THE UNIVERSITY *of* EDINBURGH

Conservation and divergence in higher order chromatin structure

Emily Victoria Chambers

The University of Edinburgh

A thesis presented for the degree of

Doctor of Philosophy

2013

The Institute of Genetics and Molecular Medicine

Western General Hospital

Crewe Road, Edinburgh, EH4 2XU

United Kingdom

Declaration

This thesis presents my original research work. Wherever the contributions of others were involved this is clearly indicated.

Emily V Chambers 2013

Acknowledgements

Firstly, I would like to acknowledge my supervisors Colin Semple and Wendy Bickmore for their guidance and support over the course of my PhD studentship. I would also like to thank all the members of the labs of Colin Semple, Wendy Bickmore and Martin Taylor for their help and advice. Particular mention goes to James Prendergast, Alison Meynert and Alida Kindt for their bioinformatic advice (and patience!). I would like to acknowledge the MRC Capacity Studentship for funding my studies over the last 4 years.

I thank my whole family for their support throughout all my studies at University and in everything that I do. A special thank you to Chris for his encouragement and motivation.

Finally, thanks to all the teachers across the years from Almondbury who sparked my enthusiasm for biology and genetics and to my parents for encouraging my interest in computing. I would not have done this without their inspiration.

Abstract

Aspects of higher order chromatin structure such as replication timing, lamina association and Hi-C inter-locus interactions have been recently studied in several human and mouse cell types and it has been suggested that most of these features of genome organisation are conserved over evolution. However, the extent of evolutionary divergence in higher order structure has not been rigorously measured across the mammalian genome, and little is known about the characteristics of any divergent loci defined. Here we generate an orthologous dataset combining multiple measurements of chromatin structure and organisation over many embryonic cell types for both human and mouse that, for the first time, allows a comprehensive assessment of the extent of structural divergence between different mammalian genomes. Comparison of orthologous regions confirms that all measurable facets of higher order structure are conserved between human and mouse, across the majority of the orthologous genome. This broad similarity is observed in spite of the substantial time since the species diverged, differences in experimental procedures among the datasets examined, and the presence of cell type specific structures at many loci. However, we also identify hundreds of regions showing consistent evidence of divergence between these species, constituting at least 10% of the orthologous mammalian genome and encompassing many hundreds of human and mouse genes. Divergent regions are enriched in genes implicated in vertebrate development, suggesting important roles for structural divergence in mammalian evolution. They are also relatively enriched for genes showing divergent expression patterns between human and mouse ES cells, implying these regions may underlie divergent regulation. Divergent regions show unusual shifts in compositional bias, sequence divergence and are unevenly distributed across both genomes. We investigate the mechanisms of divergence in higher order structure by examining the influence of sequence divergence and also many features of primary level chromatin, such as histone modification and DNA methylation patterns. Using multiple regression, we identify the dominant factors that appear to have shaped the physical structure of the mammalian genome. These data suggest that, though relatively rare, divergence in higher order chromatin structure has played important roles during evolution.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC.....Akaike information criterion
API.....Application programme interface
ChIP.....Chromatin immunoprecipitation
DAM.....DNA adenine methyltransferase
ESC .....Embryonic stem cell
FDR.....False discovery rate
Hi-C .....High resolution chromatin conformation capture
iPSC .....Induced pluripotent stem cell
LA .....Lamin association
LAD .....Lamin associating domain
LINE .....Long interspersed nuclear element
LOF .....Loss of function
LTR .....Long terminal repeat
MeDIP .....Methylated DNA immunoprecipitation
MRE.....Methylation-sensitive restriction enzyme
NPC .....Neural progenitor cell
RT.....Replication timing
SINE.....Short interspersed nuclear element
SNP .....Single-nucleotide polymorphism
TAD .....Topologically associating domain
TF .....Transcription factor

List of Publications

CHAMBERS, E. V., BICKMORE, W. A. & SEMPLE, C. A. M. 2013. Divergence of mammalian higher order chromatin structure is associated with developmental loci. PLoS Comput Biol, 9, e1003017. (See Appendix 10.5)

CHAMBERS, E. V. & SEMPLE, C. A. M. 2013. Chromatin structure and human genome evolution. eLS. John Wiley & Sons, Ltd.

CHAMBERS, E. V., KINDT, A. S. & SEMPLE, C. A. M. 2011. Opening sequence: computational genomics in the era of high-throughput sequencing. Genome Biol, 12, 310.

Chapter 1

Introduction

Topics included in this section:

- An introduction of the hierarchical layers of chromatin structure in mammals from the nucleosome to multi megabase structural domains.
- Primary level chromatin structure is subject to a number of different biochemical modifications, which alter the folding, and structure of chromatin.
- Chromatin structure can be regarded as bipolar in nature: open and accessible or compacted and inactive. These different chromatin environments harbour different properties such as replication timing, spatial positioning and expression levels.
- Comparative analyses across all facets of chromatin structure reveal correlations in structural features and domains across different species and experimental methods.
- The rate of mutation is not constant across the genome and has links to chromatin environments.
- Genes or loci that are identical in sequence but have different chromatin states, appear to be heritable and may underpin some disease states.

1.1. THE STRUCTURAL ORGANISATION OF THE EUKARYOTIC GENOME

1.1.1. THE NUCLEOSOME AND PRIMARY LEVEL CHROMATIN

The mammalian genome exists as an intricately structured three-dimensional environment comprised of linear DNA sequences that are compacted and organised in several hierarchical layers. Each layer is based on interactions between the DNA helix and proteins, and the term 'chromatin structure' covers all of these layers up to the chromosomal level. Each structural layer is subject to differing modifications and it is the relationships between the modifications at all levels of chromatin structure that create an 'epigenomic landscape'. The epigenome creates a bridge between genotype and phenotype, regulating the way the genome is expressed in different cell types, developmental stages and disease states (Goldberg et al., 2007).

At the primary level of chromatin structure, the eukaryotic nucleosome core particle is the key structural subunit. It is formed by a 147 bp section of DNA that is wrapped around an octomer of eight histone proteins This octamer is comprised of two copies of each of four histones: H2A, H2B, H3 and H4. The nucleosome is the primary unit of chromosome structure and the folding and chemical modification of long nucleosome arrays forms the basis for all higher order chromatin structures (Woodcock et al., 2006) (Figure 1.1). A linker histone (H1) is associated with the intervening stretch of DNA between nucleosomes and has a role in defining nucleosomal repeat length (Woodcock et al., 2006). Together, the folding of nucleosomes and linker DNA regions yield the 10 nm chromatin fiber, known as the 'beads on a string' array, which is hierarchically further compacted down into the chromosomes. It has been thought that an in between layer of compaction involves the formation of a 30 nm helical chromatin fiber via the addition of linker histones, but the existence of this step remains controversial (Wu et al., 2007). After discovering evidence of the 30 nm fiber from *in vitro* experiments, new experimental approaches including chromatin conformation capture and cryo-electron microscopy have failed to find evidence of the 30 nm fiber *in situ* (Fussner et al., 2011). Attempts to analyse this further have identified a family of different chromatin fibers highlighting the dynamic and complex nature of chromatin structure and the need to consider alternative models of chromatin folding (Bian

and Belmont, 2012). However, chromatin structure can be summarised as having at least three hierarchical layers. The first is the nucleosome array and the organization of nuclear processes such as transcription, the second is the higher-order organization of the chromatin fiber, and finally, the spatial arrangement of chromosomes within the cell nucleus (Misteli, 2007).

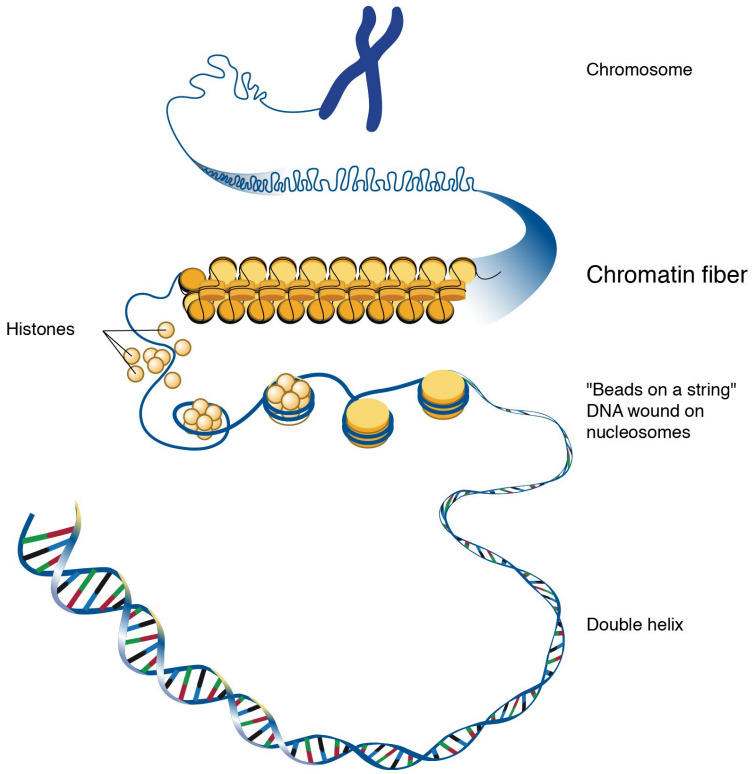

Figure 1.1 Compaction of primary level chromatin into nucleosomes. Each stage is shown from the assembly of DNA and histones into nucleosome arrays, structural organisation of the arrays and further compaction into the chromosomes (Open access image. Courtesy: National Human Genome Research Institute (NHGRI, 2010).

Genome-wide data relating to primary levels of chromatin structure, which include nucleosome occupancy, and histone modifications, in a variety of mammalian cell types are now abundant, due to the ability to map these chromatin features by combinations of new technological methods. These include Micrococcal Nuclease (MNase) digestion, which preferentially cuts linker DNA connecting two nucleosomes and is a useful tool for mapping nucleosome positioning. Chromatin Immunoprecipitation (ChIP) is a method for determining which specific proteins are associated with specific genomic regions, such as histones, transcription factors

or other DNA binding proteins. This can be coupled with high-throughput sequencing to identify which sequences of the genome are interacting with chromatin (Zhang and Pugh, 2011).

Nucleosome positioning plays a key role in chromatin organisation and gene regulation. It can regulate many DNA dependent processes, including transcription, replication and repair by physically limiting the access of binding proteins to incorporated DNA (Jiang and Pugh, 2009). Because of this, nucleosome free regions are often accessible sites of transcription factors. Some chromatin remodelling complexes are known to facilitate transcription initiation by regulating the formation and/or size of nucleosome free regions. These primarily work by either adding covalent post-translational modifications to nucleosomal histones or by moving, removing or restructuring nucleosomes in an ATP-dependent manner (Jiang and Pugh, 2009). This can include replacing core histones H2A and H3 with the variants H2A.Z and H3.3, which have been found to be enriched at nucleosome free regions at active promoters (Jin et al., 2009). It is thought that nucleosomes with these double variants disrupt the periodicity of nucleosome spacing leading to free and accessable chromatin regions for transcriptional access (Jin et al., 2009).

There are numerous histone modifications which can be incorporated into nucleosomes and alter the local properties of chromatin structure and function. They can be categorised by two main properties, those that affect interaction sites for binding proteins and those that change the charge of chromatin altering its compaction potential and can involve a variety of specialised protein complexes containing enzymes and chaperones (Sarma and Reinberg, 2005). The functional consequences of a number of histone modifications have been studied in detail. The most understood post-translational histone modifications are on the unstructured N and C-terminal tails that protrude from the nucleosome core. These harbour sites for modifications or 'marks' such as phosphorylation, methylation, acetylation and ubiquitination. Histones also contain a conserved histone-fold domain that may contain histone modifications although these are much less well understood (Tropberger and Schneider, 2013). They occur on surface of the histone octamer, close to the nucleosomal DNA, and they have the potential to alter histone-DNA interactions which can have a direct affect on chromatin dynamics (Tropberger and Schneider, 2013). The core histones that make up the nucleosome are known to be subject to many different posttranslational modifications (Tan et al., 2011).

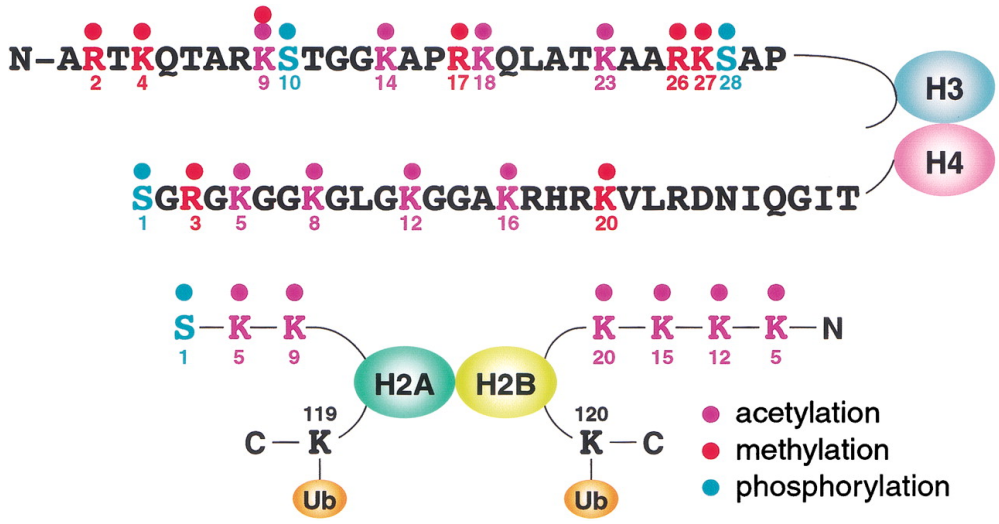

Figure 1.2 Sites of common histone tail post-translational modifications. The modifications shown include acetylation (pink), methylation (red), phosphorylation (blue), and ubiquitination (orange). Original image from (Zhang and Reinberg, 2001).

Modification of chromatin structure often results in either compaction and inactivation, creating a transcriptionally repressive chromatin environment or activating and opening, creating an accessible chromatin environment. This bipolar nature of chromatin structure is seen genome wide and can often span hundreds of kilobases. At the chromatin level, active domains correspond to high levels of histone acetylation, such as H3ac (Roh et al., 2005) and are often the earliest replicated in the cell cycle. Repressed domains have low H3ac and undergo relatively late replication (Birney et al., 2007). Acetylation of lysine residues on histones is also generally associated with activation of transcription, since this neutralizes the positive charge of the lysine residues on the nucleosome and reduces the affinity of histones for DNA, which causes the chromatin to unravel for a more accessible chromatin structure (Kouzarides, 2007). Methylation is another class of histone modification whereby a methyl group is added to the N terminal tail of histone and is more complex in its function. Monomethylations of H3K27, H3K9, H4K20 and H2BK5 are all linked to mechanisms involved in gene activation (Barski et al., 2007). Additionally, trimethylation of H3K4 is an indicator for active human promoters (Vermeulen et al., 2007). However, trimethylations of H3K27 and H3K9are linked to repression (Barski et al., 2007). For example, at polycomb repressive complex (PRC) targets, including Hox loci, H3K27 methylation

(H3K27me3) provides the binding site for the chromodomain of polycomb homologues as part of the PRC1 complex, which is required to maintain a closed chromatin state (Eskeland et al., 2010). However, defining functional chromatin environments by the type of histone marks present is far from simple. From genome-wide profiling of histone modifications, regions of chromatin have been discovered that contain both repressive H3K27 methylation and activating H3K4 tri-methylation in mouse embryonic stem cells (Bernstein et al., 2006). It is suggested that these genes, so called bivalent, keep genes in a transcriptionally 'poised' state at low expression levels. During cell differentiation, one of the modifications is preserved while the other is lost leading to silencing or activation. This can serve as a means of preserving pluripotency and maintaining tight transcriptional control (Bernstein et al., 2006). However, ChIP assays are unable to unequivocally establish the coexistence of both marks on the same allele in a given cell. Thus, it has been argued that in some cases, the observed bivalency might reflect cellular heterogeneity arising from the averaging of cells that carry either, but not both, marks at a given locus (Voigt et al., 2013).

DNA methylation, which involves the addition of a methyl group to cytosine bases, largely at CG dinucleotides, is an epigenomic feature that influences transposon silencing, X chromosome inactivation, and imprinting, however it is predominantly known for important roles in maintaining transcriptional repression (Bird, 2011). DNA methyltransferase enzymes facilitate the addition of methyl groups to DNA and vertebrates have almost genome-wide methylation apart from at CpG islands at promoter regions (Suzuki and Bird, 2008). Recently, technological advances in sequencing methods have enabled large-scale mapping of eukaryotic methylation profiles. Complex, multi-cellular organisms have been shown to have higher levels of genomic methylation; this may provide additional layers of regulation to control development in complex organisms (Zemach et al., 2010). DNA methylation patterns are variable between cell types and can alter during cell differentiation indicating they play important roles in the cell type specific expression. Similarly, polycomb group proteins form chromatin-associated complexes that act as repressors for genes involved in embryonic development and cell-fate (Bernstein et al., 2006). It is now known that these two epigenetic processes are closely linked as DNA methylation plays a critical role in genomic distribution of H3K27me3 which is important for the genomic targeting of the PRC2 polycomb complex (Lister et al., 2009, Eskeland et al., 2010). This has recently been shown to

be required for polycomb-mediated gene repression (Reddington et al., 2013).

The combined relationships between histone modifications and DNA methylation have a cumulative effect on the architecture of higher order chromatin structure. This in turn affects DNA interactions and gene expression. Given the vast number of possible combinations of known chromatin marks, fully understanding the information they encode is a huge challenge. Integrative studies combining different epigenomic datasets have shown that multiple chromatin states can affect gene expression status (Ernst et al., 2011, Ram et al., 2011). Recently, large scale, genome-wide mapping of histone modifications and related structures have emerged as an effective method for characterizing the functional consequences of chromatin structure. Large ChIP-seq datasets show strong combinatorial signals, such that groups of correlated histone marks indicate regions belonging to distinct functional classes. A study carried out on chromatin associated proteins in Drosophila defined five types of chromatin organisation within large domain like structures (Filion et al., 2010). They found closed structures that covered half the genome and further defined different aspects of open structure by the presence of particular classes of histone modifications. A ChIP study of 29 human chromatin associated proteins revealed six predominant combinations (Ram et al., 2011), but no associations were made to histone marks. Most notable among recent studies have been the efforts of the ENCODE consortium which has sought to define chromatin environments using computational methods to summarise biologically-meaningful combinations of chromatin marks (Ernst and Kellis, 2012). Using this method, there have been 25 different chromatin states defined that can be grouped into seven classes, emphasizing biological differences, these are transcription start site (TSS), promoter flanking (PF), enhancer (E), weak enhancer (WE), CTCF binding (CTCF), transcribed region (T) and repressed (R) (Hoffman et al., 2013). The ENCODE project has produced integrated maps of chromatin elements across differing resolutions, making it possible to explore chromatin states at single-nucleotide resolution. Focusing on 12 histone modifications in 46 cell types the ENCODE consortium has also found that the presence of particular combinations could accurately predict weakly and strongly activated promoter and enhancer regions (Dunham et al., 2012). Even more striking results demonstrated that combinations of histone marks at promoters could be used to quantitatively predict transcriptional output, with such combinations explaining around 90% of

expression variation across all genes in the genome (Dunham et al., 2012).

1.1.2. HIGHER ORDER CHROMATIN STRUCTURE

Higher order chromatin structure concerns the spatial conformation of nucleosome arrays and the relative accessibility of multi megabase domains of DNA sequence. Early studies of higher order chromatin structure defined a bipolar organisation of the genome whereby structure was seen as either relatively compacted and 'closed' or accessible and 'open'. One of the first investigations of higher order chromatin structure was carried out by using sucrose sedimentation and hybridization to genomic microarrays (Gilbert et al., 2004). Using this technique relatively closed and open chromatin fiber structures in the human genome were defined at a low resolution (1 Mb spaced BAC clones arrayed across the genome). High throughput methods have revolutionized the understanding of higher order chromatin structures over the past few years, portraying a variety of epigenomic landscapes across the range of relatively open and closed structures down to single base pair resolution.

From the outset, links have been made between higher order structure and gene expression. Genome wide studies directly examining higher order chromatin structure in human have indicated that protein coding genes are enriched in open chromatin where there is an accessible environment for transcriptional machinery (Gilbert et al., 2004). This open structure may provide an environment to maintain clusters of widely expressed genes together throughout evolution (Sproul et al., 2005). It has been found that chromatin environments contribute directly to expression levels of embedded genes. Identical transgenes integrated in different chromosomal regions may acquire expression levels that strongly correlate with the expression levels of the large (containing up to 80 genes) domains of integration (Gierman et al., 2007).

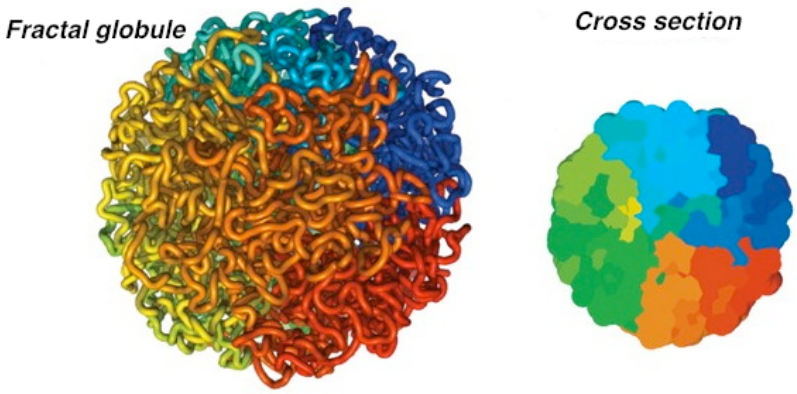

Figure 1.3 Fractal globule formation of higher order chromatin structure as reported in Lieberman-Aiden et al (2009). Regions in close proximity are visible and chromosome territories are represented by a single colour. Original image from Lieberman-Aiden et al (2009).

More recent studies of higher order structure employ Hi-C, a technique based on formaldehyde cross linking frequency, which is used to map the three-dimensional organization of chromosomes by coupling DNA proximity ligation with high throughput sequencing (de Wit and van Steensel, 2009). These studies have rediscovered the domain structures previously seen across the genome. The first Hi-C study suggested a model of genome organisation that was compartmentalised, with chromatin occupying two different types of spatial compartment (Lieberman-Aiden et al., 2009). One compartment was composed of regions of gene rich, relatively open, actively transcribed chromatin, and another contained regions with opposing features. Nuclear organisation of chromatin was reported to be consistent with an untangled 'fractal globule' conformation allowing genomic regions to loop in and out of foci for transcriptional activation (Figure 1.3; See also Figure 3.3, Figure 3.4) (Lieberman-Aiden et al., 2009). Similar structures have since been reported in other organisms (Duan et al., 2010), and at much higher resolution within the human genome (Kalhor et al., 2012, Sexton et al., 2012), although the organisation of the fractal globule remains a subject of debate (Mirny, 2011). Higher resolution Hi-C data has suggested the genome is organized into "large, megabase-sized local chromatin interaction domains", where the vast majority of regulatory interactions structures between ESC and IMR90 (lung fibroblasts) found that most of the boundary locations were shared. This indicates that the domain structure between cell types is stable but regions within each

domain take part in cell type-specific regulatory events (Dixon et al., 2012).

Figure 1.4 Hi-C interactions reveal topologically associated domains. The heatmaps represent interaction frequencies of the underlying genomic DNA (y axis) where interaction frequencies within domains are higher than between. The bars at the top represent topological domains. Interaction maps obtained from (http://chromosome.sdsc.edu/mouse/hi-c/database.php).

One of the major features of nuclear organisation rediscovered in Hi-C analyses is that each chromosome is spatially positioned in its own individual territory within the nucleus. It has been known for some time that transcriptionally active genes within each chromosome territory are often positioned at the centre of the nucleus whereas silent genes are located at the periphery (Boyle et al., 2001). These early studies involved fluorescence *in situ* hybridization (FISH), a technique that uses fluorescent probes to bind to specific regions of DNA, which can then be spatially identified and visualised using microscopy. Some genomic regions within open chromatin are visible cytologically as relatively decondensed loci that 'loop out' to a new physical position in the nucleus (Mahy et al., 2002). This is consistent with previous observations at individual loci. For example, when transcriptionally active, the entire mouse HoxB locus decondenses cytologically and loops out from its usual nuclear position within the territory (Chambeyron and Bickmore, 2004). There is now substantial experimental evidence for the looping of chromatin to facilitate interactions at a distance, such as those between transcriptional activators bound to enhancers and the transcriptional machinery at a promoter (Tolhuis et al., 2002, Cullen et al., 1993). In fact it has become clear that there are many long-range interactions both between and within chromosomes that are associated with whether genes at these loci are transcriptionally active or not (Simonis et al., 2006). This is consistent with observations that multiple active genes are often found

together at locations in the nucleus that have high local concentrations of transcriptional and mRNA-processing machinery. These are known as 'transcription factories' (Rieder et al., 2012). Some intra or interchromosomal interactions appear to be mediated by regulatory elements, perhaps via direct protein-protein interactions between loci, to promote or repress transcription by restricting the movement of genes (Fraser and Bickmore, 2007).

At the higher level of chromatin organisation there are vital genomic functions that have close relationships to chromatin environments. Replication timing refers to the coordinate replication of segments of DNA within S-phase of the cell cycle via the synchronised firing of clusters of replication origins (replicons) (Hiratani et al., 2008). These often form replication domains ranging in size from a few hundred kilobases to several megabases and undergo replication at distinct times during the cell cycle. Such domains have been characterized across both human and mouse genomes in a variety of different cell types (Hiratani et al., 2008, Hiratani et al., 2010, Ryba et al., 2010). Up to 45% of the mouse genome has been shown to have significant changes in replication timing during development highlighting the cell type specific nature of replication timing (Hiratani et al., 2010). Changes in replication timing are accompanied by a change in spatial positioning and transcription of the genomic region involved and may be a mechanism for genome wide transcriptional changes upon lineage commitment (Hiratani et al., 2010). It has been found that replication domains that replicate at different times occupy different spatial compartments within the nucleus. Regions that replicate early are more centrally located and regions that replicate later are located at the nuclear periphery (Hiratani et al., 2008). Also, early replicating regions tend to be GC and gene rich whereas late replicating regions are GC and gene poor (Woodfine et al., 2004). Late replicating domains are enriched for the histone modification H3K9 dimethylation but cells lacking H3K9me2 do not have their replication timing or their spatial positioning disrupted (Yokochi et al., 2009) In contrast, the gene Rif1, which is a telomere binding protein that binds chromatin and associates with nuclear scaffold structures during interphase, has been shown to dramatically affect replication timing profiles within cells (Yamazaki et al., 2013).

Spatial positioning of genomic locations is another important aspect of higher order chromatin organisation. Genomic regions occupy different spatial compartments within the nucleus and differ in their proximity to the nuclear

periphery, and again it seems that megabase scale domains are the unit of organisation. Nuclear lamina associating domains (LADs) are strongly associated with the nuclear periphery, contain around half of the human genome, and are between 40 Kb to 15 Mb in size (Guelen et al., 2008). Their locations and their sizes appear to be largely constant over cell types although there are domain boundaries that appear to be cell type specific (Peric-Hupkes et al., 2010). The majority of genes enriched in LADs have low transcriptional activity, which suggests a repressive chromatin environment (Peric-Hupkes et al., 2010), and this repressive state is found in LADS from human, mouse and Drosophila cells (Peric-Hupkes et al., 2010, Pickersgill et al., 2006, Guelen et al., 2008). Using experimental methods to artificially tether an active locus to the nuclear lamina has been shown to lead to reduced gene expression (Finlan et al., 2008). Genes within LADS are also enriched for repressive histone modifications such as H3K9me2, which is also true of genes within late replicating domains (Guelen et al., 2008, Ryba et al., 2010). H3K9me2 has also been shown to have an important relationship to LAD positioning in mouse cells (Kind et al., 2013).

Remarkably, given the diverse methodologies used to investigate them, significant correlations have been found among the very different facets of higher order chromatin that have been measured. There is a strong correlation between the regions that replicate together during the same temporal window of S phase, and those sequences that can be captured together by Hi-C (Ryba et al., 2010, Yaffe et al., 2010). This is consistent with the idea that genomic regions in close proximity tend to replicate at similar times and thereby define important features of chromosome organisation. This has been substantiated by more detailed analysis using fluorescence *in situ* hybridisation of specific loci (Hiratani et al., 2008, Ryba et al., 2010). There is also a strong correlation between late replicating chromosomal domains and LAD organisation, but the relationship tends to breakdown the borders of the domains and at particular genes (Peric-Hupkes et al, 2010). This is unsurprising considering the dynamic nature of higher order structure, where hundreds of genes within LADs change their genomic locations during, for example cellular differentiation, such as moving from a peripheral location to a central position where the gene becomes transcriptionally active (Peric-Hupkes et al., 2010). Similar transitions in state have also been observed for replication timing domains, where whole regions of DNA change their replication timing upon differentiation from early to late, or late to early replication (Hiratani et al., 2008, Ryba et al., 2010).

Early to late changing regions have been shown to have compact, inaccessible structure possibly as a means of stable gene silencing (Takebayashi et al., 2012). The fact that significant correlations between different facets of higher order structure are found in spite of these 'moving targets' is remarkable.

1.2. COMPARATIVE EPIGENOMICS OF PRIMARY LEVEL CHROMATIN

Evolutionary comparisons of epigenomic features between species are a powerful tool for understanding the conservation of genome organisation. The relationship between divergence in the epigenome and the divergence in the underlying DNA sequence can be studied and may shed light on regulatory features of the genome that cannot be discerned from sequence comparisons alone. A number of studies have been undertaken to compare the structures of orthologous loci between species using a variety of chromatin data. One of the first comparative epigenomic studies between human and mouse examined human 'islands' of histone acetylation and found that similar islands were present at most orthologous regions in mouse, but often with no detectable DNA sequence conservation underlying them (Roh et al., 2007). Histone modification states coupled with polycomb binding sites within human and mouse orthologous promoters have been studied genome-wide with widespread conservation of chromatin states seen between species (Ku et al., 2008). On a broader scale, patterns of enrichment for common histone modifications across the orthologous mammalian genome also appear to be generally conserved between human and mouse (Woo and Li, 2012). The first genome-wide maps of DNA methylation for over 20 eukaryotic species were completed to investigate evolutionary patterns, and again conservation of chromatin is evident (Feng et al., 2010, Zemach et al., 2010). Methylation of the gene body is a particularly ancient feature of eukaryote genomes, predating the divergence of plants and animals around 1.6 billion years ago (Zemach et al, 2010). It may have originated to prevent transcriptional initiation within the gene body, however this does not extend to transposons and repeat elements which show increased divergence in DNA methylation patters across species (Zemach et al., 2010). Non-methylated CpG islands are also present at orthologous regions within seven diverse vertebrate species indicating conservation across vast evolutionary distances. They are usually present at gene promoters but some cell type specific CpG islands are found in other regulatory sequences such as enhancers (Long et al., 2013).

Transcription factors are key elements of chromatin structure as their binding patterns are central to the expression of genes. Sequence-specific transcription factor binding sites appear to evolve rapidly in mammals, with binding events in a particular tissue shared only 10-22% of the time between human, mouse and dog genomes (Schmidt et al., 2010). The spread of transposons with integral binding sites contributes to much of this rapid divergence between human and mouse cells (Schmidt et al., 2012). In addition, the divergence of transcription factor binding patterns between human and chimpanzee greatly exceeds the level of sequence variation between the two species suggesting that divergence in chromatin structure itself may play an important role in species divergence (Kasowski et al., 2010). A more recent comparative study has examined high resolution chromatin data including eight histone modifications/variants, DNA methylation patterns and the binding patterns of four transcription factors in stem cells from three different mammals - human, mouse and pig (Xiao et al., 2012). Epigenomic maps for each species were constructed that could then be compared at orthologous regions. They again found evidence for conservation of chromatin features, such that the intensity of a histone modification, or of DNA methylation, found at a genomic region was generally correlated across species. In addition, the combinations of chromatin features predictive of gene expression levels were almost identical between species, including H3K4me3, H3K36me3, H3K27me3, and H3K27ac. As this implies, changes in chromatin structure between species were also predictive of gene expression changes. This offers a stark contrast with sequence divergence, which has been found to correlate poorly with expression divergence between species, including human and mouse (Chan et al., 2009).

1.3. COMPARATIVE EPIGENOMICS OF HIGHER ORDER CHROMATIN STRUCTURE

Conservation of higher order chromatin structure across species is less well studied than primary level chromatin features, such as histone modification patterns, but a number of studies have been carried out recently. As mentioned previously, genome wide studies of lamin associating domains (LADs) and replication timing domains, both important aspects of higher order structural organisation, have been carried out in human and mouse cell types (Hiratani et al., 2008, Guelen et al., 2008, Peric-Hupkes et al., 2010, Ryba et al., 2010). These domains adopt similar sizes between species and also display broad conservation between species. In replication timing domains, orthologous regions show coordinate replication late or early in the cell cycle. This conservation has been maintained in spite of the numerous large-scale genome rearrangements separating the two species (Yaffe et al., 2010), and is reflected in similar compositional biases in these genomes. Where domains of replication timing are divergent between species, it is thought that differences in the underlying DNA sequence may play a primary role (Pope et al., 2012).

Comparisons of mouse and human lamina interaction maps have shown that the sizes and genomic positions of these domains are strongly conserved. Constitutive LADs (cLADs), seen as constant features across differing cell types, are particularly depleted of synteny breakpoints and are characterized by long stretches of AT rich genomic sequence (Meuleman et al., 2013). The degree of cell type specificity varies among these datasets. Replication timing similarity between species is dependent on the particular cell type examined (Ryba et al., 2010). Some regions of the genome have been shown to have a constant replication profile while there are other more plastic domains that show variation across differing cell types (Hansen et al., 2010). However, the numbers and size distributions of LADs in human lung fibroblasts are reported to be similar to those seen in mouse embryonic fibroblasts, as well as several other mouse cell types (Peric-Hupkes et al., 2010). A recent Hi-C study examined mouse and human ES cells in parallel and found further evidence for megabase sized domain organisation of higher order chromatin structure. These domains were broadly stable across cell types and conserved across species (Dixon et al., 2012). As mentioned above, these 'topological' domains are

enriched in mutually interacting subregions, presumably reflecting the presence of genes and associated regulatory elements. Topological domain boundaries were also compared to LADs and replication timing domains and were found to be related, but independent, chromatin features (Dixon et al., 2012). There are also correlations between higher order chromatin structures and chromatin features at a finer scale of organisation, for example histone modifications and transcription factor binding patterns (de Wit and van Steensel, 2009). A key question arising from these findings is how are all the various chromatin features related and to what extent are they all aspects of the same entity? I present data directly addressing this question in Chapter 6. It is also increasingly apparent that correlations can be found between any published higher order chromatin dataset and underlying patterns of histone modifications, although the extent to which lower level features are responsible for higher order structure remains a subject of debate (Figure 1.5). I will consider this question further in Chapter 8.

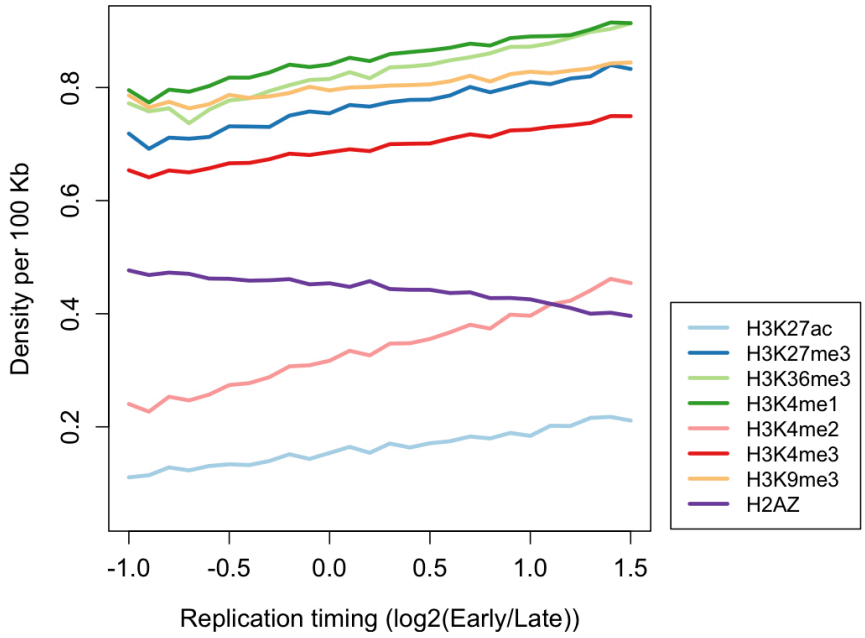

Figure 1.5 Higher order chromatin structure and histone modifications. The relationships between the densities of various histone modifications (Xiao et al., 2012) and replication timing data in human ES cells (Ryba et al., 2010). Replication timing and histone modification data is averaged over 100 Kb windows. Adapted from (Chambers and Semple, 2013).

1.4. DNA VARIATION WITHIN CHROMATIN DOMAINS

There are now many examples of the interplay of chromatin structure with the underlying DNA sequence, such as the broad compositional bias of GC rich sequence in open regions mentioned already (Yaffe et al., 2010, Meuleman et al., 2013). There are also clear examples of sequence divergence impacting chromatin, such as the spread of transposons affecting mammalian CTCF binding sites (Schmidt et al., 2012). However, it has also become clear that chromatin structure is associated with unusual patterns of genomic sequence divergence. Higher order structure has been compared to various measures of mutation rate and is consistently correlated. Divergence between human and chimpanzee sequence at intronic sites, intergenic sites and ancient repeats have indicated higher mutation rates in relatively closed domains (Prendergast et al., 2007). A similar correlation has

been observed between mutational rate and human DNA replication timing, with higher mutation rates occurring in late replicating regions (Stamatoyannopoulos et al., 2009). Single nucleotide polymorphism (SNP) density is also used to infer mutation rates as it is presumed that there has been little time for selection to act on the majority of SNPs and therefore the density of SNPs within a region should generally reflect underlying mutation rates. Again, SNP density shows a similar relationship to other mutational measures and is at a higher level in closed chromatin structures (Prendergast et al., 2007). This implies that variation in the rate of mutation across the genome has a relationship to higher order chromatin, however, the mechanisms underlying this association are not clear. It has been suggested that the silent chromatin located at the nuclear periphery may act as a 'bodyguard', absorbing mutational damage and protecting the centrally located genes (Gazave et al., 2005). An alternative theory states that open chromatin is more accessible to the DNA repair machinery, allowing any mutagenic alterations to the DNA to be repaired. This is consistent with a study, which showed homology directed repair machinery has restricted access to compact chromatin (Cummings et al., 2007).

On the other hand, other studies have observed relatively high substitution rates at small sites of particularly open chromatin structure such as DNase hypersensitive sites (Birney et al., 2007) and core promoter regions immediately upstream of transcription start sites (Taylor et al., 2006). This raises the possibility that while large open chromatin domains are resistant to mutation, there are smaller sites with unusually accessible structures that are particularly prone to mutation. Mutation rates have also been found to vary according to human nucleosome occupancy patterns, where it seems that less accessible sequences, such as those wrapped around the nucleosome cores, accumulate mutations more rapidly than sequences in the open linker regions (Prendergast and Semple, 2011). This has also been observed in studies of medaka fish (Sasaki et al., 2009) and yeast (Warnecke et al., 2008, Washietl et al., 2008). There is also significant evidence for both positive and negative selection linked to human nucleosome positioning once mutational biases are accounted for, implicating a widespread and important role for DNA sequence in the location of well-positioned nucleosomes (Prendergast and Semple, 2011). Chromatin domains possessing particular structures also differ significantly in their gene content, suggesting that higher order chromatin may play roles in the evolution of gene clustering and synteny. Relatively accessible open chromatin is

enriched for broadly expressed, 'housekeeping' genes, while more closed conformations are enriched for particular classes of noncoding RNA (ncRNA) genes (Prendergast et al., 2007). Also, particular chromatin signatures have been used to successfully predict functional ncRNAs (Guttman et al., 2009). However, chromatin structure does not always diverge in parallel with the underlying DNA sequence. It has been found that chromatin structure can be well conserved whether it is associated with genomic sequences showing accelerated or reduced substitution rates (Xiao et al., 2012).

1.5. NATURAL VARIATION IN CHROMATIN STRUCUTRE

Studies of transgenerational epigenetic inheritance have provided insights into how inherited features of chromatin structure may be capable of evolving independently of the underlying DNA sequence. This process has been studied more in plants than animals, where the study of epigenetic variants, or epialleles, over many generations is well established (Schmitz and Ecker, 2012). Studies of a collection of Arabidopsis mutation accumulation lines over 30 generations have revealed just under one sequence mutation per line per generation. In contrast the level of spontaneous generation of epialleles (variation in DNA methylation patterns) was at least four orders of magnitude greater and often resulted in significant transcriptional variation at the affected loci (Schmitz and Ecker, 2012). DNA methylation patterns have also been studied in Arabidopsis plants from several different populations and the methylomes showed marked differences suggesting heritable epialleles may be involved in adaptations to diverse environments (Schmitz et al., 2013). Until recently it was thought that the inheritance of such variations was impossible in mammals and, if it did occur was likely to be functionally irrelevant, however those views have been challenged by recent data. Genome-wide epigenomic reprogramming takes place in early embryos, where their chromatin structure is changed to allow cellular differentiation later in development. It seems that thousands of loci can escape reprogramming, maintaining their DNA methylation status, and a small proportion of these include promoters and regulatory elements (Hackett et al., 2013). There have been few studies of the inheritance of human chromatin, but it has been shown that around 10% of DNase sensitive sites and CTCF binding sites mapped are specific to individuals and are often associated with the activation or repression of neighbouring genes (McDaniell et al., 2010). These individual variations in

chromatin structure were also often observed to be inherited between parents and children, and some could not be related to underlying sequence variation (McDaniell et al., 2010). A significant degree of heterogeneity can also be present between the chromatin states of different genomic sequence alleles. We are only at the beginning of understanding epigenetic inheritance in mammals but it suggests new links between genes and the environment mediated by chromatin structure and RNA biology (Daxinger and Whitelaw, 2012).

1.6. AIMS OF INVESTIGATION

It has now been over a decade since the completion of the human genome project, but it is clear that much of its potential in the field of scientific and medical research is still to be understood. With the advent of low-cost, high-throughput sequencing technologies, there is now an abundance of genomic, transcriptomic and epigenomic data. The flood of new data and analysis techniques is causing major shifts in bioinformatics and has direct implications for this project.

The studies mentioned above provide complementary views of higher order chromatin structure. Each shows that the mammalian genome is organised into large, discrete domains of higher order chromatin with relatively open or closed conformations. These forms have opposing properties, which include levels of expression and accessibility, spatial positioning, replication timing, histone marks and mutational rates. These domains appear to be broadly similar across the different cells that have been examined, although many regions across the genome show cell type specific structure. However, the actual extent to which these datasets intersect, and how they relate to one another across cell types and species, is poorly understood. The degree of evolutionary divergence in higher order structure has not been rigorously measured across the mammalian genome until now. The relevant studies published so far have generally examined a single feature of chromatin structure in isolation. Similarly, the genomic loci underlying divergence in chromatin structure between species, and the mechanisms underlying divergence, are unknown.

In this investigation, a large number of diverse mouse and human datasets are collated to provide the most comprehensive overview of higher order chromatin structure in mammals to date. A systematic study of all orthologous regions in the mammalian genome is undertaken and the extent of conservation in higher order chromatin structure between cell types and during evolution is estimated. The

analysis identifies large tracts of structurally divergent chromatin, unevenly distributed across the genome, and containing intriguing enrichments of particular classes of genes.

Chapter 2

Methodology

Topics included in this section:

- Summary of the data used and methods of defining orthology and divergence.
- Methods involved in defining the distribution of structurally divergent regions and clusters of divergent regions.
- Techniques for identifying densities of different gene classes and functional gene enrichments.
- Identification of sequence level correlates of chromatin divergence, including base composition, repeat densities, sequence level divergence, segmental duplications and synteny.
- Multiple linear regression modelling to identify the most influential variables underlying chromatin divergence.
- Primary processing and integrative analysis of multiple ChIP-seq datasets for epigenomic comparisons of chromatin data across three species.
- Summary of programming languages, software and online resources used for this investigation.

2.1. METHODS FOR DEFINING ORTHOLOGOUS HIGHER ORDER STRUCTURES

Detailed descriptions of datasets used including cell types and experimental procedures are detailed in Chapter 3 and Table 3.1. Full details of methods developed to define structural orthology and divergence are also detailed in this chapter. A summary is included below.

2.1.1. SUMMARY OF DATASETS USED

Replication timing data in human and mouse embryonic cells were obtained from (Hiratani et al., 2008) and (Ryba et al., 2010) as log2(early replicating/late replicating) values. Full details of cell types involved can be seen in Table 3.1. Nuclear lamina association data in human and mouse embryonic cells were obtained from (Guelen et al., 2008) and (Peric-Hupkes et al., 2010). Both studies were based upon the DamID technique for labelling lamina associated sequences, where relative lamina association is represented by log2(Dam-fusion/Dam-only) values. Lastly, 100 Kb window genomic interaction probability matrix eigenvalues were defined for human lymphoblastoid cells using Hi-C data (Lieberman-Aiden et al., 2009). These values were found to reflect open and closed higher order chromatin structures positioned in different nuclear compartments. Although these data were not derived from embryonic cells it appears that many of the higher order patterns (as represented by interaction matrix eigenvectors) in Hi-C datasets are consistent between cell types (Lieberman-Aiden et al., 2009, Dixon et al., 2012).

Other Hi-C datasets (Kalhor et al., 2012) were examined to test for the presence of systematic biases that can affect Hi-C data (Yaffe and Tanay, 2011). These include the distance between restriction enzyme cut positions, the GC content of fragments, and uniqueness of short sequence reads which can cause biases within the Hi-C method (Yaffe and Tanay, 2011). However, it was concluded that the biases present in the Lieberman-Aiden et al (2009) dataset have little effect on the two compartment classification of the genome based upon these data, and therefore that the search for structurally divergent regions is unaffected. This is detailed further in Chapter 3.

2.1.2. ORTHOLOGY AND DIVERGENCE

Probe based replication timing and nuclear lamina association data

coordinates were translated to the latest human or mouse genome assembly coordinates (Human Feb. 2009 (GRCh37/hg19) and Mouse Jul. 2007 (NCBI37/mm9)) using UCSC liftOver transformations (Kent et al., 2002). For each dataset the structural data values were averaged across probes into consecutive non-overlapping 100 Kb regions, but regions represented by fewer than 10 probes were discarded as they are underrepresented by the data and potentially unreliable. This allowed comparisons between the probe based datasets and the Hi-C data, which has a fixed resolution of 100 Kb. Within each species 100 Kb regions were collated across datasets where their coordinates overlapped by 50% or more. The result was a set of 24,711 mouse and 28,786 human 100 Kb regions represented by higher order structural values from multiple datasets. Orthologous 100 Kb regions were defined as those regions with at least a 50% coordinate overlap between mouse and human genomes using reciprocal liftOver transformations. A total of 16,820 100 Kb orthologous regions, covering 54% of the human genome and 62% of the mouse genome, were defined in this way. A total of 11,966 human and 7,891 mouse regions, lacking an orthologous mapping using this protocol, were excluded as lineage specific regions. Examination of several normalisation techniques in R revealed that standard quantile normalisation procedures (R/Bioconductor limma package) (Smyth, 2005), used to normalise across different microarray experiments, were effective across the different experimental platforms and cell types here, therefore this normalisation technique was implemented across all structural datasets for all 100 Kb regions (See Chapter 3, Figure 3.8). Note that this rather aggressive form of normalisation may obscure subtler differences between datasets and is therefore likely to make our divergence results (see below) conservative. Relationships across the 100 Kb orthologous dataset and between mean higher order structure values and other chromatin features was tested using Spearman's rho non-parametric correlation tests which do not rely on any underlying structural model for the data. Hierarchical clustering of divergent regions were performed using the R package pvclust (Suzuki and Shimodaira, 2006) with 1-rho as a distance metric and heatmaps were constructed using the gplots package in R (Warnes et al., 2010).

Structurally divergent regions were defined as orthologous 100kb regions that showed a consistent difference in higher order structural values across human and mouse data. Non-parametric tests from the SAM package (Tusher et al., 2001), analogous to two class unpaired t-tests with permutation derived p-values, were

used to define divergent regions (R package samr). These tests were developed for microarray data analysis but are appropriate for other types of non-microarray derived data (Tusher et al., 2001). Full details of the method development for this technique can be found in Chapter 3. To summarise, statistical tests analogous to t-tests were carried out for each 100 Kb orthologous region, with the various normalised structural values for that region compared between species. 100,000 permutations of the test results were used to estimate the false discovery rate (FDR). The FDR threshold was set to be relatively low (FDR = 2e-04) to ensure a very low number of false positives. 1719 100 Kb regions, showing a strong and consistent difference between species were defined despite the inherently noisy, variable nature of the collated dataset. The results are necessarily bipolar with positive and negative divergent regions defined to indicate human open/mouse closed or human closed/mouse open divergence respectively. Relatively static, non-divergent regions were defined as those with p values that did not pass the FDR threshold.

The 100 Kb detectably orthologous regions defined above (using a 50% overlap threshold) will necessarily vary in the degree of similarity they show between species, it was therefore a concern that this might influence the measurement of structural divergence. Specifically it was important to show that the regions identified as structurally divergent are not simply those most poorly aligned between species at the sequence level. On closer examination the distributions of overlaps (aligned nucleotides minus gaps) were found to be very similar between structurally divergent and non-divergent regions, whether viewed in terms of the human (GRCh37/hg19) genome (divergent overlap mean = 0.80, median = 0.81; non-divergent overlap mean = 0.79, median = 0.80), or the mouse (NCBI37/mm9) genome (divergent overlap mean = 0.73, median = 0.72; non-divergent overlap mean = 0.72, median = 0.71) sequence assemblies, based upon UCSC whole genome alignments. We concluded that the estimates of structural divergence are not simply a reflection of sequence divergence.

2.2. GENOMIC DISTRIBUTION OF DIVERGENT REGIONS

2.2.1. LARGE DOMAINS OF DIVERGENT REGIONS

The degree of topological clustering among the divergent regions was formally investigated by measuring the length distribution of consecutive runs of divergent 100 Kb regions observed, relative to the distribution expected using a permutation strategy. To do this, all consecutive runs of two or more significantly

divergent regions were first identified across the orthologous human genome using Perl scripts. These domains were required to maintain the polarity of divergence (i.e. all regions involved must be either human open/mouse closed or vice versa). The loci of the orthologous divergent regions were then permuted within chromosomes 10,000 times, and length of any consecutive runs within each permuted genome was noted. The frequency with which a run of a particular length was seen in the permuted datasets was taken as an approximate p value for runs of that length in the observed dataset. Runs of divergent regions greater than or equal to 400 Kb were never seen in the permutated data ($p < 0.0001$) and therefore runs of this size were regarded as significant large divergence domains in the observed data. 159 unexpectedly large divergent domains in the human genome and 160 in the mouse were defined in the way. This strategy is likely to be conservative in detecting large regions of divergent chromatin as it does not allow for gaps, (e.g. regions that may have marginally failed to reach significance in the test for divergence above), within runs of divergent regions. Full details of the 159 large human divergent domains and 160 large mouse divergent domains can be seen in Appendix 10.1.

2.2.2. DISTRIBUTION WITHIN CHROMOSOMES

The distribution of divergent regions across chromosomes was examined to find chromosomes that were particularly enriched or depleted for higher order structural divergence. This was done by comparing the expected numbers of divergent regions, given the proportion of orthologous 100 Kb regions on each chromosome, with those observed using chi-squared tests, and we identified chromosomes of interest as those generating standardized residuals > 1.96. This was done for divergent regions within the mouse and human genomes, and for the distributions of large divergence domains (see above) in both genomes.

Enrichment or depletion of 100 Kb divergent regions within subtelomeric or pericentromeric regions was assessed using a circular permutation strategy. Each permuted dataset was generated by shifting the locations of all divergent regions on each chromosome by a random number less than the length of the chromosome. Regions assigned a shifted position greater than the final base pair of the chromosome are re-inserted at the start of the chromosome plus the number of bases by which they exceeded the final base pair. This was done for 10,000 permutations. For the purposes of the permutations, the chromosomes are regarded

as circular and maintain the degree of clustering seen among the observed divergent regions. The number of permuted datasets, n, possessing a number of divergent regions within subtelomeric (or pericentromeric) regions greater than or equal to the observed number were noted, and used to calculate approximate p-values (n/10,000) for enrichment. The significance of depletion was calculated analogously, according to the number of permuted datasets possessing the same or fewer divergent regions. Subtelomeric regions were defined as regions within 1 Mb, 5 Mb and 10 Mb of the first and final base pairs of the chromosome assemblies, and within the final base pair of the (acrocentric) mouse assemblies. Pericentromeric regions were defined as regions within 1 Mb, 5 Mb and 10 Mb of the first base pair of mouse and human chromosome q arm assemblies, and within the final base pair of human p arm assemblies. It is important to note that the density of orthologous 100 Kb regions within subtelomeric regions was not significantly different from the genome as a whole, either for human (5 Mb subtelomeric region mean density = 23.70; mean density across all genomic 5 Mb bins = 28.10) or mouse (5 Mb subtelomeric region mean density = 34.60; mean density across all genomic 5 Mb bins = 34.20). The same circular permutation approach was used to measure the enrichment or depletion of divergent regions within domains that are structurally dynamic during cellular differentiation (Hiratani et al., 2008). We also used a similar permutation strategy to compare the similarity (i.e. proximity) of domain boundaries between chromatin-mediated regulatory domains (Dixon et al., 2012) and the boundaries of divergent clusters. The median distance between divergent cluster boundaries and the nearest regulatory domain boundaries was compared to the median distance seen in 10,000 datasets that had undergone circular permutation. The proportion of datasets generating a median distance less than or equal to the observed median distance was taken as an approximate p-value. Full details of the enrichment of divergent regions within subtelomeric and pericentromeric regions can be seen in Table 4.4.

2.3. CHROMATIN STRUCTURE CORRELATES

2.3.1. GENE DENSITY AND GENE ONTOLOGY ENRICHMENTS

Gene densities were calculated per orthologous region for divergent and non-divergent datasets. This involved examining overlaps of Ensembl (Flicek et al., 2013) annotated gene lists for protein coding genes within divergent and non-divergent regions for all 100 Kb orthologous regions defined. Densities per

divergent 100 Kb region were then compared using nonparametric (Mann-Whitney) test statistics.

Functional enrichments for protein coding genes were calculated using The Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al., 2003). Ensembl gene annotations were used to describe genes in divergent regions of interest against a background of Ensembl annotated genes in all orthologous regions. This was done for all 1719 divergent regions using the total human and mouse genes present within the 16,820 orthologous 100 Kb regions as background sets for human and mouse enrichment analyses respectively, and also for the divergent regions within various clusters of interest. Enrichment of each annotation term in the set of human or mouse genes present within divergent regions was assessed using default options (p-values calculated using the hypergeometric distribution with FDR correction). Enrichment of these gene sets within cytogenetic bands was also examined as this can reflect the clustering of divergent regions. Full details of gene enrichments in divergent regions and divergent clusters are in Appendix 10.2, 10.3 and 10.4.

RNA genes were also annotated by Ensembl and different RNA gene densities were established per 100 Kb orthologous region. Observed RNA gene densities were compared to expected RNA gene densities for divergent regions using the overall density of RNA genes for all orthologous regions.

Differential gene expression datasets were obtained in order to compare structural divergence to expression divergence between human and mouse. This was examined using three different divergent expression datasets. The first expression dataset calculated differential expression in a relatively low number of genes (497 divergent and 126 conserved) within human and mouse ES cells (Cai et al., 2010). Gene lists were compiled that contained three different types of orthologous genes. These were genes upregulated in human, genes that were upregulated in mouse and genes that were conserved in expression across both species. Numbers of genes showing divergent or conserved expression across human and mouse were calculated within the regions of interest. Fisher's Exact test was then used to calculate significant odds ratios for over or under representation of differentially expressed genes in 100 Kb regions with divergent higher order chromatin. The second study contained a higher number (186 divergent, 972 conserved) of differentially expressed genes in macrophage cells (Schroder et al.,

2012). Again, Fisher's Exact test was used to determine over or under representation of divergently regulated genes in divergent 100 Kb regions. Lastly, a dataset consisting of log2 fold change measurements for 7,673 mouse-human gene pairs was constructed using reads per kilobase per million reads (RPKM) expression values for human H1 ES cells (Lister et al., 2009) and mouse E14 ES cells (Xiao et al., 2012). These were used to calculate log2(human RPKM/mouse RPKM) for all one to one orthologous mouse human Ensembl gene pairs, as an estimate of fold change in expression.

2.3.2. BASE COMPOSITION AND REPEATS

To determine GC content, alignments of human (GRCh37/hg19) and mouse (NCBI37/mm9) were obtained from UCSC MultiZ 46-way alignment blocks using the Galaxy project website (Goecks et al., 2010). The alignments were restricted to intergenic regions using Ensembl gene predictions for each of the 16,820 100kb chromatin regions. This was to counteract the GC bias in gene coding regions. Where there were overlapping alignment blocks for a region, alignments with the best quality score were kept and blocks within each region were concatenated. Perl scripts were used to define the ratio of GC content per aligned intergenic sequence in a 100kb region. Repeat content of orthologous regions was identified using UCSC RepeatMasker annotation (Smit et al., 1996) and densities of specific repeat classes were calculated per 100kb structural regions using Perl scripts.

2.3.3. SEQUENCE LEVEL DIVERGENCE ESTIMATES

Human-mouse substitution rate was measured by using UCSC chain and net pair-wise sequence alignments (GRCh37/hg19 and NCBI37/mm9) (Kent et al., 2002). Gene predictions from the Ensembl project were used to identify the intergenic sequences between genes to avoid different rates of divergence within genic regions. All intergenic human/mouse sequence alignments within each 100kb orthologous chromatin structure region were identified and concatenated to determine the overall substitution rate for each region. This was measured by using baseml from the PAML package using the REV model (Yang, 1997). All bases that overlapped a CG dinucleotide in either species were removed from the alignments to conservatively calculate non-CpG rates of divergence.

Indel rate analysis was performed using three-way alignments for human (GRCh37/hg19), mouse (NCBI37/mm9) and dog (Broad/canFam2). The alignments were extracted from MultiZ 46 way alignment blocks and restricted to intergenic

regions (outside Ensembl gene models) using Galaxy (Goecks et al., 2010). Where there were overlapping alignment blocks for a region, alignments with the best quality score were kept. Alignment blocks for each 100kb region were defined and Perl scripts were used to define and filter lineage specific insertions and deletions. An indel was only defined if a minimum of three bases flanking each side of a gap were present and gaps occurring at orthologous locations but having unequal length in different species were also excluded (Kvikstad et al., 2007). Insertion and deletion rates were then calculated as the number of indel bases per aligned intergenic human base per 100 Kb divergent and non-divergent structural region.

SNP data from the 1000 Genomes Project (Abecasis et al., 2012) was used to calculate SNP density within the 100 Kb regions of interest. To do this, VCFtools (Danecek et al., 2011), a package used for manufacturing variant call format files, was used to identify SNP densities per 100 Kb region. The low coverage SNP database was used as this contained SNPs accurately predicted between many individuals sequenced to a depth of between 3 and 5X coverage. The SNP density option within VCFtools was used to reliably find SNP densities per 100 Kb region. 1,443 Loss of function (LOF) SNPs, were independently downloaded from the 1000 Genomes Project

ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_07/low_coverage/snps/low_coverage.snps.LOF.txt and Perl scripts were used to calculate densities per 100 Kb region.

2.3.4. SEGMENTAL DUPLICATIONS AND SYNTENY

Segmental duplication data was obtained from UCSC genome browser (Kent et al., 2002) for both human and mouse. UCSC define segmental duplications as genomic regions larger than 1 Kb with at least 90% sequence homology that are present at multiple copies within a genome. Perl scripts were used to calculate overlaps between segmental duplications and the 16,820 orthologous non-divergent and divergent regions. This was also done in the 11,966 human and 7,891 mouse regions, lacking an orthologous mapping. Proportion of duplications in non-divergent compared to divergent structural regions and orthologous compared to non-orthologous regions were carried out in R using chi-squared tests.

Syntenic regions were defined as the regions that form top level whole genome alignments from the Ensembl Compara database, between human (GRCh37/hg19) and mouse (NCBI37/mm9). Lower level alignments were not used

in this investigation. Syntenic blocks were obtained using the perl application programme interface (API) to the Ensembl Compara database (Ensembl 60) (Flicek et al., 2013). Again, Perl scripts were used to define overlaps, a synteny break was defined as a syntenic block starting or ending within a single orthologous 100 Kb region. Proportion of synteny breaks in non-divergent compared to divergent structural regions and orthologous compared to non-orthologous regions were carried out in R using chi-squared tests.

2.4. LINEAR REGRESSION

Multiple linear regression was carried out for all suitable chromatin and sequence features investigated. This was done by collating each feature into a new dataset containing the 16,820 orthologous regions and each feature measure per 100 Kb region. The full list of features entered into the model are detailed in Table 2.1. Where delta (Δ) values were used, the feature entered in the model was the difference in density of that particular chromatin or sequence feature between human and mouse for each orthologous 100 Kb region.

Class	Feature	Included in Δ model?
	GC density	ΔGC
	Gene density	ΔGene density
Sequence divergence estimates	SNP density	
	Indel - deletions	
	Indel - insertions	
	Substitution rate	ΔSubstitution rate
Primary level chromatin features	H2AZ	ΔH2AZ
	H3K27ac	ΔH3K27ac
	H3K27me3	ΔH3K27me3
	H3K36me3	ΔH3K36me3
	H3K4me1	ΔH3K4me1
	H3K4me2	ΔH3K4me2
	H3K4me3	ΔH3K4me3
	H3K9me3	ΔH3K9me3
	NANOG	ΔNANOG
	p300	Δp300
	TAF1	ΔTAF1
	Oct-04	ΔOCT4
	MeDIP	ΔMeDIP
	MRE.seq	ΔMRE.seq
Repeats	DNA	ΔDNA
	LINE	ΔLINE
	LTR	ΔLTR
	Low_complexity	
	SINE	ΔSINE
	Simple	ΔSimple

Table 2.1 List of all chromatin features and sequence level variables incorporated into the multiple linear regression model.

Multiple linear regression was implemented using the glm() package in R and stepwise searches for the best models were performed according to the (generalised) Akaike Information Criterion (AIC). The aim was to find the chromatin variables with the most influence on both higher order chromatin structure and higher order chromatin structural divergence, and also to calculate to what extent chromatin structure/divergence can be explained by the features included. The AIC was used to optimise the model and identify successful combinations of variables. The AIC takes into account redundant or missing data in the model and indicates how well the model fits the data. Standardised r-squared values for the variables in the best models were then calculated using the beta coefficients.

2.5. EPIGENOMIC COMPARISONS

2.5.1. PROCESSING SEQUENCING DATA

A multispecies epigenomic dataset from embryonic and pluripotent stem cells of humans, mice, and pigs was created in a similar manner to Xiao et al 2012. In this study the orthologous genomic distributions of epigenomic features from human, mouse and pig were compared. The features included DNA methylation, from Methylation-sensitive Restriction Enzyme Sequencing (MRE-Seq) and Methylated DNA Immunoprecipitation (MeDIP-seq), which identify DNA methylation genome wide at high resolution. Also high resolution profiles of histone modifications/variants associated with repression (H3K9me3 and H3K27me3), enhancers (H3K4me1, H3K4me2, H3K27ac), promoters (H3K4me3, H2AZ) and gene bodies (H3K36me3) were acquired from chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (ChIP-seq). In addition, ChIP-seq profiles of four transcriptional regulators (NANOG, OCT4, P300, and TAF1) were also obtained. In total 15 different epigenomic features were compared across the three species. The raw data involved were all obtained from previous studies via the NCBI Sequence Read Archive (SRA) (Table 2.2).

Feature	Mark	Method	Human	Accession	Reads processed	% mapped
Repression	H3K27me3	ChiP-seq	Lister et al, 2009	GSE16256	91,076,733	70.87
	H3K9me3	ChiP-seq	Lister et al, 2009	GSE16256	193,583,168	78.64
Enhancer	H3K4me1	ChiP-seq	Lister et al, 2009	GSE16256	69,962,590	88.80
	H3K4me2	ChiP-seq	Lister et al, 2009	GSE16256	38,030,958	73.62
	H3K27ac	ChiP-seq	Lister et al, 2009	GSE16256	42,382,986	81.70
Promoter	H3K4me3	ChiP-seq	Lister et al, 2009	GSE16256	66,500,991	77.25
Gene body	H3K36me3	ChiP-seq	Lister et al, 2009	GSE16256	141,190,603	79.29
Promoter	H2AZ	ChiP-seq	Xiao et al, 2012	GSE36114	25,786,431	95.90
Methylation		MeDIP-seq	Bernstein et al, 2010	GSE16368	35,182,811	87.49
		MRE-seq	Bernstein et al, 2010	GSE16368	77,386,795	75.94
Promoter	TAF1	ChiP-seq	Encode,2012	GSE32465	32,189,724	79.21
Enhancer	P300	ChiP-seq	Encode,2012	GSE32465	53,920,750	68.22
Pluripotency	OCT4	ChiP-seq	Encode,2012	GSE32465	46,880,412	87.78
	Nanog	ChiP-seq	Encode,2012	GSE32465	53,920,750	68.22

Feature	Mark	Method	Mouse	Accession	Reads mapped	% mapped
Repression	H3K27me3	ChiP-seq	Xiao et al, 2012	GSE36114	12,932,668	88.87
	H3K9me3	ChiP-seq	Goren et al, 2010	GSE12241	31,654,344	82.84
Enhancer	H3K4me1	ChiP-seq	Xiao et al, 2012	GSE36114	23,895,406	93.29
	H3K4me2	ChiP-seq	Xiao et al, 2012	GSE36114	19,366,374	90.92
	H3K27ac	ChiP-seq	Xiao et al, 2012	GSE36114	18,351,814	92.78
Promoter	H3K4me3	ChiP-seq	Xiao et al, 2012	GSE36114	6,911,600	92.47
Gene body	H3K36me3	ChiP-seq	Xiao et al, 2012	GSE36114	31,258,590	92.07
Promoter	H2AZ	ChiP-seq	Xiao et al, 2012	GSE36114	17,641,837	88.48
Methylation		MeDIP-seq	Xiao et al, 2012	GSE36114	88,311,720	78.94
		MRE-seq	Xiao et al, 2012	GSE36114	25,728,829	96.16
Promoter	TAF1	ChiP-seq	Xiao et al, 2012	GSE36114	27,736,348	90.15
Enhancer	P300	ChiP-seq	Chen et al, 2008	GSE11431	23,450,889	69.41
Pluripotency	OCT4	ChiP-seq	Chen et al, 2008	GSE11431	17,413,416	72.47
	Nanog	ChiP-seq	Chen et al, 2008	GSE11431	11,785,618	41.77

Feature	Mark	Method	Pig	Accession	Reads processed	% mapped
Repression	H3K27me3	ChiP-seq	Xiao et al, 2012	GSE36114	10,868,857	76.41
	H3K9me3	ChiP-seq	Xiao et al, 2012	GSE36114	13,927,780	69.76
Enhancer	H3K4me1	ChiP-seq	Xiao et al, 2012	GSE36114	20,705,775	79.30
	H3K4me2	ChiP-seq	Xiao et al, 2012	GSE36114	4,322,782	79.50
	H3K27ac	ChiP-seq	Xiao et al, 2012	GSE36114	17,741,434	79.26
Promoter	H3K4me3	ChiP-seq	Xiao et al, 2012	GSE36114	21,512,279	76.91
Gene body	H3K36me3	ChiP-seq	Xiao et al, 2012	GSE36114	38,013,146	76.28
Promoter	H2AZ	ChiP-seq	Xiao et al, 2012	GSE36114	4,417,656	78.26
Methylation		MeDIP-seq	Xiao et al, 2012	GSE36114	68,357,687	68.92
		MRE-seq	Xiao et al, 2012	GSE36114	16,832,623	33.93
Promoter	TAF1	ChiP-seq	Xiao et al, 2012	GSE36114	10,700,034	77.63
Enhancer	P300	ChiP-seq	Xiao et al, 2012	GSE36114	33,455,621	77.98
Pluripotency	OCT4	ChiP-seq	Xiao et al, 2012	GSE36114	5,074,592	77.59
	Nanog	ChiP-seq	Xiao et al, 2012	GSE36114	18,575,267	77.15

Table 2.2 Chromatin features used in the three-way analysis of human, mouse and pig. The table shows values associated with the primary processing of the ChIP-seq data including numbers of reads mapped and percentages aligned.

ChIP-seq, MRE-seq and MeDIP-seq were mapped to genome assemblies GRCh37/hg19, NCBI37/mm9, and SGSC Sscrofa9.2/susScr2 using Bowtie (Langmead et al., 2009). This was incorporated into a Perl pipeline, which unpacked the files into FASTQ format, which contains DNA sequence and a corresponding quality score. The FastQC tool (Babraham Bioinformatics, 2010) was used to assess the quality of the mapped reads. The human binding site data for TAF1, P300, OCT4 and NANOG (Xiao et al., 2012) were found to have low FASTQC quality reports and a poor percentage of mapped reads (48.70, 47.24, 20.38 and 45.94 respectively). Alternative data was found from ENCODE (The ENCODE Project Consortium, 2011) that had much higher percentages of mapped reads (79.21, 68.22, 87.78, 68.22). Once the reads were aligned, SAMTools (http://samtools.sourceforge.net) was used to index and convert FASTQ files into bam and the bigwig format. The data was converted into bedgraph files for analysis of density per 100 Kb using UCSC tool bigwigtobedgraph.

2.6. SOFTWARE, ONLINE RESOURCES AND DATASETS

2.6.1. PROGRAMMING LANGUAGES AND PACKAGES

The R Project for Statistical Computing (http://www.r-project.org/) is a publicly available language and software environment for statistical calculations and graphics. Originally developed in 1996 (Ihaka and Gentlemen, 1996) to be a portable, efficient language for statistical analyses, it has now evolved to carry out high performance computing for handling complex large datasets (R Core Team, 2013). The standard R functions in conjunction with extension packages have been used for the majority of statistical analysis used in this project. These include linear regression, statistical tests, hierarchical clustering, data handling and normalisation. It has also been used to present the data graphically using various standard and extension packages.

The R extension packages used are detailed below.

- gplots is a R package containing specialised tools for plotting data. The function heatmap.2 was mainly used from this package to plot false colour diagrams for correlation matrices (Warnes et al.,

2010).

- Bioconductor; a free open development software project for the bioinformatic analysis of genomic data in R (Gentleman et al., 2004). It contains a number of R add-on packages specifically used for the analysis of microarray data. The packages used for this investigation are further detailed.
- pvclust is used for performing hierarchical clustering with accompanying p-values (Suzuki and Shimodaira, 2006). This was used for the hierarchical clustering of the 1719 divergent chromatin regions by divergence class.
- Limma, part of the Bioconductor toolset, is specialised for the analysis, linear modelling and differential expression of microarray data. In this instance, it was mainly for quantile normalisation across differing datasets (Smyth, 2005).
- Samr is used for performing SAM analysis to define divergent regions on 100 Kb orthologous dataset (Tibshirani et al., 2011).
- MASS is used to perform stepwise model selection by the Akaike Information Criterion (AIC) which measured the goodness of fit for linear models (Venables and Ripley, 2002).

The Perl programming language (www.perl.org/) is a general purpose programming language popular for bioinformatic analyses. Perl, Version 5.10 (2007), was used throughout this project to quickly collate, and parse large datasets for further analysis, usually in R. Perl pipelines were also used for the ChIP-seq analysis as mentioned above.

2.6.2. ONLINE TOOLS AND RESOURCES

The 1000 Genomes Project

The 1000 Genomes Project (www.1000genomes.org) is an international consortium set up to discover human genetic variants by aiming to sequence many individuals using next generation sequence techniques (Abecasis et al., 2012). The goal of the project is to find most genetic variants that have frequencies of at least 1% in the populations studied by sequencing many individuals. SNP data was used from the 1000 Genomes Project to analyse SNP densities in divergent and non-

divergent regions. This was done by using VCFTools (Danecek et al., 2011) to find densities of SNPs per 100 Kb region. 1443 loss of function SNPs were also downloaded from the 1000 Genomes Project.

Database for Annotation, Visualization and Integrated Discovery (DAVID)

The Database for Annotation, Visualization and Integrated Discovery (DAVID) is a web based bioinformatic resource developed by the Laboratory of Immunopathogenesis and Bioinformatics (http://david.abcc.ncifcrf.gov) (Dennis et al., 2003). It was designed to provide functional analysis of genome-scale datasets derived from high-throughput methods. For the purposes of this investigation, the DAVID functional annotation tool (v6.7) was used for gene-annotation enrichment analysis.

Galaxy

Galaxy is an open, web-based platform for biomedical analysis developed by Emory University and Penn State University (Goecks et al., 2010) (http://galaxy.psu.edu). Although Galaxy incorporates a variety of inbuilt tools for genomic analysis, the main ones used in this investigation were simple tools for genomic data manipulation. These include extraction of MAF blocks from MultiZ 46 way alignment blocks for indel analysis, restricted to intergenic regions (outside Ensembl gene models).

Bowtie

Bowtie is an open source, memory-efficient short read aligner designed for rapid alignment of large sets of short DNA sequences (reads) to large genomes (Langmead et al., 2009). It was used to do the primary processing of the three-way primary chromatin data from Xiao et al (2012). Bowtie was used to do the read mapping for the ChIP-seq histone modification data in all three species; human, mouse and pig.

SAMtools

The SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. It is easily converted to other storage formats

and easily manipulated for further analysis. SAMtools (http://samtools.sourceforge.net) is a library and software package for parsing and manipulating alignments in the SAM format. This library can convert from other alignment formats, sort and merge alignments, call SNPs and show alignments in a text-based viewer. SAMtools was used to index and convert FASTQ files into the bam format.

BEDtools

BEDtools is a suite of utilities that enable genomics tasks such as intersecting, merging and shuffling of genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF and VCF (Quinlan and Hall, 2010). It also allows for converting between different genomic file formats. The mergeBed tool was used for merging histone modification data from multiple sources. It was also used to convert bam formatted files to bed files.

FastQC

FastQC is a quality control tool for high throughput sequence data (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Babraham Bioinformatics, 2010). It reads sequence data in a variety of formats and can either provide an interactive application to review the results of several different QC checks, or create an HTML based report. It was used to independently verify the quality of the mapped reads for the chromatin feature data from Xiao et al (2012).

Sequence Read Archive (SRA)

The Sequence Read Archive (SRA) is a bioinformatics database that provides a public repository for DNA sequencing data produced from published studies. It is particularly used for short read sequences generated by High-throughput sequencing, which are typically less than 1000 base pairs in length. The SRA was the source of the ChIP-seq datasets used in Xiao et al (2012).

The Encyclopedia of DNA Elements (ENCODE)

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups. The goal of ENCODE is to build a comprehensive map of all the functional and regulatory elements in the human

genome and present the data in public databases. The pilot phase, carried out in 2007, analysed functional elements in a portion of the genome equal to about 1% (Birney et al., 2007). Since then the project has been extended and currently over 1,000 genome-wide data sets have been produced by the ENCODE project (Dunham et al., 2012). ENCODE ChIP-seq data relating to transcription factor binding sites were used in this investigation for the three-way species analysis.

Ensembl

Ensembl is a joint project between European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI) to develop a system, which produces and maintains automatic genomic annotation on selected eukaryotic genomes. The Ensembl Project produces detailed genome data for vertebrates and other eukaryotic species available online (http://www.ensembl.org) (Flicek et al., 2013). Ensembl BioMart is a highly customisable data mining tool used for downloading specific genome data. Ensembl (Release 64) was used to obtain accurate gene models and IDs for the human and mouse genomes. Top level synteny breaks were also obtained using the Perl Application Programme Interfaces (API) to the Ensembl Compara database.

UCSC Genome Bioinformatics Site

The UCSC Genome Bioinformatics Site is an online resource that contains a vast number of species' genome assemblies and annotations (http://genome.ucsc.edu) (Kent et al., 2002). It also has a library of tools for viewing and manipulating different genomes. Segmental duplication tracks were used from UCSC. Human-mouse substitution rates were determined through the use of pairwise human-mouse alignment tracks from UCSC. The web resource also provided various genomic manipulation tools such as LiftOver (discussed below) and some conversion tools. The program bigWigToBedGraph was used to convert bigWig files to ASCII bedGraph format for formatting histone modification datasets.

UCSC LiftOver

UCSC LiftOver (Kent et al., 2002) is an online utility and executable script that is part of the UCSC website. Its function is to convert genome coordinates between different assemblies with the appropriate UCSC generated chained pairwise-alignment files. Conversions are possible within species (i.e mm8 to mm9)

and between species (i.e mm9 to hg19) using whole genome sequence alignments.

Phylogenetic Analysis by Maximum Likelihood (PAML)

Phylogenetic Analysis by Maximum Likelihood (PAML) is a package for phylogenetic analyses of DNA using maximum likelihood (abacus.gene.ucl.ac.uk/software/paml.html), currently in version 4 (Yang, 2007). The program Baseml using the REV model, was used for the substitution rate analysis between human and mouse intergenic regions.

VCFtools

VCFtools is a package specifically designed for working with data from the 1000 Genomes Project (Abecasis et al., 2012). Variant Call Format (VCF) files have been developed with the advent of large-scale genotyping and are used in bioinformatics for storing gene sequence variations and SNP data. For these purposes the SNPdensity option was used to calculate the number and density of SNPs in 100 Kb bins from SNP variant call files (Danecek et al., 2011).

Chapter 3

Results: Discovery of divergent higher order chromatin structure

Topics included in this section:

- Summary of the data and methods that were involved in producing RT, LA and Hi-C data.
- Creation of a comprehensive orthologous 100 Kb higher order structure dataset across all human and mouse cell and data types.
- Assessment of the degree of conservation across all orthologous 100 Kb regions.
- Development of methodology involved in divergence metrics.
- Estimation of the degree of structural divergence between cell types and species.

3.1. INTRODUCTION

The aim of this chapter is to gain insights into the conservation and divergence of various aspects of higher order chromatin structure across human and mouse cell types. There has been a recent influx of studies that provide complementary views of higher order chromatin structure. Each shows that the mammalian genome is organised into large, discrete domains of higher order chromatin with opposing properties (levels of expression and accessibility, spatial positioning, and replication timing). These domains appear to be broadly similar across the different cells that have been examined, although many regions across the genome show cell type specific structure (Lieberman-Aiden et al., 2009, Hiratani et al., 2010, Peric-Hupkes et al., 2010). However, the actual extent to which these datasets intersect, and how they relate to one another across cell types and species, is poorly understood. Similarly, the genomic loci underlying divergence in chromatin structure between species, and the mechanisms underlying divergence, are unknown. Until recently, this type of analysis has been limited by a lack of genome-wide data for higher order chromatin structure.

In this section, a large number of diverse mouse and human datasets are collated to provide the most comprehensive overview of higher order chromatin structure in mammals to date. A systematic study of all orthologous regions in the mammalian genome is undertaken to evaluate the extent of conservation in higher order chromatin structure between cell types and during evolution.

3.2. STRUCTURAL DATA TYPES

A higher order structural orthologous dataset was collated using previously published data from five different studies. These studies generated 36 different datasets across different (though predominantly embryonic) cell types in human and mouse: replication timing (RT) (Ryba et al., 2010, Hiratani et al., 2010), nuclear lamina association (LA) (Guelen et al., 2008, Peric-Hupkes et al., 2010) and genome-wide inter-locus contact preferences (Hi-C) (Lieberman-Aiden et al., 2009). Thus, throughout this thesis higher order structure is seen in terms of these three data types: RT, LA and Hi-C. The methodology used to produce these three different data types is summarised briefly:

RT data

Embryonic, epiblast, induced pluripotent stem cells and pluripotent stem cells were labelled with Bromodeoxyuridine (BrdU). The cells were then separated into early and late S-phase fractions by flow cytometry. BrdU-labelled DNA from these cells was then immunoprecipitated, differentially labelled, and cohybridized to a human/mouse whole-genome oligonucleotide microarray (NimbleGen). This generated a replication-timing ratio (log2 (Early/Late)) for each of the tiled probes, which were positioned at every 5.8 kilobases (Kb) in mouse and 1.1 Kb in human. These replication timing ratios were then normalised and scaled to an equivalent median-absolute deviation (Hiratani et al., 2008) (Figure 3.1).

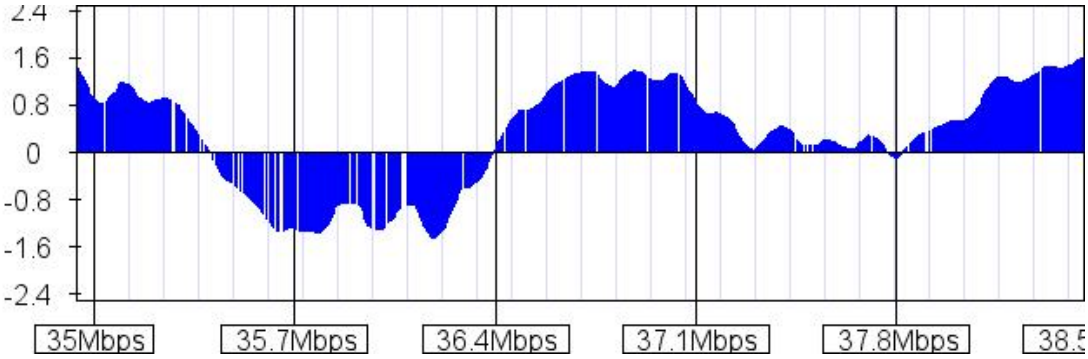

Figure 3.1 Example of replication timing domains across 3 Mb section of mouse chromosome 6 (x axis). Y axis values represent log2 (Early/Late) replication timing values. Image courtesy of www.replicationdomain.org (Weddington et al., 2008).

LA data

Genome-wide maps of nuclear lamina interactions were made for both human and mouse cell types using DamID of Lamin B1. In this method, a fusion protein is created to target genomic regions at the nuclear periphery. This is comprised of Lamin B1, which is part of the protein structure of the nuclear lamina at the periphery of the nucleus, and E. coli DNA adenine methyltransferase (Dam). The fusion protein was expressed in cultured cells and immunofluorescence and confocal microscopy confirmed that the DAM-tagged lamin B1 appears to be incorporated into the nuclear lamina. The Dam adenine-methylates GATC DNA sequences that are in close contact with the nuclear lamina. Methylated GATC was then cut by DpnI restriction endonucleases and then ligated to known sequences, which can then be amplified and hybridized to oligonucleotide microarrays. This leads to the specific selection of genomic fragments flanked by methylated GATCs,

which will be at the nuclear periphery (Vogel et al., 2007). In human lung fibroblasts, the median probe spacing was about 200 base pairs and in mouse, median intervals of 1.2 kb. In both these hybridizations, methylated DNA fragments from cells expressing un-fused Dam, which is present throughout the nucleus, were used as a reference. The relative contact frequency of each probed sequence to the nuclear lamina was taken as log2(DamID/Dam only). The data was quantile normalised across all cell types (Guelen et al., 2008).

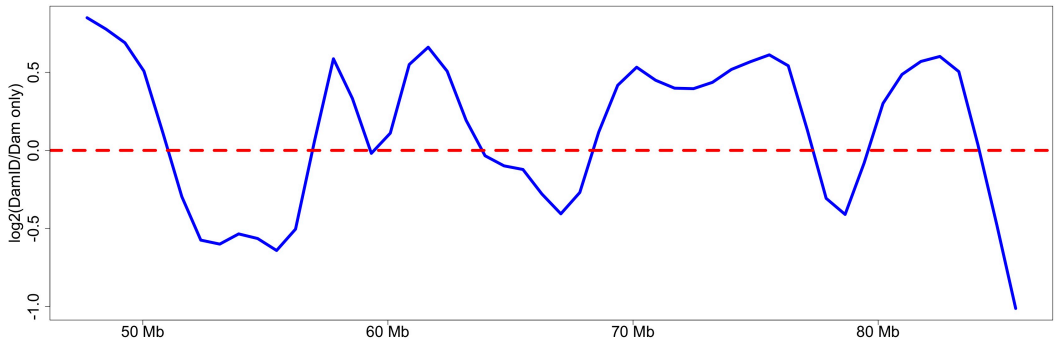

Figure 3.2 Example of smoothed lamin association data across a 30 Mb section of human chromosome 1. Y-axis values represent log2(DamID/Dam only), x-axis values are chromosomal coordinates. Data taken from Guelen et al (2008).

Hi-C data

Finally, the Hi-C method involved cross-linking lymphoblastoid cells with formaldehyde, which forms covalent links between spatially adjacent chromatin segments. The chromatin was then digested with a restriction enzyme and the ends were biotinylated and ligated. Biotinylated junctions were isolated and identified by paired-end sequencing. This provided a genome-wide contact matrix whereby the genome was divided into 1 Mb and 100 Kb regions and the matrix entry for each region defined as the number of ligation products between locus a and locus b (Figure 3.3).

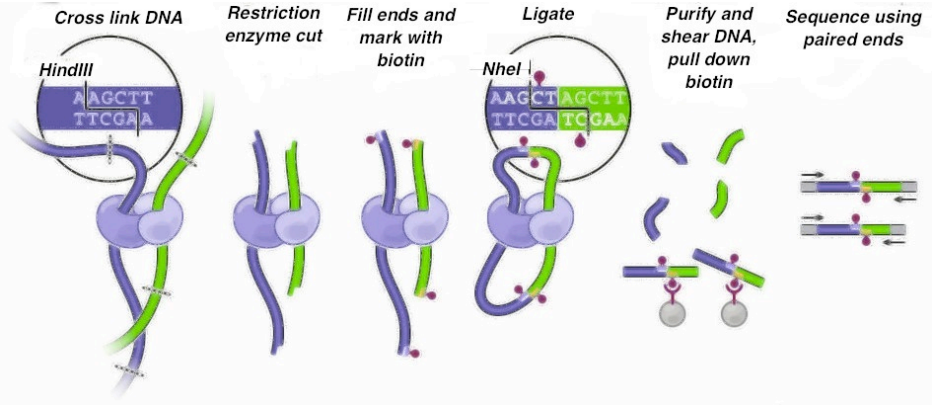

Figure 3.3 Diagrammatic overview of the Hi-C method. Adapted from (Lieberman-Aiden et al., 2009).

Most signal in these data has been found to reflect interacting loci on the same chromosome, thus the main interest was in intrachromosomal contact probabilities, which were derived from the contact matrix (Figure 3.4). These revealed two large areas in the contact matrix, where interactions within each area were enriched but enrichments between them were depleted. These 'compartments' were found to correspond to relatively active, open and relatively inactive, closed chromatin. The two compartments were found to be well defined by the eigenvectors of the contact probability matrices (Lieberman-Aiden et al., 2009) (Figure 3.4).

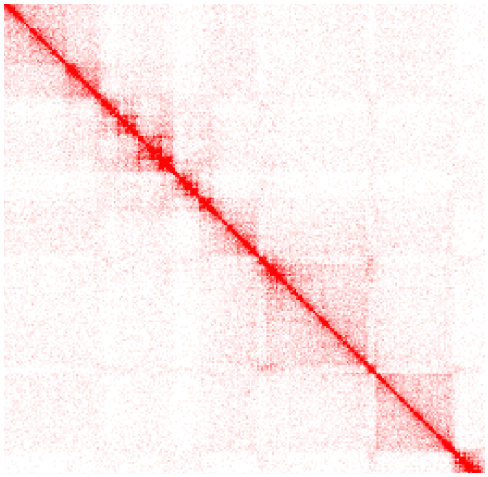

Figure 3.4 Example of Hi-C interaction matrix data. The matrix illustrates the interaction frequencies between the intrachromosomal interaction profiles of every pair of 100 Kb loci along a section of human chromosome 14. Image courtesy of The Hi-C Data Browser http://hic.umassmed.edu/.

Other data considered

Hansen et al, 2009 - A human replication timing dataset produced in human ESC, fibroblast and lymphoblastoid cell lines. A new technique - Repli-Seq is used in which BrdU labelled DNA is sorted into 6 fractions – G1, S1, S2, S3, S4, G2. Massively parallel sequencing is then used to get sequence reads, which are converted, into percent-normalised density values for a region of DNA per cell cycle phase (Hansen et al., 2010). This data was not included in this investigation due to the availability of new human replication timing data produced by the same method as the mouse replication timing data (Ryba et al., 2010).

Kalhor et al, 2012 - Re-analysis of the Hi-C method revealed the presence of systematic biases that obscure additional, finer scale structural compartments (Yaffe and Tanay, 2011). These biases include restriction enzyme cutting frequencies, GC content and the uniqueness of the DNA sequence. A high degree of Hi-C interaction frequencies are reported near restriction sites because of size selection, similarly a higher GC density near the restriction site can disrupt accurate mapping as can the uniqueness of fragment ends. There is no doubt that these biases can cause problems in interpretating of Hi-C pairwise interaction probabilities. In this investigation the focus is on eigenvectors summarising the two-compartment division between open and closed regions, rather than interaction probabilities

themselves. However, it was possible that the eigenvectors might be affected by these same biases. Consequently, an independent interaction probability map was examined which was produced for a similar lymphoblastoid cell line, using a modified Hi-C method designed to mitigate the biases inherent in previous data (Kalhor et al., 2012). When the original (Lieberman-Aiden et al., 2009) interaction data were substituted with the new, nominally unbiased data, very similar correlations with all other chromatin structure datasets were observed. It was concluded that any biases present in the Lieberman-Aiden et al (2009) dataset have little effect on a course grained, two compartment classification of the genome based upon eigenvectors, and therefore the original Lieberman-Aiden (2009) dataset was used.

The final data chosen to include in this study (detailed in Table 3.1) included 13 human and 23 mouse higher order chromatin datasets.

3.2.1. CELL TYPES

The datasets used in this investigation comprised of experimental data from a variety of different cell types. These include embryonic stem cells (ESC), epiblast derived stem cells (EpiSC), induced pluripotent stem cells (iPSC), neural progenitor cells (NPCs), fibroblasts and lymphoblasts, which are detailed briefly (Table 3.1). ESC cells are pluripotent stem cells that are able to differentiate when undergoing development. Replication timing in human and mouse and lamin association in mouse have been established across ESC cells. iPS cells are cells in which an adult differentiated cell has been reverted back to the pluripotent state, these have replication timing profiles across both human and mouse cells. Comparison of replication timing profiles between ESC and iPSC cells in mouse found that the replication timing profiles were virtually unchanged showing that replication-timing profiles in ESCs could provide a unique signature for the pluripotent state (Hiratani et al., 2008). However, replication profiles have been shown to alter more dramatically between mouse ESCs and pluripotent EpiSCs, which are only a few days older, a difference that does not correlate with changes in expression levels (Hiratani et al., 2010). In both species, replication timing profiles in ESCs was shown to undergo alteration upon differentiation to NPCs and in lymphoblast cells, which are antigen specific cells within the lymphoid tissue. Up to 45% of the mouse genome has been shown to have significant changes in replication timing during development highlighting the cell type specific nature of replication timing

(Hiratani et al., 2010). When analysing the replication timing maps of human ESCs it was found that they more closely correlate with mouse EpiSCs than ESCs, suggesting human ESCs are stabilized in a more epiblast-like epigenetic conformation (Ryba et al., 2010). This indicates that replication timing maps are a reliable indicator of the chromatin environments of the cell.

Due to the variability of chromatin structure across al the different cell types examined, robust normalisation was needed to ensure comparisons across all datasets was appropriate. This is explored further in section 3.3.2.

Author	Data Type	Species	Cell Type
Hiratani et al. 2009	Replication Timing Log2(Early/Late)	Mouse	46C-ESC, D3-ESC, TT2-ESC, iPSC D3-EPL, D3-EMB3 EpiSC5, EpiSC7 D3-EBM6, 46-NPC, TT2-NPC, D3-EMB9 Mesoderm Endoderm MEFF, MEFM, MyoBlast
Ryba et al. 2010	Replication Timing Log2(Early/Late)	Human	BG01-ESC, BG02-ESC, H7, iPSC4, iPSC5 BGO2-NPC Lymphoblast (C0202)
Peric-Hupkes et al. 2010	Lamin Association Log2(Lamin Associating/Input)	Mouse	ESC NPC Astrocytes (AC) MEF
Guelen et al. 2008	Lamin Association Log2(Lamin Associating/Input)	Human	Tig3 Human embryonic lung fibroblasts
Lieberman-Aiden et al. 2009	Eigenvector Intra-chromosomal contact probability	Human	Lymphoblast (GM06990)

Table 3.1 Details of the individual studies, cell lines and data types used.

3.3. CONSTRUCTION OF A HIGHER ORDER CHROMATIN STRUCTURE DATASET

3.3.1. SCALING UP PROBE-BASED DATA

Higher order domain structures had previously been defined in replication timing and lamin association data (Hiratani et al., 2008, Peric-Hupkes et al., 2010). For the RT data, a segmentation algorithm was used to define co-ordinately early or late replicating regions in ESC and NPC cells. This method defined domains ranging from 200 Kb to 2 Mb in size (Hiratani et al, 2010). For the LA data, a sliding

window approach with a permutation strategy was used, which yielded similar sized Lamin Associating Domains (LADs) - 80 Kb to 30 Mb (Peric-Hupkes et al., 2010). To investigate appropriate scale to examine higher order structure in this investigation, the distributions of replication timing domains and LADs were examined (Figure 3.5).

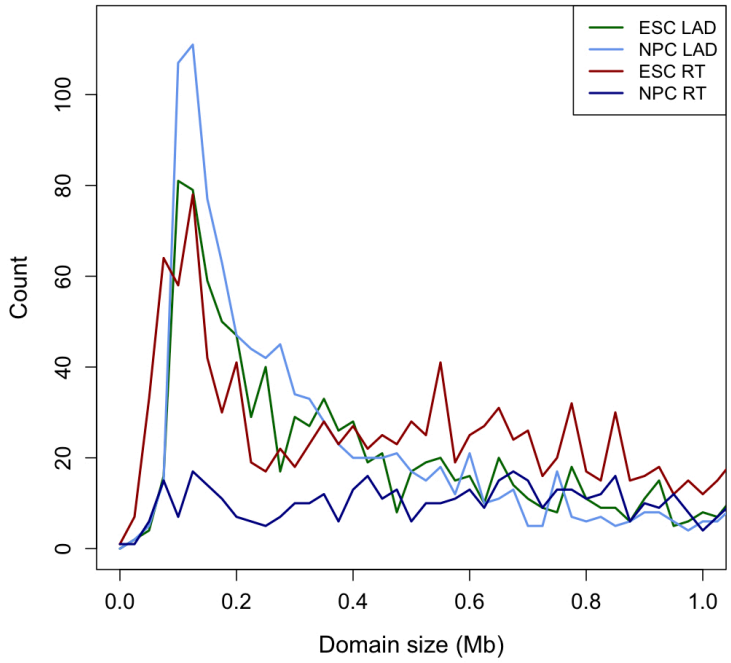

Figure 3.5 Size distribution of previously defined replication and lamin association domains (LADs). Domain sizes range from 30 Kb to 30 Mb (domains up to 1 Mb shown). ESC and NPC LAD (Peric-Hupkes et al., 2010), ESC and NPC RT domains (Hiratani et al., 2010).

It is clear from Figure 3.5 that very few domains are below 100 Kb in size with many being much larger. For this reason, a region size of 100 Kb was chosen as the scale up to examine the probe-based datasets. In addition to this, the Hi-C dataset was produced at 100 Kb resolution and could be integrated more easily at this resolution. Firstly, the probe-based data was converted to the latest human or mouse assembly coordinates (i.e. from mm8 to mm9 for mouse RT data) using the UCSC liftOver utility (See Chapter 2 Methodology). Custom perl scripts were then used to average the structural data values into consecutive non-overlapping 100 Kb regions. Regions represented by fewer than 10 probes were discarded as potentially

unreliable. This was done for each dataset and resulted in 13 human (GRCh37/hg19) and 23 mouse (NCBI37/mm9) comparable 100 Kb resolution datasets.

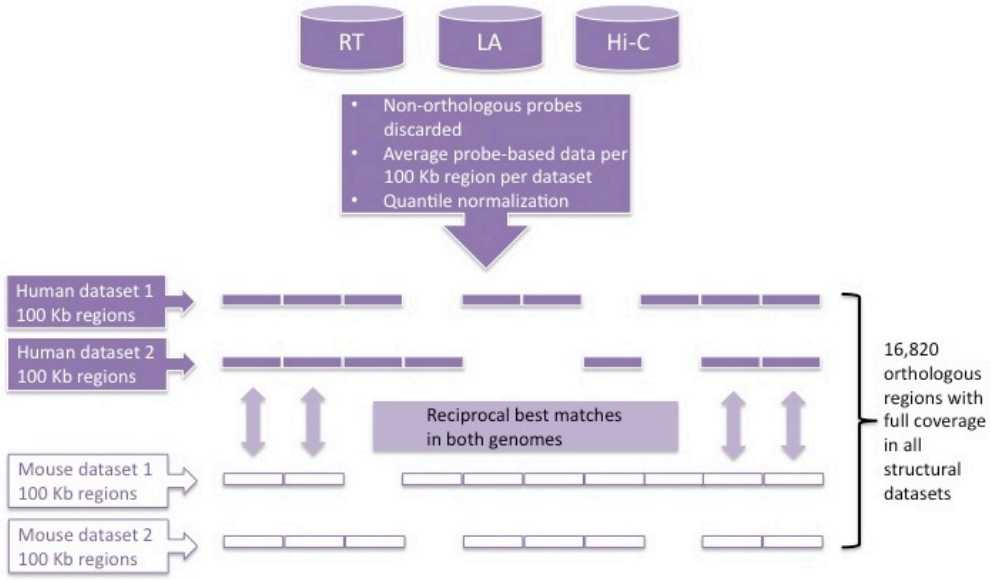

Figure 3.6 Overview of methodology. Replication timing, lamin association and Hi-C data from 36 datasets are converted to consistent genome assemblies (GRCh37/hg19 and NCBI37/mm9), averaged into 100 Kb regions and collated into 16,820 orthologous regions represented in all structural datasets.

The result was a set of 24,711 mouse and 28,786 human 100 Kb regions represented by higher order structural values from multiple datasets. To collate the human and mouse datasets together, firstly the coordinates for the human datasets were converted to the orthologous coordinates in the latest mouse assembly (GRCh37/hg19 to NCBI37/mm9) and vice versa for the mouse datasets using UCSC whole genome alignments (Kent et al., 2002). This involved pairing the USCS liftOver utility with Perl scripts to conservatively ensure correct mapping between species. Regions that did not correctly map reciprocally (i.e. forwards and backwards) between species, or that substantially changed in size (<80% or >120% of the original region) when re-mapped were discarded. A complete orthologous dataset was made in both species (i.e. one based upon human and one based upon mouse coordinates), with a 50% coordinate overlap used to define orthologous regions. The level of overlap does necessarily vary somewhat between orthologous regions, and it was a concern that this might later influence the measurement of

structural divergence. Specifically it was important to show that the regions later identified as structurally divergent were not simply those most poorly aligned (i.e. close to the 50% minimum overlap) between species at the sequence level. On closer examination, the distributions of overlaps (aligned nucleotides minus gaps) were found to be very similar between structurally divergent and non-divergent regions, whether viewed in terms of human (GRCh37/hg19) genome (divergent overlap mean = 0.80, median = 0.81; non-divergent overlap mean = 0.79, median = 0.80), or mouse (NCBI37/mm9) genome (divergent overlap mean = 0.73, median = 0.72; non-divergent overlap mean = 0.72, median = 0.71) coordinates. It was concluded that the estimates of structural divergence are not a simple reflection of sequence divergence.

The final orthologous dataset was defined as the 100 Kb regions that were successfully collated in both species and amounted to a total of 16,820 100kb orthologous regions, covering around 56% of the human genome (Figure 3.6). A total of 11,966 human and 7,891 mouse regions, lacking an orthologous mapping using this protocol, were designated putatively lineage specific regions.

3.3.2. NORMALISATION ACROSS DATASETS

The distributions of the datasets were examined to ensure direct comparisons and therefore global normalisation was appropriate. Notably, each individual dataset within the orthologous structural regions showed bimodal distributions (Figure 3.7). These values were found to reflect the relatively open and closed nuclear compartments of higher order chromatin, consistent with previous observations (Gilbert et al., 2004, Lieberman-Aiden et al., 2009, Hiratani et al., 2008, Hiratani et al., 2010, Peric-Hupkes et al., 2010).

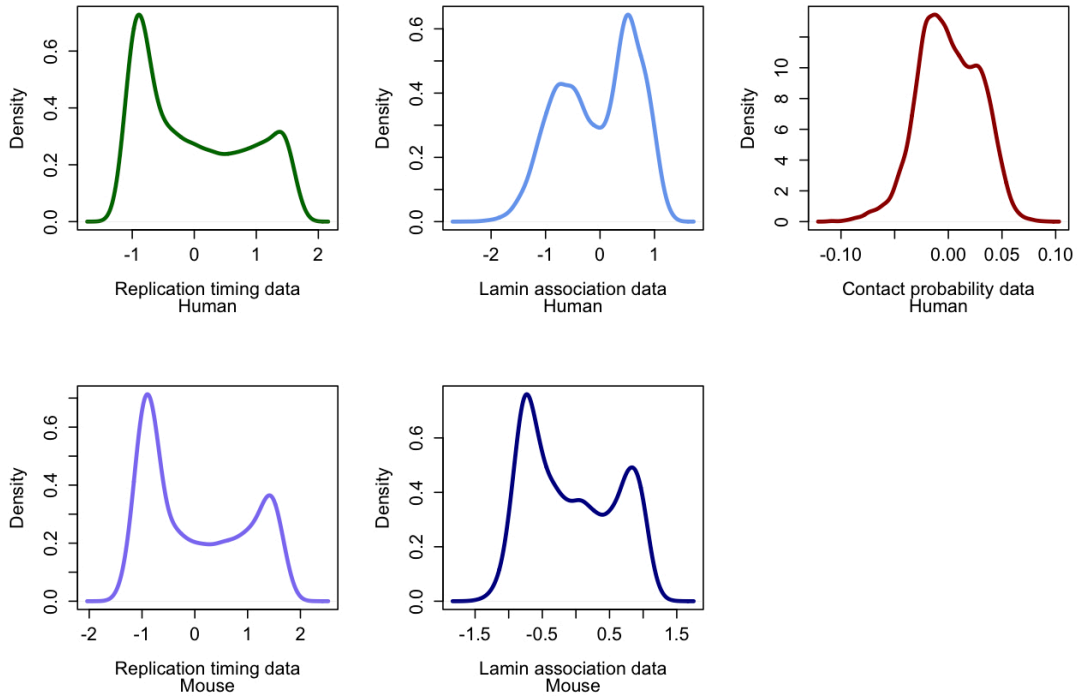

Figure 3.7 Structural data distributions. The bimodal distributions of higher order structural data before normalisation indicating two distinct populations of higher order structure across the mammalian genome. Human and mouse RT data, LA data, and human Hi-C data are shown.

The data from the Hi-C interaction method showed the weakest bimodality, with open and closed chromatin less distinctly segregated. Given the diversity of samples and methodologies across these datasets, strict normalisation must be imposed before directly comparing their features. As all the data show a degree of bimodality, global normalisation was deemed to be appropriate. Three different normalisation techniques were investigated to ensure the correct method for the data was chosen (Figure 3.8).

1. Scaled normalisation, where each dataset is centred and scaled to have a mean of 0 and a standard deviation of 1.

$$x_{\text{norm}} = \frac{x - \mu}{\sigma}$$

2. Minimum-maximum normalization, where each dataset is transformed so that each value in a dataset has its minimum value subtracted and is

divided by the range of values in the dataset. This gives uniform minimum and maximum values across all datasets of 0 and 1.

$$x_{\text{norm}} = \frac{x - min_x}{max_x - min_x}$$

3. Quantile normalisation is a common method for normalising microarray data. This method involves making two or more datasets statistically identical by having the same empirical distribution. This involves sorting the data values in order and then averaging across each ordered value. So the highest value in all cases becomes the mean of the highest values, the second highest value becomes the mean of the second highest values, etc…

After examination of the performance of these normalisation techniques (Figure 3.8), quantile normalisation was chosen as the most appropriate normalisation technique and implemented across all structural datasets for all 100 Kb regions. Transforming all the datasets so that they cover the same range of values, as in quantile normalisation, would best nullify any experimental biases present within the data and allow for more accurate comparisons across datasets.

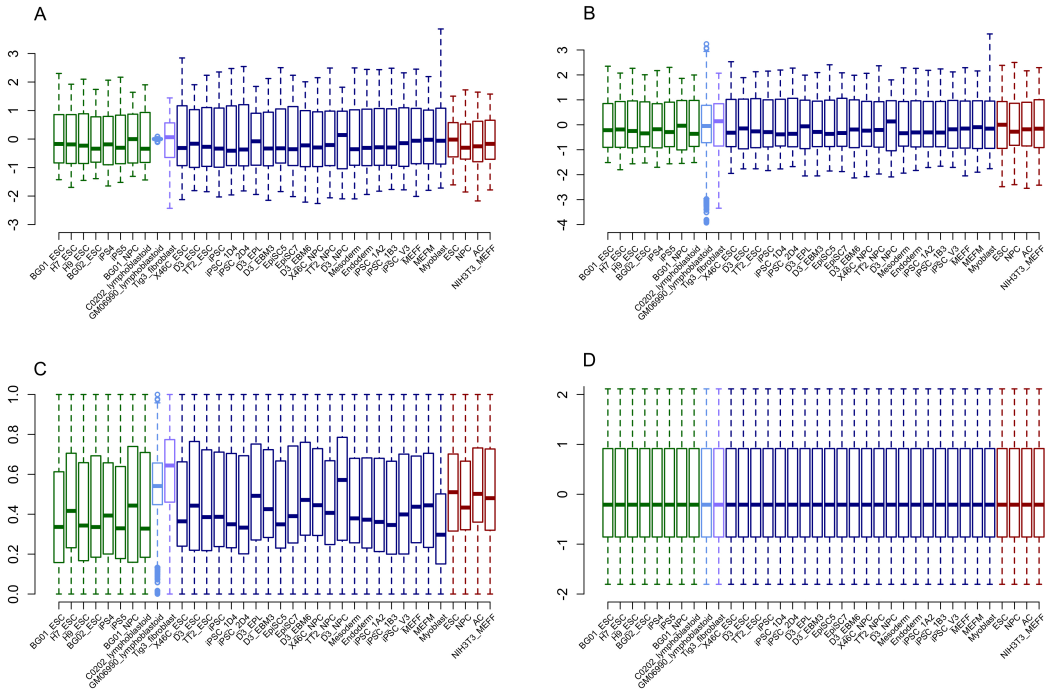

Figure 3.8 Normalisation techniques examined for appropriate scaling across all datasets. Boxplots represent the distributions of each dataset with A) No normalisation B) Scale normalisation, C) Min-max normalisation and D) Quantile normalisation. Different datasets are represented by different colours, mouse lamin interactions (green), human lamin interactions (light blue), human Hi-C (purple), mouse replication timing (purple) and human replication timing (red).

3.4. CONSERVATION AND DIVERGENCE OF HIGHER ORDER CHROMATIN STRUCTURE

3.4.1. WIDESPREAD CONSERVATION OF HIGHER ORDER CHROMATIN STRUCTURE

We initially sought to answer two related questions. Firstly, how well do these diverse datasets agree quantitatively? And secondly, what fraction of the mammalian genome can confidently be identified as structurally divergent? Similarities were expected between RT, LA and Hi-C datasets as they reflect somewhat overlapping aspects of higher order chromatin structure, but the precise extent of the correlations between them was unknown. A Spearman's Rho correlation matrix across all 36 available datasets showed that the degree of agreement is indeed strong and significant across all datasets (Rho: 0.38 to 0.98, $p <$

2.2 x 10-16), in spite of differing experimental procedures, platforms, cell types, and species. The highest agreement was observed between similar cell types from the same species, even across experimental platforms. For instance mouse RT data for a variety of ES and induced pluripotent stem cell (iPSC) types showed strong correlations (Rho: 0.7-0.9, $p < 2.2 \times 10\text{-}16$) with lamin data from mouse ES cells, and together they form a coherent cluster in the correlation matrix (Figure 3.9).

However, there are also interesting exceptions to this rule, such as the human embryonic fibroblast LA data. Although this dataset showed the weakest correlations to all other datasets, the best agreement was to the mouse fibroblast LA and RT data and not to other human cell types. The reason for this may lie in cell cycle variation: ES and iPS data may be strongly influenced by the fact that these cells are almost entirely in S phase, whereas fibroblasts divide slowly and are mainly in G0/G1. In any case it seemed that certain aspects of higher order structure in particular cell types, such as association with the nuclear periphery in fibroblasts, have been more strongly conserved than others during evolution.

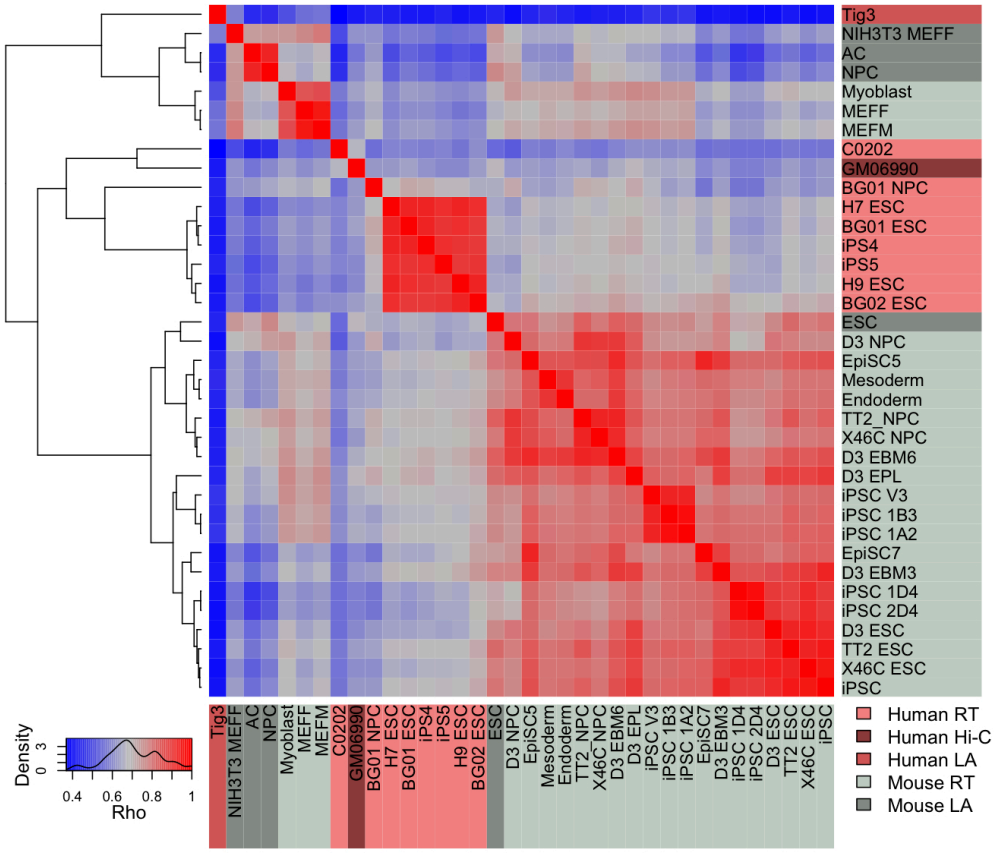

Figure 3.9 Global correlation matrix of higher order chromatin datasets. The heatmap and dendrogram show the relationships among 36 chromatin structure datasets (Spearman's rho: 0.38 to 0.98, $p < 2.2 \times 10^{-16}$). Datasets are labelled according to the experimental platform and species of origin: light grey = mouse RT, light pink = human RT, dark grey = mouse LA, medium pink = human LA, dark pink= human Hi-C.

Striking evidence of structural conservation across the mammalian genome was evident at the level of genome wide correlations (Figure 3.9). This suggests that many aspects of higher order chromatin structure have been conserved in embryonic cell types, over the ~80 million years since the divergence of rodents and primates. However, apparent divergence in higher order chromatin structure between species was also evident in specific regions. This was most simply seen as loci demonstrating a strong, consistent difference in mean normalised structure between the two species across all of the available datasets (see representative regions depicted in Figure 3.10. Although there are high correlations between many of these datasets, reflecting similar overall trends in structure, this can mask substantial variation between datasets at the level of the absolute normalised structural values for a given 100 Kb region (Figure 3.10). The critical question is therefore, which 100 Kb regions vary between species to an unexpected degree, given the extent of variation seen among all datasets? This question is addressed below using a novel divergence metric based upon permutations of the original data.

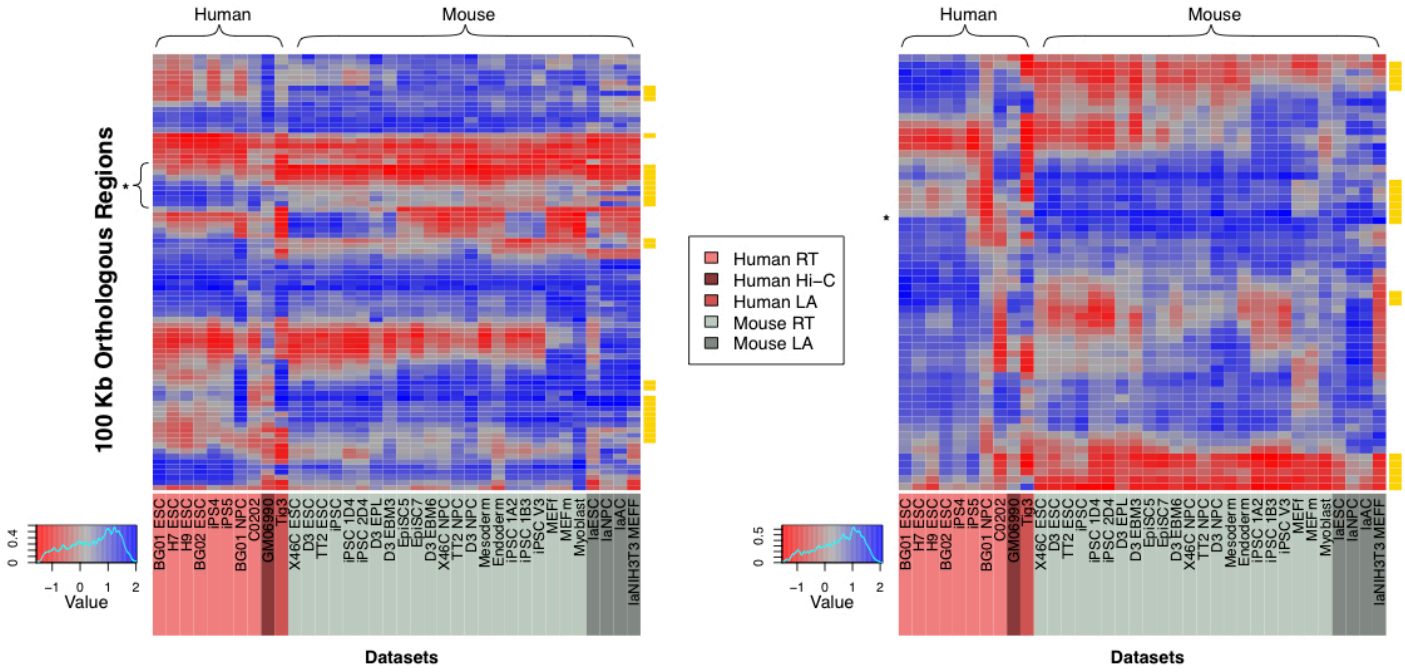

Figure 3.10 Specific human and mouse regions show significant divergence in higher-order chromatin structure. Human (pink) and mouse (grey) higher order chromatin structure across all cell types assayed, shown for two regions of the human genome: chromosome 11p15.2-15.4 (1.2-15 Mb) with the location of an OR gene cluster indicated by an asterisk (A); chromosome 7p14.3-15.3 (24-32 Mb) with the location of the HOXA gene cluster indicated by an asterisk (B). Consecutive, orthologous 100kb regions are positioned on the y-axis with heatmap colours representing relatively open (blue) and closed (red) chromatin structures. Regions displaying significantly divergent chromatin structure are highlighted in yellow.

3.4.2. DEVELOPING THE DIVERGENCE METRIC

Genomic regions were sought that showed strong and consistent structural divergence between species, across all cell types. Several techniques for identifying divergent regions were investigated:

1. Means difference

Simply subtracting the mean of the mouse chromatin values from the human mean would give a means difference distribution with the most divergent structural regions occupying the extreme tails of the distribution. The most consistent or non-divergent regions would occupy the centre of the distribution with means difference closest to 0. From this distribution the divergent regions could be selected using a quantile cut offs of 0.95 and 0.05, thereby choosing 10% of the data that is the most divergent. One caveat of this method is the fact that averaging the data in this way will be insensitive to subtle differences in chromatin structure across datasets. On the other hand an extreme value in one particular dataset may disproportionately influence the mean, and therefore also influence the means difference.

2. T-tests

The t statistic was also considered as a simple metric whereby independent t-tests between human and mouse cell types could be carried out on each 100 Kb region with divergent regions defined as those with the highest magnitude of t. In this method, t-tests were carried out on the normalised data for all human versus all mouse datasets giving a t-test statistic for each 100 Kb region. The values of t obtained showed an approximately normal distribution around a median of 0.026 (Figure 3.11). This leads to a bipolar classification of divergence at either end of the distribution, so either the human data is relatively open in structure and closed in mouse or the human data is relatively closed and the mouse is open. To define the regions where there is a significant structural difference between the species a threshold value of t was needed. One simple strategy is to use the interquartile range (IQR), the difference between the upper Q1 and lower Q3 quartiles within a dataset, to find outlier values. Outliers are classically defined as observations that fall 1.5(IQR) below Q1 or 1.5(IQR) above Q3. Similarly, extreme outliers might indicate more extreme structural divergence where values of $t < Q1\text{-}3(IQR)$, or $t > Q3+3(IQR)$. The main drawback to using such strategies to define divergent regions was the lack of any estimate of statistical significance for the divergent regions

identified. Ideally one would want a nonparametric estimate of significance and an indication of the expected false discovery rate (FDR).

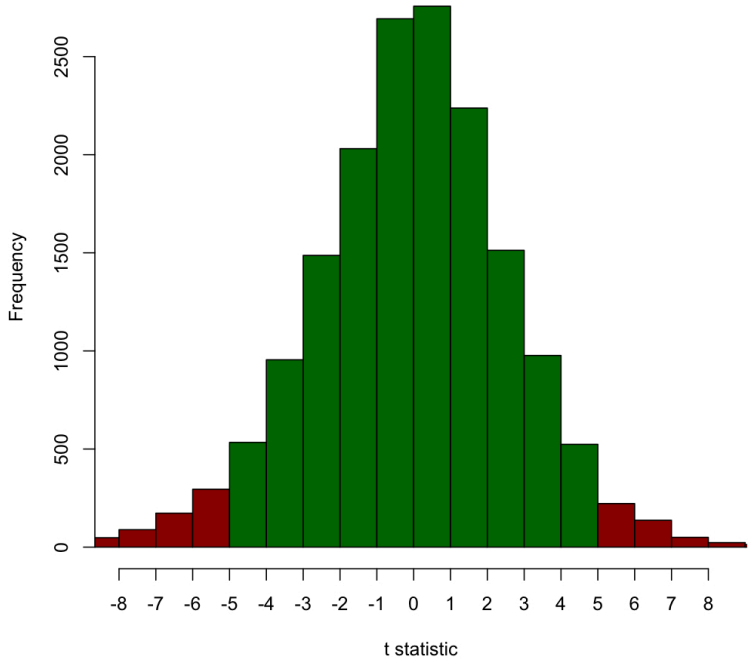

Figure 3.11 Distribution of t-test statistics of human and mouse data from each 100 Kb normalised region. The red bars show outlier, putatively divergent, regions at the ends of the distribution with t values greater than or less than threshold values based upon IQR.

3. Statistical analysis of microarrays (SAM)

The final method was chosen to combine t-tests and permutation testing by combining both together in a non-parametric test from the SAM package (Smyth, 2004). This uses both the strengths of the t test as a divergence metric and permutation testing as a method for defining divergence. This method was adopted to define divergent regions across the orthologous structural dataset and is detailed further below.

3.4.3. A PERMUTATION DERIVED DIVERGENCE METRIC

The approaches used in the SAM package (R package samr) (Tibshirani et al., 2011) were originally designed to pick out genes where the expression level is significantly different between two groups of samples. It is particularly useful when there is an expectation that some genes will have significantly different mean

expression levels between different sets of samples. For example, if looking at differential gene expression between tissue types or between different species. However it was also designed to be flexible enough to be applied to other, comparable datasets (Tusher et al., 2001).

The SAM test used here is analogous to a two class unpaired t-test with permutation derived p-values. In the two-class design, the chromatin structural values are separated into two groups by species (human and mouse) and therefore one group is considered to be "positive significant" if their mean chromatin values are significantly higher than the other. They are considered "negative significant" their mean chromatin values are significantly lower than the other. The normalised data for each 100 Kb region were permuted 100,000 times, and a test statistic *d* is computed for both the original and the permuted data for each region. The value *d* is analogous to the t-statistic in a t-test, in that it is calculated from the difference among mean chromatin structural values, scaled by a measure of variance in the data. SAM in R generates a plot of the observed versus expected (based on the permuted data) *d*-values (Figure 3.12). From this plot, the parameters of SAM can be fine-tuned to set the cut-off for significance by altering the *delta* value, which represents the vertical distance of the line on the graph where observed equals expected (Figure 3.12).

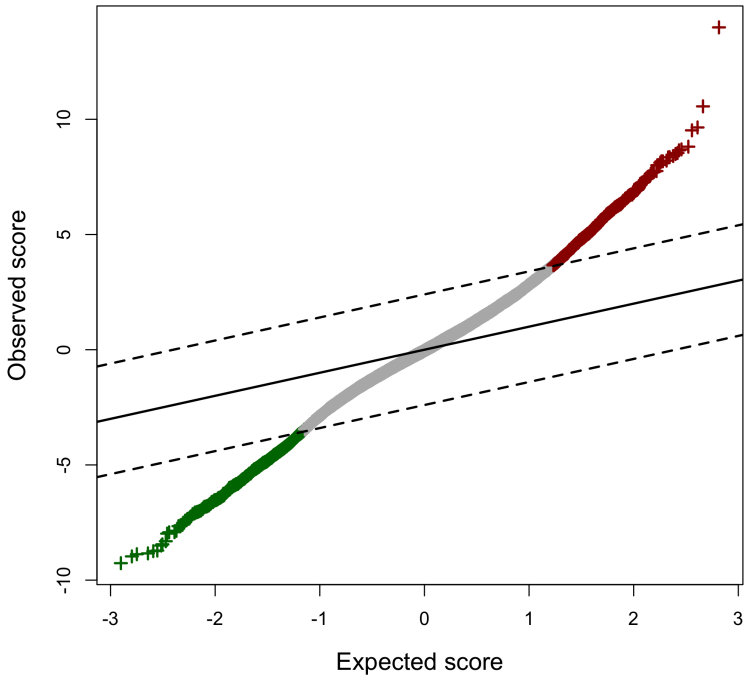

Figure 3.12 Quantifying human-mouse divergence in higher-order chromatin structure. The Q-Q plot from the two class unpaired SAM tests for each orthologous 100 Kb region. Significantly divergent regions (highlighted in green and red) generate unexpectedly extreme observed test scores relative to the expected (permutation based) scores.

The False Discovery Rate (FDR) in this instance is defined as the proportion of regions likely to have been identified by chance as being significant which is calculated as the median number of false positive divergent regions expected (given the permuted datasets), divided by the total number of divergent regions called. As SAM is interactive, it allows for the distribution of the test statistic, *d*, to be checked and then the thresholds set for significance (through the tuning parameter *delta*). The FDR threshold was set to be relatively low (FDR = 2×10^{-4}) to ensure that no false positives expected within the 1719 divergent regions found. The results are necessarily bipolar with positive and negative divergent regions called to indicate human open/mouse closed or human closed/mouse open divergence respectively. Relatively static, non-divergent regions were classed as those with p values that did not pass the FDR threshold. The ability to dynamically alter the input parameters based feedback from the plots and FDR, even before completing the analysis makes

the resulting divergent regions defined more robust.

3.5. STRUCTURAL DIVERGENCE BETWEEN SPECIES AND CELL TYPES

It was important to establish how much of the variation in divergent regions is due to differences between cell types within the two species analysed. To do this, a smaller dataset was constructed using the large volume of normalised RT data, available across both species for the ESC and NPC cell types. This allowed for a rigorous comparison of mouse and human structures across a smaller platform. It also enabled the estimation of the relative degree of divergence in structure between species compared to that seen between cell types within a particular species.

For this analysis, means difference was used as a divergence metric due to the reduced number of cell types used in this RT dataset. The structural divergence between NPC and ESC cell types were assessed within each species, and also species divergence between human and mouse datasets within each cell type (Figure 3.13). The resulting distributions of divergence showed characteristic patterns. The cell type differences within human and mouse RT data showed narrower and more peaked distributions (kurtosis 4.02, 5.20), whereas species differences within a single cell type displayed broader distributions (kurtosis 3.75, 3.69) with inflated tails of divergent regions. In other words, in data from the same experimental platform, concerning the same orthologous regions, differences between species are skewed to relatively high values. These statistically significant (Kolmogorov-Smirnov test, $p < 1.05 \times 10^{-07}$) differences in distributions strongly suggest that divergence in higher order chromatin structure between these species exceeds the divergence seen between cell types.

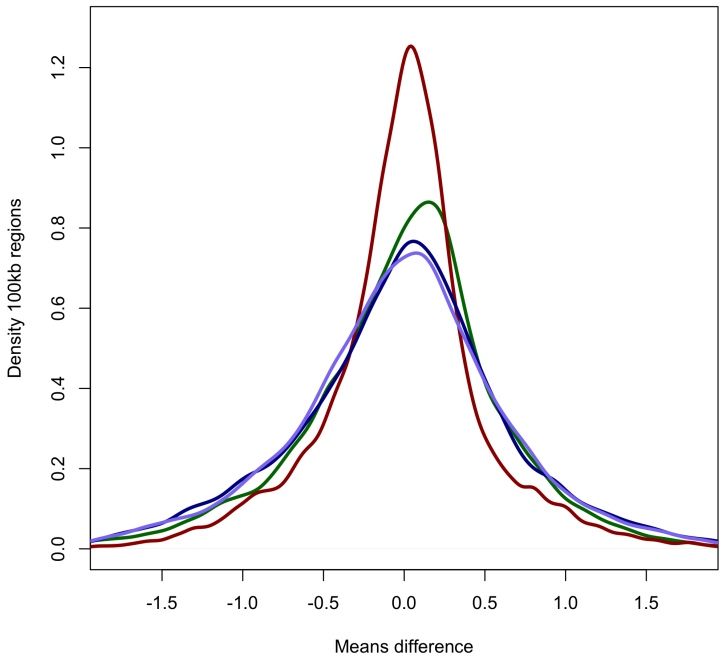

Figure 3.13 The distributions of means differences for replication timing between cell types and species. Red – cell type differences, mouse. Green – cell type differences, human. Navy, species differences, ESCs. Purple – species differences, NPCs.

The genic content of the regions implicated in differences between species and cell types showed similarities to gene enrichments in the divergent regions of the full dataset and this is discussed in Chapter 5.

Chapter 4

Results: A spectrum of divergence in chromatin structure

Topics included in this section:

- The chromosomal distribution of all 1719 structurally divergent regions.
- The discovery of spatial clustering of divergent 100 Kb regions into large divergent domains.
- The discovery that large divergent domains are unevenly distributed across chromosomes and are enriched at telomeres.
- Comparison of large divergent domains to previously established topological domains (Dixon et al., 2012).
- The demonstration that divergent regions can be clustered by their modes and patterns of divergence across all cell types.

4.1. INTRODUCTION

In this chapter, the distribution of structurally divergent regions is examined across chromosomes. We also examine their propensity to cluster spatially into large domains and the clustering of divergent regions by patterns of chromatin values. It is known that patterns of domain level chromatin structure can differ across chromosomes within a species. Human chromosome 19 is known to be significantly more open and accessible than other chromosomes corresponding to higher levels of GC composition and gene density (Castresana, 2002). Conversely, chromosome 18 is known to be relatively gene poor, more peripherally located and enriched in closed chromatin(Croft et al., 1999). It could be expected, therefore, that structural divergence will not be randomly distributed and may be at a higher density on particular chromosomes.

Chromatin structure has increasingly been shown to be organised into large domains (Ryba et al., 2010, Hiratani et al., 2010, Dixon et al., 2012). So it may be expected that some of the structurally divergent 100 Kb regions might group together spatially to form large divergent domains between human and mouse. The extent and distribution of chromatin structure divergence is examined for the first time, and we are also able to identify regions showing similar patterns of divergence.

4.2. DISTRIBUTION OF STRUCTURAL DIVERGENCE

The distribution of all structurally divergent regions, defined by SAM, was examined across all chromosomes. The expected numbers of divergent regions, given the proportion of orthologous 100 Kb regions on each chromosome, were compared with those observed using chi-squared tests, and chromosomes of interest were identified as those generating standardized residuals > 1.96 (Table 4.1, Table 4.3).

Chr	Chromosome Length	Orthologous Coverage	Ratio Covered	Observed	Expected	Chi-Square Residuals
1	2.49E+08	1.33E+08	0.53	91	136.21	**-2.12**
2	2.43E+08	1.54E+08	0.63	154	157.38	-0.14
3	1.98E+08	1.32E+08	0.67	159	134.88	0.99
4	1.91E+08	1.17E+08	0.62	118	120.15	-0.10
5	1.81E+08	1.19E+08	0.66	179	122.10	**2.32**
6	1.71E+08	1.08E+08	0.63	112	110.85	0.05
7	1.59E+08	9.71E+07	0.61	135	99.29	1.65
8	1.46E+08	9.07E+07	0.62	71	92.75	-1.20
9	1.41E+08	7.11E+07	0.5	61	72.71	-0.72
10	1.36E+08	8.60E+07	0.63	129	87.94	**1.97**
11	1.35E+08	8.11E+07	0.6	77	82.93	-0.33
12	1.34E+08	7.38E+07	0.55	55	75.47	-1.27
13	1.15E+08	6.16E+07	0.54	44	62.99	-1.30
14	1.07E+08	6.00E+07	0.56	59	61.36	-0.15
15	1.03E+08	5.12E+07	0.5	45	52.36	-0.53
16	9.03E+07	5.08E+07	0.56	53	51.95	0.07
17	8.12E+07	5.03E+07	0.62	40	51.44	-0.85
18	7.80E+07	5.14E+07	0.66	67	52.56	0.93
19	5.91E+07	1.81E+07	0.31	20	18.51	0.17
20	6.30E+07	3.77E+07	0.6	39	38.55	0.04
21	4.81E+07	1.90E+07	0.39	5	19.43	**-2.06**
22	5.12E+07	1.78E+07	0.35	7	18.20	-1.58

Table 4.1 Significant enrichment or depletion of divergent higher order chromatin across human chromosomes. Significant standardised chi-squared residuals over 1.96 in magnitude are highlighted in red (depletion) or green (enrichment) In the human genome.

Chr	Chromosome Length	Orthologous Coverage	Ratio Covered	Observed	Expected	Chi-Square Residuals
1	1.95E+08	1.30E+08	0.67	116	133.24	-0.77
2	1.82E+08	1.32E+08	0.72	144	134.78	0.39
3	1.60E+08	1.07E+08	0.67	92	109.01	-0.85
4	1.57E+08	1.04E+08	0.67	71	106.66	-1.89
5	1.52E+08	1.01E+08	0.67	106	103.69	0.11
6	1.50E+08	1.01E+08	0.68	139	103.79	1.60
7	1.45E+08	7.98E+07	0.55	104	81.60	1.16
8	1.29E+08	9.07E+07	0.70	85	92.75	-0.41
9	1.25E+08	8.71E+07	0.70	71	89.07	-1.01
10	1.31E+08	9.28E+07	0.71	89	94.90	-0.31
11	1.22E+08	9.14E+07	0.75	82	93.46	-0.61
12	1.20E+08	8.29E+07	0.69	86	84.77	0.07
13	1.20E+08	7.60E+07	0.63	125	77.72	**2.35**
14	1.25E+08	8.65E+07	0.69	85	88.45	-0.19
15	1.04E+08	7.42E+07	0.71	52	75.88	-1.49
16	9.82E+07	7.14E+07	0.73	60	73.01	-0.80
17	9.50E+07	5.80E+07	0.61	72	59.31	0.78
18	9.07E+07	7.00E+07	0.77	80	71.58	0.48
19	6.14E+07	4.53E+07	0.74	61	46.32	1.00

Table 4.2 Significant enrichment or depletion of divergent higher order chromatin across mouse chromosomes. Significant standardised chi-squared residuals over 1.96 in magnitude are highlighted in red (depletion) or green (enrichment) In the mouse genome.

Divergence was far from uniform over the genome, with several human chromosomes showing higher than expected densities of divergent regions (Figure 4.1). In both species, the distribution observed between chromosomes was significantly different to the expectation given the number of 100 Kb regions per chromosome (Chi-squared test in human $p = 4.34 \times 10^{-06}$, in mouse $p = 1.19 \times 10^{-03}$). For instance, human chromosomes 5 and 10 were found to have a 50% excess of divergent regions, while chromosomes 21 and 22 were found to have a greater than 60% depletion (Figure 4.1,Table 4.1). In the mouse genome, only chromosome 13 was found to have a significant enrichment of divergent regions, over 60% more than expected (Figure 4.1, Table 4.2). This raised the question: are divergent regions also clustered within chromosomes? That is, does the distribution of divergent regions within chromosomes reflect larger tracts of divergent chromatin?

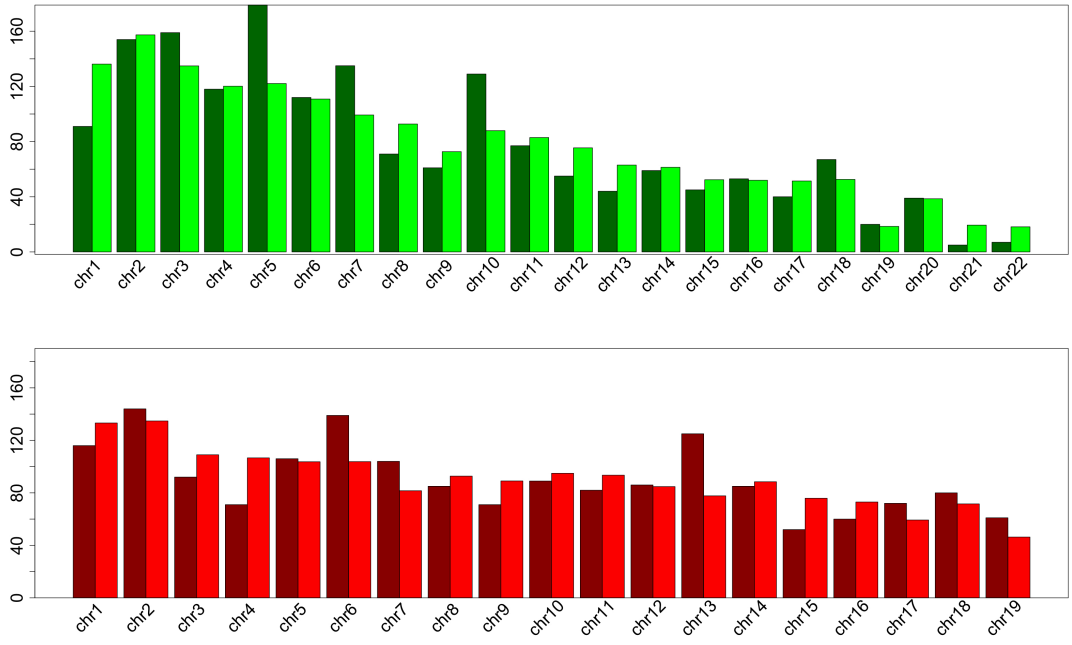

Figure 4.1 Frequency of divergent 100 Kb regions across all human (green) and mouse (red) chromosomes. The bar graph represents the observed (darker colour) and expected (lighter colour) number of divergent regions per chromosome.

4.3. CLUSTERING BY SPATIAL PROXIMITY (LARGE DOMAINS)

From a cursory examination of the data from the regions depicted in Figure 3.10, it appeared that a number of divergent 100 Kb regions were clustered together in the genome at particular loci. The degree of spatial clustering among the divergent regions was formally investigated by measuring the length distribution of consecutive runs of divergent 100 Kb regions observed, relative to the distribution expected using a permutation strategy. All consecutive runs of two or more significantly divergent regions were first identified across the orthologous human genome using Perl scripts. These clusters were required to maintain the polarity of divergence (i.e. all regions involved must be either human open/mouse closed or vice versa). The loci of the orthologous divergent regions were then permuted within chromosomes 10,000 times, and the length of any consecutive runs within each permuted genome was noted. The frequency with which a run of a particular length was seen in the permuted datasets was taken as an approximate p value for runs of that length in the observed dataset (Table 4.3). This strategy is likely to be conservative in detecting large domains of divergent chromatin as it does not allow

for gaps, (e.g. intervening regions that may have marginally failed to reach significance in the test for divergence above), within runs of divergent regions.

Size (Regions)	Observed	Expected	P-Value
1	303	1368.953	1
2	132	146.1667	0.9215
3	83	16.4976	0.0001
4	56	1.927	0.0001
5	35	0.2443	0.0001
6	29	0.0269	0.0001
7	13	0.0045	0.0001
8	11	0.0006	0.0001
9	9	0	0.0001
10	1	0	0.0001
11	1	0	0.0001
12	3	0	0.0001
13	0	0	0.0001
14	1	0	0.0001

Table 4.3 Spatial clustering of large divergent regions. The number of consecutive divergent regions indicates the size of the large domain. The expected distribution is the mean frequency of large regions in the permuted data. The frequency with which a domain of particular size was seen in the permuted datasets was taken as an approximate p value.

The clustering observed was found to be significant ($p < 1 \times 10^{-4}$). 159 unexpectedly large domains were identified in the human genome (160 in mouse) that were at least 400 Kb in size. The mean size was 800 Kb, (Appendix 10.4). The same large orthologous domains were detected in human and mouse genomes when the 100 Kb divergent regions in each genome were clustered independently, and again using the same chi-squared approach as above, large domains were not evenly distributed across chromosomes. For example human chromosomes 3 and 5 had around twice the density expected, but in contrast chromosomes 1 and 9 had around half the density expected (Figure 4.2).

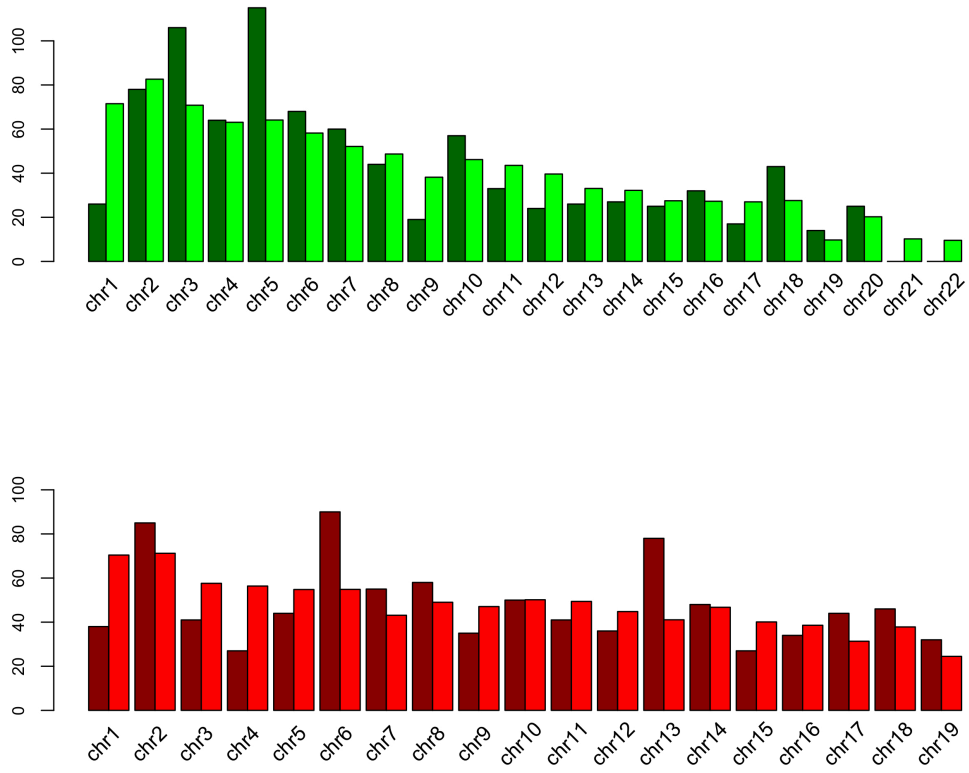

Figure 4.2 Frequency of divergent 100 Kb regions within the 159 large spatial divergent domains across all human (green) and mouse (red) chromosomes. The bar graph represents the observed (darker colour) and expected (lighter colour) number of divergent regions per chromosome.

Although the mean size of the large domains was 800 Kb, some were much larger. The three largest domains of divergent chromatin were between 2.1 and 2.7 Mb in size and were found to occupy subtelomeric regions of human chromosomes 2, 6 and 9 (Figure 4.3). However, in each case the orthologous mouse domains were not proximal to the telomeres, occupying positions long distances (80-100 Mb) away. This appeared to reflect the distribution of chromatin divergence across the human genome in general, with unexpected excesses of divergence towards the ends of some human chromosomes (Figure 4.4).

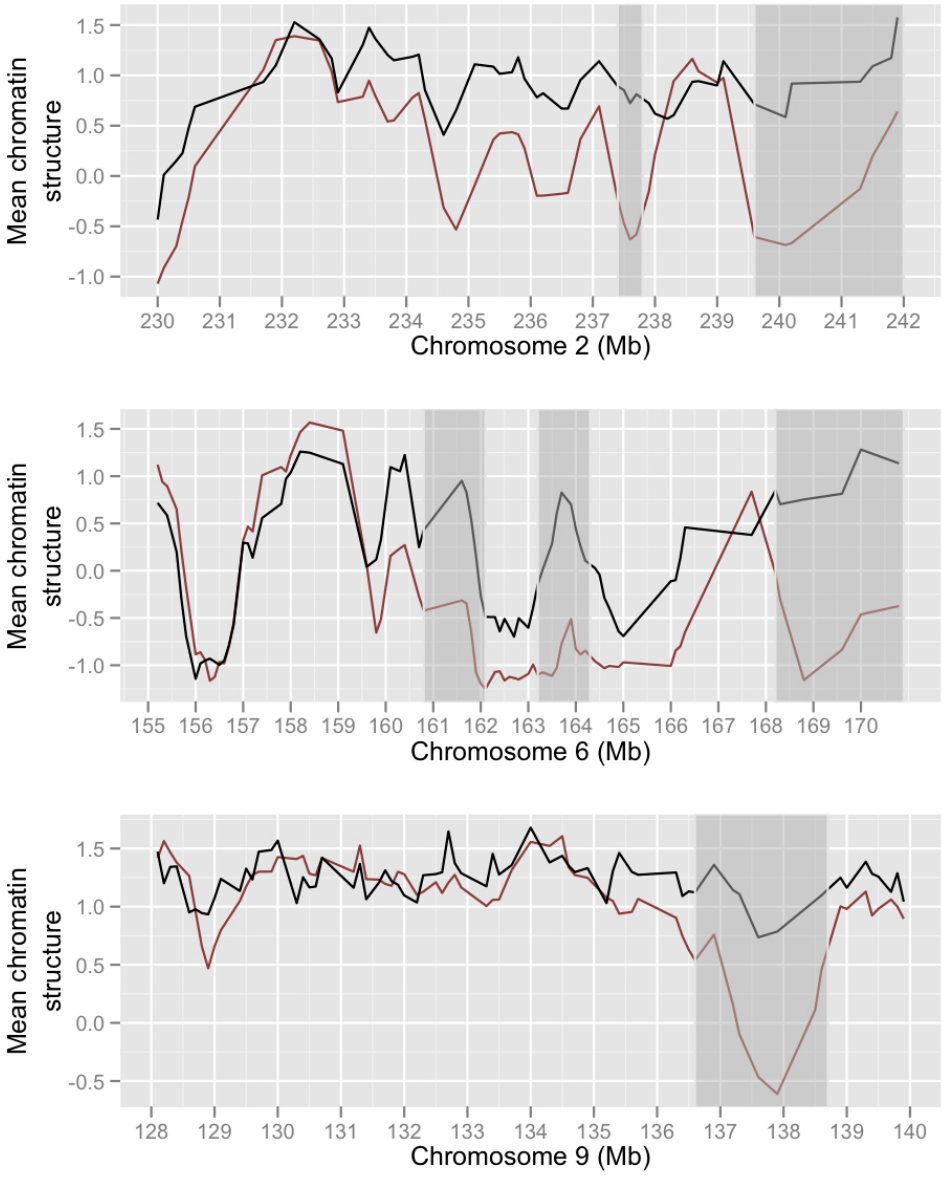

Figure 4.3 The three largest divergent domains on human chromosomes. The line plot shows mean, normalised human (black) and mouse (red) higher order chromatin structure across human chromosomes. Unexpectedly large divergent areas are highlighted in grey.

4.4. LARGE DIVERGENT DOMAINS ARE ENRICHED AT TELOMERES

To investigate this further, subtelomeric regions were designated as genomic areas within 5 Mb of the first and final base pairs of the chromosome assemblies, and within the final base pair of the acrocentric mouse assemblies. To robustly calculate enrichment or depletion of large structurally divergent domains with

subtelomeric areas, a circular permutation strategy was used. This involved revolving the positions of the large divergent domains within each circularised chromosome by a random number for 10,000 permutations. Regions assigned a new position greater than the final base pair of the chromosome with this method are re-inserted at the start of the chromosome (plus the number of bases by which they exceeded the final base pair). For the purposes of the permutations, the chromosomes are regarded as circular and maintain the degree of clustering seen among the observed divergent regions. The number of permuted datasets, n, possessing a number of divergent regions within subtelomeric (or pericentromeric) regions greater than or equal to the observed number were noted, and used to calculate approximate p-values ($n/10{,}000$) for enrichment. The significance of depletion was calculated analogously, according to the number of permuted datasets possessing the same or fewer divergent regions within the areas of interest.

	Human			Mouse		
Chr	Observed	Expected	P-value	Observed	Expected	P-value
1	7	1.98	3.00E-03	1	4.48	6.10E-02
2	10	3.32	1.00E-03	1	3.50	1.13E-01
3	13	9.74	1.58E-01	4	2.26	1.69E-01
4	6	6.77	4.61E-01	4	1.90	1.11E-01
5	2	7.55	1.20E-02	4	3.50	4.72E-01
6	9	4.55	2.70E-02	1	3.08	1.65E-01
7	11	8.33	1.88E-01	24	11.82	1.00E-03
8	0	3.75	1.50E-02	1	3.09	1.58E-01
9	11	5.31	1.20E-02	1	1.75	4.61E-01
10	14	10.93	1.79E-01	0	2.22	1.08E-01
11	8	6.33	2.73E-01	8	5.30	1.51E-01
12	3	3.53	5.24E-01	4	2.89	3.07E-01
13	6	0.86	0.00E+00	0	5.68	2.00E-03
14	1	2.11	3.56E-01	1	4.48	5.40E-02
15	4	3.10	3.80E-01	1	2.36	3.00E-01
16	5	5.85	4.56E-01	1	2.64	2.26E-01
17	7	4.83	2.10E-01	2	4.15	1.91E-01
18	25	7.82	0.00E+00	2	4.38	1.57E-01
19	1	1.97	3.74E-01	11	5.38	1.10E-02
20	6	6.19	5.69E-01			
21	1	0.59	4.71E-01			
22	2	1.14	3.15E-01			

Table 4.4 Distribution of divergent regions across telomeres. Numbers of divergent regions within human (left) and mouse (right) telomeres are indicated in the observed column. The expected distribution is the mean frequency of telomeric divergent regions in the permuted data. The frequency with which a domain of particular size was seen in the permuted datasets was taken as an approximate p value.

The density of orthologous 100 Kb regions within subtelomeric regions was not significantly different from the genome as a whole, either for human (5 Mb subtelomeric region mean density = 23.70; mean density across all genomic 5 Mb bins = 28.10) or mouse (5 Mb subtelomeric region mean density = 34.60; mean density across all genomic 5 Mb bins = 34.20). The excess of divergent regions was most pronounced within the subtelomeric regions of four human chromosomes (1, 2, 13, 18), and was also seen overall for the human genome (p = 0.016) (Table 4.4,

Figure 4.4). In contrast, most mouse subtelomeric regions showed a relative depletion of divergence, with none showing significant enrichment, and non-significant depletion over the mouse genome in general. Pericentromeric regions were also examined, similarly these were designated as regions falling within 5 Mb of the centromeres in human and for the mouse, which has telocentric chromosomes, regions falling within 5 Mb of just the p arm of each chromosome. No significant enrichment or depletion was found overall for pericentromeric regions in either species. This may be due to well-characterised differences in the chromatin structures found at human and mouse telomeres. Subtelomeric regions are known to be amongst the most rapidly evolving DNA sequences in the genome and have been subject to extensive divergence recently in the primate lineage (Linardopoulou et al., 2005). The current data suggest that the higher order chromatin structures at some primate subtelomeric regions have also been subject to dramatic change.

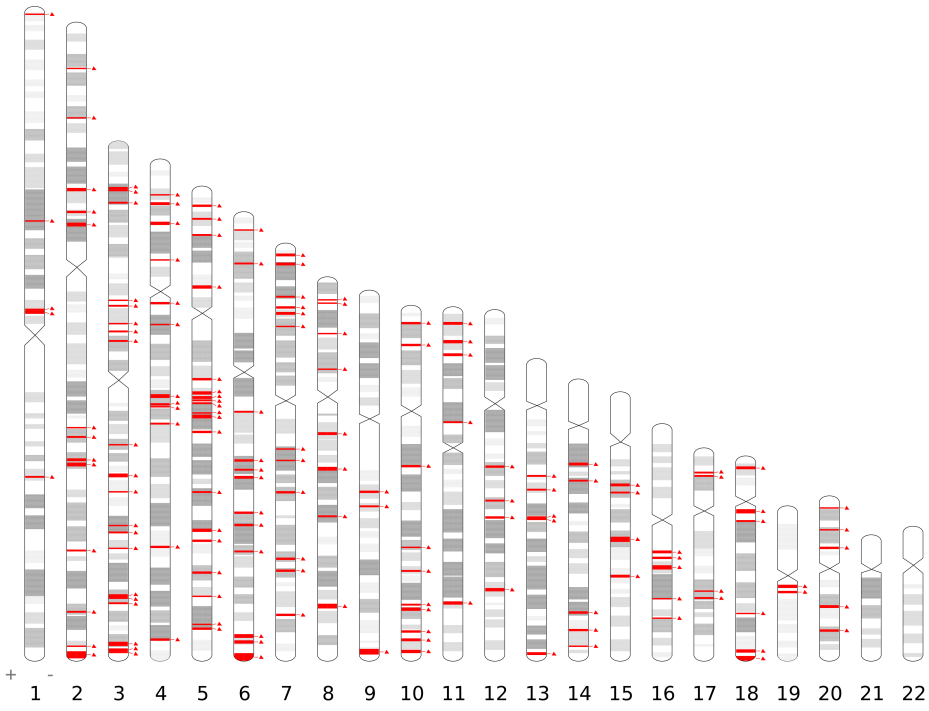

Figure 4.4 Chromosomal distribution of large divergent domains. The Ideogram shows the distribution of significantly large structurally divergent domains (red) across all human chromosomes.

The size distribution of the large divergent domains appeared similar to the

ES cell chromatin-mediated regulatory topological domains recently reported in the mouse and human genomes (Dixon et al., 2012). The topological domains had a median size of 880 Kb, similar to the 800 Kb mean size of the large divergent domains which may suggest that the domains of divergent chromatin may represent divergent regulatory domains. To investigate this further the similarity in domain boundaries between these regulatory topological domains and the divergence clusters was examined using a circular permutation approach. The median distance between divergent cluster boundaries and the nearest regulatory domain boundaries was compared to the median distance seen in 10,000 datasets that had undergone circular permutation. The proportion of datasets generating a median distance less than or equal to the observed median distance was taken as an approximate p-value. The relatively large sizes of both the topological and large divergent domains meant that the distances between all reported domain boundaries are estimates within tens or hundreds of kilobases. In the human genome, the median distance between the boundaries of divergence clusters and the nearest ES cell regulatory domain boundaries was 207,852 bp, which was somewhat less, though not significantly different ($p = 0.054$) from the expected median distance given 10,000 permuted datasets (235,581 bp). Similarly, in the mouse genome, the equivalent median distance was 260,000 bp, which is not significantly different ($p = 0.087$) from the expected distance given 10,000 permuted datasets (290,095 bp). Thus overall there is no strong association between divergent regions and these regulatory domains, which is consistent with most structural divergence being selectively neutral.

Particular genomic locations are known to change their replication timing status upon differentiation from ESC to NPC cells (Hiratani et al., 2010). These changes involve a dynamic switch in replication timing either from late in the cell cycle to early (LtoE) or early in the cell cycle to late (EtoL) upon differentiation. These changes are also coincident with repositioning of loci toward (in EtoL) or away (in LtoE) from the nuclear periphery. This suggests that significant epigenetic changes can occur to facilitate cell-type-specificity of genome organization and may be a prerequisite for large-scale transcription changes upon lineage commitment. The correspondence between the divergent domains and regions known to change replication timing during cellular differentiation was examined. Of the 1719 divergent regions, 60 overlapped these structurally dynamic regions. This represented a significant depletion compared with an expected number generated

derived from mean overlaps in 10,000 permutated datasets of 99.73 which represents a significant depletion ($p < 0.013$). Therefore, regions structurally divergent between species are not enriched for regions that alter their replication timing status upon differentiation, which have been shown to have conserved size and function between species (Ryba et al., 2010).

4.5. CLUSTERING BY DIVERGENCE TYPE

In addition to spatially grouped clusters of divergence, evidence was sought for subclasses of structurally divergent genomic regions showing unexpectedly similar patterns of divergence across all cell types. This was approached by hierarchically clustering all 1149 divergent regions according to their normalised chromatin structure values across all datasets using with 1-Rho as a distance metric. The statistical significance of clusters was assessed using multiscale bootstrap resampling, which provides a better approximation to an unbiased p-value than normal bootstrap resampling (Suzuki and Shimodaira, 2006). In total 22 significant clusters of regions, showing unexpectedly similar patterns of divergence were identified. These clusters all involved regions situated on multiple chromosomes, but despite this they often showed strikingly similar patterns of divergence across the structural data as well as stark separations of mouse and human structures. Of these 22 clusters, 4 contain genes that show significant enrichments of functional annotation (Figure 4.5), this is further investigated in Chapter 5.

From the results so far, the picture that emerges is of widespread conservation of higher order chromatin structure across the mammalian genome with a small proportion of orthologous regions showing strong evidence of divergence consistently across cell types. These regions show greater divergence between species than between cell types and are non-randomly distributed across the genome. They cluster together in larger stretches of divergent chromatin and can also be hierarchically clustered by patterns of divergence.

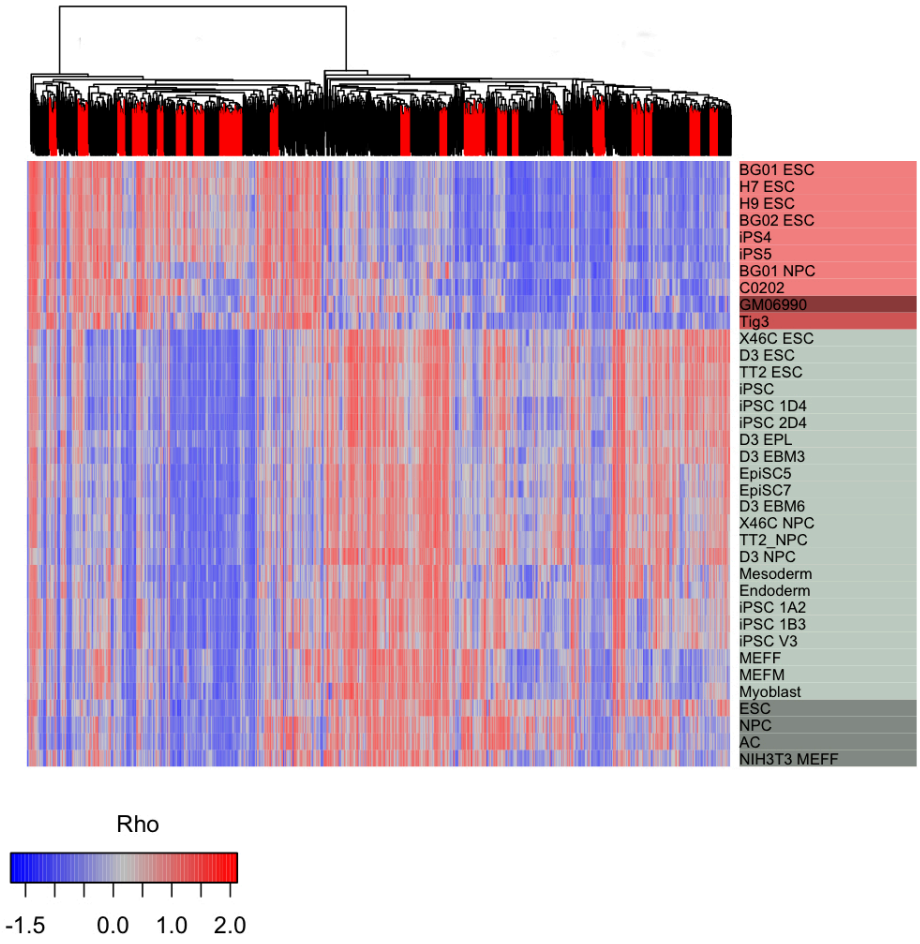

Figure 4.5 Hierarchical clustering indicates chromatin divergence subclasses. The heatmap represents open (blue) and closed (red) higher order chromatin for each 100 Kb divergent region (x-axis) over all datasets (y-axis). Datasets are labelled according to the experimental platform of origin: light grey = mouse RT, light pink = human RT, dark grey = mouse LA, medium pink = human LA, dark pink= human Hi-C. Divergent loci are clustered by structural similarity as reflected in the dendrogram and significant (unexpectedly similar) clusters are highlighted in red.

Chapter 5

Results: Divergent higher order chromatin and gene function

Topics included in this section:

- Examination of protein coding gene densities across all 100 Kb orthologous structural regions and structurally divergent regions.
- Examination of the densities of different RNA gene classes across all 100 Kb orthologous structural regions and structurally divergent regions.
- Investigation of functional enrichments within different groups of divergent and non-divergent higher order chromatin structure. These include:
- All 1719 divergent higher order chromatin regions.
- Regions identified as divergent between species and cell type within replication timing data.
- Large divergent region clusters.
- Divergent regions clustered by similarity of divergence type.
- Exploration of the correspondence between structural divergence and divergently expressed orthologous human and mouse gene pairs.

5.1. INTRODUCTION

In this chapter, the genic content and functional enrichments of regions with known higher order chromatin structure, including structurally divergent regions, are examined. From previous studies, it has been found that the density of genes across the genome is non-randomly distributed with strong correlations observed to higher order chromatin structure (Gilbert et al, 2004). A higher density of protein coding genes is observed in relatively open, active, early replicating (Craig and Bickmore, 1994), chromatin environments, possibly providing increased accessibility to transcription factors (Gilbert et al, 2004). This analysis is replicated in the current mammalian orthologous chromatin dataset with the intention of providing new insights into gene density across chromatin structure and also within the divergent chromatin regions. Enrichments of functional annotation of genes present in the different classes of divergent and non-divergent chromatin structure are also examined. To complement the gene density and functional enrichment analysis, relative densities of different RNA classes in non-divergent and divergent regions are also explored, to shed light on the types of genes that have evolved in structurally variable regions and potential mechanisms of structural regulation. Finally, we examine the expression of orthologous gene pairs using previously published data within the orthologous regions of known structure, to study the correspondence between higher order structural divergence and gene expression divergence.

5.2. GENIC CONTENT OF STRUCTURAL REGIONS

To investigate the relationships between genic content and non-divergent and divergent higher order chromatin structure, gene densities from Ensembl were compared in both species. Confirming previous results, gene density per grouped class of chromatin structure increased with structural openness (Gilbert et al., 2004) (Figure 5.1). In both species, the difference in gene density between both sets of regions was non-significant (Mann-Whitney test in human $p = 0.17$, in mouse $p = 0.52$). Human gene densities in non-divergent regions (2.34 per 100 Kb on average) were not significantly different from either human open divergent regions (2.09 per 100 Kb; Mann-Whitney $p = 0.45$), or human closed divergent regions (2.43 per 100 Kb; Mann-Whitney $p = 0.72$). Similarly, mouse gene densities in non-divergent regions (1.77 per 100 Kb) were not significantly different from either mouse open

divergent regions (1.91 per 100 Kb; Mann-Whitney p = 0.97), or mouse closed divergent regions (1.33 per 100 Kb; Mann-Whitney p = 0.51) (Figure 5.1).

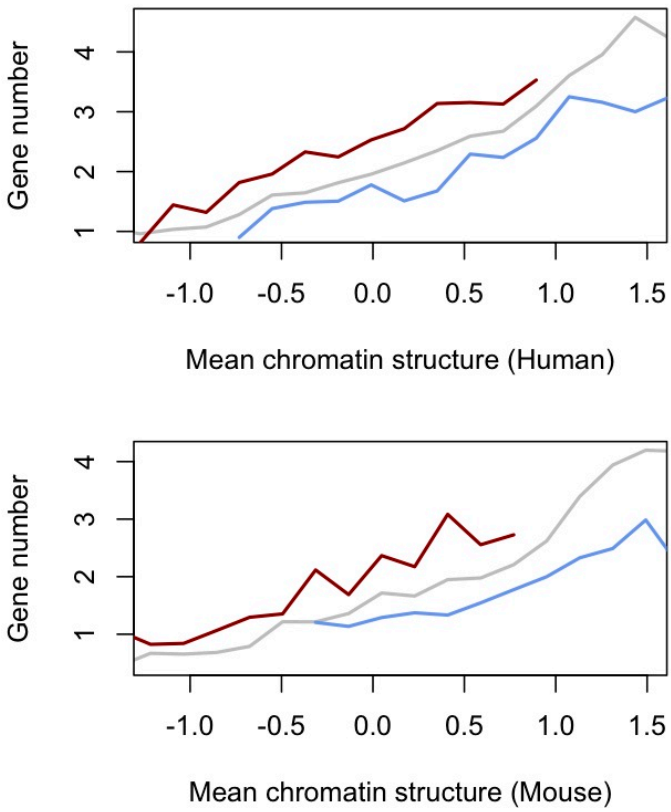

Figure 5.1 Gene densities across categorised bins of chromatin structure. Increasing values of chromatin structure across the x-axis indicate increased accessibility of chromatin structure. Gene densities are shown in non-divergent chromatin (grey), open divergent (blue) and closed divergent (red).

To investigate the genic relationship to higher order structure further, specific RNA gene classes (rRNA, snoRNA, snRNA, miRNA, lincRNA), also obtained from Ensembl, were analysed in the same way. Of all RNA classes, only lincRNAs showed significant differences between divergent and non-divergent 100 Kb regions. There were higher densities of lincRNA genes in both human (divergent mean density: 0.31 genes/Mb; non-divergent mean density: 0.20 genes/Mb; Wilcoxon $p = 1.48 \times 10^{-8}$) and mouse (divergent mean density: 0.12; non-divergent mean density: 0.09; Mann-Whitney $p = 3.68 \times 10^{-4}$) divergent (human closed/mouse open) regions. This particular class of RNA molecules is thought to regulate embryonic stem cell differentiation via the assembly of chromatin complexes and

the establishment of activating or repressive domains (Yaffe et al., 2010). The current data tentatively suggest they may also have played roles in chromatin divergence. The main caveat to this suggestion is the so far lack of complete knowledge of lincRNAs in either genome.

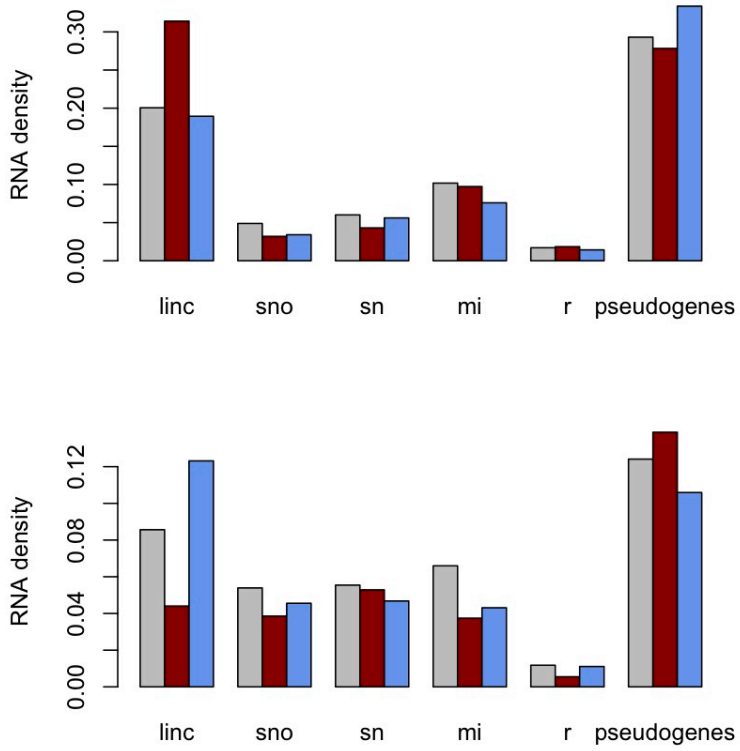

Figure 5.2 Densities (genes/Mb) of different types of RNA classes in non-divergent (grey), closed divergent (red) and open divergent (blue) regions in the human (top) and mouse (bottom) genome.

Functional enrichment was not carried out for RNA genes, as most RNA genes are poorly functionally annotated.

5.3. FUNCTIONAL ENRICHMENTS ACROSS DIVERGENT CHROMATIN

5.3.1. ALL DIVERGENT REGIONS

The 907 divergent human open/mouse closed 100 Kb regions contained 1142 human genes and 757 mouse genes, and both showed significant enrichments for multiple terms associated with olfactory receptors (ORs), implicating particular loci (Table 5.1). These loci can often be seen as enrichments for genes within particular

gene clusters. The mouse genes involved were disproportionately those located in particular OR gene clusters on chromosome 7E3 and 6B1-B2.1, while the human genes were clustered at the orthologous locations at 11p15.4 and 7q35 respectively. The 7E3 region is late replicating across all mouse cell types, but the orthologous human OR cluster at 11p15.4 is within relatively early replicated chromatin, and at least three of the human OR genes present are pseudogenes. This is consistent with the active maintenance of the repressive, late replicating chromatin state necessary for OR function (McClintock, 2010) in the mouse lineage, contrasting with OR pseudogenisation and loss of selective constraint on chromatin structure in the human lineage. Notably, recent work indicates a similar unusual primary chromatin structure (involving H3K9me3 and H4K20me3) at OR containing loci and KRAB-ZNF containing loci in the mouse genome (Magklara et al., 2011). This raised the possibility of an association between divergent higher order chromatin structures and particular classes of histone modifications. It also suggests that the repressive, relatively closed higher order chromatin structures consistently seen at this region of the mouse genome, but not evident in human cells, could have evolved as part of the regulatory landscape associated with OR gene cluster evolution in rodents.

Species	Divergence	Term	Description	Gene	FDR
Human	Human open/ Mouse closed	CYTOBAND	11p15.4	15	2.17E-07
		GO:0007606	Sensory perception of chemical stimulus	21	4.15E-06
		GO:0050877	Neurological system process	41	2.36E-04
		CYTOBAND	10p13	8	4.44E-04
		GO:0007186	G-protein coupled receptor signalling p-way	36	6.34E-04
Human	Human closed/ Mouse open	IPR001827	Homeobox protein, antennapedia type	10	7.33E-04
		CYTOBAND	18q23	6	7.52E-03
		GO:0003002	Regionalization	21	1.50E-02
		CYTOBAND	6q27	6	4.15E-02
		CYTOBAND	2q37.3	9	4.38E-02
Mouse	Human open/ Mouse closed	GO:0007606	Sensory perception of chemical stimulus	39	3.58E-15
		GO:0007608	Sensory perception of smell	34	9.10E-13
		IPR000725	Olfactory receptor	33	1.15E-12
		GO:0004984	Olfactory receptor activity	33	3.45E-12
		IPR017452	GPCR, rhodopsin-like superfamily	47	5.58E-12
Mouse	Human closed/Mouse open	GO:0003002	Regionalization	32	3.39E-06
		GO:0009952	Anterior/posterior pattern formation	27	3.97E-06
		GO:0007389	Pattern specification process	36	9.09E-06
		CYTOBAND	2 45.0 cM	9	1.89E-05
		CYTOBAND	19 D2	12	4.84E-05

Table 5.1 The top five enriched human and mouse annotation terms for protein coding genes within the 1719 divergent regions of higher order chromatin. Full list in Appendix 10.2.

Other enriched terms included those related to a protocadherin (Pcdh) gene cluster present at 5q31.3 in the human genome, and to the orthologous mouse Pcdh cluster on mouse chromosome 18qB3 (See Appendix 10.2). Recent work has shown this region adopts distinct chromatin environments in different mouse neuronal cells to coordinate Pcdh gene expression and thereby plays critical roles in establishing neuronal diversity and connectivity during development (Hirayama et al., 2012). A third cluster of genes coincides with this class of divergent regions on mouse chromosome 8D3 (and human 16q21) and is enriched for genes encoding MARVEL, a transmembrane domain involved in membrane apposition (See Appendix 10.2). The family of chemokine-like proteins containing this domain have been implicated in inflammation, immunity and development but most are not well characterised. Of the five MARVEL containing genes within the 8D3 divergent cluster, three are unstudied, but Cmtm2a and Cmtm3 are both implicated in the proliferation and development of particular testicular cells (Wang et al., 2008,

Qamar et al., 2010). The human ortholog of Cmtm3 was present in the orthologous human divergent region at 16q21 and is a known tumour suppressor gene that shows frequent inactivation via chromatin-mediated silencing in several cancers (Wang et al., 2009). It is evident that developmental gene clusters showing cell type specific regulation are unexpectedly common at regions displaying divergent higher order chromatin.

The genes within the divergent human closed/mouse open 812 orthologous regions contained 1285 human genes and 1102 mouse genes. These also showed significant enrichment for genomic regions harbouring particular gene clusters. Both human and mouse genes in these regions showed significant enrichment for terms associated with developmental genes containing Antennapedia type homeobox domains (IPR001827) (See Appendix 10.2). The genes involved are developmental genes present at the HOXA (human HOXA1-A7; Figure 2B) and HOXD (human HOXD1-4) clusters. Both clusters are implicated in multiple cancers and other disorders, and are tightly regulated by higher order chromatin environments (Wang et al., 2011, Tschopp et al., 2011). It is thought that structural divergence within the chromatin domains harbouring these clusters underlies many important innovations in the vertebrate body plan (Montavon and Duboule, 2013). Again, it seems that developmentally regulated genes are over-represented within regions of divergent chromatin. However, it is worth noting that the proportion of divergent regions generating significant functional enrichments (that is, those divergent regions possessing the genes responsible for the functional enrichments seen) was modest overall, constituting 6% of human and 11% of mouse divergent regions in total.

5.3.2. SPECIES AND CELL TYPE SPECIFIC DIVERGENT CHROMATIN

There were marked differences in the genic content of regions showing structural divergence between cell types and those divergent between species. As detailed in Chapter 3 we identified regions showing evidence for structural divergence between cell types and species using RT mouse and human datasets in matched (ES and NPC) cells.

Divergence		Term	Description	Gene	FDR
Species Differences	ESC	IPR007237	CD20/IgE Fc receptor	8	7.31E-05
		GO:0000786	Nucleosome	11	8.73E-04
		GO:0065004	Protein-DNA complex assembly	12	2.68E-03
		CYTOBAND	11q12.2	10	7.42E-03
		GO:0006334	Nucleosome assembly	11	1.04E-02
	NPC	CYTOBAND	1q42.13	12	2.73E-07
		CYTOBAND	1p36.33	14	5.30E-07
		IPR012287	Homeodomain-related	38	3.31E-06
		CYTOBAND	14q11	12	4.03E-06
		IPR001356	Homeobox	34	1.70E-04
Cell Type Differences	Human	CYTOBAND	14q11	12	1.63E-07
		CYTOBAND	7p15-p14	10	5.19E-05
		IPR012287	Homeodomain-related	30	6.62E-05
		IPR001356	Homeobox	28	8.62E-05
		IPR017970	Homeobox, conserved site	28	3.77E-04
	Mouse	CYTOBAND	11 A4	15	1.16E-04
		CYTOBAND	3 A1	11	4.91E-03
		CYTOBAND	7 24.0 cM	6	7.65E-02
		CYTOBAND	3 B	11	1.48E-01
		CYTOBAND	13 A1	10	1.90E-01

Table 5.2 The top five enriched human and mouse annotation terms for genes within regions of higher order chromatin divergent between species and between cell types. Full list in Appendix 10.3.

Regions divergent between human ESC and NPC cell types were enriched for homeodomain containing genes (IPR001356), including the HOXA cluster as before. This reflects the finding that the HOXA cluster possesses higher order chromatin structures that can be conserved across mammals and yet are variable between embryonic cells and NPC (Kim et al., 2011). However, the results also show genes at the HOXD cluster on 2q31 and a variety of other homeodomain containing genes across the genome have divergent structures between these two cell types, suggesting a widespread modulation of chromatin landscapes at such loci during neural differentiation.

There was however no detectable enrichment of homeodomain genes at the regions structurally divergent between ESC and NPC types within the mouse RT data, even though large–scale changes in higher-order chromatin conformation are

seen at HOX loci during neuronal differentiation of ES cells (Morey et al., 2007). This suggests that higher order chromatin may play different roles in development between rodent and primate lineages, and reflects differences in the exact nature of the NPCs that arise from the differentiation of human and mouse ES cells. Consistent with this, in comparisons of RT data between species, homeodomain genes were again implicated: the most divergent regions between mouse and human NPCs implicate a similar set of 34 homeodomain genes including those at HOXA and HOXD clusters (See Appendix 10.3). Again, it seems that gene clusters with functions in mammalian development appear to be a focus of structural alterations during evolution.

5.3.3. LARGE DIVERGENT DOMAINS

In Chapter 4 it was discovered that structurally divergent 100 Kb regions cluster within the mammalian genome, forming large divergent domains with mean size 800 Kb. As might be expected these large divergence domains showed similar patterns of functional enrichments as those discussed above (Table 5.3, Appendix 10.4). For example, the divergent region mentioned already at 11p15.4 containing an OR gene cluster was rediscovered as part of a larger 800 Kb domain. Similarly the divergent region containing the 7p15.2 HOXA genes was found to extend to 800 Kb, and to include neighbouring lincRNA genes such as HOTAIRM1 which is active in HOXA regulation during neurogenesis and differentiation (Lin et al., 2008). An additional 800 Kb region at 7q21.3 showing a novel functional enrichment also emerged, which contains the paraoxonase gene cluster, these genes are imprinted in the mouse genome and exhibit unusual, allele-specific expression dependent on developmental stage in human cells (Parker-Katiraee et al., 2008).

Large spatial region	Term	Description	Gene	FDR
chr11 5900000 6699999	CYTOBAND	11p15.4	15	2.05E-25
	PIRSF038651	G Protein-Coupled Olfactory Receptor	7	1.96E-07
	GO:0007608	Sensory Perception Of Smell	8	2.02E-05
	GO:0007606	Sensory Perception Of Chemical Stimulus	8	8.59E-05
	IPR000725	Olfactory Receptor	7	5.77E-05
chr16 54000000 55499999	IPR003893	Iroquois-Class Homeodomain Protein	3	1.34E-04
	IPR001356	Homeobox	3	9.32E-02
	IPR017970	Homeobox, Conserved Site	3	9.45E-02
	IPR012287	Homeodomain-Related	3	1.01E-01
	CYTOBAND	16q11.2-Q13	2	1.22E-01
chr16 66500000 66899999	IPR008253	Marvel	5	7.09E-07
	GO:0042330	Taxis	5	8.71E-04
	GO:0006935	Chemotaxis	5	8.71E-04
	GO:0005125	Cytokine Activity	5	5.41E-03
	GO:0007626	Locomotory Behaviour	5	1.04E-02
chr7 141100000 141899999	CYTOBAND	7q31.3-Q32	3	7.37E-04
	GO:0008527	Taste Receptor Activity	3	7.98E-03
	IPR007960	Mammalian Taste Receptor	3	6.12E-03
	GO:0050909	Sensory Perception Of Taste	3	9.20E-02
	GO:0007186	G-Protein Receptor Signalling Pathway	4	2.28E+00
chr7 26400000 27199999	IPR001827	Homeobox Protein, Antennapedia Type	7	4.44E-14
	CYTOBAND	7p15-P14	6	4.41E-11
	GO:0048562	Embryonic Organ Morphogenesis	7	7.55E-09
	GO:0009952	Anterior/Posterior Pattern Formation	7	9.70E-09
	GO:0048568	Embryonic Organ Development	7	2.55E-08
chr7 94500000 95299999	CYTOBAND	7q21.3	4	1.17E-05
	GO:0004063	Aryldialkylphosphatase Activity	3	3.35E-04
	IPR002640	Arylesterase	3	2.94E-04
	GO:0004064	Arylesterase Activity	3	6.69E-04
	PIRSF016435	Paraoxonase	3	1.29E-04

Table 5.3 The top five enriched human annotation terms for genes within five large clustered regions of divergent higher order chromatin. Full list in Appendix 10.4.

The large divergent domain at 16q11.2, which spans 1.5 Mb, is enriched for Iroquois-class homeodomain genes expressed during development. The genes involved here (IRX3, IRX5 and IRX6) are conserved transcription factors involved in patterning and regionalization of the vertebrate embryo, particularly in neural and cardiac tissues (Zhang et al., 2011). They have also been the focus of innovation in the patterns and timing of their expression during vertebrate embryogenesis

(McDonald et al., 2010).

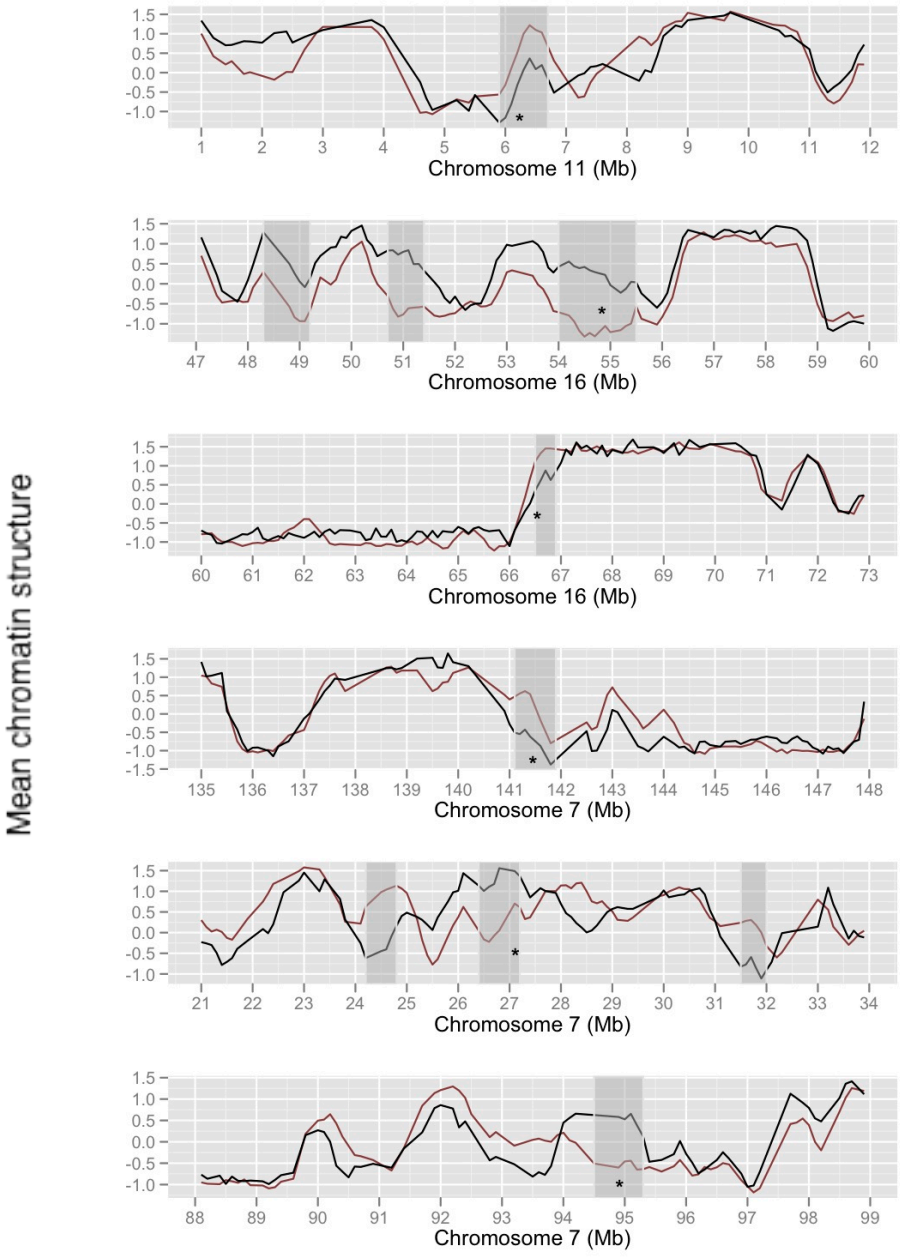

Figure 5.3 Clustering of divergent chromatin in the human genome. The line plot shows mean, normalised human (black) and mouse (red) higher order chromatin structure across human chromosomes. Unexpectedly large divergent areas are highlighted in grey. Asterisks indicate the positions of functionally enriched gene clusters listed in Table 5.3.

5.3.4. DIVERGENT REGIONS CLUSTERED BY DIVERGENCE TYPE

The final divergent subgroups examined for functional gene enrichment were groups derived from hierarchically clustering all divergent 100 Kb regions according to the similarity of their divergence patterns across all structural datasets. Many of these clusters show strong divergence between mouse and human across all available structural datasets, while others show evidence for divergence in a subset of datasets. Significant hierarchical clusters were identified using multiscale bootstrap resampling and all 24 significant clusters were found to contain regions from multiple chromosomes. Of these 24 significant clusters, 6 contained genes showing significant functional enrichments (Table 5.4). Again, many of these gene enrichments were previously seen in the enrichment analyses performed above (Figure 5.4).

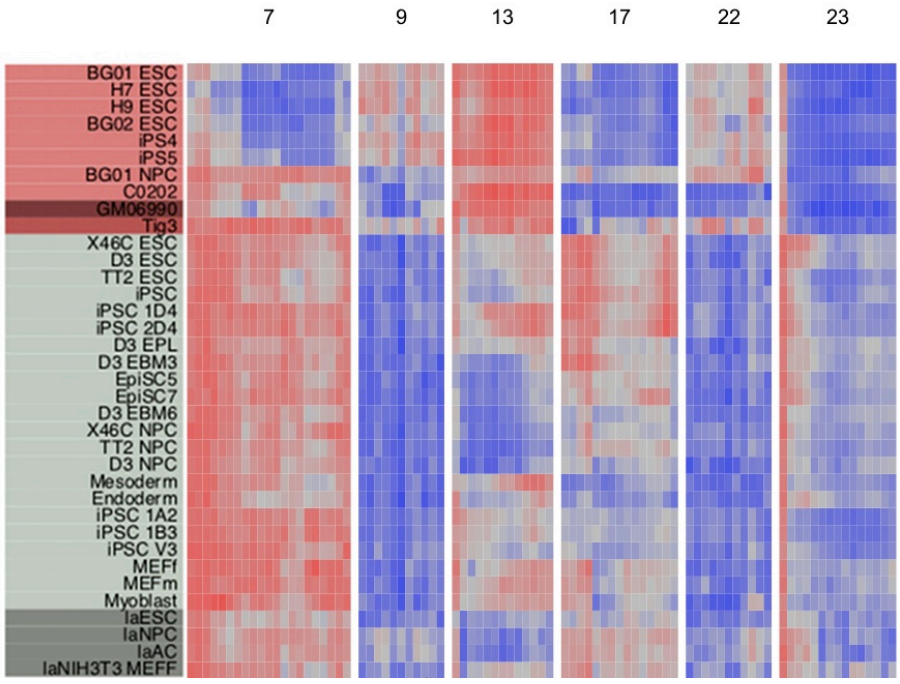

Figure 5.4 Hierarchical clusters (7, 9, 13, 17, 22 and 23) showing significant gene enrichments. The heatmap represents relatively open (blue) and closed (red) higher order chromatin for each 100 Kb divergent locus (x-axis) over all datasets (y-axis). Datasets are coloured according to experiment: light grey = mouse RT, light pink = human RT, dark grey = mouse LA, medium pink = human LA, dark pink= human Hi-C.

Cluster 13 shows notable divergence between species across LA and RT datasets and is again enriched for genes containing Iroquois-Class homeodomains.

Several divergence clusters show a shift in replication timing but little change in lamin association profiles between species. An example of this is cluster 7 (Figure 5.4), which again contains regions that are significantly enriched for OR genes.

Cluster	Term	Description	Gene	FDR
7	CYTOBAND	7q31.3-Q32	3	2.22E-04
	GO:0008527	Taste Receptor Activity	3	2.26E-02
	IPR007960	Mammalian Taste Receptor	3	3.61E-02
	GO:0050909	Sensory Perception Of Taste	3	3.00E-01
	CYTOBAND	3p14.2	2	2.93E+00
9	CYTOBAND	11q12.2	9	4.33E-17
	CYTOBAND	7p22.1	4	2.01E-03
	IPR007237	CD20/IgE Fc Receptor Beta Subunit	3	1.14E-01
	CYTOBAND	1q42.2	2	8.09E+00
	GO:0031224	Intrinsic To Membrane	13	1.76E+01
13	IPR003893	Iroquois-Class Homeodomain Protein	3	4.13E-05
	IPR017970	Homeobox, Conserved Site	3	7.37E-02
	IPR001356	Homeobox	3	7.56E-02
	IPR012287	Homeodomain-Related	3	7.76E-02
	CYTOBAND	16q11.2-Q13	2	6.71E-02
17	CYTOBAND	5q14.1	6	5.92E-10
	IPR017226	Betaine-Homocysteine S-Methyltransferase, BHMT	2	1.07E+00
	GO:0047150	Betaine-Homocysteine S-Methyltransferase Activity	2	1.21E+00
	PIRSF037505	Betaine-Homocysteine S-Methyltransferase, BHMT	2	7.74E-01
	PIRSF037505	Betaine_HMT	2	7.74E-01
22	CYTOBAND	11p15.1	5	5.35E-07
	CYTOBAND	9p24.1	2	4.46E+00
	GO:0016021	Integral To Membrane	6	3.87E+01
	GO:0031224	Intrinsic To Membrane	6	4.37E+01
23	CYTOBAND	18q11.2	8	1.14E-15
	GO:0019887	Protein Kinase Regulator Activity	2	1.66E+01
	GO:0019207	Kinase Regulator Activity	2	1.88E+01
	GO:0006869	Lipid Transport	2	3.68E+01
	GO:0010876	Lipid Localization	2	3.91E+01

Table 5.4 Annotation enrichment within hierarchical clusters of structurally divergent orthologous loci. Gene related annotation terms enriched within clusters of loci with showing similar patterns of divergence; in each case the cluster ID, annotation term ID, number of genes involved, and FDR corrected p-values are provided. Enrichments are calculated relative to the annotation found in all orthologous regions examined.

Again, it seems that structural divergence is disproportionately associated with particular developmental gene clusters, which follow tightly regulated expression patterns targeting specific cell types, and are often known to occupy

unusual chromatin environments. Many of these genes have also been implicated in developmental adaptations during vertebrate evolution and in human disease processes. This may suggest that regions of divergent chromatin structure have evolved different chromatin conformations to facilitate functional divergence at these loci. However it is not possible to exclude non-adaptive hypotheses, for example where divergence in chromatin structure is a neutral consequence of gene family or repeat expansions or other changes in the underlying genomic sequences. Indeed, since the majority of divergent regions show no detectable functional enrichments, selectively neutral divergence appears to be the most probable scenario in most cases.

5.4. CHROMATIN DIVERGENCE ASSOCIATED WITH EXPRESSION DIVERGENCE

To further investigate whether genes within divergent regions have undergone regulatory divergence, enrichment of genes showing divergent expression patterns between human and mouse cells was assessed in structurally divergent regions. An expression dataset from Cai et al (2010) was examined first, which sought significant differences in ES cell expression patterns in orthologous gene pairs. The authors compiled gene lists that contained three different types of orthologous gene pairs. Those pairs with genes upregulated in human, those with genes that were upregulated in mouse and those containing genes that were unusually conserved in expression across species. Unfortunately, the ES cell types involved were not the same as any of the ES cell types used to generate the chromatin data.

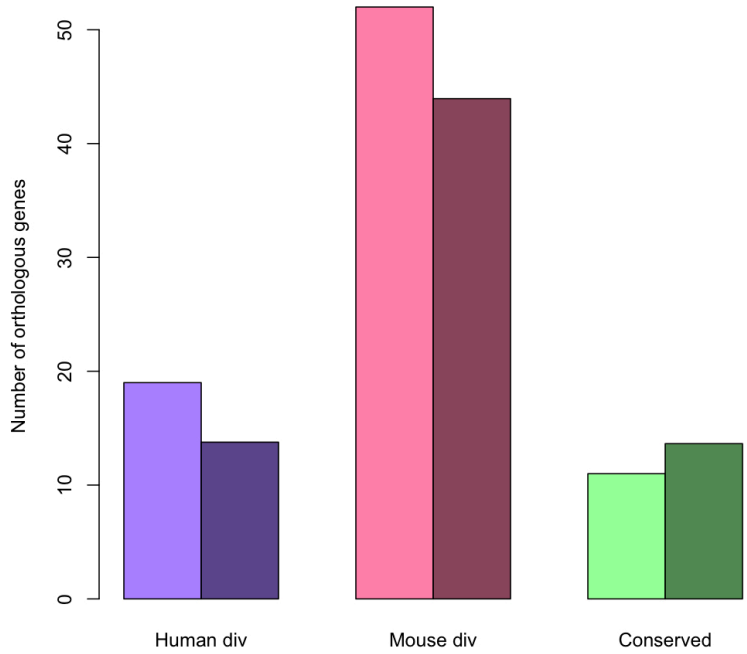

Figure 5.5 Observed (lighter colour) distribution of orthologous genes from Cai et al 2010 within human divergent regions (Human div), mouse divergent regions (Mouse div) and non-divergent regions (Conserved) compared to expected (darker colour) given the distribution of genes across all structural regions. Genes upregulated in human only (purple), genes upregulated in mouse only (pink) and genes with conserved expression (green).

Although the numbers of genes identified by Cai et al (2010) that were also present within the orthologous regions were low (497 divergent and 126 conserved), there was significant enrichment (odds ratio: 1.30; Fisher's Exact test p = 0.04) of divergently regulated genes within the 100 Kb regions of divergent higher order chromatin reported here. Genes with conserved regulation were also under-represented in divergent regions (odds ratio = 0.76; p = 0.331). These patterns were observed in spite of the fact that the data of Cai et al (2010) is based upon human and mouse embryonic cell lines that are not represented in the chromatin data studied here. Another more recent study of expression divergence between orthologous human and mouse genes has been carried out in macrophages (a cell type very different from ES cells) and identified 186 divergent and 972 conserved

gene pairs (Schroder et al., 2012). These data were examined in the same way and revealed no significant enrichment of divergently regulated genes in divergent 100 Kb regions. Indeed the genes divergently regulated in these macrophage data showed the opposite trend, and were somewhat under-represented in regions of divergent chromatin (odds ratio: 0.78; $p = 0.46$). This suggests that the relationship between higher order chromatin divergence and expression divergence is specific to embryonic cell types.

Lastly, a larger orthologous gene dataset was constructed which also measured differential expression between mouse and human ES cells (see Chapter 2 Methodology) and was based upon previous RNAseq studies (Lister et al., 2009). These data provide a higher coverage dataset consisting of log2 fold change measurements for 7,673 mouse-human gene pairs occurring within the orthologous 100 Kb structural regions. This allowed us to assess the extent of expression divergence within the two categories of divergent regions, relative to non-divergent regions (Figure 5.6). There was a striking contrast, with regions open in human but closed in mouse showing a expression divergence consistent with upregulation of human genes (non-divergent median log fold change: -0.48; divergent: -0.33; Wilcoxon $p = 0.23$), while the human closed but mouse open regions showed evidence of upregulation of mouse genes (non-divergent: -0.48; divergent: -1.00; Wilcoxon $p = 3.41 \times 10^{-6}$).

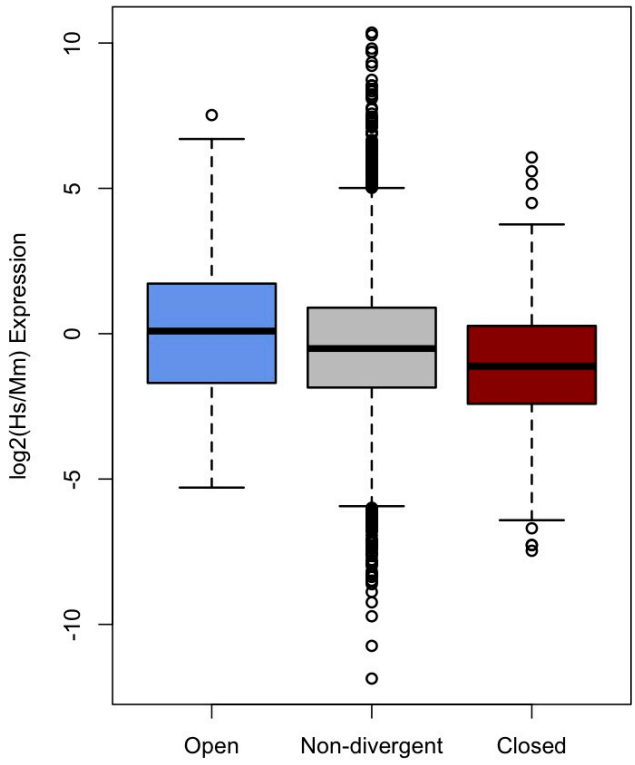

Figure 5.6 Chromatin divergence and expression divergence. Distributions of log2 fold change (log2(human/mouse expression)) for orthologous gene pairs within non-divergent regions (grey), human open/mouse closed (blue) and human closed/mouse open (red). For each plot the bottom and top of the box show the lower and upper quartiles respectively around the median, outliers outside 1.5 x interquartile range are represented as dots.

This pattern of gene expression divergence is expected within divergent regulatory domains demonstrating a respectively active or repressive environment for transcription of human genes. Again, these expression data were generated in embryonic cells similar to, but not identical to those used to derive the chromatin divergence data. It is important to note that there may be a distinct difference between the relative bipolar classification of divergent regions (human open/mouse closed and vice versa) and their absolute normalised chromatin values. For example, it is possible for a region that is relatively open in human and relatively closed in mouse to possess absolute values consistent with a closed conformation in both species. It might be expected that using such absolute values to construct more specific divergent region categories there may be an increase in the correlations to expression divergence. This was indeed the case in spite of the associated

reductions in sample sizes. Regions open in human but closed in mouse (where the absolute human value > 0 and the absolute mouse value < 0) showed a much stronger expression divergence consistent with upregulation of human genes (non-divergent median log fold change: -0.48; divergent: 5.03; Wilcoxon $p < 2.2 \times 10^{-16}$), while the opposite category (restricted to those with absolute human value < 0 and absolute mouse value > 0) showed stronger evidence of upregulation of mouse genes (non-divergent: -0.48; divergent: -4.77; Wilcoxon $p > 2.2 \times 10^{-16}$).

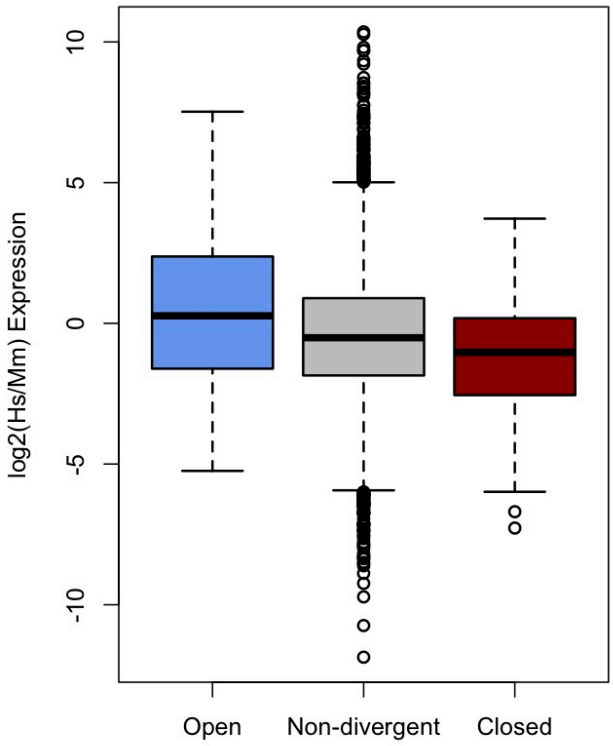

Figure 5.7 Chromatin divergence corrected for absolute open/closed values and expression divergence. Distributions of log2 fold change (log2(human/mouse expression)) for orthologous gene pairs within non-divergent regions (grey), absolute human open (>0)/absolute mouse closed (<0) (blue) and absolute human closed (<0)/absolute mouse open (>0)(red). For each plot the bottom and top of the box show the lower and upper quartiles respectively around the median, outliers outside 1.5 x interquartile range are represented as dots.

These comparisons to expression data provide independent validation of our methodology and suggest a direct link between the regions of divergent chromatin identified and the regulation of resident genes within ES cells

Chapter 6

Results: Higher order chromatin and sequence level features

Topics included in this section:

- Investigation of the relationships between higher order chromatin structure and structural divergence with DNA sequence features. Major features of genomic sequence are explored including base composition and repeat densities.
- Investigation of the relationship between higher order chromatin structure divergence and sequence divergence. Sequence divergence measures include substitution rate, SNP density and indel frequencies.
- Estimation of the proportion of segmental duplications and synteny breaks across regions of higher order chromatin and the relationships to structural divergence.

6.1. INTRODUCTION

This investigation allows for the most comprehensive study so far of the relationships between higher order chromatin structure, structural divergence and divergence at the DNA level. It has been known for some time that higher order chromatin structure has strong correlations with DNA features. For example, base composition within relatively open chromatin domains is GC rich and gene dense (Gilbert et al., 2004). However, relatively closed chromatin domains are comparatively AT rich and gene sparse. Previous studies have also suggested relationships between higher order structures and repeat content and base composition, and have shown that SINE and LINE elements tend to accumulate in GC rich regions (enriched for open chromatin) and GC poor (often closed chromatin) respectively (Versteeg et al., 2003). There are also known correlations between higher order structures and divergence at the DNA sequence level. Evolutionary rates are not constant across the human genome and it is known that substitution rates are higher in closed chromatin domains than in open domains (Prendergast et al., 2007). In this chapter, we aim to look closely at the individual relationships between DNA features including base composition and the densities of different DNA repeat classes to higher order chromatin structure and structural divergence. This will allow us to assess whether these relationships change when examining structurally divergent regions. We will also examine, several measures of sequence level divergence and their relation to higher order chromatin structure. These include substitution rates, single-nucleotide polymorphism (SNP) density, insertion and deletion (indel) density, and repeat densities.

6.2. STRUCTURAL DIVERGENCE ASSOCIATED WITH DNA COMPOSITION

6.2.1. DIVERGENCE AND GC CONTENT

It has been previously shown that higher order chromatin structure shows strong positive correlations with GC content, such that relatively open regions are more GC rich and gene dense (Gilbert et al., 2004). Total GC density in orthologous, intergenic regions was examined in both species and as expected significant correlations were found, confirming the previous observation. Human GC density is higher in open chromatin (Spearman's rho = 0.57, $p < 2.2 \times 10^{-16}$), and following this trend, mouse GC density is also enriched in open chromatin (Spearman's rho = 0.75,

$p < 2.2 \times 10^{-16}$) (Figure 6.1). For the first time, using the current data, we can also examine whether GC content is associated with divergence in higher order structure. Comparison of the percentage of GC nucleotides between divergent and non-divergent regions across all orthologous 100 Kb regions shows intriguing contrasts between the human and mouse genomes (Figure 6.1). In the human genome, there is a significant shift in human GC content between divergent and non-divergent regions, across the entire spectrum of normalised chromatin structure. Furthermore, this shift is to higher GC content (40.5%) within divergent human closed regions, and lower GC content (34.9%) within divergent human open regions, relative to non-divergent regions (37.5%; human divergent open GC versus human non-divergent GC Mann-Whitney $p < 2.2 \times 10^{-16}$; human divergent closed GC versus human non-divergent GC Mann-Whitney $p < 2.2 \times 10^{-16}$). Thus, the two divergence classes show the opposite human GC content bias to the expectation (Figure 6.1. These patterns are not seen in the GC content of the mouse genome, where there is no contradictory shift in the compositional biases of mouse sequences within divergent regions. Instead mouse divergent open regions are relatively GC rich (38.7%) and divergent closed regions are relatively GC poor (33.4%), relative to non-divergent regions (35.5%). Thus overall, divergent regions are consistent with the GC content trends seen in the mouse genome, but show a complete contrast with the GC trends in the human genome. This may reflect higher human variability in GC content, however it is not possible to disentangle cause and effect using the current data. This is fundamentally because it cannot be established whether changes in GC content drive structural change or vice versa. It is also not possible to establish which of the two species has the derived or ancestral chromatin state without an outgroup. However, these observations do suggest that chromatin divergence is often associated with unusual shifts in GC content in the human lineage, which may reflect fluctuations in mutation or selection during primate evolution.

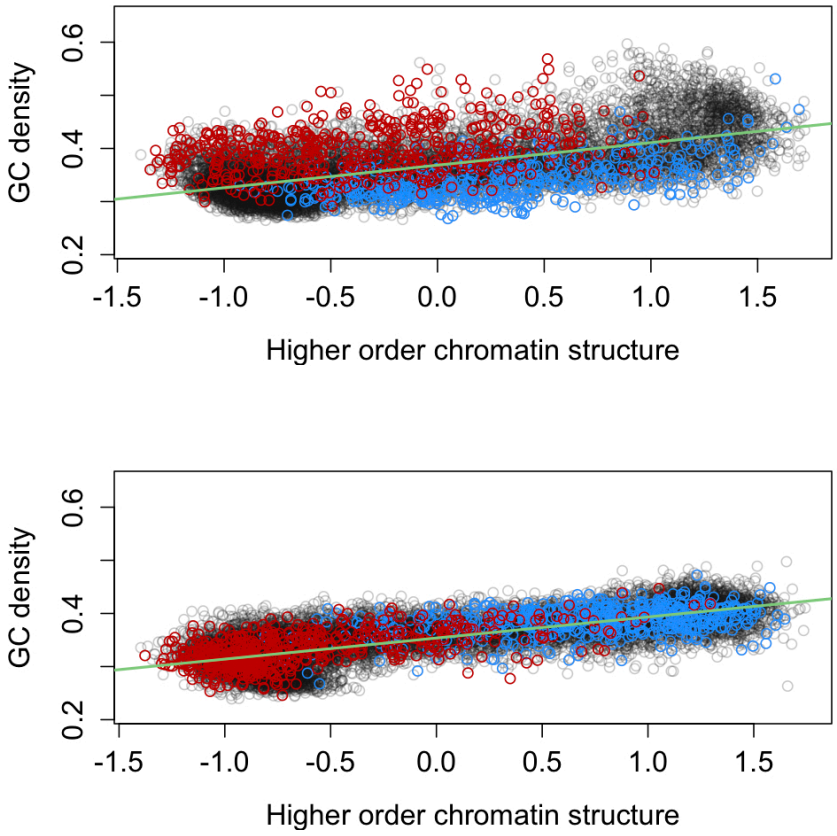

Figure 6.1 Chromatin divergence and GC content. Percentage of GC nucleotides within all 16,820 100 Kb orthologous regions across the spectrum of normalised chromatin structure value in human (top) and mouse (bottom). Three classes of regions are shown: non-divergent (grey), divergent open (blue) and divergent closed (red). The green line represents the regression line of the overall non-divergent trend.

6.2.2. DIVERGENCE AND REPEAT DENSITY

Repetitive DNA elements are widespread and thought to cover up to two thirds of the human genome (de Koning et al., 2011). With such broad coverage it is unsurprising that there are known relationships to GC content and other genomic features. One type of repeat sequence derived from transposable elements (TE), which includes long interspersed elements (LINEs) and short interspersed elements (SINEs), dominate the landscape of mammalian genomes and can affect gene expression by disrupting transcription and translation (Cordaux and Batzer, 2009).

As mentioned above previous studies have suggested a relationship between repeat content and base composition (Versteeg et al., 2003). The current data show the same trends using direct measures of higher order chromatin structure, with LINE repeat densities showing a negative correlation (Rho = -0.44 , $p < 2.2 \times 10^{-16}$), and SINE repeats a positive correlation (Rho = 0.69, $p < 2.2 \times 10^{-16}$) to both human (Figure 6.2) and mouse chromatin structure (Figure 6.3). Other major repeat classes were also assessed including DNA and long terminal repeats (LTR) which showed less consistent relationships in mouse and human chromatin structure (DNA: Rho = -0.05 Hs, 0.14 Mm $p < 9.2 \times 10^{-13}$; LTR: Rho = -0.34 Hs, -0.06 Mm $p < 4.9 \times 10^{-15}$). Simple repeats, which are repeated arrays of short runs of DNA, were examined more closely in order to establish whether particular classes of simple repeats showed a strong relationship to chromatin structure. Low complexity and simple repeat densities were found to reflect the known compositional biases of open and closed chromatin structure, with AT rich repeats showing a negative correlation to chromatin structure, reflecting the relatively AT rich nature of closed chromatin (Rho = -0.32, $p < 2.2 \times 10^{-16}$) (Figure 6.4).

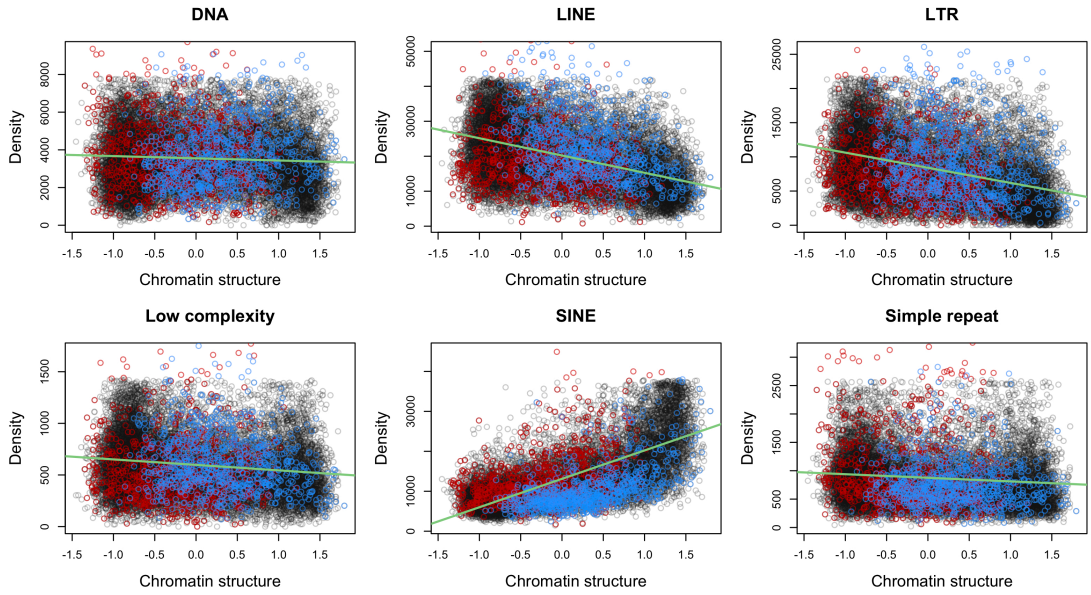

Figure 6.2 Human repeat densities (DNA, LINE, LTR, Low complexity, SINE and Simple repeat) for all orthologous 100 Kb structural regions. Divergent human open (blue) and human closed (red) regions are shown with non-divergent (black) regions. Non-divergent regression lines are shown in green.

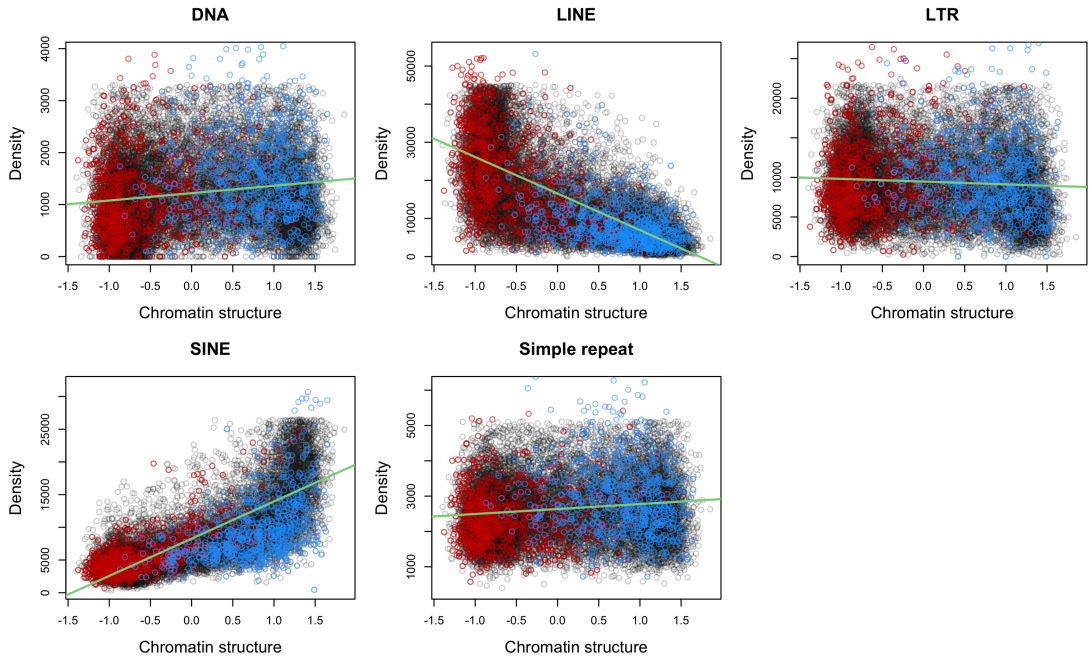

Figure 6.3 Mouse repeat densities (DNA, LINE, LTR, SINE and Simple repeat) for all orthologous 100 Kb structural regions. Divergent mouse open (blue) and mouse closed (red) regions are shown with non-divergent (black) regions. Non-divergent regression lines are shown in green.

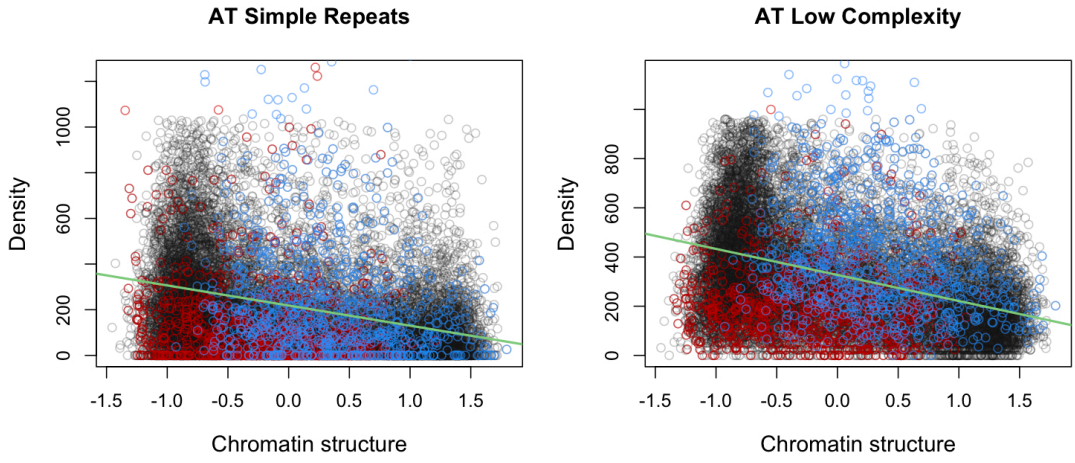

Figure 6.4 Densities of AT simple repeats and AT rich low complexity repeats in higher order chromatin structure. In each graph divergent mouse open (blue) and mouse closed (red) regions are shown with non-divergent (black) regions. Non-divergent regression lines are shown in green.

Having established the relationship between base composition, repeat

density and chromatin structure we went on to look at the relationship between these elements and structural divergence. Three classes of repeat were found to show generalised shifts (across the spectrum of higher order structure) in their densities between divergent and non-divergent regions: LINE and SINE elements, and particularly AT rich low complexity regions. Usually these shifts are seen in human repeat densities and are less clear or absent in mouse repeat data. As for DNA composition, repeat densities show shifts in divergent regions that are the opposite of the expected trend, given the overall relationships of these repeat classes to chromatin structure. For example, human SINE elements are enriched in open chromatin but relatively depleted in divergent open chromatin (Figure 6.2, Figure 6.5) (Mann-Whitney $p < 2.2 \times 10^{-16}$). However this discrepancy is not seen in mouse where divergent regions more closely follow the expected trend (Figure 6.5). As with GC content, this apparent shift in repeat densities within human divergent regions cannot yet be reliably identified as a cause or effect of structural divergence.

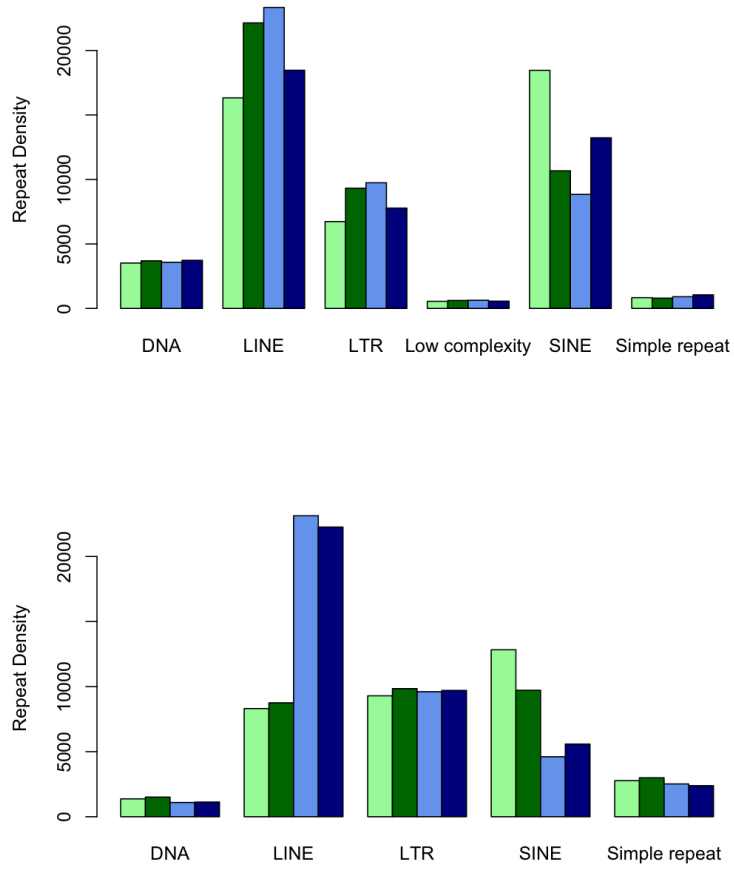

Figure 6.5 Mean repeat densities for each of the major repeat classes in human (top) and mouse (bottom). The four separate groups represent average repeat densities for non-divergent open structure (light green), divergent open structure (dark green), non-divergent closed structure (light blue) and divergent closed structure (dark blue). (Low complexity DNA data is not available in mouse UCSC RepeatMasker annotation.)

6.3. CHROMATIN STRUCTURE IS CORRELATED WITH SEQUENCE DIVERGENCE

The current data allow for a thorough investigation of the coupling between higher order chromatin structure divergence and divergence at the DNA level. It is has been shown that human-chimpanzee substitution rates are not constant across the genome, and are significantly correlated with higher order chromatin structure

(Prendergast et al., 2007), but other classes of mutation are not well studied. Pairwise sequence alignments of human and mouse were used to assess the substitution rate of intergenic regions (see Methodology) within each 100 Kb orthologous structural region (Figure 6.6). There was a significant negative correlation between human-mouse substitution rate and chromatin structure (Rho = -0.45, $p < 2.2 \times 10^{-16}$). This reaffirms the relationship seen between human-chimpanzee substitution rate and human chromatin structure albeit using much lower resolution structural data and a smaller dataset of substitutions (Prendergast et al., 2007).

The density of single nucleotide polymorphisms (SNPs) per structural region was calculated using data from the 1000 Genomes Project (Abecasis et al., 2012) and showed a similar relationship to chromatin structure (Rho = -0.33, $p < 2.2 \times 10^{-16}$), such that divergence is highest in relatively closed regions.

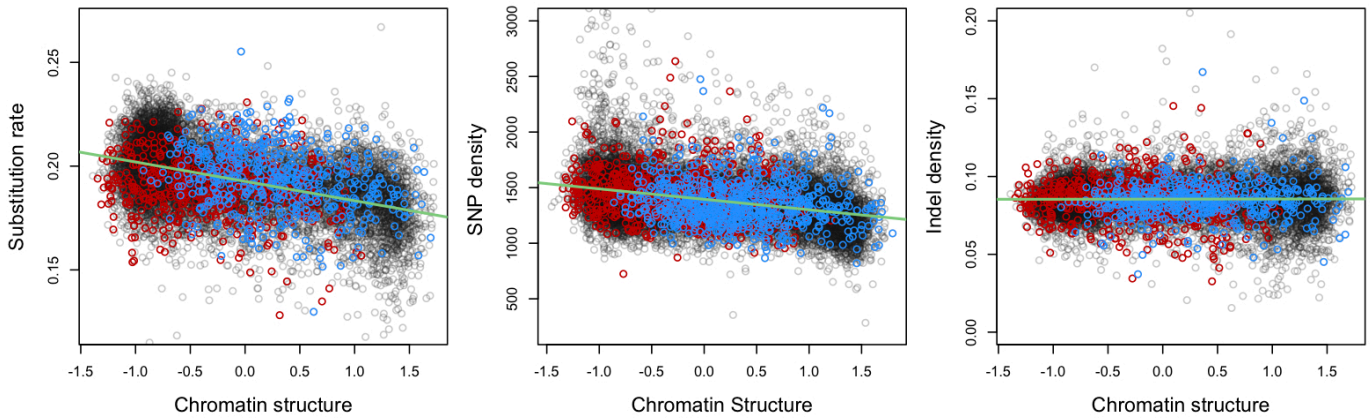

Figure 6.6 Chromatin structure and genomic sequence divergence. Intergenic substitution rates, SNP densities and indel densities are displayed for all orthologous human 100 Kb structural regions. In each graph divergent human open (blue) and human closed (red) divergent regions are shown with non-divergent (black) regions. Non-divergent regression lines are shown in green

The fact that substitution rate and SNP density show the same relationship to higher order chromatin suggests that similar mutational biases have existed across 100 million years of evolution and still operate in current human populations. We also examined 1,443 loss of function (LOF) SNPs, also from the 1000 Genomes Project, and higher order structure, which induce a stop, splice-site disruption or frame shift where they occur in the genome. Due to the low number of LOF SNPs

we averaged groups of regions by class (openness) of higher order chromatin to calculate the proportion of regions containing a LOF SNP per class. It was found that LOF SNPs had an opposing relationship to chromatin structure with most occurring in open structures (Rho = 0.8, p <4 x 10^{-34}), consistent with the higher numbers of protein coding genes in open chromatin. However, there was no clear relationship to structural divergence.

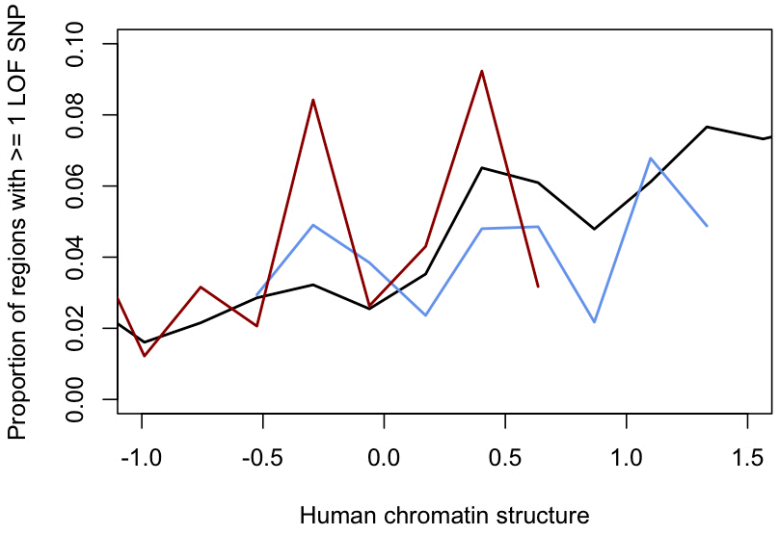

Figure 6.7 Proportion of regions containing at least one loss of function SNP (LOF) across different classes of human chromatin structure (Rho= 0.8, p <4 x 10^{-34}). Average proportions across divergent human open (blue) and human closed (red) regions are shown with non-divergent (black) regions.

Indel densities were calculated using UCSC derived whole genome alignments of human, mouse and dog. Lineage specific human and mouse indel events were inferred using the dog as an outgroup (see Methods) to calculate indel densities (number of indel bp per aligned intergenic human bp). Total indel density (total number of bases involved in insertions of deletions per aligned intergenic bp) across chromatin structure (Figure 6.6) showed no significant correlation to chromatin structure, in contrast to previous results for substitution rates. Within lineage specific indels, mouse deletions were the most prevalent type of indel event and showed a modest positive correlation to chromatin structure correlation (Rho = 0.14, p < 2.2 x 10^{-16}) (Figure 6.8). Interestingly, this positive relationship to open

chromatin structure is not apparent when comparing the frequency of mouse deletion events, rather than the proportion of deleted bases (Rho = -0.08, $p < 2.2 \times 10^{-16}$). This indicates that the number of deleted mouse bases per deletion event tends to be larger in open chromatin but smaller in closed. As expected, human indels were found to be an order of magnitude less frequent than mouse indels (mean proportion mouse indel – 0.038, mean proportion human indel – 0.004). Both mouse and human insertion rates were less frequent than deletion rates (0.003 to 0.004 human and 0.03 to 0.05 mouse) and showed negative relationships to chromatin structure as seen in the other divergence estimates (Rho = -0.15 human insertion, Rho = -0.08 mouse, $p < 2.2 \times 10^{-16}$)(Figure 6.8). However, the proportions were so infrequent that it is difficult to draw meaningful conclusions.

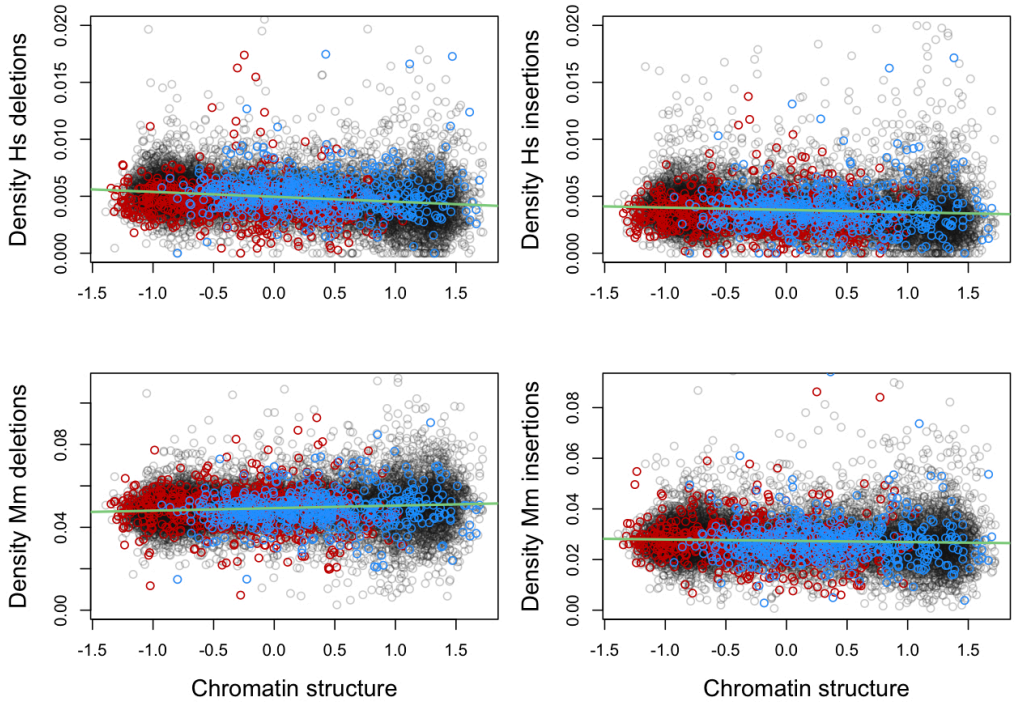

Figure 6.8 Linage specific indel densities (human deletions, human insertions, mouse deletions and mouse insertions) across higher order chromatin structure. In each graph divergent human open (blue) and human closed (red) regions are shown with non-divergent (black) regions. Non-divergent regression lines shown in green

If substitution rates, SNP densities and indel rates are important correlates of structural divergence we might expect to see a generalised shift in these rates within regions of divergent structure, relative to the expected rates seen in the non-

divergent regions. However this is not the case, instead the divergent regions show trends that are very close to the trends in non-divergent regions. Thus, although substitution, SNP and specific indel rates are significantly correlated with higher order chromatin structure they are not consistently associated with structural divergence. It follows that they are also unlikely to be major determinants of structural divergence.

6.4. SEGMENTAL DUPLICATIONS AND CONSERVATION OF SYNTENY

The mammalian genome contains numerous segmental duplications: blocks of homologous duplicated sequences that are greater than 1 Kb and map to multiple loci, sharing at least 90% sequence homology. They often contain low copy repeat sequences that can cause regions of genomic instability associated with copy number differences (Kim et al., 2008). Again, due to the relatively low number of segmental duplications, proportion of regions containing a duplication per chromatin class was used and there is a strong positive correlation to chromatin structure in both species (Rho = 0.94 in human and 0.92 in mouse $p < 3.7 \times 10^{-6}$). This indicates that there are higher densities of duplicated segments in more open chromatin domains. Human divergent regions showed no significant deviations from this trend, however this was not the case for the mouse genome. Mouse open divergent regions showed a significant ($p < 0.004$) depletion in segmental duplications (Figure 6.9).

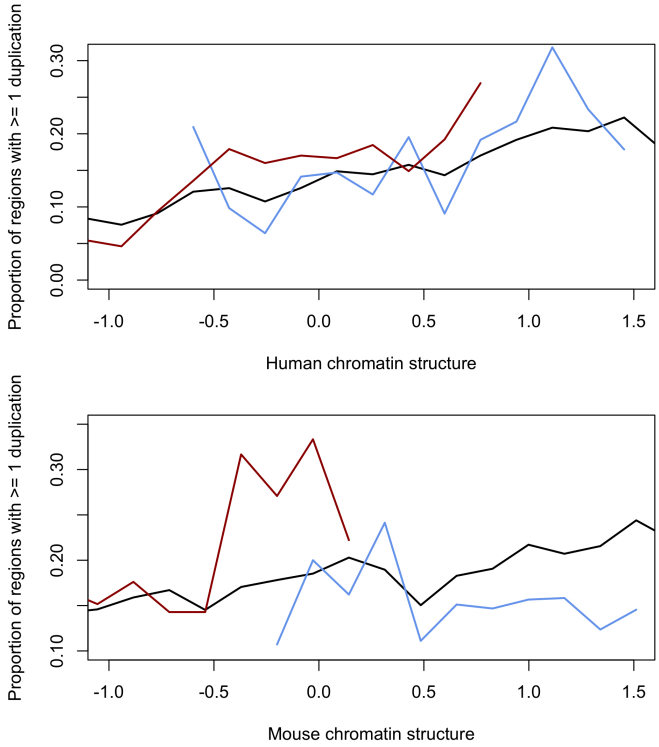

Figure 6.9 Proportion of regions containing at least one segmental duplication across human (top) and mouse (bottom) chromatin structure (Rho = 0.91 in human and 0.69 in mouse $p < 10^{-16}$). Proportions are shown across divergent human open (blue) and human closed (red) regions are shown with non-divergent (black) regions.

Segmental duplications are associated with large-scale genomic rearrangements between species, which is likely to affect patterns of sequence orthology between human and mouse. The number of regions containing a segmental duplication was compared between the orthologous 16,820 regions and the non-orthologous, lineage specific, 100 Kb regions (defined in Chapter 3) from both species. It was found that the orthologous dataset contained substantially fewer segmental duplications than the non-orthologous data (14% of orthologus regions and 27% non orthologus for human segmental duplications, 18% of orthologus regions and 40% non orthologus for mouse segmental duplications $p < 2.2 \times 10^{-16}$ (Figure 6.10). Thus it seems that segmental duplications play a more important role in sequence divergence than in structural divergence between these species.

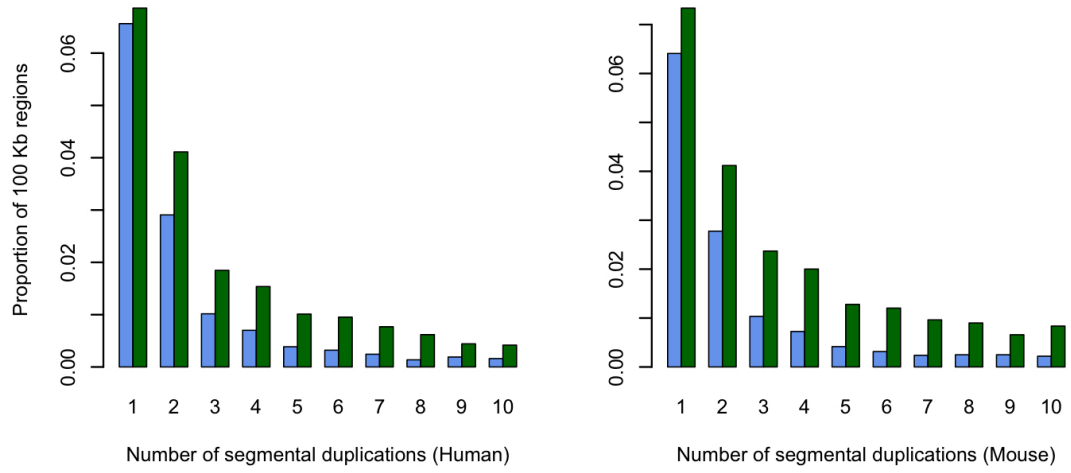

Figure 6.10 Segmental duplications in non-orthologus and orthologous regions. Proportion of regions containing different numbers of segmental duplications in human (left) and mouse (right) orthologous 16,800 regions (blue) and non-orthologous structural regions (green).

The human and mouse genomes contain many regions showing the same ordering of orthologous genes along chromosomes, that is, conservation of synteny. During evolution, genome rearrangements may separate two loci, resulting in the disruption of synteny (synteny breaks). To investigate whether chromatin structure conservation patterns are related to synteny breaks, top level synteny breaks (the primary level of alignments between human and mouse, see Chapter 2) were obtained from the Ensembl Compara database. It was found that less than 1% of all regions contained a top level synteny break. To ascertain whether synteny breaks were more prevalent in the divergent chromatin regions, chi squared tests were done on the orthologous divergent classes in both human and mouse chromatin structure. There was no significant enrichment or depletion of synteny breaks in orthologous human and mouse divergent regions. This was not surprising as synteny breaks are, by definition, non-orthologous. We then went on to look at the lineage specific human and mouse non-orthologous regions (see Chapter 3) that were not included in the 16,820 region dataset and found a marked contrast. In both species there were enrichments with 5-6% of non-orthologous regions containing a top level synteny break (p-value 9.5×10^{-11} human and 6.2×10^{-09} mouse respectively).

We conclude that higher order chromatin structure itself is associated with many aspects of the underlying DNA sequence, including measures of sequence

divergence, which complicates attempts to identify which features are significantly associated with structural divergence. The challenge is to unpick the associations due to chromatin structure itself from those of structural divergence. The approach we have adopted is to examine each sequence level correlate between divergent and non-divergent structures across the entire spectrum of observed chromatin structure. Shifts in a sequence-based variable between divergent and non-divergent classes that are seen consistently implicate that variable in structural divergence. This thinking is formalised in multiple regression analyses in Chapter 8, which also allow us to distinguish the most important combinations of variables among the many possibilities.

Chapter 7

Results: Comparative investigation of locus level chromatin

Topics included in this section:

- Construction of a three species locus level chromatin feature dataset containing histone modification, methylation and transcription factor data in human, mouse and pig.
- Analysis of the relationships among the chromatin features across each species and a comparison to the findings in Xiao et al 2012.
- Further examination of read density patterns within each species and clustering of chromatin features by type, for example, promoter associated features.
- Examination of suitability of the data for applying a divergence metric.
- Comparisons between the locus level feature data to the previous 100 Kb higher order chromatin structure data.

7.1. INTRODUCTION

One of the main caveats of the higher order chromatin structure divergence detected is that due to data availability the analysis is restricted to just two species. This makes it impossible to detect which lineage possesses the ancestral or derived structural state when examining structural divergence. The emergence of a new three species dataset containing chromatin features in human, mouse and pig (Xiao et al., 2012) allows such comparisons to be made, though for aspects of chromatin structure not examined thus far in this thesis. Rather than higher order structures the data concern locus level features such as histone modification and DNA methylation patterns. This dataset also enables a three-way examination of chromatin features in the structurally divergent regions examined so far. Key questions can be addressed about how the many different levels and facets of chromatin structure are related to one another, and also the mechanisms underlying divergence in higher order structure.

The study by Xiao et al (2012) has provided matched data for human and mouse ES cell locus level chromatin together with complementary data for an outgroup species, pig, in pluripotent stem cells. The main focus of their study was on the epigenomic conservation of chromatin features across different classes of genomic regions. They also found that conserved colocalisation of different epigenomic marks within a region can be used to discover regulatory sequences and concluded that comparative epigenomics may reveal regulatory features of the genomes under study.

Although large divergent regions of mammalian chromatin structure were reliably identified in earlier chapters, the mechanisms underlying divergence remain unknown. There are known to be strong correlations between the extent of histone modification patterns and other lower level features and the variations seen at the level of higher order structures (Zhou et al., 2011). In this chapter, the importance of a large number of lower level structural level variables in the divergence of higher order structure during mammalian evolution is directly assessed. The main dependencies between different structural levels across the genome and their relative importance are also examined.

7.2. OVERVIEW OF LOCUS LEVEL CHROMATIN DATA

Genome wide human mouse and pig data for 14 different chromatin features

was available in human H1 ESCs, mouse E14 ESCs and pig pIPC cells (Lister et al 2009; Bernstein et al, 2010, Xiao et al, 2012; ENCODE, 2012, Goren et al, 2010). These included the histone modifications H3K27ac, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3 H3K36me3 and the histone variant H2AZ. There were also ChIP-seq transcription factor binding site datasets for TAF1, OCT4, p300 and NANOG and finally DNA methylation datasets generated using MeDIP-seq and MRE-seq (Table 7.1).

Feature	Mark	Method	Human	Mouse	Pig
Repression	H3K27me3	ChIP-seq	Lister et al, 2009	Xiao et al, 2012	Xiao et al, 2012
	H3K9me3	ChIP-seq	Lister et al, 2009	Goren et al, 2010	Xiao et al, 2012
Enhancer	H3K4me1	ChIP-seq	Lister et al, 2009	Xiao et al, 2012	Xiao et al, 2012
	H3K4me2	ChIP-seq	Lister et al, 2009	Xiao et al, 2012	Xiao et al, 2012
	H3K27ac	ChIP-seq	Lister et al, 2009	Xiao et al, 2012	Xiao et al, 2012
Promoter	H3K4me3	ChIP-seq	Lister et al, 2009	Xiao et al, 2012	Xiao et al, 2012
Gene body	H3K36me3	ChIP-seq	Lister et al, 2009	Xiao et al, 2012	Xiao et al, 2012
Promoter	H2AZ	ChIP-seq	Xiao et al, 2012	Xiao et al, 2012	Xiao et al, 2012
Methylation		MeDIP-seq	Bernstein et al, 2010	Xiao et al, 2012	Xiao et al, 2012
		MRE-seq	Bernstein et al, 2010	Xiao et al, 2012	Xiao et al, 2012
Promoter	TAF1	ChIP-seq	ENCODE, 2012	Xiao et al, 2012	Xiao et al, 2012
Enhancer	P300	ChIP-seq	ENCODE, 2012	Chen et al, 2008	Xiao et al, 2012
Pluripotency	OCT4	ChIP-seq	ENCODE, 2012	Chen et al, 2008	Xiao et al, 2012
	NANOG	ChiP-seq	ENCODE, 2012	Chen et al, 2008	Xiao et al, 2012

Table 7.1 Descriptions and origins of each chromatin feature dataset similar to the data produced by Xiao et al (2012).

The raw sequence data archives for each chromatin feature in each species were downloaded from the NCBI Sequence Read Archive (SRA), converted to FASTQ format and examined for sequence read quality using FastQC. The FastQC reports for the human transcription factor datasets showed poor read mapping (<40% mapped) so alternative data was sought from ENCODE (The ENCODE Project Consortium, 2011) that showed better quality alignment (>70% mapped). The total numbers of reads and percentages of reads mapped is shown in Table 7.2.

	Human (GRCh37/hg19)		Mouse (NCBI37/mm9)		Pig (SGSCSscrofa9.2/susScr2)	
Feature	**Reads**	**% mapped**	**Reads**	**% mapped**	**Reads**	**% mapped**
H3K27me3	91,076,733	70.87	31,654,344	82.84	10,868,857	76.41
H3K9me3	193,583,168	78.64	23,895,406	93.29	13,927,780	69.76
H3K4me1	69,962,590	88.80	19,366,374	90.92	20,705,775	79.30
H3K4me2	38,030,958	73.62	18,351,814	92.78	4,322,782	79.50
H3K27ac	42,382,986	81.70	6,911,600	92.47	17,741,434	79.26
H3K4me3	66,500,991	77.25	31,258,590	92.07	21,512,279	76.91
H3K36me3	141,190,603	79.29	17,641,837	88.48	38,013,146	76.28
H2AZ	25,786,431	95.90	88,311,720	78.94	4,417,656	78.26
MeDIP-seq	35,182,811	87.49	25,728,829	96.16	68,357,687	68.92
MRE-seq	77,386,795	75.94	27,736,348	90.15	16,832,623	33.93
TAF1	32,189,724	79.21	23,450,889	69.41	10,700,034	77.63
P300	53,920,750	68.22	17,413,416	72.47	33,455,621	77.98
OCT4	46,880,412	87.78	11,785,618	41.77	5,074,592	77.59
NANOG	53,920,750	68.22	12,932,668	88.87	18,575,267	77.15

Table 7.2 Numbers of sequence reads acquired for each chromatin feature in each species. Also shown is the percentage of reads that were successfully mapped to the reference genome.

The reads were mapped to GRCh37/hg19, NCBI37/mm9 and SGSC Sscrofa9.2/susScr2 genome assemblies using Bowtie (Langmead et al., 2009). The mapped reads were converted to bedgraph files using BEDtools (Quinlan and Hall, 2010), which are designed for displaying continuous data. All data was binned to 100 Kb resolution and allow comparisons between locus level chromatin features and the higher order structure data described in earlier chapters. Read densities were averaged within the genomic intervals of the consecutive 100 Kb bins used for higher order data. This resulted in three separate 100 Kb genome wide chromatin feature sets, one for each species with 28,796 regions in the human dataset, 25,684 in the mouse and 22,637 in pig. UCSC liftOver (Kent et al., 2002) and Perl scripts were then used to reciprocally map orthologous regions between the different species. Regions that did not correctly map reciprocally (i.e. forwards and backwards) between species, or that substantially changed in size (<80% or >120% of the original region) when re-mapped were discarded. The resulting orthologous dataset contained 8,900 100 Kb regions in human, mouse and pig genomes. Quantile normalisation was imposed across all datasets as for previous analyses.

7.3. WIDESPREAD DIVERGENCE OF MAMMALIAN LOCUS LEVEL CHROMATIN

Xiao et al (2012) found that locus level features of chromatin were detectably conserved between human, mouse and pig within promoters and exons, however, they reported that the levels of conservation were more variable within intergenic and intronic regions (i.e. the vast majority of the genome), suggesting that these features may be more modestly conserved on a genome wide scale. In addition, the definition of conservation was limited to a binary measure: whether or not the same chromatin feature was found within each orthologous 200 bp sequence between species. This ignores the wide variation in read density across the genome (Figure 1.1). To address this, a global correlation matrix was constructed for all orthologous 100 Kb chromatin features to examine the levels of conservation according to read densities. The data was hierarchically clustered by similarity of read densities and the resulting heatmap showed a wide range of Spearman's Rho values ($-0.25 < Rho < 0.88$) (Figure 7.1).

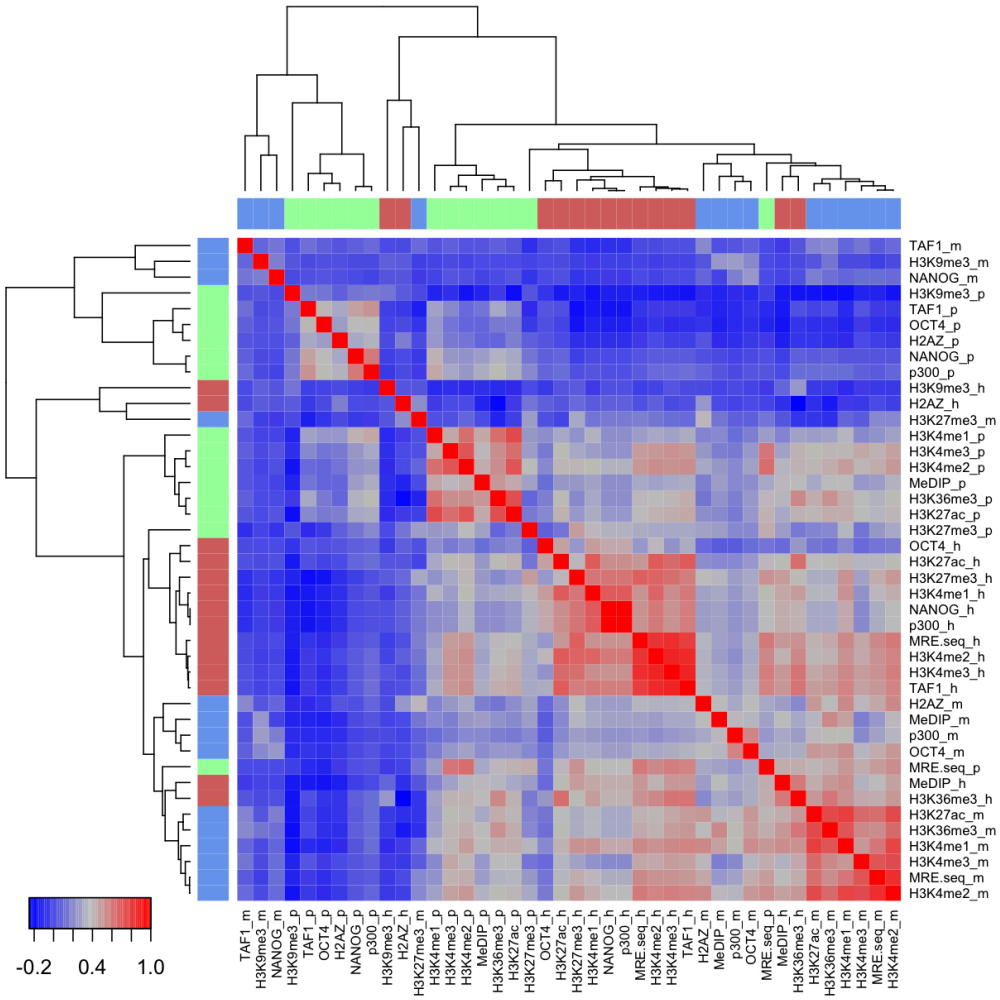

Figure 7.1 Genome wide correlation matrix of all 8,900 orthologous chromatin regions. The datasets are hierarchically clustered by similarity of genome-wide read densities. Each chromatin feature is labelled with h, m and p representing human (red), mouse (blue) and pig (green) datasets respectively. Red colours indicate positive correlation (Rho) scores and blue indicate negative.

The correlation matrix shows many, small species-specific subclusters where correlations were as high as Rho = 0.88, $p < 2.2 \times 10^{-16}$. Apart from these subclusters, the correlations were varied (Rho min -0.25, max 0.54) between chromatin features of different species but usually indicated modest (mean Rho = 0.21, $p < 2.2 \times 10^{-16}$) overall conservation at the genome-wide level. These observations are not necessarily in disagreement with Xiao et al (2012), since it is likely that the 100 Kb resolution masks locus level patterns across the data. However it is clear that the

widespread conservation seen for higher order chromatin structure is not matched by similar patterns of conservation in locus level chromatin features.

To further assess the degree of clustering within species, two separate correlation matrices were produced for the chromatin feature data in human and mouse. Mean higher order chromatin structure data for each species, (defined in Chapter 3) was included with the current data and all datasets were again hierarchically clustered according to genome-wide correlations (Figure 7.2). In both human and mouse, higher order chromatin structure showed a varying degree of correlation to the lower-order chromatin features (Rho min -0.06, max 0.82). The most highly correlated subcluster contained transcription associated histone modifications such as H3K4me2 and H3K4me3 in both species. The repressive histone modification H3K9me3 also, showed the strongest negative correlations with other chromatin features in both species (Rho -0.25 human, -0.10 mouse). It appears that while there is some degree of similarity between the two correlation matrices the patterns of clusters and inter-relationships within each species are often distinct. Histone modifications of the same classification did not show a propensity to cluster together as expected, probably due to insufficient resolution.

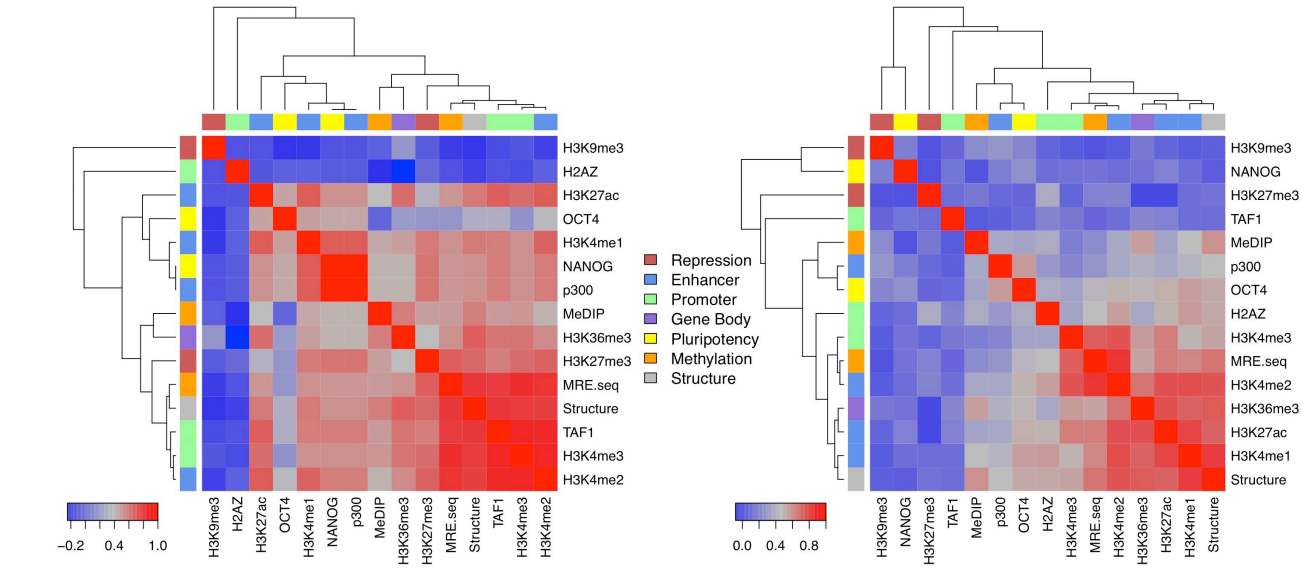

Figure 7.2 Correlation matrices for the locus level chromatin data and mean higher order structural data for human (left) and mouse (right). The coloured bar indicates the chromatin feature classification type taken from Xiao et al (2012). Red colours indicate positive correlation (Rho) scores and blue indicate negative.

7.4. LOCUS LEVEL CHROMATIN COMPARED TO HIGHER ORDER STRUCTURE

The common 100 Kb resolution of all datasets allowed the levels of histone modifications and transcription factors to be assessed across different classes of higher order chromatin structure and structural divergence in both human and mouse (Figure 7.3, Figure 7.4).

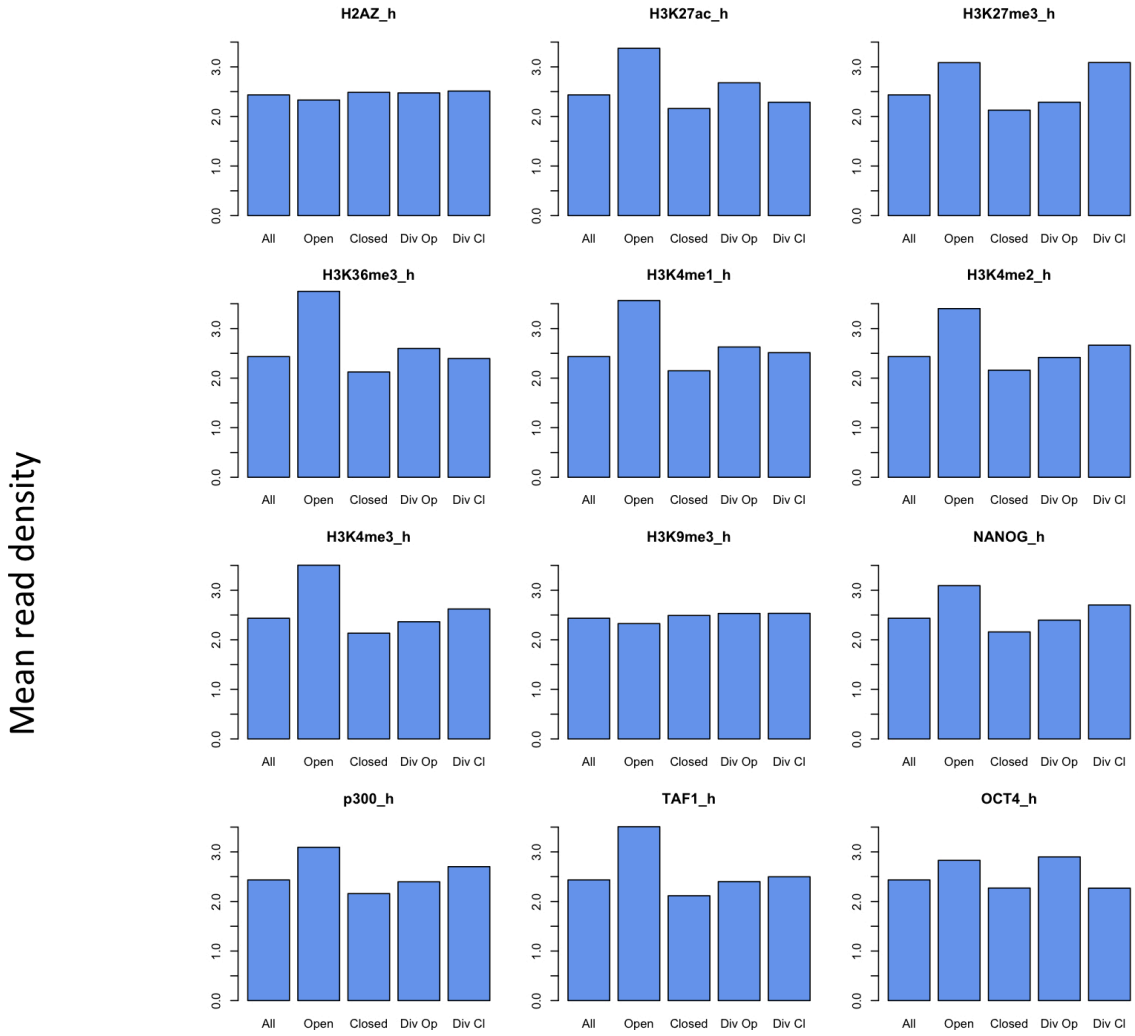

Figure 7.3 Median read densities of human histone modification and transcription factors across: All 100 Kb orthologous regions, Open chromatin structure (positive chromatin values), Closed chromatin structure (negative chromatin values), Open divergent chromatin structure, and Closed divergent chromatin structure.

It was found that some locus level chromatin features (H3K9me3, H2AZ in human, H3K9me3, NANOG, TAF1 in mouse) showed relatively little change in intensity across different classes of higher order chromatin structure and structural divergence but in others change was more pronounced (Figure 7.3, Figure 7.4). However, most showed an increase in read density within mean open chromatin structure relative to mean closed chromatin structure in both species. Given the histone modifications involved, this appears to be a result of the fact that genes and their regulatory regions are more often found in accessible (open) regions of chromatin (Black and Whetstine, 2011). This relationship was also reflected in open and closed divergent structural classes with few exceptions. Human H3K27me3, a repressive mark, was enriched in open chromatin structure but also enriched in closed divergent structure (Mann-Whitney $p < 2.2 \times 10^{-16}$), a relationship that was not reflected in the mouse dataset. H3K27me3 is known to be important for the genomic targeting of the PRC2 polycomb complex to its target gene promoters including hox gene clusters which are enriched in repressive regions (Eskeland et al., 2010).

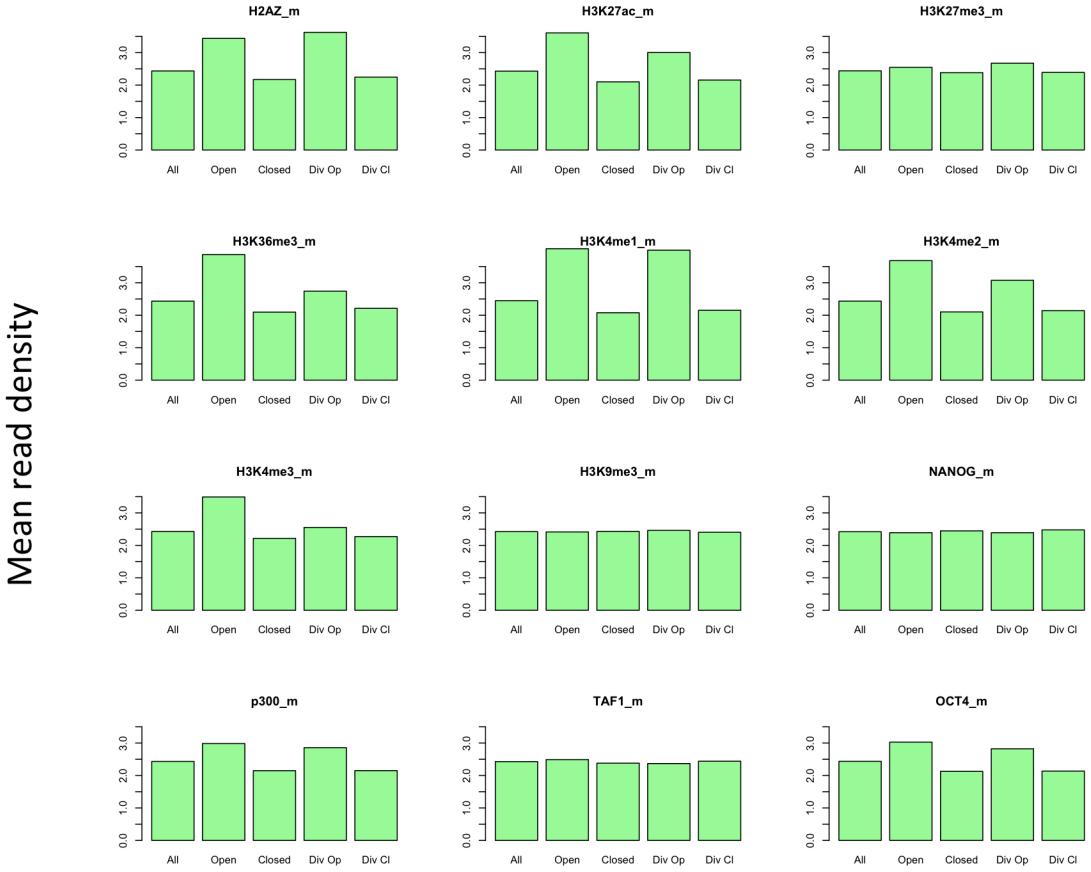

Figure 7.4 Median read densities of mouse histone modification and transcription factors across: All 100 Kb orthologous regions, Open chromatin structure (positive chromatin values), Closed chromatin structure (negative chromatin values), Open divergent chromatin structure, and Closed divergent chromatin structure.

These data provide a basis to quantitatively explore the mechanisms underlying structural divergence in higher order chromatin. To investigate the most important factors in higher order structural divergence a regression modelling approach was chosen, to account for the inter-correlated nature of locus level chromatin features discussed above. This is discussed further in Chapter 8.

Chapter 8

Results: Multiple regression modelling of chromatin structure

Topics included in this section:

- Multiple regression modelling of DNA sequence features and locus level chromatin features to predict higher order structure.
- Multiple regression modelling of DNA sequence features and locus level chromatin features to predict higher order structural divergence.
- The addition of DNA sequence divergence to models for the prediction of higher order structural divergence.

8.1. INTRODUCTION

A striking result of the analyses performed so far is that the majority of sequence level features examined show significant correlations with higher order structure. Some of these features also show intriguing shifts in these correlation patterns when comparing structurally divergent and non-divergent regions (see Chapter 6). These results suggest that structural divergence is associated with regional shifts in repeat content and compositional bias, but the overall picture remains unclear. These associations vary with the polarity of the structural divergence and it is possible that other sequence features or forms of sequence change play more minor roles. In addition, many sequence level features of the genome (including compositional bias and repeat densities) are known to be interdependent. In order to tease apart the most influential factors for higher order chromatin structure and higher order divergence between human and mouse, a multiple regression modelling approach was used. This allows us to understand the importance of each variable in situations where many variables are expected to interact to determine a dependent variable, in this case higher order structure. Also, redundancy and independence of the different sequence features can be explored and quantified. In this chapter, several multiple linear regression models are examined. For the first time we are able to combine all sequence variables studied so far and quantitatively describe the extent to which they individually and collectively explain higher order structure. We also examine the extent to which the same variables can explain higher order chromatin divergence. Lastly, we build a regression model that also includes measures of sequence divergence to explain higher order divergence, to explicitly examine the coupling between chromatin and sequence divergence during mammalian evolution.

8.2. LINEAR MODELLING OF HIGHER ORDER CHROMATIN STRUCTURE

We initially examined multiple regression models aiming to predict the normalised higher order chromatin structure measurements themselves, to discover how well sequence level variables predict structure. Linear models were created by collating all suitable factors investigated so far and using standard multiple linear regression analysis in R. Stepwise optimisation for the best models was performed according to the Akaike Information Criterion (AIC). The aim was to find the

chromatin variables with the most influence on both higher order chromatin structure and also to calculate to what extent the variables could explain chromatin structure. The AIC was used to optimise the model and identify successful combinations of variables. The R package stepAIC (Venables and Ripley, 2002) was used which works by adding and removing variables in a stepwise manner within the model to determine which combination most successfully explains the response variable. The AIC takes into account redundant or missing data in the model and indicates how well the model fits the data. The main caveat of using such optimisation methods is that where two variables are mutually dependant it can be difficult to determine which ends up in the optimised model. Q-Q plots of the standardised residuals were used to assess the suitability of the combined factors for modelling higher order structure. The linear nature of the plots demonstrated that linear relationships were present between the sequence variables and higher order structure, validating the approach (Figure 8.1).

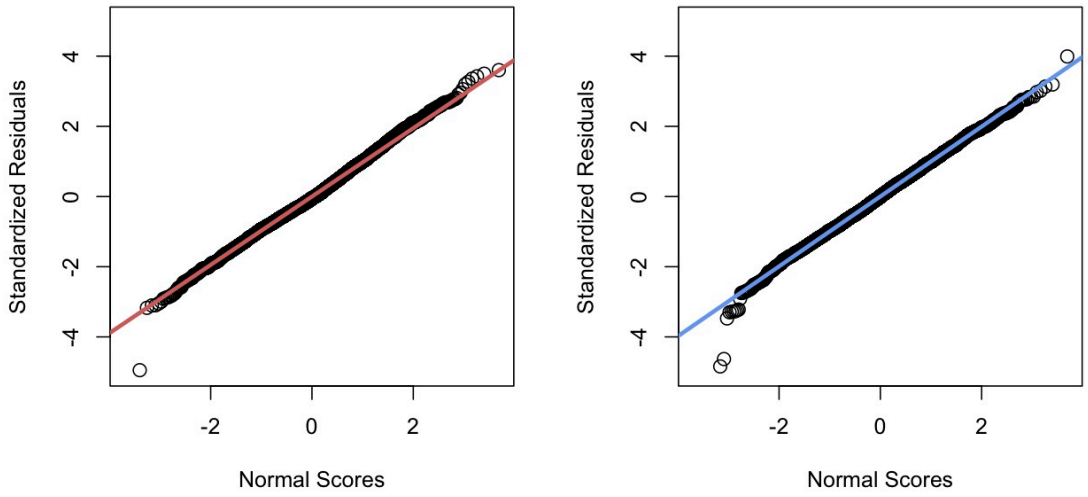

Figure 8.1 Factors affecting chromatin structure for both human and mouse models. The Q-Q plot displays the normality of the residuals in human (red) and mouse (blue). The linearity of the points suggests that the residuals are normally distributed and therefore suitable for multiple linear regression modelling.

The results suggest that linear models are reasonably successful in explaining structural variation (overall adjusted R-squared: 0.63 for human model, 0.77 for mouse) with a large number of variables contributing generally modest explanatory power (Table 8.1). Most prominent among these variables in both species were SINE element density (individual standardised R-squared: 0.252 in

human, 0.317 in mouse), gene density (R-squared: 0.111 in human, 0.097 in mouse), and GC content (R-squared: 0.138 in human, 0.229 in mouse), while other factors were somewhat less influential (Table 8.1). These data extend previous analyses (Chapter 6) showing strong pairwise correlations between chromatin structure and sequence features, and suggest that repeat content (particularly SINE repeat densities) and GC content have strong, independent associations with structure. The higher adjusted R-squared value for the mouse chromatin model than in the human model suggests that these features provide a better explanation for mouse chromatin structure than for human. This seems to be particularly driven by the higher contributions of SINE elements and GC content in the mouse model. In contrast, other variables such as gene density have a similar influence in both models.

	Feature	Coefficients	R squared
Human	SINE	3.76E-05	0.25
	Genes	5.47E-01	0.11
	GC	3.51E+00	0.14
	H3K27ac	1.66E-02	0.03
	Substitutions	-4.04E+00	0.04
	H2AZ	1.45E-02	0.01
	H3K36me3	2.20E-02	0.04
	SNP	-2.52E-02	0.02
	H3K4me1	1.01E-02	0.02
	Simple	-7.52E-05	0.00
Mouse	SINE	5.73E-05	0.32
	GC	5.80E+00	0.23
	Genes	5.16E-01	0.10
	H2AZ	3.88E-02	0.06
	LTR	1.30E-05	0.00
	Deletions	8.01E+00	0.02
	Simple repeat	-5.20E-05	-0.01
	H3K36me3	-1.04E-02	-0.01
	LINE	-3.00E-06	0.03
	MRE.seq	7.68E-03	0.01

Table 8.1 Regression coefficients and standardised r-squared values for optimised models of human and mouse higher order chromatin structure. R squared values indicate how well each chromatin feature describes chromatin structure. If a chromatin feature is absent in either species, its influence on the linear model was negligible

8.3. LINEAR MODELLING OF HIGHER ORDER CHROMATIN DIVERGENCE

Attempts to model divergence in higher order chromatin structure in the same way as chromatin structure itself were substantially less successful (overall adjusted R-squared: 0.10 in human, 0.08 in mouse). Examination of regression residuals displayed less linearity in the standardised residuals, particularly in the human model, compared to the model for chromatin structure (Figure 8.2). This indicates that the residuals contain structure that is not accounted for in the model and therefore that there is less suitability within the data for linear modelling

(Figure 8.2). The slight "S" shaped curve shown for the residuals of the model containing the divergent chromatin data indicate a bimodal distribution of residuals.

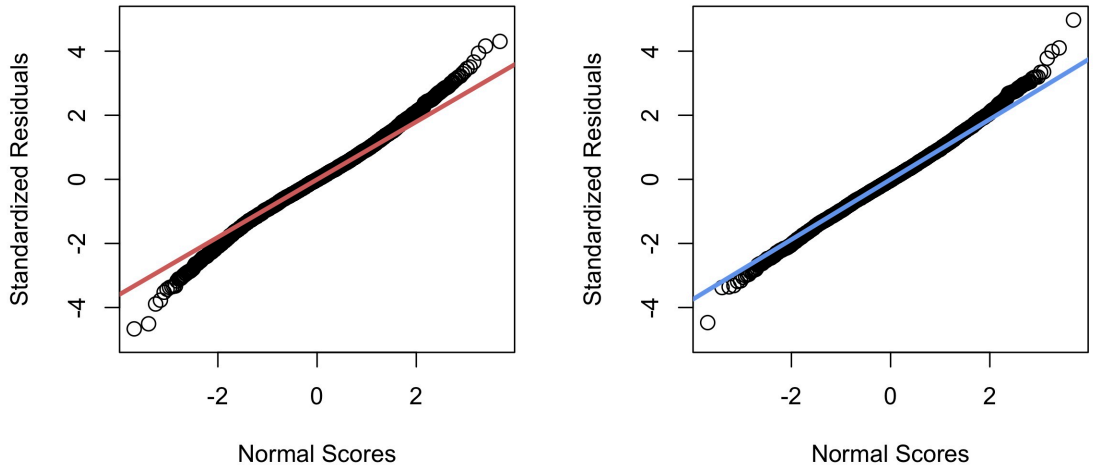

Figure 8.2 Factors affecting chromatin structure divergence for both human and mouse models. The Q-Q plot displays the normality of the residuals in human (red) and mouse (blue).

The most influential explanatory variable for modelling chromatin divergence was GC content in both species (individual standardised R-squared: 0.06 in human, 0.05 in mouse). However, it seems that there are no strong sequence level predictors of chromatin divergence among these variables. This is unsurprising given the low explanatory power of the overall model.

	Feature	Coefficients	R squared
Human	GC	-1.20E+01	0.06
	H3K27ac	8.50E-02	0.01
	H3K4me1	3.30E-02	0.00
	SNP	-4.55E-02	0.00
	OCT4	2.34E-02	0.00
	Simple	-1.47E-04	0.00
	DNA	-5.55E-05	0.00
	H3K4me3	-3.60E-02	0.00
	H3K36me3	3.74E-02	0.00
	H2AZ	2.45E-02	0.00
Mouse	GC	-9.45E+00	0.05
	Simple repeat	-1.61E-04	0.01
	Substitutions	-1.45E+01	0.01
	LINE	2.44E-05	0.03
	Gene	4.06E-01	0.00
	H2AZ	-7.23E-02	0.01
	H3K4me2	7.58E-02	0.00
	H3K27ac	-7.10E-02	0.01
	H3K36me3	4.95E-02	0.00
	MeDIP	-3.67E-02	0.00

Table 8.2 Regression coefficients and standardised r-squared values for optimised models of human and mouse structural divergence. R squared values indicate how well each chromatin feature describes chromatin divergence. If a chromatin feature is absent in either species, its influence on the linear model was negligible

8.4. MODELLING STRUCTURAL DIVERGENCE IN TERMS OF SEQUENCE DIVERGENCE

For the final regression model, examining the correspondence between divergence at the levels of sequence and chromatin, a new dataset of differences between many sequence features in human and mouse was constructed. These included differences in GC density, substitution rate, repeat densities, gene densities, histone modification densities and transcription factor binding site densities for each orthologous 100 Kb region. These were calculated by simply subtracting the mouse from the human sequence level feature value for each 100 Kb region. From this, the quantitative relationships between divergence in sequence and divergence in higher order structure in human and mouse could be directly examined. If changes in GC content or repeat densities drive structural change we

would expect to see this reflected in the new model. However, this exercise offered only a modest increase in the fit of a model to chromatin divergence (individual standardised R-squared: 0.10) over the previous model of chromatin divergence discussed above (Table 2).

Feature	Coefficients	R squared
ΔLINE	-2.94E-05	0.03
ΔH3K27ac	8.72E-02	0.02
ΔH3K4me3	-4.73E-02	0.00
ΔH2AZ	3.58E-02	0.01
ΔSimple repeat	2.27E-04	0.02
ΔSINE	3.81E-05	0.01
ΔGC	-4.36E+00	0.00
ΔGene density	6.67E-01	0.00
ΔOCT4	2.73E-02	0.00
ΔNANOG	-2.97E-02	0.00

Table 8.3 Regression coefficients and standardised r-squared values for models of divergence in orthologous DNA sequence features, represented by Δ, and chromatin structural divergence.

It can be concluded that the sequence features measured in the current investigation provide reasonable accuracy to predict higher order structure in both human and mouse. This is particularly true for GC content and specific repeat densities and shows that sequence level features and chromatin environments can be closely associated. This is expected given the previously reported associations of chromatin domains with mutational bias and the insertion preferences of repeat elements. On initial inspection, the results might also seem to suggest that structural divergence is most strongly associated with base composition. However chromatin divergence occurs within a particular spectrum of chromatin structure, so such results may be a reflection of higher order structure itself, rather than divergence. When measures of divergence in sequence features between the two species were explicitly included in the modelling there were no strong, convincing associations. For both regression models for chromatin divergence, the low R squared values meant that the explanatory power of the model was weak, implying that divergence in chromatin structure and genomic sequence are largely uncoupled.

Chapter 9

Discussion

9.1. DIVERGENT CHROMATIN IS RELATIVELY RARE IN THE GENOME

Until now, the degree of structural divergence between different mammalian species had not been measured. It has been previously established that relationships exist between different facets of higher order chromatin structure. For example large multi-megabase regions that co-ordinately replicate late in the cell cycle have been correlated to regions positioned at the periphery of the nucleus (Ryba et al., 2010). In addition, there is correspondence between regions with similar replication timing profiles and clusters of spatial proximity, reported in a Hi-C study (De and Michor, 2011). There is also evidence of conservation across species when measuring the same aspect of structure. Human and mouse replication timing profiles have been shown to be conserved, along with the stability of domain boundaries between similar cell types (Ryba et al., 2010). In a similar vein, there is a striking similarity between patterns of nuclear lamina interactions and spatial positioning between mouse and human ESCs, although there is evidence of some cell type specificity when examining conservation across different cell types (Meuleman et al., 2013). However, the quantitative extent and mechanisms of structural divergence between species had not been formally investigated prior to this study. One of the main focal points of this investigation was to carry out a comprehensive assessment of the level of agreement across all these heterogeneous datasets and, if any, the extent of divergent between different human and mouse cell types. The analyses presented extend the studies discussed so far and demonstrate the same signal across diverse datasets from studies that set out to observe nominally different aspects of structural organisation in many different embryonic cell types. It can be concluded that most measurable aspects of chromatin are remarkably well conserved across

the vast majority of the detectably orthologous genome, this is in spite of differences between species, cell types and experimental methods used.

Structural divergence was investigated by using a conservative approach which required consistent evidence of divergence between species over all cell types and all structural datasets assayed. It was found that divergent chromatin is relatively rare in the mammalian genome, representing 10.22% of orthologous 100 Kb genomic regions examined and encompassing over 170 Mb and including many hundreds of human and mouse genes. This suggests that structural divergence has played a major role in the evolution of many loci occupying these unusually divergent genomic regions. Many of the regions identified form unexpectedly large (>400 Kb) tracts of divergent chromatin, and were non-randomly distributed between and within chromosomes, and this clustering appears particularly pronounced at human subtelomeric regions. Across all types of divergent region clustering examined, the divergent regions of embryonic chromatin identified are consistently enriched for genes active in vertebrate development. These include homeodomain gene clusters, which have been implicated in evolutionary innovations to vertebrate developmental programmes, suggesting that selection may have modulated their regulation during evolution via alterations to chromatin. Specifically, HoxD clusters, important for limb development, have been previously shown to be within a 'regulatory archipelago' where expression can be tissue specific suggesting that the spatial conformation of loci may change as a result of interactions between promoters and enhancers, resulting in the creation of cell-type specific patterns of gene expression and organization (Montavon et al., 2011). Additionally, different levels of H3K27me3, important for the regulation of HoxD via polycomb targeting complexes, have been established across the developing limb. This indicates that different levels of chromatin compaction and enhancer-gene colocalisation are present across the developing limb (Williamson et al., 2012). The results from this study are consistent with this as developmental loci appear to be some of the most dynamic and most divergent across mammals. Olfactory receptors were also enriched in divergent structures; they are known to have tightly regulated expression patterns during development in mouse cells and are therefore dependent upon repressive chromatin structures (Magklara et al., 2011). This suggests that some repressive environments seen in mouse, but not in human, may have evolved as part of the regulatory landscape. Recent work has shown that there is a high degree of cell type specificity of olfactory receptor expression, where foci of

olfactory receptor genes are often found on the borders of heterochromatic regions. Cell type specificity is reported to be tightly regulated by higher order chromatin reorganisation and association to the nuclear lamina, which ensures only one olfactory receptor allele per cell, can be expressed (Clowney et al., 2012). So it seems that higher order chromatin reorganisation is critical for the tissue-specific regulation of some developmental gene expression programs.

9.2. THE MECHANISMS OF STRUCTURAL DIVERGENCE

The biological mechanisms underlying divergence in higher order chromatin structure remain unknown. Changes in the diversity or abundance of relatively rapidly evolving ncRNAs, which can mediate chromatin remodelling between cell types (Guttman and Rinn, 2012), could provide a molecular basis for divergence. However, given our present ignorance about the universe of ncRNAs this is a hypothesis for future work. Also the strong sequence-level correlates of human chromatin structure (Prendergast et al., 2007), and the unusual, lineage specific shifts in GC content seen here, suggest it is possible that sequence divergence underlies some degree of chromatin divergence. There is a strong association between mutation rate and higher order chromatin structure, such that higher mutation rates are found within more closed, gene poor, late replicating domains. This may reflect differences in the accessibility of some structures to the cellular repair machinery, or differences in their propensities to mutation itself. The fact that human-mouse substitution rate and SNP density show the same negative relationship to higher order chromatin indicates that similar mutational biases across 100 million years of evolution are still present. Consistent correlations have been observed between sequence divergence and nucleosome occupancy in studies of medaka fish (Sasaki et al., 2009), yeast (Teytelman et al., 2008) and human (Prendergast and Semple, 2011) such that nucleosome free regions undergo less sequence divergence that occupied regions and there is evidence that this reflects the action of selection (Semple and Taylor, 2009). This suggests that multiple structural levels can alter the mutation rates undergone by underlying DNA. However, the links between higher order and lower order chromatin structure and sequence divergence remain unclear. Compositional biases have previously been shown at multiple facets of the genome and genome organisation. GC rich regions are significantly enriched in genic regions, regulatory sites and open, accessible areas of chromatin (Gilbert et al., 2004). Nucleosomes have been shown to

preferentially associate with localised GC rich regions of DNA (Tillo and Hughes, 2009) and nucleosome positioning may be maintained by a balance of GC base pair gaining substitutions maintained by selection (Prendergast and Semple, 2011). Again, GC content has been shown to be one of the major correlates of chromatin structure in this investigation, underlining its importance across all the hierarchical layers of chromatin organisation. There have been no studies of how mutational bias related to nucleosome occupancy is related to the differences in mutation among higher order structures examined here.

One of the most striking features of all genomic features examined, whether based upon sequence or chromatin, was that the vast majority showed strong and significant correlations with higher order chromatin structure in both human and mouse. These observations were examined further by linear regression modelling. SINE repeat densities and DNA base composition were found to be the most important variables when defining chromatin structure. It may be relevant that SINE retrotransposons have recently been found at the boundaries of higher order chromatin domains, indicating that they may have a role in establishing the topological domain structure of the genome (Dixon et al., 2012). There were also generalised shifts in SINE and LINE repeat elements in human divergent chromatin although this could not be confirmed as an important aspect of chromatin divergence by linear modelling. Consistent with this, it has recently been suggested that transposon element derived repeat sequences, such as SINEs, are responsible for nearly half of all accessible chromatin regions, containing the majority of primate specific regulatory sequences (Jacques et al., 2013). There was also some evidence for the enrichment of ncRNAs in divergent chromatin structure in both human and mouse genomes. It is increasingly apparent that vertebrate genomes contain a large number of long non-coding RNA genes (lincRNAs) that are important for gene regulation. These have also recently been shown to be spread in a lineage specific manner by retrotransposon activity implying that the regulatory networks in which lincRNA genes act may be rapidly diverging between species (Kapusta et al., 2013). However, the lack of complete knowledge of lincRNAs in both human and mouse genomes mean that exploration of the link between lincRNAs and structural divergence must await further data.

It was not possible to link divergence in higher order structures to a variety of sequence level and locus level chromatin variables using linear modelling, and

the models constructed using such variables had low explanatory and predictive power. Overall, there was no convincing evidence for structural divergence driven by divergence in the underlying DNA sequence. However, there is clear evidence of particular patterns of sequence level divergence within particular chromatin structures. It is possible that divergent regions encounter altered mutational spectra as they assume a new conformation, giving rise over evolutionary time to the strong associations seen between chromatin structure and sequence level divergence.

9.3. LOWER ORDER AND HIGHER ORDER DIVERGENCE

In spite of data presented here, it is possible that fluctuations in locus level chromatin features such as histone modifications evolve in concert with structural divergence, but substantially more research is needed to examine this further. Locus level chromatin structure is known to have strong relationships DNA sequence and higher order structure. For example, LADs are linked with higher levels of H3K9me2 (Kind et al., 2013). Activating histone marks, H3K4me3 and H3K36me3 are known to be enriched in genomic regions that replicate early in the cell cycle (Hiratani et al., 2008). The two different structural compartments defined by Hi-C analyses also have opposing marks such that the open chromatin compartment is enriched for active histone marks, and the closed chromatin compartment was enriched for repressive histone marks, such as H3K27me3 (Lieberman-Aiden et al., 2009). H3K27me3 is needed to ensure compact chromatin environments for gene silencing (Eskeland et al., 2010). It may also be relevant that larger scale variation in chromatin structure within the mammalian genome is often associated with alterations in the spectrum of histone modifications at a region (de Wit and van Steensel, 2009), and OR gene clusters are now known to possess an unusual signature of histone modifications involving the molecular hallmarks of constitutive heterochromatin (Magklara et al., 2011). It is therefore possible that divergence in chromatin domains during evolution is caused by alterations in the constellations of histone modifications present. Although this was not seen in the linear model in Chapter 8 but this may be due to the low predictive power of the model. Recent comparative analysis have compared the extent of conservation across histone marks in three different species (Xiao et al., 2012), but as of yet, levels of divergence of histone modifications and locus level divergence compared to higher order divergence has not been measured. Using the same three-way (human, mouse and pig) dataset a similar analysis was set up to define conservation and divergence

across lower order features in order to compare findings directly to higher order structure and divergence. However it was clear that the widespread conservation seen for higher order chromatin structure is not matched by similar patterns of conservation in locus level chromatin features. This analysis may be improved in the future by focusing on particular genes or promoters present in divergent higher order chromatin to see if the patterns of epigenomic features present at divergent genes/promoters are different to those in conserved.

9.4. IMPLICATIONS OF DIVERGENT CHROMATIN STRUCTURE

This investigation was the first study to comprehensively and quantitatively investigate the degree of mammalian higher order structural divergence in the emerging field of epigenomic research. Epigenomics in the form of chromatin biology has long been an important area of science, central to genomic regulation and function. The ENCODE project consortium has stated that 80% of the genome is functional, a much higher figure than previously thought given the vast non-coding sequences in the genome (Dunham et al., 2012). This remains a controversial topic, with some researchers suggesting this to be evolutionarily impossible (Niu and Jiang, 2013) and may rest on the definition of functionality used. It can be certain, however, that the full functionality of the genome can only be realised in the context of chromatin structure. The flood of new high-resolution genome-wide data from projects such as ENCODE will undoubtedly prompt many further comparative studies of chromatin structure. The investigation also has important links to emerging medical research. As chromatin and nuclear organization is known to affect regulation it is not surprising that defects in higher-order chromatin and chromosome organization cause disease. For example, open chromatin regions have been identified within pancreatic cells (Gaulton et al., 2010) that harbour active regulatory elements. A type 2 diabetes–associated SNP has been found within a region of allele-specific open chromatin and shows allele-specific enhancer activity, which may suggest a potential disease mechanism (Gaulton et al., 2010). Histone modification profiles have been used to map regulatory elements in different cell types and have been used to analyse a wide variety disease SNPs in enhancers (Ernst et al., 2011). Chromatin state profiles such as these can be used to predict target genes whose expression may be affected by disease associated SNPs and therefore aid discovery of disease mechanisms. Chromatin structure is already well established as an area of study in cancer biology with DNA methylation and histone

modifications as the most well researched chromatin features that differ between normal cells and tumour cells in humans (Sharma et al., 2010). Aberrant DNA methylation causing gene silencing has been thought to be a mechanism for cancer cell survival (De Carvalho et al., 2012), however other studies have shown that this is not the major contributor and that the mechanisms might be more complex (Sproul et al., 2012). It is thought that unstable regulation could be at the centre of some cancer processes which then drives tumour growth (Pujadas and Feinberg, 2012). Recent studies have also found a dominant influence of higher order chromatin structures in shaping the mutational landscape of cancer genomes (De and Michor, 2011). So it is possible that anomalies in higher order chromatin structure, identified in comparative studies, may be useful indicators for cancer research.

9.5. FUTURE RESEARCH

There is extensive scope for this research to be extended with the addition of new higher resolution datasets. One of the main caveats to the comparative analysis carried out here is the inability to assign lineage specificity to chromatin divergence defined between human and mouse. Lineage specificity of DNA features such as indels and fluctuations in base composition can be readily discovered by using three-way analyses of genome-wide sequence alignment data, using the dog genome as an outgroup for example. Locus level chromatin structure features such as histone modifications, DNA methylation and transcription factor binding sites are also increasingly mapped genome-wide in a range of different species due to the emergence of ChIP-seq, MeDIP-seq and other high throughput sequencing techniques. This has enabled lineage specific comparative epigenomic analyses such as those carried out by Xiao et al (2012). Aspects of higher order chromatin structure are also now studied using high throughput methods, most notably Hi-C, revealing the importance of domain level chromatin architecture. For example, the discovery of topological domains of higher order chromatin structure categorised by high frequencies of intra-domain interactions but infrequent inter-domain interactions (Dixon et al., 2012). These domains have been further characterised by developing algorithms that can classify these domains further and associate chromatin domains with other chromatin level and sequence level features (Hu et al., 2013). Development of this sort of algorithm could further augment the investigations carried out here. Higher order structure mapped in an outgroup species such as pig

or dog would enable the designation of lineage specificity to divergent chromatin i.e. whether the divergence had occurred in the rodent or primate lineage. Lineage specific aspects of DNA might then be related to lineage specific chromatin divergence adding a new dimension to understanding the co-evolution of DNA sequence and the epigenome.

Appendix

Appendix 10.1 Full details of the 159 human large divergent domains and 160 mouse large divergent domains.

Human Chr	Start	End	SAM	Mouse Chr	Start	End	SAM
chr1	2800000	3199999	5.04	chr1	107400000	107799999	-3.99
chr1	81400000	81799999	-5.36	chr1	92000000	92399999	5.52
chr1	115100000	115699999	5.80	chr10	114000000	114399999	-4.14
chr1	115900000	116999999	5.31	chr10	124700000	125199999	-5.15
chr1	178700000	179399999	4.92	chr10	27000000	27399999	-4.69
chr10	6400000	7099999	-6.53	chr10	69600000	70199999	5.50
chr10	14700000	15399999	-5.09	chr11	46600000	46999999	-4.89
chr10	60800000	61599999	5.50	chr11	54000000	54699999	5.47
chr10	91900000	92299999	-5.26	chr11	66600000	66999999	4.38
chr10	100800000	101399999	4.38	chr11	86700000	87199999	5.16
chr10	113700000	114199999	5.07	chr11	89000000	89399999	4.57
chr10	115200000	116199999	6.03	chr12	111100000	111499999	4.30
chr10	123800000	124599999	6.73	chr12	120200000	120799999	-6.73
chr10	126900000	127699999	6.29	chr12	59000000	59499999	-5.05
chr10	131200000	132299999	4.96	chr13	13700000	14599999	-5.03
chr11	5900000	6699999	-5.48	chr13	48100000	48499999	5.22
chr11	12700000	13799999	4.87	chr13	85400000	85999999	-4.09
chr11	17900000	18799999	5.35	chr13	89800000	90199999	-4.53
chr11	43700000	44199999	4.81	chr14	11200000	11699999	-4.77
chr11	112300000	113299999	4.70	chr14	32600000	33499999	4.19
chr12	59500000	60099999	-5.15	chr14	64800000	65199999	6.55
chr12	72500000	72899999	-4.14	chr14	71100000	71499999	4.74
chr12	78800000	79499999	-6.94	chr14	72600000	73099999	-5.10
chr12	106100000	107199999	5.19	chr15	29200000	29699999	-4.89
chr13	44400000	44999999	4.57	chr16	41900000	42299999	-5.42
chr13	49600000	50099999	-5.15	chr17	52100000	52599999	-5.00
chr13	60200000	60799999	-4.75	chr17	78200000	78599999	-6.65
chr13	60900000	61399999	-4.28	chr18	15600000	16299999	-4.60
chr13	111900000	112799999	6.42	chr19	35400000	35899999	-5.26
chr14	31900000	32899999	5.85	chr19	55000000	55399999	5.07
chr14	38400000	38899999	-5.05	chr2	116600000	116999999	4.28
chr14	88500000	89199999	4.37	chr2	131500000	131899999	5.29
chr14	95300000	95999999	5.35	chr2	139000000	139399999	-5.29
chr14	101600000	101999999	4.30	chr2	93600000	93999999	4.81
chr15	35000000	35899999	-5.36	chr3	102500000	102999999	5.80
chr15	38100000	38599999	4.28	chr3	149100000	149499999	-5.36
chr15	55200000	56999999	5.78	chr3	25000000	25599999	-5.24
chr15	69700000	70699999	5.27	chr3	63200000	63599999	-4.40
chr16	48300000	49199999	4.74	chr4	153800000	154199999	5.04
chr16	50700000	51399999	5.02	chr5	142500000	143099999	6.46
chr16	54000000	55499999	5.34	chr5	19200000	19599999	-5.33
chr16	66500000	66899999	-4.31	chr5	42000000	42399999	-4.18
chr16	73900000	74299999	4.59	chr5	64800000	65199999	4.95
chr17	9100000	9699999	5.50	chr5	75200000	75699999	4.43
chr17	10400000	10899999	4.38	chr5	82100000	82499999	-5.11
chr17	54400000	54799999	4.57	chr6	100000000	100499999	5.54
chr17	56800000	57499999	5.16	chr6	25600000	26299999	-4.16
chr18	4100000	4899999	-6.01	chr6	49700000	50199999	-5.29
chr18	20300000	21799999	-5.70	chr6	64000000	64399999	-4.52
chr18	24500000	25099999	-4.60	chr6	97400000	97799999	4.97

chr18	59700000	60099999	-3.99	chr7	53800000	54399999	5.35
chr18	73700000	74699999	5.53	chr8	106700000	107099999	-4.31
chr18	76300000	77899999	5.77	chr8	110100000	110499999	4.59
chr19	30200000	31199999	4.82	chr8	30000000	30399999	-4.96
chr19	32500000	33199999	4.81	chr8	36400000	36799999	-4.64
chr2	17400000	17899999	-5.90	chr9	102900000	103299999	4.67
chr2	36200000	36599999	-6.65	chr9	91800000	92199999	-5.07
chr2	63200000	64299999	-5.93	chr1	79100000	79699999	-5.12
chr2	71900000	72599999	6.01	chr10	50100000	50999999	-5.80
chr2	76400000	77599999	-4.82	chr10	53100000	53799999	-4.52
chr2	154100000	154599999	-4.43	chr11	21100000	21999999	-5.93
chr2	157600000	158199999	-5.19	chr11	36400000	36899999	-4.47
chr2	166200000	166999999	-5.34	chr11	67600000	68099999	5.50
chr2	167800000	169099999	-5.48	chr12	11300000	11799999	-5.90
chr2	200800000	201399999	-5.03	chr13	37500000	37999999	4.62
chr2	224200000	224699999	-5.12	chr13	56100000	56699999	5.32
chr2	237400000	237799999	5.52	chr13	90800000	91299999	-4.37
chr2	239600000	241999999	5.80	chr13	91700000	93099999	-4.75
chr20	4400000	4799999	5.29	chr14	19200000	19799999	-6.44
chr20	12700000	13099999	-5.29	chr14	87700000	88199999	-4.28
chr20	19500000	20199999	5.28	chr16	30500000	30999999	5.60
chr20	41600000	42599999	4.35	chr16	75000000	75499999	-4.79
chr20	51000000	51599999	4.69	chr18	47800000	48299999	-5.08
chr3	17600000	18299999	-5.04	chr19	14300000	14799999	4.32
chr3	18600000	18999999	-5.00	chr19	43300000	43799999	4.38
chr3	23200000	23799999	-6.44	chr2	162300000	163099999	4.35
chr3	60500000	60899999	-4.77	chr2	3000000	3599999	-5.09
chr3	62500000	63099999	-4.41	chr2	53700000	54199999	-4.43
chr3	69400000	69799999	4.97	chr2	57400000	57899999	-5.19
chr3	72200000	72799999	5.54	chr3	137300000	137799999	5.94
chr3	75900000	76499999	-4.79	chr3	22800000	23399999	-4.87
chr3	115400000	115799999	-5.42	chr4	15500000	16099999	-5.50
chr3	126700000	127999999	4.47	chr4	23600000	24099999	-4.95
chr3	133300000	133699999	4.67	chr5	14300000	14799999	-4.80
chr3	146200000	146599999	-5.07	chr6	4800000	5499999	4.13
chr3	148700000	149299999	-5.30	chr6	55700000	56199999	-4.73
chr3	154900000	155299999	-4.40	chr6	62800000	63599999	-4.13
chr3	172700000	173599999	-4.95	chr7	36300000	36899999	4.81
chr3	173700000	174199999	-5.24	chr8	11800000	12499999	6.42
chr3	175800000	176299999	-4.87	chr8	89200000	89899999	4.74
chr3	190700000	192099999	-5.10	chr9	72100000	73199999	5.78
chr3	193200000	194199999	4.75	chr9	79500000	79999999	-5.61
chr3	194300000	194799999	5.60	chr1	158300000	158899999	4.92
chr4	13400000	13799999	-4.18	chr1	57400000	58099999	-5.03
chr4	16600000	17399999	-5.32	chr11	34800000	35399999	4.78
chr4	23900000	24999999	-4.53	chr12	105700000	106299999	5.35
chr4	38200000	38699999	4.95	chr12	99500000	100099999	4.37
chr4	54600000	55299999	4.43	chr13	68700000	69399999	-4.69
chr4	62800000	63199999	-5.11	chr13	77200000	77999999	-5.80
chr4	89600000	90799999	-6.00	chr14	13300000	13899999	-4.41
chr4	93000000	93699999	-4.13	chr14	76900000	77499999	4.57
chr4	94100000	94499999	-4.52	chr14	87000000	87599999	-4.75
chr4	100500000	101099999	5.94	chr15	23900000	24499999	-5.09
chr4	147300000	148099999	4.57	chr17	11900000	12699999	6.13
chr4	182600000	183299999	-5.23	chr17	51200000	51799999	-5.04
chr5	7100000	7899999	-4.69	chr17	70000000	70699999	-6.01
chr5	12200000	12799999	-4.89	chr18	82700000	83499999	5.53
chr5	18300000	18899999	-5.09	chr19	19200000	19799999	-4.74
chr5	37900000	38899999	4.59	chr2	145300000	145899999	5.28
chr5	73100000	73799999	4.73	chr2	168900000	169499999	4.69
chr5	78200000	79399999	-4.68	chr2	27100000	28099999	5.34
chr5	80000000	80899999	-4.75	chr3	25700000	26699999	-4.95
chr5	81300000	81799999	-4.37	chr4	27600000	28499999	-4.29
chr5	82400000	82899999	-4.53	chr6	21200000	21799999	-4.75
chr5	86100000	86599999	-4.09	chr6	40100000	40799999	-4.41
chr5	87000000	88399999	-5.86	chr6	51500000	52199999	5.78

chr5	93300000	93999999	-5.80
chr5	116200000	116699999	-5.08
chr5	130500000	131499999	5.47
chr5	134600000	135299999	5.32
chr5	146700000	147499999	-5.22
chr5	156000000	156399999	-4.89
chr5	166600000	167099999	-4.47
chr5	168200000	168899999	4.78
chr6	6700000	7199999	4.62
chr6	19500000	19999999	5.22
chr6	75800000	76399999	-5.61
chr6	94300000	94999999	-4.29
chr6	97900000	98399999	-4.95
chr6	100600000	101499999	-5.80
chr6	114200000	114999999	-4.87
chr6	119000000	119699999	-4.52
chr6	129100000	129599999	-4.69
chr6	160800000	162099999	6.13
chr6	163200000	164299999	4.85
chr6	168200000	170899999	8.85
chr7	4100000	4999999	6.46
chr7	7500000	8499999	-4.58
chr7	20100000	20699999	-6.73
chr7	24200000	24799999	-5.29
chr7	26400000	27199999	5.78
chr7	31500000	31999999	-4.73
chr7	78200000	78599999	-5.33
chr7	82500000	82999999	-4.80
chr7	94500000	95299999	4.13
chr7	119900000	120499999	-4.75
chr7	124400000	125099999	-4.16
chr7	141100000	141899999	-4.41
chr8	8600000	8999999	-4.64
chr8	9800000	10199999	6.55
chr8	21400000	21799999	4.74
chr8	35000000	35399999	-4.96
chr8	59300000	60099999	-5.23
chr8	72400000	73699999	-4.71
chr8	90800000	91399999	-5.50
chr8	124500000	126099999	4.53
chr9	76500000	77099999	-4.74
chr9	82100000	82599999	4.32
chr9	136600000	138699999	5.25
chr6	84100000	84699999	6.01
chr7	137700000	138399999	6.73
chr7	140400000	141099999	6.29
chr8	80500000	81199999	4.57
chr1	14400000	15499999	-4.71
chr1	93600000	95199999	5.80
chr10	108200000	108899999	-6.94
chr12	53100000	53899999	5.85
chr13	98100000	98799999	4.73
chr2	10600000	11299999	-6.53
chr2	113800000	114499999	-5.36
chr2	65600000	66299999	-5.34
chr4	6100000	6999999	-5.23
chr5	52000000	52899999	-4.53
chr6	57700000	60799999	-6.06
chr6	8100000	8899999	-4.58
chr8	49700000	50399999	-5.23
chr8	91100000	91799999	5.02
chr10	35800000	36799999	-4.87
chr15	6800000	7699999	4.59
chr16	29500000	30399999	4.75
chr18	43400000	44199999	-5.22
chr19	56300000	57099999	6.03
chr5	44900000	45699999	-5.32
chr6	88300000	89299999	4.47
chr7	112200000	112999999	-5.48
chr7	144000000	144899999	4.96
chr9	49200000	50199999	4.70
chr10	83500000	84399999	5.19
chr15	58000000	59199999	4.53
chr17	10100000	10999999	4.85
chr17	12900000	15699999	7.40
chr18	80300000	81499999	5.77
chr3	101300000	102199999	5.31
chr6	80100000	81299999	-4.82
chr7	119800000	120799999	4.87
chr7	38100000	38999999	4.82
chr9	60900000	61799999	5.27
chr2	67100000	68299999	-5.48
chr13	93500000	94599999	-4.68
chr13	83400000	84999999	-5.86
chr16	27100000	28399999	-5.10
chr8	94100000	95299999	5.34
chr18	11600000	12999999	-5.70

Appendix 10.2 The full list of significant functional gene enrichments and annotation terms within the 1719 divergent human and mouse structural regions.

Species	Divergence	Enrichment	Description	Genes	p-value	FDR
Human	Human closed/ Mouse open	IPR001827	Homeobox Protein, Antennapedia Type, Conserved Site	10	4.80E-07	7.33E-04
Human	Human closed/ Mouse open	CYTOBAND	CYTOBAND 18q23	6	5.63E-06	7.52E-03
Human	Human closed/ Mouse open	GO:0003002	Regionalization	21	8.65E-06	1.50E-02
Human	Human closed/ Mouse open	CYTOBAND	CYTOBAND 6q27	6	3.11E-05	4.15E-02
Human	Human closed/ Mouse open	CYTOBAND	CYTOBAND 2q37.3	9	3.28E-05	4.38E-02
Human	Human open/ Mouse closed	CYTOBAND	CYTOBAND 11p15.4	15	1.70E-10	2.17E-07
Human	Human open/ Mouse closed	GO:0007606	Sensory Perception Of Chemical Stimulus	21	2.50E-09	4.15E-06
Human	Human open/ Mouse closed	GO:0050877	Neurological System Process	41	1.42E-07	2.36E-04
Human	Human open/ Mouse closed	CYTOBAND	CYTOBAND 10p13	8	3.47E-07	4.44E-04
Human	Human open/ Mouse closed	GO:0007186	G-Protein Coupled Receptor Protein Signaling Pathway	36	3.81E-07	6.34E-04
Human	Human open/ Mouse closed	GO:0007608	Sensory Perception Of Smell	16	7.79E-07	1.30E-03
Human	Human open/ Mouse closed	IPR000725	Olfactory Receptor	15	1.31E-06	1.89E-03
Human	Human open/ Mouse closed	IPR017452	GPCR, Rhodopsin-Like Superfamily	24	2.46E-06	3.56E-03
Human	Human open/ Mouse closed	GO:0004984	Olfactory Receptor Activity	15	2.67E-06	3.83E-03
Human	Human open/ Mouse closed	IPR000276	7TM GPCR, Rhodopsin-Like	24	2.83E-06	4.09E-03
Human	Human open/ Mouse closed	CYTOBAND	CYTOBAND 7q35	7	4.72E-06	6.04E-03
Human	Human open/ Mouse closed	PIRSF800006	Rhodopsin-Like G Protein-Coupled Receptors	24	5.98E-06	7.12E-03
Human	Human open/ Mouse closed	GO:0007600	Sensory Perception	27	7.90E-06	1.31E-02
Human	Human open/ Mouse closed	GO:0050890	Cognition	29	1.33E-05	2.21E-02
Mouse	Human closed/ Mouse open	GO:0003002	Regionalization	32	1.96E-09	3.39E-06
Mouse	Human closed/ Mouse open	GO:0009952	Anterior/Posterior Pattern Formation	27	2.29E-09	3.97E-06
Mouse	Human closed/ Mouse open	GO:0007389	Pattern Specification Process	36	5.25E-09	9.09E-06
Mouse	Human closed/ Mouse open	CYTOBAND	CYTOBAND 2 45.0 CM	9	1.29E-08	1.89E-05
Mouse	Human closed/ Mouse open	CYTOBAND	CYTOBAND 19 D2	12	3.31E-08	4.84E-05
Mouse	Human closed/ Mouse open	IPR001356	Homeobox	30	3.48E-08	5.46E-05
Mouse	Human closed/ Mouse open	IPR012287	Homeodomain-Related	30	5.95E-08	9.34E-05
Mouse	Human closed/ Mouse open	GO:0009954	Proximal/Distal Pattern Formation	11	6.60E-08	1.14E-04
Mouse	Human closed/ Mouse open	IPR017970	Homeobox, Conserved Site	29	1.00E-07	1.57E-04
Mouse	Human closed/ Mouse open	IPR001827	Homeobox Protein, Antennapedia Type, Conserved Site	11	1.45E-07	2.27E-04

Mouse	Human closed/ Mouse open	GO:0043565	Sequence-Specific DNA Binding	48	2.58E-07	3.80E-04
Mouse	Human closed/ Mouse open	GO:0048598	Embryonic Morphogenesis	37	7.13E-07	1.23E-03
Mouse	Human closed/ Mouse open	GO:0001501	Skeletal System Development	31	2.08E-06	3.61E-03
Mouse	Human closed/ Mouse open	GO:0048705	Skeletal System Morphogenesis	20	3.88E-06	6.71E-03
Mouse	Human closed/ Mouse open	CYTOBAND	CYTOBAND 2 C3|2 45.0 CM	6	1.03E-05	1.50E-02
Mouse	Human closed/ Mouse open	CYTOBAND	CYTOBAND 17 A2	6	2.61E-05	3.81E-02
Mouse	Human closed/ Mouse open	GO:0003700	Transcription Factor Activity	53	3.01E-05	4.44E-02
Mouse	Human open/ Mouse closed	GO:0007606	Sensory Perception Of Chemical Stimulus	39	2.19E-18	3.58E-15
Mouse	Human open/ Mouse closed	GO:0007608	Sensory Perception Of Smell	34	5.80E-16	9.10E-13
Mouse	Human open/ Mouse closed	IPR000725	Olfactory Receptor	33	7.94E-16	1.15E-12
Mouse	Human open/ Mouse closed	GO:0004984	Olfactory Receptor Activity	33	2.41E-15	3.45E-12
Mouse	Human open/ Mouse closed	IPR017452	GPCR, Rhodopsin-Like Superfamily	47	3.73E-15	5.58E-12
Mouse	Human open/ Mouse closed	GO:0007186	G-Protein Coupled Receptor Protein Signaling Pathway	58	4.64E-15	7.65E-12
Mouse	Human open/ Mouse closed	IPR000276	7TM GPCR, Rhodopsin-Like	44	3.50E-13	5.18E-10
Mouse	Human open/ Mouse closed	GO:0007600	Sensory Perception	42	1.26E-12	2.07E-09
Mouse	Human open/ Mouse closed	GO:0050890	Cognition	43	1.12E-11	1.83E-08
Mouse	Human open/ Mouse closed	GO:0050877	Neurological System Process	49	1.34E-10	2.20E-07
Mouse	Human open/ Mouse closed	CYTOBAND	CYTOBAND 13 C3	11	1.29E-09	1.78E-06
Mouse	Human open/ Mouse closed	CYTOBAND	CYTOBAND 7 E3	13	1.82E-07	2.52E-04
Mouse	Human open/ Mouse closed	PIRSF003152	G Protein-Coupled Olfactory Receptor, Class II	19	2.78E-07	3.43E-04
Mouse	Human open/ Mouse closed	IPR015493	Protocadherin Beta	6	2.81E-07	4.15E-04
Mouse	Human open/ Mouse closed	GO:0007166	Cell Surface Receptor Linked Signal Transduction	63	6.49E-07	1.06E-03
Mouse	Human open/ Mouse closed	CYTOBAND	CYTOBAND 6 A3.1	7	8.00E-07	1.11E-03
Mouse	Human open/ Mouse closed	CYTOBAND	CYTOBAND 18 B3	12	8.48E-07	1.17E-03
Mouse	Human open/ Mouse closed	PIRSF800006	Rhodopsin-Like G Protein-Coupled Receptors	30	1.27E-06	1.57E-03
Mouse	Human open/ Mouse closed	IPR008253	Marvel	7	1.47E-05	2.17E-02
Mouse	Human open/ Mouse closed	GO:0016021	Integral To Membrane	137	3.56E-05	4.60E-02

Appendix 10.3 The full list of significant functional gene enrichments and annotation terms within the cell type/ species specific divergent structural regions.

Divergence	Enrichment	Description	Gene	p-value	FDR
Species Difference ESC	INTERPRO	IPR007237:CD20/IgE Fc receptor beta subunit	8	4.66E-08	7.31E-05
	GOTERM_CC_FAT	GO:0000786~nucleosome	11	6.31E-07	8.73E-04
	GOTERM_BP_FAT	GO:0065004~protein-DNA complex assembly	12	1.52E-06	2.68E-03
	CYTOBAND	11q12.2	10	5.22E-06	7.42E-03
	GOTERM_BP_FAT	GO:0006334~nucleosome assembly	11	5.95E-06	1.04E-02
	GOTERM_BP_FAT	GO:0031497~chromatin assembly	11	5.95E-06	1.04E-02
	GOTERM_BP_FAT	GO:0034728~nucleosome organization	12	6.63E-06	1.16E-02
	GOTERM_CC_FAT	GO:0032993~protein-DNA complex	12	9.33E-06	1.29E-02
Species Difference NPC	CYTOBAND	1q42.13	12	1.90E-10	2.73E-07
	CYTOBAND	1p36.33	14	3.68E-10	5.30E-07
	INTERPRO	IPR012287:Homeodomain-related	38	2.06E-09	3.31E-06
	CYTOBAND	14q11	12	2.80E-09	4.03E-06
	CYTOBAND	10q23.31	11	6.40E-08	9.21E-05
	INTERPRO	IPR001356:Homeobox	34	1.06E-07	1.70E-04
	INTERPRO	IPR017970:Homeobox, conserved site	34	1.26E-07	2.02E-04
	CYTOBAND	19q13.11	10	1.87E-07	2.69E-04
	SMART	SM00389:HOX	34	2.16E-07	2.77E-04
	CYTOBAND	7p15-p14	10	4.81E-07	6.93E-04
	GOTERM_BP_FAT	GO:0048598~embryonic morphogenesis	38	2.39E-06	4.28E-03
	GOTERM_MF_FAT	GO:0003700~transcription factor activity	79	3.02E-06	4.64E-03
	GOTERM_MF_FAT	GO:0043565~sequence-specific DNA binding	58	6.97E-06	1.07E-02
	CYTOBAND	5q31	15	1.31E-05	1.89E-02
	GOTERM_BP_FAT	GO:0031328~positive regulation of cellular biosynthetic process	61	1.39E-05	2.50E-02
	GOTERM_BP_FAT	GO:0048562~embryonic organ morphogenesis	23	1.44E-05	2.58E-02
	GOTERM_BP_FAT	GO:0009891~positive regulation of biosynthetic process	61	1.89E-05	3.39E-02
	CYTOBAND	9q22.33	7	3.41E-05	4.90E-02
Cell Type Differences Human	CYTOBAND	14q11	12	1.16E-10	1.63E-07
	CYTOBAND	7p15-p14	10	3.70E-08	5.19E-05
	INTERPRO	IPR012287:Homeodomain-related	30	4.22E-08	6.62E-05
	INTERPRO	IPR001356:Homeobox	28	2.41E-07	3.77E-04
	INTERPRO	IPR017970:Homeobox, conserved site	28	2.79E-07	4.36E-04
	SMART	SM00389:HOX	28	9.03E-07	1.13E-03
	INTERPRO	IPR001827:Homeobox protein, antennapedia type, conserved site	10	3.25E-06	5.10E-03
	GOTERM_BP_FAT	GO:0048812~neuron projection morphogenesis	21	3.11E-05	5.46E-02
Cell Type Differences Mouse	CYTOBAND	11 A4	15	7.45E-08	1.16E-04
	CYTOBAND	3 A1	11	3.16E-06	4.91E-03

Appendix 10.4 The full list of significant function gene enrichments and annotation terms within the 159 large divergent domains.

File	Cluster	Category	Term	Genes	P-value	FDR
chr11 5900000 6699999	16	CYTOBAND	CYTOBAND 11p15.4	15	3.74E-28	2.05E-25
chr2 239600000 241999999	68	CYTOBAND	CYTOBAND 2q37.3	8	1.30E-17	4.97E-15
chr7 26400000 27199999	141	IPR001827	Homeobox Protein, Antennapedia Type, Conserved Site	7	1.54E-16	4.44E-14
chr3 133300000 133699999	84	CYTOBAND	CYTOBAND 3q22.1	6	1.24E-14	2.81E-12
chr7 26400000 27199999	141	CYTOBAND	CYTOBAND 7p15-P14	6	1.16E-13	4.41E-11
chr5 78200000 79399999	111	CYTOBAND	CYTOBAND 5q14.1	6	2.20E-13	1.37E-10
chr18 20300000 21799999	49	CYTOBAND	CYTOBAND 18q11.2	6	3.34E-13	2.27E-10
chr5 146700000 147499999	121	CYTOBAND	CYTOBAND 5q32	5	1.45E-11	4.57E-09
chr7 26400000 27199999	141	GO:0048562	Embryonic Organ Morphogenesis	7	6.27E-12	7.55E-09
chr7 26400000 27199999	141	GO:0009952	Anterior/Posterior Pattern Formation	7	8.06E-12	9.70E-09
chr7 26400000 27199999	141	IPR001356	Homeobox	7	2.39E-11	1.04E-08
chr7 26400000 27199999	141	IPR017970	Homeobox, Conserved Site	7	2.49E-11	1.08E-08
chr7 26400000 27199999	141	IPR012287	Homeodomain-Related	7	3.06E-11	1.33E-08
chr7 26400000 27199999	141	GO:0048568	Embryonic Organ Development	7	2.11E-11	2.55E-08
chr7 26400000 27199999	141	GO:0048704	Embryonic Skeletal System Morphogenesis	6	3.94E-11	4.74E-08
chr7 26400000 27199999	141	GO:0003002	Regionalization	7	4.64E-11	5.58E-08
chr11 5900000 6699999	16	PIRSF038651	G Protein-Coupled Olfactory Receptor, Class I	7	2.76E-10	1.96E-07
chr7 26400000 27199999	141	GO:0048706	Embryonic Skeletal System Development	6	1.77E-10	2.13E-07
chr7 26400000 27199999	141	GO:0007389	Pattern Specification Process	7	1.86E-10	2.24E-07
chr7 26400000 27199999	141	GO:0048598	Embryonic Morphogenesis	7	4.63E-10	5.57E-07
chr16 66500000 66899999	42	IPR008253	Marvel	5	1.10E-09	7.09E-07
chr7 26400000 27199999	141	GO:0048705	Skeletal System Morphogenesis	6	1.11E-09	1.33E-06
chr17 10400000 10899999	45	CYTOBAND	CYTOBAND 17p13.1	5	1.51E-08	5.76E-06
chr7 94500000 95299999	145	CYTOBAND	CYTOBAND 7q21.3	4	3.06E-08	1.17E-05
chr7 26400000 27199999	141	IPR017995	Homeobox Protein, Antennapedia Type	4	3.34E-08	1.45E-05
chr7 26400000 27199999	141	GO:0043565	Sequence-Specific DNA Binding	7	3.29E-08	1.57E-05
chr10 115200000 116199999	12	CYTOBAND	CYTOBAND 10q25.3	4	4.21E-08	1.83E-05
chr11 5900000 6699999	16	GO:0007608	Sensory Perception Of Smell	8	1.43E-08	2.02E-05
chr3 126700000 127999999	83	CYTOBAND	CYTOBAND 3q21.3	4	5.51E-08	3.74E-05
chr11 5900000 6699999	16	IPR000725	Olfactory Receptor	7	6.10E-08	5.77E-05
chr11 5900000 6699999	16	GO:0007606	Sensory Perception Of Chemical Stimulus	8	6.09E-08	8.59E-05
chr11 5900000 6699999	16	GO:0050877	Neurological System Process	12	8.61E-08	1.21E-04
chr7 26400000 27199999	141	GO:0001501	Skeletal System Development	6	1.07E-07	1.29E-04
chr7 94500000 95299999	145	PIRSF016435	Paraoxonase	3	1.29E-06	1.29E-04
chr16 54000000 55499999	41	IPR003893	Iroquois-Class Homeodomain Protein	3	4.26E-07	1.34E-04
chr11 43700000 44199999	19	CYTOBAND	CYTOBAND 11p11.2	4	3.57E-07	1.36E-04
chr7 26400000 27199999	141	GO:0043009	Chordate Embryonic Development	6	1.30E-07	1.57E-04

chr7 26400000 27199999	141	GO:0009792	Embryonic Development Ending In Birth Or Egg Hatching	6	1.37E-07	1.65E-04
chr7 26400000 27199999	141	GO:0003700	Transcription Factor Activity	7	3.52E-07	1.68E-04
chr7 94500000 95299999	145	IPR002640	Arylesterase	3	5.11E-07	2.94E-04
chr11 5900000 6699999	16	GO:0004984	Olfactory Receptor Activity	7	2.83E-07	3.17E-04
chr14 95300000 95999999	33	CYTOBAND	CYTOBAND 14q32.13	3	1.47E-06	3.32E-04
chr7 94500000 95299999	145	GO:0004063	Aryldialkylphosphatase Activity	3	4.22E-07	3.35E-04
chr7 94500000 95299999	145	GO:0004064	Arylesterase Activity	3	8.44E-07	6.69E-04
chr7 141100000 141899999	148	CYTOBAND	CYTOBAND 7q31.3-Q32	3	1.69E-06	7.37E-04
chr11 5900000 6699999	16	GO:0050890	Cognition	10	5.60E-07	7.90E-04
chr16 66500000 66899999	42	GO:0006935	Chemotaxis	5	8.79E-07	8.71E-04
chr16 66500000 66899999	42	GO:0042330	Taxis	5	8.79E-07	8.71E-04
chr11 5900000 6699999	16	IPR017452	GPCR, Rhodopsin-Like Superfamily	8	9.52E-07	9.00E-04
chr11 5900000 6699999	16	IPR000276	7TM GPCR, Rhodopsin-Like	8	1.01E-06	9.50E-04
chr7 26400000 27199999	141	GO:0030528	Transcription Regulator Activity	7	4.24E-06	2.03E-03
chr5 130500000 131499999	119	CYTOBAND	CYTOBAND 5q31.1	3	1.18E-05	2.68E-03
chr11 5900000 6699999	16	GO:0007600	Sensory Perception	9	2.38E-06	3.36E-03
chr7 26400000 27199999	141	GO:0006355	Regulation Of Transcription, DNA-Dependent	7	3.48E-06	4.19E-03
chr7 26400000 27199999	141	GO:0051252	Regulation Of RNA Metabolic Process	7	3.95E-06	4.75E-03
chr16 66500000 66899999	42	GO:0005125	Cytokine Activity	5	6.40E-06	5.41E-03
chr7 141100000 141899999	148	IPR007960	Mammalian Taste Receptor	3	1.02E-05	6.12E-03
chr7 141100000 141899999	148	GO:0008527	Taste Receptor Activity	3	9.84E-06	7.98E-03
chr16 66500000 66899999	42	GO:0007626	Locomotory Behavior	5	1.04E-05	1.04E-02
chr7 26400000 27199999	141	GO:0006350	Transcription	7	9.09E-06	1.09E-02
chr7 26400000 27199999	141	GO:0003677	DNA Binding	7	2.37E-05	1.14E-02
chr7 94500000 95299999	145	GO:0019439	Aromatic Compound Catabolic Process	3	1.67E-05	1.64E-02
chr11 5900000 6699999	16	PIRSF800006	Rhodopsin-Like G Protein-Coupled Receptors	8	2.66E-05	1.89E-02
chr16 66500000 66899999	42	CYTOBAND	CYTOBAND 16q22.1	4	3.19E-05	1.99E-02
chr7 94500000 95299999	145	IPR011042	Six-Bladed Beta-Propeller, TolB-Like	3	4.68E-05	2.69E-02
chr3 193200000 194199999	92	CYTOBAND	CYTOBAND 3q29	3	5.23E-05	2.87E-02
chr7 26400000 27199999	141	PIRSF002612	Homeotic Protein Hox A5/D4	3	9.63E-05	4.61E-02
chr7 26400000 27199999	141	GO:0045449	Regulation Of Transcription	7	4.06E-05	4.89E-02

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Divergence of Mammalian Higher Order Chromatin Structure Is Associated with Developmental Loci

Emily V. Chambers, Wendy A. Bickmore, Colin A. Semple*

MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom

Abstract

Several recent studies have examined different aspects of mammalian higher order chromatin structure – replication timing, lamina association and Hi-C inter-locus interactions — and have suggested that most of these features of genome organisation are conserved over evolution. However, the extent of evolutionary divergence in higher order structure has not been rigorously measured across the mammalian genome, and until now little has been known about the characteristics of any divergent loci present. Here, we generate a dataset combining multiple measurements of chromatin structure and organisation over many embryonic cell types for both human and mouse that, for the first time, allows a comprehensive assessment of the extent of structural divergence between mammalian genomes. Comparison of orthologous regions confirms that all measurable facets of higher order structure are conserved between human and mouse, across the vast majority of the detectably orthologous genome. This broad similarity is observed in spite of many loci possessing cell type specific structures. However, we also identify hundreds of regions (from 100 Kb to 2.7 Mb in size) showing consistent evidence of divergence between these species, constituting at least 10% of the orthologous mammalian genome and encompassing many hundreds of human and mouse genes. These regions show unusual shifts in human GC content, are unevenly distributed across both genomes, and are enriched in human subtelomeric regions. Divergent regions are also relatively enriched for genes showing divergent expression patterns between human and mouse ES cells, implying these regions cause divergent regulation. Particular divergent loci are strikingly enriched in genes implicated in vertebrate development, suggesting important roles for structural divergence in the evolution of mammalian developmental programmes. These data suggest that, though relatively rare in the mammalian genome, divergence in higher order chromatin structure has played important roles during evolution.

Citation: Chambers EV, Bickmore WA, Semple CA (2013) Divergence of Mammalian Higher Order Chromatin Structure Is Associated with Developmental Loci. PLoS Comput Biol 9(4): e1003017. doi:10.1371/journal.pcbi.1003017

Editor: Andrey Rzhetsky, University of Chicago, United States of America

Received October 8, 2012; **Accepted** February 18, 2013; **Published** April 4, 2013

Funding: CAS and WAB are supported by UK Medical Research Council (MRC) core financial support. EVC is supported by a MRC Capacity Building Studentship. WAB is also supported by ERC advanced grant 249956. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Colin.Semple@igmm.ed.ac.uk

Introduction

Chromatin structure plays critical roles in genome functions such as transcription, replication and repair, it can mediate human disease processes [1] and is implicated in ageing [2]. The primary level of eukaryotic chromatin structure involves the DNA sequence wrapped around nucleosomes and the covalent modification of histones within the nucleosomes. Interactions between nucleosomes give rise to secondary structures, which may include a 30 nm chromatin fibre, and which vary in their degree of compaction across the genome [3]. Multiple higher levels of topological organisation, further structuring the genome, are also known to exist but their precise nature and their inter-relationships are the subjects of intense study and debate [4].

Genome-wide data relating to primary levels of chromatin structure (nucleosome occupancy, histone modifications etc) in a variety of mammalian cell types are abundant, due to the ability to profile these chromatin features by combinations of MNase digestion, chromatin immunoprecipitation and high-throughput sequencing [5]. However, the diversity of higher order structure across the genome is less well studied. An early genome-wide survey of higher order chromatin structure in the human genome discovered an undulating landscape of domains from hundreds of kilobases to many megabases in size; some relatively accessible or 'open' and others adopting a spectrum of more 'closed' condensed structures [3]. The most open domains corresponded to regions of relatively high gene density, replicating early in the cell cycle, and they may create an environment that facilitates transcriptional activation [6]. In contrast, more closed regions were relatively late replicating and gene poor. Replication timing profiles measured across the genome in multiple human and mouse cell types have also revealed the presence of domains on a similar scale, ranging from a few hundred kilobases to several megabases, that show coordinated replication timing during the cell cycle [7,8]. Other studies have examined different facets of higher order chromatin structure and organisation. Genomic regions interacting with tagged nuclear lamina components, and hence considered to be located at the nuclear periphery, have been mapped across the human and mouse genomes [9,10]. These lamina-associated domains (LADs) are relatively late replicating, gene poor regions from 40 Kb to 15 Mb in length and harbour genes with low transcriptional activity [10]. Overall LADs encompass around 40% of the genome and their locations and extent appear to be largely similar over cell types [10]. More recently, 3C-type

Author Summary

The mammalian genome is organised into large multi-megabase domains defined by their physical structure, or higher order chromatin structure. Although these structures are believed to be well conserved between species, there have been few studies attempting to quantify such conservation, or identify divergent structures. We find that regions showing clear evidence of divergence in higher order chromatin structure encompass at least 10% of the mammalian genome, and include many hundreds of genes whose regulation may have been affected. At least some of these genes have been directly implicated in evolutionary innovations to vertebrate developmental programmes, so divergent regions may have been disproportionately important during evolution. In addition, we show that divergent regions occur in large stretches of more than 2 Mb in the human genome and are enriched towards telomeres at the ends of human chromosomes. This may reflect shifts in the nuclear organisation and regulatory functions of chromatin domains between human and mouse.

physical contact maps, based on cross-linking frequencies, have been used to infer the spatial proximities and 3D- architecture between all possible 1 Mb segments of the human genome [11–13]. A familiar pattern of two spatial compartments within the nucleus also emerged from these data. One compartment composed of regions of gene rich, open, actively transcribed chromatin, and another containing regions with opposing features. These broad patterns emerge at the genome wide level, in spite of many regions that adopt cell type specific structures.

Remarkably, given the diverse methodologies used to investigate them, significant correlations have been found among some of these coarse-grained facets of higher-order genome organisation and function. There is a strong overlap between the sequences that replicate together during the same temporal window of S phase, and those sequences that can be captured together by Hi-C [12,14], consistent with the idea that genomic regions in close proximity tend to replicate at similar times and thereby define important features of chromosome organisation. These may well equate to the replication foci visible in the nucleus [15]. It has long been known that globally late replication tends to occur at the nuclear periphery [16,17] and this has been substantiated by more detailed analysis using fluorescence in situ hybridisation (FISH) of specific loci [7,14]. There is also a correlation between late replicating chromosomal domains and LADs [10] but it is not absolute and the relationship tends to breakdown at LAD borders and at particular genes. Moreover, such correlations present a moving target as genomic patterns of replication timing domains and LADs change upon differentiation and re-programming [8,10]. We also lack a comprehensive view of how genome-wide chromatin structure varies across cell types. Although cell type specific structures are clearly present, it seems that the higher order domains reflected in replication timing and Hi-C data remain largely unchanged over a variety of cell types and throughout the cell cycle [18,19]. Key questions in chromatin structure and nuclear organisation therefore relate to the ontology of the various structural domains that are known – namely how are they related and to what extent are they all aspects of the same entity?

Until recently there has been a lack of comparable, genome-wide chromatin structure data across species and comparative studies have therefore generally examined a single feature of chromatin structure in isolation. Ku et al [20] studied genome-wide Polycomb binding sites and histone modification data in mouse and human embryonic stem (ES cells) within orthologous promoter regions. They stressed the widespread conservation of chromatin states between species, with more than half of promoters showing the same state. Similarly, regions across the orthologous mammalian genome that are enriched for common histone modifications appear to be broadly conserved between human and mouse [21]. In contrast, sequence-specific transcription factor binding patterns appear to evolve rapidly in mammals, with binding events in a particular tissue shared only 10–22% of the time between human, mouse and dog genomes [22]. Higher order chromatin structures are generally assumed to show much less divergence, although detailed studies are rare. The numbers and size distributions of LADs in human lung fibroblasts are reported to be similar to those seen in mouse embryonic fibroblasts, as well as several other mouse cell types [10]. However it is not clear how the extent of divergence between cell types compares with divergence between species, or which genomic regions are involved in either. Replication timing appears generally conserved between human and mouse within large genomic regions showing conserved synteny, but notably less so than between orthologous human and mouse promoters [14]. This conservation has been maintained in spite of the numerous large-scale genome rearrangements separating the two species [23]. It also appears that the similarity in replication timing between species is heavily dependent on the particular cell type examined [14]. On the other hand, Hi-C data has suggested that the mouse and human genomes are separated into largely conserved, megabase sized interaction domains, that are similar between cell types [24].

The studies mentioned above provide complementary views of higher order chromatin structure. Each shows that the mammalian genome is organised into large, discrete domains of higher order chromatin with opposing properties (levels of expression and accessibility, spatial positioning, and replication timing). These domains appear to be broadly similar across the different cells that have been examined, although many regions across the genome show cell type specific structure [8,10,14]. However, the actual extent to which these datasets intersect, and how they relate to one another across cell types and species, is poorly understood. Similarly, the genomic loci underlying divergence in chromatin structure between species, and the mechanisms underlying divergence, are unknown. Here we collate a large number of diverse mouse and human datasets to provide the most comprehensive overview of higher order chromatin structure in mammals to date. We undertake a systematic study of all orthologous regions in the mammalian genome and document the extent of conservation in higher order chromatin structure between cell types and during evolution. Our analysis identifies large tracts of structurally divergent chromatin, unevenly distributed across the genome, and containing intriguing enrichments of particular classes of genes.

Results/Discussion

We conducted our analyses on 36 genome-wide datasets that measure three aspects of higher order chromatin structure and function in mouse and human: replication timing (RT) [7,14], nuclear lamina association (LA) [9,10] and genome-wide inter-locus contact preferences (Hi-C) [11,13]. The datasets were all generated using embryonic or pluripotent cells, with the exception of the Hi-C data (see Methods). All probe-based data were mapped to the latest genome assemblies using UCSC whole

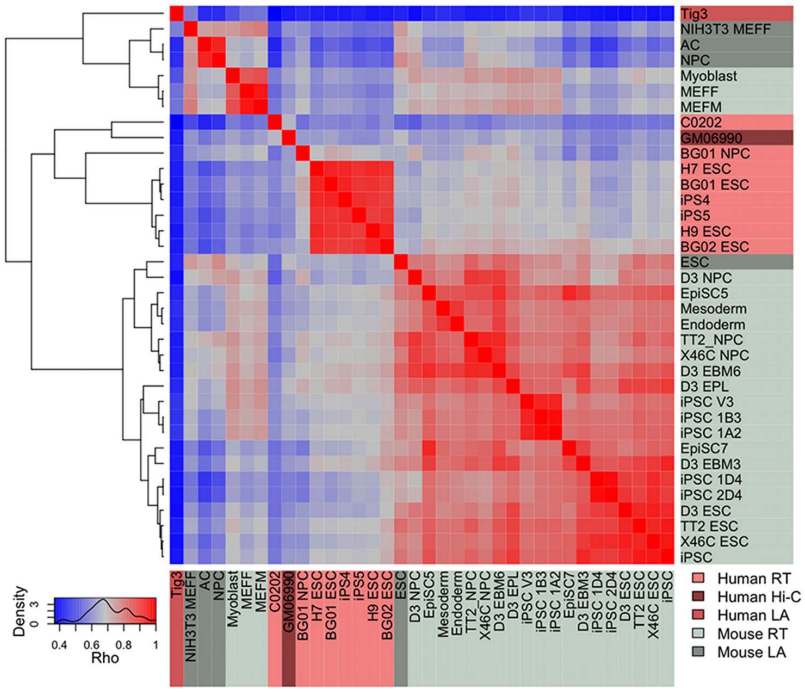

Figure 1. Global correlation matrix of higher order chromatin datasets. The heatmap and dendrogram show the relationships among 36 chromatin structure datasets (Spearman's rho: 0.38 to 0.98, p<1e-16). Datasets are labelled according to the experimental platform of origin: light grey = mouse RT, light pink = human RT, dark grey = mouse LA, medium pink = human LA, dark pink = human Hi-C.
doi:10.1371/journal.pcbi.1003017.g001

genome alignment data (hg19 and mm9), averaged into consecutive non-overlapping 100 Kb regions and collated by their genomic coordinates separately for human and mouse. Orthologous 100 Kb regions were identified conservatively by requiring reciprocal best match overlaps, both at the probe level and 100 Kb region level, between human and mouse genomes (see Methods). This resulted in 16,820 100 Kb regions represented in all higher order structure datasets in both mouse and human genomes. These orthologous regions encompass 54% of the human genome and 62% of the mouse genome. The distributions of the higher order data were examined to ensure global normalisation and scaling was appropriate and quantile normalisation was imposed across all datasets (see Methods). Prior to normalisation all primary datasets showed bimodal distributions with two peaks representing two distinct populations of higher order structure across the mammalian genome (Figure S1), consistent with previous observations [3,8,10,11]. We then addressed two related questions. Firstly, how well do these diverse datasets agree quantitatively? And secondly, what fraction of the mammalian genome can confidently be identified as structurally divergent?

Widespread conservation of mammalian higher order chromatin structure

Significant correlations were expected between replication timing (RT), lamin association (LA) and interlocus contact patterns (Hi-C) as they appear to reflect somewhat overlapping aspects of higher order chromatin structure [10,14,23]. The degree of agreement overall among the 36 datasets is indeed strong and significant (Spearman's Rho: 0.38 to 0.98, p<1e-16). In spite of differing experimental procedures, platforms, cell types, and species, moderate to strong positive correlations are ubiquitously observed (Figure 1). The highest agreement is usually observed between similar cell types from the same species, even across experimental platforms. For instance mouse RT data for a variety of ES and induced pluripotent stem cell (iPSC) types show strong correlations (Rho: 0.7–0.9, p<2.2e-16) with LAD data from mouse ES cells, and together they form a coherent cluster in the correlation matrix (Figure 1). However, there are also interesting exceptions to this rule, such as the human embryonic fibroblast LA data. Although this dataset shows the weakest correlations to all other datasets, the best agreement is to the mouse fibroblast LA and RT data and not to other human cell types. The reason for this may lie in cell cycle variation: ES and iPS data may be strongly influenced by the fact that these cells are almost entirely in S phase, whereas fibroblasts divide slowly and are mainly in G0/G1. In any case it seems that certain aspects of higher order structure in particular cell types, such as association with the nuclear periphery in fibroblasts, have been more strongly conserved than others during evolution.

Striking evidence of structural conservation across the mammalian genome is evident when examining contiguous stretches of orthologous regions (Figure 2). This suggests that many aspects of higher order chromatin structure have been conserved in embryonic cell types, over the ~80 million years since the divergence of rodents and primates. However apparent divergence

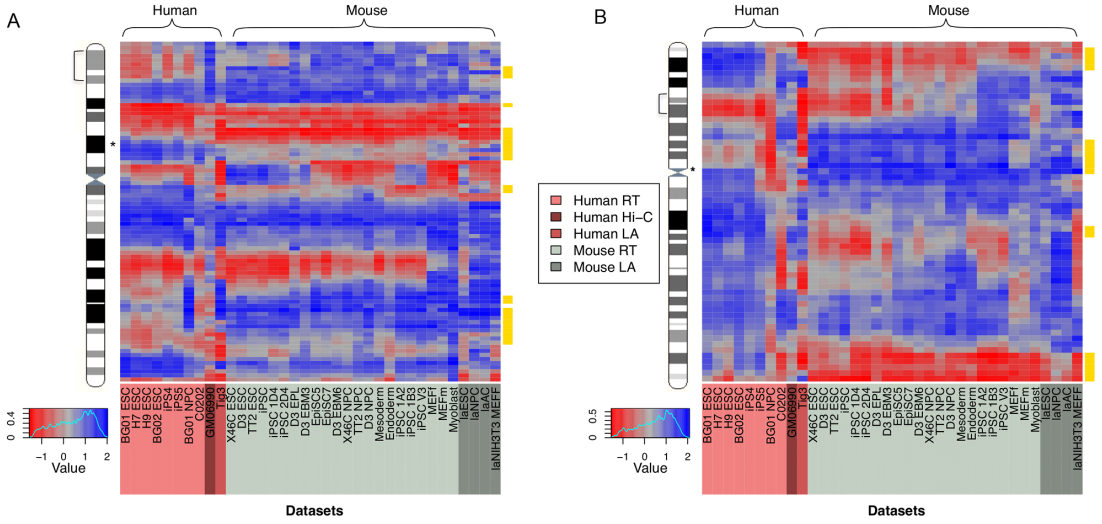

Figure 2. Specific human and mouse regions show significant divergence in higher-order chromatin structure. Human (pink) and mouse (grey) higher order chromatin structure across all cell types assayed, shown for two regions of the human genome: chromosome 11p15.2–15.4 (1.2–15 Mb) with the location of an OR gene cluster indicated by an asterisk (A); chromosome 7p14.3–15.3 (24–32 Mb) with the location of the HOXA gene cluster indicated by an asterisk (B). In each case the chromosome ideogram indicates the region expanded in the heatmaps with a square bracket. Consecutive, orthologous 100 kb regions are positioned on the y-axis with heatmap colours representing relatively open (blue) and closed (red) chromatin structures. Regions displaying significantly divergent chromatin structure are highlighted in yellow.
doi:10.1371/journal.pcbi.1003017.g002

in higher order chromatin structure between species is also evident in specific regions. This is most simply seen as loci demonstrating a strong, consistent difference in mean normalised structure between the two species across all of the available datasets (see representative regions depicted in Figure 2). Although there are high correlations between many of these datasets, reflecting similar overall trends in structure as we traverse chromosomes, this can mask substantial variation between datasets at the level of the absolute normalised structural values for a given 100 Kb region (Figure 2). The critical question is therefore, which 100 Kb regions vary between species to an unexpected degree, given the extent of variation seen among all datasets? This is the question we address below using a novel divergence metric based upon permutations of the original data.

We systematically sought genomic regions showing strong and consistent structural divergence between species, across all cell types, using non-parametric tests for each orthologous 100 Kb region (see Methods). The resulting p values were conservatively thresholded to ensure a low false discovery rate (FDR) and robust results. We defined two broad categories of regions based upon their levels of divergence: divergent regions (generating significant p-values passing the FDR threshold) and relatively static non-divergent regions (nonsignificant) (Figure 2; Figure S2). Viewed in this way divergence is necessarily bipolar, containing regions with mean structure values that are relatively open in human but closed in mouse, and vice versa. Such estimates of structural divergence are likely to be inherently conservative, since they depend upon strong consistent evidence for divergence over multiple cell types and experimental platforms. The divergent regions were found to constitute 10.22% (1,719 out of 16,820) of the orthologous regions examined, and possessed a similar (Mann-Whitney test in human $p = 0.17$, in mouse $p = 0.52$) protein-coding gene density to non-divergent regions. Human gene densities in nondivergent regions (2.34 per 100 Kb on average) were not significantly different from either human open divergent regions (2.09 per 100 Kb; Mann-Whitney $p = 0.45$), or human closed divergent regions (2.43 per 100 Kb; Mann-Whitney $p = 0.72$). Similarly, mouse gene densities in nondivergent regions (1.77 per 100 Kb) were not significantly different from either mouse open divergent regions (1.91 per 100 Kb; Mann-Whitney $p = 0.97$), or mouse closed divergent regions (1.33 per 100 Kb; Mann-Whitney $p = 0.51$). The distribution of divergent regions was far from uniform over the genome, with several chromosomes showing higher than expected densities (see Methods; Chi-squared test in human $p = 4.34e\text{-}06$, in mouse $p = 1.19e\text{-}03$). For instance, human chromosomes 5 and 10 were found to have a 50% excess of divergent regions, while chromosomes 21 and 22 were found to have a greater than 60% depletion. This raises the question: does the distribution of divergent regions within chromosomes reflect larger tracts of divergent chromatin?

Divergent chromatin is clustered within chromosomes

Cursory examination of these data (e.g. the regions depicted in Figure 2), suggests that a number of divergent 100 Kb regions are clustered in the genome at particular loci. We formally investigated the degree of clustering by measuring the length distribution of consecutive runs of divergent 100 Kb regions observed, relative to the distribution expected using a permutation strategy (see Methods). The clustering observed was found to be highly significant, and we identified 159 unexpectedly large (at least 400 Kb; $p<1e\text{-}04$) clusters of divergent regions with a median size of 800 Kb (Figure 3; Table S2). The same large orthologous clusters were detected in human and mouse genomes when the 100 Kb divergent regions in each genome were clustered (Figure S3), but were not evenly distributed across all chromosomes, for example human chromosomes 3 and 5 had around twice the density expected, but in contrast chromosomes 1 and 9 had around half the density expected. The size distribution of

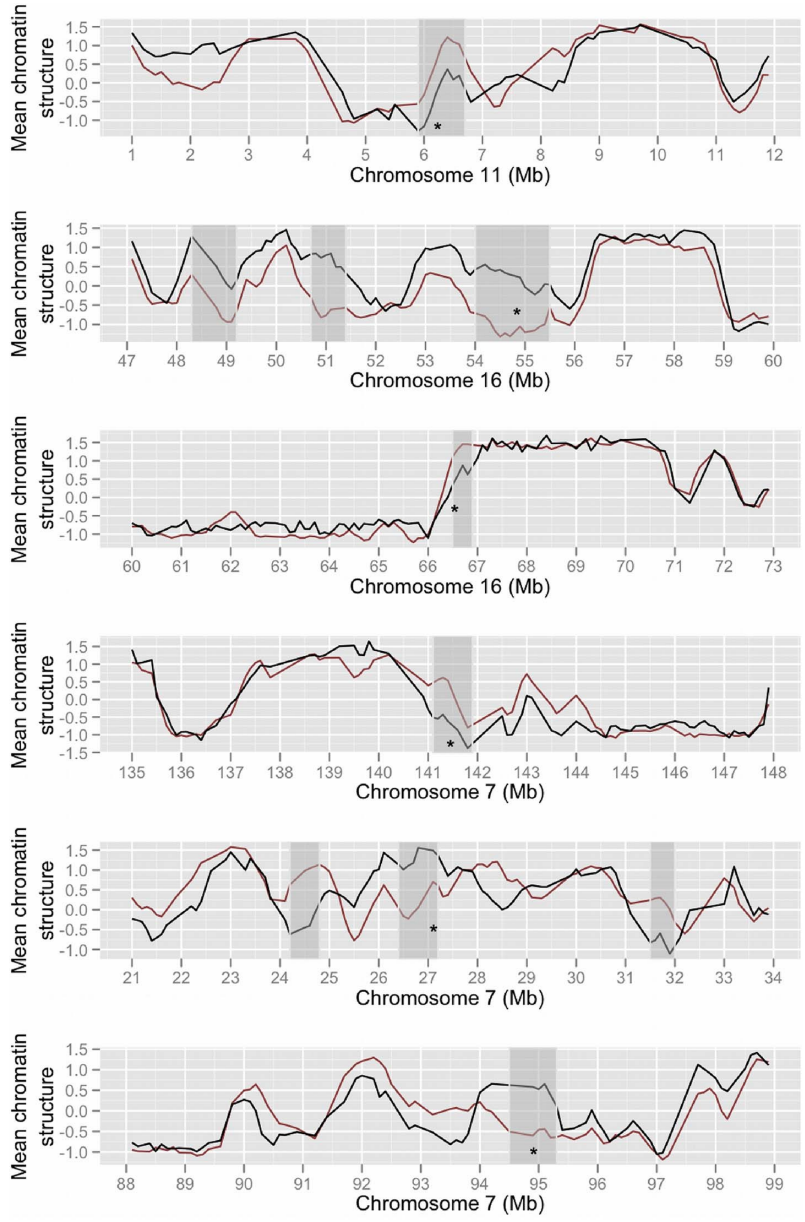

Figure 3. Clustering of divergent chromatin in the human genome. The line plot shows mean normalised human (black) and mouse (red) higher order chromatin structure values across human chromosomes. Unexpectedly large divergent areas are highlighted in grey. Asterisks indicate the positions of functionally enriched gene clusters listed in Table 2.
doi:10.1371/journal.pcbi.1003017.g003

divergence clusters appeared similar to the ES cell chromatin-mediated regulatory domains (median size 880 Kb) recently reported in the mouse and human genomes [24], suggesting that these stretches of divergent chromatin may represent divergent regulatory domains. We therefore examined the similarity in domain boundaries between these regulatory domains and the divergence clusters using a permutation approach (see Methods). An important caveat is the resolution of these datasets, which means that all reported domain boundaries are estimates within tens or hundreds of kilobases. In the human genome the median distance between the boundaries of divergence clusters and the nearest ES cell regulatory domain boundaries was 207,852 bp, which is somewhat less, though not significantly different ($p = 0.054$) from the expected median distance given 10,000 permuted datasets (235,581 bp). Similarly, in the mouse genome, the equivalent median distance was 260,000 bp, which is not

significantly different ($p = 0.087$) from the expected distance given 10,000 permuted datasets (290,095 bp). We conclude that overall there is no strong association between divergent regions and these regulatory domains, which is consistent with most structural divergence being selectively neutral. We also examined the correspondence between the divergent clusters and regions known to be structurally variable during cellular differentiation from ES cells [7]. Of the 1719 divergent regions, 60 overlapped these structurally dynamic regions, compared with an expected number (mean overlaps in 10,000 permutated datasets) of 99.73 which represents a significant depletion ($p<0.013$).

The three largest (2.1–2.7 Mb) regions of divergent chromatin were found to occupy subtelomeric regions of human chromosomes 2, 6 and 9 (Figure S4), but in each case the orthologous mouse regions were long distances (80–100 Mb) from mouse telomeres. This was found to reflect the distribution of chromatin divergence across the human genome in general, with unexpected excesses of divergence towards the ends of some human chromosomes (Figure S5; Table S3). This excess was most pronounced within the subtelomeric regions (within 5 Mb of the ends of each chromosome sequence assembly) of 4 human chromosomes (1, 2, 13, 18), and was also seen overall for the human genome ($p = 0.016$). In contrast most mouse (5 Mb) subtelomeric regions showed a relative depletion of divergence, with none showing significant enrichment, and (nonsignificant) depletion over the mouse genome in general. (No significant enrichment or depletion was found overall for pericentromeric regions in either species.) There are well-characterised differences in the chromatin structures found at human and mouse telomeres, and mammalian telomere biology appears to have been a focus for evolutionary adaptation [25]. Subtelomeric regions are known to be amongst the most rapidly evolving DNA sequences in the genome and have been subject to extensive divergence recently in the primate lineage [26]. The current data suggest that the higher order chromatin structures at some primate subtelomeric regions have also been subject to dramatic change.

Higher order chromatin structure itself is known to show strong positive correlations with GC content, such that relatively open regions are more GC rich and gene dense, and this is also seen here (Figure 4; Human GC density versus chromatin structure Spearman's $rho = 0.57$, $p<2.2e-16$; Mouse GC density versus chromatin structure Spearman's $rho = 0.75$, $p<2.2e-16$). Similarly, the human genome shows greater variability in GC content overall than in the mouse, consistent with the poor conservation of mammalian isochore structure in rodents [27]. The current data allow us to ask, for the first time, whether GC content is also associated with divergence in higher order structure. Comparison of the percentage of GC nucleotides between divergent and nondivergent regions across all orthologous 100 Kb regions shows intriguing contrasts between the human and mouse genomes (Figure 4). In the human genome there is a significant shift in human GC content between divergent and nondivergent regions, across the entire spectrum of normalised chromatin structure. Furthermore, this shift is to higher GC content (40.5%) within divergent human closed regions, and lower GC content (34.9%) within divergent human open regions, relative to nondivergent regions (37.5%; human divergent open GC versus human nondivergent GC Mann-Whitney $p<2.2e-16$; human divergent closed GC versus human nondivergent GC Mann-Whitney $p<2.2e-16$). Thus the two divergence classes show the opposite human GC content bias to the expectation e.g. although open chromatin in the human genome is relatively GC rich (Figure 4), divergent regions that are open in human actually tend to be GC poor. These patterns are not seen in the GC content of the mouse genome, where there is no contradictory shift in the compositional biases of mouse sequences within divergent regions. Instead mouse divergent open regions are relatively GC rich (38.7%) and divergent closed regions are relatively GC poor (33.4%), relative to nondivergent regions (35.5%). Correspondingly there is no global shift in mouse GC content between divergent and nondivergent regions (Figure 4). Thus overall, divergent regions are consistent with the GC content trends seen in the mouse genome, but show a complete contrast with the GC trends in the human genome. The magnitude of the human GC content shift varies between chromatin categories, as reflected in the varying separation between divergent and nondivergent regression lines (Figure 4). Further examination of these data suggests that the largest shifts are seen for regions towards the extreme ends (i.e. unusually open or closed) of the spectrum of chromatin structure categories (Table S1). It is not possible to disentangle cause and effect using the current data, to establish that changes in GC content drive structural change or vice versa. It is also not possible to establish which species has the derived or ancestral chromatin state. However, these observations do suggest that chromatin divergence is often associated with unusual shifts in GC content in the human lineage, which may reflect fluctuations in mutation or selection during primate evolution.

Chromatin divergence is associated with gene expression divergence in embryonic cells

If genes within divergent regions have undergone regulatory divergence we might expect to see some evidence of this in appropriate expression data. Although perfectly matched expression data is not available, the present data are mainly derived from embryonic cell types and previous studies have examined genome-wide regulatory divergence in human and mouse ES cells. Cai et al (2010) [28] sought significant differences in time-course expression patterns between mouse and human ES cells to rigorously measure regulatory divergence across orthologous genes. They were able to compile classes of genes showing either conserved regulation or divergent regulation in either mouse or human. We examined the distribution of these gene classes across all regions of divergent and nondivergent chromatin. Although the numbers of genes identified by Cai et al (2010) [28] that were also present within the orthologous regions studied here were modest (497 divergent and 126 conserved), we found enrichment (odds ratio: 1.30; Fisher's Exact test $p = 0.04$) of divergently regulated genes within the 100 Kb regions of divergent higher order chromatin reported here. Genes with conserved regulation were also under-represented in divergent regions (odds $ratio = 0.76$; $p = 0.331$). These patterns were observed in spite of the fact that the data of Cai et al (2010) [28] is based upon human and mouse embryonic cell lines that are not represented in the chromatin data studied here. Another more recent study of expression divergence between human and mouse genes, examined expression over a time course in specialised immune (macrophage) cells induced by exposure to bacterial lipopolysaccharide, and reported significant results for larger numbers (186 divergent, 972 conserved) of orthologous gene pairs [29]. We examined these data in the same way and found no significant enrichment of divergently regulated genes in divergent 100 Kb regions. Indeed the genes divergently regulated in these macrophage data showed the opposite trend, and were somewhat under-represented in regions of divergent chromatin (odds ratio: 0.78; $p = 0.46$). This suggests that the correspondence between chromatin divergence and expression divergence is specific to embryonic cell types.

We also constructed a larger dataset measuring differential expression between mouse and human ES cells for orthologous

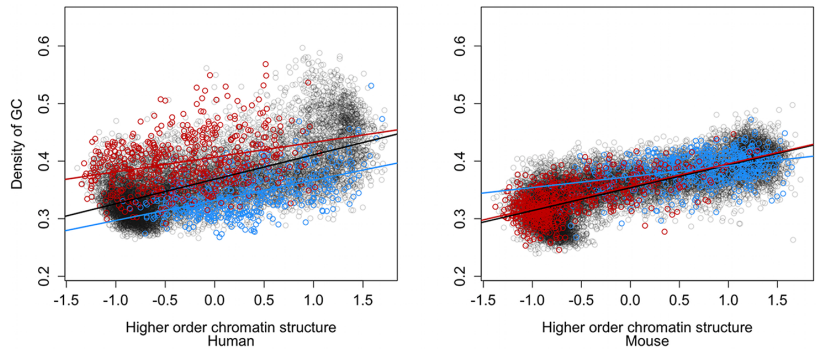

Figure 4. Chromatin divergence and GC content. Percentage of GC nucleotides within all 16,820 100 Kb orthologous regions across the spectrum of mean normalised chromatin structure values. The GC content and higher order structure values for human (left panel) are compared with the GC content and higher order structure values for mouse (right panel). Three classes of regions are shown with their least squares regression lines: nondivergent (grey), divergent open (blue) and divergent closed (red). Note that the bipolar classification of orthologous divergent regions (see text) means that human divergent open regions correspond to mouse divergent closed regions, and vice versa.
doi:10.1371/journal.pcbi.1003017.g004

gene pairs (see Methods), based upon previous RNAseq studies [30,31]. These data provide a higher coverage dataset consisting of log fold change measurements for 7,673 gene pairs occurring within the orthologous 100 Kb regions studied here. This allowed us to examine the extent of expression divergence within the two possible bipolar categories of divergent regions, relative to

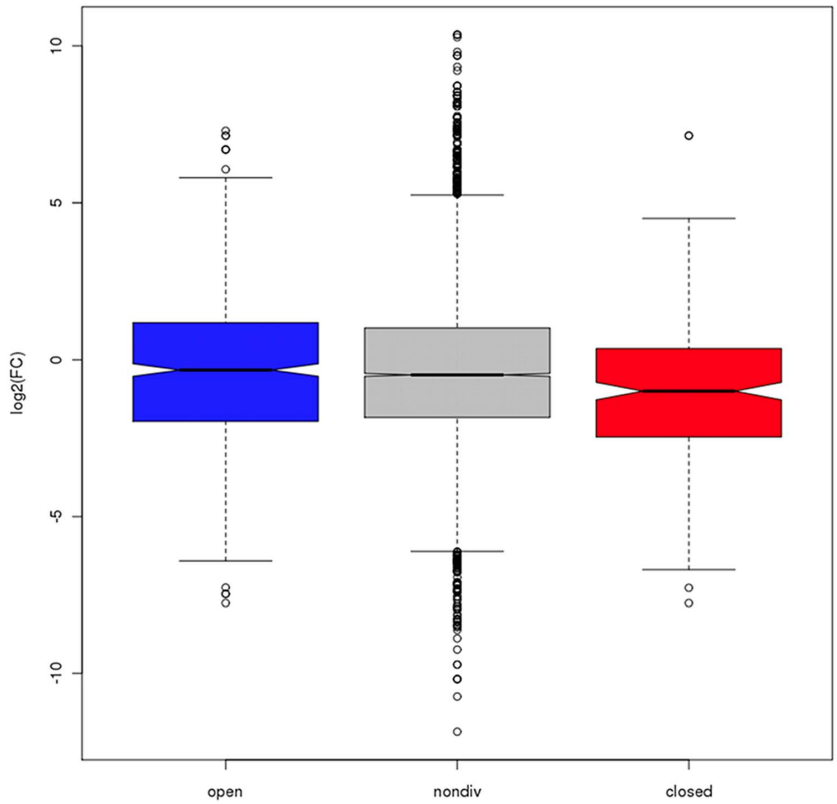

Figure 5. Chromatin divergence and expression divergence. Distributions of log2 fold change (log2(human/mouse expression)) for orthologous gene pairs within nondivergent regions (grey) and two classes of divergent regions: open in human but closed in mouse (blue), closed in human but open in mouse (red). For each class the bottom and top of the box show the lower and upper quartiles respectively around the median, and the width of the notches is proportional to the interquartile range.
doi:10.1371/journal.pcbi.1003017.g005

Table 1. The top 5 enriched human and mouse annotation terms for genes within divergent regions of higher order chromatin.

Annotation	Divergent regions	Term	Description	Gene #	P	FDR
Human	Human open/Mouse closed	CYTOBAND	11p15.4	15	1.70E-10	2.17E-07
		GO:0007606	Sensory perception of chemical stimulus	21	2.50E-09	4.15E-06
		GO:0050877	Neurological system process	41	1.42E-07	2.36E-04
		CYTOBAND	10p13	8	3.47E-07	4.44E-04
		GO:0007186	G-protein coupled receptor protein signaling pathway	36	3.81E-07	6.34E-04
Human	Human closed/Mouse open	IPR001827	Homeobox protein, antennapedia type	10	4.80E-07	7.33E-04
		CYTOBAND	18q23	6	5.63E-06	7.52E-03
		GO:0003002	Regionalization	21	8.65E-06	1.50E-02
		CYTOBAND	6q27	6	3.11E-05	4.15E-02
		CYTOBAND	2q37.3	9	3.28E-05	4.38E-02
Mouse	Human open/Mouse closed	GO:0007606	Sensory perception of chemical stimulus	39	2.19E-18	3.58E-15
		GO:0007608	Sensory perception of smell	34	5.80E-16	9.10E-13
		IPR000725	Olfactory receptor	33	7.94E-16	1.15E-12
		GO:0004984	Olfactory receptor activity	33	2.41E-15	3.45E-12
		IPR017452	GPCR, rhodopsin-like superfamily	47	3.73E-15	5.58E-12
Mouse	Human closed/Mouse open	GO:0003002	Regionalization	32	1.96E-09	3.39E-06
		GO:0009952	Anterior/posterior pattern formation	27	2.29E-09	3.97E-06
		GO:0007389	Pattern specification process	36	5.25E-09	9.09E-06
		CYTOBAND	2 45.0 cM	9	1.29E-08	1.89E-05
		CYTOBAND	19 D2	12	3.31E-08	4.84E-05

doi:10.1371/journal.pcbi.1003017.t001

nondivergent regions (Figure 5). We found a striking contrast, with regions open in human but closed in mouse showing a expression divergence consistent with upregulation of human genes (non-divergent median log2 fold change: −0.48; divergent: −0.33; Wilcoxon p = 0.23), while the opposite category (closed in human, open in mouse) showed evidence of upregulation of mouse genes (nondivergent: −0.48; divergent: −1.00; Wilcoxon p = 3.41e-06). This is the pattern of gene expression divergence expected within divergent regulatory domains conferring a respectively permissive or repressive environment for transcription of human genes. Again, these expression data were generated in embryonic cells similar to, but not identical to those used to derive the chromatin divergence data. It is important to note that there is a distinction between the relative bipolar classification of divergent regions (human open/mouse closed and vice versa) and their absolute normalised chromatin values. Thus, it is possible for a region that is relatively open in human and relatively closed in mouse to possess absolute values consistent with a closed conformation in both species. One might expect that using such absolute values to construct more specific divergent region categories might increase the differences seen (Figure 5). This was indeed the case in spite of the associated reductions in sample sizes. Regions open in human but closed in mouse (where the absolute human value > 0 and the absolute mouse value < 0) showed a stronger expression divergence consistent with upregulation of human genes (non-divergent median log2 fold change: −0.48; divergent: 5.03; Wilcoxon p<2.2e-16), while the opposite category (restricted to those with absolute human value<0 and absolute mouse value>0) showed stronger evidence of upregulation of mouse genes (nondivergent: −0.48; divergent: −4.77; Wilcoxon p>2.2e-16). These comparisons to expression data provide independent validation of our methodology and suggest a direct link between the regions of divergent chromatin identified and the regulation of resident genes.

Regions of divergent chromatin structure harbour developmental gene clusters

Using standard enrichment analyses, we identified over-representation of particular functional classes of genes in the divergent orthologous regions, and the results establish interesting themes. The 907 divergent 100 Kb regions relatively open in human (but closed in mouse) contain 1142 human genes and 757 mouse genes, and both show significant enrichments for multiple terms associated with olfactory receptors (ORs) at particular loci (seen as enrichments for genes mapping to particular cytogenetic bands) (Table 1; Table S2). The mouse genes involved are disproportionately those located in particular OR gene clusters on chromosome 7E3 and 6B1-B2.1, while the human genes are clustered at the orthologous locations at 11p15.4 (Figure 2A) and 7q35 respectively, within extended regions of conserved synteny. Mouse OR genes have been shown to exhibit tightly regulated expression patterns during development, dependent upon repressive chromatin structures spanning clusters of OR genes [32], including histone modifications associated with constitutive heterochromatin [33]. This raises the intriguing possibility of an association between divergent higher order chromatin structures and particular histone modifications. It also suggests that the repressive, relatively closed higher order chromatin structures consistently seen at this region of the mouse genome, but not evident in human cells, could have evolved as part of the regulatory landscape associated with OR gene cluster evolution in rodents.

Other enriched terms include those related to a protocadherin (Pcdh) gene cluster present at 5q31.3 in the human genome, and to

the orthologous mouse Pcdh cluster on mouse chromosome 18qB3 (Table S2). Recent work has shown this region adopts distinct chromatin architectures in different mouse neuronal cell types to affect Pcdh gene expression and thereby plays critical roles in establishing neuronal diversity and connectivity during development [34]. A third cluster of genes coincides with this class of divergent regions (open in human, closed in mouse) on mouse chromosome 8D3 (and human 16q21) and is enriched for genes encoding MARVEL, a transmembrane domain involved in membrane apposition. The family of chemokine-like proteins containing this domain have been implicated in inflammation, immunity and development but most are not well characterised. Of the five MARVEL containing genes within the 8D3 divergent cluster, three are unstudied, but Cmtm2a and Cmtm3 are both implicated in the proliferation and development of particular testicular cells [35,36]. The human ortholog of Cmtm3 is present in the orthologous human divergent region at 16q21 and is a known tumour suppressor gene that shows frequent inactivation via chromatin-mediated silencing in several cancers [37].It seems that developmental gene clusters showing cell type specific regulation are unexpectedly common at regions displaying divergent higher order chromatin. Other clusters of genes, enriched at other divergent regions are also present in the results but lack sufficient functional annotation to generate significant enrichment results after multiple testing corrections (Table S2).

The genes within the divergent 812 orthologous human closed (mouse open) regions contain 1285 human genes and 1102 mouse genes. These also showed significant enrichment for genomic regions harbouring particular gene clusters. Both human and mouse genes in these regions show significant enrichment for terms associated with developmental genes containing Antennapedia type homeobox domains (IPR001827). The genes involved are exemplar developmental genes present at the HOXA (human HOXA1-A7; Figure 2B) and HOXD (human HOXD1-4) clusters. Both clusters are implicated in multiple cancers and other disorders, and are tightly regulated via higher order chromatin domains [38,39]. It is thought that structural divergence within the chromatin domains harbouring these clusters underlies many important innovations in the vertebrate body plan [40]. Other, relatively poorly studied, homeodomain containing genes at other loci are also present within this class of (human closed, mouse open) divergent regions (Table S2). Again, it seems that

Table 2. The top 5 enriched human annotation terms for genes within large regions of divergent higher order chromatin.

Cluster	Term	Description	Gene #	P	FDR
chr11:5900000–6699999	CYTOBAND	11p15.4	15	3.74E-28	2.05E-25
	PIRSF038651	G Protein-Coupled Olfactory Receptor, Class I	7	2.76E-10	1.96E-07
	GO:0007608	Sensory Perception Of Smell	8	1.43E-08	2.02E-05
	GO:0007606	Sensory Perception Of Chemical Stimulus	8	6.09E-08	8.59E-05
	IPR000725	Olfactory Receptor	7	6.10E-08	5.77E-05
chr16:54000000–55499999	IPR003893	Iroquois-Class Homeodomain Protein	3	4.26E-07	1.34E-04
	IPR001356	Homeobox	3	2.96E-04	9.32E-02
	IPR017970	Homeobox, Conserved Site	3	3.00E-04	9.45E-02
	IPR012287	Homeodomain-Related	3	3.21E-04	1.01E-01
	CYTOBAND	16q11.2-Q13	2	3.88E-04	1.22E-01
chr16:66500000–66899999	IPR008253	Marvel	5	1.10E-09	7.09E-07
	GO:0042330	Taxis	5	8.79E-07	8.71E-04
	GO:0006935	Chemotaxis	5	8.79E-07	8.71E-04
	GO:0005125	Cytokine Activity	5	6.40E-06	5.41E-03
	GO:0007626	Locomotory Behavior	5	1.04E-05	1.04E-02
chr7:141100000–141899999	CYTOBAND	7q31.3-Q32	3	1.69E-06	7.37E-04
	GO:0008527	Taste Receptor Activity	3	9.84E-06	7.98E-03
	IPR007960	Mammalian Taste Receptor	3	1.02E-05	6.12E-03
	GO:0050909	Sensory Perception Of Taste	3	9.69E-05	9.20E-02
	GO:0007186	G-Protein Coupled Receptor Protein Signaling Pathway	4	2.43E-03	2.28E+00
chr7:26400000–27199999	IPR001827	Homeobox Protein, Antennapedia Type, Conserved Site	7	1.54E-16	4.44E-14
	CYTOBAND	7p15-P14	6	1.16E-13	4.41E-11
	GO:0048562	Embryonic Organ Morphogenesis	7	6.27E-12	7.55E-09
	GO:0009952	Anterior/Posterior Pattern Formation	7	8.06E-12	9.70E-09
	GO:0048568	Embryonic Organ Development	7	2.11E-11	2.55E-08
chr7:94500000–95299999	CYTOBAND	7q21.3	4	3.06E-08	1.17E-05
	GO:0004063	Aryldialkylphosphatase Activity	3	4.22E-07	3.35E-04
	IPR002640	Arylesterase	3	5.11E-07	2.94E-04
	GO:0004064	Arylesterase Activity	3	8.44E-07	6.69E-04
	PIRSF016435	Paraoxonase	3	1.29E-06	1.29E-04

doi:10.1371/journal.pcbi.1003017.t002

developmentally regulated genes are over-represented within regions of divergent chromatin. However, it is worth noting that the proportion of divergent regions generating significant functional enrichments (that is, those divergent regions possessing the genes responsible for the functional enrichments seen) is modest overall, constituting 6% of human and 11% of mouse divergent regions in total.

Most RNA genes are poorly functionally annotated which makes analogous enrichment analyses impossible, but we did examine the densities of the main RNA gene classes (rRNA, snoRNA, snRNA, miRNA, lincRNA) in structurally divergent regions. Only the lincRNA class showed significant differences, with higher densities of both human (divergent mean density: 0.31 genes/Mb; nondivergent mean density: 0.20 genes/Mb; Wilcoxon $p = 1.48e-08$) and mouse (divergent mean density: 0.12; non-divergent mean density: 0.09; Mann-Whitney $p = 3.68e-04$) lincRNA genes found in divergent (human closed/mouse open) regions. These molecules are thought to regulate ES cell differentiation via the assembly of chromatin complexes and the establishment of activating or repressive domains [23]. The present data suggest they may also have played roles in chromatin divergence.

As expected the large divergence clusters showed similar patterns of functional enrichments as those discussed above (Table 2; Table S2). For example, the divergent regions mentioned already at 11p15.4 (containing an OR gene cluster) and 16q12.2 (containing an IRX gene cluster) were found to extend across 800 Kb and 1.5 Mb respectively. Similarly the divergent region containing the 7p15.2 HOXA genes was found to encompass 800 Kb, and to include neighbouring lincRNA genes such as HOTAIRM1 which is active in HOXA regulation during neurogenesis and differentiation [41]. An additional region at 7q21.3 showing a novel functional enrichment also emerged, which contains the paraoxonase gene cluster (Table 2), these genes are imprinted in the mouse genome and exhibit unusual, allele-specific expression dependent on developmental stage in human cells [42]. Again, it seems that structural divergence is disproportionately associated with particular developmental gene clusters, which follow tightly regulated expression patterns targeting specific cell types, and are often known to occupy unusual chromatin environments. Many of these genes have also been implicated in developmental adaptations during vertebrate evolution and in human disease processes. This may suggest that regions of divergent chromatin structure have evolved different chromatin conformations to facilitate functional divergence at these loci. However it is not possible to exclude non-adaptive hypotheses, for example where divergence in chromatin structure is a neutral consequence of gene family or repeat expansions or other changes in the underlying genomic sequences. Indeed, since the majority of divergent regions show no detectable functional enrichments, selectively neutral divergence appears to be the most likely scenario in most cases.

Conclusions

Individual studies of various aspects of higher order chromatin structure have suggested widespread conservation across the mammalian genome, in spite of many interesting structural differences between cell types [10,14,23]. The comprehensive analyses presented here are consistent with this, and demonstrate the same signal across diverse datasets from studies that set out to observe nominally different aspects of structural genome organisation in many different embryonic cell types. We conclude that most measurable aspects of chromatin are conserved across the vast majority of the detectably orthologous genome. However, using a conservative approach (requiring consistent evidence of divergence between species over all cell types and all structural datasets assayed) we also observe divergent chromatin structure at 10.22% of orthologous 100 Kb genomic regions examined, encompassing over 170 Mb and including many hundreds of human and mouse genes. This suggests that structural divergence has played a major role in the evolution of many loci occupying these unusual genomic regions. Many of the regions identified form unexpectedly large tracts of divergent chromatin, nonrandomly distributed between and within chromosomes, and this clustering appears particularly pronounced at human subtelomeric regions. Overall the divergent regions of embryonic chromatin identified are significantly enriched for genes active in vertebrate development. These include homeodomain gene clusters, which have been implicated in evolutionary innovations to vertebrate developmental programmes, suggesting that selection may have modulated their regulation during evolution via alterations to chromatin. Consistent with this we find that genes showing evidence of regulatory divergence between human and mouse are over-represented within regions of divergent higher order chromatin structure.

The mechanisms underlying divergence in higher order chromatin structure remain unknown, but one may speculate that alterations at lower levels of chromatin are likely to be involved. For example, changes in the diversity or abundance of relatively rapidly evolving ncRNAs, which can mediate chromatin remodelling between cell types [43], could provide a molecular basis for divergence. Also the strong sequence-level correlates of human chromatin structure [44,45] and the unusual, lineage specific shifts in GC content seen here, suggest it is possible that sequence divergence underlies chromatin divergence. It may also be relevant that larger scale variation in chromatin structure within the mammalian genome is often associated with alterations in the spectrum of histone modifications at a region. For example, human LADs are reported to show enrichments of H3K9 and H3K27 methylation [46], and OR gene clusters are now known to possess an unusual signature of histone modifications involving the molecular hallmarks of constitutive heterochromatin [33]. It is therefore possible that divergence in chromatin domains during evolution is caused by alterations in the constellations of histone modifications present. However, definitive evidence of the mechanisms underlying evolutionary divergence in higher order chromatin structure will require substantial future investigations.

Methods

Higher order chromatin structure data

All cell types and datasets, and their abbreviations are listed in Table S5. Replication timing data in human and mouse embryonic cells were obtained from Hiratani et al [7], and Ryba et al [14] as log2(early relicating/late replicating) values. Nuclear lamina association data in human and mouse embryonic cells were obtained from Guelen et al [9] and Peric-Hupkes et al [10]. Both studies were based upon the DamID technique for labelling lamina associated sequence, where relative lamina association is represented by log2(Dam-fusion/Dam-only) values. Finally, 100 Kb window genomic interaction probability matrix eigenvalues were defined for human lymphoblastoid cells using Hi-C by Lieberman-Aiden et al [11]. These values were found to largely reflect two relatively open and closed nuclear compartments of higher order chromatin. Although these data were not derived from embryonic cells it appears that many of the higher order patterns (as represented by interaction matrix eigenvectors) in Hi-C datasets are consistent between cell types [11,24]. Re-analysis of

these interaction data has revealed the presence of systematic biases that afflict the Hi-C method, obscuring additional, finer scale structural compartments [12]. Although our analysis only concerns the course grained, two-compartment division between open and closed regions (since we use eigenvalues of interaction matrices not interaction probabilities themselves) we were concerned that our results might be affected by these biases. Consequently we examined an independent genomic interaction map produced for a similar lymphoblastoid cell line using a modified Hi-C method designed to mitigate the biases inherent in previous data [13]. When the original [11] interaction data were substituted with these new, nominally unbiased [13] data we observed very similar correlations with all other chromatin structure datasets. We conclude that the biases present in the Lieberman-Aiden et al [11] dataset have little effect on a course grained, two compartment classification of the genome based upon these data, and therefore that our search for structurally divergent regions is unaffected.

Orthology and divergence

Probe based replication timing and nuclear lamina association data coordinates were translated to the latest human or mouse genome assembly coordinates (hg19 and mm9) using reciprocal liftOver transformations to ensure accurate remapping [47]. Probes failing to map reciprocally to overlapping coordinates between mouse and human genomes were discarded as unreliable. For each dataset the structural data values were averaged across probes into consecutive non-overlapping 100 Kb regions, but regions represented by fewer than 10 probes were discarded as potentially unreliable. This allowed comparisons between the probe based datasets and the Hi-C data, which has a fixed resolution of 100 Kb. Within each species 100 Kb regions were collated across datasets where their coordinates overlapped by 50% or more. The result was a set of 24,711 mouse and 28,786 human 100 Kb regions represented by higher order structural values from multiple datasets. Orthologous 100 Kb regions were defined as those regions with at least a 50% coordinate overlap between mouse and human genomes using reciprocal liftOver transformations. A total of 16,820 100 Kb orthologous regions, covering 54% of the human genome and 62% of the mouse genome, were defined in this way. A total of 11,966 human and 7,891 mouse regions, lacking an orthologous mapping using this protocol, were designated putatively lineage specific regions. As expected, lineage specific regions were highly enriched for segmental duplications, repeats and duplicated gene families, whereas orthologous regions were relatively rich in protein coding genes [48]. Examination of several techniques revealed that standard quantile normalisation procedures (R/Bioconductor limma package) [49] used to normalise across different microarray experiments were effective across the different experimental platforms and cell types here, therefore this normalisation technique was implemented across all structural datasets for all 100 Kb regions (Figure S1; Figure S7). The normalised structural data and chromosome coordinates for all 16,820 orthologous regions are provided in Table S6.

Structurally divergent regions were defined as orthologous 100 kb regions that showed a consistent difference in higher order structural values across human and mouse data. Non-parametric tests from the SAM package [50], analogous to two class unpaired t-tests with permutation derived p-values, were used to assess divergence (R package samr). These tests were developed for microarray data analysis but are appropriate for other types of non-microarray derived data [50]. The approach was developed to identify unusual genes that show a strong and consistent expression difference between treatments, given many variable replicate measurements. In the present case we identify unusual 100 Kb regions, showing a strong and consistent difference between species, given the many variable measurements of chromatin structure. In both cases the aim is to identify significant differences between states (treatments, species) for the measured entities (genes, 100 Kb regions) given a number of inherently noisy, variable observations. The permutation approach ensures that the observed variability in the observations is accounted for in the significance of the test result. Tests were carried out for each 100 Kb orthologous region, with the various normalised structural values for that region compared between species. 100,000 permutations of the normalised structure dataset were used to estimate the false discovery rate (FDR), defined in this instance as the median number of false positive divergent regions expected (given the permuted datasets), divided by the total number of divergent regions called. The FDR threshold was set to be relatively low (FDR = 2e-04) to ensure that less than 1 false positive was expected within the 1719 divergent regions found. The results are necessarily bipolar with positive and negative divergent regions called to indicate human open/mouse closed or human closed/mouse open divergence respectively. Relatively static, nondivergent regions were classed as those with p values that did not pass the FDR threshold. The mean normalised structure values for 100 Kb regions, over all of the available datasets in a species, were calculated as a useful guide to trends in structure across chromosomes and the genome overall.

The 100 Kb detectably orthologous regions defined above (using a 50% overlap threshold) will necessarily vary in the degree of similarity they show between species, it was therefore a concern that this might influence the measurement of structural divergence. Specifically it was important to show that the regions identified as structurally divergent are not simply those most poorly aligned between species at the sequence level. On closer examination the distributions of overlaps (aligned nucleotides minus gaps) were found to be very similar between structurally divergent and nondivergent regions, whether viewed in terms of the human (hg19) genome (divergent overlap mean = 0.80, median = 0.81; nondivergent overlap mean = 0.79, median = 0.80), or the mouse (mm9) genome (divergent overlap mean = 0.73, median = 0.72; nondivergent overlap mean = 0.72, median = 0.71) sequence assemblies, based upon UCSC whole genome alignments. We concluded that our estimates of structural divergence are not a simple reflection of sequence divergence.

Distribution and gene content of divergent regions

We examined the distribution of divergent regions across chromosomes by comparing the expected numbers, given the proportion of orthologous 100 Kb regions on each chromosome, with those observed using chi-squared tests, and identified chromosomes of interest as those generating standardized residuals>1.96. To define divergence clusters (i.e. clustered groups of divergent 100 Kb regions) we first identified all consecutive runs of significantly divergent regions across the orthologous human (and separately the mouse) genome, and the observed distribution of their lengths. Consecutive runs were required to maintain the polarity of divergence (i.e. all regions involved must be either human open/mouse closed or vice versa). We then permuted the divergence data among orthologous 100 Kb regions within chromosomes 10,000 times, and noted the length distributions of consecutive runs within each permuted genome. The frequency with which a run of n consecutive divergent 100 Kb regions was seen in the permuted datasets was taken as an approximate p value for runs of length n in the observed dataset. Observed runs of

divergent regions greater than or equal to 400 Kb were never seen in the permutated data ($p<0.0001$) and were taken to be significant divergence clusters. This strategy is likely to be conservative in detecting large regions of divergent chromatin as it does not allow for gaps (e.g. regions that may have marginally failed to reach significance in the test for divergence above) within runs of divergent regions. 159 large divergent regions were discovered at the same, orthologous locations in the human and mouse genomes (Table S3). An additional 1.4 Mb divergent region (at chr18: 11600000–12999999) was found in the mouse genome that lacked a reciprocally orthologous human region.

Enrichment or depletion of 100 Kb divergent regions within subtelomeric or centromeric regions was assessed using a circular permutation strategy [51] to preserve the observed degree of clustering, over 10,000 permuted datasets. Each permuted dataset was generated by shifting the locations of all divergent regions on each chromosome by a random number (less than the length of the chromosome). Regions assigned a shifted position greater than the final base pair of the chromosome are reassigned to the start of that chromosome (plus the number of bases by which they exceeded the final base pair). Thus the permutations regard chromosomes as circularised, and thereby maintain the degree of clustering seen among the observed divergent regions. The number of permuted datasets, n, possessing a number of divergent regions within subtelomeric (or centromeric) regions greater than or equal to the observed number were noted, and used to calculate approximate p-values (n/10,000) for enrichment. The significance of depletion was calculated analogously, according to the number of permuted datasets possessing the same or fewer divergent regions. Subtelomeric regions were defined as regions within 1 Mb, 5 Mb and 10 Mb of the first and final base pairs of the chromosome assemblies, and within the final base pair of the (acrocentric) mouse assemblies. Centromeric regions were defined as regions within 1 Mb, 5 Mb and 10 Mb of the first base pair of mouse and human chromosome q arm assemblies, and within the final base pair of human p arm assemblies. It is important to note that the density of orthologous 100 Kb regions within subtelomeric regions was not significantly different from the genome as a whole, either for human (5 Mb subtelomeric region mean density = 23.70; mean density across all genomic 5 Mb bins = 28.10) or mouse (5 Mb subtelomeric region mean density = 34.60; mean density across all genomic 5 Mb bins = 34.20). The same circular permutation approach was used to measure the enrichment or depletion of divergent regions within domains that are structurally dynamic during cellular differentiation [7]. We also used a similar permutation strategy to compare the similarity (i.e. proximity) of domain boundaries between chromatin-mediated regulatory domains [24] and the boundaries of divergent clusters. The median distance between divergent cluster boundaries and the nearest regulatory domain boundaries was compared to the median distance seen in 10,000 datasets that had undergone circular permutation. The proportion of datasets generating a median distance less than or equal to the observed median distance was taken as an approximate p-value.

Gene densities were calculated per Mb for divergent and nondivergent datasets and tested using nonparametric (Mann-Whitney/Wilcoxon test) statistics. Functional enrichments for protein coding genes were calculated using DAVID [52] using the total human and mouse genes present within the 16,820 orthologous 100 Kb regions as background sets for human and mouse enrichment analyses respectively. Enrichment of each annotation term in the set of human or mouse genes present within divergent regions was assessed using default options (p-values calculated using the hypergeometric distribution with FDR correction). Enrichment of these gene sets within cytogenetic bands was also examined as this can reflect the clustering of divergent regions. Both protein coding and RNA genes were annotated by Ensembl (http://www.ensembl.org) and include lincRNAs predicted according to combinations of histone modifications and complementary EST and cDNA data. RPKM expression values for human H1 ES cells [30] and mouse E14 ES cells [31] were used to calculate log2(human RPKM/mouse RPKM) for all one to one orthologous mouse human Ensembl gene pairs, as an estimate of fold change in expression.

Supporting Information

Figure S1 Structural data distributions. The bimodal distributions of higher order structural data for all orthologous 100 Kb regions before normalisation with two peaks representing two distinct populations of higher order structure across the mammalian genome. Human and mouse RT data, LA data, and human Hi-C data are shown.
(JPG)

Figure S2 Quantifying human-mouse divergence in higher-order chromatin structure. The Q-Q plot from the two class unpaired SAM tests (see Methods) for each orthologous 100 Kb region. Significantly divergent regions (highlighted in green and red) generate unexpectedly extreme observed test scores relative to the expected (permutation based) scores.
(JPG)

Figure S3 Distribution of mammalian divergence clusters. Large human divergent regions (red) are shown with the orthologous positions of large mouse (blue) divergent regions in the human genome.
(JPG)

Figure S4 The three largest divergence clusters on human chromosomes. The line plot shows mean normalised human (black) and mouse (red) higher order chromatin structure across human chromosomes. Unexpectedly large divergent areas are highlighted in grey.
(JPG)

Figure S5 Distribution of structural divergence across the human and mouse genomes. The occurrence of divergent orthologous 100 Kb regions across human (top panel) and mouse (bottom panel) chromosomes. In each species the divergent regions found to be relatively open (blue) or relatively closed (red) within that species are indicated.
(JPG)

Figure S6 Enriched functional classes within divergent regions. The relationships between enriched GO terms for genes within divergent 100 Kb regions, related terms are coloured similarly and the areas ascribed to each term reflect the significance of their enrichment.
(JPG)

Figure S7 Structural data distributions after normalisation. The identical bimodal distributions of higher order structural data across all orthologous 100 Kb regions, after quantile normalisation. Representative datasets of human (BG01) and mouse (iPSC V3) RT data, human (Tig3) and mouse (NIH3T3) LA data, and human Hi-C data (GM06990) are shown, both separately and together (All).
(JPG)

Table S1 GC content and structural divergence. Percentage of GC nucleotides within all 16,820 100 Kb orthologous

regions across the spectrum of normalised chromatin structure values as in Figure 4. The GC content difference between divergent and nondivergent regions is shown for each binned category of higher order structure, together with the significance of the difference according to Mann-Whitney tests.
(DOCX)

Table S2 Full functional enrichment results. Functional enrichment results for all classes and clusters of divergent regions.
(XLS)

Table S3 Full divergent region details. All divergent orthologous regions discovered.
(XLS)

Table S4 Enrichment of divergence clusters at subtelomeric regions. Results of permutation tests (see Methods) assessing the significance of observed relative to expected numbers of divergence clusters at a variety of proximities (1 Mb, 5 Mb, 10 Mb) to telomeres in human and mouse genomes. Significant ($p<0.05$) enrichments (labelled E) or depletions (labelled D) in observed relative to expected numbers are highlighted in yellow.
(XLS)

Table S5 Cell types and datasets. Details of the cell lines, data types and embryonic stages in this study.
(DOC)

Table S6 Full orthologous region details. Structural data for all orthologous regions examined.
(CSV)

Acknowledgments

We would like to thank Martin Taylor, Chris Ponting, James Prendergast, and three anonymous reviewers for their helpful comments and discussions regarding this project.

Author Contributions

Conceived and designed the experiments: CAS. Performed the experiments: EVC. Analyzed the data: EVC. Contributed reagents/materials/analysis tools: EVC CAS. Wrote the paper: EVC WAB CAS.

References

1. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, et al. (2009) Parental origin of sequence variants associated with complex diseases. Nature 462: 868–874.
2. Pegoraro G, Misteli T (2009) The central role of chromatin maintenance in aging. Aging 1: 1017–1022.
3. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, et al. (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell 118: 555–566.
4. Woodcock CL, Ghosh RP (2010) Chromatin higher-order structure and dynamics. Cold Spring Harb Perspect Biol 2: a000596.
5. Zhang Z, Pugh BF (2011) High-resolution genome-wide mapping of the primary structure of chromatin. Cell 144: 175–186.
6. Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet 6: 775–781.
7. Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, et al. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol 6: e245.
8. Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, et al. (2010) Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res 20: 155–169.
9. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, et al. (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453: 948–951.
10. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SWM, Solovei I, et al. (2010) Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell 38: 603–613.
11. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293.
12. Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 43: 1059–1065.
13. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol 30: 90–98.
14. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, et al. (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res 20: 761–770.
15. Gilbert DM (2001) Nuclear position leaves its mark on replication timing. J Cell Biol 152: F11–15.
16. O'Keefe RT, Henderson SC, Spector DL (1992) Dynamic organization of DNA replication in mammalian cell nuclei: spatially and temporally defined replication of chromosome-specific alpha-satellite DNA sequences. J Cell Biol 116: 1095–1110.
17. Dimitrova DS, Gilbert DM (1999) The spatial position and replication timing of chromosomal domains are both established in early G1 phase. Mol Cell 4: 983–993.
18. Baker A, Audit B, Chen CL, Moindrot B, Leleu A, et al. (2012) Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. PLoS Comput Biol 8: e1002443.
19. Moindrot B, Audit B, Klous P, Baker A, Thermes C, et al.(2012) 3D chromatin conformation correlates with replication timing and is conserved in resting cells. Nucleic Acids Res 40: 9470–81.
20. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, et al. (2008) Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet 4: e1000242.
21. Woo YH, Li W-H (2012) Evolutionary conservation of histone modifications in mammals. Mol Biol Evol 29: 1757–1767.
22. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, et al. (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328: 1036–1040.
23. Yaffe E, Farkash-Amar S, Polten A, Yakhini Z, Tanay A, et al. (2010) Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. PLoS Genet 6: e1001011.
24. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376–380.
25. Gomes NMV, Ryder OA, Houck ML, Charter SJ, Walker W, et al. (2011) Comparative biology of mammalian telomeres: hypotheses on ancestral states and the roles of telomeres in longevity determination. Aging Cell 10: 761–768.
26. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, et al. (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 437: 94–100.
27. Costantini M, Cammarano R, Bernardi G (2009) The evolution of isochore patterns in vertebrate genomes. BMC Genomics 10: 146.
28. Cai J, Xie D, Fan Z, Chipperfield H, Marden J, et al. (2010) Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells. PLoS Comput Biol 6: e1000707.
29. Schroder K, Irvine KM, Taylor MS, Bokil NJ, Le Cao K-A, et al. (2012) Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. Proc Natl Acad Sci USA 109: E944–953.
30. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315–322.
31. Xiao S, Xie D, Cao X, Yu P, Xing X, et al. (2012) Comparative epigenomic annotation of regulatory DNA. Cell 149: 1381–1392.
32. McClintock TS (2010) Achieving singularity in mammalian odorant receptor gene choice. Chem Senses 35: 447–457.
33. Magklara A, Yen A, Colquitt BM, Clowney EJ, Allen W, et al. (2011) An epigenetic signature for monoallelic olfactory receptor expression. Cell 145: 555–570.
34. Hirayama T, Tarusawa E, Yoshimura Y, Galjart N, Yagi T (2012) CTCF Is Required for Neural Development and Stochastic Expression of Clustered Pcdh Genes in Neurons. Cell Rep 2: 345–357.
35. Wang Y, Li T, Qiu X, Mo X, Zhang Y, et al. (2008) CMTM3 can affect the transcription activity of androgen receptor and inhibit the expression level of PSA in LNCaP cells. Biochem Biophys Res Commun 371: 54–58.
36. Qamar I, Gong E-Y, Kim Y, Song C-H, Lee HJ, et al. (2010) Anti-steroidogenic factor ARR19 inhibits testicular steroidogenesis through the suppression of Nur77 transactivation. J Biol Chem 285: 22360–22369.
37. Wang Y, Li J, Cui Y, Li T, Ng KM, et al. (2009) CMTM3, located at the critical tumor suppressor locus 16q22.1, is silenced by CpG methylation in carcinomas and inhibits tumor cell growth through inducing apoptosis. Cancer Res 69: 5194–5201.
38. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, et al. (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472: 120–124.

39. Tschopp P, Fraudeau N, Béna F, Duboule D (2011) Reshuffling genomic landscapes to study the regulatory evolution of Hox gene clusters. Proc Natl Acad Sci USA 108: 10632–10637.
40. Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, et al. (2011) A regulatory archipelago controls Hox genes transcription in digits. Cell 147: 1132–1145.
41. Lin M, Pedrosa E, Shah A, Hrabovsky A, Maqbool S, et al. (2011) RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. PLoS ONE 6: e23356.
42. Parker-Katiraee L, Bousiaki E, Monk D, Moore GE, Nakabayashi K, et al. (2008) Dynamic variation in allele-specific gene expression of Paraoxonase-1 in murine and human tissues. Hum Mol Genet 17: 3263–3270.
43. Guttman M, Rinn JL (2012) Modular regulatory principles of large non-coding RNAs. Nature 482: 339–346.
44. Prendergast JGD, Semple CAM (2011) Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. Genome Res 21: 1777–1787.
45. Prendergast JGD, Campbell H, Gilbert N, Dunlop MG, Bickmore WA, et al. (2007) Chromatin structure and evolution in the human genome. BMC Evol Biol 7: 72.
46. de Wit E, van Steensel B (2009) Chromatin domains in higher eukaryotes: insights from genome-wide mapping studies. Chromosoma 118: 25–36.
47. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.
48. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, et al. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7: e1000112.
49. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.
50. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98: 5116–5121.
51. Cabrera CP, Navarro P, Huffman JE, Wright AF, Hayward C, et al. (2012) Uncovering networks from genome-wide association studies via circular genomic permutation. G3 2: 1067–75.
52. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.

References

ABECASIS, G. R., AUTON, A., BROOKS, L. D., DEPRISTO, M. A., DURBIN, R. M., HANDSAKER, R. E., KANG, H. M., MARTH, G. T. & MCVEAN, G. A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

BABRAHAM BIOINFORMATICS 2010. FastQC - Quality Control Tool for High Throughput Sequence Data.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-37.

BERNSTEIN, B. E., MIKKELSEN, T. S., XIE, X., KAMAL, M., HUEBERT, D. J., CUFF, J., FRY, B., MEISSNER, A., WERNIG, M., PLATH, K., JAENISCH, R., WAGSCHAL, A., FEIL, R., SCHREIBER, S. L. & LANDER, E. S. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell,* 125**,** 315-26.

BIAN, Q. & BELMONT, A. S. 2012. Revisiting higher-order and large-scale chromatin organization. *Curr Opin Cell Biol,* 24**,** 359-66.

BIRD, A. 2011. The dinucleotide CG as a genomic signalling module. *J Mol Biol,* 409**,** 47-53.

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., 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., HACKERMULLER, 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.

BLACK, J. C. & WHETSTINE, J. R. 2011. Chromatin landscape: methylation beyond transcription. *Epigenetics,* 6**,** 9-15.

BOYLE, S., GILCHRIST, S., BRIDGER, J., MAHY, N., ELLIS, J. & BICKMORE, W. 2001. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum Mol Genet,* 10**,** 211-9.

CAI, J., XIE, D., FAN, Z., CHIPPERFIELD, H., MARDEN, J., WONG, W. H. & ZHONG, S. 2010. Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells. *PLoS Comput Biol,* 6**,** e1000707.

CASTRESANA, J. 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Res,* 30**,** 1751-6.

CHAMBERS, E. V., BICKMORE, W. A. & SEMPLE, C. A. 2013. Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS Comput Biol,* 9**,** e1003017.

CHAMBERS, E. V. & SEMPLE, C. A. M. 2013. Chromatin Structure and Human Genome Evolution. *eLS.* John Wiley & Sons, Ltd.

CHAMBEYRON, S. & BICKMORE, W. A. 2004. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev,* 18**,** 1119-30.

CHAN, E. T., QUON, G. T., CHUA, G., BABAK, T., TROCHESSET, M., ZIRNGIBL, R. A., AUBIN, J., RATCLIFFE, M. J., WILDE, A., BRUDNO, M., MORRIS, Q. D. & HUGHES, T. R. 2009. Conservation of core gene expression in vertebrate tissues. *J Biol,* 8**,** 33.

CLOWNEY, E. J., LEGROS, M. A., MOSLEY, C. P., CLOWNEY, F. G., MARKENSKOFF-PAPADIMITRIOU, E. C., MYLLYS, M., BARNEA, G., LARABELL, C. A. & LOMVARDAS, S. 2012. Nuclear aggregation of olfactory receptor genes governs their monogenic expression. *Cell,* 151**,** 724-37.

CORDAUX, R. & BATZER, M. A. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet,* 10**,** 691-703.

CRAIG, J. M. & BICKMORE, W. A. 1994. The distribution of CpG islands in mammalian chromosomes. *Nat Genet,* 7**,** 376-82.

CROFT, J. A., BRIDGER, J. M., BOYLE, S., PERRY, P., TEAGUE, P. & BICKMORE, W. A. 1999. Differences in the localization and morphology of chromosomes in the human nucleus. *J Cell Biol,* 145**,** 1119-31.

CULLEN, K. E., KLADDE, M. P. & SEYFRED, M. A. 1993. Interaction between transcription regulatory regions of prolactin chromatin. *Science,* 261**,** 203-6.

CUMMINGS, W. J., YABUKI, M., ORDINARIO, E. C., BEDNARSKI, D. W., QUAY, S. & MAIZELS, N. 2007. Chromatin structure regulates gene conversion. *PLoS Biol,* 5**,** e246.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G. & DURBIN, R. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-8.

DAXINGER, L. & WHITELAW, E. 2012. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Rev Genet,* 13**,** 153-62.

DE CARVALHO, D. D., SHARMA, S., YOU, J. S., SU, S. F., TABERLAY, P. C., KELLY, T. K., YANG, X., LIANG, G. & JONES, P. A. 2012. DNA methylation screening identifies driver epigenetic events of cancer cell survival. *Cancer Cell,* 21**,** 655-67.

DE KONING, A. P., GU, W., CASTOE, T. A., BATZER, M. A. & POLLOCK, D. D. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet,* 7**,** e1002384.

DE, S. & MICHOR, F. 2011. DNA replication timing and long-range DNA interactions

predict mutational landscapes of cancer genomes. *Nat Biotechnol,* 29**,** 1103-8.

DE WIT, E. & VAN STEENSEL, B. 2009. Chromatin domains in higher eukaryotes: insights from genome-wide mapping studies. *Chromosoma,* 118**,** 25-36.

DENNIS, G., JR., SHERMAN, B. T., HOSACK, D. A., YANG, J., GAO, W., LANE, H. C. & LEMPICKI, R. A. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol,* 4**,** P3.

DIXON, J. R., SELVARAJ, S., YUE, F., KIM, A., LI, Y., SHEN, Y., HU, M., LIU, J. S. & REN, B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature,* 485**,** 376-80.

DUAN, Z., ANDRONESCU, M., SCHUTZ, K., MCILWAIN, S., KIM, Y. J., LEE, C., SHENDURE, J., FIELDS, S., BLAU, C. A. & NOBLE, W. S. 2010. A three-dimensional model of the yeast genome. *Nature,* 465**,** 363-7.

DUNHAM, I., KUNDAJE, A., ALDRED, S. F., COLLINS, P. J., DAVIS, C. A., DOYLE, F., EPSTEIN, C. B., FRIETZE, S., HARROW, J., KAUL, R., KHATUN, J., LAJOIE, B. R., LANDT, S. G., LEE, B. K., PAULI, F., ROSENBLOOM, K. R., SABO, P., SAFI, A., SANYAL, A., SHORESH, N., SIMON, J. M., SONG, L., TRINKLEIN, N. D., ALTSHULER, R. C., BIRNEY, E., BROWN, J. B., CHENG, C., DJEBALI, S., DONG, X., ERNST, J., FUREY, T. S., GERSTEIN, M., GIARDINE, B., GREVEN, M., HARDISON, R. C., HARRIS, R. S., HERRERO, J., HOFFMAN, M. M., IYER, S., KELLLIS, M., KHERADPOUR, P., LASSMANN, T., LI, Q., LIN, X., MARINOV, G. K., MERKEL, A., MORTAZAVI, A., PARKER, S. C., REDDY, T. E., ROZOWSKY, J., SCHLESINGER, F., THURMAN, R. E., WANG, J., WARD, L. D., WHITFIELD, T. W., WILDER, S. P., WU, W., XI, H. S., YIP, K. Y., ZHUANG, J., BERNSTEIN, B. E., GREEN, E. D., GUNTER, C., SNYDER, M., PAZIN, M. J., LOWDON, R. F., DILLON, L. A., ADAMS, L. B., KELLY, C. J., ZHANG, J., WEXLER, J. R., GOOD, P. J., FEINGOLD, E. A., CRAWFORD, G. E., DEKKER, J., ELINITSKI, L., FARNHAM, P. J., GIDDINGS, M. C., GINGERAS, T. R., GUIGO, R., HUBBARD, T. J., KELLIS, M., KENT, W. J., LIEB, J. D., MARGULIES, E. H., MYERS, R. M., STARNATOYANNOPOULOS, J. A., TENNEBAUM, S. A., WENG, Z., WHITE, K. P., WOLD, B., YU, Y., WROBEL, J., RISK, B. A., GUNAWARDENA, H. P., KUIPER, H. C., MAIER, C. W., XIE, L., CHEN, X., MIKKELSEN, T. S., et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

ERNST, J. & KELLIS, M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods,* 9**,** 215-6.

ERNST, J., KHERADPOUR, P., MIKKELSEN, T. S., SHORESH, N., WARD, L. D., EPSTEIN, C. B., ZHANG, X., WANG, L., ISSNER, R., COYNE, M., KU, M., DURHAM, T., KELLIS, M. & BERNSTEIN, B. E. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* 473**,** 43-9.

ESKELAND, R., LEEB, M., GRIMES, G. R., KRESS, C., BOYLE, S., SPROUL, D., GILBERT, N., FAN, Y., SKOULTCHI, A. I., WUTZ, A. & BICKMORE, W. A. 2010. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell,* 38**,** 452-64.

FENG, S., COKUS, S. J., ZHANG, X., CHEN, P. Y., BOSTICK, M., GOLL, M. G., HETZEL, J., JAIN, J., STRAUSS, S. H., HALPERN, M. E., UKOMADU, C., SADLER, K. C., PRADHAN, S., PELLEGRINI, M. & JACOBSEN, S. E. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A,* 107**,** 8689-94.

FILION, G. J., VAN BEMMEL, J. G., BRAUNSCHWEIG, U., TALHOUT, W., KIND, J.,

WARD, L. D., BRUGMAN, W., DE CASTRO, I. J., KERKHOVEN, R. M., BUSSEMAKER, H. J. & VAN STEENSEL, B. 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell,* 143**,** 212-24.

FINLAN, L. E., SPROUL, D., THOMSON, I., BOYLE, S., KERR, E., PERRY, P., YLSTRA, B., CHUBB, J. R. & BICKMORE, W. A. 2008. Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet,* 4**,** e1000039.

FLICEK, P., AHMED, I., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S., GIL, L., GARCIA-GIRON, C., GORDON, L., HOURLIER, T., HUNT, S., JUETTEMANN, T., KAHARI, A. K., KEENAN, S., KOMOROWSKA, M., KULESHA, E., LONGDEN, I., MAUREL, T., MCLAREN, W. M., MUFFATO, M., NAG, R., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., PRITCHARD, E., RIAT, H. S., RITCHIE, G. R., RUFFIER, M., SCHUSTER, M., SHEPPARD, D., SOBRAL, D., TAYLOR, K., THORMANN, A., TREVANION, S., WHITE, S., WILDER, S. P., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., DUNHAM, I., HARROW, J., HERRERO, J., HUBBARD, T. J., JOHNSON, N., KINSELLA, R., PARKER, A., SPUDICH, G., YATES, A., ZADISSA, A. & SEARLE, S. M. 2013. Ensembl 2013. *Nucleic Acids Res,* 41**,** D48-55.

FRASER, P. & BICKMORE, W. 2007. Nuclear organization of the genome and the potential for gene regulation. *Nature,* 447**,** 413-7.

FUSSNER, E., CHING, R. W. & BAZETT-JONES, D. P. 2011. Living without 30nm chromatin fibers. *Trends Biochem Sci,* 36**,** 1-6.

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. *Nat Genet,* 42**,** 255-9.

GAZAVE, E., GAUTIER, P., GILCHRIST, S. & BICKMORE, W. A. 2005. Does radial nuclear organisation influence DNA damage? *Chromosome Res,* 13**,** 377-88.

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A. J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J. Y. & ZHANG, J. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol,* 5**,** R80.

GIERMAN, H. J., INDEMANS, M. H., KOSTER, J., GOETZE, S., SEPPEN, J., GEERTS, D., VAN DRIEL, R. & VERSTEEG, R. 2007. Domain-wide regulation of gene expression in the human genome. *Genome Res,* 17**,** 1286-95.

GILBERT, N., BOYLE, S., FIEGLER, H., WOODFINE, K., CARTER, N. P. & BICKMORE, W. A. 2004. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell,* 118**,** 555-66.

GOECKS, J., NEKRUTENKO, A. & TAYLOR, J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol,* 11**,** R86.

GOLDBERG, A. D., ALLIS, C. D. & BERNSTEIN, E. 2007. Epigenetics: a landscape takes shape. *Cell,* 128**,** 635-8.

GUELEN, L., PAGIE, L., BRASSET, E., MEULEMAN, W., FAZA, M. B., TALHOUT, W., EUSSEN, B. H., DE KLEIN, A., WESSELS, L., DE LAAT, W. & VAN STEENSEL, B.

2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature,* 453**,** 948-51.

GUTTMAN, M., AMIT, I., GARBER, M., FRENCH, C., LIN, M. F., FELDSER, D., HUARTE, M., ZUK, O., CAREY, B. W., CASSADY, J. P., CABILI, M. N., JAENISCH, R., MIKKELSEN, T. S., JACKS, T., HACOHEN, N., BERNSTEIN, B. E., KELLIS, M., REGEV, A., RINN, J. L. & LANDER, E. S. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature,* 458**,** 223-7.

GUTTMAN, M. & RINN, J. L. 2012. Modular regulatory principles of large non-coding RNAs. *Nature,* 482**,** 339-46.

HACKETT, J. A., SENGUPTA, R., ZYLICZ, J. J., MURAKAMI, K., LEE, C., DOWN, T. A. & SURANI, M. A. 2013. Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science,* 339**,** 448-52.

HANSEN, R. S., THOMAS, S., SANDSTROM, R., CANFIELD, T. K., THURMAN, R. E., WEAVER, M., DORSCHNER, M. O., GARTLER, S. M. & STAMATOYANNOPOULOS, J. A. 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A,* 107**,** 139-44.

HIRATANI, I., RYBA, T., ITOH, M., RATHJEN, J., KULIK, M., PAPP, B., FUSSNER, E., BAZETT-JONES, D. P., PLATH, K., DALTON, S., RATHJEN, P. D. & GILBERT, D. M. 2010. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res,* 20**,** 155-69.

HIRATANI, I., RYBA, T., ITOH, M., YOKOCHI, T., SCHWAIGER, M., CHANG, C. W., LYOU, Y., TOWNES, T. M., SCHUBELER, D. & GILBERT, D. M. 2008. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol,* 6**,** e245.

HIRAYAMA, T., TARUSAWA, E., YOSHIMURA, Y., GALJART, N. & YAGI, T. 2012. CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. *Cell Rep,* 2**,** 345-57.

HOFFMAN, M. M., ERNST, J., WILDER, S. P., KUNDAJE, A., HARRIS, R. S., LIBBRECHT, M., GIARDINE, B., ELLENBOGEN, P. M., BILMES, J. A., BIRNEY, E., HARDISON, R. C., DUNHAM, I., KELLIS, M. & NOBLE, W. S. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res,* 41**,** 827-41.

HU, M., DENG, K., QIN, Z., DIXON, J., SELVARAJ, S., FANG, J., REN, B. & LIU, J. S. 2013. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol,* 9**,** e1002893.

IHAKA, R. & GENTLEMEN, R. 1996. R: A language for data analysis and graphics. *J. Comp. Graph. Stat.,* 5**,** 299-314.

JACQUES, P. E., JEYAKANI, J. & BOURQUE, G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet,* 9**,** e1003504.

JIANG, C. & PUGH, B. F. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet,* 10**,** 161-72.

JIN, C., ZANG, C., WEI, G., CUI, K., PENG, W., ZHAO, K. & FELSENFELD, G. 2009. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat Genet,* 41**,** 941-5.

KALHOR, R., TJONG, H., JAYATHILAKA, N., ALBER, F. & CHEN, L. 2012. Genome

architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol,* 30**,** 90-8.

KAPUSTA, A., KRONENBERG, Z., LYNCH, V. J., ZHUO, X., RAMSAY, L., BOURQUE, G., YANDELL, M. & FESCHOTTE, C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet,* 9**,** e1003470.

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.

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 Res,* 12**,** 996-1006.

KIM, P. M., LAM, H. Y., URBAN, A. E., KORBEL, J. O., AFFOURTIT, J., GRUBERT, F., CHEN, X., WEISSMAN, S., SNYDER, M. & GERSTEIN, M. B. 2008. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res,* 18**,** 1865-74.

KIM, Y. J., CECCHINI, K. R. & KIM, T. H. 2011. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proc Natl Acad Sci U S A,* 108**,** 7391-6.

KIND, J., PAGIE, L., ORTABOZKOYUN, H., BOYLE, S., DE VRIES, S. S., JANSSEN, H., AMENDOLA, M., NOLEN, L. D., BICKMORE, W. A. & VAN STEENSEL, B. 2013. Single-cell dynamics of genome-nuclear lamina interactions. *Cell,* 153**,** 178-92.

KOUZARIDES, T. 2007. Chromatin modifications and their function. *Cell,* 128**,** 693-705.

KVIKSTAD, E. M., TYEKUCHEVA, S., CHIAROMONTE, F. & MAKOVA, K. D. 2007. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput Biol,* 3**,** 1772-82.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

LIEBERMAN-AIDEN, E., VAN BERKUM, N. L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J., DORSCHNER, M. O., SANDSTROM, R., BERNSTEIN, B., BENDER, M. A., GROUDINE, M., GNIRKE, A., STAMATOYANNOPOULOS, J., MIRNY, L. A., LANDER, E. S. & DEKKER, J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science,* 326**,** 289-93.

LIN, Z., MA, H. & NEI, M. 2008. Ultraconserved coding regions outside the homeobox of mammalian Hox genes. *BMC Evol Biol,* 8**,** 260.

LINARDOPOULOU, E. V., WILLIAMS, E. M., FAN, Y., FRIEDMAN, C., YOUNG, J. M. & TRASK, B. J. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature,* 437**,** 94-100.

LISTER, R., PELIZZOLA, M., DOWEN, R. H., HAWKINS, R. D., HON, G., TONTI-FILIPPINI, J., NERY, J. R., LEE, L., YE, Z., NGO, Q. M., EDSALL, L., ANTOSIEWICZ-BOURGET, J., STEWART, R., RUOTTI, V., MILLAR, A. H., THOMSON, J. A., REN, B. & ECKER, J. R. 2009. Human DNA methylomes at base

resolution show widespread epigenomic differences. *Nature,* 462**,** 315-22.

LONG, H. K., SIMS, D., HEGER, A., BLACKLEDGE, N. P., KUTTER, C., WRIGHT, M. L., GRUTZNER, F., ODOM, D. T., PATIENT, R., PONTING, C. P. & KLOSE, R. J. 2013. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife,* 2**,** e00348.

MAGKLARA, A., YEN, A., COLQUITT, B. M., CLOWNEY, E. J., ALLEN, W., MARKENSCOFF-PAPADIMITRIOU, E., EVANS, Z. A., KHERADPOUR, P., MOUNTOUFARIS, G., CAREY, C., BARNEA, G., KELLIS, M. & LOMVARDAS, S. 2011. An epigenetic signature for monoallelic olfactory receptor expression. *Cell,* 145**,** 555-70.

MAHY, N. L., PERRY, P. E. & BICKMORE, W. A. 2002. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *J Cell Biol,* 159**,** 753-63.

MCCLINTOCK, T. S. 2010. Achieving singularity in mammalian odorant receptor gene choice. *Chem Senses,* 35**,** 447-57.

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.

MCDONALD, L. A., GERRELLI, D., FOK, Y., HURST, L. D. & TICKLE, C. 2010. Comparison of Iroquois gene expression in limbs/fins of vertebrate embryos. *J Anat,* 216**,** 683-91.

MEULEMAN, W., PERIC-HUPKES, D., KIND, J., BEAUDRY, J. B., PAGIE, L., KELLIS, M., REINDERS, M., WESSELS, L. & VAN STEENSEL, B. 2013. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res,* 23**,** 270-80.

MIRNY, L. A. 2011. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research,* 19**,** 37-51.

MISTELI, T. 2007. Beyond the sequence: cellular organization of genome function. *Cell,* 128**,** 787-800.

MONTAVON, T. & DUBOULE, D. 2013. Chromatin organization and global regulation of Hox gene clusters. *Philos Trans R Soc Lond B Biol Sci,* 368**,** 20120367.

MONTAVON, T., SOSHNIKOVA, N., MASCREZ, B., JOYE, E., THEVENET, L., SPLINTER, E., DE LAAT, W., SPITZ, F. & DUBOULE, D. 2011. A regulatory archipelago controls Hox genes transcription in digits. *Cell,* 147**,** 1132-45.

MOREY, C., DA SILVA, N. R., PERRY, P. & BICKMORE, W. A. 2007. Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation. *Development,* 134**,** 909-19.

NHGRI. 2010. *National Human Genome Research Institute* [Online]. Available: http://www.genome.gov/ [Accessed].

NIU, D. K. & JIANG, L. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Commun,* 430**,** 1340-3.

PARKER-KATIRAEE, L., BOUSIAKI, E., MONK, D., MOORE, G. E., NAKABAYASHI, K. & SCHERER, S. W. 2008. Dynamic variation in allele-specific gene expression of Paraoxonase-1 in murine and human tissues. *Hum Mol Genet,* 17**,** 3263-70.

PERIC-HUPKES, D., MEULEMAN, W., PAGIE, L., BRUGGEMAN, S. W., SOLOVEI, I., BRUGMAN, W., GRAF, S., FLICEK, P., KERKHOVEN, R. M., VAN LOHUIZEN, M., REINDERS, M., WESSELS, L. & VAN STEENSEL, B. 2010. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell,* 38**,** 603-13.

PICKERSGILL, H., KALVERDA, B., DE WIT, E., TALHOUT, W., FORNEROD, M. & VAN STEENSEL, B. 2006. Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat Genet,* 38**,** 1005-14.

POPE, B. D., CHANDRA, T., BUCKLEY, Q., HOARE, M., RYBA, T., WISEMAN, F. K., KUTA, A., WILSON, M. D., ODOM, D. T. & GILBERT, D. M. 2012. Replication-timing boundaries facilitate cell-type and species-specific regulation of a rearranged human chromosome in mouse. *Hum Mol Genet,* 21**,** 4162-70.

PRENDERGAST, J. G., CAMPBELL, H., GILBERT, N., DUNLOP, M. G., BICKMORE, W. A. & SEMPLE, C. A. 2007. Chromatin structure and evolution in the human genome. *BMC Evol Biol,* 7**,** 72.

PRENDERGAST, J. G. & SEMPLE, C. A. 2011. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res,* 21**,** 1777-87.

PUJADAS, E. & FEINBERG, A. P. 2012. Regulated noise in the epigenetic landscape of development and disease. *Cell,* 148**,** 1123-31.

QAMAR, I., GONG, E. Y., KIM, Y., SONG, C. H., LEE, H. J., CHUN, S. Y. & LEE, K. 2010. Anti-steroidogenic factor ARR19 inhibits testicular steroidogenesis through the suppression of Nur77 transactivation. *J Biol Chem,* 285**,** 22360-9.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

R CORE TEAM 2013. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.* Vienna, Austria.

RAM, O., GOREN, A., AMIT, I., SHORESH, N., YOSEF, N., ERNST, J., KELLIS, M., GYMREK, M., ISSNER, R., COYNE, M., DURHAM, T., ZHANG, X., DONAGHEY, J., EPSTEIN, C. B., REGEV, A. & BERNSTEIN, B. E. 2011. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell,* 147**,** 1628-39.

REDDINGTON, J. P., PERRICONE, S. M., NESTOR, C. E., REICHMANN, J., YOUNGSON, N. A., SUZUKI, M., REINHARDT, D., DUNICAN, D. S., PRENDERGAST, J. G., MJOSENG, H., RAMSAHOYE, B. H., WHITELAW, E., GREALLY, J. M., ADAMS, I. R., BICKMORE, W. A. & MEEHAN, R. R. 2013. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol,* 14**,** R25.

RIEDER, D., TRAJANOSKI, Z. & MCNALLY, J. G. 2012. Transcription factories. *Front Genet,* 3**,** 221.

ROH, T.-Y., CUDDAPAH, S. & ZHAO, K. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Development,* 19**,** 542-552

ROH, T. Y., WEI, G., FARRELL, C. M. & ZHAO, K. 2007. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res,* 17**,** 74-81.

RYBA, T., HIRATANI, I., LU, J., ITOH, M., KULIK, M., ZHANG, J., SCHULZ, T. C., ROBINS, A. J., DALTON, S. & GILBERT, D. M. 2010. Evolutionarily conserved

replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res,* 20**,** 761-70.

SARMA, K. & REINBERG, D. 2005. Histone variants meet their match. *Nat Rev Mol Cell Biol,* 6**,** 139-49.

SASAKI, S., MELLO, C. C., SHIMADA, A., NAKATANI, Y., HASHIMOTO, S., OGAWA, M., MATSUSHIMA, K., GU, S. G., KASAHARA, M., AHSAN, B., SASAKI, A., SAITO, T., SUZUKI, Y., SUGANO, S., KOHARA, Y., TAKEDA, H., FIRE, A. & MORISHITA, S. 2009. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science,* 323**,** 401-4.

SCHMIDT, D., SCHWALIE, P. C., WILSON, M. D., BALLESTER, B., GONCALVES, A., KUTTER, C., BROWN, G. D., MARSHALL, A., FLICEK, P. & ODOM, D. T. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell,* 148**,** 335-48.

SCHMIDT, D., WILSON, M. D., BALLESTER, B., SCHWALIE, P. C., BROWN, G. D., MARSHALL, A., KUTTER, C., WATT, S., MARTINEZ-JIMENEZ, C. P., MACKAY, S., TALIANIDIS, I., FLICEK, P. & ODOM, D. T. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science,* 328**,** 1036-40.

SCHMITZ, R. J. & ECKER, J. R. 2012. Epigenetic and epigenomic variation in Arabidopsis thaliana. *Trends Plant Sci,* 17**,** 149-54.

SCHMITZ, R. J., SCHULTZ, M. D., URICH, M. A., NERY, J. R., PELIZZOLA, M., LIBIGER, O., ALIX, A., MCCOSH, R. B., CHEN, H., SCHORK, N. J. & ECKER, J. R. 2013. Patterns of population epigenomic diversity. *Nature,* 495**,** 193-8.

SCHRODER, K., IRVINE, K. M., TAYLOR, M. S., BOKIL, N. J., LE CAO, K. A., MASTERMAN, K. A., LABZIN, L. I., SEMPLE, C. A., KAPETANOVIC, R., FAIRBAIRN, L., AKALIN, A., FAULKNER, G. J., BAILLIE, J. K., GONGORA, M., DAUB, C. O., KAWAJI, H., MCLACHLAN, G. J., GOLDMAN, N., GRIMMOND, S. M., CARNINCI, P., SUZUKI, H., HAYASHIZAKI, Y., LENHARD, B., HUME, D. A. & SWEET, M. J. 2012. Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proc Natl Acad Sci U S A,* 109**,** E944-53.

SEMPLE, C. A. & TAYLOR, M. S. 2009. Molecular biology. The structure of change. *Science,* 323**,** 347-8.

SEXTON, T., KURUKUTI, S., MITCHELL, J. A., UMLAUF, D., NAGANO, T. & FRASER, P. 2012. Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. *Nat Protoc,* 7**,** 1335-50.

SHARMA, S., KELLY, T. K. & JONES, P. A. 2010. Epigenetics in cancer. *Carcinogenesis,* 31**,** 27-36.

SIMONIS, M., KLOUS, P., SPLINTER, E., MOSHKIN, Y., WILLEMSEN, R., DE WIT, E., VAN STEENSEL, B. & DE LAAT, W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet,* 38**,** 1348-54.

SMIT, A., HUBLEY, R. & GREEN, P. 1996. RepeatMasker Open-3.0.

SMYTH, G. K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol,* 3**,** Article3.

SMYTH, G. K. 2005. *Limma: linear models for microarray data*, Springer.

SPROUL, D., GILBERT, N. & BICKMORE, W. A. 2005. The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet,* 6**,** 775-81.

SPROUL, D., KITCHEN, R. R., NESTOR, C. E., DIXON, J. M., SIMS, A. H., HARRISON, D. J., RAMSAHOYE, B. H. & MEEHAN, R. R. 2012. Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol,* 13**,** R84.

STAMATOYANNOPOULOS, J. A., ADZHUBEI, I., THURMAN, R. E., KRYUKOV, G. V., MIRKIN, S. M. & SUNYAEV, S. R. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet,* 41**,** 393-5.

SUZUKI, M. M. & BIRD, A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet,* 9**,** 465-76.

SUZUKI, R. & SHIMODAIRA, H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics,* 22**,** 1540-2.

TAKEBAYASHI, S., RYBA, T. & GILBERT, D. M. 2012. Developmental control of replication timing defines a new breed of chromosomal domains with a novel mechanism of chromatin unfolding. *Nucleus,* 3**,** 500-7.

TAN, M., LUO, H., LEE, S., JIN, F., YANG, J. S., MONTELLIER, E., BUCHOU, T., CHENG, Z., ROUSSEAUX, S., RAJAGOPAL, N., LU, Z., YE, Z., ZHU, Q., WYSOCKA, J., YE, Y., KHOCHBIN, S., REN, B. & ZHAO, Y. 2011. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell,* 146**,** 1016-28.

TAYLOR, M. S., KAI, C., KAWAI, J., CARNINCI, P., HAYASHIZAKI, Y. & SEMPLE, C. A. 2006. Heterotachy in mammalian promoter evolution. *PLoS Genet,* 2**,** e30.

TEYTELMAN, L., EISEN, M. B. & RINE, J. 2008. Silent but not static: accelerated base-pair substitution in silenced chromatin of budding yeasts. *PLoS Genet,* 4**,** e1000247.

THE ENCODE PROJECT CONSORTIUM 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol,* 9**,** e1001046.

TIBSHIRANI, R., CHU, G., NARASIMHAN, B. & LI, J. 2011. SAM: Significance Analysis of Microarrays. *R package version 2.0.*

TILLO, D. & HUGHES, T. R. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics,* 10**,** 442.

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. *Mol Cell,* 10**,** 1453-65.

TROPBERGER, P. & SCHNEIDER, R. 2013. Scratching the (lateral) surface of chromatin regulation by histone modifications. *Nat Struct Mol Biol,* 20**,** 657-61.

TSCHOPP, P., FRAUDEAU, N., BENA, F. & DUBOULE, D. 2011. Reshuffling genomic landscapes to study the regulatory evolution of Hox gene clusters. *Proc Natl Acad Sci U S A,* 108**,** 10632-7.

TUSHER, V. G., TIBSHIRANI, R. & CHU, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A,* 98**,** 5116-21.

VENABLES, W. N. & RIPLEY, B. D. 2002. *Modern Applied Statistics with S*, Springer.

VERMEULEN, M., MULDER, K. W., DENISSOV, S., PIJNAPPEL, W. W., VAN SCHAIK, F. M., VARIER, R. A., BALTISSEN, M. P., STUNNENBERG, H. G., MANN, M. & TIMMERS, H. T. 2007. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell,* 131**,** 58-69.

VERSTEEG, R., VAN SCHAIK, B. D., VAN BATENBURG, M. F., ROOS, M., MONAJEMI, R., CARON, H., BUSSEMAKER, H. J. & VAN KAMPEN, A. H. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res,* 13**,** 1998-2004.

VOGEL, M. J., PERIC-HUPKES, D. & VAN STEENSEL, B. 2007. Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nat Protoc,* 2**,** 1467-78.

VOIGT, P., TEE, W. W. & REINBERG, D. 2013. A double take on bivalent promoters. *Genes Dev,* 27**,** 1318-38.

WANG, K. C., YANG, Y. W., LIU, B., SANYAL, A., CORCES-ZIMMERMAN, R., CHEN, Y., LAJOIE, B. R., PROTACIO, A., FLYNN, R. A., GUPTA, R. A., WYSOCKA, J., LEI, M., DEKKER, J., HELMS, J. A. & CHANG, H. Y. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature,* 472**,** 120-4.

WANG, Y., LI, J., CUI, Y., LI, T., NG, K. M., GENG, H., LI, H., SHU, X. S., LIU, W., LUO, B., ZHANG, Q., MOK, T. S., ZHENG, W., QIU, X., SRIVASTAVA, G., YU, J., SUNG, J. J., CHAN, A. T., MA, D., TAO, Q. & HAN, W. 2009. CMTM3, located at the critical tumor suppressor locus 16q22.1, is silenced by CpG methylation in carcinomas and inhibits tumor cell growth through inducing apoptosis. *Cancer Res,* 69**,** 5194-201.

WANG, Y., LI, T., QIU, X., MO, X., ZHANG, Y., SONG, Q., MA, D. & HAN, W. 2008. CMTM3 can affect the transcription activity of androgen receptor and inhibit the expression level of PSA in LNCaP cells. *Biochem Biophys Res Commun,* 371**,** 54-8.

WARNECKE, T., BATADA, N. N. & HURST, L. D. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet,* 4**,** e1000250.

WARNES, G. R., BOLKER, B., BONEBAKKER, L., GENTLEMAN, R., HUBER, W., LIAW, A., LUMLEY, A., MAECHLER, M., MAGNUSSON, A., MOELLER, S., SCHWARTZ, M. & B, V. 2010. *Various R programming tools for plotting data. R package version 2.8.0.*

WASHIETL, S., MACHNE, R. & GOLDMAN, N. 2008. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet,* 24**,** 583-7.

WEDDINGTON, N., STUY, A., HIRATANI, I., RYBA, T., YOKOCHI, T. & GILBERT, D. M. 2008. ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics,* 9**,** 530.

WILLIAMSON, I., ESKELAND, R., LETTICE, L. A., HILL, A. E., BOYLE, S., GRIMES, G. R., HILL, R. E. & BICKMORE, W. A. 2012. Anterior-posterior differences in HoxD chromatin topology in limb development. *Development,* 139**,** 3157-67.

WOODCOCK, C. L., SKOULTCHI, A. I. & FAN, Y. 2006. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Res,* 14**,** 17-25.

WOODFINE, K., FIEGLER, H., BEARE, D. M., COLLINS, J. E., MCCANN, O. T., YOUNG, B. D., DEBERNARDI, S., MOTT, R., DUNHAM, I. & CARTER, N. P. 2004. Replication timing of the human genome. *Hum Mol Genet,* 13**,** 191-202.

WU, C., BASSETT, A. & TRAVERS, A. 2007. A variable topology for the 30-nm chromatin fibre. *EMBO Rep,* 8**,** 1129-34.

XIAO, S., XIE, D., CAO, X., YU, P., XING, X., CHEN, C. C., MUSSELMAN, M., XIE, M., WEST, F. D., LEWIN, H. A., WANG, T. & ZHONG, S. 2012. Comparative epigenomic annotation of regulatory DNA. *Cell,* 149**,** 1381-92.

YAFFE, E., FARKASH-AMAR, S., POLTEN, A., YAKHINI, Z., TANAY, A. & SIMON, I. 2010. Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet,* 6**,** e1001011.

YAFFE, E. & TANAY, A. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet,* 43**,** 1059-65.

YAMAZAKI, S., HAYANO, M. & MASAI, H. 2013. Replication timing regulation of eukaryotic replicons: Rif1 as a global regulator of replication timing. *Trends Genet,* 29**,** 449-60.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci,* 13**,** 555-6.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24**,** 1586-91.

YOKOCHI, T., PODUCH, K., RYBA, T., LU, J., HIRATANI, I., TACHIBANA, M., SHINKAI, Y. & GILBERT, D. M. 2009. G9a selectively represses a class of late-replicating genes at the nuclear periphery. *Proc Natl Acad Sci U S A,* 106**,** 19363-8.

ZEMACH, A., MCDANIEL, I. E., SILVA, P. & ZILBERMAN, D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science,* 328**,** 916-9.

ZHANG, S. S., KIM, K. H., ROSEN, A., SMYTH, J. W., SAKUMA, R., DELGADO-OLGUIN, P., DAVIS, M., CHI, N. C., PUVIINDRAN, V., GABORIT, N., SUKONNIK, T., WYLIE, J. N., BRAND-ARZAMENDI, K., FARMAN, G. P., KIM, J., ROSE, R. A., MARSDEN, P. A., ZHU, Y., ZHOU, Y. Q., MIQUEROL, L., HENKELMAN, R. M., STAINIER, D. Y., SHAW, R. M., HUI, C. C., BRUNEAU, B. G. & BACKX, P. H. 2011. Iroquois homeobox gene 3 establishes fast conduction in the cardiac His-Purkinje network. *Proc Natl Acad Sci U S A,* 108**,** 13576-81.

ZHANG, Y. & REINBERG, D. 2001. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev,* 15**,** 2343-60.

ZHANG, Z. & PUGH, B. F. 2011. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell,* 144**,** 175-86.

ZHOU, V. W., GOREN, A. & BERNSTEIN, B. E. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet,* 12**,** 7-18.