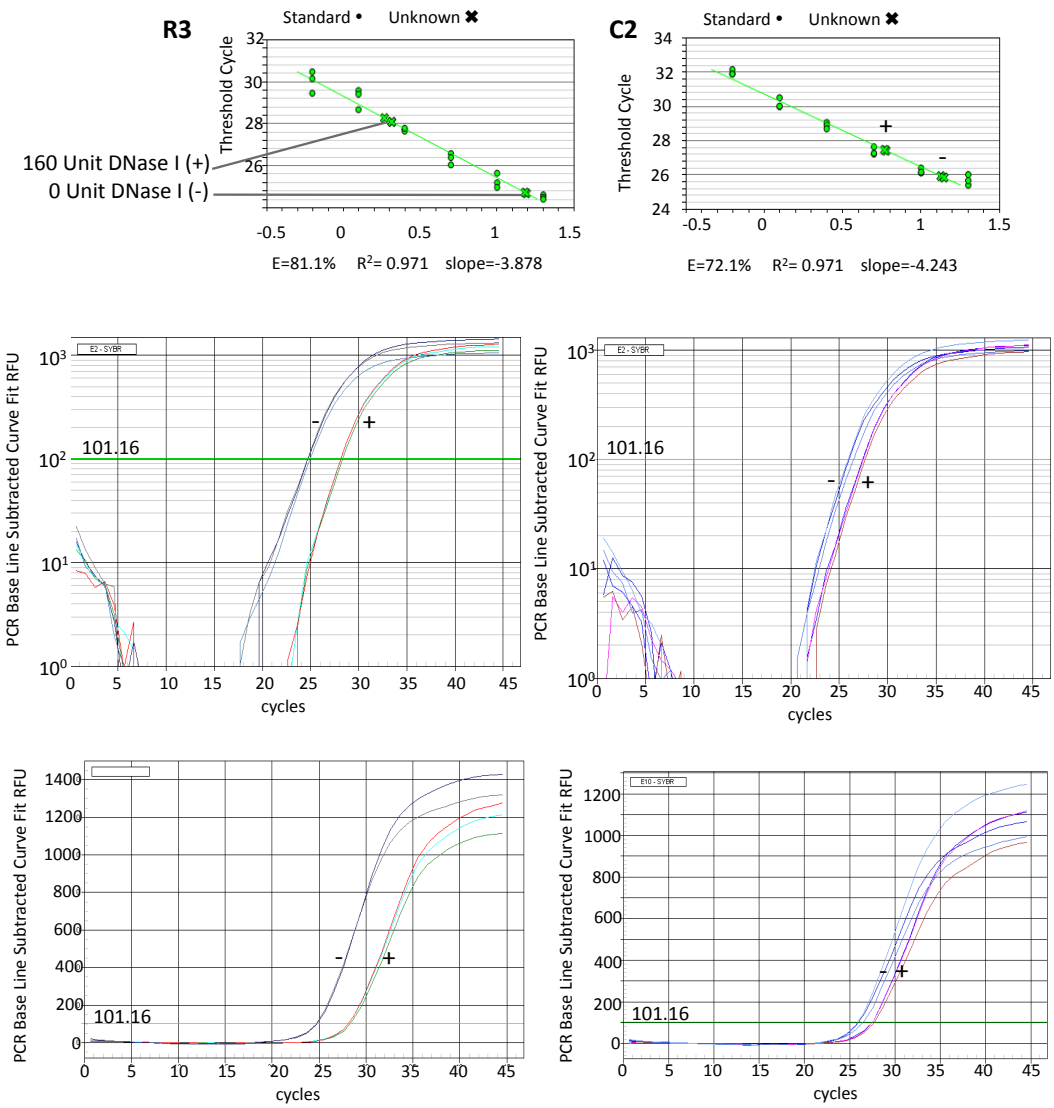


Figure 3-19 - SYBR-green qPCR profiles of R3 and C2 primer pairs. Standard curves of 2- fold dilutions of C3H/HeJ-1 genomic DNA of both primer pairs are generated, and 0/160 U DNase I treated amplicons are quantified relative to standard curves. Amplification curves show florescence vs cycle number. Threshold was set at 101.16. Left hand panel shows DNA amplicon from 0 U DNase I treated sample amplified ~3 PCR cycles in advance of 160 U DNase I treated template. Significant less difference (~1 cycle) was observed in 0 and 160 U DNase I treated samples with C2.

Based on qPCR R3/C2 0r R3/C3 ratio results (Figure 3-18), 8 samples, which had the best declining trend in their mid-phase of digestion, were initially chosen (those which depicted by star: A/J-3, AKR/J-1, BALBc/J-3, C57BL/6J-1, C3H/HeJ-1, DBA/2J-1, LP/J-2) for Solexa compatible library preparation. Figure 3-20 shows the outline of library preparation (see Methods). All the libraries passed library quality control except LP/J-2, which was inadequate in terms of quantity for sequencing. These libraries were sequenced by GAII platform.

To include replicates, I prepared libraries from all other samples and made new libraries from LP/J-2, which failed the quality check last time and CBA/J-2 (preliminary DNase-seq analysis indicated the library was overdigested compared to other libraries); a new library was prepared from adjusted phases of digestion). The libraries, which passed quality check before or after improving concentration with Savant ISS 110 SpeedVac Concentrator were submitted for sequencing on a HiSeq platform. As part of quality testing, contamination of libraries with adapters had been tested by Agilent technologies 2200 Tape Station Software A.01.01. The DBA/2J-3 library was found to be contaminated with adaptors (23.80% of integrated area) and was therefore made again. A total of 21 libraries were sequenced by HiSeq platform: 14 DNase-seq libraries from terminally differentiated erythroblasts, and 2 libraries from C57BL/6J MEFs. For details see Table 3-3.

Table 3-3 - Selected phases of DNase I digestion for DNase-seq library preparations.

Sample ID		Threshold	R3 or *Gapdh* efficiency % *	C2 or C3 efficiency %†	Selected phases (R3/C2, R3/C3, or *Gapdh*/C3 ratios)				GAII	HiSeq
A/J	1	166.84	108.1	83.3	1 (0.8)	2 (0.7)	4 (0.5)	8 (0.6)		x
	2	258.37	108.7	69.9	1 (0.7)	2 (0.7)	4 (0.8)	8 (0.6)		x
	3	252.73	90.2	80.5	1 (0.8)	2 (0.7)	4 (0.6)	8 (0.5)	x	
AKR/J	1	955.40	95.0	66.0	1 (0.8)	2 (0.8)	4 (0.6)	8 (0.5)	x	
	2	783.04	93.1	74.5	0.5 (0.7)	1 (0.7)	2 (0.6)	4 (0.6)		
	3	587.04	75.7	57.6	1 (1.0)	2 (0.8)	4 (0.8)	8 (0.7)		x
BALBc/J	1	513.40	106.7	72.7	1 (1.0)	2 (0.5)	4 (0.7)	8 (0.4)		
	2	666.38	83.2	50.1	0.5 (0.7)	1 (0.8)	2 (1.0)	8 (0.5)		x
	3	616.53	83.1	52.2	1 (1.0)	2 (1.0)	4 (0.8)	8 (0.6)	x	
C57BL/6J	1	120.06	99.5	105.5	1 (0.9)	2 (0.7)	4 (0.8)	8 (0.5)	x	
	2	79.30	70.1	64.4	1 (1.0)	2 (0.8)	4 (0.6)	8 (0.6)		x
	3	102.26	66.4	61.6	0.5 (1.3)	1 (1.4)	2 (0.8)	4 (0.6)		x
C3H/HeJ	1	102.61	81.0	72.0	1 (0.9)	2 (0.8)	4 (0.6)	8 (0.4)	x	
	2	112.48	94.6	93.4	1 (0.6)	2 (0.3)	4 (0.3)	8 (0.7)		x
	3	151.34	101.9	93.3	0.5 (1.1)	1 (0.6)	2 (0.5)	4 (0.7)		
DBA/2J	1	631.00	80.5	51.2	1 (1.2)	2 (0.8)	4 (0.7)	8 (0.5)	x	
	2	735.85	79.8	50.1	1 (0.8)	2 (0.8)	4 (0.9)	8 (0.7)		x
	3	746.85	58.8	57.8	0.5 (1.0)	1 (1.5)	2 (0.9)	4 (0.5)		
CBA/J	1	154.32	104.4	105.4	0.5 (0.9)	1 (0.9)	2 (0.8)	4 (0.4)		x
	2	518.01	69.9	93.3	1 (1.1) 0.5 (0.9)	2 (0.9) 1 (1.1)	4 (0.6) 2 (0.9)	8 (0.4) 4 (0.6)	x	
	3	486.52	93.6	88.9	0.5 (0.9)	1 (1.3)	2 (0.9)	4 (0.7)		
	4	579.44	74.6	67.8	1 (0.63)	2 (0.59)	4 (0.62)	8 (0.56)		
	5	630.47	82.3	51.8	0.5 (0.8)	1 (0.6)	2 (0.7)	8 (0.7)		
	6	86.79	101.2	60.3	0.5 (0.9)	1 (0.7)	4 (0.4)	8 (0.5)		x
LP/J	1	220.38	60.5	107.5	0.5 (1.0)	1 (2.3)	2 (1.6)	4 (0.7)		
	2	730.57	87.6	57.2	1 (1.1)	2 (0.8)	4 (0.9)	8 (0.8)		
	3	1161.95	94.4	57.8	2 (0.7)	4 (0.7)	8 (0.6)	16 (0.4)		x
	4	121.24	31.7	27.1	1 (0.89)	2 (0.88)	8 (0.85)	16 (0.7)		x
	5	103.99	29.5	20.7	1 (0.83)	2 (0.87)	4 (0.82)	8 (0.73)		
	6	369.32	29.4	20.6	2 (0.89)	4 (0.86)	8 (0.83)	16 (0.81)		
C57BL/6J MEF	1	119.18	42.5	48.2	0.5 (0.74)	4 (0.78)	8 (0.75)	16 (0.52)		x
	2	280.28	16.6	26.9	0.5 (1.0)	2 (0.7)	4 (0.7)	8 (0.6)		
	3	662.16	79.4	81.2	0.5 (0.8)	1 (0.6)	4 (0.7)	16 (0.5)		X

* *Gapdh* is positive DHS of C57BL/6J MEFs.

† C3 is control for AKR/J and C57BL/6J. Another C3 primer set designed for MEFs.

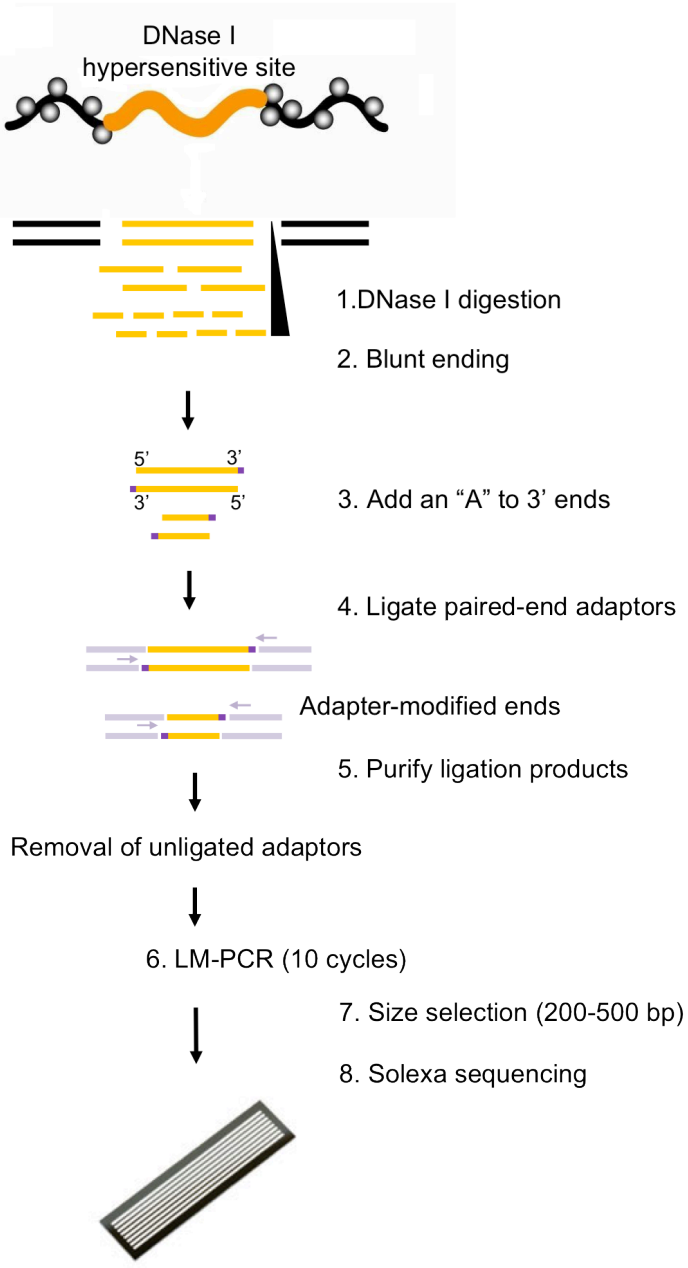


Figure 3-20 - DNase-seq protocol. Outline of DNase-seq protocol.

3.5 PREPARING DNASE-SEQ LIBRARY FROM C57BL/6J MEFS

To identify erythroid specific DHSs vs ubiquitous DHSs, I prepared DNase-seq data from another tissue for comparison with DHS data from terminally differentiated erythroblasts. Initially, primary fibroblasts cells of inbred mice were chosen for this purpose. Fibroblasts were derived from the tails of given strains and stored at liquid nitrogen. Standard protocols were used to revive

frozen cells and culture them in DMEM + 10% FBS medium to obtain enough number of cells for DNase I assay (minimum number of 5 x 10^7 – 1 x 10^8 cells). Since an adequate number of cells was not obtained, I decided to use cell cultures from mouse embryonic fibroblasts. I received 2 batches of cultured C57BL/6J MEFs (4 x 10^7 and 2 x 10^7 cells) from Prof Higgs' group (Dr Jacqueline Solane-Stanely prepared the cells). Dr Ben Davies generated MEFs from C57BL6/J, which was revived and cultured to obtain 3rd batch of C57BL/6J MEFs to perform DNase I assay. To enhance the number of cells I used 175 ml triple flasks on passage IV and 7.9 x 10^7 cells obtained from which 5.9 x 10^7 cells used for DNase I assays. Since the number of cells of second batch was inadequate to perform conventional DNase I assay (minimum number of 5 x 10^7 cells is required), only 0ice, 0.5, 1, 2, 4, 8, and 16 reactions were prepared for DNase I assay (0_{37}, 32 and 64 reactions were omitted from the experiment) (Figure 3-21).

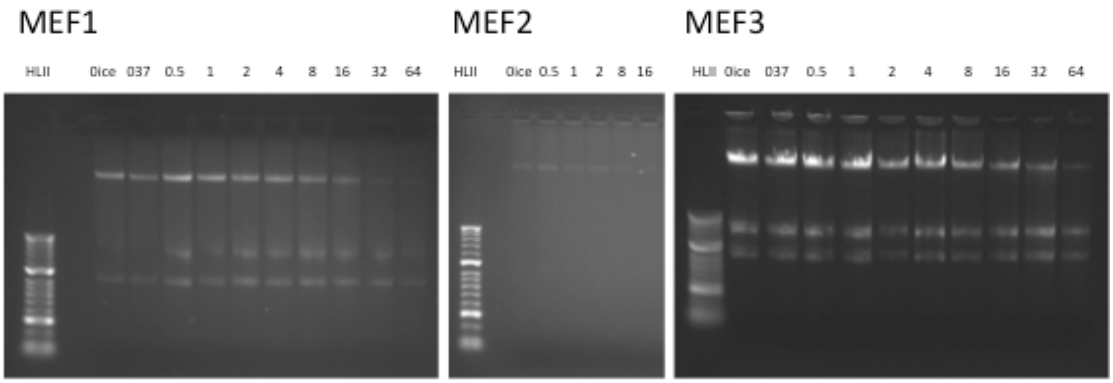


Figure 3-21 - DNase I assays on C57BL/6J MEFs.

3.5.1 DESIGNING PRIMERS FOR SYBR-GREEN QPCR DHS SENSITIVE AND CONTROL PRIMER PAIRS FOR C57BL/6J MEFS

SYBR green qPCR recently published guidelines were applied to design 3 sets of primer pairs for DHS negative amplicons (control-1, control-2, control-3) and 2 primer pairs for DHS positive amplicons of two housekeeping genes *Gapdh* and beta-actin in MEFs (see Methods). The initial DNase-seq data from C57BL/6J also used to improve primer design by choosing sequences under the peak in *Actb* and *Gapdh* loci, which indicate the presence of active regulatory sites. Appendix 12 summarizes the details of designed primers.

3.5.1.1 Assessing the efficiency of MEF qPCR primer pairs

The sensitivity and specificity of these primers was initially tested by a modified Touchdown PCR (see Appendix 3). A single band at the expected size was observed for each primer pairs (Figure 3-22).

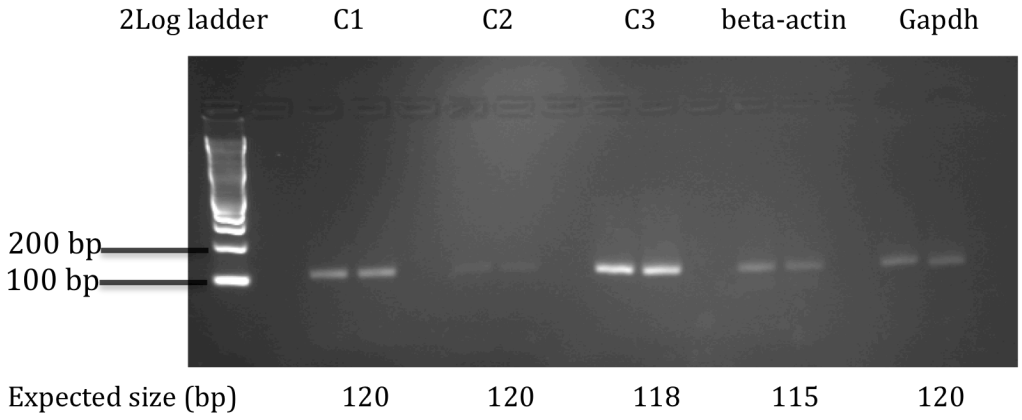


Figure 3-22 - Touchdown PCR products of MEF qPCR primer pairs. The PCR products of two replicates run on 3% TAE gel. C57BL/6J genomic DNA was used as a template. All 5 primer pairs (3 controls: C1-C3 and 2 DHS positive primer pairs: *Actb* and *Gapdh*) showed single bands at expected size indicating their sensitivity and specificity.

3.5.1.2 Checking the efficiency of primers C3 and *Gapdh* selected from C1-3, *Actb* and *Gapdh* on reaction 0ice and 16

A similar strategy was considered to assess the efficiencies of DHS sensitive and control primer pairs for C57BL/6J MEFs. Standard curves of 6 dilution series (dilution factor=2) were generated (20, 10, 5, 2.5, 1.25, 0.625 ng/µl). To improve the precision of efficiency and R^2 calculations of primers, 3 and 4 dilution series were used for the analysis. The efficiency of 2 DHS positive primer pairs (*Actb* and *Gapdh*) and three DHS negative control primer pairs (C1-C3) were tested on C57BL/6J MEF-3-0ice and C57BL/6J MEF-3-16 as untreated and moderately DNaseI-treated templates. 5 ng of DNA template was used in each reaction. Similar threshold values were used to compare results in different plates (616.79 for untreated template vs 611.17 for moderately DNase I-treated template).

Figure 3-23 shows the standard curve analysis of two DHS positive primer pairs (*Actb* and *Gapdh*) on untreated (C57BL/6J MEF-3-0ice) and moderately DNaseI-treated templates (C57BL/6J MEF-3-16). Considering primer efficiencies and R^2 values of primers on untreated and moderately treated templates, the *Actb* is a more efficient primer pair than *Gapdh* at both conditions.

Figure 3-24 illustrates the standard curve analysis of three DHS negative control primer pairs (C1-C3) on untreated (C57BL/6J MEF-3-0ice) and moderately DNase I–treated (C57BL/6J MEF-3-16) templates. Considering efficiency and R^2 values, all three negative primer pairs (C1-C3) have better overall efficiencies than positive primer pair (*Actb* and *Gapdh*). Among three negative controls, C3 showed higher efficiency and R^2 values including 4 dilution series in calculations; however, C1 and C2 also showed desirable efficiency and comparable efficiencies in untreated and moderately treated conditions.

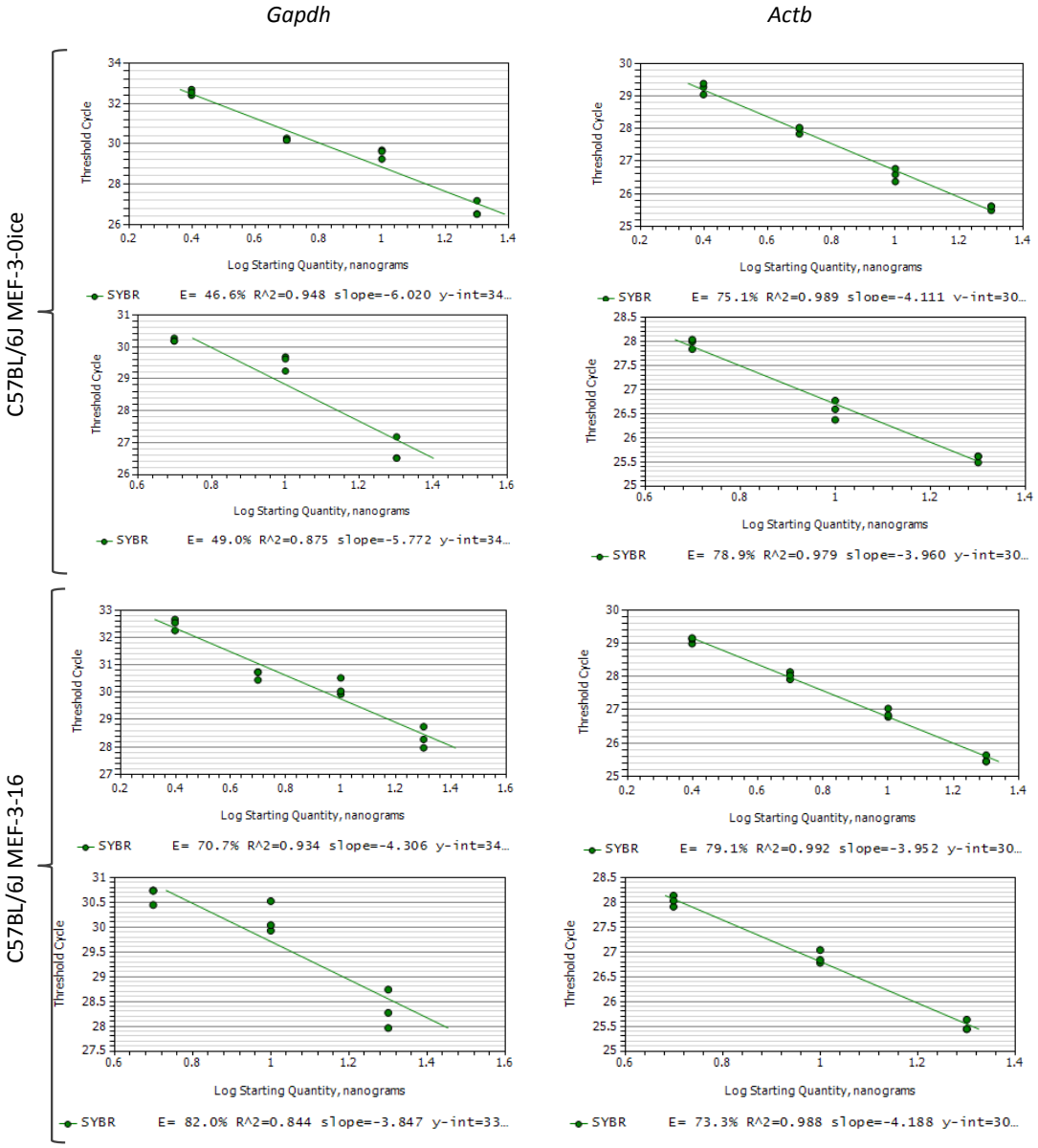


Figure 3-23 - Assessment of *Gapdh* and *Actb* SYBE green qPCR primer pairs on untreated and moderately DNase I treated templates. 3 and 4 dilution series were used for efficiency calculations. Threshold of untreated template (C57BL/6J MEF-3-0ice) set at 611.17, and threshold of moderately DNase I treated template (C57BL/6J MEF-3-16) set at 616.79.

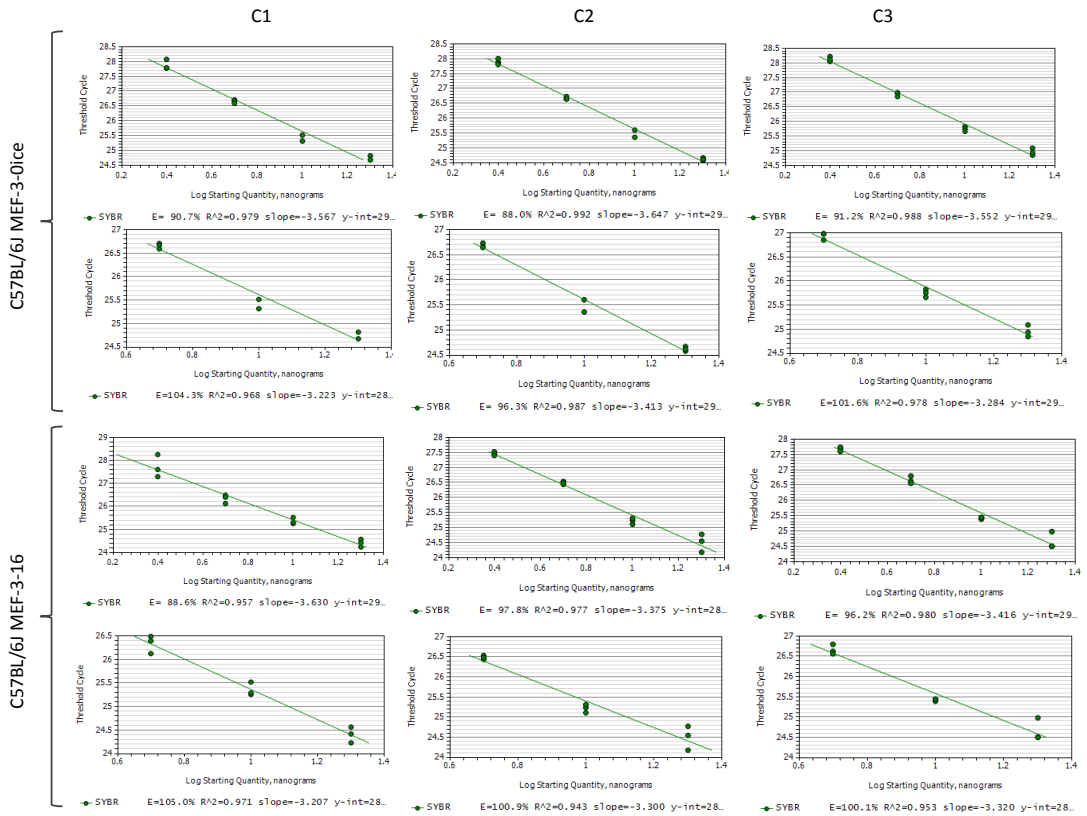


Figure 3-24 - Assessment of DHS negative SYBR green qPCR primer pairs (C1-C3) on untreated and moderately DNase I treated templates. 3 and 4 dilution series were used for efficiency calculations. Threshold of untreated template (C57BL/6J MEF-3-0ice) set at 611.17, and threshold of moderately DNase I treated template (C57BL/6J MEF-3-16) set at 616.79.

Another critical factor is the melting curve analysis, a measure of the specificity of primer pairs. Melting curve analysis shows that the *Gapdh* primer pair generated a more specific product (significantly narrower peak) compared to the *Actb* primer pairs. Therefore, *Gapdh* was selected as a positive control primer pair for further SYBR green qPCR analysis on C57BL/6J MEF DNase I assays.

Primer pairs C1-C3 all generated narrow peaks. C3 had higher efficiency and R^2 values on both untreated and treated templates and was selected as a negative control for further SYBR green qPCR analysis on C57BL/6J MEF DNase I assays. Figures 3-25 and 3-26 show the melting curve analysis of DHS positive

and negative control primer pairs on both untreated (C57BL/6J MEF-3-0ice) and DNase I treated (C57BL/6J MEF-3-16) templates. Table 3-4 summarizes the efficiency and R^2 values of given primers at untreated and DNase I-treated templates including 3 and 4 dilution series.

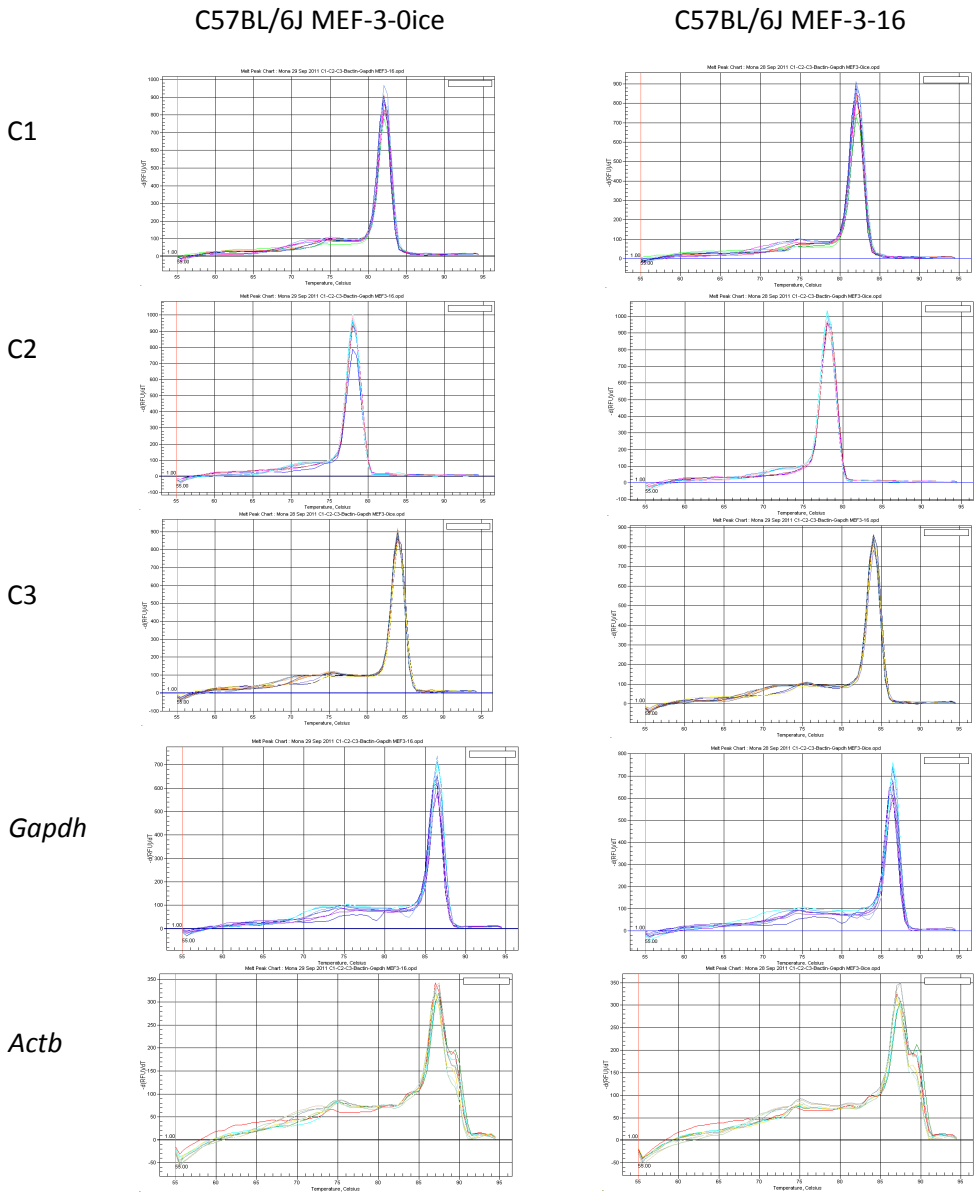


Figure 3-25 - Melting curve analysis of DHS positive and negative control SYBR green qPCR primer pairs on untreated and moderately DNase I treated C57BL/6J MEF templates. All three negative control primer pairs (C1, C2, and C3) have narrow peaks indicating of specificity. Between DHS positive primer pairs (*Gapdh* and Actb), beta-actin has a wide peak, indicating lack of specificity of this primer pair.

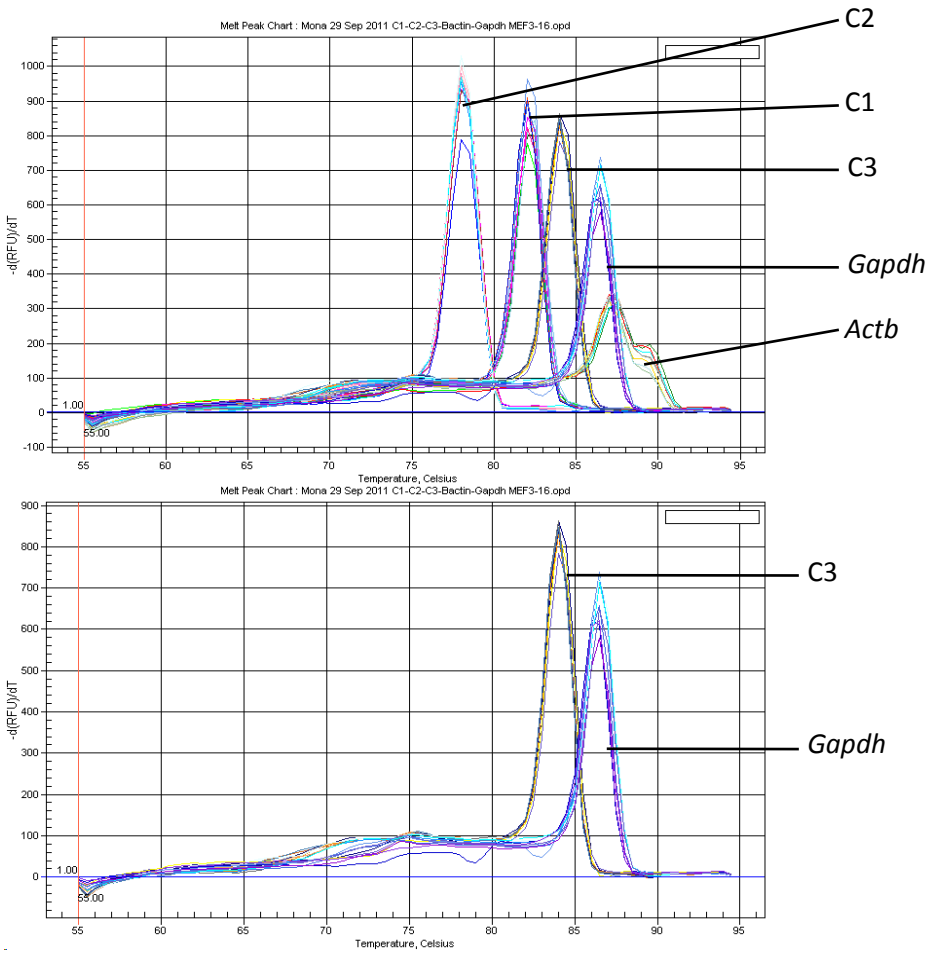


Figure 3-26 - Comparing melting curve peaks of SYBR green qPCR primer pairs on DNase I treated template (C57BL/6J MEF-3-16). C3 and *Gapdh*, which have narrow peaks and peaks at close proximity chosen for further qPCR quantification of DNase I digestions.

Table 3-4 - Summary of SYBR green qPCR DHS positive and negative primer pairs' efficiencies on untreated and moderately DNase I treated C57BL/6J MEF template

Template	Threshold	Primer set	4 dilution series (20, 10, 5, 2.5 ng/μl)			3 dilution series (20, 10, 5 ng/μl)		
			Efficiency (%)	R^2	Slope	Efficiency (%)	R^2	Slope
C57BL/6J MEF-3-0ice	611.17	C1	90.7	0.979	-3.57	104.3	0.968	-3.22
		C2	88.0	0.992	-3.65	96.3	0.987	-3.41
		C3	91.2	0.988	-3.55	101.6	0.978	-3.28
		Gapdh	46.6	0.948	-6.02	49.0	0.875	-5.77
		Actb	75.1	0.989	-4.11	78.9	0.979	-3.96
C57BL/6J MEF-3-16	616.79	C1	88.6	0.957	-3.63	105.0	0.971	-3.21
		C2	97.8	0.977	-3.38	100.9	0.943	-3.30
		C3	96.2	0.980	-3.42	100.1	0.953	-3.32
		Gapdh	70.7	0.934	-4.31	82.0	0.844	-3.84
		Actb	79.1	0.992	-3.95	73.3	0.988	-4.18

Figure 3-27 shows the profile of C3 and *Gapdh* primer pairs on untreated (C57BL/6J MEF-3-0ice) vs moderately DNase I treated template (C57BL/6J MEF-3-16). The 0 U DNase I treated template amplifies ~1 PCR cycle prior to 160 U DNase I treated template with *Gapdh* (DHS positive primer pair). Almost less than half of this effect was observed with C3 (DHS negative) primer pairs.

DNase I digestion was quantified on C57BL/6J MEF 1-3 DNase I assays by qPCR using C3 as negative control primer pair and *Gapdh* as a DHS positive primer pair. Standard curves for both primer pairs (*Gapdh* and C3) were included in each qPCR plate. 5 ng of template was loaded on each well. Starting quantities were estimated by Bio-Rad iQ5 Optical System Software for increasing quantities of DNase I in each DNase I assay. Figure 3-28 shows *Gapdh*, C3 quantities and *Gapdh*/C3 ratios in three MEF DNase I assays. *Gapdh*/C3 rations

used to select mid-phase of DNaseI digestion for Solexa sequencing compatible library preparation from MEF1-3 DNase I assays. Details of *Gapdh* and C3 efficiencies, threshold values and *Gapdh*/C3 ratios of selected phases of DNase I assay selected for library preparation are summarized in Table 3-3.

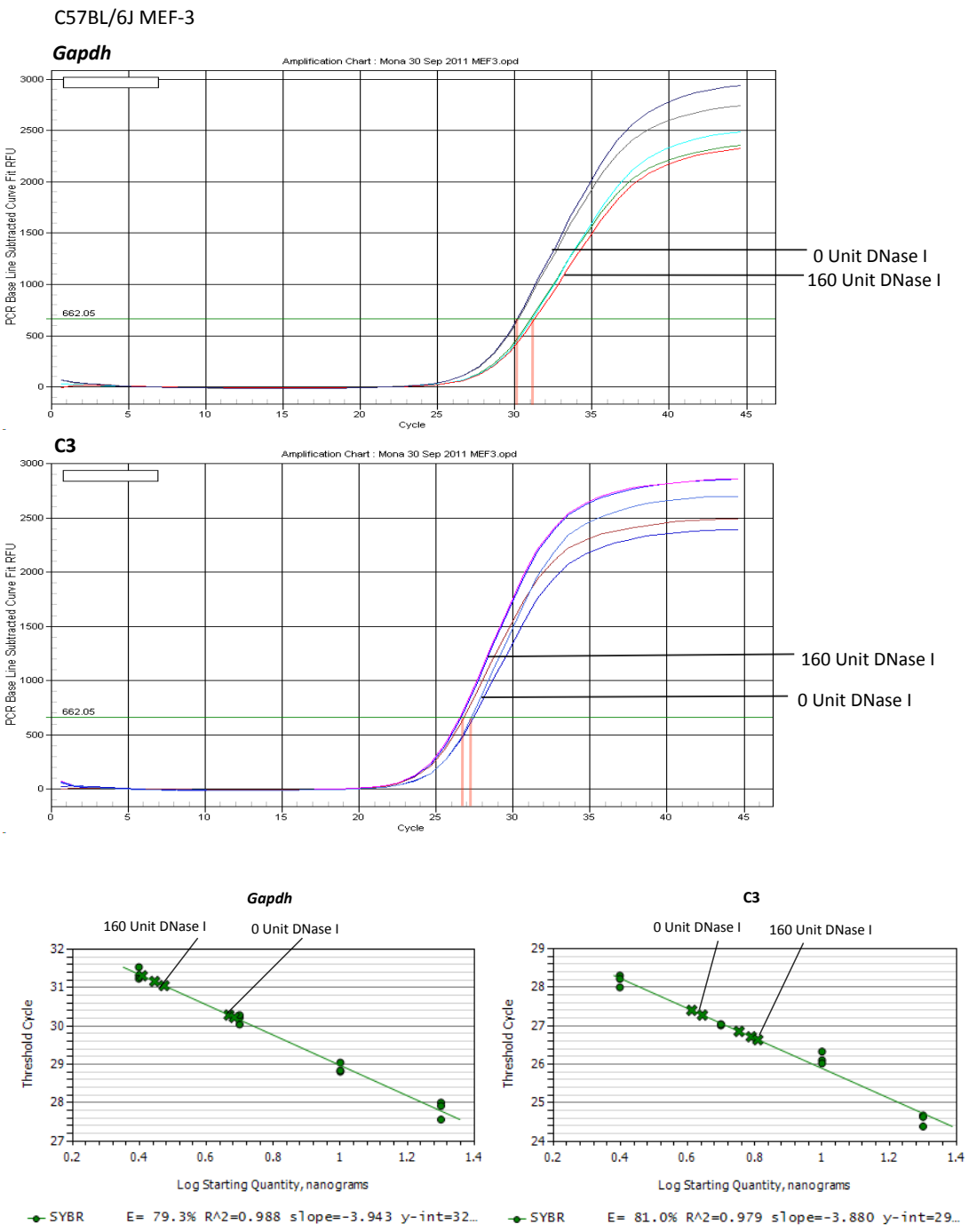


Figure 3-27 - SYBR-green qPCR profiles of *Gapdh* and C3 primer pairs. Standard curves of 2-fold dilutions of C57BL/6J MEF-3-0ice DNA template of both primer pairs are generated. Amplicon curves show florescent vs cycle number. Threshold was set at 662.05. Efficiency of both primers are comparable. Untreated template amplifies ~1 cycle prior to moderately DNase I treated template with DHS positive primer pair (*Gapdh*); however, the difference between Ct values of untreated and moderately treated template with C3 primer pair is significantly less (~ less than 0.5 cycle).

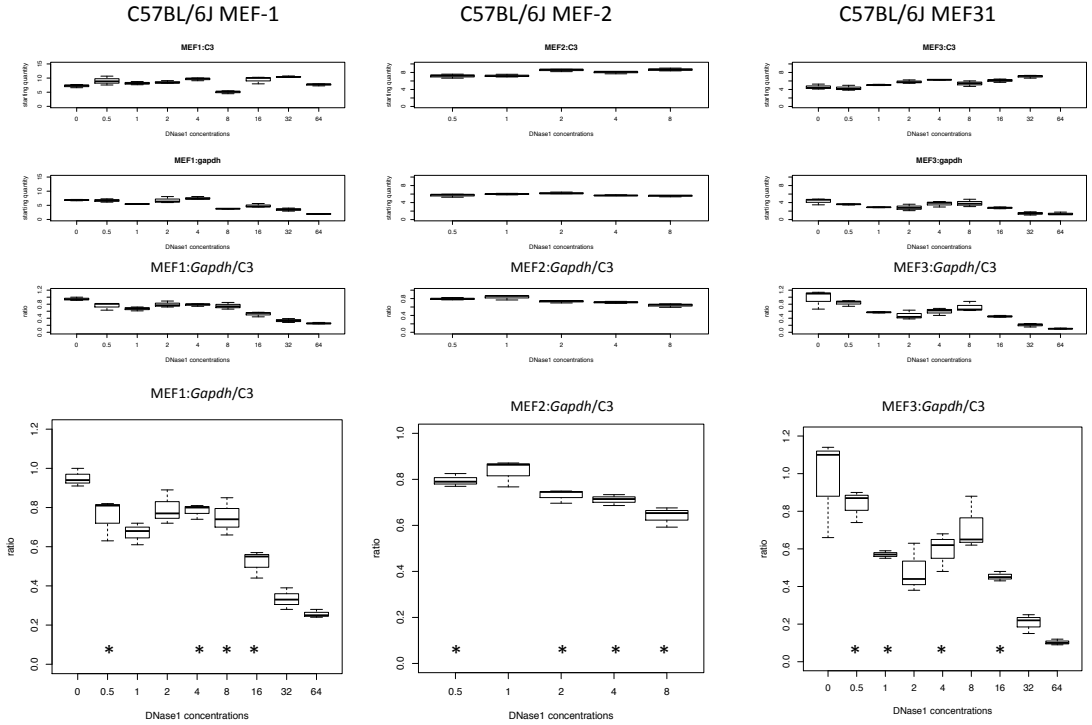


Figure 3-28 - Quantification of DNase I digestion by qPCR in C57BL/6J MEF DNase I assays. Each column represents one biological replicate. The first and second row in each column represent estimated starting quantities of the template (Y axis) in samples with increasing titration of DNase I (X axis) for DNase I negative and positive primer pairs respectively (i.e. C3 and *Gapdh*). Different concentrations of DNase I labelled by the volume of 10 U/µl DNase I used per 1 ml of reaction volume (0.5, 1, 2, 4, 8, 16, 32, 64). Absences of some phases of digestion in replicate 2 was due to inadequate number of cells for complete set of phases as described earlier. The third row shows *Gapdh*/C3 ratio. Final row is the magnification of *Gapdh*/C3 ratios to enhance the details. Stars show selected phases for library preparations.

3.6 SUMMARY

In this chapter I have developed a DNase-seq protocol based on an existing DNasechip protocol to identify regulatory elements on a genome wide scale. I used human erythroid immortalized cell line (K562) and mouse primary erythroid cells to develop this protocol. The efficiency of this method was assessed by comparing the DNase chip results with known regulatory sites at alpha and beta globin loci. I used qPCR to quantify DNaseI digestion. Mid phase of DNase I digestions were used for library preparations. Individualized assays designed for erythroid and fibroblast tissue. I generated DNase-seq libraries for terminally differentiated erythroid cells from 8 inbred mouse strains and C57BL/6J MEFs. I included replicates for each strain to be able to assess within strain variations and compare them with between strain variations.

CHAPTER 4 : DNASE-SEQ DATA ANALYSIS

4.1 INTRODUCTION

In this chapter, I will present analysis of DNase I hypersensitivity data obtained from the eight inbred mouse strains to study genomic regulatory elements in eight given mouse strains. In addition to data from erythroid tissue, I also had data from two strains from mouse embryonic fibroblasts. Initial analyses were aimed at determining the quality of the data, to assess the amount of variation both within and between strains, and to determine if there were observable differences between tissues, which would permit an assessment of the number of DNase I hypersensitive sites that were specific to erythroid tissue.

Before proceeding to a genome-wide analysis, I needed to establish that the data were of high quality, and that there were at least some positions where variation between strains in DNase I hypersensitive sites was detectable. To achieve my aims of assessing quality and variation, I relied upon pre-existing information about the regulatory regions surrounding the globin genes. I set out to inspect these, and other regions in the genome. The data presented in this chapter is largely qualitative in nature. The following chapter provides a comprehensive quantitative description of variation in in DNase I hypersensitive sites.

4.2 DNASE I SEQUENCING DATA SETS

I start by presenting quality control data for all the sequences we obtained. These results establish that the sequence was of good quality. I briefly discuss these quality scores, before turning to a more detailed discussion of the datasets I obtained.

The DNase I data sets included 7 GAII single end data sets (A/J-3, AKR/J-1, BALBc/J-3, C3H/HeJ-1, C57BL/6J-1, CBA/J-2, and DBA/2J-1) and 14 HiSseq pair end data sets (A/J-1, A/J-2, AKR/J-3, BALBc/J-2, C3H/HeJ-2, C57BL/6J-2, C57BL/6J-3, CBA/J-1, CBA/J-6, DBA/2J-2, LP/J-3, LP/J-4, C57BL/6J MEF-1, and C57BL/6J MEF-3).

GAII data sets were sequenced to an average depth of 30 million uniquely mapping sequencing tags (51 bp) per sample. For HiSeq data sets these value was ~ 159 million mapping sequencing tags (51 bp) per replicate. Table 4-1 and 4-2 summarize sequencing details of DNase I libraries on GAII and HiSeq platforms, including number of reads, mappable percentage, quality measures (Q20), and library insert sizes.

Base quality scores were assessed using a *Phred* base-specific scoring. This score is an accurate method to assign accuracy to each base call in DNA sequencing. It ranges from 4-60 and higher values indicate higher quality. *Phred* quality score 20 (Q20), for instance, shows the probability of a wrong base call in sequencing is 1 in 100. In other words, the accuracy of base call in this case is 99%. Commonly used Q20 counts bases with *Phred* quality score 20 and above, which sometimes referred to as “high quality bases”. Q30 includes base calls with quality score of 30 and above (probability of wrong call is one in 1000 or

the base call accuracy is 99.9%) (Ewing et al., 1998). Table 4-1 and 4-2 shows Q20 in our data sets.

Sequence alignment was performed by using STAMPY (Lunter and Goodson, 2011) to align reads from each data set to pseudo genomes created from each of eight strains by incorporating SNPs from relevant strain into MGSC37 mouse reference (see Methods: sequence alignment section). The data filtered for mapping quality and repetitive regions.

Table 4-1 - Summary of sequencing quality features (GAII). DNase-seq data have been mapped to the reference genome (C57BL/6)

Sample		Length (bp)	Yield (Million read)	Yield (Mb)	Yield (Mb Q20)	Mapped to C57BL/6 (%)
GAII	A/J-1	51	38.23	1949.50	1888.23	77.9
	AKR/J-1	51	41.73	2128.33	2015.44	90.6
	BALBc/J-1	51	30.78	1569.54	1497.55	90.9
	C3H/HeJ-1	51	31.22	1592.10	1525.72	75.2*
	C57BL/6J-1	51	32.32	1648.24	1563.76	63.2*
	CBA/J-1	51	36.93	1883.18	1830.39	77.3
	DBA/2J-1	51	32.89	1677.32	1627.20	76.6

* Unexpected low percentage of C3H/HeJ-1 and C57BL/6J-1 mapped to C57BL/6 (the 1st two sequencing data) were due to problems with the alignment pipeline and the raw data were realigned prior to further analysis from these two strains after amending the pipeline.

Table 4-2 - Summary of sequencing quality features (HiSeq)

Sample		Length (bp)	Insert ±MAD*	Yield (Million read)	Yield (Mb)	Yield (MbQ20)	Mapped to BL6 (%)
HiSeq	A/J-1	51	126 ± 60	165.37	8433.91	8182.63	86.4
		51	127 ± 59	165.37	8433.91	7918.38	85.4
	A/J-2	51	150 ± 60	202.92	10348.94	10081.54	96.4
		51	150 ± 59	202.92	10348.94	9411.67	93.8
	AKR/J-3	51	123 ± 43	202.29	10316.90	9902.02	96.0
		51	123 ± 43	202.29	10316.90	8940.22	92.1
	BALBc/J-2	51	159 ± 50	113.13	5769.72	5662.03	96.5
		51	159 ± 50	113.13	5769.72	5536.78	95.5
	C3H/HeJ-2	51	126 ± 51	146.42	7467.59	7101.95	93.2
		51	127 ± 51	146.42	7467.59	6706.05	90.4
	C57BL/6J-2	51	159 ± 74	67.94	3464.84	3386.30	90.2
		51	160 ± 73	67.94	3464.84	3297.29	89.4
	C57BL/6J-3	51	150 ± 68	85.28	4349.46	4232.31	87.5
		51	151 ± 67	85.28	4349.46	4130.60	86.5
	CBA/J-1	51	168 ± 64	119.72	6105.92	5997.35	90.4
		51	168 ± 63	119.72	6105.92	5798.62	89.3
	CBA/J-6	51	149 ± 52	184.76	9422.98	8896.48	96.1
		51	148 ± 52	184.76	9422.98	7192.33	86.3
	DBA/2J-2	51	162 ± 61	193.16	9851.07	9555.90	93.8
		51	162 ± 60	193.16	9851.07	8968.54	91.7
	LP/J-3	51	145 ± 60	166.75	8504.18	8251.55	90.2
		51	145 ± 60	166.75	8504.18	7770.53	88.2
	LP/J-4	51	131 ± 51	179.18	9138.25	8785.48	95.7
		51	131 ± 51	179.18	9138.25	8383.77	93.7
	C56BL/6J MEF-1	51	119 ± 39	197.14	10053.91	9611.30	89.3
		51	119 ± 39	197.14	10053.91	8617.67	86.3
	C57BL/6J MEF-3	51	122 ± 42	206.51	10531.92	10025.99	94.7
		51	122 ± 41	206.51	10531.92	9190.10	92.8

*MAD stands for median absolute deviation (a robust estimator of variance)

4.3 SEQUENCING RESULTS FROM C57BL/6J-1 AND C3H/HEJ-1

I decided to begin my experiments by sequencing hypersensitive sites for just two strains C3H/HeJ and C57BL6/J. I wanted to ensure that I could generate good quality data before embarking on sequencing hypersensitive sites for the remaining six strains.

I size selected fragments (200-500 bp) according to a standard protocol, and sequenced the libraries on an Illumina genome analyzer (GAII platform). Single end reads were mapped back to the reference genome (C57BL/6J, build 37). Those regions of the genome to which multiple reads map, above background levels, are expected to indicate the position of DHS sites. To investigate the sensitivity and specificity of our data I compared my DNase I seq data from the first two strains at two well-studied gene loci: alpha and beta globin loci where regulatory elements have been studied accurately by a gold standard Southern blotting technique. Comparing peaks with known regulatory elements in the well-studied alpha and beta globin loci revealed that the quality of the DNaseI hypersensitive sites produced by DNase-seq is high and reproducible (Figure 4-1). The names of previously annotated DNaseI hypersensitive sites (DHS) are shown colored as ubiquitous (blue), erythroid (red) and black (CTCF)(Kowalczyk et al., 2012). DHS-peaks are aligned with the position of previously described DHS sites by vertical lines, color- coded as described. Two previously unpublished DHS (marked by *) are associated with CTCF binding sites.

The figure is notable in showing that my experiment identified all known DHS sites. In other words, at this locus, my assay has achieved 100% sensitivity

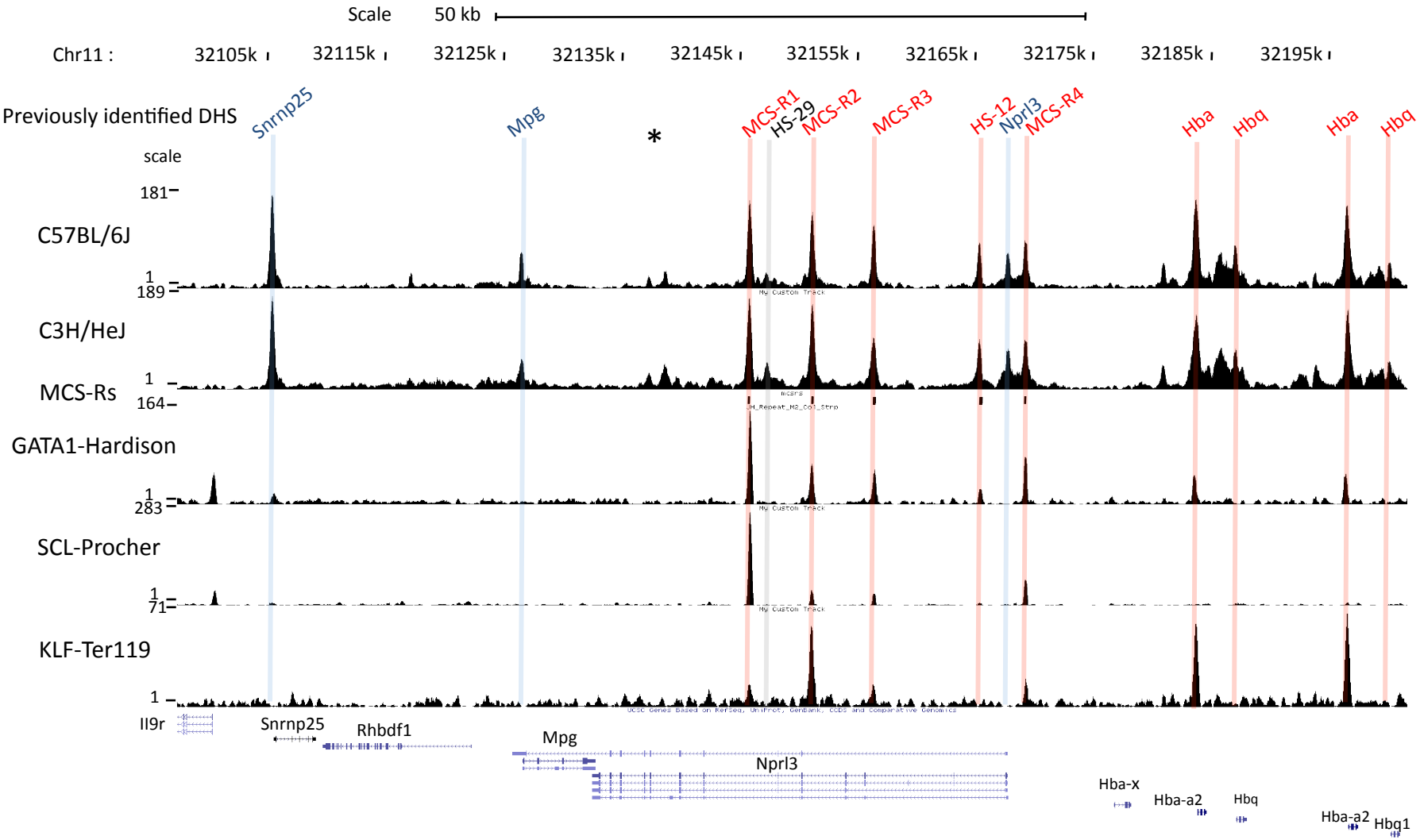


Figure 4-1 A snapshot of C57BL6/J and C3H/HeJ DNaseI-seq data over the alpha globin locus. The top two tracks are the DNaseI data in the two strains (C57BL6/J and C3H/HeJ respectively) then the annotation for the globin *cis* elements; then the three tracks of erythroid transcription factor ChIP-seq (GATA1, SCL and KLF) and UCSC/Refseq gene annotations.

and very high specificity. No false negative and false positive at these loci, indicating the high quality of my data.

The next question I addressed was whether DHS varies between C57BL/6J and C3H/HeJ. Figure 4-2 – 4-5 shows a few examples of how the data looked like at various loci. Overall, visual analysis of C3H/HeJ and C57BL/6J indicates: at many sites the DHS pattern was similar in both strains (e.g. figures 4-2 and 4-3); however, differential DHS observed in various sites (e.g. figures 4-4 and 4-5).

Figure 4-2 shows a snapshot of DNase-seq C57BL/6J-1 and C3H/HeJ-1 GAII data sets on chromosome 2 (3,188,167-3,437,637 bp), which represent how the data looks like at many loci, there are spiky peaks over a clear background signal which are easily identified by visual inspection at both strains and therefore easy to identify variation of active regulatory elements.

As you can see in this figure, all the obvious DHS peaks are aligned with the promoters of genes in this locus. Figure 4-3 shows another example of reproducible results in both strains in beta globin locus, which is specifically active in haematopoesis. Known beta globin DHSs (HS-1-6) are shown by vertical red lines aligned with DNase-seq data. Previously published GATA1, erythroid transcription factor, ChIP-Seq data is included in this figure as an additional indication of active regulatory sites.

Figure 4-4 shows an example where there is a differential DHS between C57BL/6J-1 and C3H/HeJ-1 on chromosome 8 (87,112,238-87,112,814bp). Among three adjacent DHS peaks, the most prominent one is absent in C3H/HeJ. Previously published SCL (erythroid transcription factor) ChIP-Seq data is included in the figure to provide additional evidence on active erythroid

regulatory elements. A clear peak on SCL data at the same position of this peak (vertical blue dashed line) indicates this DHS peak in C57BL/6J belongs to active regulatory elements and suggests an explanation of the function of this element.

Figure 4-5 illustrates another example of variable DHS between C3H/HeJ and C57BL/6J on chromosome 19. In this case, multiple adjacent DHS peaks are absent in C57BL/6J (32,416,859-32,465,363bp), marked by stars within the yellow box. Previously published GATA1 (erythroid transcription factor) ChIP-Seq data is included in this figure; suggests an explanation for the presence of DHS in C3H/HeJ and indicates that the peaks in C3H/HeJ are indeed real peaks.

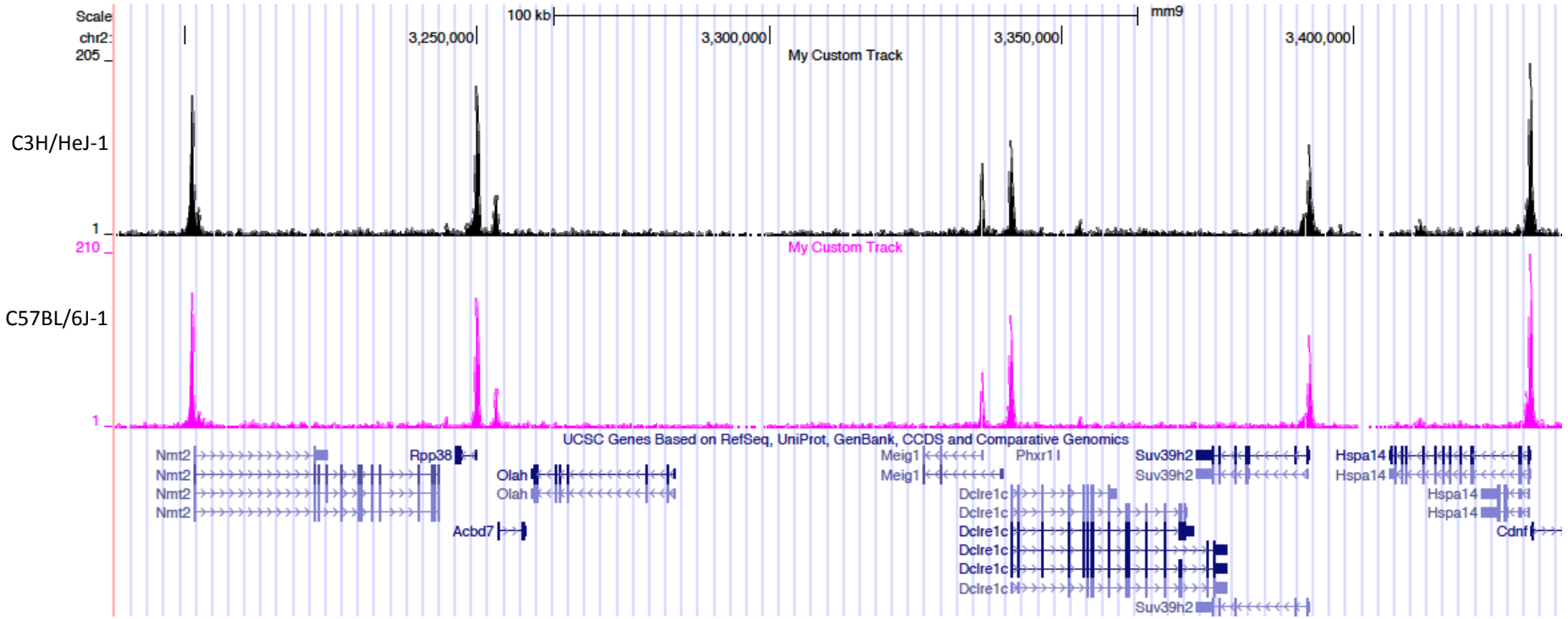


Figure 4-2 - A snapshot of C57BL/6J-1 and C3H/HeJ-1 DNase-seq GAII data on chromosome 2. Clear spiky DHS reproducible peaks observed in both strains. There is no variation in DHS in this locus between the two strains. Overall DHS peaks are aligned with promoters of UCSC genes shown in blue. Presented DNase-seq data have been mapped to the reference genome (C57BL/6) and C57BL/6J-1 DNase-seq data is presented in pink.

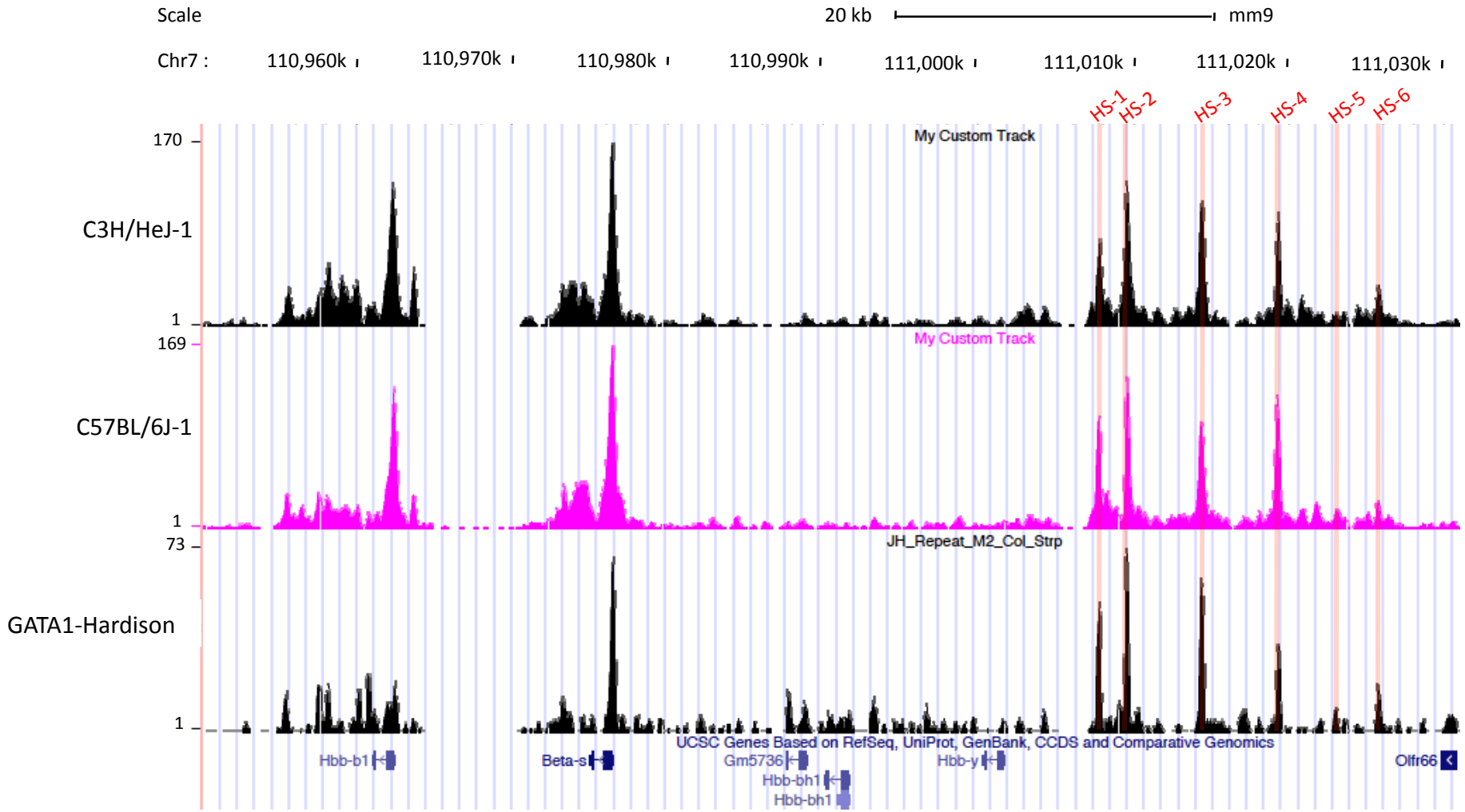


Figure 4-3 - A snapshot of C57BL/6J and C3H/HeJ DNaseI-seq data over beta globin locus on chromosome 7. C3H/HeJ-1 and C57BL/6J-1 DNase-seq GAII data followed by previously published GATA1 ChIP-seq data and UCSC gene annotations. Vertical red lines display known beta globin regulatory hypersensitive sites (HS-1-6). DNase-seq data presentation is as described in Figure 4-2.

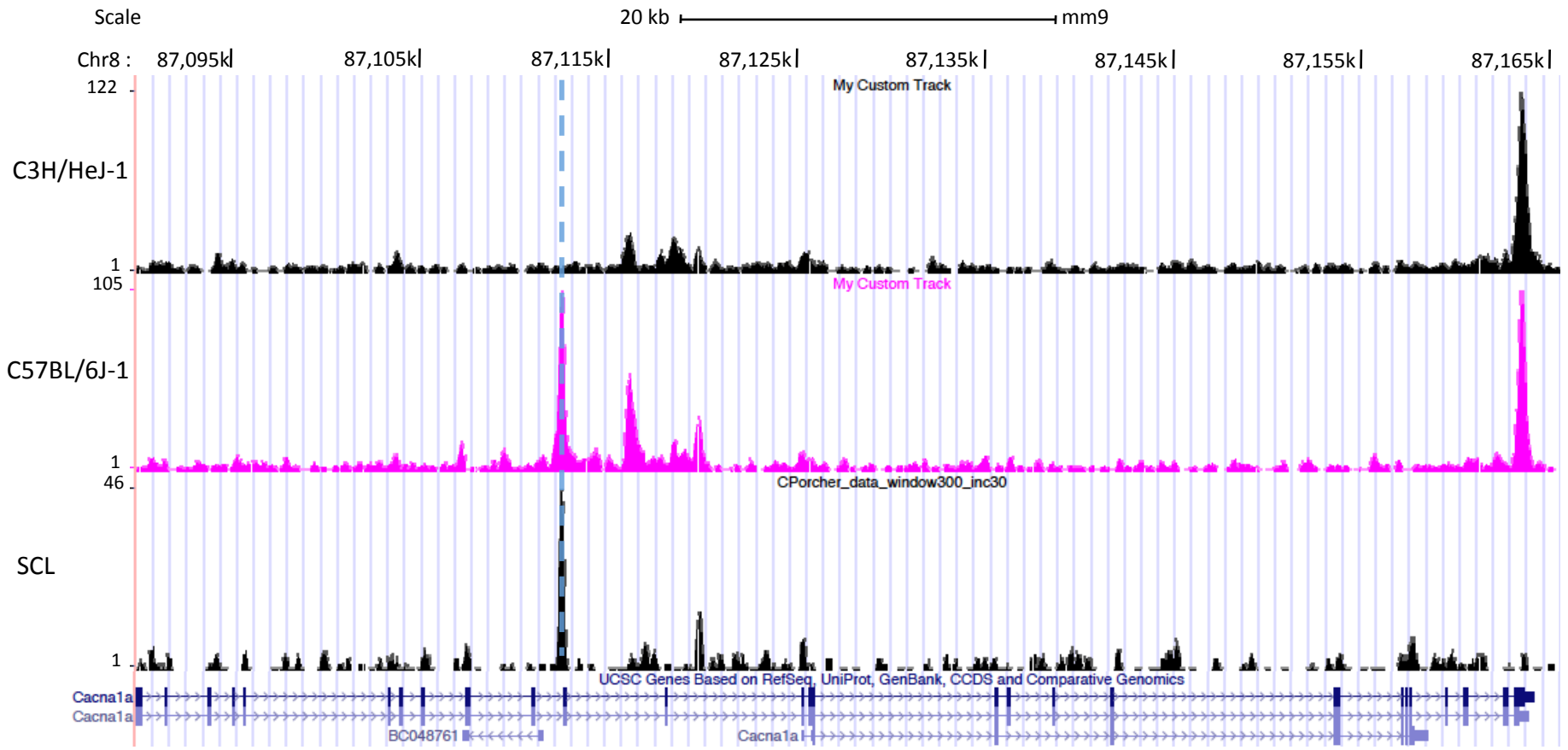


Figure 4-4 - An example of differential DHS between C3H/HeJ-1 and C57BL/6J on chromosome 8. An obvious DHS peak (blue dashed line) is absent in C3H/HeJ-1. SCL ChIP-Seq data included providing further evidence on transcriptionally active regulatory elements. DNase-seq data presentation is as described in Figure 4-2

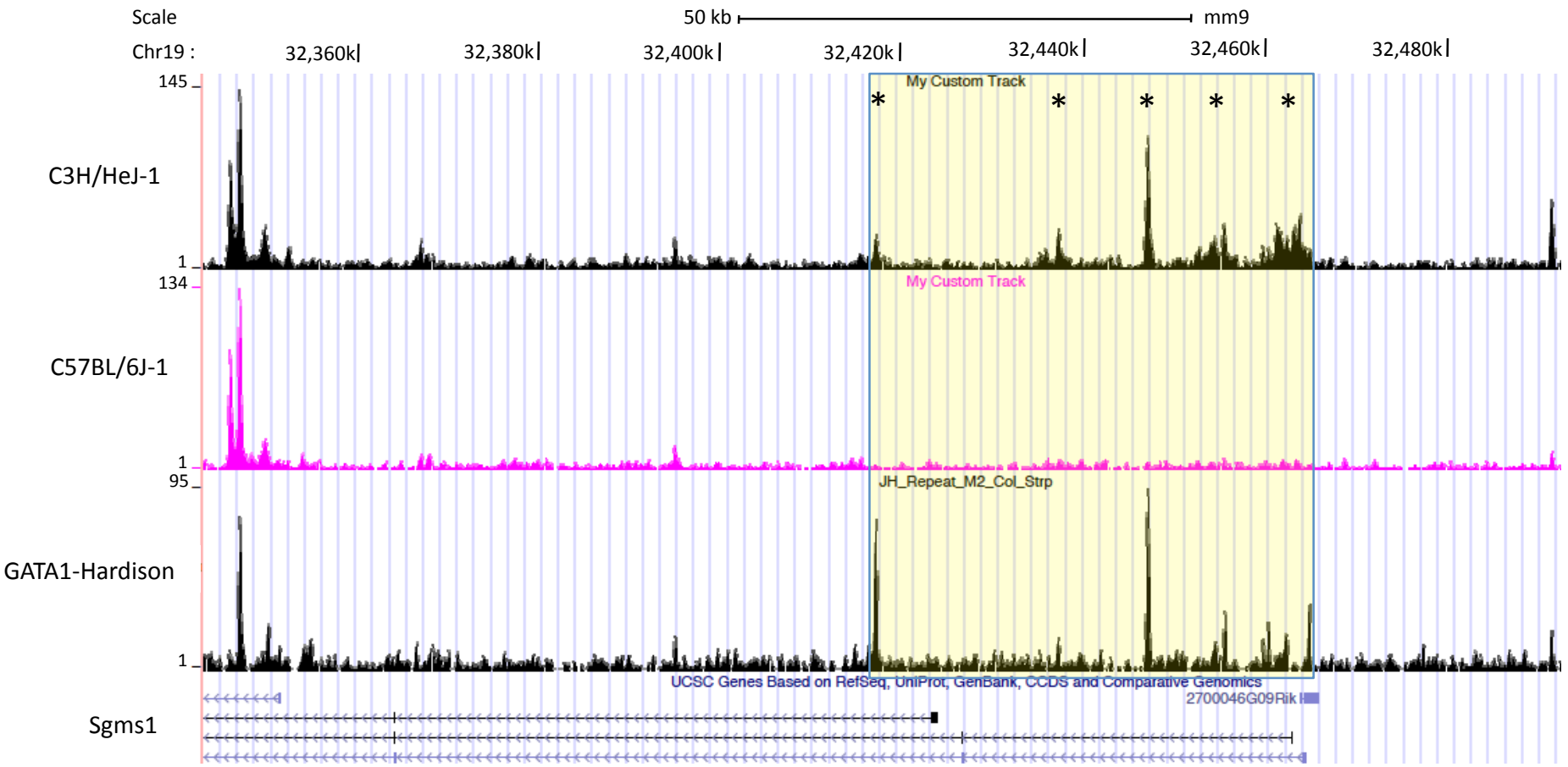


Figure 4-5 - An example of differential DHS between C3H/HeJ-1 and C57BL/6J on chromosome 19. Multiple DHS peaks in C3H/HeJ-1 are absent in C57BL/6J-1 (starred within yellow box). Previously published GATA1 ChIP-seq data provides additional evidence on active erythroid regulatory elements. DNase-seq data presentation is as described in Figure 4-2.

4.4 DNASE-SEQ DATA ON GAII PLATFORM

Given the encouraging preliminary results of C57BL/6J-1 and C3H/HeJ-1, I prepared libraries from the other 6 inbred strains. DNase-seq data was obtained from all other strains except LP/J, which failed QC for sequencing by GAII platform. I again addressed the sensitivity and specificity of the DHS data on GAII platform to determine if it was as good as that observed for C3H/HeJ and C57BL/6J. Visual inspection of GAII data showed highly comparable reproducible data for 7 inbred strains. Figures 4-6 and 4-7 show GAII data at alpha and beta globin loci of 7 inbred strains. All the known regulatory sites produced a clear signal in all sequenced inbred strains.

Overall CBA had the noisiest background compared to other strains. BALBc/J, C3H/HeJ, C57BL/6J, and DBA/2J had the lowest background noise and A/J and AKR/J were in between. In retrospect, I looked at qPCR analysis and the selected DNaseI digestion phases of CBA/J, which I made the DNase-seq library from, and I realized why we had the noisiest background for CBA/J. There was an indication of over-digestion on 4^{th} phase I used for the library preparation. Therefore, I prepared a new DNase-seq library from modified selected digestion phases of this sample; however this failed sequencing QC and ultimately has not been sequenced.

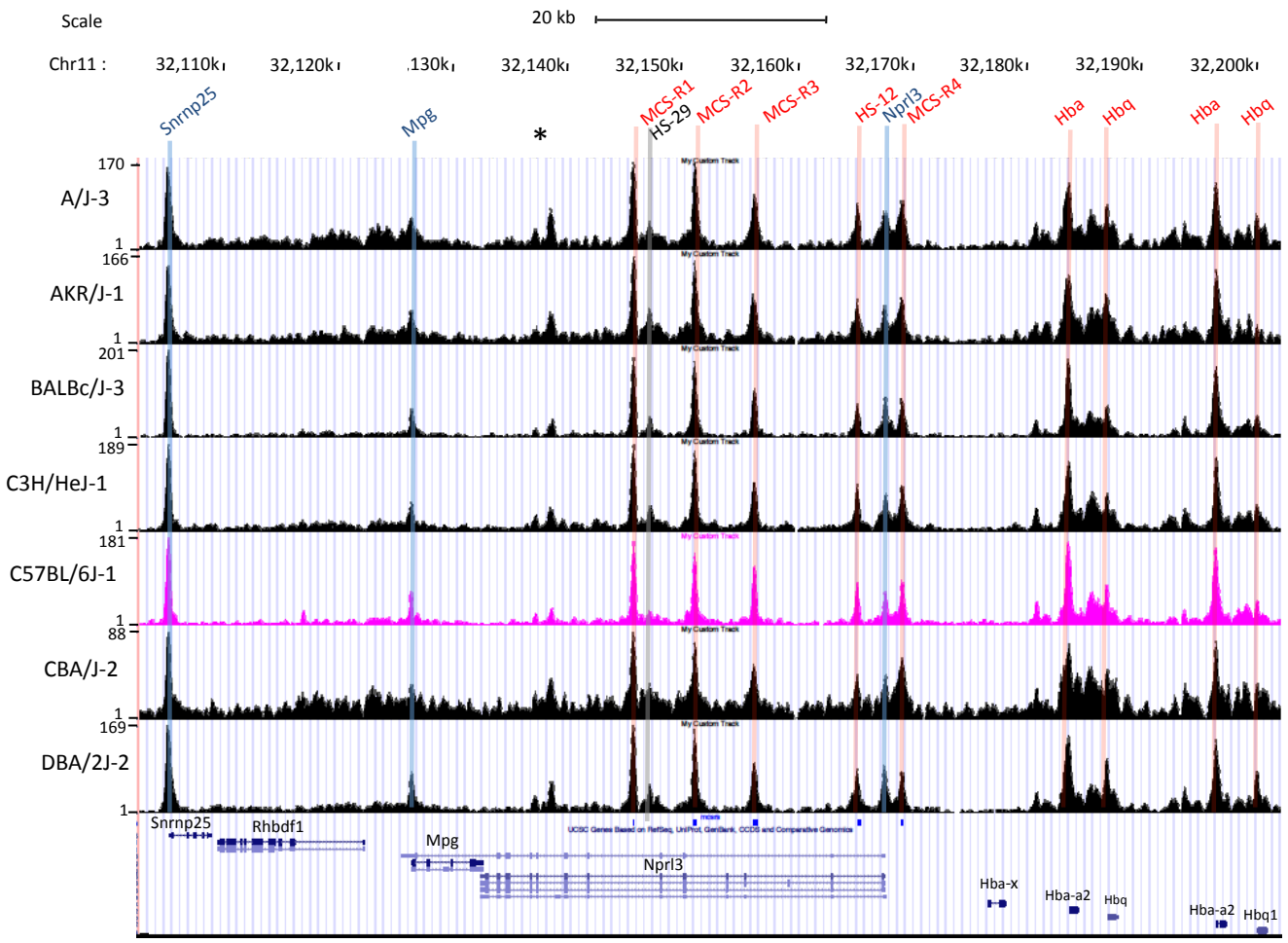


Figure 4-6 - GAII DNase-seq data of seven inbred strains at alpha globin locus. DNase-seq data presentation is as described in Figure 4-2.

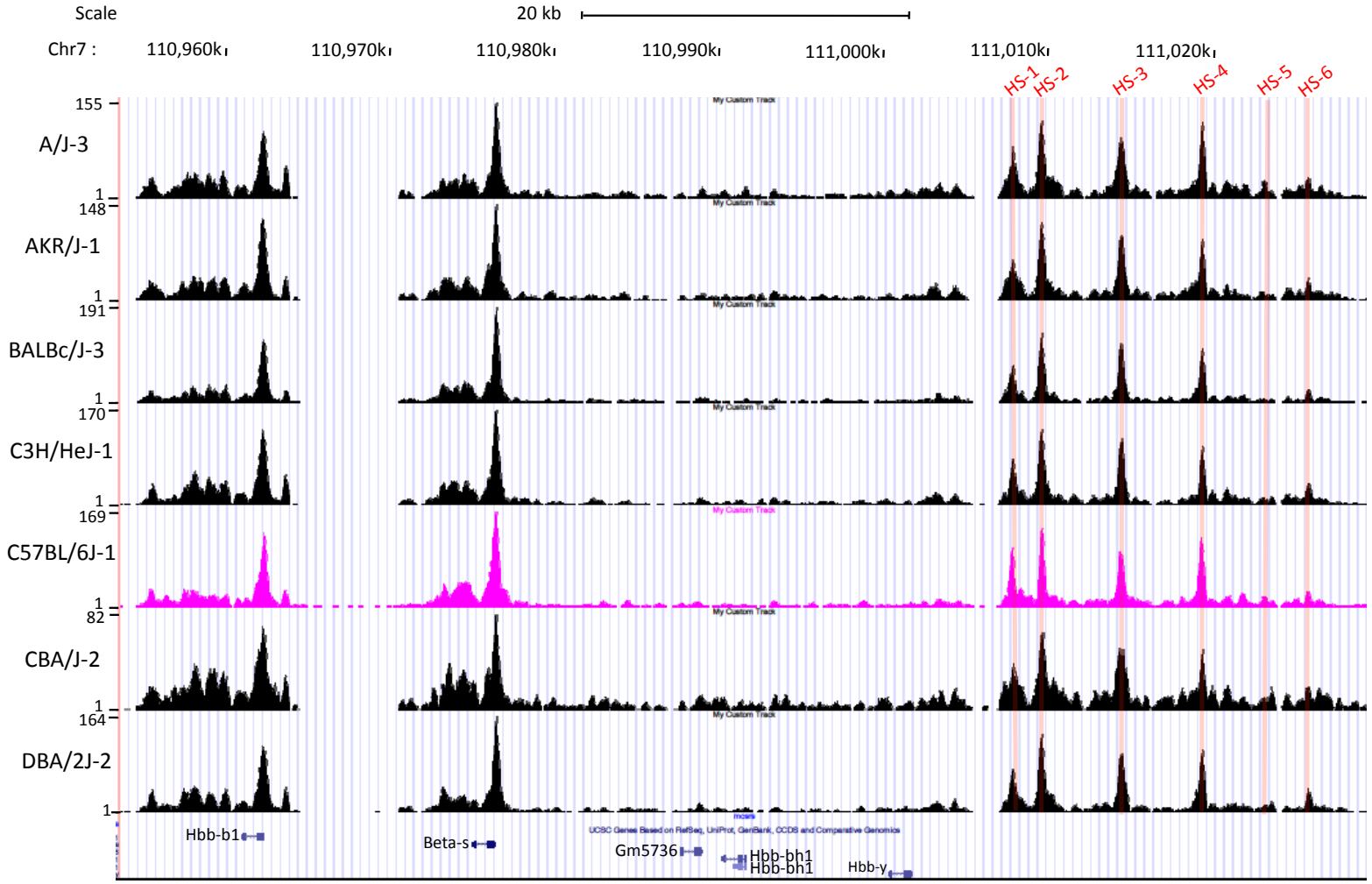


Figure 4-7 GAII DNase-seq data of seven inbred mouse strains on beta globin locus. DNase-seq data presentation is as described in Figure 4-2.

4.4.1 DIFFERENTIAL DHS PEAKS WITH GAII PLATFORMS

The subsequent question was to ask whether we could find examples of differential DHS peaks in different inbred strains. Two clearly defined categories were when single or multiple adjacent peaks were gained or lost in one or more strains. The feature observed in more than a half of peaks was the base of calling a variable DHS as gain or loss. For instance of a DHS peak has been observed in minority of strains (among 7 GAII DNase-seq data), it defined as "gain" and if it has been observed in majority of strains, it defined as "loss".

Figure 4-8 and 4-9 represent two examples of variable DHS peaks in 6 inbred strains we obtained GAII DNase-seq data from. BALBc/J DNase-seq GAII data does not shown in these figures but the peaks reproduced what we observed in the majority of peaks (i.e. no variable peaks has been observed in BALBc/J strain in these loci).

Figure 4-8 shows an example of a "gain": emergence of a single spiky DHS peak on C57BL/6J on chromosome 19 (5,405,866-5,407,652 bp). However it is not always easy to decide whether a DHS peak is present at a locus in visual analysis. For example, in this case, the DHS peak (aligned with a light blue vertical line) is present in C57BL/6J and is clearly absent in A/J, C3H/H3J, DBA/2J, and BALBc/J; however, it is not definitely clear in CBA/J and AKR/J whether there are small DHS peaks in this locus or tiny spikes are artifacts raised by noisier background (marked by "?").

In order to determine whether a peak in the sequence data represented a true DHS, I incorporated chromatin marks that are associated with regulatory elements in the genome. Known DHS sites can broadly be divided into three types: enhancers, promoters and CTCF binding sites. Enhancers are

characterized as regions enriched for H3K4Me1 and promoters for H3K4me3 enrichment. I used H3K4me3 and H3K4me1 ChIP-Seq data from terminally differentiated erythroblasts (C57BL6/J Ter119+ cells), and CTCF ChIP-seq data of C57BL6 Ter119+ cells. .

The variable DHS peak in figure 4-8 aligns with the promoter of BC085508 UCSC (Mus musculus RIKEN cDNA 4930481A15 non-coding gene). High H3K4me3 and low H3K4me1 ChIP-Seq signal at this locus indicates this peak is a meaningful signal and functions as an active promoter regulatory element.

Figure 4-9 shows an example of a different type of variable DHS peaks: "loss" of multiple DHSs on chromosome 19 (32,416,859-32,465,363bp). Five adjacent peaks are only absent in C57BL/6J and present in all other six inbred strains (BALBc/J DNase-seq GAII is not shown here but it contains all the peaks). This example was shown in figure 4-5 in C3H/HeJ and C57BL/6J, here we can see the DNase-seq data from the other GAII data sets.

Figure 4-9 includes USCS gene annotations and chromatin marks (H3K4me3 and H3K4me1 ChIP-Seq data of terminally differentiated erythroblasts (C57BL/6 Ter119+ cells) as well as previously published GATA1 ChIP-Seq data, as an indication of erythroid active regulatory elements. There is no obvious peak at H3K4me3 and H3K4me1 lanes suggesting there is no promoter and enhancer activity at these regulatory sites for the given strain (C57BL/6) which agrees with what expect from DHS pattern in C57BL/6J, but clear peaks at GATA1 lane suggest these peaks are meaningful signals and could indicate regulatory activity as aligned with GATA1 ChIP-Seq peaks indicating

being transcription factor binding sites. Therefore, ChIP-Seq data can be informative provided that regulatory site is open and active in given strain.

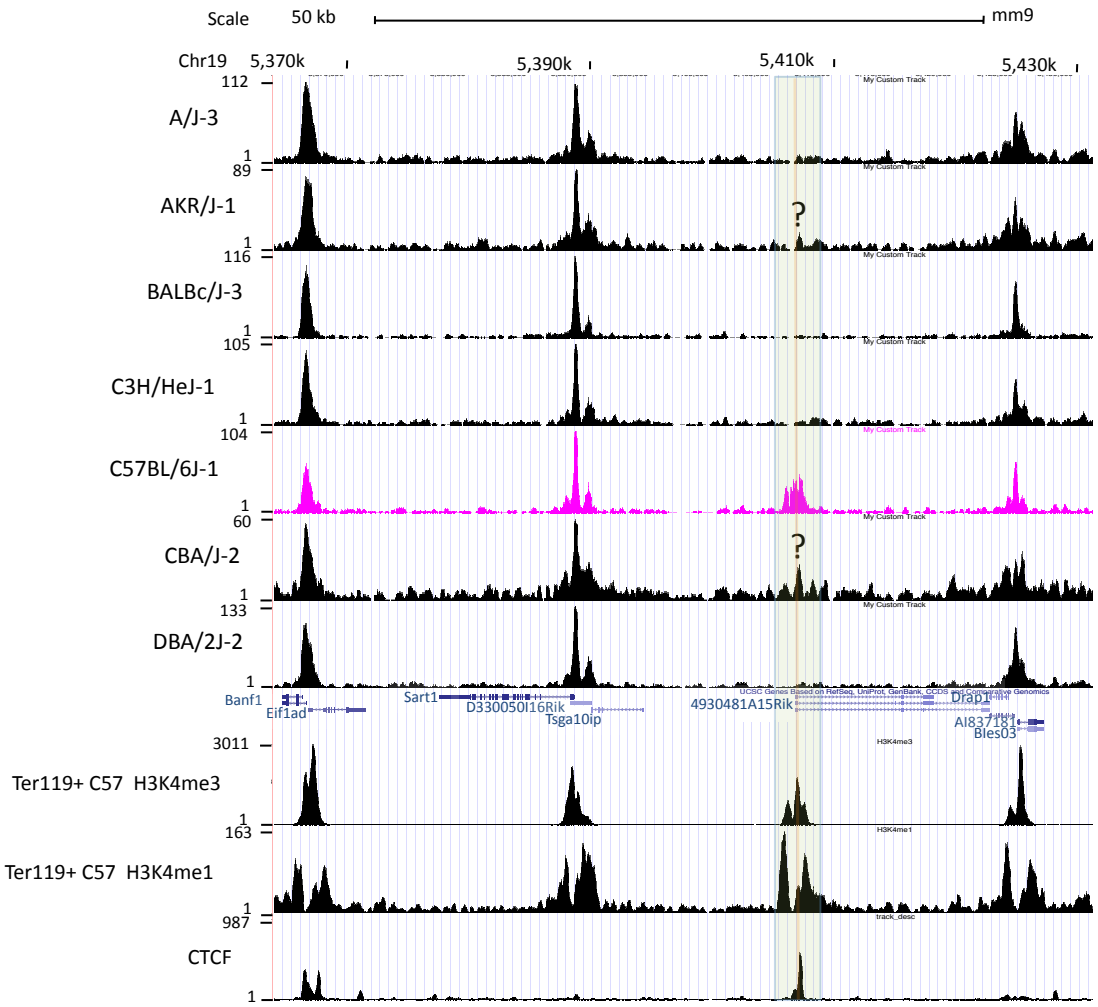


Figure 4-8 - An example of DHS differential peak in seven inbred strains (A/J, AKR/J, BLBC/J, C3H/HeJ, C57BL/6J, CBA/J and DBA/2J) over chromosome 19 (coordinates: 5,405,786-5,408,255 bp reference genome): DHS gain in C57BL/6J, small spiky protrusions also observed in CBA/J and AKR/J strains that could be due to less prominent DHS site. This DHS site also has chromatin signature for promoters and CTCF binding site and aligned with the promoter of 4930481A15Rik gene (the vertical orange line passes through the given promoter).

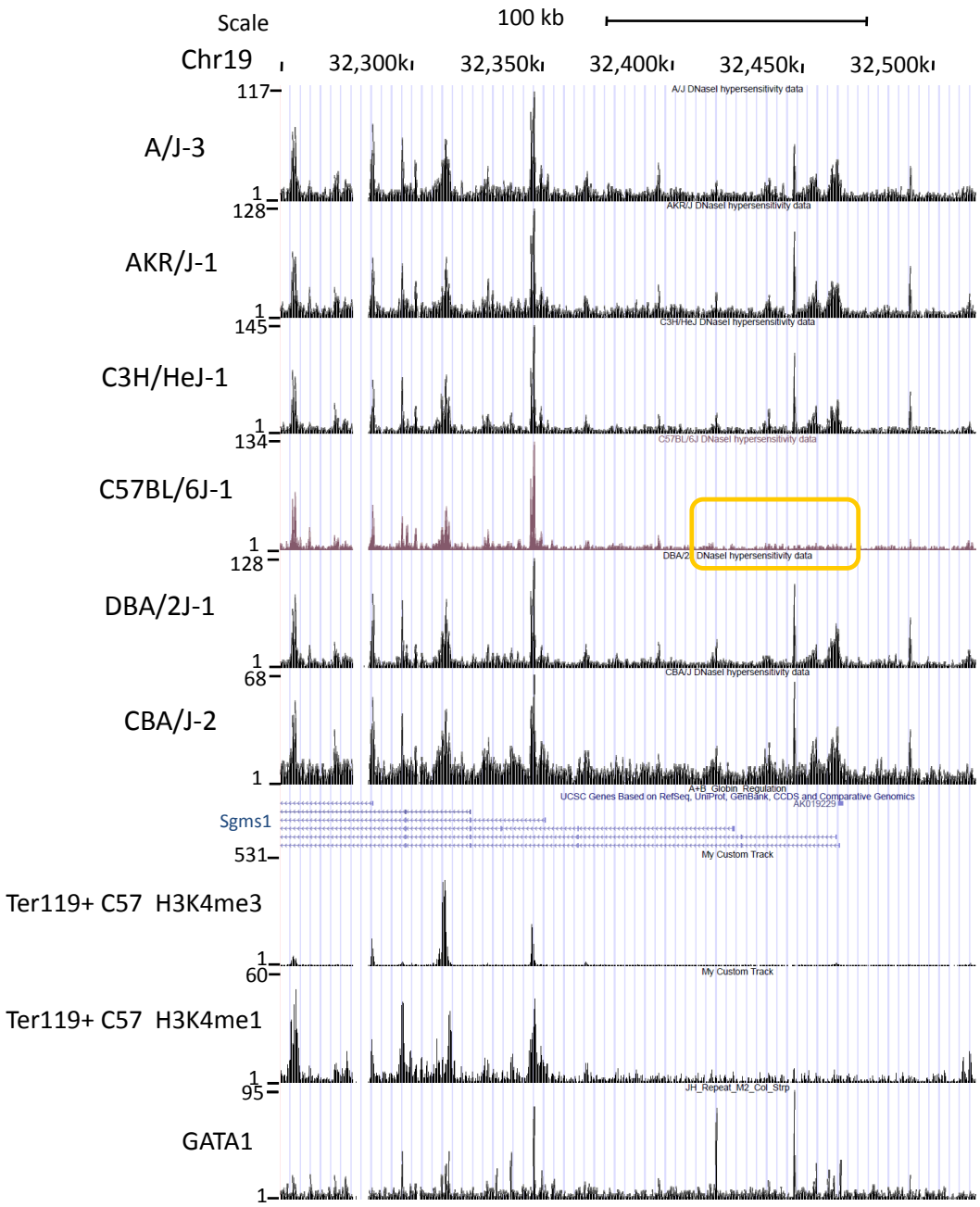


Figure 4-9 - An example of DHS differential peaks in six inbred strains (A/J, AKR/J, C3H/HeJ, C57BL/6J, CBA/J and DBA/2J) over chromosome 19: Multiple DHSs loss in C57BL/6J

4.5 WITHIN STRAIN VARIATION

My results so far indicated that the DNase-seq data are of high quality. I was able to detect known regulatory elements at the globin loci and elsewhere in the genome putative DHS peaks coincided with known chromatin marks of enhancers, promoters and CTCF binding sites. I had also detected regions where there are differences between strains. The quality of the data made it unlikely that these differences could be due to artifacts but it remained possible that they arose because of some environmental perturbation. In order to establish whether the variation has a genetic basis, I needed to compare data sets taken from the same strain. These had to be true biological replicates, not technical replicates. Therefore I set out to repeat the construction of libraries appropriate for the detection of DNase I hypersensitive sites. In the case of variable peaks, I could test the consistency and reliability of DNase-seq data by comparing the sequencing results of the replicates.

During the processing of the first set of sequencing data, our sequencing core obtained Illumina HiSeq machines. These are an upgrade of their sequencing platform, able to deliver longer reads, and more sequencing data. The use of this machine would increase the sensitivity of my analyses. Therefore I decided to run the new libraries I had prepared on the new machines, although this introduced a potential confound in comparisons with data obtained from the older GAII platform. I begin by describing the comparison between results from the two platforms and then turn to the analysis of within strain variation.

Tables 4-1 and 4-2 summarize the sequencing features and quality details of both GAII and HiSeq platforms. Overall, HiSeq generated more than five times more mapping sequencing tags (~ 159 vs ~ 35 million read) and almost five

times more average yield (MbQ20) (~1700 Mb vs ~7600 Mb with *Phred* quality score 20 and above) than GAII sequencing. Figure 4-10 summarizes what we could observe in detailed visual inspection of GAII vs HiSeq DNase-seq data. These are examples of three typical visual findings in comparing these two types of data sets. All these examples are selected from chromosome 1. In all three figures each color represents data from replicates of the identical strain and GAII data is depicted by star (*).

Figure 4-10a shows an example of false within strain variability. An absolutely clear peak identified by HiSeq platform is absent in all GAII data sets. Since LP/J is unique in having both replicates data sets on HiSeq platform if we visually compare LP/J data with other inbred strains GAII data sets we probably introduce many false positive variable peaks both in within strain and between strains domains. Figure 4-10b is another example of false positive within and between strains variable DHS. If we compare GAII data sets in seems that this DHS is absent or at least less obvious in CBA/J-2 (which is on GAII platform). However when we back up these GAII data sets with HiSeq replicates, it is quite obvious that this peak is present in all strains and when we look closely we could recognize the similar pattern within scarce reads on GAII platforms. Due to scarcity of the reads on GAII platforms at least for some inbred strains on this locus, it was initially picked up as a variable DHS examples. Including HiSeq DNase-seq data from the replicates revealed that this is not a true variable site. Figure 4-10c shows an example where DHS is clearly recognized by both platforms. However, there are considerable differences in the size of the peaks in GAII and HiSeq platforms as the pile of reads at DHSs in much higher at HiSeq.

In summary, Figure 4-10 shows how the high overall mean coverage in HiSeq vs GAII data sets leads to visual false positive variable DHS peak detection both within and between strains. This problem could be resolved with adjusting and normalizing the data. To avoid introducing false positives into my initial visual inspection of DNase-seq data, I tested whether the DHS variation within and between strains is backed up by available replicates on HiSeq platform from the same strain. I decided to rely mainly on HiSeq data sets for my initial DNase-seq visual analysis, since this offered higher resolution of DHS peak detection.

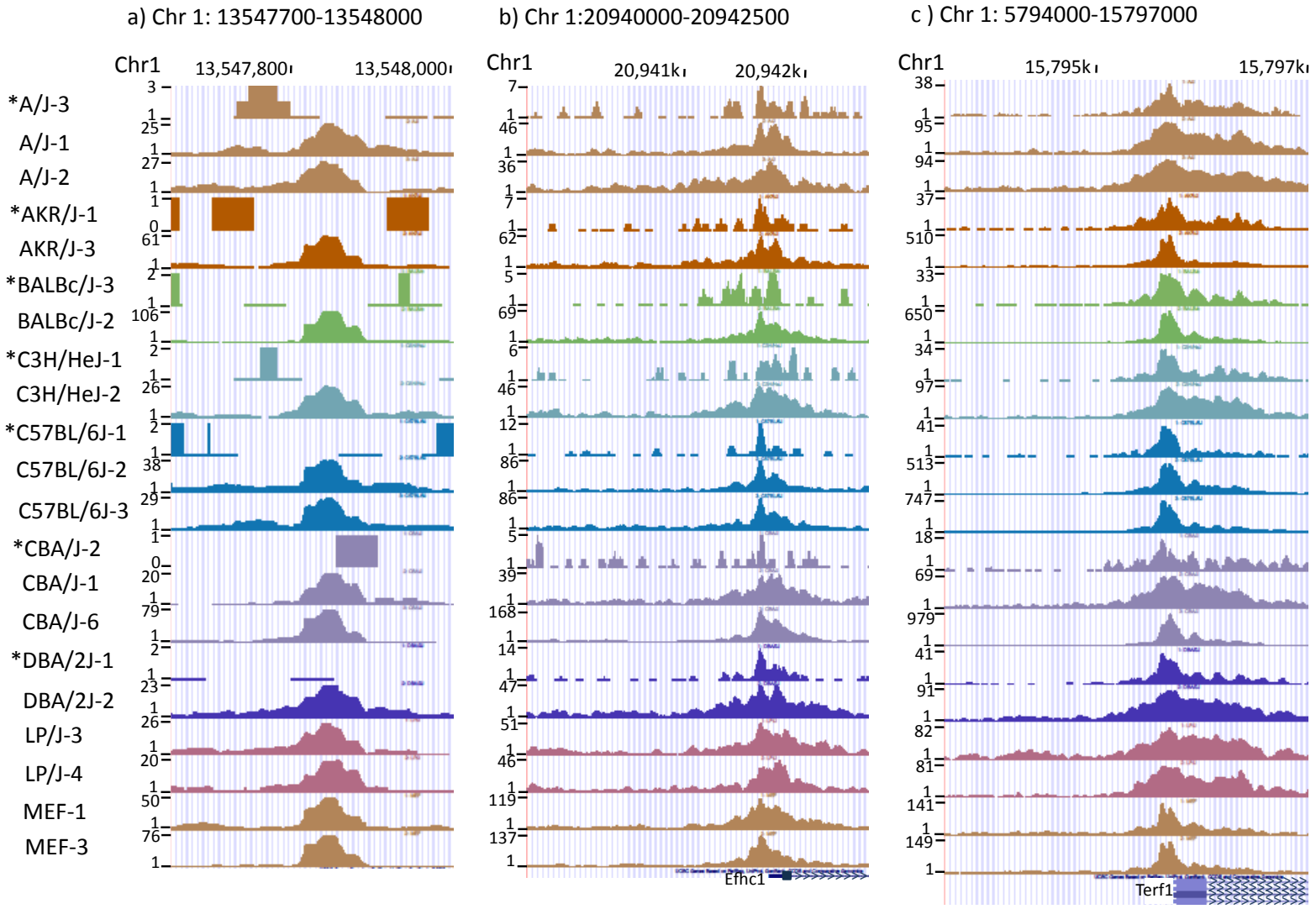


Figure 4-10 - GAII vs HiSeq DNase I-seq data. Three examples of DNase I sequencing data on GAII and HiSeq platforms on chromosome 1. Each colour represents data from replicates of the identical strain. GAII data were depicted by *. Panel a) is an example of visual false variable DHS site where an obvious DHS site seems absent in all GAII lanes. Panel c) is an example where the peak pattern is pretty similar between GAII and HiSeq however the peaks of HiSeq data had higher values and panel b) is a case between a) and c).

To understand how much variation we could observe within HiSeq DNase-seq replicates, DHS peaks of HiSeq replicates were visually inspected along the total length of chromosomes one, 14 – 19, and some random loci in the genome. There was no indication of variable DHS within HiSeq replicates of the same strain. Figure 4-11 shows an example of what I observed overall in my visual analysis regarding within strain DHS variability over 100 kb on chromosome 7 on four inbred strains (A/J, C57BL/6J, CBA/J, and LP/J). All eight strains included in our visual analysis, only four have been chosen for the sake of space to generate this figure.

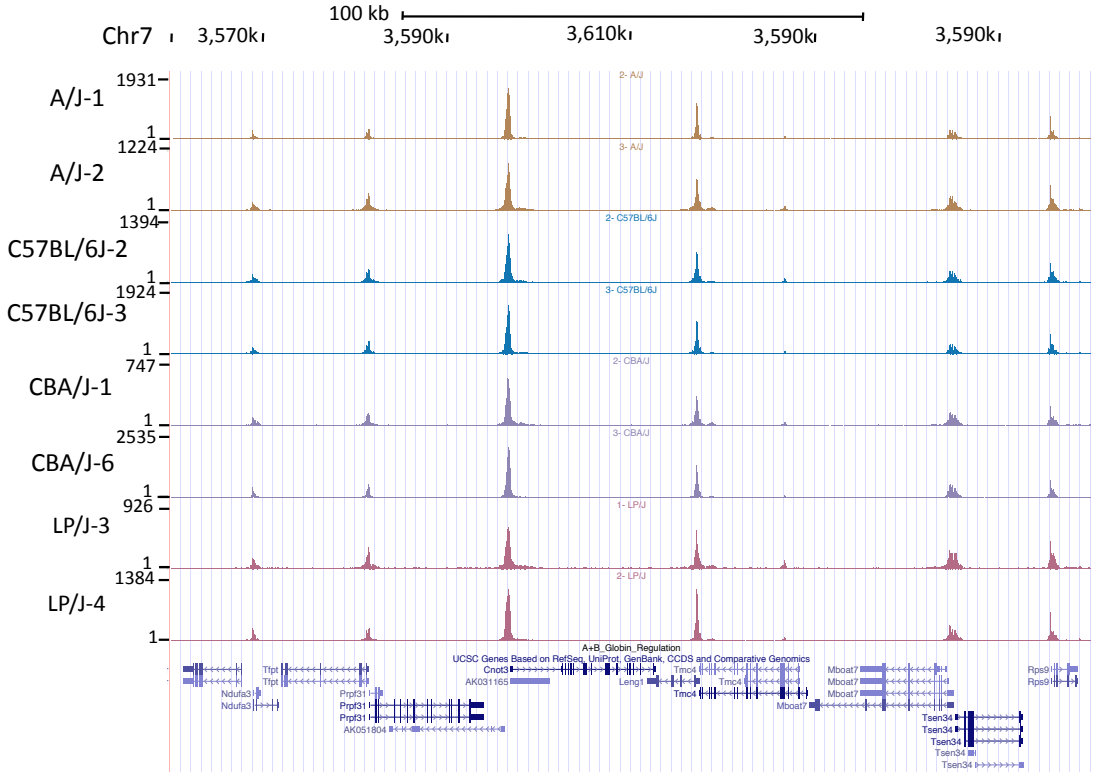


Figure 4-11 - A snapshot of within strain DHS variability in replicates (A/J, C57BL/6J, CBA/J and LP/J) over 100 kb on Chromosome 7.

4.6 VARIABLE SITES

I next turned to the issue of the variability between strains, using the data from multiple strains. My aim here was to produce a set of manually annotated peaks that could be used to calibrate a genome-wide automated analysis (as described in the next chapter).

Initially detailed visual inspection of chromosomes 14-19 revealed two main categories of changes between given inbred strains: qualitative and quantitative changes. Qualitative changes, which were more obvious to identify, could be classified as: gain or loss of single or cluster/multiple peaks. Figures 4-12 and 4-13 show two examples of gain and loss of DHS peaks in replicate analyses of the C57BL/6J strain. I include one replicate per strain for the sake of simplicity of figures. In addition, I included C57BL/6 CTCF ChIP-Seq data as a supportive and informative evidence for this variable DHS in Figure 4-12. GATA1 previously published ChIP-Seq data are also included in Figure 4-13.

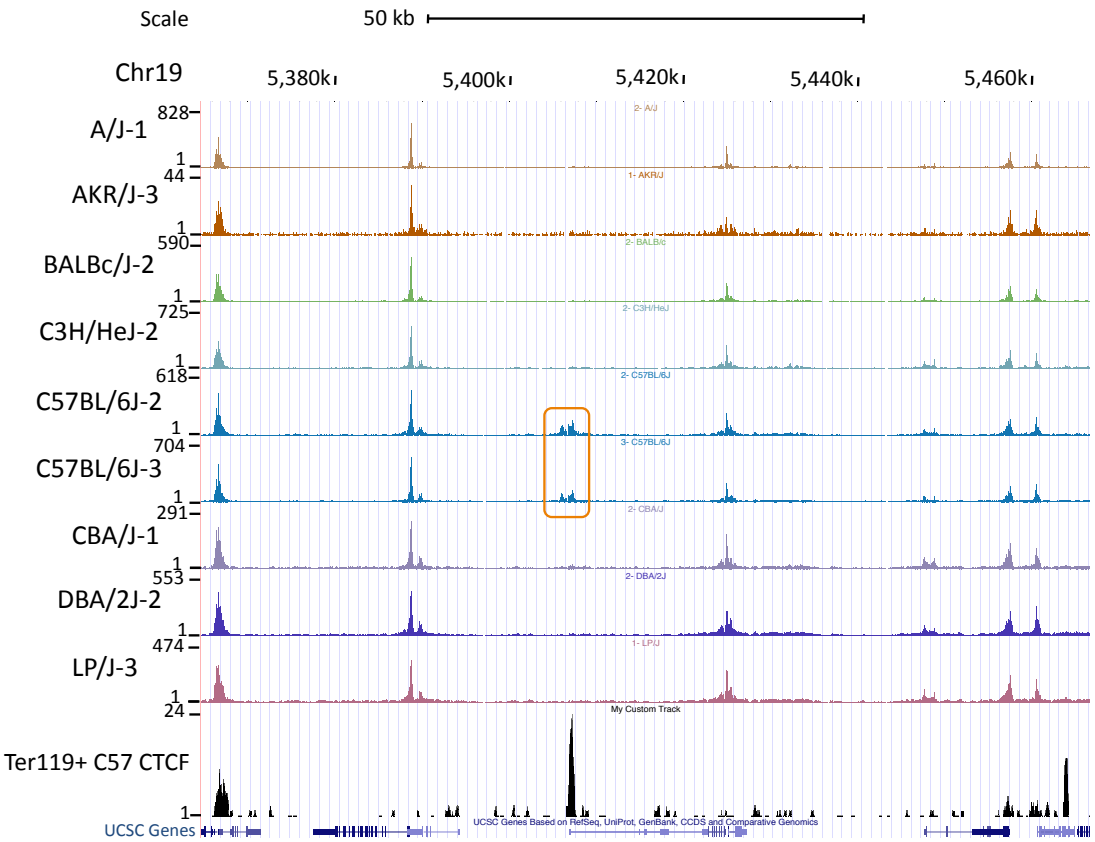


Figure 4-12 - An example of discrete DHS peak variation (gain) in eight inbred mouse strains (A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J and LP/J) over chromosome 19. DHS "gain" was observed in C57BL/6J. There is no within strains variation at this site (only two C57BL/6J replicates are shown here). CTCF track provides further evidence that there is a regulatory site in C57 in this site.

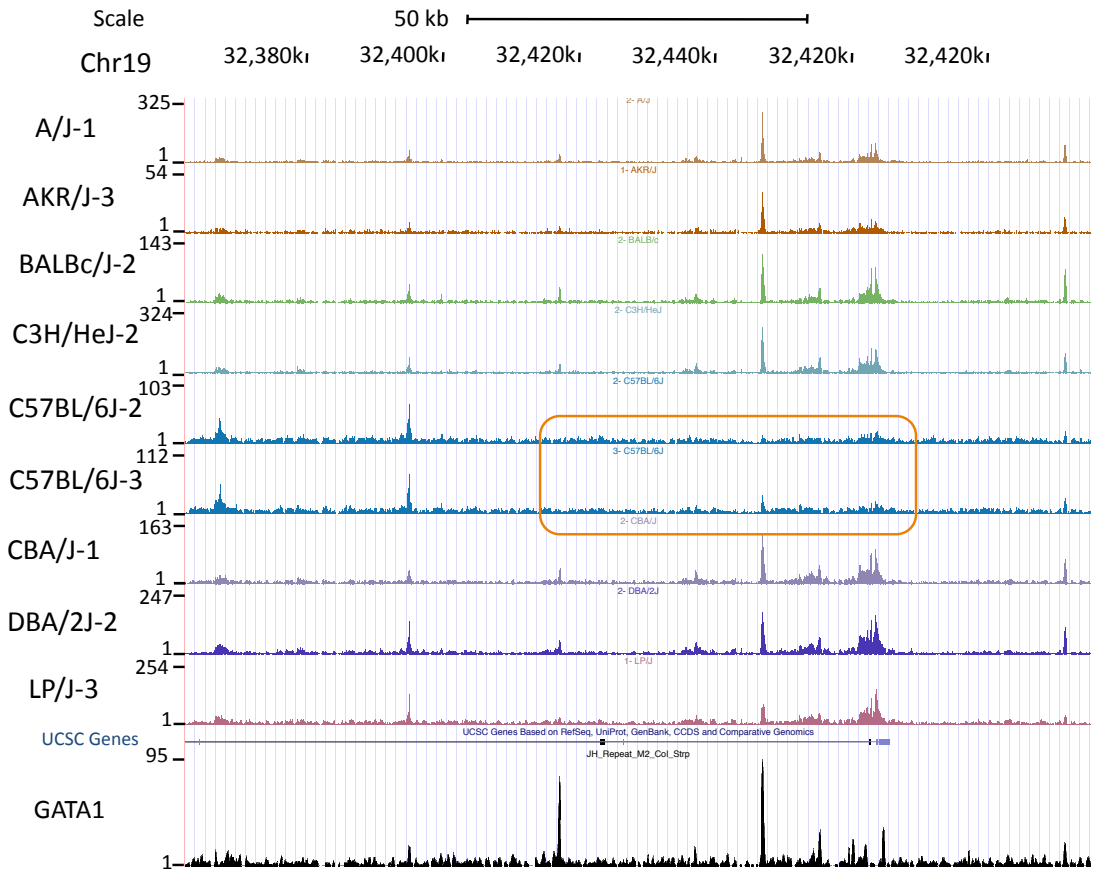


Figure 4-13 - An example of discrete DHS peak variation (loss) in eight inbred mouse strains (A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J and LP/J) over chromosome 19. DHS peak "loss" was observed in C57BL/6J strain. There is no within strain DHS variation at this site (DHS data of 2 C57BL/6J replicates is shown). GATA1 track provides further evidence for existence of regulatory element at this site.

I divided variable peaks into three different categories: I) Major variable peaks, where I was the most certain of observing a differential DHS peaks between inbred strains. II) Minor variable peaks, where I could see a variation but I was less certain in calling the variable peaks; and III) questionable cases, where in the majority of cases I could see variable DHSs, but when I zoomed in I could detect more or less similar morphological DHS peaks almost in all inbred strains (and absence of DHS peak in few exceptional cases in few inbred strains) with a very diverse peak height. I visually detected 318 confidently called variable DHS peaks on chromosome 14-19. I also manually called all the DHS peaks (including variable and non-variable DHS peaks) on the first 4.2 Mb of chromosome 19, again for use in calibrating automated calls. I could detect visually 64 DHS peaks. In fact no peak was observed in 1st 3.2 Mb of chromosome 19, so these peaks were observed over a ~1 Mb span on chromosome 19. These peaks revealed variation in size and configuration as well as the density of peaks. Some peaks appeared in close proximity and it was a challenge to consider them as separate peaks or as a cluster of peaks. Morphologically, some peaks had spiky configuration, while others form a more complicated configuration with double, triple, or multiple adjacent peaks forming a cluster of peaks.

Although I observed quantitative variation on DHS peak heights within and more prominently between inbred strains, I do not comment on the variation of DHS peak heights as it could be fundamentally affected by lack of normalizing and adjusting the data at this stage.

4.7 TISSUE SPECIFIC DHS: ERYTHROID VS MEFS

We expect to observe variation in active regulatory sites (DHSs) among different tissues. To study erythroid specific DHS vs erythroid non-specific or ubiquitous DHSs, we decided to generate genome-wide DHS data from mouse fibroblasts as a non-erythroid tissue. Initially primary fibroblasts obtained from tails of each inbred strain, cells retrieved and cultured. Since a large number of cells is required for DNase I assay ($5 \times 10^7 - 10^8$ cells) and we could not achieve the minimum number of required cells within limited number of passages to optimize the quality of prepared samples, we decided to use mouse embryonic fibroblasts (MEF). These have better proliferative characteristics compare to primary fibroblasts. MEFs were generated from C57BL/6J strain and used for cell culture and DNase I assay and DNase-seq library preparation (see Methods). Two HiSeq DNase-seq data sets generated from two replicates of MEF cell cultures.

Panel a) of Figure 4-14 shows a locus on chromosome 15 where you can see both erythroid specific DHS candidates as well as ubiquitous DHS peaks. Two morphologic categories of variable DHS peaks can be observed in the panel a: spiky single variant DHS and multiple/clustered variant DHS. Light yellow shadow boxes indicate the territory of DHSs, which were positive in erythroid replicates and absent in embryonic fibroblasts replicates. Erythroid non-specific DHSs are clearly present in DNase-seq obtained from both terminally differentiated erythroblasts and MEFs.

If a DHS is present in erythroid replicates and MEF replicates, it is definitely not erythroid specific and can be categorized as ubiquitous. In

contrast, when a peak is present in any of them, the site could be shared with some other tissues, which were not included in our study.

Figure 4-14 panel b) shows C57BL/6J MEF and C57BL/6J DNase-seq data over alpha globin locus. Since we know all the erythroid specific and ubiquitous DHSs of this locus, we can use our prior knowledge to assess how sensitive and specific our assay was to depict all known tissue specific and ubiquitous DHSs by comparing erythroid and embryonic fibroblasts replicates DNase-seq HiSeq data at this locus. As you can see, all the known erythroid specific DHSs are present in DNase-seq from terminally differentiated erythroid replicates and absent in C57BL/6J MEF replicates and all the known ubiquitous DHSs are present in both C57BL/6J erythroid and C57BL/6J MEF replicates. In this figure we can also see few examples of DHSs, which are present in C57BL/6J MEF replicates and absent in erythroid replicates that probably are related to fibroblast active regulatory elements. However there is a possibility that these DHSs are shared with some other tissues other than terminally differentiated erythroblasts (depicted by stars).

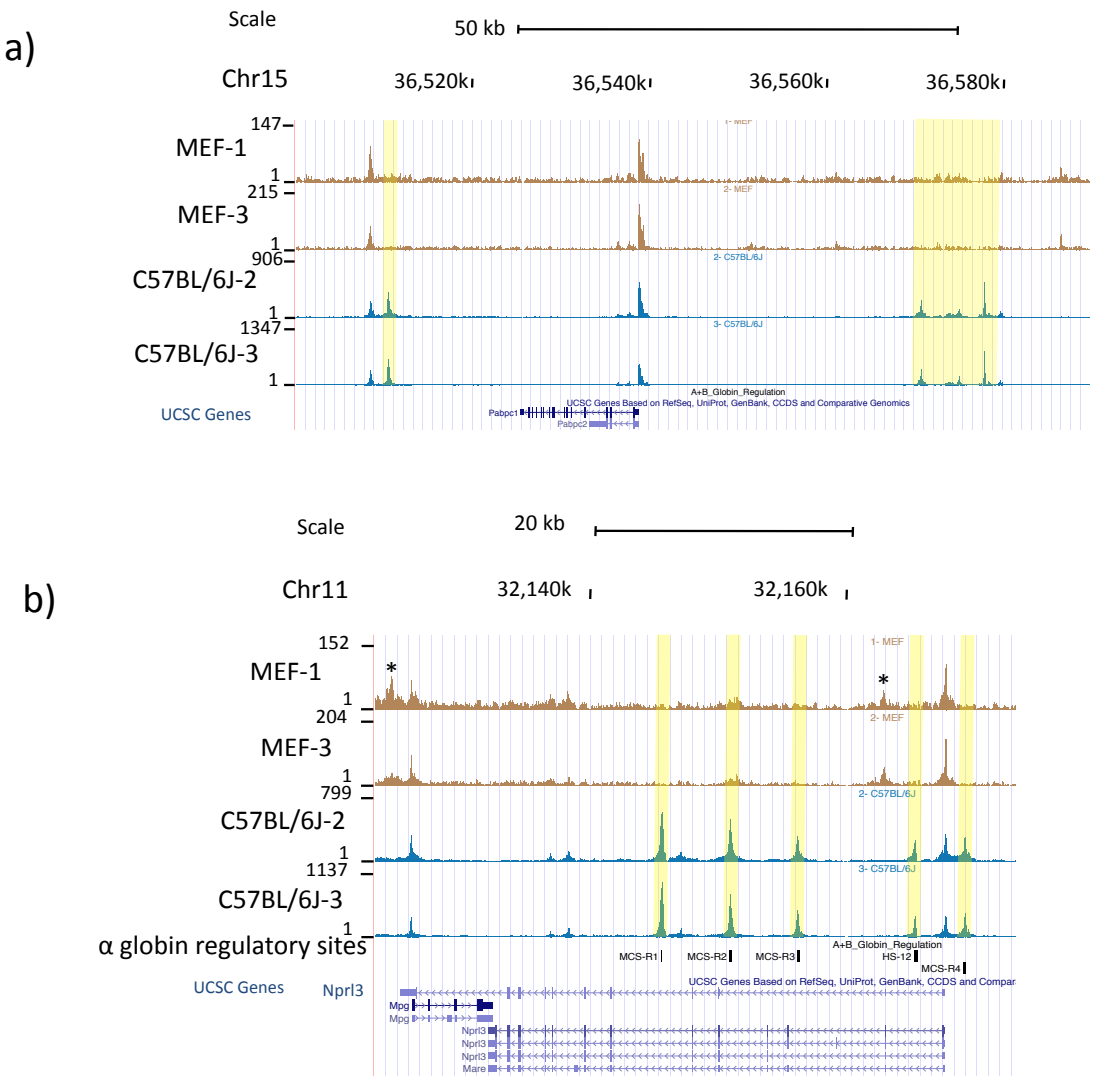


Figure 4-14 - Snapshots of DHSs in C57BL/6J erythroid cells and MEFs: a comparison between erythroid-specific and ubiquitous DHSs. a) Erythroid specific (yellow shadowed boxes) vs ubiquitous DHSs on chromosome 15. b) Erythroid specific vs nonspecific DHSs at alpha globin locus. Two DHS peaks are differentially identified in MEFs (marked by "*").

4.8 CONCLUSION

In this chapter I have described the initial phases of inbred mice DNase-seq data analysis. Comparing my DHS data to known regulatory sites at alpha and beta globin loci and chromatin regulatory marks (H3K4me3, H3K4me1 and CTCF) indicate that my data are of high quality, sensitivity and specificity. I included biological replicates for each inbred strain and I observed DHS within strain consistency and no indication of DHS qualitative variation within strains. I performed detailed visual inspection of chromosomes 14-19 and I identified several examples of between strains variation of DHSs. In some cases I observed a presence/absence difference (the gain or loss of a single peak, or cluster of peaks). This type of variation is obvious and easy to identify. In many case, DHS variation was quantitative which is not as straightforward to identify as discretely variable peaks. I observed more complex morphological changes in some cases of variable DHS peaks (variable shape). DHSs vary between different tissues due to tissue-specific gene regulation. As expected, I observed reproducible differences between the DHSs in terminally differentiated erythroid cells and mouse embryonic fibroblasts. These differences can be used to assign erythroid specificity to DHSs.

CHAPTER 5 : GENOME-WIDE ANALYSIS OF SITES OF DNASE HYPERSENSITIVY

5.1 INTRODUCTION

In this chapter I describe the development and application of an automated pipeline to analyse DNase-seq data sets on a genome wide scale. After initial visual examination of DHS data (as described in the previous chapter) determined that the quality of the sequence was high, the next task was to analyse all DHS across the genome. These results would then form the basis of an analysis of the extent and nature of DHS variation among the mouse inbred strains.

At the start of this work, I assumed that there existed variation in the number and form of DHS within the inbred strains. I assumed further that this DHS variation was due to changes in underlying genomic sequence. A related consequence would be alteration of transcription from a gene associated with the DHS (where the DHS acted as a promoter or enhancer). In the simplest model I expected to find sequence variation that disrupted a know motif, such as a transcription factor binding site. To test these ideas I set out to identify a genome wide set of DHS, which I would then be able to relate to sequence variation and other genomic features indicating the likely function of the site (such as its designation as a promoter or an enhancer).

My analysis faced a number of challenges. First, I had to generate a set of peaks in such a way that I could estimate the false positive and negative calls. I had established (as described in the previous chapter) that the data were of high quality. My 'gold standard' for assessing the quality of a set of peaks would be a

set of manually identified peaks, and the well-defined active regulatory elements at the alpha globin locus. Using these gold standard peaks, I aimed to adjust the parameters of a genome-wide peak caller to find an optimal set of peaks. Once this was achieved, my next task was to establish an analysis pipeline to detect variation in peaks and to associate this variation with other genomic features, including sequence variants, to inform me about the causes, and the consequences of DHS variation. This chapter describes this analysis and its results.

5.2 GENOME-WIDE PEAK CALLER

My initial task was to identify a suitable peak caller. My visual inspection of the mapped reads confirmed that, outside of unequivocal peaks, there is considerable variation in the extent to which reads map back to the genome. Unless the peak caller takes this background into account, in some regions of the genome with high background it may mistakenly call peaks, and in regions of the genome with low background it may miss peaks. After discussion within my group, we decided that available software was not appropriate (they typically over-called peaks) so Dr Richard Mott wrote a new peak caller.

The peak caller worked as follows. The evidence for a DHS was measured by the normalized count of sequence reads mapping to the site. *Xj* defined as the count of aligned reads covering *j* genomic position. Data were adjusted for local variation in the coverage of aligned reads (W_j) by subtracting deviation of overall mean coverage from local mean coverage (d_j) from X_j. Figure 5-1 panel a) shows a snapshot of DHS HiSeq data at part of the alpha globin locus on chromosome 11.

Figure 5-1 shows an example of the variable background signals we observed at DHS sites, where we observed higher background signals (due to higher local read coverage) than overall mean coverage signal at peaks at both sides as well as an example of adjacent DHS with lower local background signal (the peak in the middle). This figure illustrates how important it is to take this feature (variable background signals), which is due to local variation in the coverage of aligned reads, into consideration while automatically measuring DHS peak height signals and calling peaks. If we set the background signal based on regions with lower background coverage we overestimate calling the peaks in local high coverage reads and vice versa. Therefore, data were adjusted for the local coverage of aligned reads at different slots (windows) of data with similar background signal. Panel b) shows a close up view of the first two peaks on the left hand side of panel a) on DBA/2J DNase-seq data. The black and red line present DHS local read coverage of left and right DHS peaks respectively and the blue line present overall mean coverage (μ).

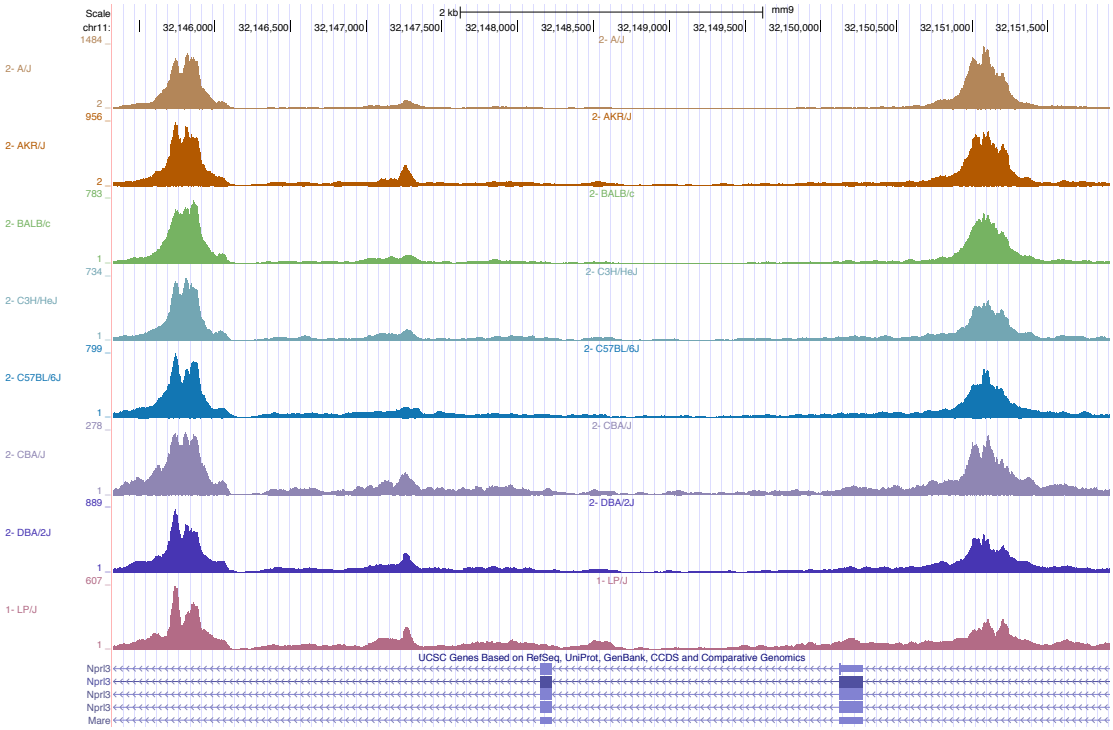
It is clear that if we choose high coverage background peaks (black line) as a universal background threshold, we underestimate the peak heights of lower background coverage peaks (e.g. right peak) and if select a universal low coverage background (red line) we overestimate peak heights of high background coverage peaks (e.g. left peak). To overcome variable background signal in various regions of DHS data, we adjusted the peak heights to local read coverage on separate slots of data (windows). Figure 5-2 schematically shows how data adjustment calculated based on number of aligned reads to position j in the genome (X_j) and deviation of overall mean coverage from local mean coverage (d_j).

In order to make GAII and HiSeq data comparable and mitigate the effect of peaks with very high values (probably raised from unresolved repetitive loci), data normalized by transforming W_j to Y_j log score which is independent of overall mean coverage (μ) and parameter p which controls stringency of peak detection (see Methods: peak detection and analysis section).

After adjusting for local variation in the coverage of aligned reads, we called DHS peaks by searching for regions with high numbers of aligned reads (hereafter peak heights)(Karlin and Altschul, 1990). Peaks called by identifying high-scoring segments, where adjusted normalized counts are at least p (stringency parameter) times of overall coverage (μ).

To choose the most appropriate values for window width and stringency parameters for automated DHS peak detection, I examined different window width and stringency parameter values and tested at what values automated peak calling results agreed with our visual observations. To calibrate automated peak calling parameters I benefited from a list of well-defined active regulatory elements (DHSs) on alpha globin locus and my set of 318 confident visually called DHS peaks on chromosomes 14-19. Table 5-1 summarizes a set of known DHSs on alpha globin locus on chromosome 11. Window and stringency values were optimized to detect all known alpha globin DHS in our automated peak calling.

a)



b)

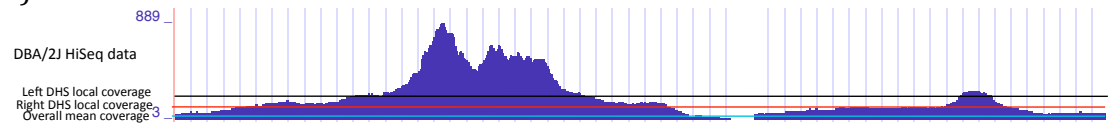


Figure 5-1 - Variable background signals at different DHSs on chromosome 11. a) DNaseI HiSeq data of inbred strains on chromosome 11 part of alpha globin locus. DHS peaks at both sides have higher local background signal to overall mean coverage in compare with the peak in the middle. If we do not adjust for local coverage of aligned reads and set a universal threshold we will overestimate peak heights for DHS peaks with higher local coverage. And if we set the threshold based on the coverage at high local coverage peaks we will underestimate DHS signal at lower coverage sites. Therefore, for performing unbiased analysis we adjusted DHS data for local coverage variations. b) Close up of first two peaks on the left in panel a) of DBA/2J DHS data. Black line presents left DHS local read coverage, red line presents right DHS local read coverage and blue line shows overall mean coverage. If we set our universal background threshold on high coverage background peaks (black line) we underestimate the peak heights of lower background coverage peaks (right peak) and if set it on universal low coverage background (red line) we underestimate peak heights of high background coverage peaks (left peak). Therefore we adjusted the peak heights to local read coverage on separate slots of data (windows).

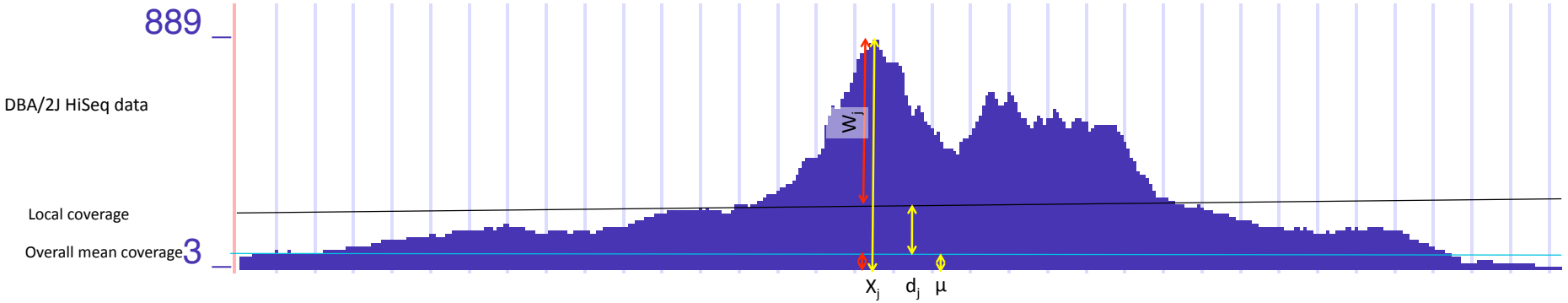


Figure 5-2 - Adjusting read counts for local coverage variation: $W_j = X_j - d_j$. W_j is adjusted count of aligned reads for local coverage variation. X_j is the count of aligned reads covering j genomic position and d_j is deviation of local mean coverage (black line) from overall mean coverage (blue line). In other words W_j can be measured by subtracting local mean coverage from X_j and adding overall mean coverage (μ).

Table 5-1 – List of known DNase I hypersensitivity sites on alpha globin locus (reference genome)

DHS name	Chromosome	Start	Stop
Mpg_Promoter	chr11	32126343	32126570
MCS-R1	chr11	32145705	32145786
HS-29_CTCF	chr11	32147137	32147347
MCS-R2	chr11	32150980	32151218
MCS-R3	chr11	32156216	32156487
HS-12	chr11	32165247	32165511
Nprl3_promoter	chr11	32167614	32167722
MCS-R4	chr11	32169035	32169192
HBA1_Promoter	chr11	32183580	32183710
HBQ1_Promoter	chr11	32186849	32186987
HBA2_Promoter	chr11	32196397	32196527
HBQ2_Promoter	chr11	32199954	32200092

Figure 5-3 shows snapshots of automated DHS peak calling on two different settings on chromosome one. Y axis represents the number of reads and X axis location on chromosome one on Mb. Green dots present start and red dots present stop of DHS peaks. The upper set of dots belong to p=6.5 and window width=100,000 bp and the lower set of dots belong to p=3.9 and window width= 550 bp. As we expect with higher stringency parameter we automatically called less DHS peaks. We observed examples of broader DHS peaks and small peaks are missed in this setting (like the first DHS peak on the left hand side of this figure). An incident of calling two adjacent DHS peaks as one broader DHS peak marked by star (*) in this figure. The same peak is called as two distinct adjacent DHS peaks in the lower setting (lower stringency parameter and smaller window

width). In this setting we called more DHS peaks but generally DHS peaks are too narrow as adjacent DHS peaks often called as separate DHS peaks.

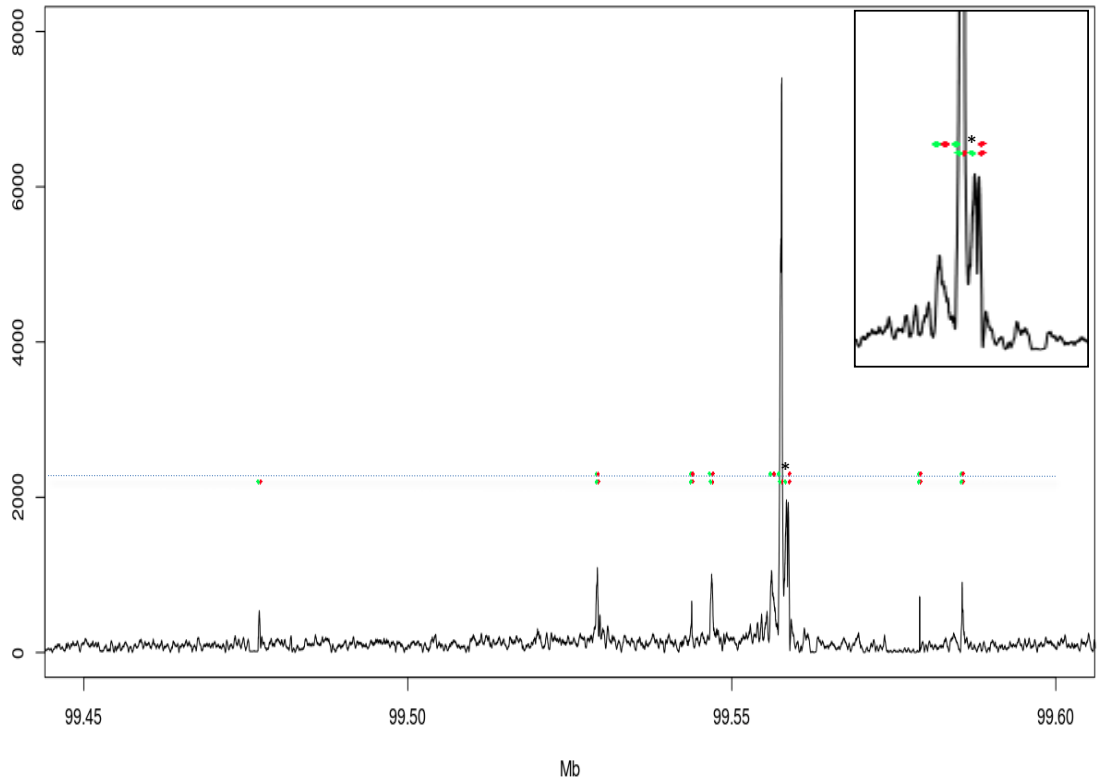


Figure 5-3 – Automated DHS peak calling on chromosome one with two different parameter setting. Y axis shows number of reads and X axis show the location chromosome one on Mb. Green and red dots present start and stop of DHS peaks respectively in *p*=6.5 and window=100,000 bp (upper row of dots) and *p*=3.9 and window=550 bp (lower row of dots). Dashed blue line separate two set of peaks. Overall upper set called less DHS peaks as expected by more stringent parameter and incidents of broader DHS peaks (one example marked by star in this figure). Lower sets detected more DHS peaks and incidents of narrower peaks (the first DHS peak on the left is detected with lower stringency parameter set and missed by higher stringency parameter set). At the upper right corner of the figure the magnified view of starred peak is presented for better visibility.

As described earlier we used well-defined set of active regulatory elements (DHSs) on alpha globin locus as a guide for calibrating peak calling parameters. Figure 5-4 shows the comparison between automated peak calling and known DHS sites on chromosome 11 at part of alpha globin locus. Y axis shows number of reads and X axis shows location on chromosome 11 on Mb. Green and black dots present start and stop of automated called DHS peaks respectively with p=3 and window width= 5000 bp. Paired red dots present start and stop of known DHS at part of alpha globin locus. It is evident that these parameters required more titration as false negative and false positive DHS peaks observed based on definite known DHS sites on alpha globin locus. Ideally, we aimed to detect all known regulatory sites or minimum number of false negatives and minimum number of false positives.

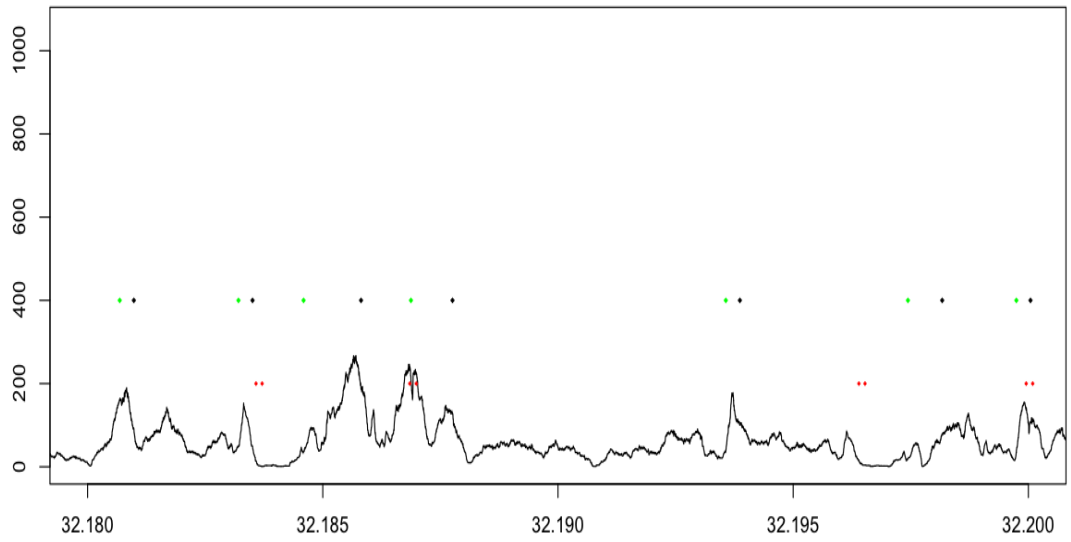


Figure 5-4 - Comparison between automated DHS peak calling (*p*=3, window=5000 bp) and known DHS sites at alpha globin locus on chromosome 11. Y axis shows number of reads, x axis shows the location on chromosome 11 on Mb. Green and black dots show start and stop of automated DHS called with above parameters. Paired red dots present start and stop of known DNase I hypersensitivity sites on part of alpha globin locus. These parameters led to number of false positive and negative DHS peak detection based on comparison with known set of active regulatory elements at this locus.

Out of various testing parameters for *p* and window width, at *p*=2 and window width=800 bp, automated peak calling identified 47 out of 56 confident visually called DHS peaks on chromosome 14 and 11 out of 12 known DHSs on alpha globin locus. Figures 5-5 and 5-6 show snapshots of automated DHS peaks on chromosome 14 and 11 with above parameters. Using these parameters for automated detection of DHS peak across the genome: 291 DHS peaks matched from set of 318 manually called peaks.

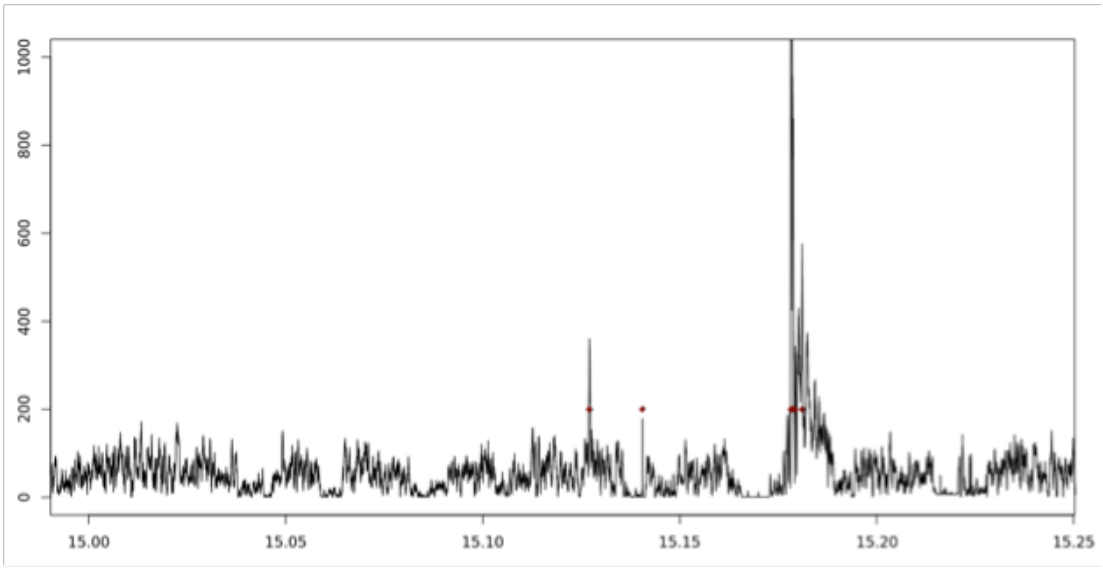


Figure 5-5 - Automated DHS peak calling on chromosome 14. Y axis represents number of reads and X axis represents location in Mb. Red dots show automated DHS peak called at *p*=2 and window width=800 bp. These parameters captured all prominent DHS peaks.

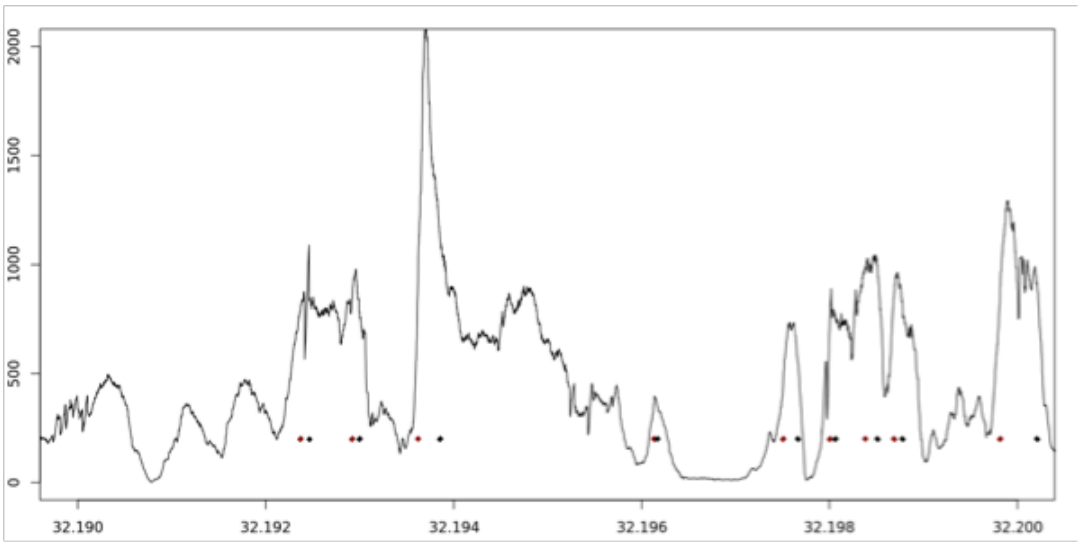


Figure 5-6 - Automated DHS peak calling on chromosome 11. Y axis shows the number of reads and X axis shows the location in Mb. Red and black dots represent beginning and end of automated called DHS peaks respectively at *p*=2 and window width=800 bp.

Figure 5-7 shows a snapshot of automated DHS peak calling on chromosome 19. Three panels are showing the same region but in different magnification. From Top to the bottom we zoomed in as it covers less distance. Upper panels shows DHS-seq data reads and DHS automated peak calls. Green and red diamonds present start and stop of DHS peaks. Due to compaction, green diamonds almost fully obliterated by red diamonds. Middle and lower panels show the same information on 21-22 Mb and 21.6 – 21.85 Mb on chromosome 19 respectively. As clearly evident in the lower panel, there are visually detectable DHS, which are not called automatically (marked by star (*)). We used a confident set of visually called DHS to calibrate our automated DHS calling parameters to minimize the difference between visually and automated DHS peak calling. Eventually *p*=3.9 and window width=550 bp parameters were chosen for genome-wide automatic DHS peak calling.

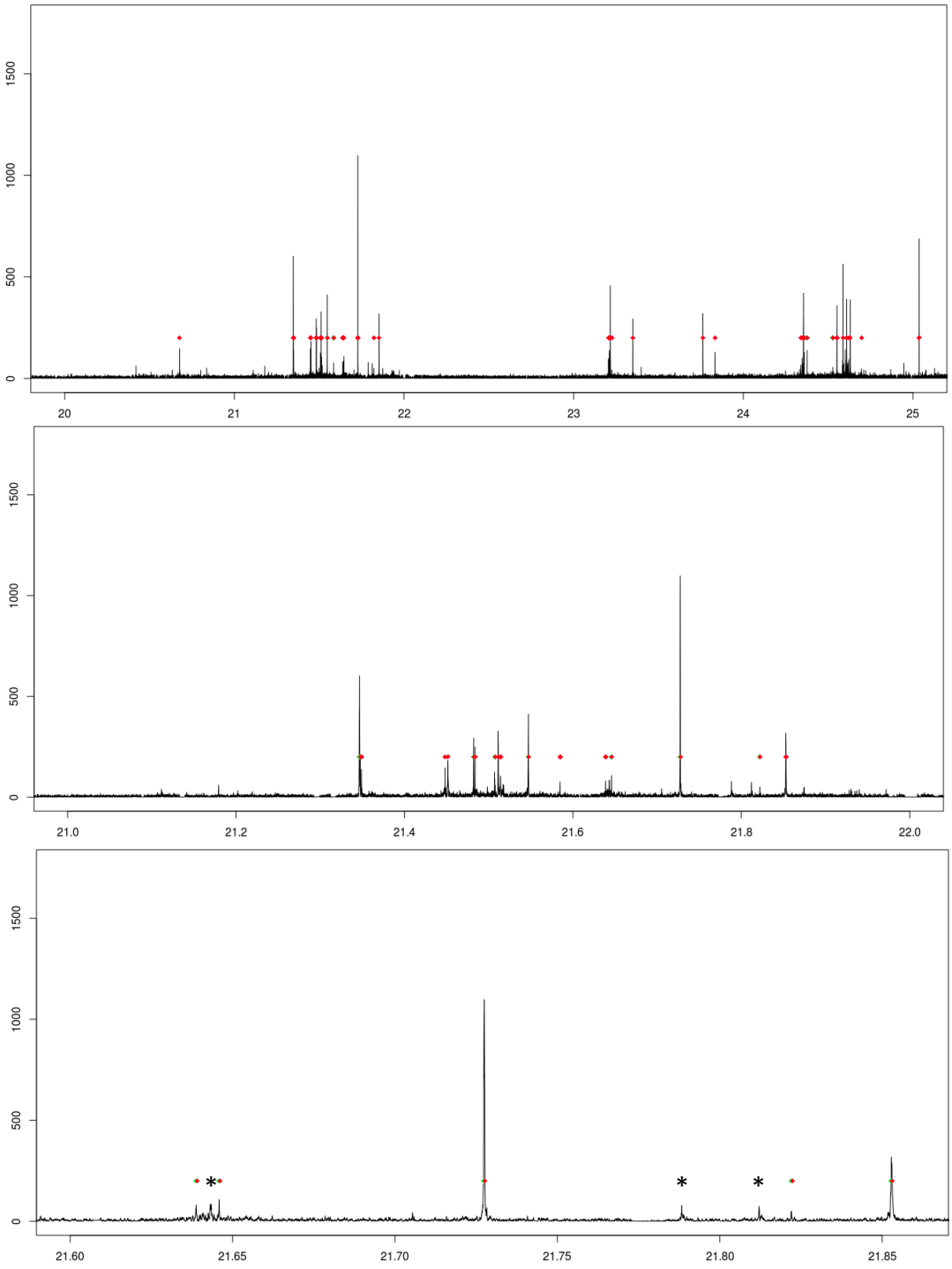


Figure 5-7 – Automated DHS peak calling on chromosome 19. Y-axis shows number of reads and X-axis location on Mb. Top panel shows DHS on 20-25 Mb. Middle and lower panels show automated peak calling on 21-22 Mb and 21.60-21.85 Mb for better visibility. Green and red diamonds show start and stop of automated DHS called. Due to display compaction, green diamonds almost fully covered by red diamonds. As clearly seen in lower panel there are incidences of visible DHS peaks, which failed to be called automatically (*).

5.3 COMPARING DHS PEAKS WITH ALPHA AND BETA GLOBIN KNOWN REGULATORY ELEMENTS

To validate the sensitivity and specificity of our DNase-seq data sets, we compared the result of our DNase-seq data sets with known regulatory elements over alpha and beta globin loci. As a plus to our previous analysis this time we took advantages of using normalized counts of aligned reads and each data set aligned to its specific strain reference sequence.

Figure 5-8 shows the result of this comparison for the eight inbred strains over a 100 kb region containing the previously well characterized alpha globin gene cluster on chromosome 11 (Flint et al., 1997). DHS peaks for each strain are identical and agree on the location of previously mapped DHS (the position of known DHS are shown by vertical lines)(Kowalczyk et al., 2012). Similar analysis for beta globin locus showed compatible results (Figure 5-9).

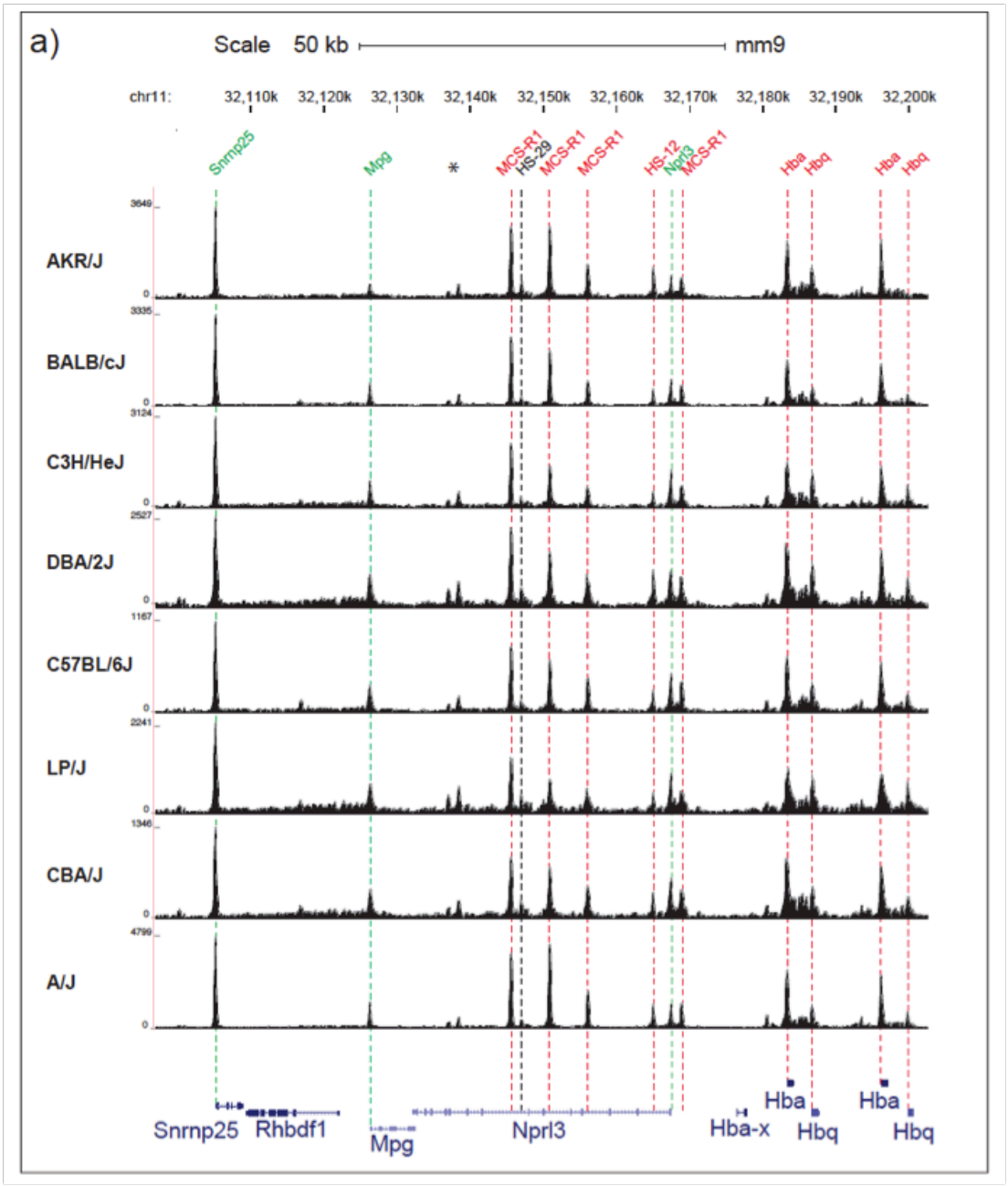


Figure 5-8 - DNase I Hypersensitive sites at the alpha globin cluster on chromosome 11. a) Chromosomal positions in kb on chromosome 11 are shown for MGSC37 mouse reference genome. The names of previously annotated DNaseI hypersensitive sites (DHS) are shown coloured as ubiquitous (green), erythroid (red) and black (CTCF) (Kowalczyk et al., 2012). Two previously unpublished DHS (marked *) are associated with CTCF binding sites. The density of aligned DHS-seq reads in a moving 300bp window, with a 30bp increment is shown for each mouse strain. DHS-seq peaks are aligned with the position of each previously described DHS site by dashed lines, colour coded as described. The position of Refseq annotated genes are named and shown below in blue.

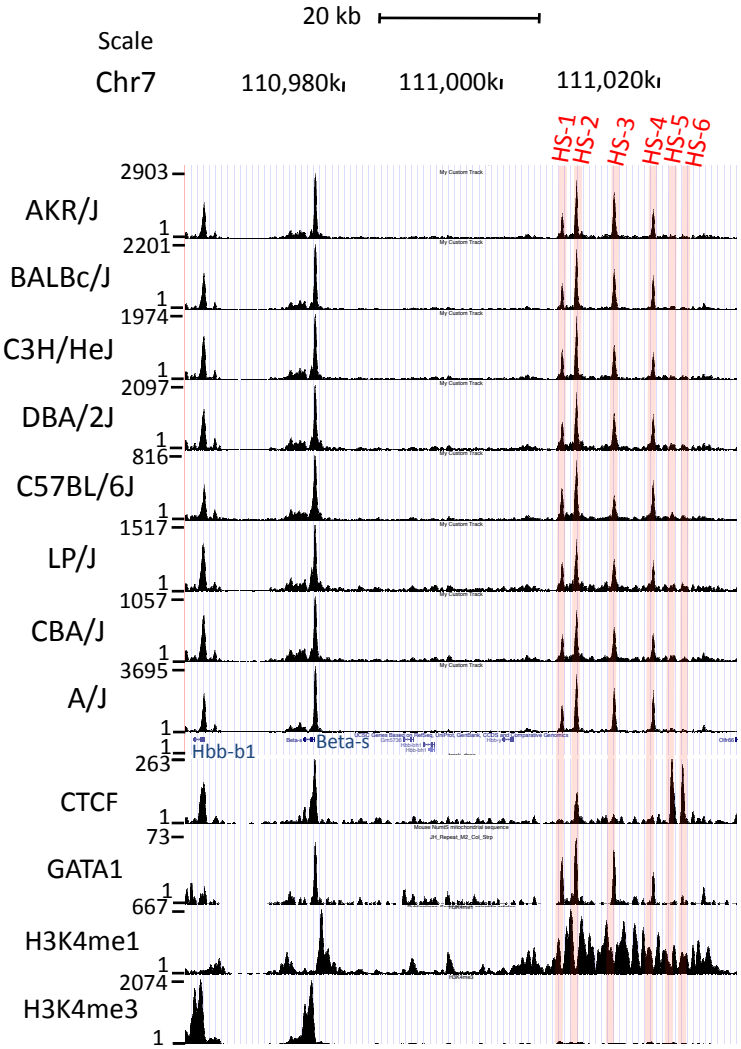


Figure 5-9 - DNase I hypersensitive sites at beta globin cluster on chromosome 7. DHS data of eight inbred mouse strains is shown at beta globin locus and CTCT, GATA1, H3K4me1, H3K4me3 tracks also provide additional evidence on regulatory sites. Beta globin locus control regulatory sites HS1-6 is marked by vertical red lines. Prominent peaks were observed at HS-1-4 and smaller peaks at HS-5 and HS-6.

5.4 DHS CHARACTERISTICS

I pooled DNase-seq data from all 19 experiments. Across all 19 experiments there were a total of 36,693 DHS peaks that covered 9.1 Mb (0.29% of the genome); 25,700 peaks (6.3 Mb) lay within mappable (single copy) regions of the genome and we used this subset for further analyses. Figure 8 5-10 shows a histogram of interval sizes (DHS peak width) of automated called DHSs (Mean 396 ± SD193.5).

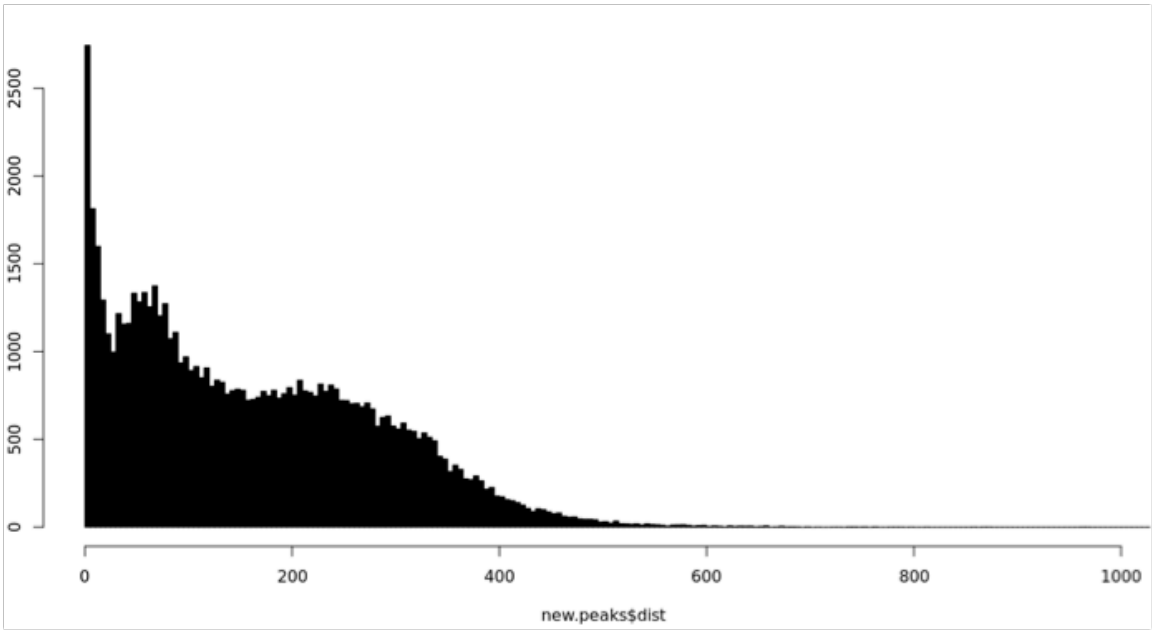


Figure 5-10 - Histogram of interval sizes of DHS peaks. The majority of peaks are below 500 bp wide (Mean 396 ± SD193.5).

Figure 5-11 shows an incidence of exceptionally big DHS on chromosome 17 (coordinates: chr17:39,979,923 - 39,985,797 bp, peak interval size: 5,874 bp). This DHS peak is present in all inbred strain with identical morphology in all strains. Although this peak is not among variable DHS peaks, which were interesting for us to understand the cause and effects of variable chromatin

among inbred mouse strains, the exceptional size of this peak raised our interest to study this peak individually in more depth. To understand the nature of this extraordinary big DHS peak, first I asked if this peak is erythroid specific or is also present in other cell types by looking at the DHS data I generated from MEFs (Mouse Embryonic Fibroblasts from C57BL/6J strain). The result is this huge DHS site is present in both 2 replicates of C57BL/6J MEFs as well as all 8 inbred strains terminally differentiated erythroid cells and therefore is not erythroid-specific. Similar peaks are also observed at promoter (H3K4me3), enhancer (H3K4me1) and actively transcribed gene bodies (H3K36me3) chromatin signatures ChIP-Seq data and CTCF ChIP-Seq data from C57BL/6 strain and erythroid transcription factor ChIP-Seq data (GATA1, Klf, Scl) at this site. The only gene annotated to date within the span of this peak is a 110 bp mouse Mir715 (microRNA) gene (coordinates: chr17: 39,981,081 - 39,981,190). This DHS peak is not in a gene rich region of the genome and the first adjacent annotated genes to it are on the right and left of this gene are *Pgk2* (on the right) and *Esp15* (on the left) ~ 3.6 and 2 Mb apart from the borders of this DHS peak. Our RNA-Seq data obtained from terminally differentiated erythroid cells indicates evidences of RNA poly(A)$^+$ and poly(A)$^-$ expression at this locus (C3H/HeJ strain tracks is presented here). Comparison of our DHS data with human ENCODE DNase I hypersensitivity publically available tracks on UCSC genome browser at this site indicates similar DHS peaks were also identified in DHS data obtained from erythroid and non-erythroid tissues (including liver, heart, Bcells, brain tissues) in Mouse ENCODE project with similar extraordinary width (~6.4 kb in erythroid cells and similar results in other DHS data).

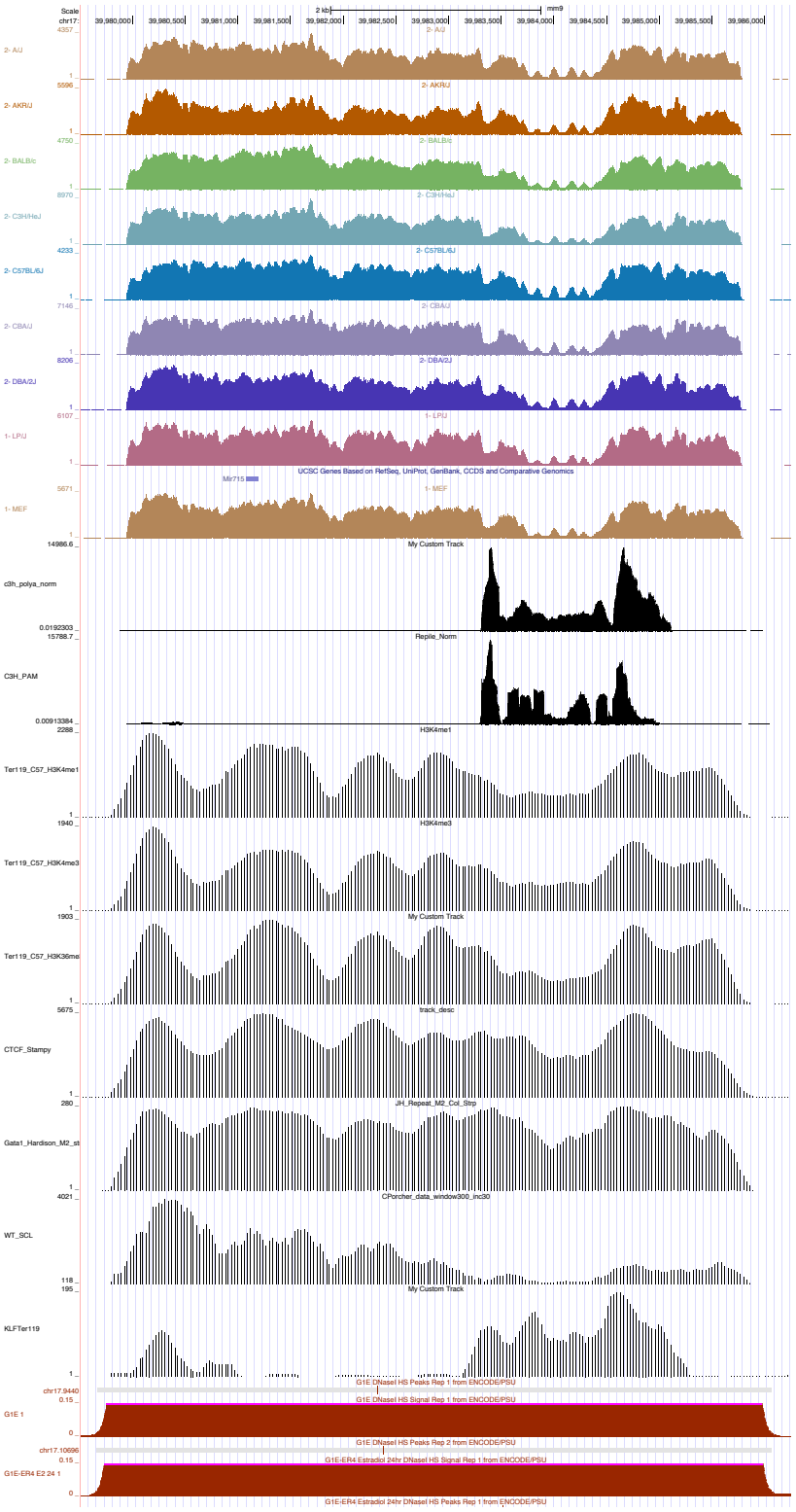


Figure 5-11 - A giant DHS on chr17:39,979,781-39,985,956. Tracks displayed in this figure are DHS data from terminally differentiated erythroid cells from 8 inbred strains and C57BL/6J MEF, RNASeq poly $(A)^+$ and poly$(A)^-$ data from C3H/HeJ erythroid cells, histone marks (H3K4me1, H3k4me3, and H3K36me3), CTCF and erythroid TF (GATA1, Scl and KLF) ChIP-Seq data and two erythroid DHS data from Mouse ENCODE project on UCSC genome browser.

5.5 DETECTION OF VARIABLE PEAKS

I had generated DHS data from terminally differentiated erythroblasts using 19 mice from eight inbred strains (three each of C57BL/6J, A/J, and CBA/J mice, and two each of AKR/J, C3H/HeJ, DBA/2, BALBc/J, and LP/J strains). The first replicate of each strain, except LP/J, was sequenced on a GAII platform and the rest were sequenced on HiSeq platforms (7 GAII and 12 HiSeq erythroid DNase-seq data sets). An analysis of variability had to take into account these different levels of variation, between batch and between strains.

Reads mapped back to the genome do not have a simple Gaussian distribution so it is not appropriate to detect variation using standard analysis of variance, since this assumes an underlying normal distribution. Instead we used a method that models the distribution as a negative binomial. We used the software package DESeq for R to carry out this analysis (performed by Drs Richard Mott and Jonathan Flint).

At a false discovery rate (FDR) of 10% we found 2,530 variable peaks (9.8%), and 1,397 peaks at a 1% FDR (~5.4%). I checked the automated detection of variable peaks by visual inspection of peaks on chromosomes 14 to 19 and confirmed all were detected in the automated analysis. I refer to these 1,397 DHS as the set of variable DHS peaks.

We found that across a wide range of parameter values the proportion of variable peaks detected automatically remained approximately constant (Figure 5-12). X-axis shows the window size in kb and the Y-axis represents the percentage of variable sites. Each color assigned to a specific stringency parameter (p) presented in the logo of the figure. This figure shows as expected the percentage of variable sites slightly increases with the higher stringency

parameters (*p*) as the graph shifts clearly up as *p* increases, but window size alteration had no effect on proportion of variable sites and it remained almost constant along X-axis (with variable window sizes (0-100 kb).

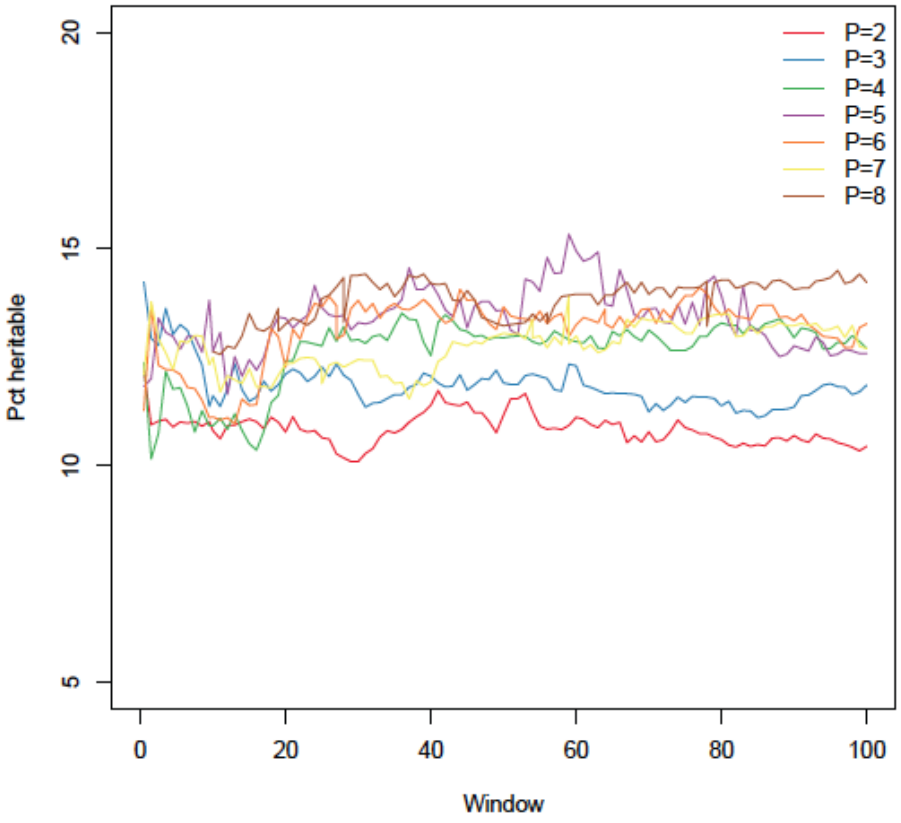


Figure 5-12 - Effect of altering window (kb) and stringency parameter (*p*) on the proportion of variable sites. Y-axis shows percentage of detected variable DHS peaks and X-axis shows window size (kb). Increase in *p* values led to higher proportion of automatic variable site detection but window size alterations do not have any obvious effect.

5.6 CLASSIFICATION OF VARIABLE PEAKS

I identified three visually recognizable types of variable DHS peaks (Figures 5-13 - 5-15): at 173 cases of variable DHS peaks (12.3%) there was a loss (or gain) of a peak (Category I Figure 5-13) in one or more strain(s); at 23 cases (1.6%), variable peaks occurred in clusters within 10 kb of each other (Category II Fig 5-14); the remaining majority 1201 variable DHSs ~ 84% showed variation in size and/or shape (Category III Figure 5-15).

This classification initially was suggested based on morphological distinct characteristics of variable peaks, the beauty of this classification is that it nicely agrees with the nature of variation: qualitative vs quantitative variation. Category I addresses the qualitative variation of a single DHS peak, where the main variation occurs in the form of "loss" (i.e. absence of a detected DHS in other strains in less than a half of strains) or "gain" (i.e. emergence of a new DHS in less than half of strains). Category III refers to a quantitative (continuous) variation of DHS peak heights. In contrast of type I, there is no qualitative change, DHS peaks are present in all inbred strains, but the intensity of signal varies between strains beyond the variation occurs within replicates of the same strain. Category II refers to variation of cluster of DHS peaks within the vicinity of 10 kb. In contrast with type I and III which both refer to variation of a single DHS, as multiple DHS peaks are involved in category II, any of individual peaks in the cluster could be categorized as category I or III. So this category of variable peaks might involve both quantitative/qualitative variations so maybe we could describe them as "compound" or "complex" regarding the nature of variation. This classification could represent spectrum of variable DHS where at one end we see qualitative variation of a single peak (category I), in the middle we have

qualitative and/or quantitative variation of cluster of peaks (category II), and at the other end of this spectrum we have single variable DHS with qualitative variation (change in shape or size) (category III).

Alternatively, we called these categories as “Discrete Unique” (category I) where we observed discrete variation in a single DHS. “Discrete Compound” (category II) where we observed discrete change in at least one of the DHS peaks within a cluster of peaks, and “Continuous” (category III) where we observed quantitative variation in variable DHS peaks. In summary, minority of variable peaks (196 peaks, 14%) displayed discrete variation between in one or more strain with loss (or gain) of a peak in one or more strain, and the majority of variable peaks (86%) showed continuous variation (variation in peak height or shape).

Figure 5-13 shows two examples of Category I variable DHS peaks, where a clear single spiky DHS is absent in at least one inbred strain. Panel a) shows an example of category I variant DHS peak on chromosome 17. The chromosome location is showed on top following by individual data sets from inbred strain (small scale on Y axis shows the intensity of signal in each data set). For the sake of space readability and simplicity of the figures only the data from one replicate per strain displayed in figures. There was no within strain variability and the data was robust and reproducible within replicates of the same strain. UCSC gene annotation presented in blue, next there are two chromatin mark ChIP-Seq data sets from C57BL/6 terminally differentiated erythroblasts: H3K4me1 (enhancers are predominantly H3K4me1 modified) and H3K4me3 (promoters are predominantly H3K3me3 modified). Since variable DHS peaks are present in C57BL/6J strain in these two examples, ChIP-Seq data

(which has been obtained from the same strain) is informative in these cases. The presentation of data remains similar in all 6 examples showed in figures 5-13 - 5-15. Panel a) shows a clear single spiky variable DHS peak on chromosome 17 (coordinates: 85,175,568 - 85,176,695 bp), which is only absent in DBA/2J and present in all other seven inbred strains. H3K4me1 and H3K4me3 chromatin marks show prominent peaks, which add further evidence to validate this peak as an active regulatory site and suggest an enhancer like function for this site. This site aligns in intronic locus of Lrpprc gene. The adjacent non-variable DHS peak aligns with the promoter of this gene. The comparison of configuration of chromatin mark ChIP-Seq data is consistent with what we expect to observe in enhancers and promoters (i.e. big H3K4me1 peak and small H3K4me3 peak for enhancer-like elements and big H3K4me3 peak and small H3K4me1 peak for promoters). Panel b) shows another example of spiky single variable DHS peak on chromosome 10 (coordinates: 121,005,821 - 121,006,767 bp). This variable DHS peak is absent in 3 inbred strains (BALBc/J, DBA/2J, and A/J) and present in other 4 inbred mouse strains. Chromatin mark ChIP-Seq data add further validation evidence and in parallel to gene annotation suggest the function of this active variable regulatory site. This site, which resides in intronic region of Tbk1 gene, also could be an enhancer-like element based on the positional property and the configuration of H3K4me1 and H3K4me1 ChIP-Seq peaks.

Figure 5-14 shows 2 examples of category II (cluster/multiple/ compound) variable DHS peaks. Panel a) shows an example of cluster variable DHS site on chromosome 8 (cluster coordinates: 87,112,158 - 87,120,136 bp). Two adjacent peaks out of four DHS peaks within this cluster picked as variable

peaks (selected within red shadow box). If we analyse them individually, the one on the left (coordinates: 87,112,369 - 87,112,763 bp) is only present in C57BL/6J and is similar peaks in category I, but in contrast to peaks in category one, it occurs in the proximity of other variable DHS peaks. The variable DHS on the right (coordinates: 87,115,803 - 87,116,629 bp) is present in all 8 inbred strains; however, size of this peak varies among inbred strains (similar to category variable DHS peaks in category III). Position of these peaks to gene annotations and H3K4me1 (enhancer) and H3K4me3 (promoter) chromatin marks, suggests both of these variable DHS peaks are intronic enhancer-like variable active regulatory sites. Panel b) shows another example of cluster variable DHS peaks on chromosome 5 (cluster coordinates: 114,422,812 - 114,444,750 bp). This cluster includes 12 DHS adjacent peaks. There are multiple variable peaks in this cluster, which marked by red shadow box. Overall, DHS profile is clearly different in C3H/HeJ and CBA/J inbred strains and include a complex variation in size and "loss" of DHS in these two strains. Chromatin marks add additional support for the presence of active regulatory elements at this locus.

Figure 5-15 shows two examples of category III variable DHS peaks. Panel a) shows an example of category III variable DHS peak (variation of size/shape) on chromosome 9 (coordinates: 88,494,068 - 88,494,709 bp). A clear spiky single DHS peak is present in all eight inbred strains but the size varies between strains. The size is significantly smaller in DBA/2J and CBA/J in compare with other inbred strains. Panel b) shows another example of category III variable DHS peak on chromosome 10 (coordinates: 82,930,135 - 82,930,236 bp). Shape and size of DHS peak is similar in 3 inbred strains (BALBc/J, DBA/2J,and A/J) and different from other inbred strains. Presence of peaks at chromatin marks

(H3K4me1 and H3K4me3 ChIP-Seq) data add further support to the presence of active regulatory sites at these loci.

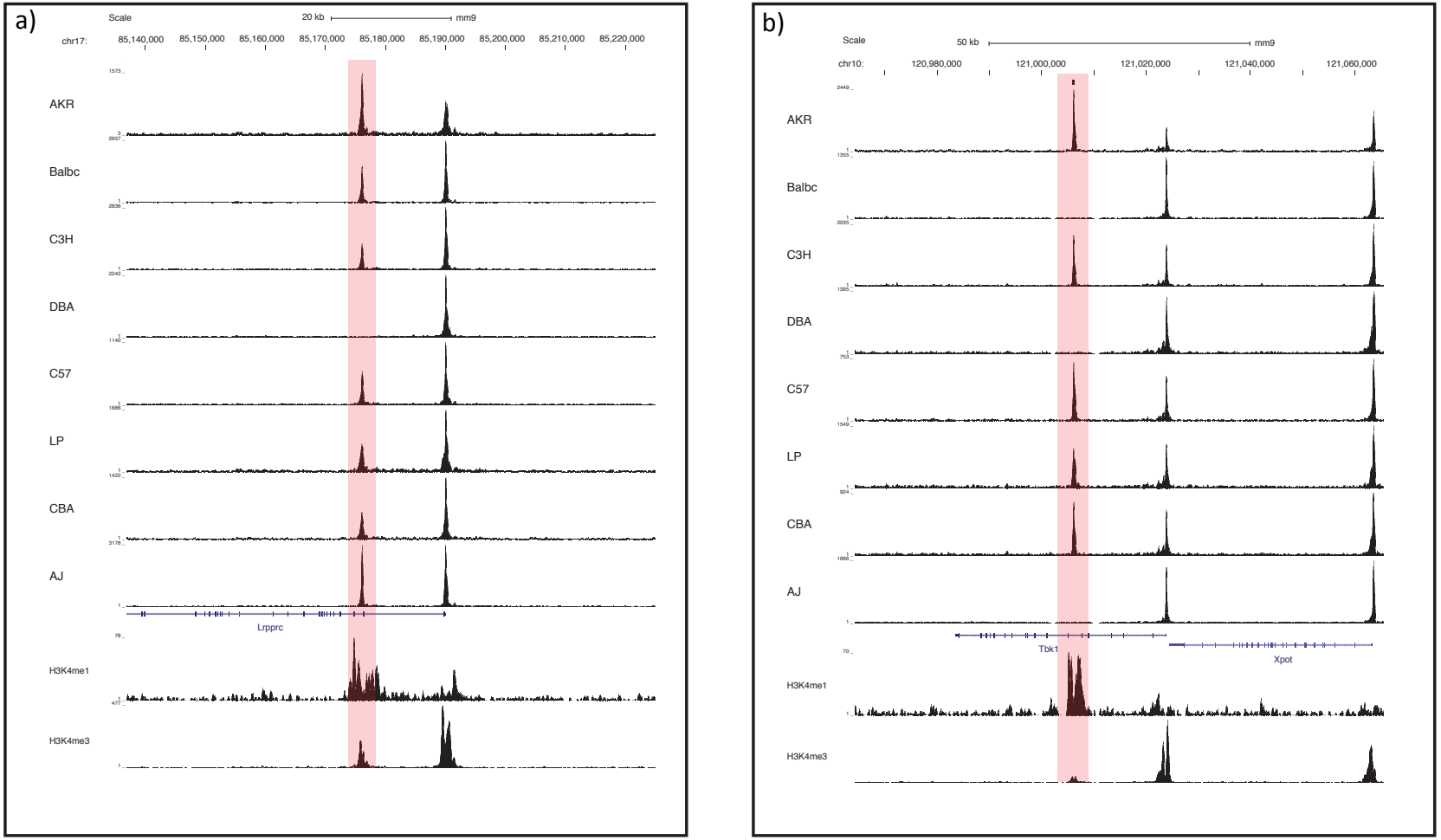


Figure 5-13 - Loss of DHS in at least one strain

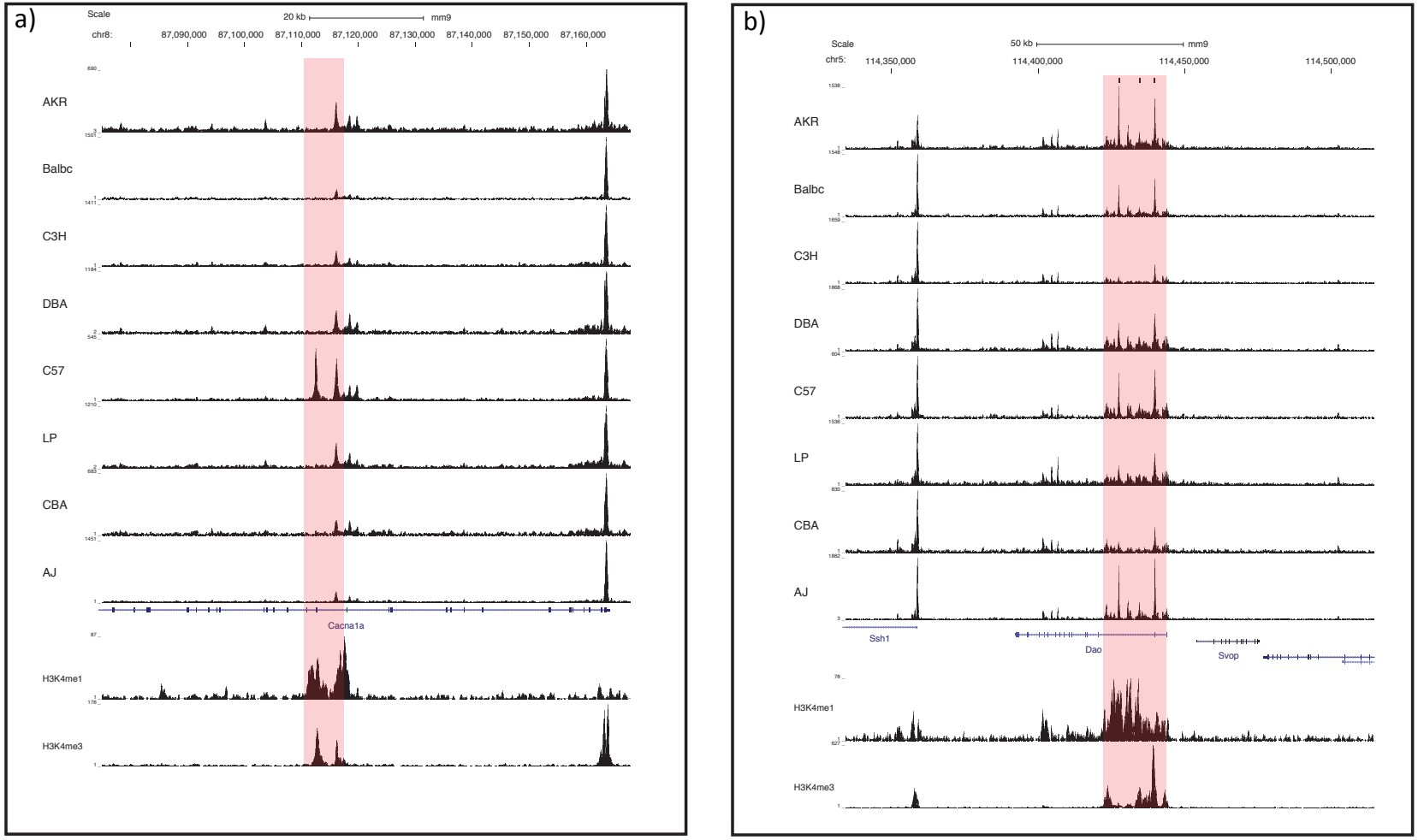
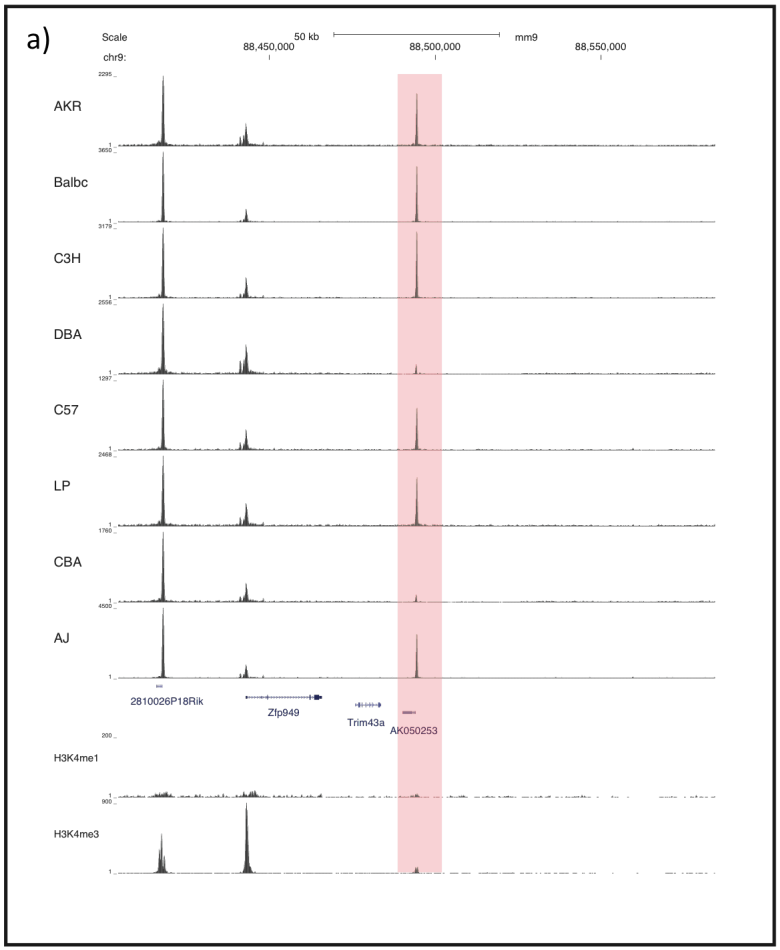


Figure 5-14 - Clustered variable peaks



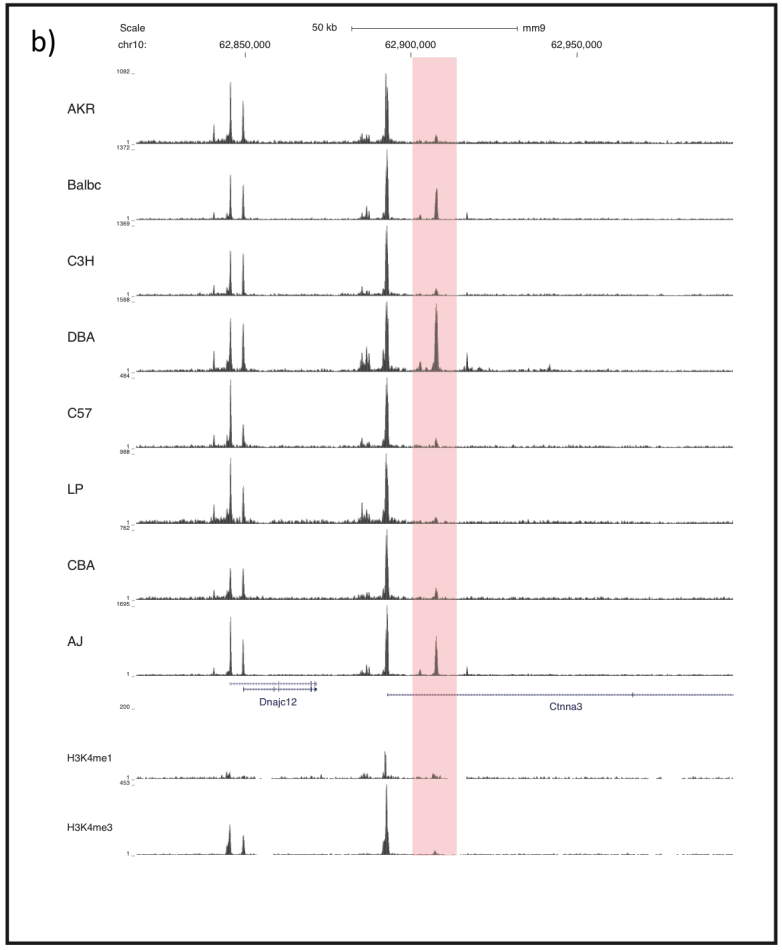


Figure 5-15 - Variation in size and shape

5.7 COMPARISON OF VARIABLE AND NON VARIABLE PEAKS

In order to investigate the characteristics of variable regulatory sites in more depth, we compared the 1,397 variable DHS peaks to 2,849 with least evidence for variation (P-value > 0.97) and noted the following features of variable DHS peaks:

First, they contain more sequence variants: 75% contain one or more previously identified sequence variants within 100 bp of the edges of the peak compared to 58.3% of non-variable DHS (P = 4.3e-27, Fisher exact test).

Second, they are less likely to coincide with conserved regions (using the multiple alignment of 30 vertebrate species and measurements of evolutionary conservation, UCSC browser track) of the mouse genome (61.5% compared to 71.4%, P = 1.07e-10 (Fisher exact test)).

Third, variable DHS were more likely to be associated with chromatin that has the signature of an enhancer (20.7 % compared to 9.7%, P = 5.6e-22 (Fisher exact test)), rather than a promoter (25.5% compared to 52.2%, P=3.1e-63 (Fisher exact test)) (here enhancers were classified as DHS overlapping a genomic region enriched for H3K4me1 and promoters for H3K4me3 enrichment in terminally differentiated mouse erythroblasts, using data from (Kowalczyk et al., 2012).

It is possible that our comparison between variable and non-variable peaks is biased by the choice of an extremely non-variable set of peaks. Therefore we compared variable peaks with all other peaks (excluding the most variable) and with the 16,299 peaks where analysis of variation gave a P value of 0.5 or greater. The results, presented in table 5-2 shows similarly highly significant differences in the same features as those obtained from the most non-variable

peaks. Note that as the set of peaks becomes less variable (as indicated by an increasing P-value in column one of the table) the relative enrichment in promoter elements, and conservation increases, as does the relative impoverishment in enhancers, and sequence variant. This is consistent with the hypothesis that the 2,849 DHS with least evidence for variation (P-value > 0.97) are simply an extreme, rather than an atypical set.

Table 5-2 - Characteristics of variable DHS peaks

Peak type	Number of peaks	Variants (%)	LogP	Conservation (%)	LogP	Enhancer (%)	LogP	Promoter (%)	LogP
Most variable peaks (P value < 0.008)	1,396	75.00		61.53		20.63		25.50	
All peaks (P value > 0.008)	24,303	61.05	25.02	70.04	10.04	13.91	10.13	46.22	52.01
All peaks (P value > 0.5)	16,299	59.11	28.95	71.37	12.09	10.78	20.85	50.74	68.33
Least variable peaks (P value > 0.97)	2,849	58.25	26.37	71.43	9.98	9.72	20.84	52.12	62.07

The table shows features of the most variable peaks (identified at FDR threshold of 1%, equivalent to an adjusted P value of 0.008), compared to non-variable peaks. The features shown are the percentage of sites containing sequence variants (Variants), the percentage of sites showing sequence conservation (Conservation), the percentage of sites with the signature of an enhancer (Enhancer) and those with the signature of a promoter (Promoter). The P value is that obtained from estimating peak differences between strains (modeled as a negative binomial distribution). Results are shown for comparisons between the most variable peaks and three sets of non-variable peaks, defined at different stringencies (P values > 0.008, 0.5 and 0.97). The significance of the comparison is shown as the negative logarithm (base 10) of the P value of a Fisher exact test.

5.8 THE RELATIONSHIP BETWEEN SEQUENCE VARIATION AND DHS PEAK VARIATION

A central aim of my work was to relate variation in DHS with sequence variation. With a set of variable peaks I was in a position to address this aim. To do so, we took the set of sequence variants in the eight inbred strains and identified these within the regions containing the variable peaks. The assumption is that a sequence variant, which is responsible for the peak variation, will follow the same strain distribution pattern as the DHS variation. For example if there is peak in strains A/J and C3H/HeJ, but not in the other six strains, we expect to find a sequence variant present in A/J and C3H/HeJ but not in the other six strains. One way to represent a strain distribution pattern (SDP) is as a binary vector, with 1 and 0 representing the presence and absence of a DHS, or a sequence variant. We can then express the relationship between DHS and sequence variation as a correlation between these two vectors. I begin this section by presenting specific examples where I identify the relationship between DHS and sequence. I then go on to describe the results of a genome wide comparison, using the correlation between SDPs.

Figures 5-16 - 5-18 once more show examples of the three categories of variable DNase hypersensitivity sites. The position of a DHS missing in one or more strains is highlighted in red. The position of annotated Refseq genes are named and shown in blue. In addition to DHS data in each inbred, they also include the strain specific genotype of the DNA underlying the variable peaks in the lower section with a gray bar representing the portions which is identical to the current mm9 build; however differences are represented by black lines. The top part shows chromosomal positions in Kb (MGSC37 mouse reference), while

the lower section gives the chromosomal positions in bp over the highlighted DHS. The density of aligned DHS-seq reads in a moving 300bp window, with a 30bp increment is shown for each mouse strain.

Figure 5-16 shows the identical example of category I variable DHS peak (i.e. "peaks are entirely absent in one or more strains") that we observed in Figure 5-13 panel b). As described before, this DHS peak on chromosome 10 (coordinates: 121,005,821 - 121,006,767 bp) is completely absent in BALBc/J, A/J, and DBA/2J.

This is a very nice example of variable DHS peaks, in which the DHS strain distribution pattern is in complete agreement with the underlying genotype strain distribution pattern. It is clearly evident in the lower section of Figure 5-16 that the strain specific genotype of these three inbred strains is identical and varies from the other five inbred (AKR/J, C3H/HeJ, C57BL/6J, LP/J, and CBA/J). Any of these 9 sequence variants, represented with black lines, could be considered as a candidate causal sequence variant for variant DHS as the strain distribution pattern vector is 100% compatible with DHS strain distribution pattern. They could be described as reciprocal identical (DHS SDP for this site (i): 10111010 considering the order of strains as displayed in this figure and genotype SDP for all these sequence variants are: 01000101). In other words where we have this sequence variants the variant DHS peak (i) is lost.

Figure 5-17 shows a new example of category II variable DHS peaks (i.e. "clusters of variable peaks") on chromosome one (cluster coordinates: 194,819,390 - 194,863,527 bp). This category is harder to interpret as we would see multiple variable peaks within the vicinity of 10 kb from one another. Two major peaks within the cluster highlighted with deeper red for showing their

strain specific underlying genotype variation in lower section (ii, and iii) (coordinates: 194,827,602 - 194,828,563 bp and 194,860,480 - 194,863,156 bp respectively). Unlike the previous example in Figure 5-16, where we could see a very definite association between DHS strain distribution pattern and underlying genotypic strain distribution pattern, we could not fully explain relate variation in DHS SDP and to underlying genotype SDP; however, in case (ii) we could observe moderate similarities between genotype SDP of AKR/J and BALBc/J which expressed more similarity on their DHS peak morphology. The same applies to CBA/J and DBA/2J. Overall, these two variable DHS peaks (ii and iii) are two examples where DHS and genotype SDPs are not consistent with each other.

Figure 5-18 shows a new example of category III variant DHS peaks (i.e. "peaks vary in shape or size") on chromosome 7 (coordinates: 30,808,864 - 30,809,463 bp). Peaks in this category are similar to category I as they are single and not multiple (like we classify as category II), but the variation varies from what we categorized in type I. Here no DHS peak is absent in inbred strains but we could observe quantitative variation in the size of DHS peaks in most cases, or DHS peaks morphology varies among inbred strains in minority of cases (which by nature is a qualitative change). As in most cases of this category variation is in the form of continuous quantitative change it is less obvious to categories variation in 0 and 1 as we easily done for the example of category I in Figure 5-16; however, if we arbitrarily set a threshold to divide DHS peaks into more obvious peaks or bigger and assign "1" to those and less obvious peaks or smaller ones and assign "0" to them in this example (we could do the same if we can recognize two various shape of DHS among inbred strains), "11001101"

could summarize DHS strain distribution pattern in this example according to the order of strains as presented in this figure. To find candidate causative sequence variants we investigated genotype strain distribution patterns of sequence variants in the area under the variant peak and eight sequence variants, which had "identical" SDP with DHS SDP (i.e. "11001101")"were marked by green arrows. In this example in all these eight sequence variants, we observed taller peaks associated with the sequence variants and smaller peaks when the sequence variant is absent, which suggest this variants affected chromatin structure at this site.

5.8.1.1 Candidate causal variants for DHS variation identification

To identify the probable causal variants for DHS variation, we correlated sequence and DHS peak variation. However, it should be noted that large shared haplotype blocks in the inbred strains used here typically make it impossible to unequivocally identify single causative variants because many variants may share the same SDP. Bearing in mind these caveats, we assessed the correlation between sequence and DHS strain distribution patterns at all variable peaks. 503/1397 (36%) of DHS peaks are associated with a sequence variant located within the peak or 100bp either side at an FDR of 5%.

We chose to search for variants within 100bp of the DHS to take into account the fact that nucleosome occupancy at the edge of the sites is not entirely fixed. Ideally we would like to look further, but our ability to do so is limited by the small number of haplotypes we are examining in the eight inbred strains (there are only 127 possible combinations of di-allelic variants in 8

strains). The further we extend our search out from the DHS, the greater the chance that we will find an association by chance. Figure 5-19 shows a linear increase in the percentage of associated sequence variants (up to to 70% at 100 kb). We are unable to tell which of these represent true associations.

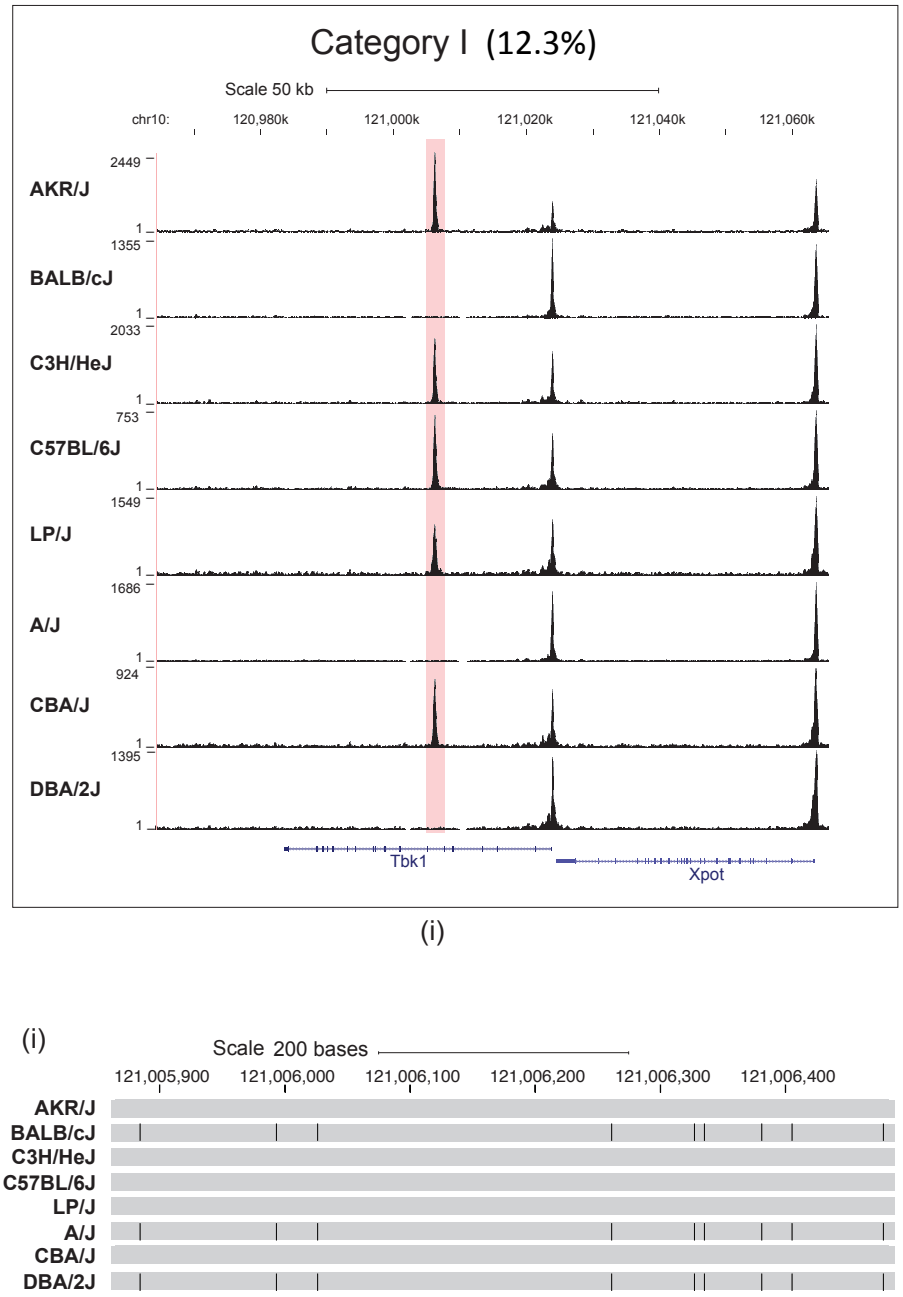


Figure 5-16 – Category I variable DHS: peaks entirely absent in one or more strains. The position of a DHS missing in one or more strains is highlighted in red. The position of annotated Refseq genes are named and shown in blue. The strain specific genotype of the DNA underlying the variable peaks is shown below with a grey bar representing the portions identical to the current mm9 build; differences are represented as black lines. The top part shows chromosomal positions in kb (MGSC37 mouse reference), while the lower section gives the chromosomal positions in bp over the highlighted DHS. The density of aligned DHS-seq reads in a moving 300bp window, with a 30bp increment is shown for each mouse strain.

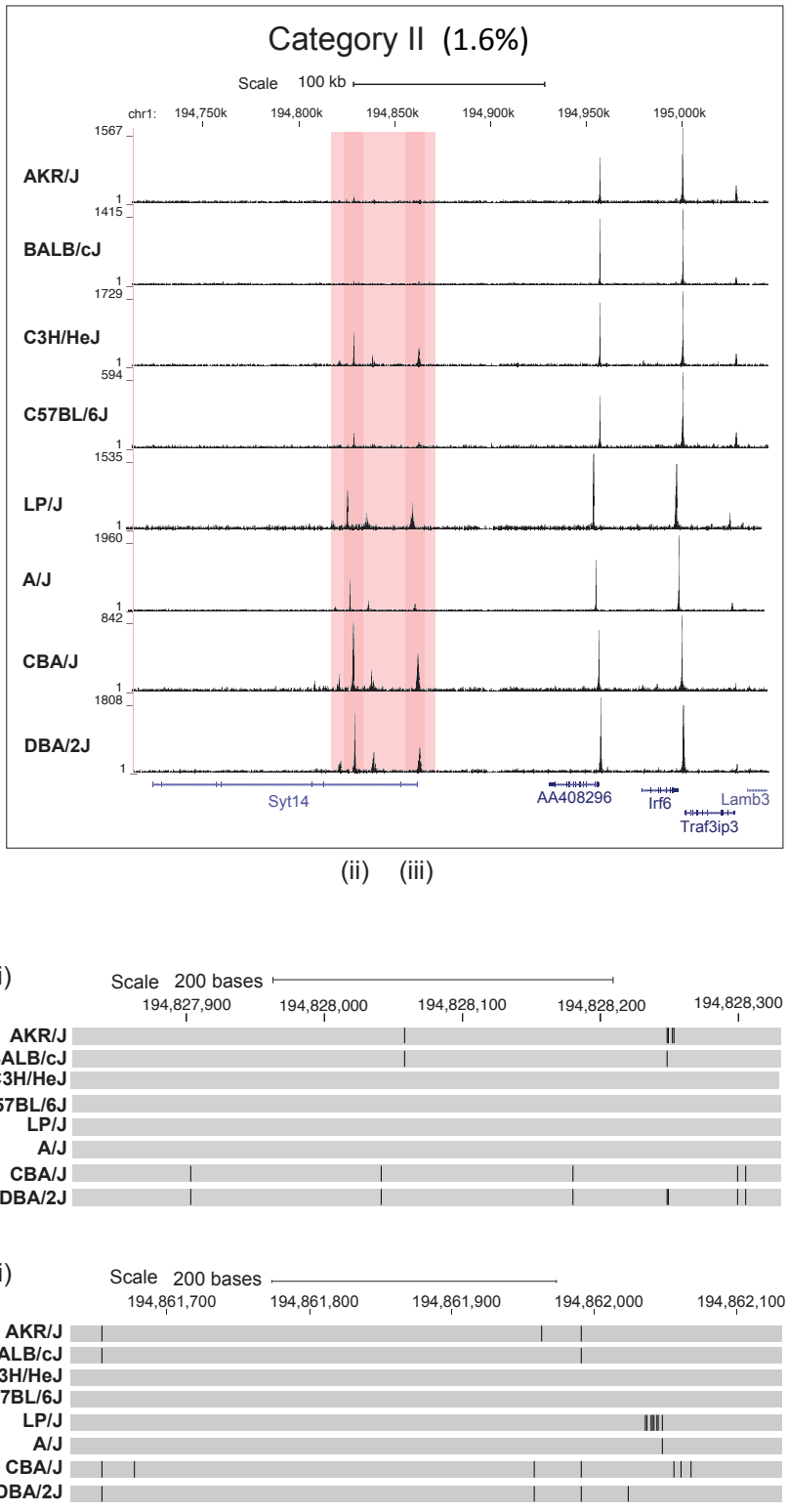


Figure 5-17 - Category II variable DHS: clusters of variable peaks. Positions of DHS sites that vary together with the underlying strain differences and the structure of top part and lower section are displayed as described for figure 17. The lower section gives the chromosomal positions in bp over the deeply red highlighted DHS (ii and iii).

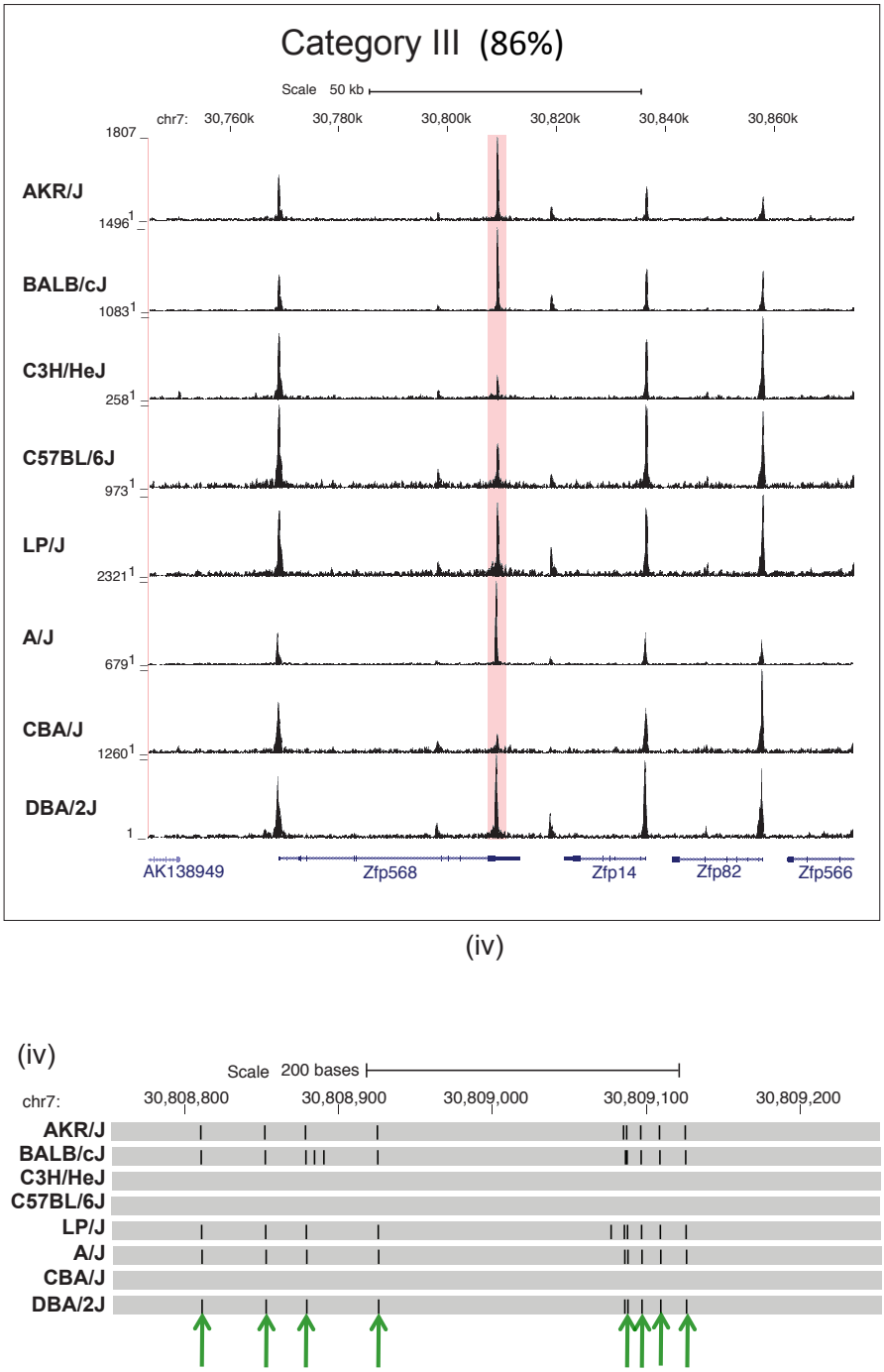


Figure 5-18 – Category III variable DHS: peaks vary in shape or size. The position of DHS sites together with underlying strain differences and the structure of top part and lower section are displayed as described for figure 17. Candidate causative sequence variants for the variant DHS (iv) are marked by green arrows.

We asked next whether sequence variants potentially causal for peak variation are more likely to be found for discretely varying peaks (the set of 196 peaks where there is either total loss or gain of a peak in at least one strain). Of the 196 selected DHS, 113 (57%) contain a sequence variant within 100bp of the peak whose strain distribution pattern correlates with that of the variable DHS (at an FDR of 5%). This is significantly different from the set of continuously varying peaks (P-value 1.18e-8, Fisher exact test). Four discretely varying DHS coincide with a structural variant (one deletion and three insertions). Table 5-3 summarizes the details of these discrete variable DHS peaks.

Table 5-3- details of 4 structural variants underlying variable DHS peaks

No	Category	Variable DHS Peak Coordinate			Variant Type	Variant SDP*	DHS SDP*
		Chr	Start	Stop			
1	I	4	62068853	62069152	Deletion	11110110	00001001
2	I	7	71432287	71432586	Insertion/ SINE/SINE	00000010	11111101
3	I	13	112806496	112806796	Insertion/ SINE/SINE	10000000	01111111
4	I	13	114633545	114633844	Insertion/ IAPTypeI/ ERV	00000011	11111100

* Strain Distribution Pattern of eight inbred strains in the following order: A/J, AKR/J, BALBc/J, C3H/HeJ, C57BL/6J, CBA/J, DAB/2J, LP/J.

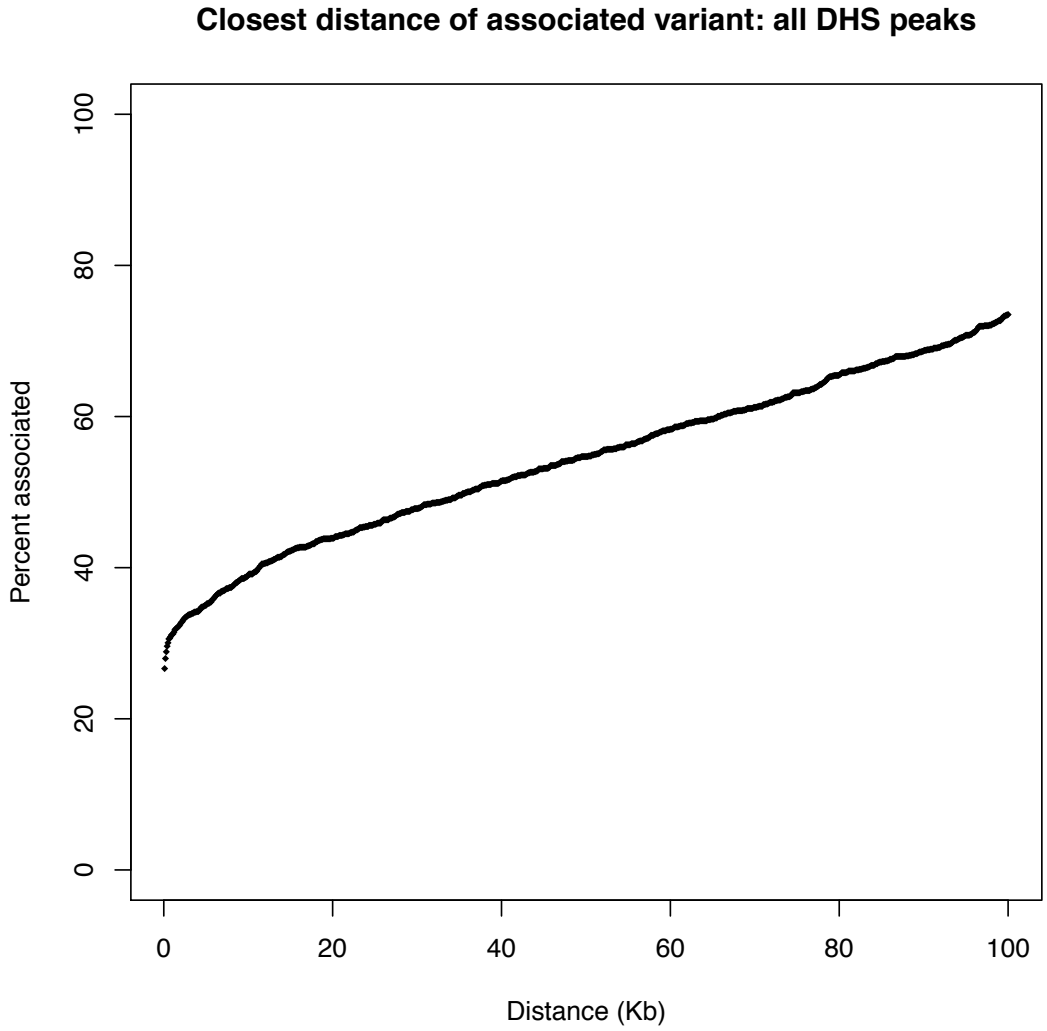


Figure 5-19 - Closest distance from a variable DHS peak to a significantly associated sequence variant. The horizontal axis shows the distance in kilobases (Kb) from each of the 1,397 variable DHS to a sequence variant that is associated at a 5% FDR. The P value is calculated by associating the strain distribution pattern of the DHS with the strain distribution sequence variant. We report here the distance to the closest variant (because of the haplotype structure of the inbred mouse genome there are often large haplotype blocks with identical strain distribution patterns). The vertical axis shows the percentage of DHS peaks for which an association is found.

5.9 CORRELATION OF DHS VARIATION AND TRANSCRIPTION FACTOR BINDING SITES (TFBSS) VARIATION

Next question was to figure out the probable mechanisms through which these regulatory sequence variants (i.e. variable DHS associated with underlying sequence variation) render their effects (possibly as variation in gene expression and phenotypic variations). One probable could be through alterations in TFBSs. A regulatory sequence variant could possibly alter a TFBS motif or might be able to alter a previously neutral site for a TF as a novel motif to bind to.

To test this, we examined known TFBSs motifs of erythroid transcription factors (TFs including: GATA1, KLF1, NFe2, Tal1) (see Methods). 165 of these 503 variant DHS peaks (12% of the total) contained one or more variants that disrupted a known motif for transcription factor binding site in one or more strains. Among discretely varying DHS peaks 42/196 (21%) contain a known TFBS.

5.10 CONCLUSION

Automated analysis of 19 genome-wide DNase-seq data from terminally differentiated erythroid cells from eight inbred mouse strains showed 1,397 out of total 25,700 automatically called DHS peaks (~ 5.4%) were variable (FDR 1%). We tested if DHS variation could be attributed to underlying sequence variation by comparing within and between strains DHS variation. Correlation testing between sequence variation and DHS peak variation showed 36% (503/1,397) of variable peaks contain a sequence variant with compatible strain distribution pattern. Our findings appear to indicate that about 65% of variable sites are not correlated with sequence variation. However sequence variation showed higher correlation with discretely variable peaks: 113/196 (57%) discretely variable peaks contain a candidate sequence variant.

In order to understand the effect of DHS variation we examined if variable DHS containing candidate sequence variants disrupt TFBSs. 165 of these 503 variable DHS peaks (12%) contain a sequence variant that disrupts TFBSs. This might explain the mechanism through which variable DHS render their effect on downstream gene expression in subgroup of variable peaks. Further consequences of variable DHSs on gene expression (RNA expression profiles) and phenotypic variation will be discussed in the next chapter.

CHAPTER 6 : FUNCTIONAL CONSEQUENCES OF DHS VARIATION

So far I have shown that DHS varies between inbred strains and I have studied how much of this variation relates to underlying genetic variation on a genome-wide scale. In this chapter I discuss the functional consequences of alteration in DHS.

Given that DHS mark regulatory regions of the genome, one obvious hypothesis is that changes in DHS would be reflected in changes in transcription. Thus if we examined a region of the genome with a DHS site present in one strain but absent in another, I expected that we would see an appropriate change in the transcription of a nearby gene. For example if chromatin over the promoter at a gene is in an open conformation, then transcription is permissible, but if the chromatin is closed (as indicated by the absence of a DHS site), then transcription will not occur. Assessing the consequences of DHS variation at enhancer elements is likely to be more complex, since the association between an enhancer and its cognate gene is not straightforward (sometimes the enhancer lies megabases away from the gene it regulates) but I still expected to be able to detect some changes in the expression of nearby genes.

In order to maximize the chances of observing an associated change, I used both Poly A- and Poly A+ subsets of RNA. Nascent pre-mRNA transcripts undergo multiple processing events during their maturation. The addition of a poly(A) tail, added to the 3' end of almost all eukaryotic mRNAs, is important for mRNA stability, nucleocytoplasmic export, and translation. The poly A- fraction represents the unprocessed fraction, and provides a more dynamic view of the RNA that the cell transcribes (Yang et al., 2011, Moore and Proudfoot, 2009, Shatkin and Manley, 2000, Kowalczyk et al., 2012).

A second issue I sought to address was the relationship between variable DHS and quantitative trait loci (QTLs). It was possible that I could detect a pathway relating sequence variation to a change in a DHS, which in turn would alter transcription and finally a change in a phenotype. To answer this question I had access to a 843 QTLs mapped in a heterogeneous stock, a set of quasi outbred animals derived from the same eight inbred strains for which I had DHS data. These data would allow me to examine the association between the different layers of information, from sequence to phenotype.

6.1 RNA-SEQ LIBRARY PREPARATION AND DATA ANALYSIS

To study RNA expression in inbred terminally differentiated erythroblasts I used RNA-Seq. Total RNA extracted from spleens of phenylhydrazine treated female mice of the same eight inbred strains (i.e. A/J, AKR/J, BALBc/J, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J, LP/J) and within the same age range in four biological replicates. Total RNA extraction was performed using QIAGEN RNeasy Mini Kit (replicated 1-3) and an in-house Trizol method (replicate 4). To have an indication of animal response to phenylhydrazine treatment Ter119+ cells was counted from a fraction of harvested spleens.

Table 6-1 summarizes the frequency of Ter119+ cells (%) in cells harvested from the spleens of treated mice and gives the RNA Integrity Number (RIN) measured by an Agilent bioanalyzer. This reflects the quality of total RNA obtained from the samples. Commonly an RIN cutoff 7 is used to select total RNA samples to proceed. As I planned to apply poly-A selection and globin depletion steps downstream on our samples, I decided to begin with a higher quality total RNA, so RNA integrity number (RIN) => 9 was chosen as cutoff.

Since globin is highly expressed in terminally differentiated erythroblasts, globin depletion is a critical step to avoid masking the expression of other genes in these cells by globin expression. I initially prepared three replicates per inbred strain samples. With the exception of LP/J-2 (INR=6.8), all samples had RIN>7 but only two samples (C3H/HeJ-2 and CBA/J-2) fulfill RIN>9 criterion; therefore, we extracted total RNA from the 4th replicate with an optimized RNA extraction in-house Trizol method. Eventually samples with the highest quality (higher RIN) were obtained suitable for library preparation (marked by * in Table 6-1). Total RNA was split into poly(A)$^+$ and poly(A)$^-$ RNA fractions. RNA-Seq libraries were prepared from both fractions. Globin depletions was performed on poly(A)$^+$ fraction and ribosomal transcript depletion on poly(A)$^-$ prior to library preparation (see methods). Poly(A)$^+$ and poly(A)$^-$ libraries were sequenced on Illumina HiSeq2000 and GA-IIx respectively.

The quality of Poly(A)$^+$ RNA samples was checked prior to sequencing. This assessment included measuring contamination with ribosomal RNA. C3H/HeJ-2 sample, which was the only submitted sample, which prepared by different method (QIAGEN RNeasy Mini Kit) showed 10-15% contamination with ribosomal RNA. Since the ribosomal RNA contamination level of C3H/HeJ-4 was even higher, eventually C3H/HeJ-2 proceeded for sequencing.

RNA-Seq read alignments were filtered for quality. Table 6-2 summarizes the quality features of poly(A)+ RNA samples. Sequence reads overlapping with transcripts were counted based on transcript and exon coordinates of mouse reference assembly mm9 available on UCSC. RNA reads were mapped to n = 26347 (out of 37681) transcripts.

Table 6-1 – Frequency of Ter119+ cells (%) harvested from spleens of phenylhydrazine treated female mice from eight inbred strains and RNA Integrity Number (RIN) of total RNA obtained from these samples.

	Ter119+ % (RIN of total RNA extracted from these samples)			
	Replicate			
	1	2	3	4
A/J	77 (6.7)	78 (7.6)	78 (7.4)	*99 (>9)
AKR/J	88 (7.4)	83 (7.1)	89 (7.3)	* 83 (9.9)
BALBc/J	78 (7.5)	80 (8)	79 (7)	* 84 (9.6)
C3H/HeJ	75 (8.6)	*56 (9.4)	72 (8.8)	86 (>9)
C57BL/6J	62 (8.6)	75 (7.5)	74 (8)	*100 (>9)
CBA/J	+ (8.7)	79 (9)	72 (8.4)	*79 (>9)
DBA/J	+ (7.9)	+ (7.3)	+ (8.1)	* 86 (9.7)
LP/J	87 (7.5)	78 (7.6)	85 (6.8)	*82 (>9)
+ Missing data				
* Those selected eventually for RNA-seq				
Total RNA extraction done by QIAGEN RNeasy Mini Kit (replicated 1-3) and in house Trizol method (replicate 4)				

Table 6-2 – Summary of sequencing quality features (HiSeq2000) of poly(A)+ RNA samples

Sample	QC analysis done by sequencing facility					Number of reads (Million read)	MAPQ>=15§ Mapped to BL6 Million read (%)
	Length (bp)	Insert ±MAD*	Yield (Million read)	Yield (MbQ20)	Mapped to BL6 (%)		
A/J	100	162 ± 37	27.34	2640.15	99.0	54.7	31.4 (57.5)
AKR/J	100	163 ± 39	25.46	2463.22	99.3	50.9	33.1 (65.0)
C3H/HeJ	100	162 ± 36	26.00	2510.86	99.2	52.0	31.1 (59.7)
BALBc/J	100	163 ± 40	26.92	2605.47	99.2	53.8	35.5 (66.0)
DBA/2J	100	163 ± 41	21.83	2111.67	99.2	43.7	28.9 (66.3)
LP/J	100	165 ± 46	24.22	2344.50	99.2	48.4	34.5 (71.3)
C57BL/6J	100	164 ± 41	20.94	2025.60	99.2	41.9	26.1 (62.3)
CBA/J	100	164 ± 44	23.27	2253.45	99.3	46.5	32.7 (70.2)

* MAD stands for median absolute deviation (a robust estimator of variance)

§ MAPQ >= 15: reads with mapping quality greater or equal to 15 were mapped to C57BL/6 genome and used for RNA poly(A)+ analysis

6.2 RELATIONSHIP BETWEEN DHS AND RNA EXPRESSION VARIATION IN ERYTHROBLASTS

Using RNA-Seq, genome-wide RNA abundance and inter-strain variability was estimated in erythroblasts from the same eight strains used for the DHS analysis. We hypothesized that changes in both RNA expression as well as RNA processing are more likely to be found associated with variable DHS than with non-variable DHS.

We identified 3,472 poly(A)+ transcripts that varied between strains (using an FDR of 5%) (see Methods). We then asked whether transcripts that vary in abundance are equally likely to be found close to a variable as to a non-variable DHS. If variable DHS have no relation to variable transcripts, then the distances from variable transcript to variable and to non-variable DHS should not differ. Note that this analysis makes no assumption about what that distance might be. We calculated the shortest distance from the start of each variable transcript to variable DHS and to non-variable DHS (the 2,849 least variable). Within 10 kb of any variable poly(A)$^+$ transcript there are a total of 181 variable DHS (~13% of the total 1,397 variable DHSs) compared to 85 non-variable (~3% of the total 2,849 least variable DHSs) (P=3.4e-91, Fisher exact test). These 181 variable vs 85 non-variable DHS peaks significantly enriched in enhancers (17.7% vs 5.8%) and deficient in promoters (28.2% compared to 57.0%), a pattern consistent with that seen in the complete set of 1,397 variable DHS. Figure 6-1 shows the proximity of DNase I hypersensitive sites to variable RNA. X axis shows the distance from DHS to variable RNA in kb. Y axis shows the frequency of variable DHS (in red) and non variable DHS (in blue). Within the vicinity of 10 kb distance any variable poly(A)+ RNA transcript, there is

significantly higher frequency of variable DHSs vs non- variable DHSs; however increasing the distance between variable mRNA and DHS completely fades this difference. For instance at 30 kb distance slightly higher frequency of variable DHS vs non- variable DHS observed.

Proximity of variable RNA and variable DHS does not mean the two events are correlated so we tested whether this was so. We found that quantitative variation in peak height at 54 of these 181 variable DHS lying within 10 kb of variable poly(A)+ RNA transcripts was significantly ($P < 0.05$, uncorrected) correlated with variation in RNA abundance. If we assume that the effect of a variable DHS on a transcript is limited to transcripts within 10 kb,

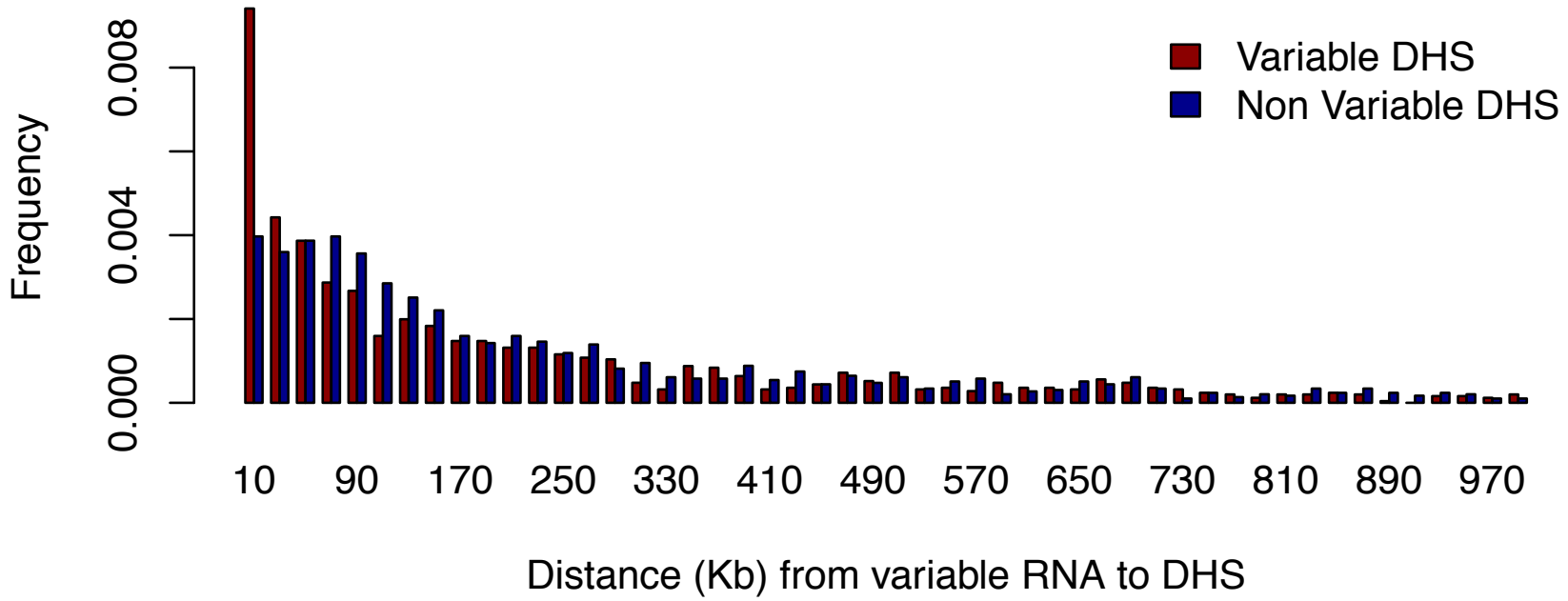


Figure 6-1 -Proximity of DNase I hypersensitive sites to variable RNA. The histogram shows the relative abundance of DHS within 1 Mb of the transcriptional start site of variable RNA species. The horizontal scale is in kb. Variable RNA refers to the 3,472 poly(A)$^+$ transcripts that varied between strains (FDR of 5%). Variable DHS refer to the 1,396 variable peaks (as described in the text), and non-variable refers to 2,849 with least evidence for variation.

then this result indicates that at most ~ 4% (54) of the 1,396 variable DHS affect transcription. This is a conservative estimate since the analysis does not take into account the number of false discoveries (5% of 181 = 36). Figure 6-2 shows an example where DHS variation and RNA poly(A)+ variation were significantly correlated on chromosome 13. DHS variation in A/J and CBA/J strongly correlated with variation of Emb gene expression.

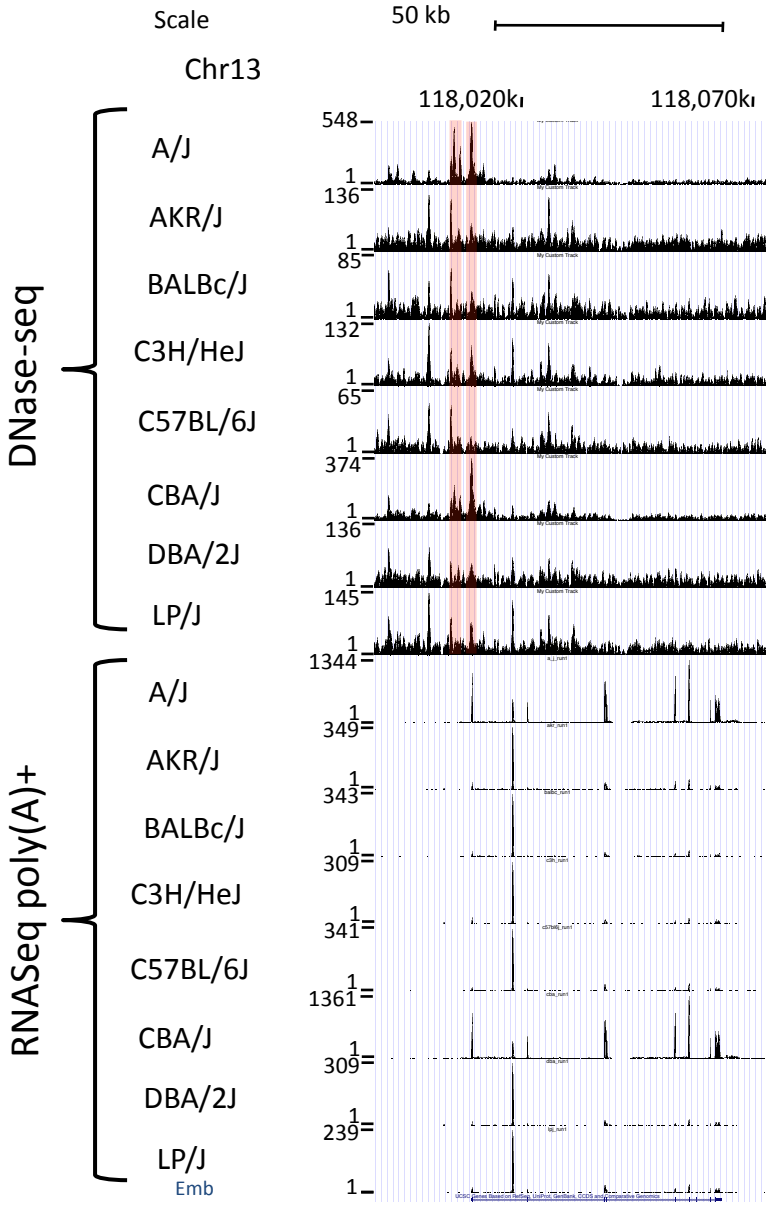


Figure 6-2 – An example of correlation between DHS variation and RNA poly (A)+ variation on chromosome 13. Two variable DHSs one at Emb gene promoter and one at its upsteam (highlighted in red) correlate with RNA expression variation (Emb gene) in A/J and CBA/J. DNase-seq raw data (unnormalized unadjusted) have been presented here.

Our findings appear to indicate that over 95% of variable DHS are not related to readily detectable variation in gene expression. One explanation for this finding is that we failed to detect the effects on transcription because they lie outside the 10 kb window we examined. Another explanation for this finding is that effects of some variable DHSs on transcription are subtle, and will be more evident with more manifest changes in DHS, such as at loci where a DHS is absent in a strain, or where a strain has a novel DHS. Furthermore, our assessment of RNA was limited to the poly(A)$^+$ fraction.

A more dynamic and comprehensive assessment of RNA expression may be obtained by analyzing poly(A)$^-$ RNA which represents nascent transcripts both from genes and regulatory elements. Assessing the Poly A- fraction might reveal a broader association between DHS and transcription.

We therefore analyzed the set of DHS for which an effect on transcription is likely most easily observed, namely that set of 196 peaks with either total loss or gain of a peak in at least one strain (this is the set of DHS with discrete, rather than continuous variation including 22 DHS with discrete compound variable peaks (Category II)). We visually inspected 300 kb either side of the peak for poly(A)$^-$ transcriptional changes and analyzed the relationship between poly(A)$^-$ transcripts and the 196 DHS. Figure 6-3 shows an example of where we observed consistent variation between DHS cluster and poly (A)$^-$ transcription in inbred mouse strains on chromosome one. It is clearly seen that DHS peak cluster and poly(A)$^-$ transcripts are both absent in AKR/J and BALBc/J strains and present in other inbred strains. At 14 loci (7%), of the 196 selected DHS, the gain or loss of a DHS was associated with a change in gene expression (Table 6-3).

For example on chromosome 13, at the Emb gene (embigin, a trans-membrane protein of the immunoglobulin super-family class of cell adhesion molecules) the occurrence of two DHSs (i and ii), one at the transcriptional start site (ii) and another 3,600bp upstream (i), are associated with transcription of the gene (Figure 6-4). For the sake of simplicity and presenting a less crowded figure DHS and poly(A)$^-$ data of four inbred strains included in this figure. C57BL/6J and LP/J as two examples of absence of DHS peaks and RNA transcript at this locus and CBA/J and A/J as two and only examples of presence of DHS and poly(A)$^-$ transcripts within eight inbred strains at this locus. At the bottom of the figure strain specific genotypes underlying DHS peaks displayed at all inbred strains. These four strains are labeled in red and starred. The candidate DNA sequence variants associated with the presence or absence of DHS peaks (marked by green arrows) are clearly correlated with the presence or absence of RNA transcripts (Figure 6-4). Four candidate sequence variants assigned for (i) variable DHS peak and three for (ii) variable DHS peak based on having identical sequence distribution pattern (SDP). Overall, approximately 7% of discretely variable DHS are associated with changes in transcription (either poly(A)$^+$ or poly(A)$^-$).

Table 6-3 - Summary of 14 cases of discretely variable DHS peaks and correlated poly(A)$^-$ RNA transcripts

Variable DHS case	Chromosome	Coordinates: Start-Stop	Variable Peak Category	Evidence of correlated variable poly(A)$^-$RNA expression within 300,000 bp either side of discretely variable DHS
1	1	194827926-194828226	II	Syt14 expression is absent in AKR/J and BALBc/J where the DHS cluster is absent (lose)
2	3	125678886-125679185	I	Ugt8a has lower expression in AKR/J, DBA/2J, C57BL/6J, and LP/J where DHS peak is absent
3	3	153352879-153353185	I	St6galnac3 is not expressed in LP/J where the DHS is absent (lose)
4	5	23356214-23356513	II	AKR/J, BALBc/J, DBA/2J expressed AI506816 whereas C3H/HeJ, C57BL/6J,CBA/J expressed ENSMUSG00000073777
5	5	109136787-109137087	I	Fgfrl1 is expressed in C57BL/6J where the DHS is present (gain)
6	10	107797855-107798155	II	Pawr expression is absent in C57BL/6J where DHS cluster is absent (lose)
7	11	58030239-58030738	I	Irgm2 lower gene expression in C57BL/6J where the DHS is absent (Stimulatory DHS)
8	12	107144916-107145215	I	Papola higher expression in BALBc/J where the DHS is absent (inhibitory DHS)
9	13	118009311-118009610	II	Emb is expressed in CBA/J and A/J where DHS cluster is present (gain)
10	14	70977245-70977745	II	Nudt18 is expressed in BALBc/J, LP/J and A/J identical with DHS Strain Distribution Pattern
11	14	51575743-51576042	I	Pnp2 expression (gain) and Pnp higher expression in C57BL/6J where the DHS peak is present
12	15	79804245-79804545	I	Apobec3 gene is more expressed in C57BL/6J and DHS is absent in C57BL/6J and LP/J
13	17	85333345-85333645	I	Lrpprc lower expression in DBA/2J where the DHS is absent (Stimulatory DHS)
14	18	67408499-67408799	II	Mppe1 expression is absent in BALBc/J and A/J where DHS cluster is absent (lose)

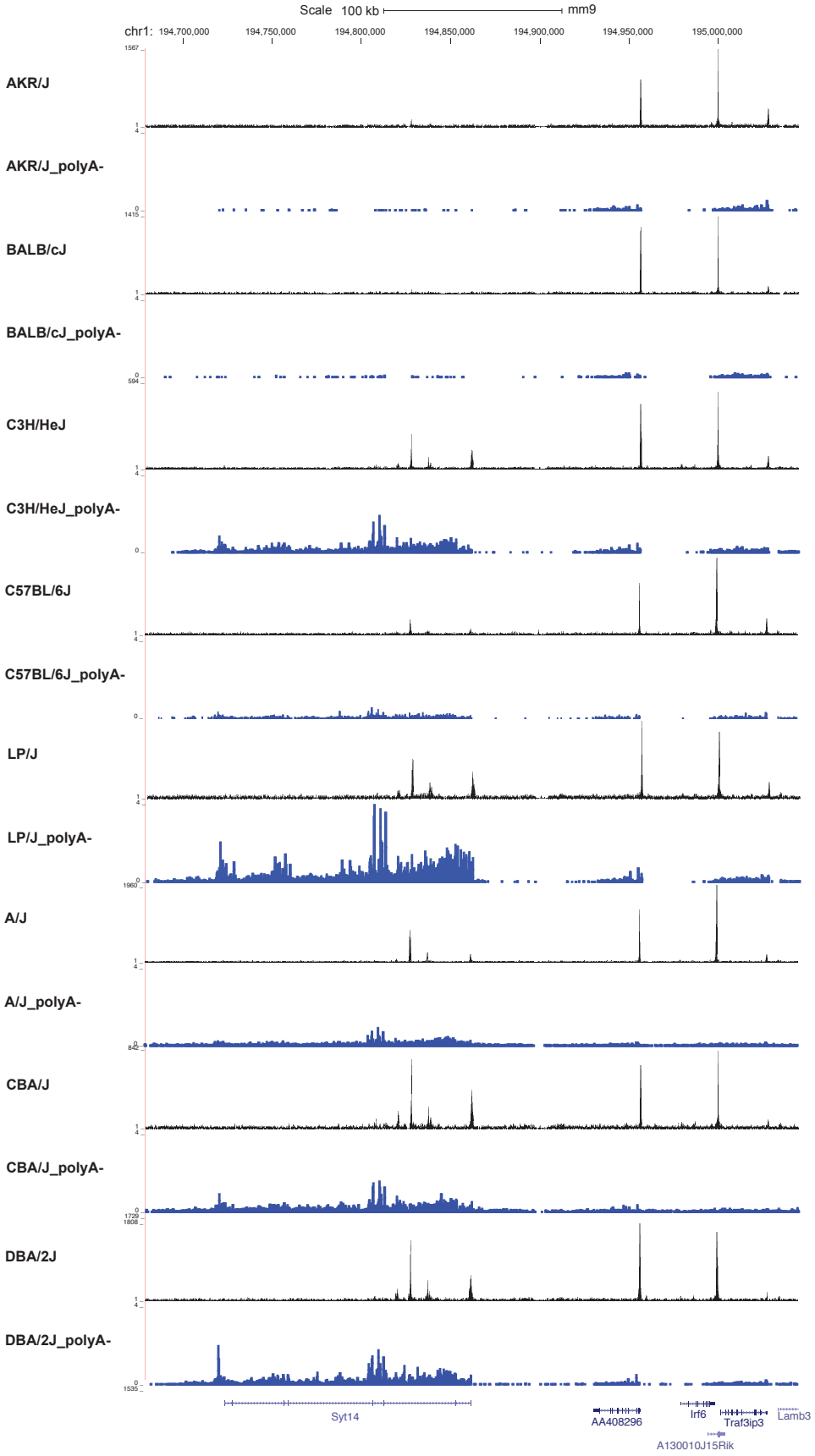


Figure 6-3 - Relationship between variation in transcription and a variable cluster of DHS peaks. Chromosomal positions are shown in kilobases. The density of aligned DHS-seq reads in a moving 300bp window, with a 30bp increment is shown for each mouse strain. The polyA$^-$ transcription from each strain is displayed (in blue) as the aligned read depth per base pair normalized per million of aligned reads.

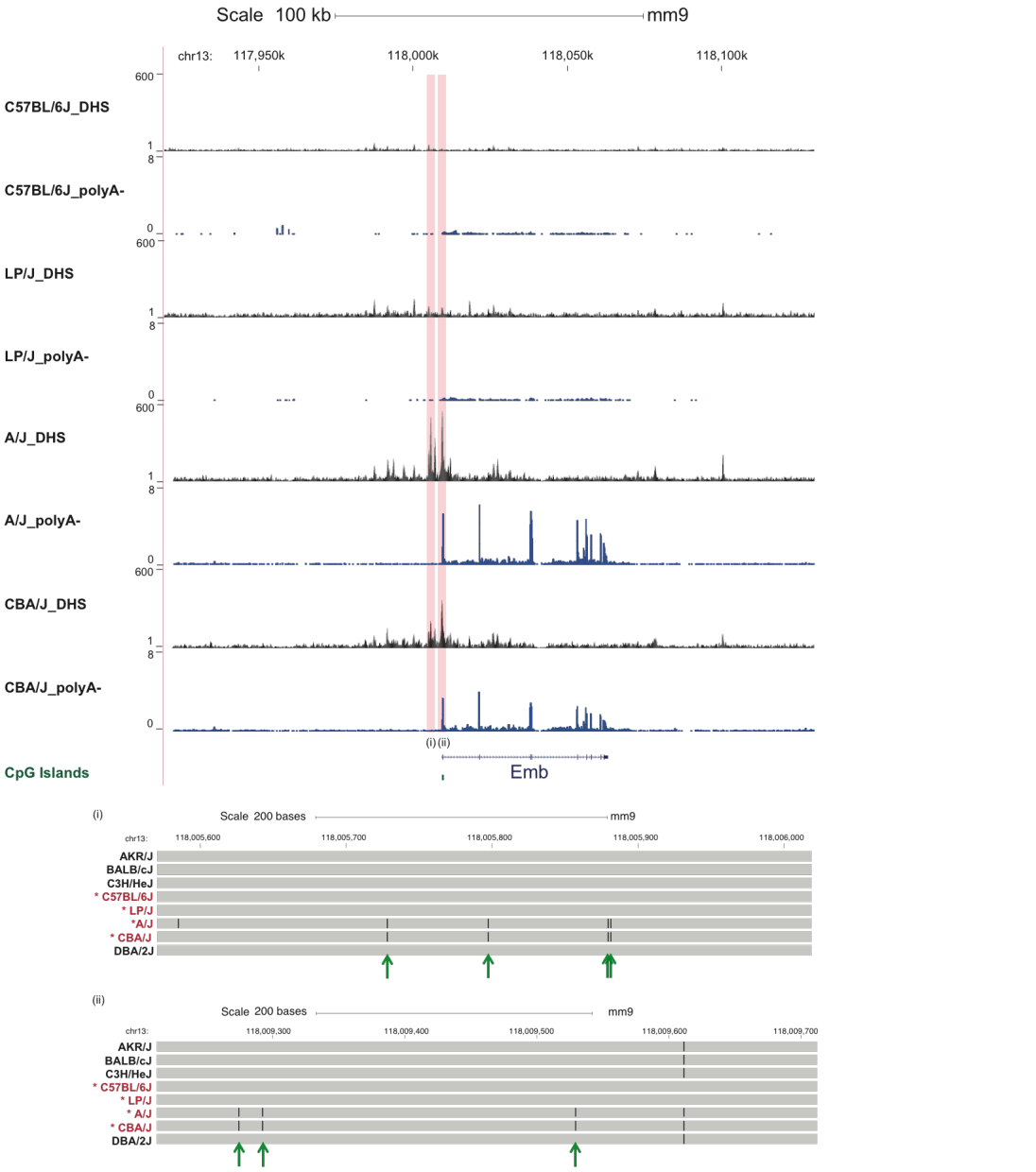


Figure 6-4 - Relationship between variable DNase I hypersensitive sites, underlying sequence and transcript variation. Scale and position in the MGSC37 mouse reference are shown as before. The density of aligned DNase-seq reads in a moving 300bp window, with a 30bp increment, is shown for the mouse strains C57BL/6J, LP/J, A/J and CBA/J. The poly(A)$^-$ transcription from each strain is displayed (in blue) as the aligned read depth per base pair normalized per million of aligned reads. The two variable DHS (one at the Emb gene promoter and one upstream of the promoter) are highlighted in red. The strain specific genotype of the DNA underlying the variable peaks is shown below (regions indicated by roman numerals) with a grey bar representing the portions identical to the current mouse genome build; differences are represented as black lines. Green arrows indicate presence of sequence variants correlate with variable DHS peaks and variable RNA transcripts.

Our results do not exclude the possibility that some DHS alter transcription without the presence of a sequence variant. We searched for examples by re-examining the entire set of 36,693 DHS. We identified 42 additional discretely variable (compound discrete/class II) peaks that did not appear to contain correlated sequence variants. In one case, we found a DHS present only in one strain (DBA/2J) associated with both poly(A)$^+$ and poly(A)$^-$ transcription (Figure 6-5). The novel transcript contains a re-activated LTR repeat element. No variants were detected consistent with the strain distribution pattern of the DHS (Figure 6-5) (failure to find variants could not be attributed to low sequence coverage or read mapping errors). Thus, in some circumstances, changes in both DHS and expression can occur in the absence of local sequence variation.

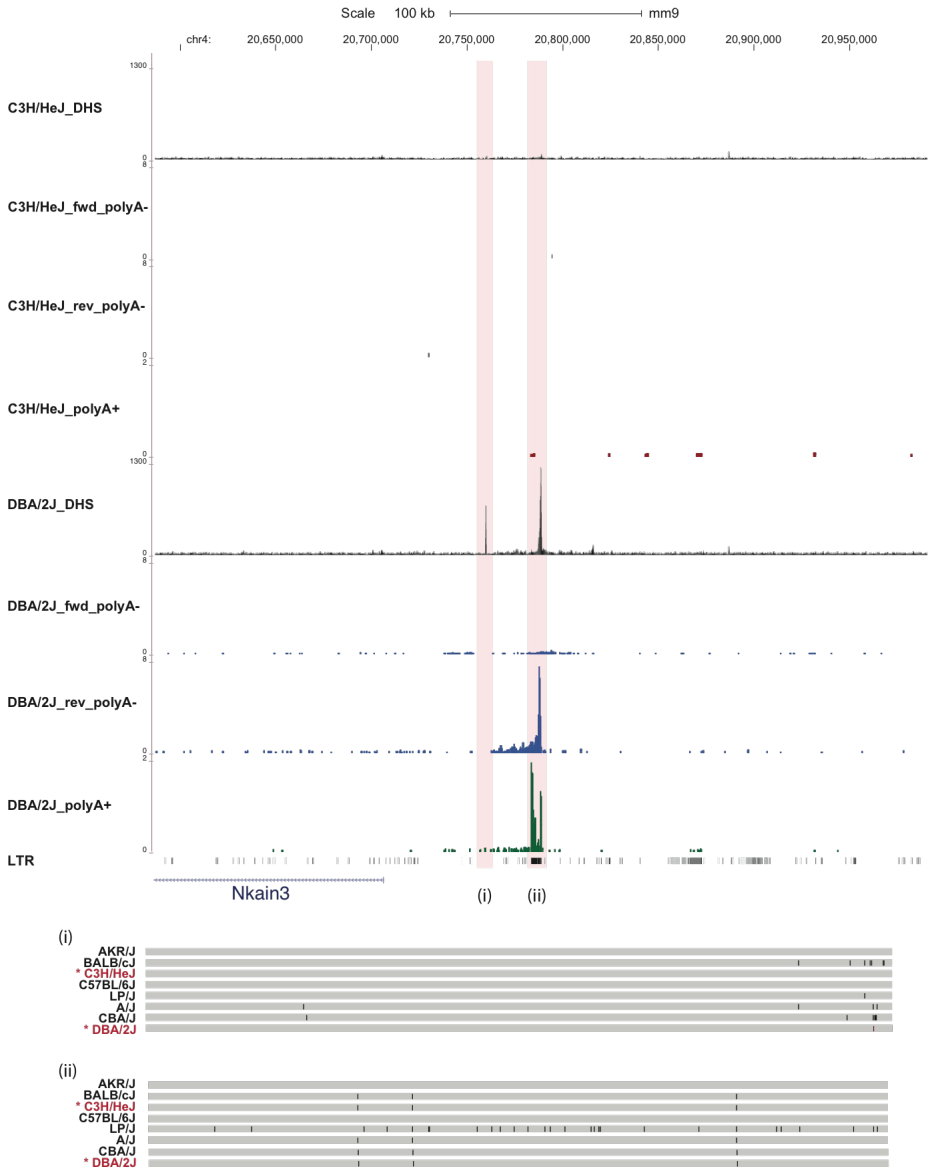


Figure 6-5 - Some variable DNase I hypersensitive sites have no underlying sequence variation. The figure shows one example on chromosome 4 where a DHS, present only in strain DBA/2J, is associated with both poly(A)- (blue) and poly(A)+ (green) transcription, in the absence of sequence variation consistent with the strain distribution of the DHS. The chromosomal position in is shown in kb. The density of aligned DHS-seq reads in a moving 300bp window, with a 30bp increment is shown for strains C3H/HeJ and DBA/2J. The strain specific genotype of the DNA underlying the variable peaks is shown below with a grey bar representing the portions identical to the current mm9 build; differences are represented as black lines. Note that there are no differences between C3H/HeJ and DBA/2J.

6.3 RELATIONSHIP BETWEEN DHS AND PHENOTYPIC VARIATION

Finally we investigated the relationship between DHS and complex traits phenotypic variation. We found that sequence variants underlying variable DHS are more likely to influence phenotypes than variants in non-variable DHS, using a test that discriminates between variants that could be functional and those that are not (Keane et al., 2011, Yalcin et al., 2005). The eight strains used in the DHS experiments were the progenitors of a population in which 843 quantitative trait loci (QTL) were mapped (Valdar et al., 2006). We took advantage of a prior analysis in which variants in the genomes of the eight strains analyzed here had been tested for causal involvement at each of 843 quantitative trait loci (QTLs) for over 100 different phenotypes (Yalcin et al., 2005, Valdar et al., 2006). For all sequence variants within a QTL that segregate between the eight strains, we tested if the SDP of the variant, when imputed across the mapping population of HS mice produced genotypes that were associated with the trait. The results provide data similar to that from human genome wide association studies, namely a set of positions in the genome that are candidate functional variants. We term these candidate variants.

6.3.1 TESTING THE ENRICHMENT OF CANDIDATE VARIANTS IN VARIABLE/NON-VARIABLE DHS SITES WITHIN QTLS: DISCRIMINATING FUNCTIONAL VARIANTS

We asked if variable DHS sites within the QTL are enriched for candidate variants (as defined above). Within QTLs, we identified a total of 2,813 variants for 1,397 variable DHSs and 3,408 variants for 2,849 non-variable DHSs. There were 199 candidate variants within the variable DHS peaks and 188 within the

non-variable DHS peaks. The numbers of non-candidate variants in the two types of peaks are 2,614 and 3,220 respectively. Therefore, functional variants are enriched in variable DHSs within QTLs (P = 0.0131, Fisher exact test) (Table 6-4).

Table 6-4 – Analysis of candidate functional variants at DHSs

		Variable DHSs (1,397)	Non-variable DHSs (2,849)
QTLs with	At least one functional variants	74	76
	Non-functional variants	720	1074
	No variants	602	1,699
Total functional variants		199 *	188
Total non-functional variants		2,614	3,220
Total variant count		2,813	3,408
* Significant enrichment of functional variants at variable DHSs vs non-variable sites (P= 0.01310 Fisher exact test)			

This indicates that variants lying at variable DHSs are more likely to have an effect on phenotypes compared to variants at non-variable DHSs.

6.4 CONCLUSION

In this chapter I addressed the functional consequences of variable DHS in terminally differentiated erythroblasts in eight inbred mouse strains by looking at RNA expression data (RNA-Seq) and phenotypic variation (QTL) data.

If we assume the effect of variable DHS is limited to 10 kb, 4% of variable DHS (54 out of 1,397) correlated with variation in RNA abundance. Our results indicate that over 95% of the variable peaks did not show detectable variation in gene expression. Among 196 discretely variable DHS peaks (which we expected to observe more prominent effect in gene expression) 14 cases (7%) are associated with variation in RNA expression (either poly(A)$^+$ or poly(A)$^-$) and in these cases underlying DNA sequence clearly correlates with the presence/absence or variation in RNA abundance. We also showed an example where both DHS and gene expression variation occurred in the absence of DNA sequence variation. In terms of the relationship between variable DHS and phenotypic variation, we showed variable DHS are enriched for sequence variants that contribute to complex traits in mice. These findings indicate that DNA sequence variants exist that influence variation in chromatin structure (DHS), gene expression and complex traits.

CHAPTER 7 : DISCUSSION

In this thesis I have described the results of an analysis of genome-wide DHS obtained from eight inbred mouse strains (A/J, AKR/J, BALBc/J, C57BL/6J, C3H/HeJ, CBA/J, DBA/2J, and LP/J). My aim was to use this information to understand more about the relationship between sequence variation, regulatory element variation and phenotypic variation. I found that among the eight inbred mouse strains about 5% of DHSs are reproducibly variable between strains (at an FDR of 1%). By comparing variable DHS peaks to non-variable peaks, we found that variable peaks are more likely to occur in regions bearing the chromatin signature of enhancers rather than promoters, are more likely to contain sequence variants and less likely to occur in conserved genomic regions. Surprisingly we observed that only 36% of variable DHS sites can be explained by a sequence variant either within the site, or 100bp either side. This observation appears to rule out simple models in which a sequence variant locally alters chromatin conformation.

I then looked at the functional consequences of site variation. On the assumption that changes in a gene regulatory element would alter gene regulation, I analyzed the amount of transcript produced from each gene, and expected to see changes in transcriptional abundance associated with changes in DHS. However changes in DHS result in relatively limited detectable changes in RNA expression: only 7% of discretely variable peaks are correlated with changes in transcriptional activity. Finally we find that sequence variants within variable DHS are more likely to be associated with complex traits than sequence variants lying in non-variant DHS.

7.1 VARIABLE DHSS PREVALENCE

I showed ~5% of DHSs are variable among inbred mice terminally differentiated erythroblasts. My estimate of the proportion of DHS sites that is variable is consistent with those obtained from human lymphoblastoid cell lines (McDaniell et al., 2010, Kasowski et al., 2010). However I reported fewer sites than identified in ENCODE cell lines (ranging from 84,201 to 266,618 total sites per cell line)(Thurman et al., 2012). This may be because there are fewer DHS in primary differentiated cells than in immortalized cell lines; and that we chose to concentrate our analysis on high confidence peaks, reducing the chances that our detection of variable peaks would include false positive peaks.

7.2 MOLECULAR BASIS OF DHS VARIATION

A simple model to explain the molecular basis of DHS variation could be that local sequence variants contribute to chromatin variation marked by DHS variation and consequently affect gene expression. Two previous studies have estimated individual variation in regulatory regions in human lymphoblastoid cell lines: 7% of DHS displayed inter-individual variation (McDaniell et al., 2010) and 25% of binding sites for RNA polymerase II (sites that will appear in surveys of DHS) (Kasowski et al., 2010). Both studies identified underlying sequence variation as an important contributor, but could not determine whether sequence variation was the major contributor. For example, the total fraction of significant binding differences coinciding with genetic variations was 26% for RNA polymerase II (Kasowski et al., 2010), leaving the possibility open that non-genetic causes are an important cause of variation at regulatory sites.

My results indicate sequence variation accounts for DHS variation in one third of cases (503 out of 1,397 variable DHSs (36%)). In other words, almost

65% of variable sites are not correlated with sequence variation. Even when we limit our observations to instances where a site is lost or gained (i.e. 196 discretely variable DHS peaks) at a third of peaks we are still unable to attribute this alteration to a variant lying under the peak or within 100bp. Therefore, the simple model attributing variable DHS to underlying sequence variation cannot fully explain all variable DHSs.

There are a number of possible explanations for observing variable DHS in the absence of local sequence variants. First, in some cases the variable peaks we identified might not truly be variable. This clearly is not true for the 196 sites with discrete changes, but may account for some of the non-discretely varying sites. Second, failure to find a sequence variant can be attributed to missing variants in the mouse strain sequence; some variant classes (in particular insertion deletion polymorphisms between 10 ad 100bp) remain poorly detected. Third, the relevant sequence variation might lie within distal regulatory elements, which interact with the variant DHS. Such interactions, in which widely separated regions of the genome come into contact with each other via regulatory elements (Vernimmen et al., 2007, Wang et al., 2005, Tolhuis et al., 2002), would explain the occurrence of clustered sites (e.g. Figure 5-14). Our study, like others, is not sufficiently powered to detect this source of genetic variation (Degner et al., 2012). Fourth, there may be non-DNA based causes for the DHS variation. One possibility is the epigenetic effects reported for variable expression of the agouti locus (Morgan et al., 1999, Youngson and Whitelaw, 2008). Heritable epigenetic changes (such as DNA methylation) may also alter chromatin structure and DHS variation in the absence of sequence variation (Stadler et al., 2011, Thurman et al., 2012). Another is that environmental factors

might be responsible. For example variation in patterns of maternal care (specifically the way in which a mother rat cares for her young) have been shown to be inherited independently of DNA variation (Weaver et al., 2004). However it should be noted that the use of inbred mice minimized the environmental effect on phenotypic variation.

We identified a potential molecular mechanism for DHS variation (the disruption of a TFBS or a structural variant coinciding with a DHS) at 165 sites (12% of the total). While this analysis cannot unequivocally identify which variants are causal, it is consistent with recently published ENCODE results indicating the importance of sequence variation at a proportion of DHS (Bernstein et al., 2012, Thurman et al., 2012). However in one case we observed the presence of a DHS in a single strain, associated with a novel transcript, without an underlying sequence change, indicating that sequence variation at the DHS alone is an insufficient explanation both for DHS change and its functional consequences. Chromatin state also could be determined by interacting non-genetic factors (including epigenetic signature) or sequence variants within trans-acting elements (TFs) (Sladek and Hudson, 2006), which did not have enough power to assess in our experiments.

7.3 DHS VARIATION AND GENE EXPRESSION

We studied the relationship between DHS and gene expression variation. Local gene expression variation is an expected consequence at some variable DHS. A recent genome-wide survey of DHS in lymphoblastoid cell lines confirmed the importance of sequence variation as a source of DHS variation, and reported that 16% of DHS associated with local sequence variants were also associated with variation in transcript abundance at neighboring genes (Degner et al., 2012).

Potentially therefore these 16% might contribute to phenotypic variation. We asked if variable expressed genes are enriched at variable DHS sites. However the functional consequences of DHS variation on local gene expression appear relatively limited. We found that less than 10% of variable DHS are likely to have an effect on transcription (within 10 kb). However we cannot exclude the possibility that the effect of DHS site variation is broader than we have observed, having an effect only in different environments or tissues to those assayed, or that the effects are too subtle to have been detected. Also our results indicate that DHS variation at promoters is more constrained than at enhancers. Presumably variation at enhancers, and others types of DHS peaks, is less likely to have phenotypic consequences.

It is possible that the difficulties finding functional correlates of site variation might be due to the fact that site variation reflects some unknown but consistent measurement artifact. The bulk of DHS variation we observe is continuous, so that we use a probability cut-off to determine what is variable and what is not. To counter this problem, we analyzed separately the 196 sites with discrete, rather than continuous variation. These DHS show total loss or gain of a site in at least one strain. Our analyses of the discrete sites support results from the entire set of variable sites, namely only a small fraction (approximately 7%) are associated with changes in transcription (either poly(A)+ or poly(A)-). A third of 196 DHS lack sequence variants that can explain strain distribution pattern correlates with that of the variable DHS.

7.4 VARIABLE DHS AND QTL ANALYSIS

Many of the genetic signals identified through genome wide association studies (GWAS) for complex traits and disease susceptibility reside within DHS

(Bernstein et al., 2012, Maurano et al., 2012), but whether these signals preferentially exist at variable DHS is not known. Finally we used our DHS data in parallel with virtually all sequence variants could possibly contribute to DHS variation identified through almost complete sequence of all given eight inbred mouse strains (Keane et al., 2011, Yalcin et al., 2011) and 843 QTLs identified from 2,000 HS mice descended from these eight strains (Valdar et al., 2006) to address whether sequence variants underlying inbred mice variable DHSs contribute to complex phenotypes. Prof Jonathan Flint and Prof Richard Mott performed this analysis. Functional sequence variants are indeed enriched in DHS. We have shown that sequence variants at variable DHS are more likely to contribute to phenotypic variation than sequence variants within non-variable DHS, consistent with the finding that DHS are enriched with signals from human genome wide association studies (Maurano et al., 2012). Our data suggest that searching for causal variants of complex traits will profit by focusing within variable DHS, but such an approach will have to survey many tissues, presumably at different developmental time points and under a variety of environmental circumstances, to find a substantial fraction of the sequence variants involved in complex traits.

7.5 CONCLUSION

Regulatory sites of the genome affect gene expression and complex traits, including disease susceptibility. Variable regulatory sites are potentially interesting because they are a likely cause of phenotypic variation, providing a bridge between sequence and transcriptional variation. We identified chromatin accessible regions of the genome where DNA is not wrapped up in chromatin (hence potentially active regulatory sites) marked by DNase I hypersensitivity in eight inbred strains of mice in terminally differentiated erythroblasts. We compared sites that vary among strains (~ 5%) and compared them to non-variable sites. We showed that more than half of variable sites could not be attributed to local sequence variation. Functional consequences (in terms of readily detectable changes in gene expression) are associated with less than 10% of discretely variable DHS. We showed that variable sites are enriched for sequence variants contributing to complex traits in mice.

An original article entitled “causes and consequences of chromatin variation between inbred mice” has been published from this thesis:

HOSSEINI, M., GOODSTADT, L., HUGHES, J. R., KOWALCZYK, M. S., DE GOBBI, M., OTTO, G. W., COPLEY, R. R., MOTT, R., HIGGS, D. R. & FLINT, J. 2013. Causes and Consequences of Chromatin Variation between Inbred Mice. PLoS Genet, 9, e1003570.

SUPPLEMENTARY MATERIAL

Appendix 1- Solutions for DNase I assay

High Salt Nulear Lysis Buffer

50 mM KCl
10 mM $MgSO_4$
3 mM DTT
5 mM Hepes
0.05% NP40

Lysis Buffer

10 mM Tris
10 mM EDTA
10 mM NaCl
0.5% SDS
add Proteinase K prior to use

RSB

10 mM NaCl
10 mM Tris pH7.4
3 mM $MgCl_2.6H_2O$

Other requirements

PBS (0.9 M)
DNaseI RNase-free (Roche 04716728001)
Complete Protease Inhibitor Cocktail Tablets (Roche Diagnostics 11836145001):
1 tab in 2 ml water (25X)
DNase, RNase-free Sigma Water

Appendix 2 - Details of three sets of DHS positive and negative control qPCR primer pairs for terminally differentiated erythroblasts

	Primer ID	Sequence	Size (bp)	F/R	Chr	Coordinates
1	control1-F	CCTTGAAGCCATGGAAGAAC	20	+	11	32229687 32229706
2	control1-R	CCTCCTGCTCTGGAACTGAG	20	- +	11 8	32229687 32229706 86614976 86614995
3	control2-F	TCCCCTGTCATCCTCACAAC	20	+	11	32366439 32366458
4	control2-R	GAGCCTCAGGCACAGAGAAC	20	-	11	32366578 32366597
5	control3-F	AAAGCTGGACCATTTTGGTG	20	+	15	85646457 85646476
6	control3-R	ACGAGGGCTACATCATCAGG	20	-	15	85646584 85646603
7	DHSmcs-r2_F	TGGGATCTATGGAGATGCTTG	21	+	11	32151057 32151077
8	DHSmcs-r2_R	ATGGGAAAACAGCGTCAGAG	20	-	11	32151198 32151217
9	DHSmcs-r3_F	GCCTGGTGTTGCCTCTTATC	20	+	11	32156308 32156327
10	DHSmcs-r3_R	CTTCCCCAAACCACCCTATC	20	-	11	32156447 32156466
11	DHSminus14_F	CTCCTGGCTCCTCTGGAAG	19	+	11	32167441 32167460
12	DHSminus14_R	CCTACGCGAGTTCCTCTCAG	20	-	11	32167564 32167583

SNPs are presented by red letters. There is a single SNP in C2 forward primer in AKR/J (there is A base instead of C).

Appendix 3- Modified Touchdown PCR protocol (Korbie and Mattick, 2008)

PCR setup 1 for each primer set	
Template DNA 2ng/μl	25 μl
PCR buffer 10X	5 μl
dNTP mix 8 mM	5 μl
$MgCl_2$ 25 mM	3 μl
Forward primer 10 pM/μl	2.5 μl
Reverse primer 10 pM/μl	2.5 μl
HotStar Taq DNA Polymerase (Quigen)	0.3 μl
Milli Q water	Up to 50 μl

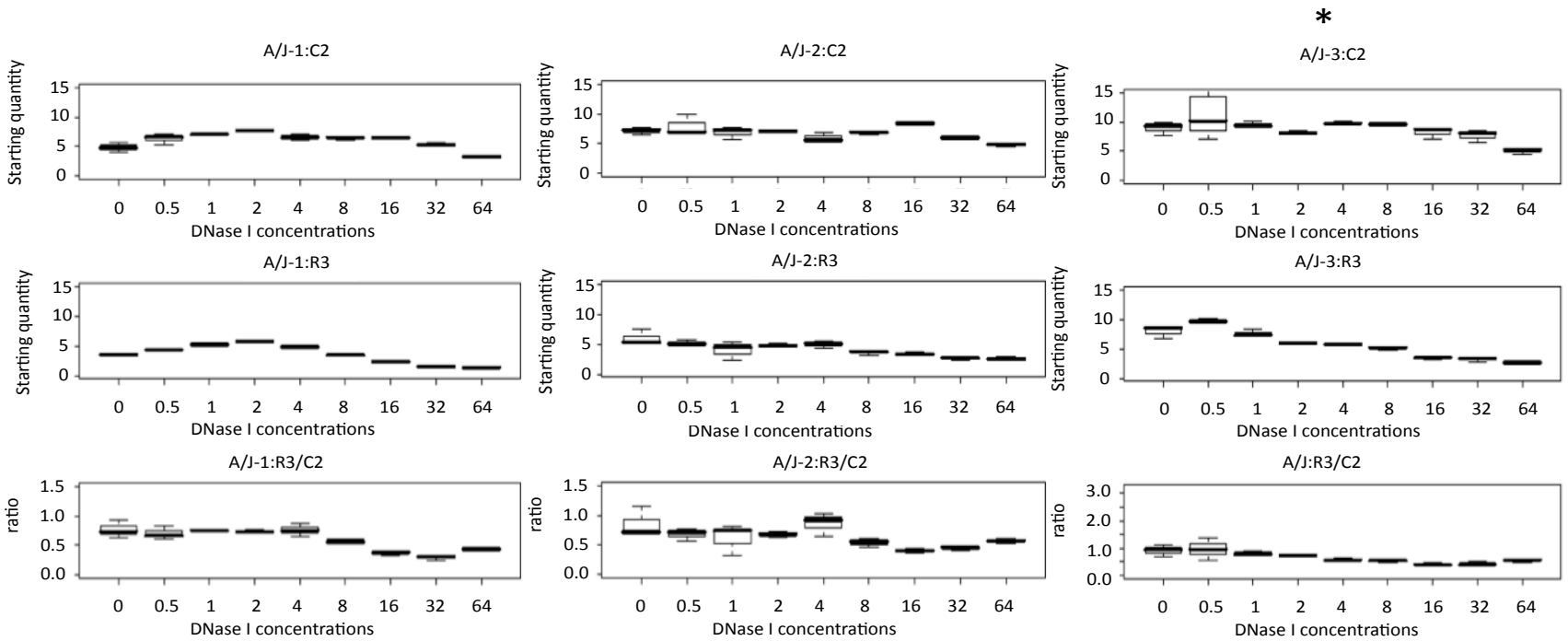
This setup was used for erythroid qPCR primer pairs' general assessment

PCR setup 2 for each primer set	
Template DNA 10 ng/μl	1 μl
PCR buffer 10X	1 μl
dNTP mix 8 mM	1 μl
$MgCl_2$ 25 mM	0.6 μl
Forward primer 10 pM/μl	0.5 μl
Reverse primer 10 pM/μl	0.5 μl
HotStar Taq DNA Polymerase (Quigen)	0.06 μl
Milli Q water	Up to 10 μl

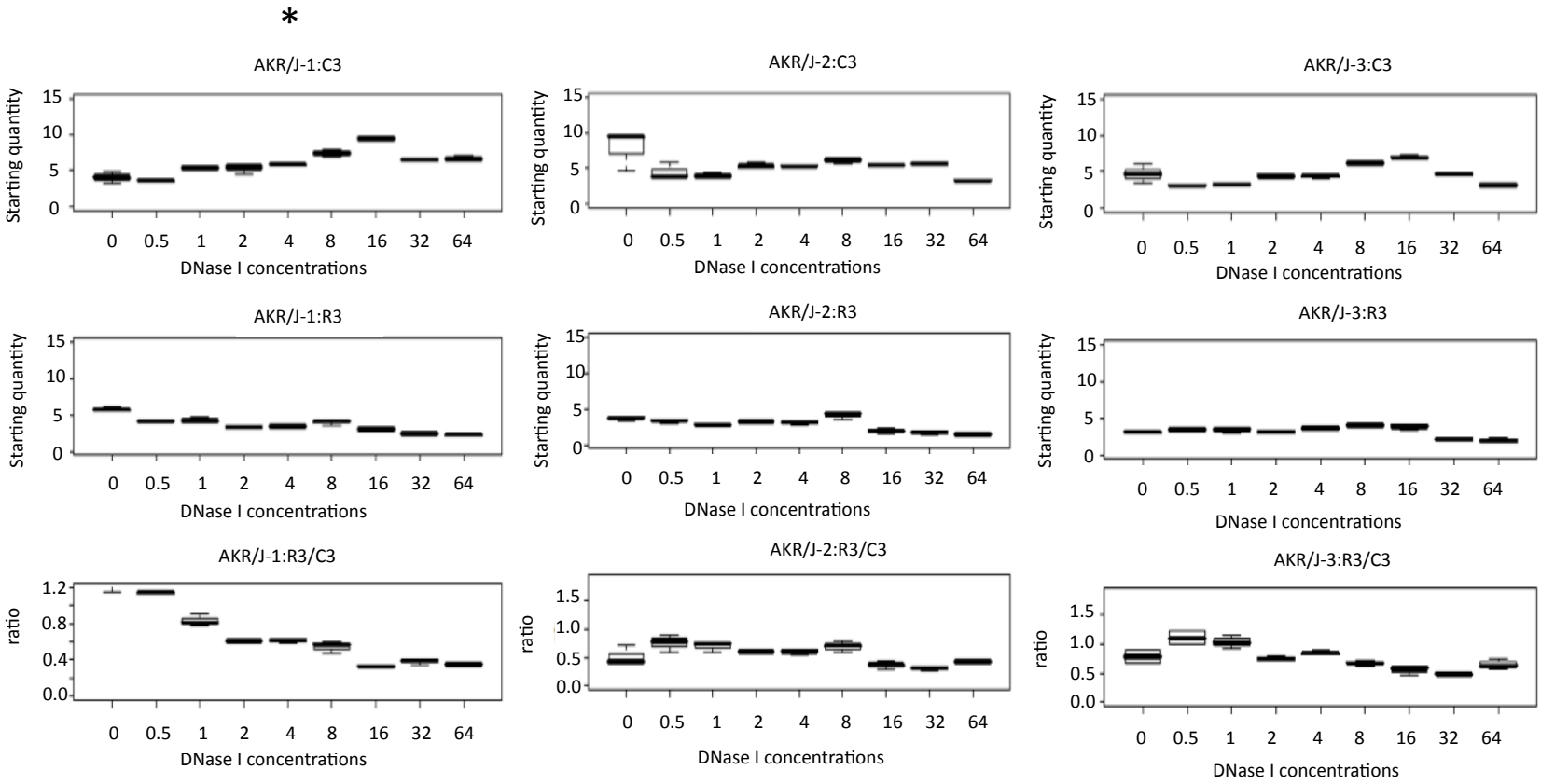
This setup was used for MEF qPCR primer pairs' general assessment

Phase1	Step	Temperature	Time
1	Denature	95° C	10 min
2	Denature	95° C	30 s
3	Anneal	65° C (-0.5° C per cycle)	30 s
4	Elongate	72° C	1 min *
Repeat step 2-4 (13 times)			
Phase2	Step	Temperature	Time
5	Denature	95° C	30 s
6	Anneal	55° C	30 s
7	Elongate	72° C	1 min *
Repeat step 5-7 (27 times)			
Termination	Step	Temperature	Time
8	Elongate	72° C	7 min
9	Hold	4° C	Forever
* For MEF primer pairs Touchdown PCR it was reduced to 30 s.			

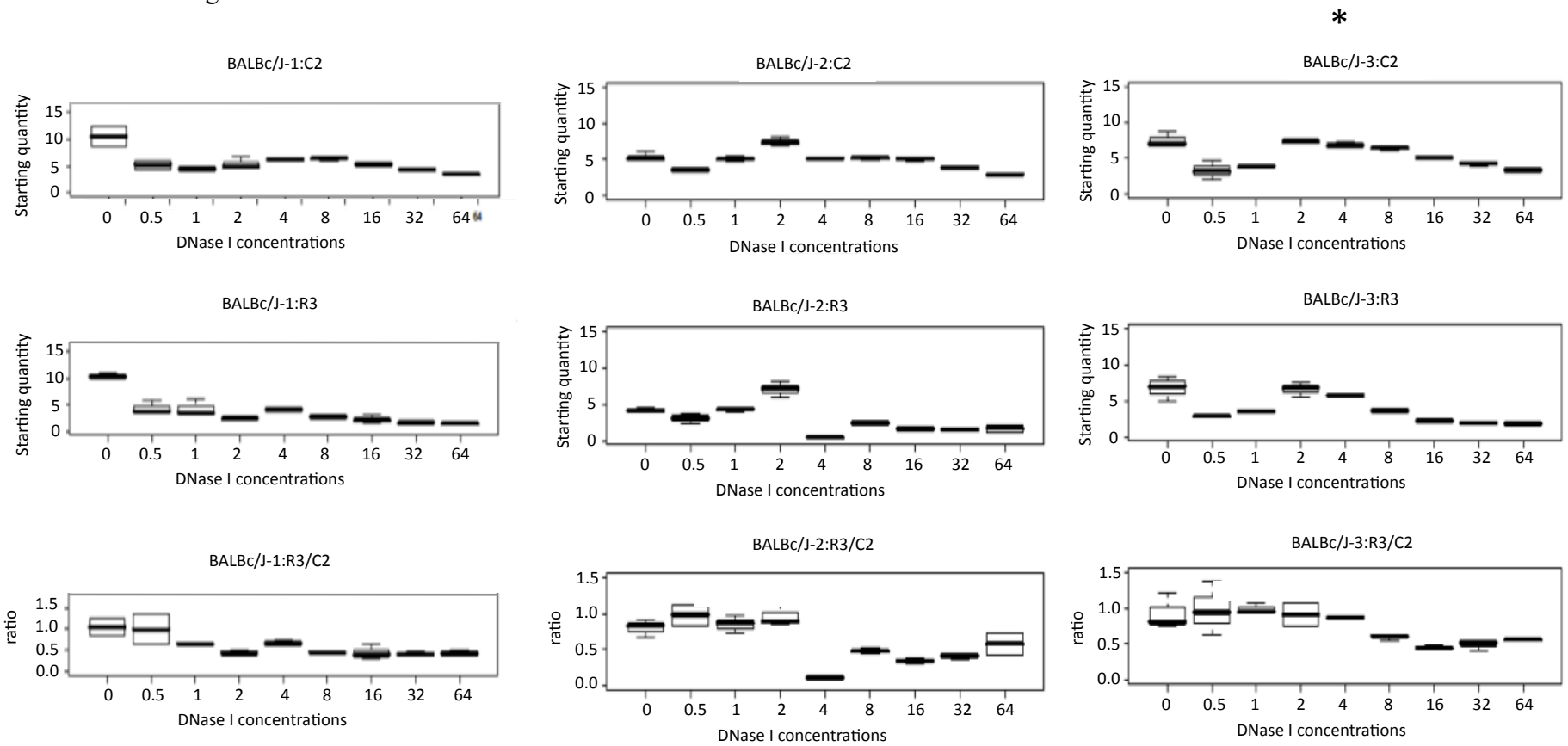
Appendix 4 – Quantification of DNase I digestion by qPCR in three replicates of A/J strain. The presentation of data is as described on Figure 3-18.



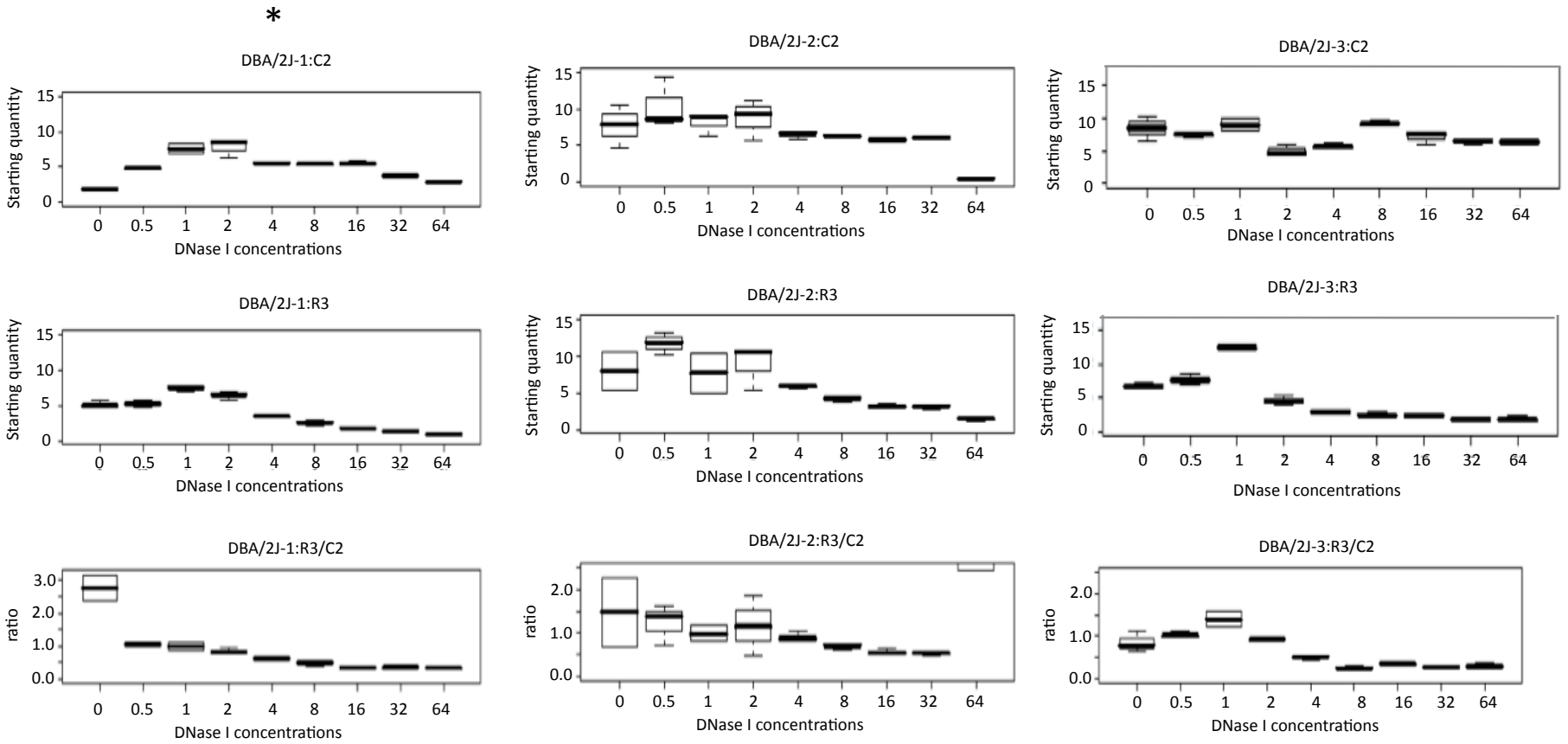
Appendix 5 – Quantification of DNase I digestion by qPCR in three replicates of AKR/J strain. The presentation of data is as described on Figure 3-18.



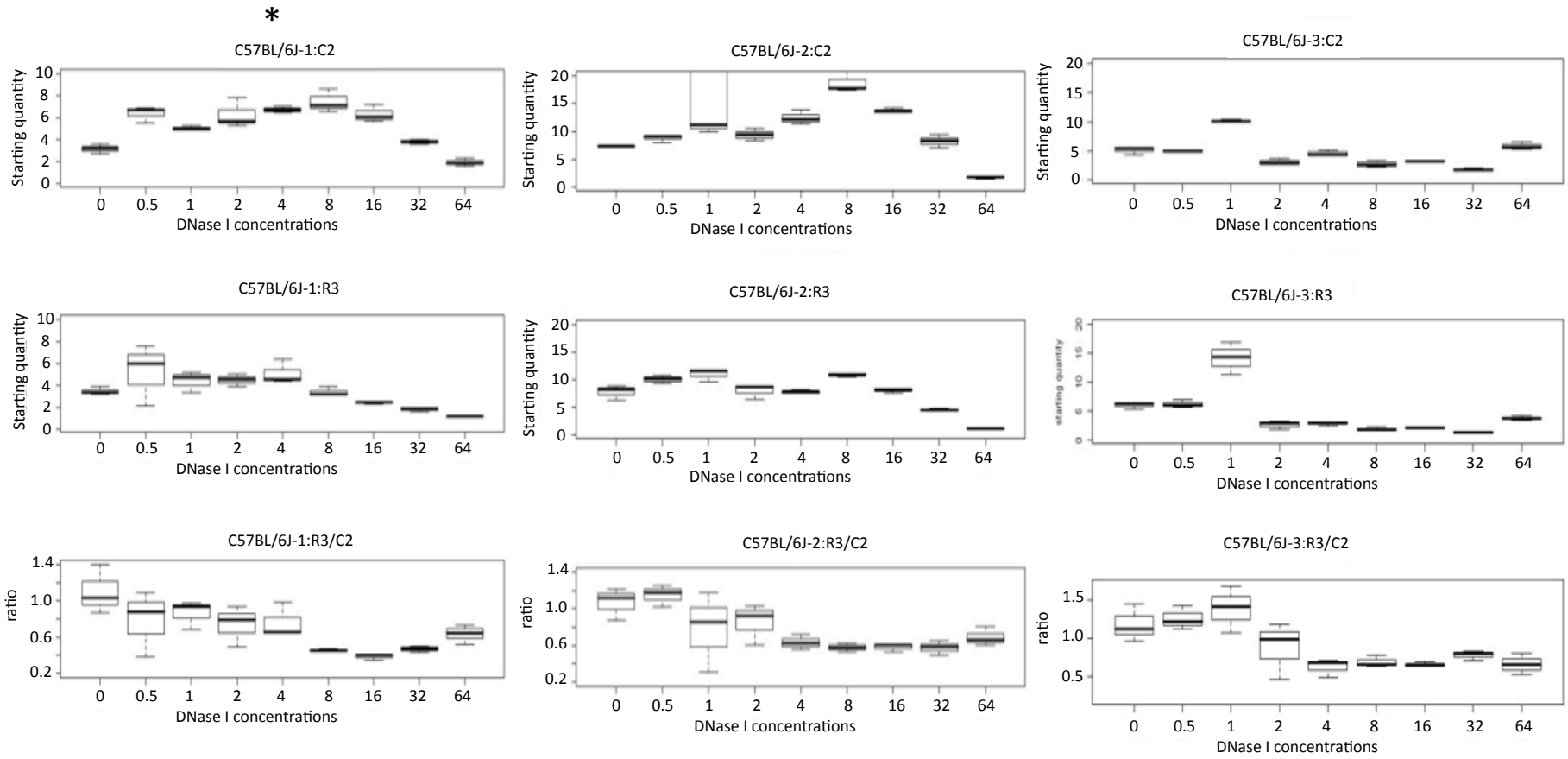
Appendix 6 – Quantification of DNase I digestion by qPCR in three replicates of BALBc/J strain. The presentation of data is as described on Figure 3-18.



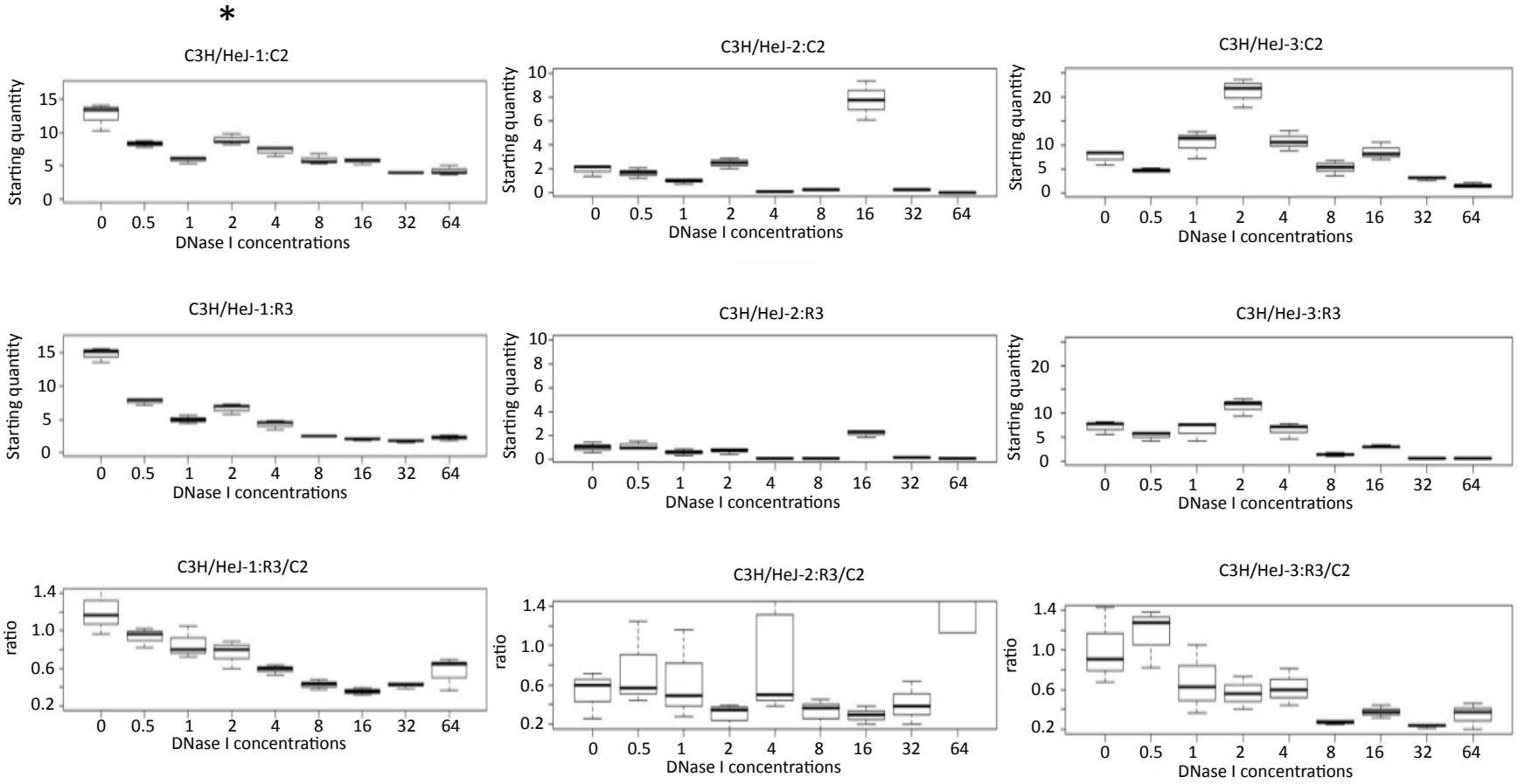
Appendix 7 – Quantification of DNase I digestion by qPCR in three replicates of DBA/2J strain. The presentation of data is as described on Figure 3-18.



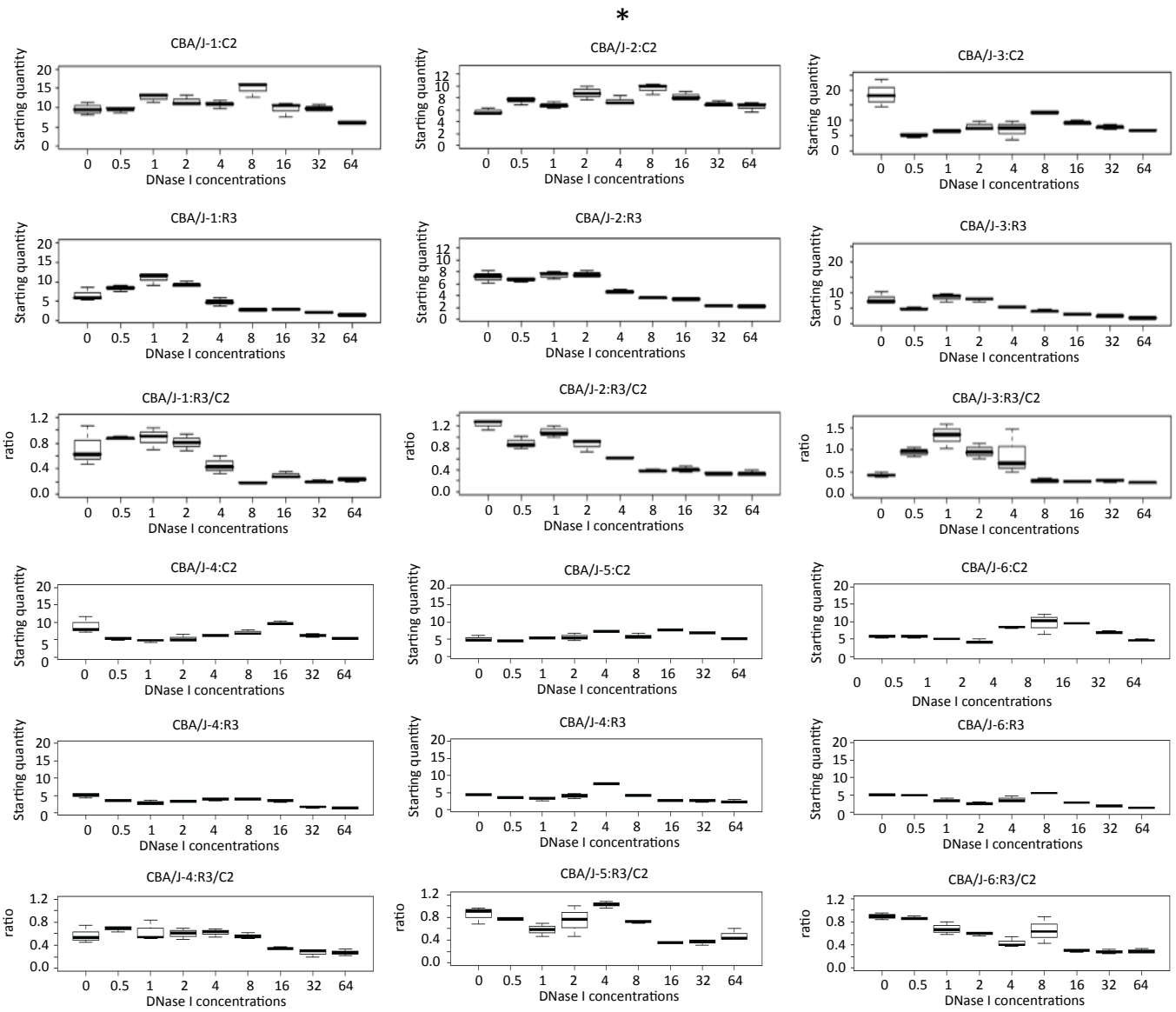
Appendix 8 – Quantification of DNase I digestion by qPCR in three replicates of C57BL/6J strain. The presentation of data is as described on Figure 3-18.



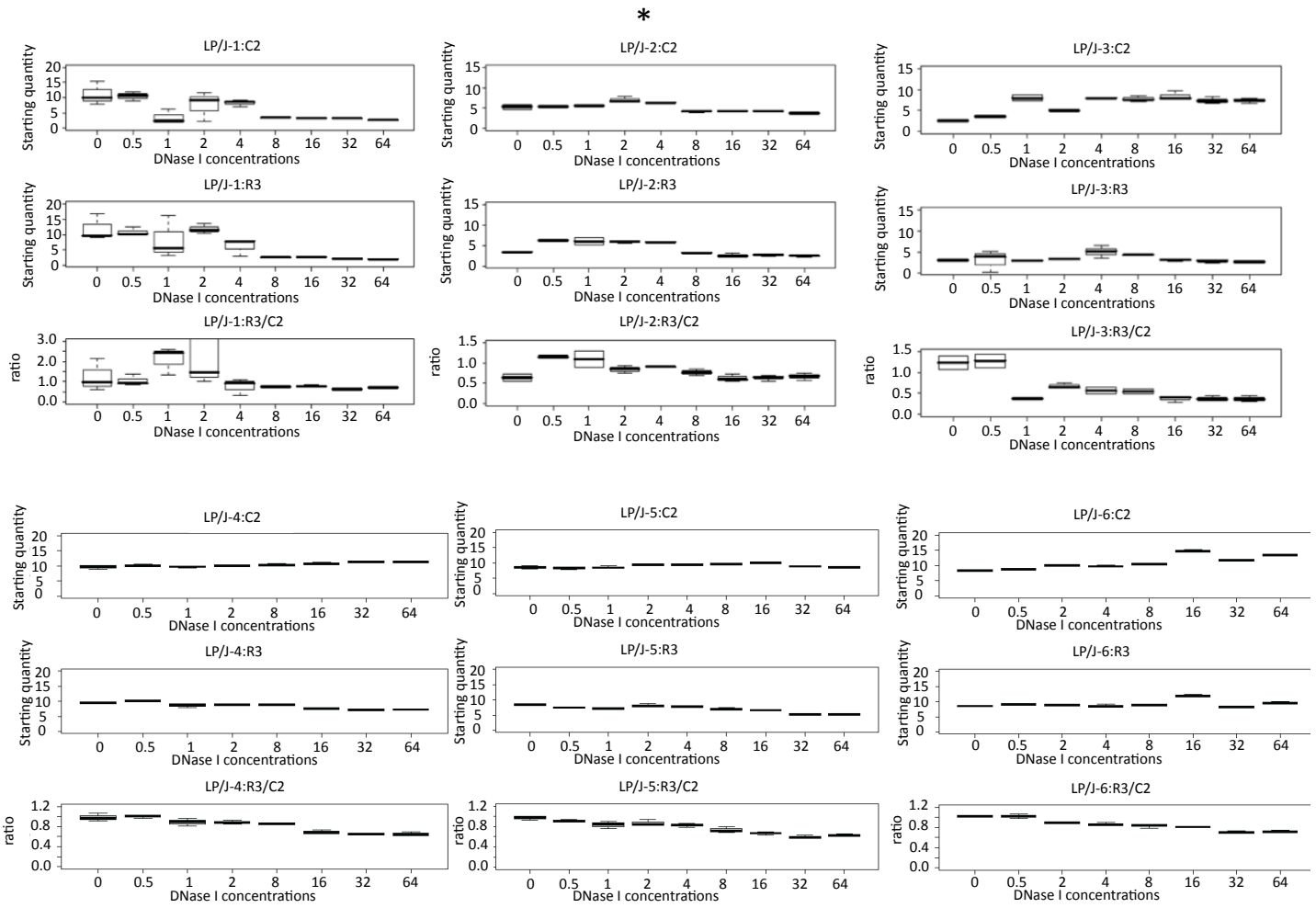
Appendix 9 – Quantification of DNase I digestion by qPCR in three replicates of C3H/HeJ strain. The presentation of data is as described on Figure 3-18.



Appendix 10 – Quantification of DNase I digestion by qPCR in six replicates of CBA/J strain. The presentation of data is as described on Figure 3-18.



Appendix 11 – Quantification of DNase I digestion by qPCR in six replicates of LP/J mice. The presentation of data is as described on Figure 3-18.



Appendix 12 - SYBR green qPCR DHS positive and negative primer pairs for C57BL/6J MEFs

Primer ID	Sequence	Chr	Strand	Coordinates	Span (bp)	Product Size (bp)	BLAT	LookSeq
C1-F	GCAGACCCACGACTCTCCTC	11	+	32229492-32229511	20	120	unique	No SNP
C1-R	CCCACACACAGGGAACAACCT	11	-	32229591-32229611	21	120	unique	No SNP
C2-F	ACCCAGGCTACAGTGGTTTCA	11	+	32366347-32366367	21	120	unique	No SNP
C2-R	GAATGGAAGTTGTGAGGATGACAGG	11	-	32366442-32366466	25	120	unique	No SNP
C3-F	ATGTTGGGTTTGGGTCATTGTGTAA	15	+	85647207-85647231	25	118	unique	No SNP
C3-R	CAGTTTATTCGCCATCGGTTCTCC	15	-	85647301-85647324	24	118	unique	No SNP
beta-actin-F	CCAAGCTCAGGGGACAAAGGAAGC	5	+	143667483-143667506	24	115	unique	No SNP
beta-actin-R	GTTTGGGGAGGGGGCGGAGG	5	-	143667578-143667597	20	115	unique	No SNP
Gapdh-F	TGCTTATCCAGTCCTAGCTCAAGGG	6	+	125115704-125115728	25	120	unique	No SNP
Gapdh-R	CAAAAATGAGGCGGGTCCAAAGAGA	6	-	125115799-125115823	25	120	unique	No SNP

REFERENCES

2012a. *The Human Gene Mutation Database (HGMD) at the Institute of Medical Genetics in Cardiff* [Online]. Available: http://www.hgmd.cf.ac.uk/ac/hahaha.php [Accessed 16 Sep 2012].

2012b. *Online Mendelian Inheritance in Man (OMIM) Database at Center for Medical Genetics Johns Hopkins University* [Online]. Available: http://omim.org/statistics/entry [Accessed 15 September 2012].

ALBERT, I., MAVRICH, T. N., TOMSHO, L. P., QI, J., ZANTON, S. J., SCHUSTER, S. C. & PUGH, B. F. 2007. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature,* 446**,** 572-576.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Genome Biol,* 11**,** R106.

ANDERSSON, L. C., NILSSON, K. & GAHMBERG, C. G. 1979. K562 - Human Erythroleukemic Cell-Line. *International Journal of Cancer,* 23**,** 143-147.

ANTONARAKIS, S. E., IRKIN, S. H., CHENG, T. C., SCOTT, A. F., SEXTON, J. P., TRUSKO, S. P., CHARACHE, S. & KAZAZIAN, H. H. 1984. Beta-Thalassemia in American Blacks - Novel Mutations in the Tata Box and an Acceptor Splice Site. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences,* 81**,** 1154-1158.

AXEL, R. & GAREL, A. 1977. Structure of Ovalbumin Gene in Chromatin. *Ann N Y Acad Sci,* 286**,** 135-146.

BAILEY, T. L., BODEN, M., BUSKE, F. A., FRITH, M., GRANT, C. E., CLEMENTI, L., REN, J., LI, W. W. & NOBLE, W. S. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res,* 37**,** W202-8.

BARSKI, A., CUDDAPAH, S., CUI, K. R., ROH, T. Y., SCHONES, D. E., WANG, Z. B., WEI, G., CHEPELEV, I. & ZHAO, K. J. 2007a. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-837.

BARSKI, A., CUDDAPAH, S., CUI, K. R., ROH, T. Y., SCHONES, D. E., WANG, Z. B., WEI, G., CHEPELEV, I. & ZHAO, K. J. 2007b. Response: Mapping nucleosome positions using ChIP-Seq data. *Cell,* 131**,** 832-833.

BELL, A. C., WEST, A. G. & FELSENFELD, G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell,* 98**,** 387-396.

BERNSTEIN, B. E., BIRNEY, E., DUNHAM, I., GREEN, E. D., GUNTER, C. & SNYDER, M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

BIRD, A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev,* 16**,** 6-21.

BIRNEY, E., STAMATOYANNOPOULOS, J. A., DUTTA, A., GUIGO, R., GINGERAS, T. R., MARGULIES, E. H., WENG, Z., SNYDER, M., DERMITZAKIS, E. T., THURMAN, R. E., KUEHN, M. S., TAYLOR, C. M., NEPH, S., KOCH, C. M., ASTHANA, S., MALHOTRA, A., ADZHUBEI, I., GREENBAUM, J. A., ANDREWS, R. M., FLICEK, P., BOYLE, P. J., CAO, H., CARTER, N. P., CLELLAND, G. K., DAVIS, S., DAY, N., DHAMI, P., DILLON, S. C., DORSCHNER, M. O., FIEGLER, H., GIRESI, P. G., GOLDY, J., HAWRYLYCZ, M., HAYDOCK, A., HUMBERT, R., JAMES, K. D., JOHNSON, B. E., JOHNSON, E. M., FRUM, T. T., ROSENZWEIG, E. R., KARNANI, N., LEE, K., LEFEBVRE, G. C.,

NAVAS, P. A., NERI, F., PARKER, S. C. J., SABO, P. J., SANDSTROM, R., SHAFER, A., VETRIE, D., WEAVER, M., WILCOX, S., YU, M., COLLINS, F. S., DEKKER, J., LIEB, J. D., TULLIUS, T. D., CRAWFORD, G. E., SUNYAEV, S., NOBLE, W. S., DUNHAM, I., DENOEUD, F., REYMOND, A., KAPRANOV, P., ROZOWSKY, J., ZHENG, D., CASTELO, R., FRANKISH, A., HARROW, J., GHOSH, S., SANDELIN, A., HOFACKER, I. L., BAERTSCH, R., KEEFE, D., DIKE, S., CHENG, J., HIRSCH, H. A., SEKINGER, E. A., LAGARDE, J., ABRIL, J. F., SHAHAB, A., FLAMM, C., FRIED, C., HACKERMUELLER, J., HERTEL, J., LINDEMEYER, M., MISSAL, K., TANZER, A., WASHIETL, S., KORBEL, J., EMANUELSSON, O., PEDERSEN, J. S., HOLROYD, N., TAYLOR, R., SWARBRECK, D., MATTHEWS, N., DICKSON, M. C., THOMAS, D. J., WEIRAUCH, M. T., GILBERT, J., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* 447**,** 799-816.

BLANCHETTE, M., KENT, W. J., RIEMER, C., ELNITSKI, L., SMIT, A. F., ROSKIN, K. M., BAERTSCH, R., ROSENBLOOM, K., CLAWSON, H., GREEN, E. D., HAUSSLER, D. & MILLER, W. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res,* 14**,** 708-15.

BOYLE, A. P., DAVIS, S., SHULHA, H. P., MELTZER, P., MARGULIES, E. H., WENG, Z., FUREY, T. S. & CRAWFORD, G. E. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell,* 132**,** 311-322.

BOYLE, A. P. & FUREY, T. S. 2009. High-resolution mapping studies of chromatin and gene regulatory elements. *Epigenomics,* 1**,** 319-329.

BOYLE, A. P., SONG, L., LEE, B.-K., LONDON, D., KEEFE, D., BIRNEY, E., IYER, V. R., CRAWFORD, G. E. & FUREY, T. S. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research,* 21**,** 456-464.

BUCHER, P. 1990. Weight Matrix Descriptions of 4 Eukaryotic Rna Polymerase-Ii Promoter Elements Derived from 502 Unrelated Promoter Sequences. *Journal of Molecular Biology,* 212**,** 563-578.

BULGER, M. & GROUDINE, M. 2002. TRAPping enhancer function. *Nature Genetics,* 32**,** 555-556.

CHEN, X. Y., HOFFMAN, M. M., BILMES, J. A., HESSELBERTH, J. R. & NOBLE, W. S. 2010. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics,* 26**,** i334-i342.

CHORLEY, B. N., WANG, X., CAMPBELL, M. R., PITTMAN, G. S., NOUREDDINE, M. A. & BELL, D. A. 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: Current and developing technologies. *Mutation Research-Reviews in Mutation Research,* 659**,** 147-157.

COCKERILL, P. N. 2011. Structure and function of active chromatin and DNase I hypersensitive sites. *Febs Journal,* 278**,** 2182-2210.

COLLINS, F. S., LANDER, E. S., ROGERS, J. & WATERSTON, R. H. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-945.

CRAWFORD, G. E., DAVIS, S., SCACHERI, P. C., RENAUD, G., HALAWI, M. J., ERDOS, M. R., GREEN, R., MELTZER, P. S., WOLFSBERG, T. G. & COLLINS, F. S. 2006a. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods,* 3**,** 503-9.

CRAWFORD, G. E., HOLT, I. E., MULLIKIN, J. C., TAI, D., GREEN, E. D., WOLFSBERG, T. G. & COLLINS, F. S. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 992-997.

CRAWFORD, G. E., HOLT, I. E., WHITTLE, J., WEBB, B. D., TAI, D., DAVIS, S., MARGULIES, E. H., CHEN, Y. D., BERNAT, J. A., GINSBURG, D., ZHOU, D. X., LUO, S. J., VASICEK, T. J., DALY, M. J., WOLFSBERG, T. G. & COLLINS, F. S. 2006b. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research,* 16**,** 123-131.

DE GOBBI, M., VIPRAKISIT, V., HUGHES, J. R., FISHER, C., BUCKLE, V. J., AYYUB, H., GIBBONS, R. J., VERNIMMEN, D., YOSHINAGA, Y., DE JONG, P., CHENG, J. F., RUBIN, E. M., WOOD, W. G., BOWDEN, D. & HIGGS, D. R. 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science,* 312**,** 1215-1217.

DEGNER, J. F., PAI, A. A., PIQUE-REGI, R., VEYRIERAS, J. B., GAFFNEY, D. J., PICKRELL, J. K., DE LEON, S., MICHELINI, K., LEWELLEN, N., CRAWFORD, G. E., STEPHENS, M., GILAD, Y. & PRITCHARD, J. K. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature,* 482**,** 390-4.

DEKOK, Y. J. M., VOSSENAAR, E. R., CREMERS, C., DAHL, N., LAPORTE, J., HU, L. J., LACOMBE, D., FISCHELGHODSIAN, N., FRIEDMAN, R. A., PARNES, L. S., THORPE, P., BITNERGLINDZICZ, M., PANDER, H. J., HEILBRONNER, H., GRAVELINE, J., DENDUNNEN, J. T., BRUNNER, H. G., ROPERS, H. H. & CREMERS, F. P. M. 1996. Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4. *Human Molecular Genetics,* 5**,** 1229-1235.

DERMITZAKIS, E. T. & CLARK, A. G. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Mol Biol Evol,* 19**,** 1114-1121.

DERRIEN, T., ESTELLE, J., MARCO SOLA, S., KNOWLES, D. G., RAINERI, E., GUIGO, R. & RIBECA, P. 2012. Fast computation and applications of genome mappability. *PLoS One,* 7**,** e30377.

ELGIN, S. C. R. 1984. Anatomy of Hypersensitive Sites. *Nature,* 309**,** 213-214.

EMISON, E. S., MCCALLION, A. S., KASHUK, C. S., BUSH, R. T., GRICE, E., LIN, S., PORTNOY, M. E., CUTLER, D. J., GREEN, E. D. & CHAKRAVARTI, A. 2005. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature,* 434**,** 857-863.

EPSTEIN, D. J. 2009. Cis-regulatory mutations in human disease. *Briefings in Functional Genomics & Proteomics,* 8**,** 310-316.

EWING, B., HILLIER, L., WENDL, M. C. & GREEN, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research,* 8**,** 175-185.

FEINGOLD, E. A., GOOD, P. J., GUYER, M. S., KAMHOLZ, S., LIEFER, L., WETTERSTRAND, K., COLLINS, F. S., GINGERAS, T. R., KAMPA, D., SEKINGER, E. A., CHENG, J., HIRSCH, H., GHOSH, S., ZHU, Z., PATE, S., PICCOLBONI, A., YANG, A., TAMMANA, H., BEKIRANOV, S., KAPRANOV, P., HARRISON, R., CHURCH, G., STRUHL, K., REN, B., KIM, T. H., BARRERA, L.

O., QU, C., VAN CALCAR, S., LUNA, R., GLASS, C. K., ROSENFELD, M. G., GUIGO, R., ANTONARAKIS, S. E., BIRNEY, E., BRENT, M., PACHTER, L., REYMOND, A., DERMITZAKIS, E. T., DEWEY, C., KEEFE, D., DENOEUD, F., LAGARDE, J., ASHURST, J., HUBBARD, T., WESSELINK, J. J., CASTELO, R., EYRAS, E., MYERS, R. M., SIDOW, A., BATZOGLOU, S., TRINKLEIN, N. D., HARTMAN, S. J., ALDRED, S. F., ANTON, E., SCHROEDER, D. I., MARTICKE, S. S., NGUYEN, L., SCHMUTZ, J., GRIMWOOD, J., DICKSON, M., COOPER, G. M., STONE, E. A., ASIMENOS, G., BRUDNO, M., DUTTA, A., KARNANI, N., TAYLOR, C. M., KIM, H. K., ROBINS, G., STAMATOYANNOPOULOS, G., STAMATOYANNOPOULOS, J. A., DORSCHNER, M., SABO, P., HAWRYTYCZ, M., HUMBERT, R., WALLACE, J., YU, M., NAVAS, P. A., MCARTHUR, M., NOBLE, W. S., DUNHAM, I., KOCH, C. M., ANDREWS, R. M., CLELLAND, G. K., WILCOX, S., FOWLER, J. C., JAMES, K. D., GROTH, P., DOVEY, O. M., ELLIS, P. D., WRAIGHT, V. L., MUNGALL, A. J., DHAMI, P., FIEGLER, H., LANGFORD, C. F., CARTER, N. P., VETRIE, D., SNYDER, M., EUSKIRCHEN, G., URBAN, A. E., et al. 2004. The ENCODE (ENCyclopedia of DNA elements) Project. *Science,* 306**,** 636-640.

FELSENFELD, G. 1992. Chromatin as an Essential Part of the Transcriptional Mechanism. *Nature,* 355**,** 219-224.

FELSENFELD, G., BOYES, J., CHUNG, J., CLARK, D. & STUDITSKY, V. 1996. Chromatin structure and gene expression. *Proceedings of the National Academy of Sciences of the United States of America,* 93**,** 9384-9388.

FELSENFELD, G. & GROUDINE, M. 2003. Controlling the double helix. *Nature,* 421**,** 448-453.

FENG, W. C., SOUTHWOOD, C. M. & BIEKER, J. J. 1994. Analyses of Beta-Thalassemia Mutant-DNA Interactions with Erythroid Kruppel-Like Factor (Eklf), an Erythroid Cell-Specific Transcription Factor. *Journal of Biological Chemistry,* 269**,** 1493-1500.

FLINT, J., THOMAS, K., MICKLEM, G., RAYNHAM, H., CLARK, K., DOGGETT, N. A., KING, A. & HIGGS, D. R. 1997. The relationship between chromosome structure and function at a human telomeric region. *Nature Genetics,* 15**,** 252-257.

FRANSZ, P. & DE JONG, H. 2011. From nucleosome to chromosome: a dynamic organization of genetic information. *Plant Journal,* 66**,** 4-17.

FU, Y. T., SINHA, M., PETERSON, C. L. & WENG, Z. P. 2008. The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome. *PLoS genetics,* 4**,** e1000138.

GARDINERGARDEN, M. & FROMMER, M. 1987. Cpg Islands in Vertebrate Genomes. *Journal of Molecular Biology,* 196**,** 261-282.

GAULTON, K. J., NAMMO, T., PASQUALI, L., SIMON, J. M., GIRESI, P. G., FOGARTY, M. P., PANHUIS, T. M., MIECZKOWSKI, P., SECCHI, A., BOSCO, D., BERNEY, T., MONTANYA, E., MOHLKE, K. L., LIEB, J. D. & FERRER, J. 2010. A map of open chromatin in human pancreatic islets. *Nature Genetics,* 42**,** 255-9.

GERSHENZON, N. I. & IOSHIKHES, I. P. 2005. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics,* 21**,** 1295-1300.

GIRESI, P. G., KIM, J., MCDANIELL, R. M., IYER, V. R. & LIEB, J. D. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res,* 17**,** 877-85.

GRICE, E. A., ROCHELLE, E. S., GREEN, E. D., CHAKRAVARTI, A. & MCCALLION, A. S. 2005. Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. *Human Molecular Genetics,* 14**,** 3837-3845.

GROSS, D. S. & GARRARD, W. T. 1988. Nuclease Hypersensitive Sites in Chromatin. *Annu Rev Biochem,* 57**,** 159-197.

HARDY, J. & SINGLETON, A. 2009. CURRENT CONCEPTS Genomewide Association Studies and Human Disease. *New England Journal of Medicine,* 360**,** 1759-1768.

HARK, A. T., SCHOENHERR, C. J., KATZ, D. J., INGRAM, R. S., LEVORSE, J. M. & TILGHMAN, S. M. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature,* 405**,** 486-489.

HEGMANN, J. P. & POSSIDENTE, B. 1981. Estimating Genetic Correlations from Inbred Strains. *Behavior Genetics,* 11**,** 103-114.

HEINTZMAN, N. D., STUART, R. K., HON, G., FU, Y. T., CHING, C. W., HAWKINS, R. D., BARRERA, L. O., VAN CALCAR, S., QU, C. X., CHING, K. A., WANG, W., WENG, Z. P., GREEN, R. D., CRAWFORD, G. E. & REN, B. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics,* 39**,** 311-318.

HESSELBERTH, J. R., CHEN, X., ZHANG, Z., SABO, P. J., SANDSTROM, R., REYNOLDS, A. P., THURMAN, R. E., NEPH, S., KUEHN, M. S., NOBLE, W. S., FIELDS, S. & STAMATOYANNOPOULOS, J. A. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods,* 6**,** 283-289.

HEWISH, D. R. & BURGOYNE, L. A. 1973. Chromatin Sub-Structure - Digestion of Chromatin DNA at Regularly Spaced Sites by a Nuclear Deoxyribonuclease. *Biochem Biophys Res Commun,* 52**,** 504-510.

HINDORFF, L. A., MACARTHUR, J., WISE, A., JUNKINS, H. A., HALL, P. N., KLEMM, A. K. & MANOLIO, T. A. *A Catalog of Published Genome-Wide Association Studies* [Online]. Available: www.genome.gov/gwastudies [Accessed 17 Sep 2012].

HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S. & MANOLIO, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 9362-9367.

HOGAN, G. J., LEE, C. K. & LIEB, J. D. 2006. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS genetics,* 2**,** 1433-1450.

HOSSEINI, M., GOODSTADT, L., HUGHES, J. R., KOWALCZYK, M. S., DE GOBBI, M., OTTO, G. W., COPLEY, R. R., MOTT, R., HIGGS, D. R. & FLINT, J. 2013. Causes and Consequences of Chromatin Variation between Inbred Mice. *PLoS Genet,* 9**,** e1003570.

HUGHES, J. R., CHENG, J. F., VENTRESS, N., PRABHAKAR, S., CLARK, K., ANGUITA, E., DE GOBBI, M., DE JONG, P., RUBIN, E. & HIGGS, D. R. 2005. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 9830-9835.

IOSHIKHES, I. P. & ZHANG, M. Q. 2000. Large-scale human promoter mapping using CpG islands. *Nature Genetics,* 26**,** 61-63.

JANSEN, A., VAN DER ZANDE, E., MEERT, W., FINK, G. R. & VERSTREPEN, K. J. 2012. Distal chromatin structure influences local nucleosome positions and gene expression. *Nucleic Acids Research,* 40**,** 3870-3885.

JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. & WOLD, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science,* 316**,** 1497-1502.

JOHNSON, S. M., TAN, F. J., MCCULLOUGH, H. L., RIORDAN, D. P. & FIRE, A. Z. 2006. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res,* 16**,** 1505-16.

JONES, P. L., VEENSTRA, G. J. C., WADE, P. A., VERMAAK, D., KASS, S. U., LANDSBERGER, N., STROUBOULIS, J. & WOLFFE, A. P. 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics,* 19**,** 187-191.

JONES, S. J. M. 2006. Prediction of genomic functional elements. *Annual Review of Genomics and Human Genetics,* 7**,** 315-338.

KAGEY, M. H., NEWMAN, J. J., BILODEAU, S., ZHAN, Y., ORLANDO, D. A., VAN BERKUM, N. L., EBMEIER, C. C., GOOSSENS, J., RAHL, P. B., LEVINE, S. S., TAATJES, D. J., DEKKER, J. & YOUNG, R. A. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature,* 467**,** 430-435.

KARLIN, S. & ALTSCHUL, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring methods. *Proceedings of the National Academy of Sciences, USA,* 87**,** 2264-2268.

KASOWSKI, M., GRUBERT, F., HEFFELFINGER, C., HARIHARAN, M., ASABERE, A., WASZAK, S. M., HABEGGER, L., ROZOWSKY, J., SHI, M., URBAN, A. E., HONG, M. Y., KARCZEWSKI, K. J., HUBER, W., WEISSMAN, S. M., GERSTEIN, M. B., KORBEL, J. O. & SNYDER, M. 2010. Variation in transcription factor binding among humans. *Science,* 328**,** 232-5.

KEANE, T. M., GOODSTADT, L., DANECEK, P., WHITE, M. A., WONG, K., YALCIN, B., HEGER, A., AGAM, A., SLATER, G., GOODSON, M., FURLOTTE, N. A., ESKIN, E., NELLAKER, C., WHITLEY, H., CLEAK, J., JANOWITZ, D., HERNANDEZ-PLIEGO, P., EDWARDS, A., BELGARD, T. G., OLIVER, P. L., MCINTYRE, R. E., BHOMRA, A., NICOD, J., GAN, X., YUAN, W., VAN DER WEYDEN, L., STEWARD, C. A., BALA, S., STALKER, J., MOTT, R., DURBIN, R., JACKSON, I. J., CZECHANSKI, A., GUERRA-ASSUNCAO, J. A., DONAHUE, L. R., REINHOLDT, L. G., PAYSEUR, B. A., PONTING, C. P., BIRNEY, E., FLINT, J. & ADAMS, D. J. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature,* 477**,** 289-294.

KEENE, M. A., CORCES, V., LOWENHAUPT, K. & ELGIN, S. C. R. 1981. Dnase-I Hypersensitive Sites in Drosophila Chromatin Occur at the 5' Ends of Regions of Transcription. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences,* 78**,** 143-146.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Research,* 12**,** 996-1006.

KING, D. C., TAYLOR, J., ELNITSKI, L., CHIAROMONTE, F., MILLER, W. & HARDISON, R. C. 2005. Evaluation of regulatory potential and

conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Research,* 15**,** 1051-1060.

KING, D. C., TAYLOR, J., ZHANG, Y., CHENG, Y., LAWSON, H. A., MARTIN, J., CHIAROMONTE, F., MILLER, W. & HARDISON, R. C. 2007. Finding cis-regulatory elements using comparative genomics: Some lessons from ENCODE data. *Genome Research,* 17**,** 775-786.

KNIGHT, J. C. 2005. Regulatory polymorphisms underlying complex disease traits. *Journal of Molecular Medicine-Jmm,* 83**,** 97-109.

KORBIE, D. J. & MATTICK, J. S. 2008. Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nat Protoc,* 3**,** 1452-1456.

KOWALCZYK, M. S., HUGHES, J. R., GARRICK, D., LYNCH, M. D., SHARPE, J. A., SLOANE-STANLEY, J. A., MCGOWAN, S. J., DE GOBBI, M., HOSSEINI, M., VERNIMMEN, D., BROWN, J. M., GRAY, N. E., COLLAVIN, L., GIBBONS, R. J., FLINT, J., TAYLOR, S., BUCKLE, V. J., MILNE, T. A., WOOD, W. G. & HIGGS, D. R. 2012. Intragenic Enhancers Act as Alternative Promoters. *Molecular cell,* 45**,** 447-458.

KULKARNI, M. M. & ARNOSTI, D. N. 2005. cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster. *Molecular and Cellular Biology,* 25**,** 3411-3420.

LEE, W., TILLO, D., BRAY, N., MORSE, R. H., DAVIS, R. W., HUGHES, T. R. & NISLOW, C. 2007. A high- resolution atlas of nucleosome occupancy in yeast. *Nature Genetics,* 39**,** 1235-1244.

LEMON, B. & TJIAN, R. 2000. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev,* 14**,** 2551-2569.

LENHARD, B., SANDELIN, A., MENDOZA, L., ENGSTROM, P., JAREBORG, N. & WASSERMAN, W. W. 2003. Identification of conserved regulatory elements by comparative genome analysis. *Journal of biology,* 2**,** 13.

LETTICE, L. A., HEANEY, S. J. H., PURDIE, L. A., LI, L., DE BEER, P., OOSTRA, B. A., GOODE, D., ELGAR, G., HILL, R. E. & DE GRAAFF, E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics,* 12**,** 1725-1735.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, Q., PETERSON, K. R., FANG, X. & STAMATOYANNOPOULOS, G. 2002. Locus control regions. *Blood,* 100**,** 3077-86.

LODISH, H., BERK, A., MATSUDAIRA, P., KAISER, C. A., KRIEGER, M., SCOTT, M. P., L., Z. S. & DARNELL, J. 2004. Molecular structures of genes and chromosomes. *Molecular Cell Biology.* Fifth ed. New York: W. H. Freeman and Company.

LOHR, D. & LOPEZ, J. 1995. Gal4/Gal80-Dependent Nucleosome Disruption Deposition on the Upstream Regions of the Yeast Gal1-10 and Gal80 Genes. *Journal of Biological Chemistry,* 270**,** 27671-27678.

LU, Q. & RICHARDSON, B. 2004. DNaseI hypersensitivity analysis of chromatin structure. *Methods in molecular biology (Clifton, N.J.),* 287**,** 77-86.

LUDWIG, M. Z., BERGMAN, C., PATEL, N. H. & KREITMAN, M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature,* 403**,** 564-567.

LUNTER, G. & GOODSON, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res,* 21**,** 936-9.

MANOLIO, T. A. 2010. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine,* 363**,** 166-176.

MANOLIO, T. A., BROOKS, L. D. & COLLINS, F. S. 2008. A HapMap harvest of insights into the genetics of common disease. *Journal of Clinical Investigation,* 118**,** 1590-1605.

MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics.* Palo Alto: Annual Reviews.

MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H. Z., BRODY, J., SHAFER, A., NERI, F., LEE, K., KUTYAVIN, T., STEHLING-SUN, S., JOHNSON, A. K., CANFIELD, T. K., GISTE, E., DIEGEL, M., BATES, D., HANSEN, R. S., NEPH, S., SABO, P. J., HEIMFELD, S., RAUBITSCHEK, A., ZIEGLER, S., COTSAPAS, C., SOTOODEHNIA, N., GLASS, I., SUNYAEV, S. R., KAUL, R. & STAMATOYANNOPOULOS, J. A. 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science,* 337**,** 1190-1195.

MCARTHUR, M., GERUM, S. & STAMATOYANNOPOULOS, G. 2001. Quantification of DNasel-sensitivity by real-time PCR: Quantitative analysis of DNasel-hypersensitivity of the mouse beta-globin LCR. *Journal of Molecular Biology,* 313**,** 27-34.

MCBRIDE, D. J., BUCKLE, A., VAN HEYNINGEN, V. & KLEINJAN, D. A. 2011. DNaseI Hypersensitivity and Ultraconservation Reveal Novel, Interdependent Long-Range Enhancers at the Complex Pax6 Cis-Regulatory Region. *Plos One,* 6**,** e28616.

MCDANIELL, R., LEE, B. K., SONG, L., LIU, Z., BOYLE, A. P., ERDOS, M. R., SCOTT, L. J., MORKEN, M. A., KUCERA, K. S., BATTENHOUSE, A., KEEFE, D., COLLINS, F. S., WILLARD, H. F., LIEB, J. D., FUREY, T. S., CRAWFORD, G. E., IYER, V. R. & BIRNEY, E. 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science,* 328**,** 235-9.

MCGAUGHEY, D. M., VINTON, R. M., HUYNH, J., AL-SAIF, A., BEER, M. A. & MCCALLION, A. S. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. *Genome Research,* 18**,** 252-260.

MCGHEE, J. D., WOOD, W. I., DOLAN, M., ENGEL, J. D. & FELSENFELD, G. 1981. A 200-Base Pair Region at the 5' End of the Chicken Adult Beta-Globin Gene Is Accessible to Nuclease Digestion. *Cell,* 27**,** 45-55.

MCKNIGHT, S. L. & KINGSBURY, R. 1982. Transcriptional Control Signals of a Eukaryotic Protein-Coding Gene. *Science,* 217**,** 316-324.

MOORE, M. J. & PROUDFOOT, N. J. 2009. Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation. *Cell,* 136**,** 688-700.

MORGAN, H. D., SUTHERLAND, H. G. E., MARTIN, D. I. K. & WHITELAW, E. 1999. - Epigenetic inheritance at the agouti locus in the mouse. - 23**,** - 318.

NAGY, P. L., CLEARY, M. L., BROWN, P. O. & LIEB, J. D. 2003. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proceedings of the National Academy of Sciences of the United States of America,* 100**,** 6364-6369.

NEPH, S., STERGACHIS, A. B., REYNOLDS, A., SANDSTROM, R., BORENSTEIN, E. & STAMATOYANNOPOULOS, J. A. 2012a. Circuitry and dynamics of human transcription factor regulatory networks. *Cell,* 150**,** 1274-86.

NEPH, S., VIERSTRA, J., STERGACHIS, A. B., REYNOLDS, A. P., HAUGEN, E., VERNOT, B., THURMAN, R. E., JOHN, S., SANDSTROM, R., JOHNSON, A. K., MAURANO, M. T., HUMBERT, R., RYNES, E., WANG, H., VONG, S., LEE, K., BATES, D., DIEGEL, M., ROACH, V., DUNN, D., NERI, J., SCHAFER, A., HANSEN, R. S., KUTYAVIN, T., GISTE, E., WEAVER, M., CANFIELD, T., SABO, P., ZHANG, M., BALASUNDARAM, G., BYRON, R., MACCOSS, M. J., AKEY, J. M., BENDER, M. A., GROUDINE, M., KAUL, R. & STAMATOYANNOPOULOS, J. A. 2012b. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature,* 489**,** 83-90.

NOONAN, J. P. & MCCALLION, A. S. 2010. Genomics of Long-Range Regulatory Elements. *Annual Review of Genomics and Human Genetics, Vol 11.* Palo Alto: Annual Reviews.

OHLSSON, R., BARTKUHN, M. & RENKAWITZ, R. 2010. CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin. *Chromosoma,* 119**,** 351-360.

ONG, C. T. & CORCES, V. G. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics,* 12**,** 283-293.

OZSOLAK, F., SONG, J. S., LIU, X. S. & FISHER, D. E. 2007. High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol,* 25**,** 244-248.

PASTINEN, T. & HUDSON, T. J. 2004. Cis-acting regulatory variation in the human genome. *Science,* 306**,** 647-650.

PAUL, D. S., NISBET, J. P., YANG, T. P., MEACHAM, S., RENDON, A., HAUTAVIITA, K., TALLILA, J., WHITE, J., TIJSSEN, M. R., SIVAPALARATNAM, S., BASART, H., TRIP, M. D., GOTTGENS, B., SORANZO, N., OUWEHAND, W. H. & DELOUKAS, P. 2011. Maps of Open Chromatin Guide the Functional Follow-Up of Genome-Wide Association Signals: Application to Hematological Traits. *PLoS genetics,* 7**,** e1002139.

PIQUE-REGI, R., DEGNER, J. F., PAI, A. A., GAFFNEY, D. J., GILAD, Y. & PRITCHARD, J. K. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research,* 21**,** 447-455.

QUAIL, M. A., KOZAREWA, I., SMITH, F., SCALLY, A., STEPHENS, P. J., DURBIN, R., SWERDLOW, H. & TURNER, D. J. 2008. A large genome center's improvements to the Illumina sequencing system. *Nature Methods,* 5**,** 1005-1010.

R-DEVELOPMENT-CORE-TEAM 2004. *A language and environment for statistical computing.,* Vienna, R Foundation for Statistical Computing.

REN, B., ROBERT, F., WYRICK, J. J., APARICIO, O., JENNINGS, E. G., SIMON, I., ZEITLINGER, J., SCHREIBER, J., HANNETT, N., KANIN, E., VOLKERT, T. L., WILSON, C. J., BELL, S. P. & YOUNG, R. A. 2000. Genome-wide location and function of DNA binding proteins. *Science,* 290**,** 2306-9.

RICHMOND, T. J. & DAVEY, C. A. 2003. The structure of DNA in the nucleosome core. *Nature,* 423**,** 145-150.

ROH, T. Y., CUDDAPAH, S., CUI, K. R. & ZHAO, K. J. 2006. The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 15782-15787.

ROTH, S. Y. & ALLIS, C. D. 1992. Chromatin Condensation - Does Histone H1 Dephosphorylation Play a Role. *Trends Biochem Sci,* 17**,** 93-98.

ROZEN, S. & SKALETSKY, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.),* 132**,** 365-86.

SABO, P. J., HUMBERT, R., HAWRYLYCZ, M., WALLACE, J. C., DORSCHNER, M. O., MCARTHUR, M. & STAMATOYANNOPOULOS, J. A. 2004. Genome-wide identification of DNasel hypersensitive sites using active chromatin sequence libraries. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 4537-4542.

SABO, P. J., KUEHN, M. S., THURMAN, R., JOHNSON, B. E., JOHNSON, E. M., CAO, H., YU, M., ROSENZWEIG, E., GOLDY, J., HAYDOCK, A., WEAVER, M., SHAFER, A., LEE, K., NERI, F., HUMBERT, R., SINGER, M. A., RICHMOND, T. A., DORSCHNER, M. O., MCARTHUR, M., HAWRYLYCZ, M., GREEN, R. D., NAVAS, P. A., NOBLE, W. S. & STAMATOYANNOPOULOS, J. A. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods,* 3**,** 511-8.

SANDELIN, A., ALKEMA, W., ENGSTROM, P., WASSERMAN, W. W. & LENHARD, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research,* 32**,** D91-D94.

SARAGOSTI, S., MOYNE, G. & YANIV, M. 1980. Absence of Nucleosomes in a Fraction of Sv40 Chromatin between the Origin of Replication and the Region Coding for the Late Leader Rna. *Cell,* 20**,** 65-73.

SCHMID, C. D. & BUCHER, P. 2007. ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell,* 131**,** 831-832.

SCHMITT, S., PRESTEL, M. & PARO, R. 2005. Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev,* 19**,** 697-708.

SCHONES, D. E. & ZHAO, K. 2008. Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics,* 9**,** 179-191.

SCHREIBER, S. L. & BERNSTEIN, B. E. 2002. Signaling network model of chromatin. *Cell,* 111**,** 771-778.

SHATKIN, A. J. & MANLEY, J. L. 2000. The ends of the affair: Capping and polyadenylation. *Nature Structural Biology,* 7**,** 838-842.

SIEPEL, A., BEJERANO, G., PEDERSEN, J. S., HINRICHS, A. S., HOU, M. M., ROSENBLOOM, K., CLAWSON, H., SPIETH, J., HILLIER, L. W., RICHARDS, S., WEINSTOCK, G. M., WILSON, R. K., GIBBS, R. A., KENT, W. J., MILLER, W. & HAUSSLER, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research,* 15**,** 1034-1050.

SLADEK, R. & HUDSON, T. J. 2006. Elucidating cis- and trans-regulatory variation using genetical genomics. *Trends in Genetics,* 22**,** 245-250.

SMALE, S. T. & KADONAGA, J. T. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem,* 72**,** 449-479.

SPIVAK, J. L., TORETTI, D. & DICKERMA.HW 1973. Effect of Phenylhydrazine-Induced Hemolytic-Anemia on Nuclear-Rna Polymerase-Activity of Mouse Spleen. *Blood,* 42**,** 257-266.

SPROUL, D., GILBERT, N. & BICKMORE, W. A. 2005. The role of chromatin structure in regulating the expression of clustered genes. *Nature Reviews Genetics,* 6**,** 775-781.

STADLER, M. B., MURR, R., BURGER, L., IVANEK, R., LIENERT, F., SCHOLER, A., WIRBELAUER, C., OAKELEY, E. J., GAIDATZIS, D., TIWARI, V. K. & SCHUBELER, D. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature,* 480**,** 490-495.

STALDER, J., LARSEN, A., ENGEL, J. D., DOLAN, M., GROUDINE, M. & WEINTRAUB, H. 1980. Tissue-Specific DNA Cleavages in the Globin Chromatin Domain Introduced by Dnaase-I. *Cell,* 20**,** 451-460.

STRAHL, B. D. & ALLIS, C. D. 2000. The language of covalent histone modifications. *Nature,* 403**,** 41-45.

STRAKA, C. & HORZ, W. 1991. A Functional-Role for Nucleosomes in the Repression of a Yeast Promoter. *Embo Journal,* 10**,** 361-368.

SUTO, R. K., CLARKSON, M. J., TREMETHICK, D. J. & LUGER, K. 2000. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. *Nature Structural Biology,* 7**,** 1121-1124.

TAGLE, D. A., KOOP, B. F., GOODMAN, M., SLIGHTOM, J. L., HESS, D. L. & JONES, R. T. 1988. Embryonic Epsilon-Globin and Gamma-Globin Genes of a Prosimian Primate (Galago-Crassicaudatus) - Nucleotide and Amino-Acid Sequences, Developmental Regulation and Phylogenetic Footprints. *Journal of Molecular Biology,* 203**,** 439-455.

TAKAI, D. & JONES, P. A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 3740-3745.

TATE, P. H. & BIRD, A. P. 1993. Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics & Development,* 3**,** 226-31.

THORNTON, B. & BASU, C. 2011. Real-Time PCR (qPCR) Primer Design Using Free Online Software. *Biochemistry and Molecular Biology Education,* 39**,** 145-154.

THURMAN, R. E., RYNES, E., HUMBERT, R., VIERSTRA, J., MAURANO, M. T., HAUGEN, E., SHEFFIELD, N. C., STERGACHIS, A. B., WANG, H., VERNOT, B., GARG, K., JOHN, S., SANDSTROM, R., BATES, D., BOATMAN, L., CANFIELD, T. K., DIEGEL, M., DUNN, D., EBERSOL, A. K., FRUM, T., GISTE, E., JOHNSON, A. K., JOHNSON, E. M., KUTYAVIN, T., LAJOIE, B., LEE, B. K., LEE, K., LONDON, D., LOTAKIS, D., NEPH, S., NERI, F., NGUYEN, E. D., QU, H. Z., REYNOLDS, A. P., ROACH, V., SAFI, A., SANCHEZ, M. E., SANYAL, A., SHAFER, A., SIMON, J. M., SONG, L. Y., VONG, S., WEAVER, M., YAN, Y. Q., ZHANG, Z. C., ZHANG, Z. Z., LENHARD, B., TEWARI, M., DORSCHNER, M. O., HANSEN, R. S., NAVAS, P. A., STAMATOYANNOPOULOS, G., IYER, V. R., LIEB, J. D., SUNYAEV, S. R., AKEY, J. M., SABO, P. J., KAUL, R., FUREY, T. S., DEKKER, J., CRAWFORD, G. E. & STAMATOYANNOPOULOS, J. A. 2012. The accessible chromatin landscape of the human genome. *Nature,* 489**,** 75-82.

TOLHUIS, B., PALSTRA, R. J., SPLINTER, E., GROSVELD, F. & DE LAAT, W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell,* 10**,** 1453-1465.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. & PACHTER, L. 2010. Transcript

assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol,* 28**,** 511-5.

TSANKOV, A. M., THOMPSON, D. A., SOCHA, A., REGEV, A. & RANDO, O. J. 2010. The Role of Nucleosome Positioning in the Evolution of Gene Regulation. *Plos Biology,* 8.

TURNER, B. M. 2000. Histone acetylation and an epigenetic code. *Bioessays,* 22**,** 836-845.

URETA-VIDAL, A., ETTWILLER, L. & BIRNEY, E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nature Reviews Genetics,* 4**,** 251-262.

VALDAR, W., SOLBERG, L. C., GAUGUIER, D., BURNETT, S., KLENERMAN, P., O COOKSON, W., TAYLOR, M. S., RAWLINS, J. N. P., MOTT, R. & FLINT, J. 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics,* 38**,** 879-887.

VARSHAVSKY, A. J., SUNDIN, O. H. & BOHN, M. J. 1978. Sv40 Viral Mini-Chromosome - Preferential Exposure of Origin of Replication as Probed by Restriction Endonucleases. *Nucleic Acids Research,* 5**,** 3469-3477.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. Q. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J. H., MIKLOS, G. L. G., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, C., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z. M., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W. M., GONG, F. C., GU, Z. P., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z. X., KETCHUM, K. A., LAI, Z. W., LEI, Y. D., LI, Z. Y., LI, J. Y., LIANG, Y., LIN, X. Y., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B. X., SUN, J. T., WANG, Z. Y., WANG, A. H., WANG, X., WANG, J., WEI, M. H., WIDES, R., XIAO, C. L., YAN, C. H., et al. 2001. The sequence of the human genome. *Science,* 291**,** 1304-51.

VERNIMMEN, D., DE GOBBI, M., SLOANE-STANLEY, J. A., WOOD, W. G. & HIGGS, D. R. 2007. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *Embo Journal,* 26**,** 2041-2051.

VISEL, A., AKIYAMA, J. A., SHOUKRY, M., AFZAL, V., RUBIN, E. M. & PENNACCHIO, L. A. 2009a. Functional autonomy of distant-acting human enhancers. *Genomics,* 93**,** 509-513.

VISEL, A., BLOW, M. J., LI, Z. R., ZHANG, T., AKIYAMA, J. A., HOLT, A., PLAJZER-FRICK, I., SHOUKRY, M., WRIGHT, C., CHEN, F., AFZAL, V., REN, B., RUBIN, E. M. & PENNACCHIO, L. A. 2009b. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature,* 457**,** 854-858.

WANG, Q. B., CARROLL, J. S. & BROWN, M. 2005. Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Molecular cell,* 19**,** 631-642.

WASSERMAN, W. W. & SANDELIN, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics,* 5**,** 276-287.

WATERSTON, R. H., LINDBLAD-TOH, K., BIRNEY, E., ROGERS, J., ABRIL, J. F., AGARWAL, P., AGARWALA, R., AINSCOUGH, R., ALEXANDERSSON, M., AN, P., ANTONARAKIS, S. E., ATTWOOD, J., BAERTSCH, R., BAILEY, J., BARLOW, K., BECK, S., BERRY, E., BIRREN, B., BLOOM, T., BORK, P., BOTCHERBY, M., BRAY, N., BRENT, M. R., BROWN, D. G., BROWN, S. D., BULT, C., BURTON, J., BUTLER, J., CAMPBELL, R. D., CARNINCI, P., CAWLEY, S., CHIAROMONTE, F., CHINWALLA, A. T., CHURCH, D. M., CLAMP, M., CLEE, C., COLLINS, F. S., COOK, L. L., COPLEY, R. R., COULSON, A., COURONNE, O., CUFF, J., CURWEN, V., CUTTS, T., DALY, M., DAVID, R., DAVIES, J., DELEHAUNTY, K. D., DERI, J., DERMITZAKIS, E. T., DEWEY, C., DICKENS, N. J., DIEKHANS, M., DODGE, S., DUBCHAK, I., DUNN, D. M., EDDY, S. R., ELNITSKI, L., EMES, R. D., ESWARA, P., EYRAS, E., FELSENFELD, A., FEWELL, G. A., FLICEK, P., FOLEY, K., FRANKEL, W. N., FULTON, L. A., FULTON, R. S., FUREY, T. S., GAGE, D., GIBBS, R. A., GLUSMAN, G., GNERRE, S., GOLDMAN, N., GOODSTADT, L., GRAFHAM, D., GRAVES, T. A., GREEN, E. D., GREGORY, S., GUIGO, R., GUYER, M., HARDISON, R. C., HAUSSLER, D., HAYASHIZAKI, Y., HILLIER, L. W., HINRICHS, A., HLAVINA, W., HOLZER, T., HSU, F., HUA, A., HUBBARD, T., HUNT, A., JACKSON, I., JAFFE, D. B., JOHNSON, L. S., JONES, M., JONES, T. A., JOY, A., KAMAL, M., KARLSSON, E. K., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature,* 420**,** 520-562.

WEAVER, I. C. G., CERVONI, N., CHAMPAGNE, F. A., D'ALESSIO, A. C., SHARMA, S., SECKL, J. R., DYMOV, S., SZYF, M. & MEANEY, M. J. 2004. Epigenetic programming by maternal behavior. *Nature Neuroscience,* 7**,** 847-854.

WEI, C. L., WU, Q., VEGA, V. B., CHIU, K. P., NG, P., ZHANG, T., SHAHAB, A., YONG, H. C., FU, Y. T., WENG, Z. P., LIU, J. J., ZHAO, X. D., CHEW, J. L., LEE, Y. L., KUZNETSOV, V. A., SUNG, W. K., MILLER, L. D., LIM, B., LIU, E. T., YU, Q., NG, H. H. & RUAN, Y. J. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell,* 124**,** 207-219.

WEINTRAUB, H. & GROUDINE, M. 1976. Chromosomal Subunits in Active Genes Have an Altered Conformation. *Science,* 193**,** 848-856.

WEINTRAUB, H., PALTER, K. & VANLENTE, F. 1975. Histones H2a, H2b, H-3, and H4 Form a Tetrameric Complex in Solutions of High Salt. *Cell,* 6**,** 85-110.

WENDT, K. S., YOSHIDA, K., ITOH, T., BANDO, M., KOCH, B., SCHIRGHUBER, E., TSUTSUMI, S., NAGAE, G., ISHIHARA, K., MISHIRO, T., YAHATA, K., IMAMOTO, F., ABURATANI, H., NAKAO, M., IMAMOTO, N., MAESHIMA, K., SHIRAHIGE, K. & PETERS, J. M. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature,* 451**,** 7796-801.

WINGENDER, E., DIETZE, P., KARAS, H. & KNUPPEL, R. 1996. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Research,* 24**,** 238-241.

WITTKOPP, P. J. & KALAY, G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics,* 13**,** 59-69.

WU, C. 1980. The 5' Ends of Drosophila Heat-Shock Genes in Chromatin Are Hypersensitive to Dnase-I. *Nature,* 286**,** 854-860.

XI, H. L., SHULHA, H. P., LIN, J. M., VALES, T. R., FU, Y. T., BODINE, D. M., MCKAY, R. D. G., CHENOWETH, J. G., TESAR, P. J., FUREY, T. S., REN, B., WENG, Z. P. & CRAWFORD, G. E. 2007. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS genetics,* 3**,** 1377-1388.

YALCIN, B., FLINT, J. & MOTT, R. 2005. Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics,* 171**,** 673-681.

YALCIN, B., WONG, K., AGAM, A., GOODSON, M., KEANE, T. M., GAN, X., NELLAKER, C., GOODSTADT, L., NICOD, J., BHOMRA, A., HERNANDEZ-PLIEGO, P., WHITLEY, H., CLEAK, J., DUTTON, R., JANOWITZ, D., MOTT, R., ADAMS, D. J. & FLINT, J. 2011. Sequence-based characterization of structural variation in the mouse genome. *Nature,* 477**,** 326-9.

YANG, L., DUFF, M. O., GRAVELEY, B. R., CARMICHAEL, G. G. & CHEN, L. L. 2011. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol,* 12.

YOUNGSON, N. A. & WHITELAW, E. 2008. Transgenerational Epigenetic Effects. *Annual Review of Genomics and Human Genetics,* 9**,** 233-257.

YUAN, G. C., LIU, Y. J., DION, M. F., SLACK, M. D., WU, L. F., ALTSCHULER, S. J. & RANDO, O. J. 2005. Genome-scale identification of nucleosome positions in S-cerevisiae. *Science,* 309**,** 626-630.