THE UNIVERSITY
of EDINBURGH

Nucleosome Positioning Dynamics in Evolution and Disease

Zhenhua Hu

Doctor of Philosophy

The University of Edinburgh

2016

Declaration of Authorship

I declare that the thesis has been composed by myself, the work presented in it is my own, and helps from other people are clearly acknowledged. I confirm that the work has not been submitted for any other degree or qualification at this University or any other institution.

Zhenhua Hu

September, 2016

Abstract

Nucleosome positioning is involved in a variety of cellular processes, and it provides a likely substrate for species evolution and may play roles in human disease. However, many fundamental aspects of nucleosome positioning remain controversial, such as the relative importance of underlying sequence features, genomic neighbourhood and trans-acting factors.

In this thesis, I have focused on analyses of the divergence and conservation of nucleosome positioning, associated substitution spectra, and the interplay between them. I have investigated the extent to which nucleosome positioning patterns change following the duplication of a DNA sequence and its insertion into a new genomic region within the same species, by assessing the relative nucleosome positioning between paralogous regions in both the human (using *in vitro* and *in vivo* datasets) and yeast (*in vivo*) genomes. I observed that the positioning of paralogous nucleosomes is generally well conserved and detected a strong rotational preference where nucleosome positioning has diverged. I have also found, in all datasets, that DNA sequence features appear to be more important than local chromosomal environments in nucleosome positioning evolution, while controlling for *trans*-acting factors that can potentially confound inter-species comparisons.

I have also examined the relationships between chromatin structure and DNA sequence variation, with a particular focus on the spectra of (germline and somatic) substitutions seen in human diseases. Both somatic and germline substitutions are found to be enriched at sequences coinciding with nucleosome cores. In addition, transitions appear to be enriched in germline relative to somatic substitutions at nucleosome core regions. This difference in transition to transversion ratio is also seen at transcription start sites (TSSs) genome wide. However, the contrasts seen between somatic and germline mutational spectra do not appear to be attributable to alterations in nucleosome positioning between cell types. Examination of multiple human nucleosome positioning datasets shows conserved positioning across TSSs and strongly conserved global phasing between 4 cancer cell lines and 7 non-cancer cell lines. This suggests that the particular mutational profiles seen for somatic and germline cells occur upon a common landscape of conserved chromatin structure.

I extended my studies of mutational spectra by analysing genome sequencing data from various tissues in a cohort of individuals to identify human somatic mutations. This allowed

an assessment of the relationship between age and mutation accumulation and a search for inherited genetic variants linked to high somatic mutation rates. A list of candidate germline variants that potentially predispose to increased somatic mutation rates was the outcome.

Together these analyses contribute to an integrated view of genome evolution, encompassing the divergence of DNA sequence and chromatin structure, and explorations of how they may interact in human disease.

Lay Summary

Nucleosome positioning defines the locations and patterns of nucleosome occupancy across the genome and is a universal feature of eukaryotic chromatin structure. The properties of this fundamental layer of chromatin organisation have important implications since nucleosomes prevent bound DNA from being accessed by a wide range of cellular machinery. Nucleosome positioning has been shown to regulate a variety of cellular processes, and contributes to species evolution and human diseases.

Due to its functional importance, nucleosome positioning itself has to be precisely regulated. Certain factors are known to regulate how and where nucleosomes are organised in the genome, including the differential affinities of particular DNA sequences and physical barriers provided by proteins. However, the relative importance of these different factors to the fine tuning of nucleosome positioning remains controversial.

In this thesis, I have studied the relationships between chromatin structure and underlying DNA sequences, by exploring the factors that affect nucleosome positioning and the impacts of nucleosome positioning on the patterns of mutations seen in DNA sequences. I have investigated the extent to which nucleosome positioning patterns change following the duplication of a DNA sequence and its insertion into a new genomic region within the same genome, a process called segmental duplication which creates two regions of high sequence identity (paralogous regions) for each duplication event. I assessed the differences in nucleosome positioning between paralogous regions in both the human (using datasets resulting from in vitro and in vivo experiments) and yeast (using in vivo data) genomes. I observed that most nucleosomes tend to occupy similar places following the duplication of paralogous regions, but detected a strong preference for positional shifts in multiples of 10 bp in the regions nucleosome positioning has diverged. The 10 bp preference is related to the increased binding affinity between nucleosome core histones and DNA sequences that show 10 bp periodicity with respect to AA/TT/AT dinucleotides. I have also found, in all datasets, that DNA sequence composition is more important in nucleosome positioning evolution than features of local chromosomal environments, such as whether the duplication event happens between different chromosomes or within the same chromosome and whether the chromatin states of duplicated regions are the same or not. Importantly, the analysis of duplicated

regions within the same cells controls for the influence of trans-acting factors (such as transcription factors) that have potentially confounded previous inter-species comparisons.

I have also examined the relationships between nucleosome positioning and the patterns of germline and somatic DNA substitutions seen in human diseases, including cancers. Both somatic and germline substitutions were found to be enriched in nucleosome occupied DNA sequences (at the nucleosome core). In addition, specific transition substitutions of A<->T or C<->G appear to be enriched in germline (inherited from parents) sequence variants relative to somatic variants (not inherited from parents but acquired in somatic tissues) at nucleosome core regions. This difference in mutational spectra is also seen at transcription start sites (TSSs) genome wide. Thus the different profiles observed for germline and somatic mutations suggest that different underlying molecular mechanisms are involved. However, the contrasts seen between somatic and germline mutational profiles do not appear to be attributable to alterations in nucleosome positioning between cell types, as indicated by the broadly conserved nucleosome positioning I have identified between 4 cancer cell lines and 7 non-cancer cell lines. This suggests that the particular mutational profiles seen for somatic and germline cells occur upon a common landscape of conserved chromatin structure.

Mutations in somatic tissues accumulate as people age, as well as in diseases such as cancers. Using large collections of genomic sequencing data I detected somatic mutations in normal somatic tissues to explore how age and germline variants affect mutation accumulation in normal somatic tissues. I have shown, for the first time, that DNA mutations have accumulated in normal human cells as a linear function of age and also identified a list of potential germline variants that appear to accelerate the accumulation of DNA mutations in somatic normal tissues.

Together these analyses contribute to an integrated view of genome evolution, encompassing the fundamental relationships between DNA sequence and chromatin structure, and explorations of how they may interact in human disease.

Acknowledgements

Firstly, I would like to acknowledge and thank my supervisors Colin Semple, James Prendergast and Martin Taylor for their guidance and support throughout the course of my PhD. I feel extremely lucky to have three supervisors and their student friendly supervision styles and considerations hugely mitigate my stress during my whole PhD. I would also like to thank the members of my thesis committee panel, Susan Farrington and Richard Meehan for their helpful suggestions, and the members of Evogen for all their help and advice. I particularly thank Alison Meynert for her help with variant calling and Bioinformatics advice.

Special sorry goes to James for my bombardment of all sorts of stupid questions (at least 10 per day) during the first two years of my PhD, ranging from the basic usage of the Linux operating system, to programming languages like PERL and R, and to many basic concepts in computational biology and bioinformatics.

Finally, I would like to thank all my family members, especially my mother and grandmother, and friends for their support.

Contents

List of Figures

List of Tables

Abbreviations

TF	Transcription factor
TSS	Transcription start site
NDR	Nucleosome depleted region
MNase	Micrococcal Nuclease
TTS	Transcription termination site
Pol II	RNA polymerase II
Indel	Insertion/deletion
DHS	DNase I hypersensitive site

Chapter 1: Introduction

1.1 Chromatin structure and the nucleosome

1.1.1 Different layers of chromatin structure in eukaryotes

The eukaryotic genome is packed into the nucleus in the form of chromatin, a DNA-protein complex (Kornberg 1977; Igo-Kemenes et al. 1982), and compaction of genomic DNA can be achieved at multiple levels, corresponding to the different layers of chromatin structure. The primary level involves DNA packaged into nucleosomes, generating a 10 nm fibre and the resulting nucleosome arrays look like "beads on a string" under an electron microscope (Thoma et al. 1979; Richmond and Davey 2003). Nucleosomes are further condensed into a 30 nm chromatin fibre through interaction with nearby nucleosomes and through linker histone H1 binding to DNA sequences that lie between successive nucleosomes and core histones (Tremethick 2007; van Steensel 2011; Fudenberg and Mirny 2012). However, there are debates on this pervasive presence of the 30 nm fibre (Tremethick 2007; Eltsov et al. 2008; Maeshima and Eltsov 2008; Maeshima et al. 2010; Fussner et al. 2011). The highest levels of chromatin structure determine the spatial organization of the genome in nucleus. For example, the human genome can be arranged into functionally distinct domains at mega-base scales, and show different degrees of accessibility to trans-acting factors. One of the indicators for genome accessibility is sensitivity to DNase I cleavage, and regions that show high sensitivity are called DNase hypersensitive sites (DHSs) (Birney et al. 2007; Bell et al. 2011; Thurman et al. 2012: 1). The genome accessibility is usually lower in heterochromatin than in euchromatin; heterochromatin is preferentially distributed at the periphery of the nucleus and interacts with nuclear lamins while euchromatin is more likely to locate in the centre, thus away from the periphery (Van Bortle and Corces 2012).

1.1.2 Nucleosome structure and its biological importance

Nucleosomes form the basic repeating unit of chromatin structure and provide the basis for all other higher orders of chromatin structure. The discovery and confirmation of the

existence of nucleosome and that nucleosome is the repeating unit of chromatin, as summarised in Kornberg (1977), came from different lines of evidences. Several studies using X-ray diffraction technique observed the multiple levels of folding of genomic DNA (Luzzati et al. 1961; Bram and Ris 1971; Pardon and Wilkins 1972), and the work from Olins and Olins (1974) showed that chromatin appeared as chains of particles under the electron microscope. Based on the finding from X-ray diffraction and biologicals results (Kornberg and Thomas 1974), Kornberg proposed that nucleosome was the repeating unit of chromatin, made of "eight histone molecules and about 200 DNA base pairs" (Kornberg 1974). Subsequently, the crystal structure of nucleosome core particle with different resolutions was provided by Klug and his colleagues (Finch et al. 1977; Richmond et al. 1984).

A nucleosome consists of 147 bp of DNA wrapped ~1.7 times around an octameric core complex of histone proteins, made up of two copies for each of H2A, H2B, H3 and H4 canonical histones (Luger et al. 1997; Luger et al. 1997; Richmond and Davey 2003; Cutter and Hayes 2015). The interaction between DNA and core histones is not static; rather nucleosomes undergo constant unwrapping and rebinding from core histones, with the time required for DNA at the edge of the histone octamer to unwrap and rebind far shorter than that observed for DNA at the centre position (Polach and Widom 1995). Linker histones bind linker DNA between adjacent nucleosomes, interact with core histones, and play important roles in further chromatin folding. Another function for the linker histone is regulation of nucleosome spacing, and the difference in the length of linker DNA seen across different species and cell types correlates with the linker histone expression levels. For example, the linker DNA is ~20 bp in yeast and ~50 bp in human (Valouev et al. 2011; Brogaard et al. 2012), and the abundance of linker histone Hho1p in yeast is about 1 molecule per 4 ~ 40 nucleosomes while the abundance of linker histone H1 in human is approximately 1 molecule per nucleosome (Bates and Thomas 1981; Freidkin and Katcoff 2001; Downs et al. 2003). Nucleosome spacing is shorter and correspondingly H1 linker histone expression is lower in human primary granulocytes than CD4+ T cells (Valouev et al. 2011).

In addition to compaction of DNA, nucleosomes decrease the accessibility of DNA to *trans*-acting proteins, including transcription factors (TFs), chromatin remodelling complexes, polymerases, and DNA mutagens and repair enzymes. The accessibility of linker DNA is much higher than that of nucleosomal DNA, and DNA located at nucleosome entry/exit points is more accessible than DNA close to nucleosome centre (dyad). Motifs that sit on major grooves that face towards the histone octamer surface are generally not available for *trans*-acting factors while those that sit on the major grooves exposed to the solvent are

more accessible. Because of the reduced accessibility of the nucleosomal DNA, the nucleosome provides an extra layer for the regulation of genome functions and nucleosome positioning has been observed to be involved in a variety of nuclear processes that use DNA as template, including gene transcription, DNA replication, and DNA damage and repair (Li et al. 2007; Clapier and Cairns 2009; Rando and Winston 2012; Hughes and Rando 2014).

Nucleosomes are generally thought to have a repressive effect on gene expression through blocking the access of *cis*-regulatory sites to TFs and RNA polymerase, and thus preventing the assembly of bulk transcription machinery at promoters which are immediately upstream of transcription start sites (TSSs) during transcription initiation; nucleosomes could also block the passage of the RNA polymerase during transcription elongation (Bondarenko et al. 2006; Jiang and Pugh 2009; Kulaeva et al. 2010; Radman-Livaja and Rando 2010; Valouev et al. 2011; Hughes and Rando 2014). Indeed, ubiquitously expressed promoters are usually depleted of nucleosomes and enriched with unstable nucleosomes which are easily digested by MNase, and display a nucleosome depleted region (NDR) at promoters, such as in growth genes (ribosomal genes) in yeast and housekeeping genes in human (Sekinger et al. 2005; Weiner et al. 2010; Vavouri and Lehner 2012). The local chromatin architecture at promoters has been demonstrated to strongly affect both the baseline gene expression level and the ability to change gene expression under environmental change (expression plasticity) in both yeast and human species, leading to differential strategies for gene expression regulation (Tirosh and Barkai 2008). In addition, Vaillant et al. (2010) have also observed a novel strategy for transcription regulation in yeast mediated by the nucleosome ordering pattern in coding regions. Interestingly, nucleosomes can also activate gene expression by different mechanisms as summarised in Hughes and Rando (2014).

1.1.3 Determinants of nucleosome positioning

Unlike sequence specific DNA binding proteins such as TFs, core histones do not have sequence-specific DNA binding domains and the wrapping of DNA around the nucleosome octamer is mediated by the direct charge-charge interactions of core histones with DNA backbone phosphates at minor grooves (Luger et al. 1997; Davey et al. 2002; Cutter and Hayes 2015). The physical properties of DNA thus have important roles in determining the overall affinity between histone octamer and nucleosomal DNA, such as the ability of DNA to bend and curve around the histone octamer surface. Indeed, even though the histone can bind to almost any DNA sequence, nucleosome formation has significant sequence preferences and the affinity between DNA sequences and the histone octamer can vary over

more than three orders of magnitude (Struhl and Segal 2013). Similarly, even though nucleosomes can occupy all possible positions along a stretch of DNA sequences, positions with the minimum free energy for DNA-histone interactions are occupied at the highest frequencies by nucleosomes (Kornberg 1981; Drew and Travers 1985; Zhurkin 1985; Kornberg and Stryer 1988; Lowary and Widom 1997; Becker 1999; Thåström et al. 1999; Widom 2001). A change in the properties of either histones or DNA, such as DNA methylation and the replacement of canonical histones with variants and histones carrying post-translational modifications can also affect intrinsic histone-DNA interaction affinity or provide docking sites for the recruitment of other *trans*-acting factors like chromatin remodelling complexes, thus leading to differential positioning of nucleosomes on a given genome. Transcription factors have long been observed to be able to compete with the nucleosomes for DNA with variable abilities, and notably a subset of special TFs called "pioneer factors" can invade the nucleosomal DNA (Tims et al. 2011; Zaret and Carroll 2011; Struhl and Segal 2013).

The pattern of the genome-wide nucleosome organization on a given genome can be summarised and described by nucleosome occupancy, nucleosome positioning (translational and rotational), and nucleosome phasing (Albert et al. 2007; Mavrich et al. 2008; Hughes et al. 2012; Hughes and Rando 2014). There is a long standing dispute about whether local (*cis*-acting) DNA sequences or (*trans*-acting) proteins play the dominant role in nucleosome positioning (Zhang et al. 2009; Tirosh et al. 2010; Zhang et al. 2010; Hughes et al. 2012). Several studies have attempted to dissect the effects of DNA sequences from that of *trans*-acting proteins with one approach being the comparison of nucleosome positioning *in vivo* and *in vitro* using experimental reconstructions of nucleosome formation by mixing histone proteins and genomic DNA (Kaplan et al. 2009; Zhang et al. 2009; Valouev et al. 2011). Tirosh et al. (2010) compared inter-species difference in nucleosome positioning between *S. cerevisiae and S. paradoxus* with that of their hybrid. Differences maintained in the hybrid being inferred to be largely the result of *cis*-acting DNA sequences, and not *trans*-acting proteins that are expected to be distinct in different species. Other efforts have combined genetic and evolutionary approaches to study the relative contributions of *cis* and *trans* determinants. For example, Hughes et al. (2012) introduced yeast artificial chromosomes (YAC) containing genomic DNA from other yeast species into *S. cerevisiae*, and compared the nucleosome positioning between these yeast artificial chromosomes and the same regions in their native, donor species. Schones et al. (2008) have suggested that nucleosome positioning at non-regulatory regions is largely determined by DNA sequence preferences but that at regulatory regions it is mainly governed and regulated by *trans*-acting factors.

It is now accepted that nucleosome organization is not random and is determined by the interplay of sequence preferences, *trans* factors and statistical positioning principles (Kornberg and Stryer 1988; Segal et al. 2006; Mavrich et al. 2008; Struhl and Segal 2013; Hughes and Rando 2014).

1.1.3.1 Sequence preference in nucleosome occupancy

Various *in vivo* genome-wide nucleosome positioning maps of differential resolutions in yeast have observed a decrease in nucleosome occupancy at promoters and at transcription termination sites (TTSs), with much stronger depletion in promoter regions, compared to that over coding regions (Yuan et al. 2005; Mavrich et al. 2008; Shivaswamy et al. 2008; Kaplan et al. 2009; Fan et al. 2010; Brogaard et al. 2012). For example, different groups have studied the average nucleosome positioning pattern around genes and observed a classical nucleosome depletion region (NDR) of ~150 bp just upstream of transcription start sites, immediately flanked by two well positioned nucleosomes with the downstream nucleosome called +1 nucleosome and the upstream nucleosome called -2 nucleosome; while in a subset of genes, including some stress genes in yeast and tissue specific genes in human, a nucleosome called -1 nucleosome was observed to occupy the region between -2 and +1 nucleosomes that coincides with NDR (Jansen and Verstrepen 2011; Hughes and Rando 2014). In addition, nucleosome depletion has been observed at many of the transcription factor binding sites (Segal et al. 2006; Kaplan et al. 2009).

The observed phenomena mentioned above caused efforts to dissect the relative importance of DNA sequence and other factors in genome wide nucleosome organization patterns. Since it is hard if not impossible to disentangle the effects of *trans* mechanisms from that of *cis*-acting DNA sequence preference, several labs have contrasted *in vivo* genome-wide nucleosome positioning maps to *in vitro* maps. These maps are generated by assembling recombinant histone octamers of canonical histones purified from other species on genomic DNA, which only reflect the role of DNA sequence preference in nucleosome occupancy (Segal et al. 2006; Kaplan et al. 2009; Zhang et al. 2009; Valouev et al. 2011; Struhl and Segal 2013). For example, similar global nucleosome occupancy patterns observed both *in vivo* and *in vitro*, with correlations as high as 0.74 in genome-wide per base pair nucleosome occupancy and almost a perfect correlation (0.98) in 5mer nucleosome occupancy, and the accuracy of separation of nucleosome enriched regions from nucleosome depleted regions *in vivo* based on the *in vitro* data is high (Segal et al. 2006; Kaplan et al. 2009).

One of the mechanisms for low nucleosome occupancy are intrinsic histone-DNA interaction preferences, as observed in HIS3-PET56 and DED1 promoter regions in yeast (Sekinger et al. 2005). Sequence analysis found that the low preference for nucleosome occupancy in yeast promoters is mainly shaped by enrichment of the antinucleosomal AT rich DNA sequence such as poly(dA:dT) homo-polymeric stretches, which are found in the promoters of about 95% of yeast genes (Mavrich et al. 2008). Most of the genes with poly(dA:dT) rich promoters are growth genes, which are usually deprived of TATA boxes and show high baseline expression level and low transcriptional plasticity (Tirosh and Barkai 2008). Another class of genes are stress genes which feature low basal transcription levels and high transcription plasticity; promoters of genes in this class commonly possess TATA boxes and show deficiencies of poly(dA:dT) stretches, concurrent with high nucleosome occupancy and more fuzzily localized and evenly distributed nucleosomes (Mavrich et al. 2008; Tirosh and Barkai 2008; Rando and Winston 2012).

Homo-polymeric poly(dA:dT) stretches are intrinsically stiff and have low binding affinity with histone octamers (McCall et al. 1985; Nelson et al. 1987; Suter et al. 2000; Segal and Widom 2009; Struhl and Segal 2013). The affinity of poly(dA:dT) to histone octamers has been thought to be length dependent, and ploy(dA:dT) stretches with length of 4bp or greater have low affinity with histone octamers (Anderson and Widom 2001; Scipioni et al. 2004; Thåström et al. 2004; Jansen and Verstrepen 2011). The depletion of nucleosomes due to poly(dA:dT) is independent of transcription activity and is observed *in vivo* under different growth conditions, including YPD, ethanol and galactose (Sekinger et al. 2005; Kaplan et al. 2009). Raveh-Sadka et al. (2012) managed to alter nucleosome organization in promoters and their transcription levels through manipulating poly(dA:dT) tracts. Hughes et al. (2012) found that the NDR has been maintained in gene promoters of foreign species after being introduced into the budding yeast and maintenance is through poly(dA:dT) tracts.

In contrast, GC rich sequences are intrinsically nucleosome favouring. Human promoters are generally GC rich and are occupied by nucleosomes (Hughes and Rando 2009; Tillo et al. 2010; Valouev et al. 2011). High nucleosome occupancy in human promoters has been observed *in vitro* and in a subset of unexpressed genes *in vivo* which feature the absence of CpG island in their promoters (Valouev et al. 2011; Vavouri and Lehner 2012). More than 50% of human genes contain a CpG island in their promoter and most of the genes in this class are constitutively expressed genes such as housekeeping genes (Vavouri and Lehner 2012). CpG promoters are less likely to contain a TATA box while human genes that lack CpG island promoters are more likely to contain a TATA box and associate with tissue-

specific genes (Davuluri et al. 2001; Carninci et al. 2006; Saxonov et al. 2006; Vavouri and Lehner 2012). Interestingly, CpG promoters have a constitutive NDR which seems to be independent of transcription activity, and thus can still be observed in unexpressed genes (Vavouri and Lehner 2012). The low nucleosome occupancy in CpG promoters has been shown to be linked to the assembly of CpG islands into unstable nucleosomes, preoccupation and pausing of RNA polymerases, and high levels of CTCF binding (Tirosh and Barkai 2008; Ramirez-Carrozzi et al. 2009; Vavouri and Lehner 2012). Nonetheless, sequence preference is at least partially responsible for the assembly of CpG island into unstable nucleosomes: tetramers of high GC content associate with high nucleosome core coverage, such that the tetramers of 100% GC content are associated with the highest nucleosome core occupancy. The CpG dinucleotide tetramer (CGCG) shows a 30% reduction in nucleosome occupancy *in vitro* (Valouev et al. 2011); in addition, the non-CpG island sequences outperform the CpG island sequences in associating with purified recombinant histone octamers from *Xenopus laevis*. Interestingly the famous Widom601 CpG island sequence that has periodic A/T dinucleotides also shows greater affinity with histone octamers than CpG island sequences that lack properly phased A/T dinucleotides, suggesting that these histone-DNA interaction preferences are also affected by the rotational positioning of the A/T dinucleotides at the histone octamer surface (Lowary and Widom 1997; Lowary and Widom 1998; Thåström et al. 1999; Ramirez-Carrozzi et al. 2009). A/T dinucleotides are bendable and their periodic arrangement along the nucleosome DNA, featuring repeating occurrence every ~10.2 bp, has been shown to be favourable for nucleosome positioning (Struhl and Segal 2013). However, it has also been shown that the sequence-directed nucleosome occupancy signals are mainly realized through that of poly (dA:dT) and GC content, rather than that of rotational preference of dinucleotides (Tillo and Hughes 2009).

1.1.3.2 Sequence bias in nucleosome positioning

Beside the involvement in nucleosome occupancy, the DNA sequence preference has also been shown to play important roles in nucleosome positioning which is similar to but distinct from nucleosome occupancy (Pugh 2010; Struhl and Segal 2013; Figure **1.1**). Nucleosome occupancy (or density) describes the fraction of cells in a population in which a given genomic region is occupied by a nucleosome. Nucleosome positioning with respect to a given genomic region can be represented by the consensus primary position plus the average deviation from the primary position such as standard deviation (Pugh 2010). A nucleosome occupying the same genomic region with small standard deviation is said to be well

positioned. In extreme cases, a nucleosome is said to be perfectly positioned if the nucleosome binds to the same genomic region in every cell in the population, and a nucleosome taking up variable positions with no preference in the cell population is defined as poorly positioned or fuzzily positioned (Mavrich et al. 2008; Struhl and Segal 2013). Regions with similar occupancy level might be covered by well or poorly positioned nucleosomes and the indiscriminate usage of nucleosome occupancy and nucleosome positioning underlies some disputes on whether the DNA sequence preference is the primary determinant of genomic nucleosome organization or not (Segal et al. 2006; Kaplan et al. 2009; Zhang et al. 2009; Kaplan et al. 2010; Pugh 2010; Zhang et al. 2010; Struhl and Segal 2013).

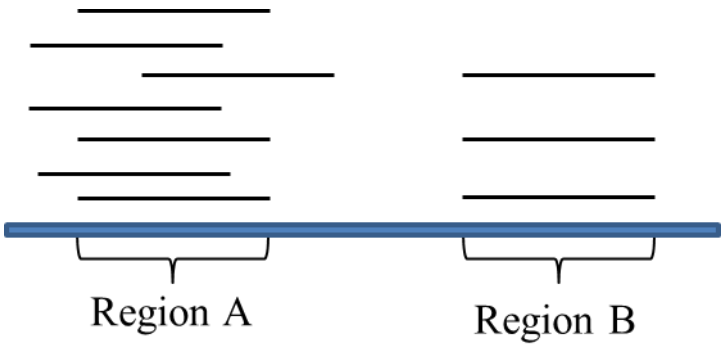


Figure 1.1. Illustration of the concept of nucleosome positioning and nucleosome occupancy. Region A and B are of 147 bp on a genome. Region A: partial positioning but higher occupancy; Region B: perfect positioning but lower occupancy. Each black horizontal line represents a nucleosome core which is 147 bp.

Two related concepts with respect to nucleosome positioning are translational positioning and rotational positioning (Figure 1.2), and they both describe the preference of certain positions over others. While translational positioning defines the specific genome region of 147 bp that the histone octamer binds, rotational positioning describes a set of translational positions separated by distances of multiple complete helical turns that are more preferable than other positions (Thåström et al. 1999; Albert et al. 2007; Brogaard et al. 2012; Gaffney et al. 2012).

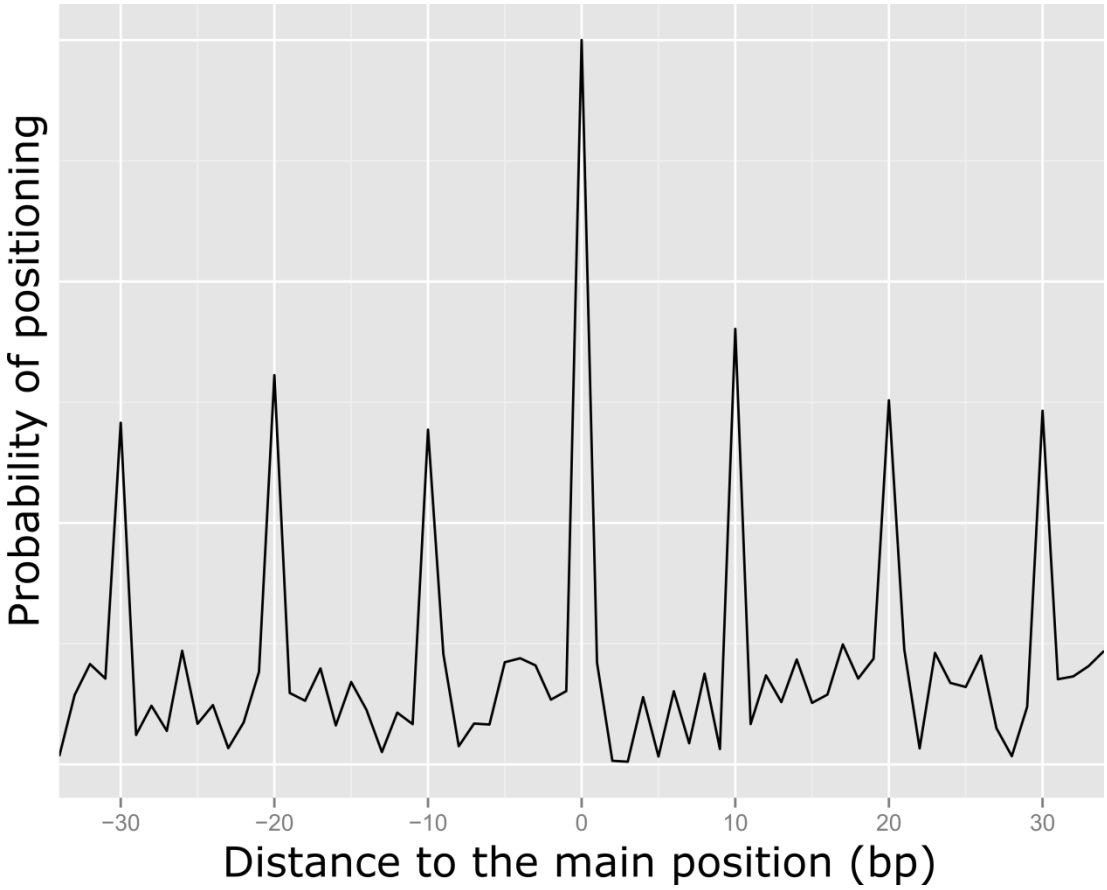


Figure 1.2. Illustration of the concept of rotational positioning and translational positioning. The dyad of a nucleosome in a given genomic region can occupy every position allowed, but with different probabilities, and this phenomenon is called translational positioning. The position of highest probability to be occupied is called main position and positions around it are called alternative positions. Rotational positioning means that not all alternative positions are of equal probabilities to be occupied by a dyad, and those with distances of multiple of ~10.5 bp to the main position have higher probability to be occupied than other alternative positions. The structural basis for rotational positioning is the periodic organization of dinucleotides on nucleosome core.

Tirosh et al. (2010) estimated that about 70% of inter-species difference in nucleosome positioning is due to DNA sequences. They suggested that the divergence of inter-species nucleosome positioning could be explained by the divergence in AT-rich nucleosome disfavouring but not the nucleosome favouring sequences. In addition, these authors found that nucleosome positioning divergence was able to propagate to multiple adjacent nucleosomes either side of the NDR but could not cross the NDR and concluded that the NDR, rather than the +1 nucleosome serves as the barrier, in direct contradiction with what was found by Zhang et al. (2009) that the strongly positioning of the +1 nucleosome is due to

the binding of the transcription factors and the assembly of the transcriptional machinery in promoters. Valouev et al. (2011) observed that sequence driven nucleosome positioning signals are involved in the translational positioning of nucleosomes at many genomic sites. From highly positioned nucleosomes, these authors also extracted another type of sequence-directed nucleosome positioning signal that is different from the 10 bp dinucleotide frequency for rotational positioning: "container sites". These sites are characterised by the central GC-rich nucleosome favouring sequence and boundaries set out by the AT-rich nucleosome repelling sequences. Using more accurate nucleosome positioning data obtained from lymphoblastoid cell lines from 7 human individuals, Gaffney et al. (2012) observed non-random translational positioning for about 84% of nucleosomes, with stronger translational positioning than expected by chance and 8.7% of total nucleosomes detected seen to be strongly positioned. One example, a 76 kb region near the centromere on chromosome 12 has been observed to contain phased arrays of highly positioned nucleosomes, featuring high GC content in core and high AT content in linker (Gaffney et al. 2012).

DNA methylation involves the addition of a methyl group to the cytosine (C) nucleotide, and this feature of the epi-genome has also been shown *in vitro* to increase the stability of nucleosomes (Collings et al. 2013).

The sequence basis for rotational positioning has been linked to the rotational preference of nucleosomal DNA, which is characterised by a distinct 10 bp dinucleotide frequency: AA/TT/AT/TA dinucleotides are favoured at minor grooves that are ~10 bp apart from each other, and face inwards towards and are in direct contact with the histone cores; set off by 5 bp, GG/CC/GC/CG dinucleotides are favoured at major grooves that face towards the histone cores; and A/T dinucleotide has high bendability and the nucleosome formation is more stable when it is in direct contact with histone (Segal et al. 2006; Albert et al. 2007; Brogaard et al. 2012; Gaffney et al. 2012; Cutter and Hayes 2015). This 10 bp periodicity has been observed in the dinucleotide frequency along the nucleosomal DNA by different groups. Kaplan et al. (2009) observed that nucleosomal DNA has the characteristic 10 bp periodicity in yeast nucleosomes both *in vivo* and *in vitro*, with the signal much stronger *in vitro*; a similar difference between *in vivo* and *in vitro* in terms of rotational positioning and the rotational preference of nucleosomal DNA has also been observed by Zhang et al. (2009). In addition, based on a nucleosome positioning map with base pair accuracy in yeast, Brogaard et al. (2012) observed that more stable nucleosomes in yeast are associated with stronger signal for periodic dinucleotide frequency, and also the corresponding rotational positioning that features the 10 bp periodicity in the distance between nucleosome centres in

the redundant map. Valouev et al. (2011) observed a much stronger signal of 10 bp periodicity in nucleosomal DNA *in vitro* than in human granulocytes. Gaffney et al. (2012) observed the rotational preference of nucleosomal DNA and the rotational positioning, indicated by a clear 10 bp periodicity in both the dinucleotide frequency, the DNase I cleavage frequency and the MNase midpoint distribution. Taken together, these studies suggest that rotational positioning of nucleosomes is primarily determined by the rotational preference of the DNA helix on the histone octamer surface in both yeast and humans.

1.1.3.3 *Trans*-determinants of nucleosome organization

Evidence mentioned above clearly shows the importance of DNA sequence preferences in nucleosome organization. However, it is also clear that nucleosome organization *in vivo* cannot be perfectly recapitulated by that *in vitro*, strongly arguing for the role of *trans*-acting factors in determining *in vivo* genome-wide nucleosome positioning (Hughes and Rando 2014). For example, reduced nucleosome occupancy was observed in yeast promoters *in vitro*, but the depletion is much more profound *in vivo* (Hughes and Rando 2009; Kaplan et al. 2009; Zhang et al. 2009). The opposite nucleosome occupancy pattern exists in human promoters between *in vitro* and *in vivo*, especially for CpG promoters where nucleosome occupancy *in vitro* is 5 times as high as *in vivo* (Valouev et al. 2011; Vavouri and Lehner 2012). Nucleosome depletion at yeast terminators, with a well-positioned nucleosome just upstream of transcription termination sites, has proven not to be intrinsic but instead related to the transcriptional mechanism (Mavrich et al. 2008; Shivaswamy et al. 2008; Fan et al. 2010). In addition, the most important determinant of cell identity is the differential expression profiles of all genes in a given genome since the sequence of the genome is the same for all cells with different identities, a particular cell identity must be encoded by *trans*-acting factors that fine-tune gene expression by regulating nucleosome positioning (Thurman et al. 2012; Hughes and Rando 2014).

Choi and Kim (2009) observed stably positioned nucleosomes at sites that show high propensities for DNA binding and periodic occurrence of dinucleotides. These sites also coincide with TATA boxes and transcription terminator sites in the variably expressed genes, suggesting that altered transcription activity is realized by evicting or repositioning nucleosomes from their original sites through *trans* mechanisms. Unlike in human CpG promoters, nucleosome occupancy in non-CpG promoters is dependent on the transcription rate. Unexpressed non-CpG promoters display high nucleosome occupancy which is consistent with high nucleosome affinity of GC rich sequence, whereas promoters become

depleted of nucleosomes in actively transcribed genes, showing the classical NDR, and during transcription activation, nucleosomes have to be cleared from promoters to allow the assembly of the transcription machinery either through nucleosome eviction or translocation which inevitably involves trans-acting factors (Clapier and Cairns 2009; Ramirez-Carrozzi et al. 2009; Vavouri and Lehner 2012; Hughes and Rando 2014).

1.1.3.4 Importance of transcription factors

Transcription factors (TFs) are proteins that read signals for transcription initiation encoded in *cis*-regulatory sequences located in core promoters immediately upstream of TSSs and position transcription machinery over TSSs by associating with components of transcription preinitiation complexes (PIC) that control gene expression (Zaret and Carroll 2011). Though the accessibility of nucleosomal DNA is significantly reduced for most TFs, a special class of TFs have strong DNA-binding activity and can invade nucleosome embedded DNA motifs, open up local chromatin either by the direct competition against nucleosomes or by recruiting chromatin remodelling complexes, and enhance the binding of other TFs and the assembly of bulk transcription machinery. Examples include the pioneer factors such as general TFs Abf1 and Reb1 in yeast and FoxA (the FoxA family includes FoxA1, FoxA2, and FoxA3) and GATA TFs in human (Kaplan et al. 2008; Zhang et al. 2009; Zaret and Carroll 2011; Rando and Winston 2012). Abf1 and Reb1 binding sites are intrinsically nucleosome favouring; however, Abf1 and Reb1 expel nucleosome formations in the vicinity of their binding sites and their loss leads to increased nucleosome occupancy *in vivo* (Kaplan et al. 2009). The power of pioneer factors to open closed chromatin was supported by the crystal structure of the FoxA1 pioneer factor, which revealed a DNA binding domain (DBD) showing high similarity with that of the linker histone that could also displace linker histone, and a C-terminal domain which binds to the core histones; FoxA1 can open up local chromatin by disrupting the interaction between nearby nucleosomes (Clark et al. 1993; Ramakrishnan et al. 1993; Cirillo et al. 2002; Zaret and Carroll 2011).

It has long been verified that TFs compete against nucleosomes for binding to DNA sequence, and the direct competition mechanism is consistent with the observation in yeast that the transcription factor binding sites are preferentially located at the entry/exit sites than dyad positions of nucleosomes and the entry/exist sites have far shorter exposure time compared to sites close to dyads (Polach and Widom 1995; Albert et al. 2007; Tims et al. 2011). The NDR in CLN2 promoter has eight conserved transcription factor binding sites and lacks the homo-polymeric poly(dA:dT) stretch; the complete NDR is maintained when

promoter sequences containing all of the eight conserved factor binding sites is inserted into new nucleosome favouring sites, but lost after some of the conserved transcription factor binding sites were mutated, clearly supporting the role of TFs in nucleosome positioning (Bai et al. 2011). The authors also noticed that TFs seem to work in synergy with poly(dA:dT) in the establishment of the NDR in yeast promoters, indicated by the observation that the NDRs are most enriched with nucleosome-depleting factors binding sites and poly(dA:dT). This is consistent with findings that nucleosome depletion at many of the transcription factor binding sites is encoded intrinsically into the yeast genome, as revealed by *in vitro* genome-wide nucleosome reconstruction (Kaplan et al. 2009; Bai et al. 2011). The dominance of *trans*-acting proteins over sequence preferences was also seen in the comparison of nucleosome positioning patterns around NRSF and CTCF protein binding sites between *in vitro* and *in vivo* data (Valouev et al. 2011). The NRSF binding site is within nucleosome favouring sequence and was occupied *in vitro* by nucleosomes through the "container site" mechanism mentioned above. In contrast, NRSF bound to DNA *in vivo* and occluded nucleosomes from their original sites, leading to that nucleosomes were well positioned both upstream and downstream of the bound NRSF protein (Valouev et al. 2011). Gaffney et al. (2012) have also confirmed the nucleosome excluding role of CTCF and also notified a similar effect for other TFs like GABP and C-fos.

1.1.3.5 RNA polymerase and the transcription machinery in nucleosome organization

To transcribe a gene, the transcription initiation complex (PIC) has to assemble at the core promoter in the vicinity of TSSs. During transcription elongation, RNA polymerase has to move along the DNA which is embedded in nucleosomes. Nucleosomes act as a barrier in both transcription initiation and elongation, and thus histone-DNA interaction has to be disrupted, either through nucleosome eviction or sliding, to allow for the assembly of PIC at promoters during transcription activation and the passage of RNA polymerase during transcription elongation (Weiner et al. 2010; Hughes and Rando 2014). RNA polymerase has been tested experimentally *in vitro* showing that it can invade DNA located at the edges of the histone octamer, and loosen the binding of DNA to the histone octamer, which is followed by subsequent rebinding of the histone octamer to DNA upstream (Studitsky et al. 1994; Studitsky et al. 1997; Kulaeva et al. 2010). Nucleosome organization in both yeast and human species correlates with transcriptional activity: genes that are highly active display greater nucleosome depletion at promoters, terminator sites, and possess densely packaged nucleosomes at coding regions showing shorter inter-nucleosomal spacing (Albert et al. 2007;

Mavrich et al. 2008; Shivaswamy et al. 2008; Valouev et al. 2011; Vavouri and Lehner 2012; Hughes and Rando 2014).

A study on the transcription coupled regulation of nucleosome positioning compared the positions of individual nucleosomes at promoters before and after heat shock (Shivaswamy et al. 2008) and found that genes usually react to transcriptional perturbation by changing nucleosome occupancy and positioning, and that binding sites for TFs that mediate transcriptional activation become more accessible. RNA polymerase II has been observed to evict nucleosomes from promoters and shift nucleosomes upstream during transcriptional elongation, and the loss of function for RNA polymerase resulted in a decreased NDR width and downstream nucleosome shifting, a nucleosome positioning pattern which is more similar to that observed *in vitro* in yeast (Weiner et al. 2010). Schones et al. (2008) found that the extent to which nucleosomes are depleted at core promoter regions in CD4+ T cells is directly dependent on RNA polymerase II (Pol II) binding level. Stalled promoters which feature relatively high levels of Pol II binding but very low gene expression levels display significantly reduced nucleosome occupancy which is similar to that of active genes. In addition, a subset of repressed genes showed a significant decrease in Pol II binding, accompanied by significantly increased nucleosome occupancy at promoters, suggesting that changes in transcription levels might be achieved by fine-tuning nucleosome organization at promoters and variable levels of Pol II binding (Schones et al. 2008).

Interestingly, the positioning of the +1 nucleosome of stalled promoters is different from that in elongation promoters, and the difference seems not to be dependent on the amount of Pol II binding but the function of the bound Pol II. Promoters activated by TCR signalling show the same positioning of the +1 nucleosome as in expressed genes, concurrent with the conversion of hypo-phosphorylated Pol II to ser5-phosphorylated Pol II (Schones et al. 2008). This relationship between the +1 nucleosome position and TSS activation is observed in human and many other species, suggesting that the transcriptional machinery itself is a general mechanism for transcription coupled fine-tuning of nucleosome positioning (Jiang and Pugh 2009; Zhang et al. 2009; Valouev et al. 2011; Hughes et al. 2012).

1.1.3.6 Chromatin remodelling complexes

Chromatin remodelling complexes consist of multiple protein subunits and can be divided into to four different families: SWI/SNF, ISWI, CHD and INO80 (Clapier and Cairns 2009). All four families of remodelling complexes contain a catalytic ATPase subunit of

SWI2/SNF2 subfamily and non-catalytic attendant subunits. Each ATPase subunit contains an ATPase domain that is split into DExx and HELICc, and unique flanking domains. ATPase domain is conserved in all four families and across eukaryotes, and each family is distinguished from others by unique domains within and flanking the ATPase domain (Clapier and Cairns 2009). The ATPase domain uses the energy generated from dialysing ATP to disrupt the histone-DNA interaction and translocate DNA, and distinct combinations of non-catalytic subunits, as well as flanking domains in the ATPase subunits, detect unique targeting signals and regulate the remodelling activities of individual complexes (Clapier and Cairns 2009; Längst and Manelyte 2015). The targeting signals include DNA sequence/structure, RNA molecules, histone variant and post-translational modification, and TFs (Längst and Manelyte 2015). Eventual outcomes of remodelling activity can be nucleosome sliding, partial or complete nucleosome eviction, and nucleosome restricting such as the replacement of canonical H2A/H2B by H2A.Z/H2B (Hughes and Rando 2014).

SWI/SNF family remodelling complexes preferentially target acetylated lysine residues on the histone tails through bromo-domains, and can slide or evict nucleosomes but have no role in chromatin assembly (Hassan et al. 2002; Kassabov et al. 2003). The SWI/SNF remodelling complex has been shown to regulate nucleosome positioning at promoters of yeast ribosomal genes and non-CpG promoters in humans (Shivaswamy et al. 2008; Ramirez-Carrozzi et al. 2009). The SANT domain together with SLIDE domain in the ATPase subunit of ISWI family remodellers can bind to linker DNA and unmodified H4 tails, and thus many complexes from this family have a role in nucleosome spacing and chromatin assembly (Strohner et al. 2004; Clapier and Cairns 2009; Längst and Manelyte 2015). Zhang et al. (2009) used the chromatin assembly factor ACF to assemble nucleosome arrays on the yeast genome and found that both the translational and rotational positioning were diminished compared to the *in vitro* reconstruction by salt dialysis, suggesting that chromatin remodelling complexes can overcome the sequence preference and translocate nucleosomes to sequences that are less nucleosome favourable. The chromo-domain defines the CHD family complexes, and the Chd1 complex specifically interacts with methylated histone tails through the chromo-domain to regulate gene expression in yeast (Pray-Grant et al. 2005; Marfella and Imbalzano 2007). The INO80 complex has been found to be important in the dynamics of the H2A.Z histone variant at yeast promoters and the deletion of the INO80 complex disrupts the link between H2A.Z level and transcriptional activity (Papamichos-Chronakis et al. 2011; Längst and Manelyte 2015).

1.1.3.7 Histone modifications and histone variants

The four core histones consist of an N-terminal tail domain and a C-terminal histone fold domain (Cutter and Hayes 2015). The post-translational modifications affect either domain, and can either directly affect the interaction affinity with DNA and/or function indirectly by providing docking signals for the recruitment of other effect proteins, including TFs and chromatin remodelling complexes (Taverna et al. 2007; Musselman et al. 2012; Zentner and Henikoff 2013; Tessarz and Kouzarides 2014). Acetylation neutralizes the positive charge of lysine residues, and thus interferes with the electrostatic interaction between positively charged lysine residue and the negatively charged phosphates on the DNA backbone, leading to reduced stability of nucleosome formation (Hong et al. 1993). Histone acetylation is involved in a variety of cellular processes (Xu et al. 2005; Hu et al. 2011; Vavouri and Lehner 2012; Zentner and Henikoff 2013; Tessarz and Kouzarides 2014). Methylation does not affect the charge of the lysine residue and up to three methyl groups can be added to the lysine residue. In addition to lysine residues, histones can also be mono- or di-methylated on arginine residues (Zentner and Henikoff 2013; Tessarz and Kouzarides 2014). While histone acetylation is usually associated with transcriptional activation (Vavouri and Lehner 2012), histone lysine methylation can be associated with both transcription activation and repression (Barski et al. 2007; Valouev et al. 2011; Thurman et al. 2012; Vavouri and Lehner 2012).

Histone variant H2A.Z containing nucleosomes have been shown to be preferentially located at promoters in both yeast and human species, and the H2A.Z level is associated with transcription rate, suggesting the H2A.Z containing nucleosomes are less stable (Albert et al. 2007; Barski et al. 2007; Ramirez-Carrozzi et al. 2009; Vavouri and Lehner 2012). H3.3/H2A.Z containing nucleosomes are observed to be preferentially located at active promoters and other cis-regulatory sites in humans (Jin et al. 2009a).

1.1.3.8 Nucleosome phasing and "statistical positioning"

In contrast to nucleosome occupancy and positioning, the mechanisms by which eukaryotic genomes are compacted into regularly ordered arrays of nucleosomes (nucleosome phasing) are exclusively regulated by *trans*-acting factors, including linker histones and chromatin remodelling complexes, and positioning constraints such as the "statistical principle" (Kornberg 1981; Kornberg and Stryer 1988; Mavrich et al. 2008; Zhang et al. 2009; Valouev et al. 2011; Zhang et al. 2011). Kornberg and his colleague (Kornberg 1981; Kornberg and Stryer 1988) suggested that nucleosome formation on DNA

was decided by the difference in the sequence specificity of DNA for nucleosome histones and other non-histone regulatory proteins and proposed the statistical model for nucleosome position and distribution (Figure 1.3). In this model, nucleosomes cannot form on DNA associated with other proteins that set the boundary or barrier for the nucleosomes, whose positions are then restricted to the regions of DNA between the boundaries. Nucleosomes close to the boundary are well positioned and show clear periodicity in flanking regions up to 1 kb (equivalent to about 5 nucleosomes), but this decreases and eventually disappears as the distance increases. A study from Fu et al. (2008) has revealed that the insulator CTCF can actually affect nucleosome positioning in this way. Nucleosomes cannot overlap with each other, and if the distance between the boundaries is less than 166 bp nucleosomes cannot form inside this region; thus the periodic pattern of nucleosome positioning can be affected by the concentration of core histones and the length of available DNA sequences (Kornberg 1981; Kornberg and Stryer 1988; Mavrich et al. 2008; Vaillant et al. 2010). The barriers producing a boundary effect can be sequence-specific binding to DNA by *cis*-acting proteins including TFs and CTCF (Fu et al. 2008; Cuddapah et al. 2009; Valouev et al. 2011), the transcription coupled nucleosome depleted regions located at both 5' and 3' termini of genes (Kornberg and Stryer 1988; Mavrich et al. 2008; Vaillant et al. 2010), and the strongly positioned nucleosomes observed at "container sites" (Valouev et al. 2011; Gaffney et al. 2012). The nucleosome phasing around "container sites" was only observed *in vivo* but not *in vitro*, which confirms the role of *trans*-acting factors and the "statistical principle" in establishing the regular spacing between nucleosomes (Valouev et al. 2011).

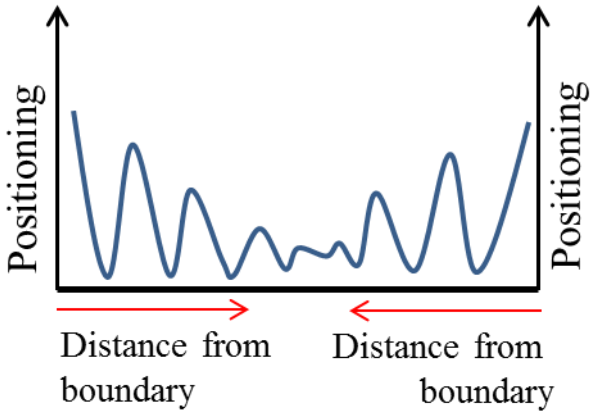


Figure 1.3. Illustration of the concept of "statistical positioning". The probability of a nucleosome positioning at a certain place was shown as vertical arrows and the linear distance from the boundaries at two ends of a region was shown as horizontal arrows.

1.1.4 Personal summary

Though some labs used different approaches and came to different conclusions in terms of whether *cis* or *trans* factors are the major determinants of nucleosome positioning (Tirosh et al. 2010; Hughes et al. 2012), my opinion on this is that both the DNA sequences and *trans* factors such as TFs are essential. While DNA sequence determines the baseline nucleosome organization in a given genome other factors including TFs, regulate the genome functions, such as the cell identity, through affecting nucleosome positioning.

1.2 Mutation spectra and chromatin structure

1.2.1 Central importance of mutations in the evolutionary process and human diseases

Mutation provides raw materials required for evolution to occur both at organism level and cell level. While mutation contributes to biological diversity on earth, it also contributes to human heritable diseases, since mutations that occur in germ cells are passed on from generation to generation (germline mutations). In non-heritable disease, mutations can accumulate in the genomes of dividing cells in somatic tissues as people age (somatic mutations). Mutations may have functional consequences by affecting gene expression patterns (mutation in ci-regulatory sites), and/or protein structural changes (such as loss of function and gain-of-function mutations in coding regions), which underlie human diseases including cancer. Cancer can be regarded as a disease of the genome and is the result of an

evolutionary process within populations of cells (Crespi and Summers 2005; Jones et al. 2008; Ye et al. 2009; Heng et al. 2010; Heng et al. 2011). Driver mutations in a certain subset of important genes like tumour suppressor genes (loss-of-function mutations) and/or oncogenes (gain-of-function mutations) confer proliferation and survival advantages to carrier cells over their neighbours. As a result of this evolutionary process, positively selected cells with increased fitness are clonally expanded and acquire increased ability to survive and invade proximal and distal organs through blood and lymphatic vessels, leading to cancer development and progression (Ye et al. 2009; Heng et al. 2010; Talavera et al. 2010; Chaffer and Weinberg 2011; Lawrence A. Loeb 2011; Valastyan and Weinberg 2011). Passenger mutations happen by chance with no effect on cell fitness and only expand in cell populations because of driver mutations (Pon and Marra 2015). Talavera et al. (2010) showed that mutations are not evenly distributed across coding sequences, with driver mutations concentrated in conserved functional domains while the passenger mutations are more evenly dispersed. Indeed, the main goal of The Cancer Genome Atlas (TCGA) consortium was to catalogue the driver mutations, identifying targets for the delivery of better personalized medical service.

1.2.2 Variations in mutation spectra

Due to the central importance of mutations in both evolutionary biology and the aetiology of human diseases, great efforts have been focused on understanding the mechanisms underlying variation in the mutation rate. The distribution of mutations of different types is not random and variation in mutation rates is a general feature across the genome (Baer et al. 2007). The rate of mutations, both germline and somatic, vary from chromosome to chromosome and from region to region within a given chromosome measured at different scales, ranging from single nucleotide to mega-base scales (Gaffney and Keightley 2005; Hellmann et al. 2005; Spencer et al. 2006; Hodgkinson and Eyre-Walker 2011; Hodgkinson et al. 2012; Schuster-Böckler and Lehner 2012; Makova and Hardison 2015; Polak et al. 2015). It has been observed that germline mutation rate is higher in sex chromosome than autosomal chromosome and higher in male than female, due to the phenomenon of male mutation bias where, compared to the eggs in female, sperm cells in male have increased numbers of cell divisions (Crow 2000; Conrad et al. 2011; Wilson Sayres and Makova 2011). Since DNA replication is not a perfect process the increased cell divisions lead to a higher accumulation of mutations in sperm. Kong et al. (2012) found that the rate of *de novo* germline mutations could be explained by the father's age when children were conceived, showing a linear relationship with an increase of two mutations per year. Genomic context also affects the mutation rate. At the lowest single base pair scale, the mutation rate is affected by the neighbouring nucleotide composition, exemplified by the increased rate of C->T base changes (both germline and somatic) at CpG sites, due to the methylation of cytosine (C) in the context of CpG dinucleotide and conversion of methylated cytosine to thymine (T) through deamination (Hwang and Green 2004; Lee et al. 2010; Pleasance et al. 2010; Chapman et al. 2011; Hodgkinson and Eyre-Walker 2011). In contrast, hydroxymethylated cytosines were observed to preferentially mutate to guanine (G) (Supek et al. 2014), although interestingly mutation rate is significantly decreased at promoter CpG islands (Polak and Arndt 2008; Cohen et al. 2011; Hodgkinson and Eyre-Walker 2011). Sequence comparison between human and chimpanzee has shown that the SNPs in human are excessively enriched at the sites where there are also chimpanzee SNPs and the base substitution rate increases around insertions and deletions (Indels) (Tian et al. 2008; Hodgkinson and Eyre-Walker 2011; Johnson and Hellmann 2011).

In addition to the variation in the mutation rate across the genome, heterogeneity has been also observed between cancer cells and normal somatic cells, and also between different cancer types (Greenman et al. 2007; Salk et al. 2010; Lawrence et al. 2013). Various studies have clearly shown that mutation spectra are tumour specific and show differential degrees

of heterogeneity across different cancer types, different tumours of a given cancer type, and different cell populations within a single tumour (Salk et al. 2010; Lawrence A. Loeb 2011; Gerlinger et al. 2012; Helleday et al. 2014). In the past few years, increasing efforts have been made to explain these varying mutation spectra (Nik-Zainal et al. 2012; Ludmil B. Alexandrov et al. 2013; Helleday et al. 2014). For example, based on a non-negative matrix factorization algorithm (NMF), the landscapes of somatic mutations seen in different tumours can be considered to be the total sums of products of the exposure rates and signatures of individual mutational processes, which in turn are determined by distinct mutagenesis pathways and deficits in DNA repair systems (Ludmil B. Alexandrov et al. 2013; Helleday et al. 2014).

1.2.3 Chromatin structure and mutation rates

Every nuclear process involving genomic DNA happens in the context of chromatin, and both genome landscape features and chromatin structure have been observed to be closely associated with regional variation in mutation rates (Makova and Hardison 2015). For example, Hellmann et al. (2005) have found that genomic features, including GC and CpG content, recombination rates, and the distance to the centromere and telomere, are associated with both sequence diversity within human species and the sequence divergence between human and chimpanzee. Hodgkinson et al. (2012) have observed significant variations in somatic mutation density at mega-base scales, which are correlated with both genomic and epi-genomic features, including GC content, nucleosome occupancy, and other factors. DNA replication timing has also been found to be correlated with the rate of single nucleotide substitutions in cancer genomes (Liu et al. 2013).

Chromatin structure, from three-dimensional nuclear organisation, to nucleosome positioning and occupancy at primary level, have proved to be important determinants of the variation in mutation spectra (Makova and Hardison 2015). Human-chimpanzee sequence divergence has been found to be relatively higher in regions of closed chromatin than that of open chromatin (Prendergast et al. 2007). Analyses to assess the impact of higher order chromatin structure on somatic mutational spectra at 1 Mb scale in human cancers have shown that the somatic mutation rate is mainly determined and shaped by chromatin organization (Schuster-Böckler and Lehner 2012; Polak et al. 2015). Schuster-Böckler and Lehner (2012) found that even though other features, including GC content, are associated with mutational rate at 1 Mb scale, the single feature H3K9me3 that is associated with heterochromatin can account for more than 40% of the variation seen in mutation rate and

combined features related to chromatin organization explain more than 55% of the variation in somatic mutation rates. In addition, when regressing the somatic mutation rate at 1 Mb scale against the genomic and epi-genomic features from the same cell type, Polak et al. (2015) found that chromatin accessibility and modification, together with replication timing, explain up to 86% of the variation in mutation rates across 173 genomes from 8 cancer types.

The relationship between the openness of chromatin (accessibility) and base substitution rate often seems to be contradictory between studies: some have found that the base mutation rate is elevated in closed heterochromatin while others found that the base mutation rate is actually elevated in the open chromatin (Birney et al. 2007; Haygood et al. 2007; Prendergast et al. 2007; Taylor et al. 2008: 11). For example, while Prendergast et al. (2007) observed higher human-chimpanzee sequence divergence in closed chromatin and Thurman et al. (2012) found the mutation rate at open chromatin (DHS) was lower than 4-fold degenerate sites, previous work has suggested that primate promoter regions are subject to higher mutation rates than other regions of the genome (Taylor et al. 2006; Taylor et al. 2008). The observation of higher mutation rates in both closed and open chromatin was confirmed in two multivariate analyses: canonical correlation analysis (CCA) and hidden Markov model genome segmentation analysis (Ananda et al. 2011; Don et al. 2013). On one hand, the base substitution rate was observed to be elevated at regions harbouring nuclear lamina binding sites in closed chromatin; on the other hand, along with insertions and deletions, base substitutions were seen to be highly elevated at open chromatin (Ananda et al. 2011; Don et al. 2013). The base substitution rate has also been shown to be elevated at sites surrounding insertions and deletions (Tian et al. 2008) and the intervals with elevated insertion, deletion and substitution rates were located in open chromatin with reduced lamina interactions and enriched with DHS and H3K4me1 marks, making up about 8% of the genome. In contrast, intervals characterised by mildly elevated deletion and substitution rates were associated with closed chromatin featuring high number of nuclear lamina binding sites and low DHS and H3K4me1 levels, making up 18% of the genome.

At the primary level of chromatin structure, the linkage between nucleosome positioning and sequence divergence is well supported (Gilbert et al. 2004; Gazave et al. 2007; Prendergast et al. 2007; Washietl et al. 2008; Semple and Taylor 2009; Prendergast and Semple 2011). Higasa and Hayashi (2006) found that the distribution of human SNPs shows a 146bp periodicity around CpG but not nonCpG TSSs. Also, Sasaki et al. (2009) observed a 200bp periodicity of mutation rates around TSSs in Medaka fish and found that this periodicity is associated with nucleosome positioning. However, Tolstorukov et al. (2011) have not found any significant correlation between nucleosome occupancy level and SNP

frequency around TSSs in humans. Several groups have noted a higher rate of sequence variation at nucleosome cores relative to linker DNA, which could be explained by differential mutation rates, or by alterations in the modes and strength of selection between core and linker regions. Prendergast and Semple (2011) have shown that both mutation rates and patterns of selection observed in the human lineage are correlated with nucleosome positioning. Furthermore the direction and strength of selection observed was predicted to maintain the optimal variation in local GC content for nucleosome positioning. In yeast, the variation in GC content between nucleosome core and linker regions has been linked to mutational bias, and the presence of nucleosomes has been shown to preferentially suppress certain types of spontaneous single base mutations (Chen et al. 2012; Xing and He 2015).

1.2.4 Personal summary

Mutations is of central importance of mutations in the evolutionary process and human diseases. In addition, mutational spectra is linked with different levels of chromatin structure, both at mega-base and nucleosome level. Thus a potential confounding factor to affect the mutational spectra across different cells is whether global nucleosome positioning is massively altered across cells.

1.3 Goals of investigation

In this thesis, I am interested in nucleosome positioning dynamics in evolution and disease, and the interplay between mutational spectra and nucleosome structure.

I firstly focused on nucleosome positioning evolution in the human genome. To better understand the role of DNA sequence in nucleosome positioning dynamics *in vivo*, Chapter 2 is focused on the effect of genetic changes on nucleosome locations between paralogous regions, where a genomic segment is copied and inserted into a new genomic region within the same sample (using *in vitro* and *in vivo* datasets). The levels of *trans*-acting factors are inherently controlled for by comparing paralogous regions *in vivo*, and all broad features of the cellular environment (the expression of chromatin binding proteins etc.) should be consistent between regions. However duplicated region pairs will share different levels of sequence similarity, allowing us to assess the role of sequence changes in altering nucleosome structure. Also comparisons of duplicons between *in vivo* and *in vitro* samples, where *trans* factors are completely absent, allows one to study the roles of trans factors in

positioning. In Chapter 3, I directly compared nucleosome positioning evolution between paralogous regions in human and yeast genomes side by side to investigate whether the sequence related mechanisms in nucleosome positioning evolution were conserved between these two eukaryotes which are separated by over 1 billion years.

In Chapter 4, I investigated nucleosome positioning in human diseases and the interplay between mutational spectra and nucleosome structure. Previous evidence suggests that the positioning of nucleosomes and the patterns of histone modifications or variants they carry may differ between cancer and normal cell lines (Fraga et al. 2005; Barski et al. 2007; Esteller 2007; Lin et al. 2007; Zhang et al. 2008; Jin et al. 2009a; Chodavarapu et al. 2010; Portela and Esteller 2010; Brait and Sidransky 2011; Wilson and Roberts 2011; Collings et al. 2013). Thus I directly compared nucleosome positioning across 4 cancer and 7 non-cancerous cell lines, aiming to investigate whether the nucleosome organization seen in cancer cell lines is altered compared to that in non-cancerous cell lines. I then analysed the interplay of mutational spectra (germline and somatic) with nucleosome structure and concluded that the profiles seen for somatic and germline cells depend upon a common landscape of conserved chromatin structure, since nucleosome positioning was conserved across all 11 cell lines examined.

Finally in Chapter 5, I extended our studies of mutational spectra by comparing genome sequencing data from various tissues in a cohort of individuals to identify human somatic mutations. This allowed me to explore the relationships between age and mutation number and to seek inherited genetic variants linked to high somatic mutation rates. I identified a list of candidate germline variants that potentially predispose to increased somatic mutation rates. Together these analyses contribute to an integrated view of genome evolution, encompassing the divergence of DNA sequence and chromatin structure, and how they may interact in human disease.

Chapter 2: Nucleosome Positioning Dynamics in Evolution

2.1 Introduction

Nucleosome positioning is involved in a variety of cellular processes and provides a likely important substrate for species evolution; the positions of nucleosomes have been shown to be relatively well conserved between species, in particular around promoters (Jiang and Pugh 2009; Hughes and Rando 2014). Nucleosome positioning is generally considered to be determined by the combined effects of DNA sequence preference, steric constraints from neighbouring nucleosomes which can be explained by statistical positioning, and trans-acting factors (Kornberg and Stryer 1988; Segal et al. 2006; Mavrich et al. 2008; Struhl and Segal 2013). However, many fundamental aspects of nucleosome positioning remain controversial, such as the relative importance of underlying sequence features, genomic neighbourhood and *trans*-acting factors (Kaplan et al. 2008; Zhang et al. 2009; Stein et al. 2010; Tirosh et al. 2010; Hughes et al. 2012). For example, homo-polymeric poly(dA:dT) sequences that are intrinsically nucleosome repelling are enriched in yeast promoters, and associated with the classical nucleosome depleted region (NDR) upstream of transcription start sites (TSSs); in contrast, the NDR observed in human promoters is generally GC rich in sequence and GC rich sequences have been shown to be intrinsically nucleosome favouring, directly arguing for the formation of NDR in human promoters by *trans*-acting factors (Tillo et al. 2010; Valouev et al. 2011; Struhl and Segal 2013).

Investigations into nucleosome positioning can be categorised into single-locus scale, large-scale, and genome-wide scale (Jansen and Verstrepen 2011). Large-scale studies of nucleosome positions in both yeast and human genomes started with the chromatin immunoprecipitation followed by DNA microarray (ChIP-chip) technology (Bernstein et al. 2004; Lee et al. 2004; Yuan et al. 2005; Heintzman et al. 2007; Lee et al. 2007; Ozsolak et al. 2007). However, only thanks to the advance in next generation sequencing (NGS) technologies in the past decade, which all feature low per-base cost and massively parallel sequencing ability (Mardis 2008; Liu et al. 2012; Koboldt et al. 2013), the genome-wide nucleosome positions maps have been generated routinely to study the patterns and

determinants of positioning of bulk nucleosomes and/or subsets of functionally distinct nucleosomes, including H2A.Z and H3.3/H2A.Z containing nucleosomes which are enriched at promoters, and H3K4me3 containing nucleosomes around TSSs (Albert et al. 2007; Barski et al. 2007; Jin et al. 2009b; Thurman et al. 2012). Bulk chromatin *in vivo* or *in vitro* (resulted from the assembly of purified histone octamers on the genomic DNA are fragmented by either physical shearing, like sonication, or enzymatic digestion, such as micrococcal nuclease (MNase). Nucleosomal DNA is protected from digestion while linker DNA is preferentially digested by MNase (Chung et al. 2010; Jansen and Verstrepen 2011; Allan et al. 2012). Mono-nucleosome sized DNA fragments are extracted and purified from PCR gel. In chromatin immunoprecipitation followed by next generation sequencing (ChIP-Seq), the genomic DNA is firstly covalently cross-linked to core histones and fragmented by physical or nuclease methods as mentioned above, then nucleosome bound DNA fragments are then pulled down and enriched by chromatin immunoprecipitation (ChIP) with antibodies against the specific residues in the core histones which can be either residues of the canonical histone or variant, or post-translationally modified residues (Barski et al. 2007; Anon 2011; Tollefsbol 2011). Enriched and purified nucleosome-protected DNA fragments are finally sequenced by different NGS sequencing platforms. The ChIP-Seq protocol was widely applied in Encode project to identify and annotate the functional elements including chromatin states (Consortium 2004).

The genome-wide pattern of nucleosome positions in a given genome can be described by related but distinct entities, including nucleosome occupancy, nucleosome transitional and rotational positioning, and nucleosome spacing. Empowered by the high throughput of the NGS technologies and due to its biological importance, great interest has been focused on linking the specific aspects of nucleosome positioning with individual determinants, aiming to better understand the mechanisms governing nucleosome positioning in detail. For example, building on the knowledge from biochemical studies that poly(dA:dT) has low affinity with core histones, (Raveh-Sadka et al. 2012) have demonstrated that manipulating a poly(dA:dT) tract can cause change in gene expression in a predictable way, corroborating former claims from the same group on the role of DNA sequences in nucleosome positioning (Anderson and Widom 2001; Segal et al. 2006; Kaplan et al. 2008; Segal and Widom 2009). The causal effect of RNA polymerase on regulating nucleosome occupancy and transitional positioning has been demonstrated in a study from Weiner et al. (2010) which compared both nucleosome occupancy and transitional positioning around promoter and TSS regions before and after the induction of the loss of function of RNA polymerase. The central importance of chromatin remodelling complexes in nucleosome positioning has been shown

by Zhang et al. (2011). In the *in vitro* nucleosome reconstruction by mixing genomic DNA and purified histones plus the whole cell extract (WCE), the Adenosine triphosphate (ATP), but no other nucleoside triphosphates, is required to well match the *in vivo* nucleosome organization pattern around 5' end of genes (Zhang et al. 2011), confirming the ability of chromatin remodelling complex, using energy produced by the hydrolysis of ATP to evict or slide nucleosomes from their original positions, in organizing nucleosomes at 5' of genes alone (Clapier and Cairns 2009). In addition, the sufficiency of chromatin remodelling complexes in uniformly placing nucleosomes into an array has been shown to be independent of statistical positioning, indicated by the observation of the regular spacing between nucleosomes with low histone concentration while the high nucleosome concentration was required by statistic principle (Zhang et al. 2011).

However, disentangling effects of individual factors on nucleosome positioning is far from simple and straightforward when it comes to differentiating whether the nucleosome positioning is primarily determined by *cis*-acting DNA sequence preferences, as shown by the contradicting conclusions reached by different studies after comparing genome-wide nucleosome positioning maps in yeast generated by different approaches (Kaplan et al. 2008; Zhang et al. 2009; Tirosh et al. 2010; Hughes et al. 2012). Both Kaplan et al. (2008) and Zhang et al. (2009) have contrasted *in vitro* and *in vivo* genome-wide nucleosome maps but came to opposite conclusions. A joint review by these studies' authors, published in 2013 (Struhl and Segal 2013), attributed the discrepancy to the fact that, nucleosome occupancy was chosen to represent the genome-wide nucleosome positioning in Kaplan et al. (2008); while in Zhang et al. (2009), nucleosome positioning was strictly interpreted to be the exact nucleosome translational positioning (Kaplan et al. 2010; Pugh 2010; Zhang et al. 2010). The comparison of *in vitro* and *in vivo* nucleosome maps has been used in other studies to deduce the relative contribution of DNA sequence to nucleosome positioning. The comparison of *in vitro* and *in vivo* human nucleosome positioning maps from three primary blood cell lines enabled Valouev et al. (2011) to deduce a sequence preference driven nucleosome positioning signal called the "container site", which is distinct from the rotational preference and was formed by the synergic effects of GC rich sequence in core and AT rich sequence in linker. The "container site" was later observed in a region of 76 kb around the centromere on chromosome 12 in Gaffney et al. (2012). Other efforts have combined genetic and evolutionary approaches to study the relative contributions of *cis* and *trans* determinants. For example, Tirosh et al. (2010) compared the inter-species difference in nucleosome positioning between *S. cerevisiae and S. paradoxus* with that of their hybrid. Differences maintained in the hybrid were inferred to be largely the result of *cis*-acting DNA

sequences, rather than the presence of *trans*-acting chromatin binding proteins that are expected to be distinct in different species. These authors estimated that about 70% of inter-species divergence in nucleosome positioning is due to DNA sequences, which is mainly realised by the anti-nucleosomal 5-mers of AT nucleotides rather than the GC rich nucleosome favouring sequences. The features chosen to represent nucleosome positioning include both nucleosome occupancy and nucleosome translational positioning, thus arguing for the role of DNA sequence in nucleosome translational positioning in contradiction to Zhang et al. (2009). A different experimental approach (Hughes et al. 2012) introduced yeast artificial chromosomes (YACs) containing genomic DNA from other yeast species into *S. cerevisiae* cells, and compared the nucleosome positioning between these yeast artificial chromosomes and the same regions in their native, donor species. In contrast to the findings of Tirosh et al. (2010), they observed that the only feature that is invariant between native and donor species is the nucleosome depletion due to the poly(dA:dT), but other features, including nucleosome spacing and the translational positioning of the +1 nucleosome, differed notably between the same DNA sequences in YACs and their native species, suggesting trans- acting factors are in fact the primary determinants of nucleosome translational positioning.

One limitation of these previous studies is that the levels of *trans*-acting chromatin binding factors are expected to vary widely between cell types, replicates and organisms. To better understand the role of DNA sequence in *in vivo* nucleosome positioning, ideally one would investigate the effect of sequence changes on nucleosome locations within the same organism and within the same sample of cells. To achieve this, we have investigated the extent to which nucleosome positioning patterns change following the duplication of a DNA sequence and its insertion into a new genomic region within the same species, by assessing the relative nucleosome positioning between paralogous regions in the human genome (using *in vitro* and *in vivo* datasets). Within *in vivo* samples, broad features of the global cellular environment (the expression of chromatin binding proteins etc.) should be consistent and controlled for, but duplicated regions will share different levels of sequence similarity. The effects of changes in the DNA sequence between duplicons, including the birth or death of binding sites (motifs), can therefore be studied while controlling for global levels of the relevant *trans* factors. Comparisons to duplicons within *in vitro* samples, where *trans* factors are completely absent, allows one to distinguish divergence in nucleosome positioning where trans factors play no role at all.

The studies of duplicated regions have already provided numerous insights into genome evolution (Zheng 2008; Lorente-Galdos et al. 2013; Prendergast et al. 2014). Duplication

events, transferring copies of one genomic region to a new location, being considered a key driving force for the creation of new genes in eukaryotic genomes (Wolfe and Shields 1997; Koszul et al. 2004; Dehal and Boore 2005; Aarts et al. 2014). It has been estimated that approximately 5% of the human genome has arisen from a recent duplication event, corresponding to around 150 Mb of sequence (Marques-Bonet et al. 2009). Although the evolving DNA sequences of sister duplicons have been explored in detail since the publication of the yeast genome (LangkjAEr et al. 2003; Papp et al. 2003; Kellis et al. 2004), their epi-genomic divergence is under studied. Recent work has shown that human DNA methylation patterns appear to be well conserved following a duplication event (Prendergast et al. 2014), but Zheng (2008) have also observed that histone modifications show asymmetric patterns between duplicons. However, the evolution of nucleosome positioning, though fundamental to chromatin structure, has not been studied between duplicated sister regions across the human genome.

In this chapter, we have focused on the analyses of the divergence and conservation of nucleosome positioning, and investigated the potential roles of both DNA sequence features and local chromosomal environments in nucleosome positioning evolution, by investigating the correspondence between the positions of *in vivo* and *in vitro* nucleosomes, and also by comparing genome-wide nucleosomes between human paralogous regions *in vitro* and *in vivo*. We observed significant correspondences between the positions of *in vivo* and *in vitro* nucleosomes and detected the DNA composition and local chromatin state related nucleosome positioning signals. In addition, the positioning of paralogous nucleosomes is generally well conserved and we detected a strong rotational preference where nucleosome positioning has diverged. Though both DNA sequence features and local chromosomal environments are significantly associated with nucleosome positioning evolution between paralogous regions, DNA sequence features appear to be more important than local chromosomal environments both *in vivo* and *in vivo*.

2.2 Materials and Methods:

2.2.1 Bioinformatics tools used

2.2.1.1 Bowtie1 and Bowtie2

Though there are several short read aligners available, we chose Bowtie1 and Bowtie2 in this study (Langmead et al. 2009; Langmead and Salzberg 2012). Bowtie1 and Bowtie2 are both ultrafast and memory efficient in mapping reads to long sequences such as the human genome. The pre-built human genomes (hg18) were downloaded from the Bowtie1 website (http://bowtie-bio.sourceforge.net/manual.shtml#paired-end-alignment). Bowtie1 was used to map single-end reads (33 bp) in colour space generated for the *in vitro* human dataset and Bowtie2 was used to map paired-end reads (25 bp each) for the *in vivo* human sample respectively. We used Bowtie1 for mapping *in vitro* reads because only Bowtie1 supports alignment of reads in colour space.

2.2.1.2 iNPS

Improved nucleosome-positioning algorithm (iNPS) was used to called nucleosomes *in vitro* (Chen et al. 2014). To infer nucleosome positions, iNPS extends each MNase-seq tag from the 5' end by 150 bp towards the 3' position, the middle 75 bp is then selected to infer the nucleosomes position, iNPS then detects the inflection start and end points, which represent the borders of nucleosomes, from the smoothed nucleosome occupancy profile, and defines dyads as the midpoints of the paired inflection points. Nucleosomes called by iNPS have a 10 bp resolution, an uncertainty of 10 bp in nucleosome's real positions.

2.2.1.3 PERL and R

PERL (version 5.14.1, https://www.perl.org/) is a high-level and general purpose programming language and has been used for the daily data handling and manipulation. Data manipulation for statistical analysis and all the statistical analyses were carried out in R (version 3.1.0, https://www.rproject.org/) using its standard functions and extension packages including Reshape and Plyr for data manipulation and ggplot2 for data visualization (Wickham 2014; Wickham 2015; Wickham and Chang 2015: 2).

2.2.2 Summary of datasets used

2.2.2.1 The paralogous regions

The list of paralogous segments (sequence identity ≥ 90%; length ≥ 1 kb), that are enriched in telomere and centromere related regions and comprise 159 Mb of the human genome (based on hg18), was obtained from http://humanparalogy.gs.washington.edu/ (She et al. 2004). We identified a total of 53,626 pairs of paralogous regions in the human genome, with a median average size of 3 kb. The size distribution is plotted in Figure 2.1.

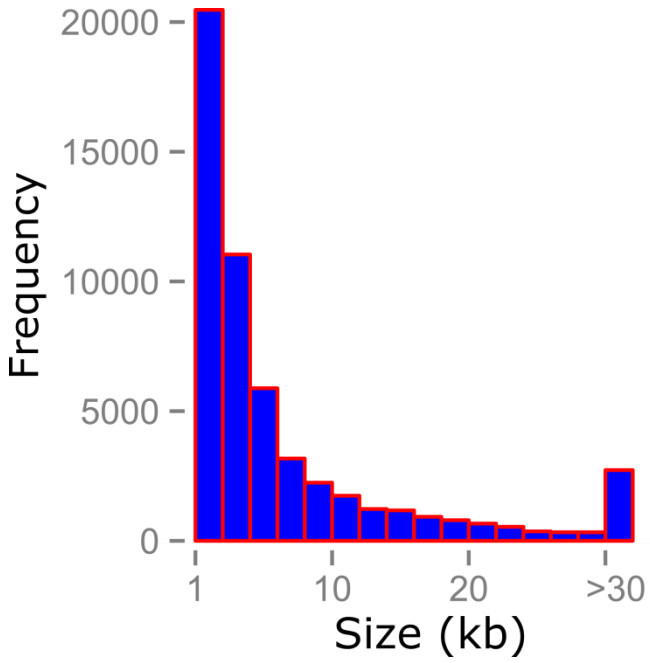


Figure 2.1. Size distributions of average paralogous duplicons in the human genome.

2.2.2.2 Sequencing reads for human in vivo sample

The paired-end sequencing reads to infer human nucleosome positions *in vivo* were obtained from (Gaffney et al. 2012). The authors studied genome-wide nucleosome organization in lymphoblastoid cell lines belonging to seven Yoruba individuals from the HapMap project, generated by MNase digestion of chromatin, size selection of mono-nucleosomal DNA fragments, and both single-end and paired-end sequencing. The data included in this analysis were based on the paired-end reads

2.2.2.3 Sequencing reads for human in vitro sample

The *in vitro* nucleosomes were generated through the reconstruction of recombinant *Xenopus* histones onto human genomic DNA derived from neutrophil granulocytes (Valouev et al. 2011). Briefly, purified naked genomic DNA was sheared by sonication and fragments of length ranging from 850 bp to 2000 bp were selected to be mixed with approximately equal molar of histone octamers to ensure that on average one nucleosome was found every 850 bp. The resulting nucleosomes in six replicate experiments were then digested with MNase, and mono-nucleosomal DNA fragments were retrieved by size selection and sequenced from one-end (single-end sequencing) using the SOLiD technology to generate reads of 35 bp in colour space.

2.2.2.4 HOMER genome-wide motifs

The list of genome-wide motifs predicted by the HOMER program was downloaded from the HOMER website (http://homer.salk.edu/homer/data/motifs/homer.KnownMotifs.hg19.bed.gz, (Heinz et al. 2010). The genome-wide sequence scanning and motif prediction was based on the known motifs for different factors in the HOMER motif database which is collated from different sources. Motifs that are deposited in the HOMER database include transcription factor motifs from ChIP-Seq data, promoter related motifs which are preferentially enriched in promoters as compared against random genomic regions, and general factor motifs which are observed to bind to factors in different studies but lacking information on the direct association between DNA binding domains and motifs. Generally, screening for binding motif for a particular factor consists of two main steps: generating a position-specific weight matrix (PSWM) where each row (corresponding to one position) represents the observed frequencies of four nucleotides A, T, C, and G based on DNA sequences bound by a given factor. Query sequences are compared against these PSWM to assess how likely the query sequence matches the PSWM for a particular factor (Stormo 2000; Heinz et al. 2010).

2.2.2.5 Chromatin state data

Chromatin state data for the lymphoblastoid cell line GM12878 was downloaded from the UCSC genome browser (Kent et al. 2002; Ernst et al. 2011). The study generated density maps for nine histone marks by ChIP-Seq and an input control from whole cell extract (WCE) of 9 cell lines including GM12878, resulting in 90 maps in total. Multivariate Hidden

Markov Model (HMM) was applied to learn the chromatin state for each 200 bp interval across the human genome in 9 cell lines based on the combinations of the signal density from 9 histone marks and input control. Consecutive intervals of 200bp with the same state were concatenated to generate final chromatin state maps for each cell line. A total of 15 chromatin states were inferred (Table 2.1), including states corresponding to promoters, enhancers, insulators, actively transcribed regions, and repressed and inactive domains (Ernst and Kellis 2010; Ernst et al. 2011).

Table 2.1. 15 chromatin states inferred in (Ernst et al. 2011)

Chromatin states	**Function domains**
State 1	Active Promoter
State 2	Weak Promoter
State 3	Inactive/poised Promoter
State 4	Strong enhancer
State 5	Strong enhancer
State 6	Weak/poised enhancer
State 7	Weak/poised enhancer
State 8	Insulator
State 9	Transcriptional transition
State 10	Transcriptional elongation
State 11	Weak transcribed
State 12	Polycomb-repressed
State 13	Heterochromatin; low signal
State 14	Repetitive/Copy Number Variation
State 15	Repetitive/Copy Number Variation

2.2.2.6 Uniqueness data

Uniqueness data (hg18) based on reads of 35 bp (wgEncodeDukeUniqueness35bp table) was downloaded from UCSC genome browser to gauge the uniqueness (or mappability) of a given genomic position in the sense that whether a read can be unambiguously mapped back to its originating genomic position (Kent et al. 2002; Ernst et al. 2011; Derrien et al. 2012). The mappability of a given genome region depends on its sequence complexity, read length, and the number of mismatches allowed during mapping. We chose the uniqueness data for 35 bp reads to approximately match the length of 33bp for single-end reads *in vitro*. Though reads are of 25 bp *in vivo*, it is far less of a problem for paired-end reads since both reads for a nucleosome fragment have to be uniquely mapped to two positions of 147 bp apart.

2.2.3 Analysis procedure

2.2.3.1 Reads mapping

To obtain nucleosome positions in the human genome, paired-end sequencing reads were obtained from (Gaffney et al. 2012) for the *in vivo* dataset, and single-end sequencing reads from (Valouev et al. 2011) for the *in vitro* analysis (Table 2.2). Sequence similarity is high between paralogous regions by definition, and therefore it was important to only analyse those regions where it was possible to unambiguously determine the locations of nucleosomes. For this reason when mapping reads to the genome with Bowtie2 for *in vivo* and Bowtie1 for *in vitro* datasets (Langmead et al. 2009; Langmead and Salzberg 2012), no mismatches in the entire read length (25 bp for paired-end reads and 33 bp for single-end reads) were allowed and only uniquely mapped reads were kept. Uniquely mapped reads from different sequencing runs were combined prior to determining nucleosome positions. The *in vivo* dataset was restricted to those reads derived from fragments of exactly 147 bp, and therefore expected to provide base-pair resolution nucleosome locations. The usage of the strictest parameters for mapping reads in our study resulted in ~45 million uniquely mapped and properly paired fragments of 147 bp *in vivo* from 16 replicate experiments and ~408 million unambiguously mapped single-end reads *in vitro* from 6 replicates, compared to a total of ~124 million and ~996 million reads respectively in their corresponding original studies.

Table 2.2. Summary of human *in vivo* and *in vitro* datasets

Datasets	Sequencing strategy	Read length	Total reads	Nucleosomes called	Source
in vivo	Paired-end	25 bp	90,243,756	3,039,065	Gaffney et al. 2012
in vitro	Single-end	33 bp	408,543,428	11,947,675	Valouev et al. 2011

2.2.3.2 Nucleosome positioning inference

To determine dyad positions *in vivo*, genome-wide midpoint profile was first generated, and a 150 bp sliding window with a 10 bp step was run on the midpoint profile, with the locations in each window with the highest fragment midpoint count identified, and annotated as nucleosome dyad if containing at least 4 midpoints. Consecutive dyad calls were not

allowed to overlap by more than 40 bp, by checking whether there was a dyad called in the upstream 107 bp interval, as done in Brogaard et al. (2012). I identified 3,039,065 nucleosomes *in vivo* in total by this method. Due to the single-end sequencing for *in vitro* dataset, nucleosome positions were determined using the iNPS software described above using default parameters (Chen et al. 2014).with a total of 11,947,675 nucleosomes called at a resolution of 10 bp *in vitro*.

2.2.3.3 Defining paralogous nucleosomes

Nucleosomes located in each duplicated region were selected and assigned a relative position with respect to the start of the corresponding duplicated region in both *in vivo* and *in vitro* datasets. Reciprocally closest nucleosome pairs on each duplicate were defined as paralogous nucleosomes, as illustrated by a dummy example shown in Figure 2.2, for nucleosome A1 on paralog A, B1 is the corresponding nucleosome on paralog B that is closest to the location of A1; in turn the nucleosome on paralog A which is closest to B1 is A1. In this case, A1 and B1 are said to be reciprocally closest to each other and are defined as paralogous nucleosomes, and for the same reasoning nucleosomes A4 and B3 are also considered as a paralogous pair. However, though the closest nucleosome for A2 is B2 on paralog B, A2 and B2 are not reciprocally closest to each other because the closest nucleosome for B2 on paralog A is A3; instead, A3 and B2 are considered as paralogous nucleosomes.

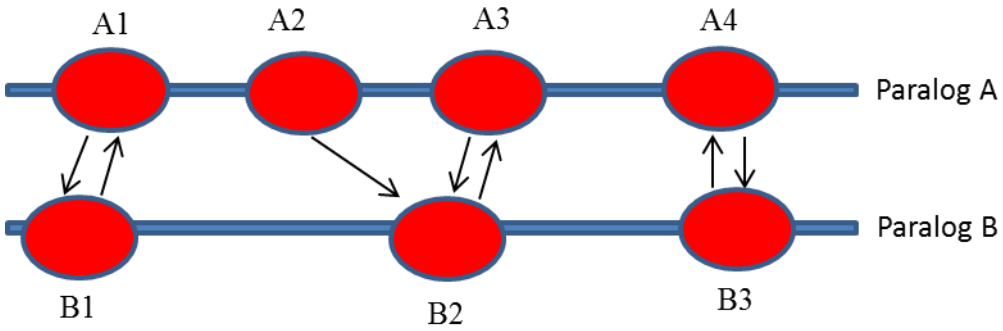


Figure 2.2. Schematic description on defining paralogous nucleosomes.

Due to the rounding strategy adopted by iNPS that each position is rounded down (in a 5 prime direction) to the nearest 10n+1, all dyad positions in vitro end either with a 1 or 6, and there is risk of introducing an extra 10 bp bias between paralogous nucleosomes if one duplicon is aligned on the plus strand while the other is on a minus strand for a given pair of paralogous regions. As a result each nucleosome call had to be shifted 10 bp 3 prime if the containing duplicon was on the minus strand and the other on the plus strand. This is explained schematically in Figure 2.3.

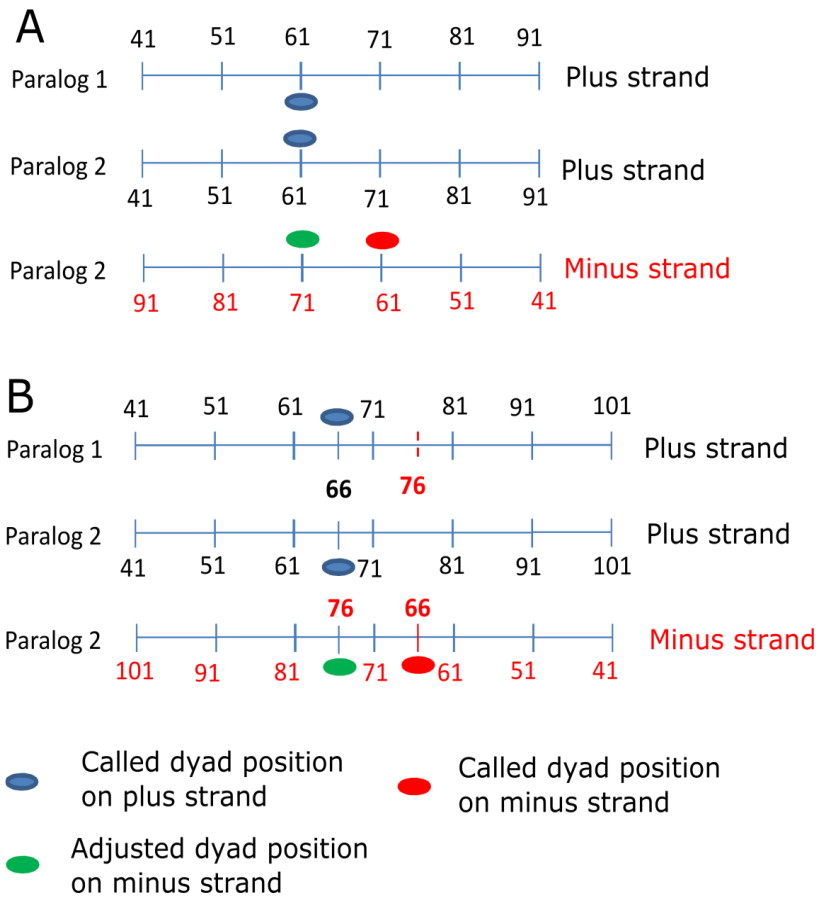


Figure 2.3. Correction of the inflated divergence of *in vitro* nucleosome positioning between paralogous regions due to the iNPS rounding approach. Correction of the inflated divergence of in vitro nucleosome positioning between paralogous regions due to the iNPS rounding approach. The dyad position of a nucleosome is defined as the midpoint of the inflection start and end positions by iNPS. Due to the strategy adopted by iNPS that each position is rounded down (in a 5 prime direction) to the nearest 10n+1, the inflection start position is rounded down to genomic coordinate 41 if the position is in the range [41, 50] (A and B). Similarly, the inflection end position is rounded down to genomic coordinate 81 if in the range [81, 90]

(A) and to 91 if in the range [91, 100] (B). Where the duplicons are inverted (paralog 2 on minus strand) this will lead to the divergence in nucleosome positioning being inflated by 10 bp, as indicated by the distance between nucleosomes coloured by blue and red. To correct this bias, 10 bp was added to the dyad position if the nucleosome is on the minus strand and its paralogous counterpart is on the plus strand, indicated by nucleosomes with green colours.

2.2.3.4 Permutation to assess the conservation in nucleosome positioning

To show whether there was conservation in nucleosome positioning between paralogous regions, I also compared the observed shift against the permuted shift for both in vivo and in vitro datasets (for Figure 2.9A page 59, page 46 of main text). In each pair of duplicons, I firstly retrieved reciprocally closest nucleosomes between duplicons, and the distance was calculated as the observed shift. To obtain permuted shift, I chose one duplicon by random as reference and randomly permute all nucleosomes while keeping the inter-nucleosome distance in the other duplicon. A random number between 1 and the duplicon length was generated and added to the genomic position of each nucleosome to be permuted. If the permuted genomic position of a nucleosome was X bp larger than the higher end point of the duplicon, it was projected at the X bp from the lower end point of the duplicon by considering the duplicon as circular. I then retrieved reciprocally closest nucleosomes between reference duplicon and permuted duplicon, and the distance was calculated as permuted shift. I repeated this permutation process for 1000 times and compared the distribution of observed shifts against that of the 97.5th percentile of permutation.

To show whether there was conservation in positioning between in vivo and in vitro nucleosomes (for Figure 2.5A, page 54, page 41 of main text), I adopted a similar permutation strategy as between paralogous nucleosomes mentioned above. Due to the size of human genome, I randomly chose a 1 Mb region from each chromosome and compared the observed shift between in vivo and in vitro nucleosomes against that from permutation. To obtain the permuted shift, I randomly permuted all in vitro nucleosomes and calculated the distance between in vivo and permuted in vitro nucleosomes as the permuted shift.

2.2.3.5 Defining nucleosome core and linker regions

The core region of nucleosomes was defined as from -73 to +73 bp relative to the reference dyad, which is the centre position of the nucleosomes located on the reference copy that was randomly selected from each pair. The reason for randomly choosing a

member from each duplicated region as the reference copy is due to the constraint of sample size; for example, the ancestral copy of only a very small proportion of duplicons (~ 2%) could be determined in the human genome, and even then the ancestral nucleosome position would be unknown. This is because the nucleosome on the ancestral duplicon may have moved more than its paralog. Linker regions have been estimated at approximately 50bp in humans (Valouev et al. 2011), and so 50bp at each side of the core regions was annotated as linker region. In addition, we set the maximum shift in humans as 100 bp, as similarly done in Kaplan et al. (2008).

2.2.3.6 Permutation analysis to detect the correlation in positioning between proximal paralogous nucleosome pairs

To assess whether the shifts between proximal paralogous nucleosome pairs are correlated or not, the observed frequency of the difference in shifts between proximal paralogous nucleosome pairs was compared to that expected by chance. Shifts between paralogous nucleosome pairs as well as at other nearby paralogous pairs (up to 1 kb) were first recorded and the observed shifts of proximal pairs were then shuffled genome wide and the frequency of individual differences in shifts between nearby nucleosome pairs recalculated, this being repeated 1000 times. If the observed frequency of nucleosome pairs shifting by distance X was greater than the 97.5 percentile of permutations it was deemed significant. Because of the 10 bp resolution of nucleosome positions in the *in vivo* sample due to iNPS, the count of proximal paralogous pairs were smoothed using a 10 bp moving average by R package "caTools" (R Development Core Team 2009; Tuszynski 2014) .

2.2.3.7 Multiple linear regression and relative importance test

The relative contributions of different DNA sequence related features and local chromosomal environment variables to the divergence in nucleosome positioning evolution between paralogous regions were assessed by multiple linear regression (glm() package in R) in all datasets. To find factors with the most influence on divergence in positioning between paralogous regions, step functions, according to Akaike information criterion (AIC), were applied for model selection in R (Venables and Ripley 2002; R Development Core Team 2009). Additionally, the relative importance of different variables to the divergence in nucleosome positioning was investigated by relative importance analysis using the "lmg" method from "relaimpo" R package, which assesses the proportion of variations (relative

importance) in the dependent variable explained by individual regressors in the linear model (Lindeman and Merenda 1980; Ulrike 2006; Groemping and Matthias 2013).

2.2.3.8 The divergence in the occupancy of dyads around motifs in human genome

Motif locations (in hg19) were obtained from HOMER (http://homer.salk.edu/homer/motif/, (Heinz et al. 2010) and lifted to hg18 using default parameters with the UCSC genome browser utility liftOver (Kent et al. 2002). Only relatively short motifs (≤ 50 bp; 242 motifs in total) were kept and analysed. For any given motif, we recorded the midpoint positions of all of its occurrences on both duplicons in each pair of the paralogous regions. For each recorded midpoint position, if there was only one occurrence of this particular motif in the 1001 bp window centred at the midpoint position but none in the 1001bp window centred at the projected midpoint position on the paralogous duplicon then the recorded midpoint position is annotated as motif maintained and correspondingly the projected midpoint position as motif lost. To exclude the possibility that any occupancy changes was due to the region being unmappable to in the original ChIP-seq experiments that defined the motifs, we calculated the number of unique sites in the 101 bp window centred at the recorded and projected midpoints using the wgEncodeDukeUniqueness35bp table (hg18) from the UCSC genome browser respectively (Kent et al. 2002; Derrien et al. 2012). Only pairs with at least 50 unique sites in the 101 bp window at both duplicons were kept and analysed. Positions between -20 bp to +20 bp relative to motif midpoints (observed and projected) were selected and analysed. For each motif, I then calculated a dyad frequency, from all occurrences of this particular motif, at each position selected (-20 to +20 bp relative to motif midpoints) as the ratio of the observed number of dyads to the total number of sites that can be uniquely mapped. Since the number of nucleosomes called *in vivo* is far less than *in vitro* (~3 million *in vivo* compared to ~12 million *in vitro*) and to make the analyses comparable between *in vivo* and *in vitro*, at each position relative to any given observed or projected motif separately both *in vivo* and *in vitro*, I calculated a nucleosome dyad occupancy score, defined as $\log_2(\frac{N}{M})$, where N represents the dyad frequency at each relative position of a given observed or projected motif and M presents the median value of the dyad frequencies at all positions selected from 242 motifs. The difference in dyad occupancy scores between maintained and lost motifs was tested using both the Student's *t*-test and Mann-Whitney test *in vivo* and *in vitro*.

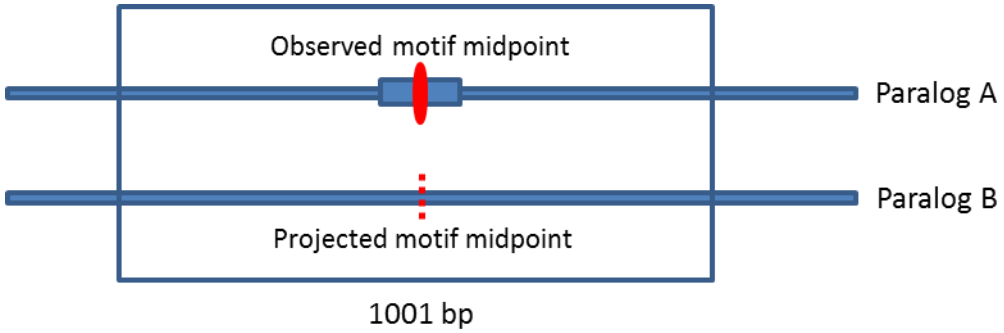


Figure 2.4. Schematic drawing to show that the DNA motif is absent in one duplicon. This can be due to the loss of the motif in one duplicon (such as paralog B) or birth of the motif in the other duplicon (such as paralog A). To decreased the influence on the nucleosome occupancy of proximal occurrence from the same motif, a window of 1001 bp was centered at the midpoint of a given observed motif (on Paralog A) and the occupancy scores of nucleosome dyads from positions -20 to +20 relative to the midpoits of observed (solid red eclipse) and projected (red dashed line) motifs were contrasted in a given pair of paralogous regions only if 1) the observed motif is the only occurrence in this 1001 bp window; and 2) there are at least 50 positions that can be uniquely mapped in each window of 101 bp centered at observed and projected motif points respectively.

2.3 Results

2.3.1 Translational positioning between *in vitro* and *in vivo* nucleosomes

A widely accepted approach to study the importance of DNA sequence in nucleosome positioning has been the comparison of nucleosome positions maps generated *in vitro* and *in vivo* (Kaplan et al. 2008; Zhang et al. 2009; Valouev et al. 2011). To directly investigate the divergence of nucleosome positioning between *in vitro* and *in vivo* human samples, I retrieved 2,472,523 pairs of nucleosomes meeting the maximum inter-dyad distance of 100 bp, as done in Kaplan et al. (2008). *In vivo* nucleosome positions (at up to 1 bp resolution) were called from a previous study that sequenced MNase digested DNA obtained from 5 lymphoblastoid cell lines (Gaffney et al. 2012). The *in vitro* nucleosome positions (10 bp resolution) were determined from a second study that reassembled recombinant *Xenopus* histones onto human genomic DNA derived from neutrophil granulocytes (Valouev et al. 2011).

I observed a symmetry in the distribution of the signed distance around +4 bp, rather than 0 bp as may be expected, due to the rounding strategy used by iNPS such that nucleosome dyad ends with either a 1 or 6 (Figure 2.5A). In addition, it clearly showed conservation in positioning between *in vivo* and *in vitro* nucleosomes (Figure 2.5A). To test whether the observed conservation in nucleosome positioning was significant, I compared the distribution of observed shifts against that from permutation based on a random 10 Mb region from each chromosome. As shown in Figure 2.5B, compared to permutation, the divergence in positioning between *in vivo* and *in vitro* nucleosomes was significantly lower. However, the direct comparison of translational positions *in vitro* against *in vivo* human samples might lack the power to detect DNA sequence driven nucleosome positioning signals, due to the potential technological variations in the generation of *in vitro* and *in vivo* nucleosome positions maps.

Due to the 10 bp resolution issue associated with *in vitro* nucleosome positions map, I was unable to assess the nucleosome rotational positioning by searching for the classic ~10 bp periodicity in the distance between *in vivo* and *in vitro* nucleosomes; instead I tested whether another set of sequence driven nucleosome positioning signal called “container site” by Valouev et al. (2011) was associated with the divergence of translational positioning between *in vitro* and *in vivo* nucleosomes, by comparing the GC content in nucleosome core and AT content in linker regions between nucleosomes showing low and high relative

distances. We divided nucleosome pairs into two equal sized groups based on the median distance of 39 bp.

Figure 2.6 clearly shows that both the GC content in core and AT content in linker are significantly negatively correlated with the distance between the translational positions, suggesting that nucleosomes coinciding with sequences that are rich in GC in core and in AT in linker are more stable consistent with previous findings (Valouev et al. 2011; Gaffney et al. 2012).

Among all factors that affect the nucleosome positioning is the local chromosomal environment (Thurman et al. 2012). By taking advantage of a recent genome-wide chromatin state map from the GM12878 human cell line that closely matches the cell line from which the *in vivo* nucleosomes were derived (Ernst et al. 2011), I compared the distribution of the signed distances between *in vitro* and *in vivo* nucleosomes in the context of individual chromatin states. The nucleosome positioning between *in vitro* and *in vivo* human samples was found to be least conserved at active promoters (chromatin state 1) but most conserved at repetitive/CNV regions (chromatin state 14 and 15), as shown by the detailed distribution of signed distances (Figure 2.7) and point estimation of unsigned distances in context of each chromatin state (Figure 2.8). It is consistent with previous studies suggesting that, in human, nucleosome positioning at cis-regulatory regions such as active promoters is mainly realized by trans-acting factors including chromatin remodelling complex and transcription factors while that at non-regulatory regions is primarily determined by DNA sequences (Schones et al. 2008).

Finally, both effects of chromatin states and DNA sequences on the nucleosome positioning between *in vitro* and *in vivo* human samples were confirmed by the multiple linear regression analysis where the relative contributions of different factors were tested simultaneously (Table 2.3).

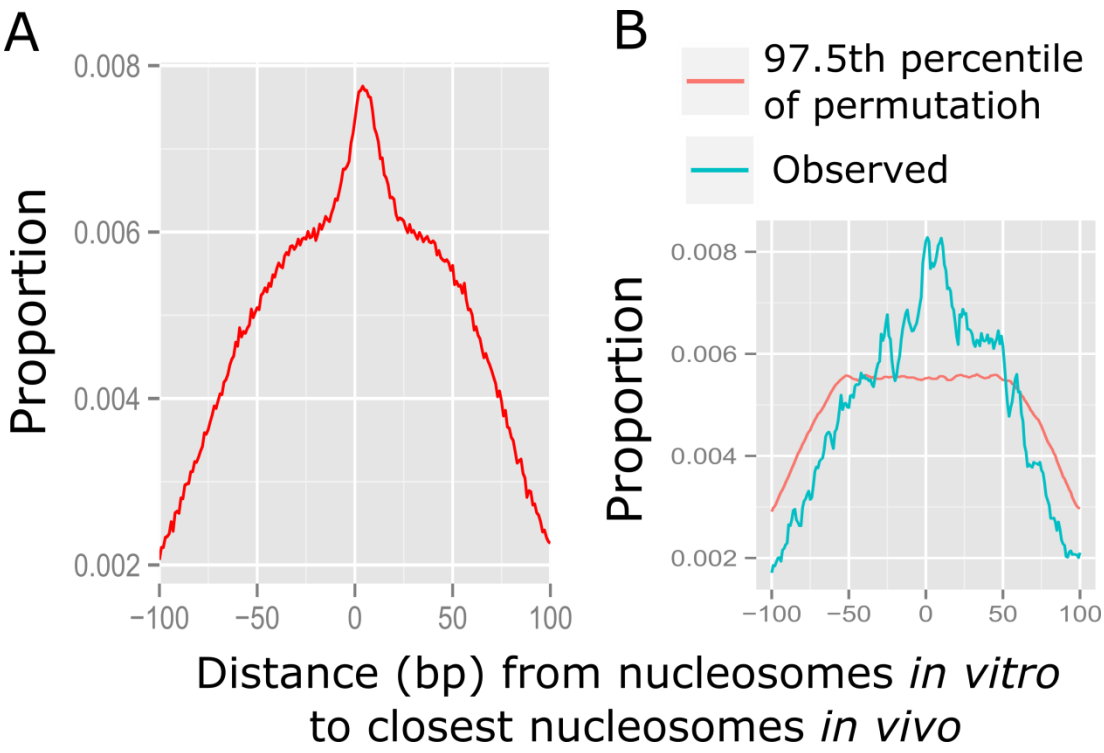


Figure 2.5. Comparison of translational positioning between *in vitro* and *in vivo* nucleosomes. (A). Distribution of signed differences in positions between *in vitro* and *in vivo* nucleosomes. Negative numbers in the X axis correspond to the situation that the *in vivo* nucleosome was located upstream of the *in vitro* nucleosome in each reciprocally closest *in vitro* and *in vivo* nucleosome pair. (B). Comparison of the distributions of observed signed differences in positions and that from permutation. The permutation is based on a randomly seletced 1 Mb region on each chromosome. Blue lines are observed signed differences in positions while red lines are permutated signed differences in positions (based on 1000 permutations) assuming independence between the positions of nucleosomes *in vivo* and *in vitro*.

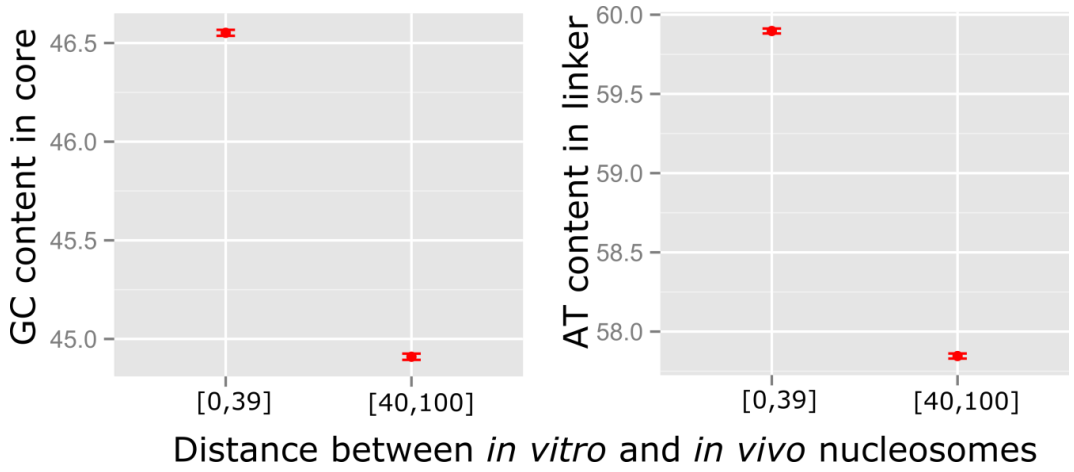


Figure 2.6. Translational nucleosome positioning is associated with underlying sequence composition. Nucleosome pairs were classified into two approximately equal sized groups based

on the median size of their observed distances. The mean and 95% confidence interval for both GC content in core and AT content in linker were plotted for the nucleosomes pairs with the "low" and "high" distances respectively. P values were obtained by the Mann-Whitney test: p < 2.2e-16 for both GC content in core and AT content in linker.

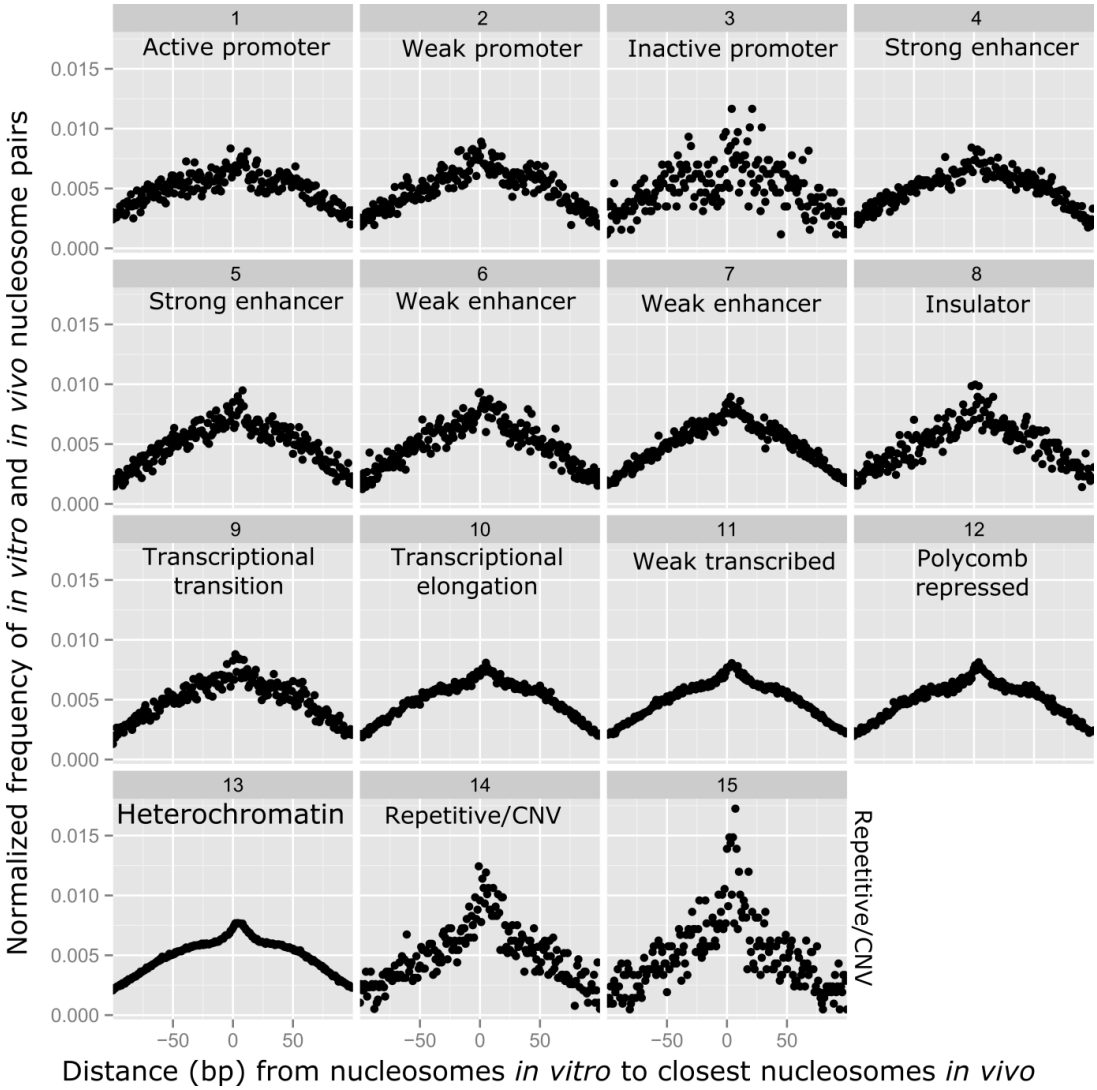


Figure 2.7. Comparison of translational positioning between *in vitro* and *in vivo* nucleosomes in different chromatin states. Distribution of signed differences in positions between *in vitro* and *in vivo* nucleosomes was plotted in each chromatin state. Numbers (1-15) above each figure represent individual chromatin states and corresponding functional domains were also labelled below. CNV: copy number variation.

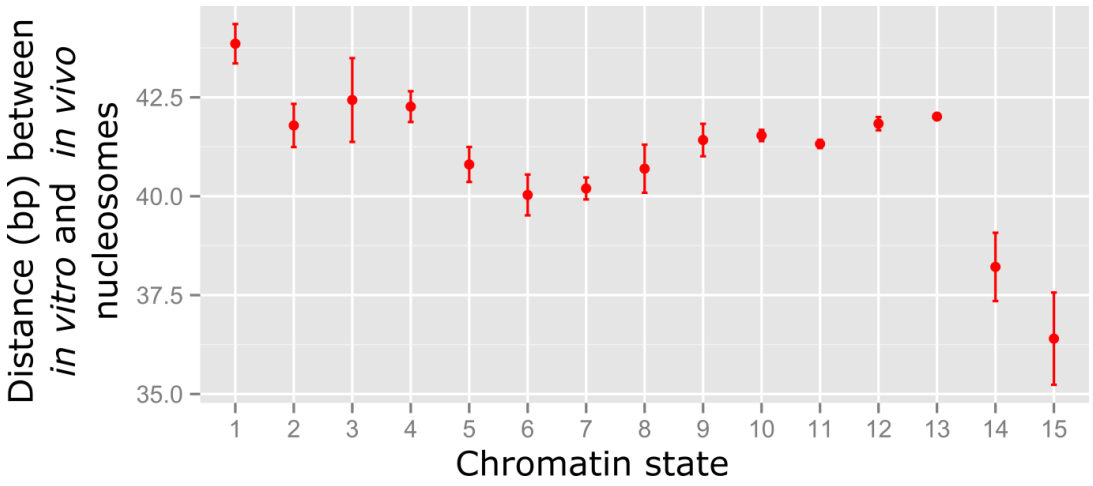


Figure 2.8. Comparison of translational positioning between *in vitro* and *in vivo* nucleosomes in the context of individual chromatin states. The mean and 95% confidence interval for estimated average unsigned distances were plotted within each chromatin state. Kruskal-Wallis rank sum test: p < 2.2e-16. The post-hoc tests (105 tests in total) following kruskall-Wallis test were done by "posthoc.kruskal.nemenyi.test" function from R package "PMCMR" (Pohlert 2015). Post-hoc tests were significant for comparisons of chromatin state 1 (Active promoter) vs. all other states except chromatin state 3 (Inactive promoter, p=0.58). Comparisons of chromatin states 14 and 15 (both correspond to Repetitive/CNV regions) to all other chromatin states were significant while the comparison between chromatin state 14 and 15 was not (0.40). The corresponding functional domain for each chromatin state could be found in Table 2.1 and Figure 2.7.

Table 2.3. Relative effects of chromatin state and DNA compositions on positioning between *in vitro* and *in vivo* nucleosomes by multiple linear regression.

	Coefficient	**p value**
Chromatin State 2 (Weak promoter)	-0.724154	0.04
chromatin State 3 (Inactive/poised Promoter)	-0.519146	0.36
chromatin State 4 (Strong enhancer)	0.09858	0.73
chromatin State 5 (Strong enhancer)	-0.804442	0.01
chromatin State 6 (Weak/poised enhancer)	-1.632703	3.29e-06
chromatin State7 (Weak/poised enhancer)	-1.224527	9.70e-06
chromatin State 8 (Insulator)	-1.834445	1.50e-06
chromatin State 9 (Transcriptional transition)	-0.031133	0.92
chromatin State 10 (Transcriptional elongation)	0.392668	0.12
chromatin State 11 (Weak transcribed)	0.154951	0.53
chromatin State 12 (Polycomb-repressed)	-0.16329	0.52
chromatin State 13 (Heterochromatin; low signal)	0.122605	0.61
chromatin State 14 (Repetitive/Copy Number Variation)	-6.783459	<2.00e-16
chromatin State 15 (Repetitive/Copy Number Variation)	-9.360009	<2.00e-16
GC content in core	-0.990708	<2.00e-16
AT content in linker	-0.991987	<2.00e-16

Note: the reference level of chromatin states is chromatin state.1 (active promoter) and all of other chromatin states were tested against chromatin state 1 which corresponds to "Active promoter". P values for chromatin state levels (2-15) show the probability to observe such a difference from

chromatin state 1 (active promoter) on nucleosome positioning divergence by chance under the null hypothesis that there is really no difference. Correspondingly, p value for GC and AT content means the probability to observe an effect on nucleosome positioning divergence by chance if there is really no effect.

2.3.2 Translational positioning between paralogous nucleosomes within both *in vitro* and *in vivo* human samples

2.3.2.1 Nucleosome positioning is generally well conserved following duplication events

To investigate the positioning of nucleosomes following the duplication of a DNA sequence and its insertion into a new genomic region, I characterised the positioning of nucleosomes at paralogous segments of the human genome using publically available *in vivo* and *in vitro* datasets (Valouev et al. 2011; Gaffney et al. 2012). Since the focus was a comparison between homologous regions, a very stringent mapping approach was used in which no mismatches were permitted when mapping reads back to the reference genome. I identified 14,023 and 167,013 paralogous nucleosome pairs in a total of 53,626 segmental duplicates that are ≥ 90% in sequence identity and ≥ 1 kb in length *in vivo* and *in vitro* respectively.

If chromosomal location or aspects of the broader chromatin environment are the primary determinants of nucleosome positioning we would expect to see little conservation in the positioning of nucleosomes following a duplication event in both the *in vivo* and *in vitro* samples. However, 35% (4,900 pairs) of the *in vivo* paralogous pairs of nucleosomes displayed no change in position following a duplication event (Figure 2.9A); whereas *in vitro,* 25% (42,263 pairs) of paralogous pairs of nucleosomes were observed to have shifted by less than 10 bp, the approximate resolution of the algorithm used to determine nucleosome positioning in this dataset (Chen et al. 2014). The median shift between paralogous nucleosomes was 20bp *in vivo* and 27bp *in vitro;* interestingly the median distance between *in vitro* and *in vivo* nucleosomes was 39 bp while the sequence is identical, suggesting that the comparison of paralogous regions within the same cells could indeed reduce the confounding effects of *trans*-acting factors and/or control for the technical variations in generation of nucleosome positions maps from different samples. I conclude that following the duplication of a region, nucleosomes in general assemble at broadly similar locations on the original and duplicated copies with a third assembling at identical positions *in vivo*. The broad similarity between *in vivo* and *in vitro* datasets suggests an important relative contribution of local DNA sequence to nucleosome positioning.

Nucleosomes are able to assemble at approximately the same location following the duplication of DNA sequences and their insertion into a new genomic location even in the absence of *trans*-acting factors.

2.3.2.2 Nucleosome positioning divergence shows strong periodicity in vivo

The *in vivo* nucleosome map used in this study allowed me to examine nucleosome divergence in detail, and investigate how nucleosome positioning diverges. As shown in Figure 2.9B, paralogous nucleosomes were observed to be preferentially shifted by a multiple of a complete DNA helical turn. The DNA double helix makes one complete turn every ~10.4 bp and it has been shown previously that the DNA sequences underlying nucleosomes have a distinct 10 bp dinucleotide periodicity (Albert et al. 2007; Mavrich et al. 2008; Brogaard et al. 2012; Gaffney et al. 2012). Consequently shifts of complete helical turns between paralogous nucleosomes would be expected to maintain this dinucleotide periodicity relative to the dyad. As far as I am aware, this is the first time that a strong rotational preference in human nucleosome positioning evolution has been observed. Thus nucleosomes do not therefore appear to simply drift from their original locations; rather they preferentially relocate a full helical turn, which argues for important constraints imposed by the DNA helix on nucleosome positioning evolution.

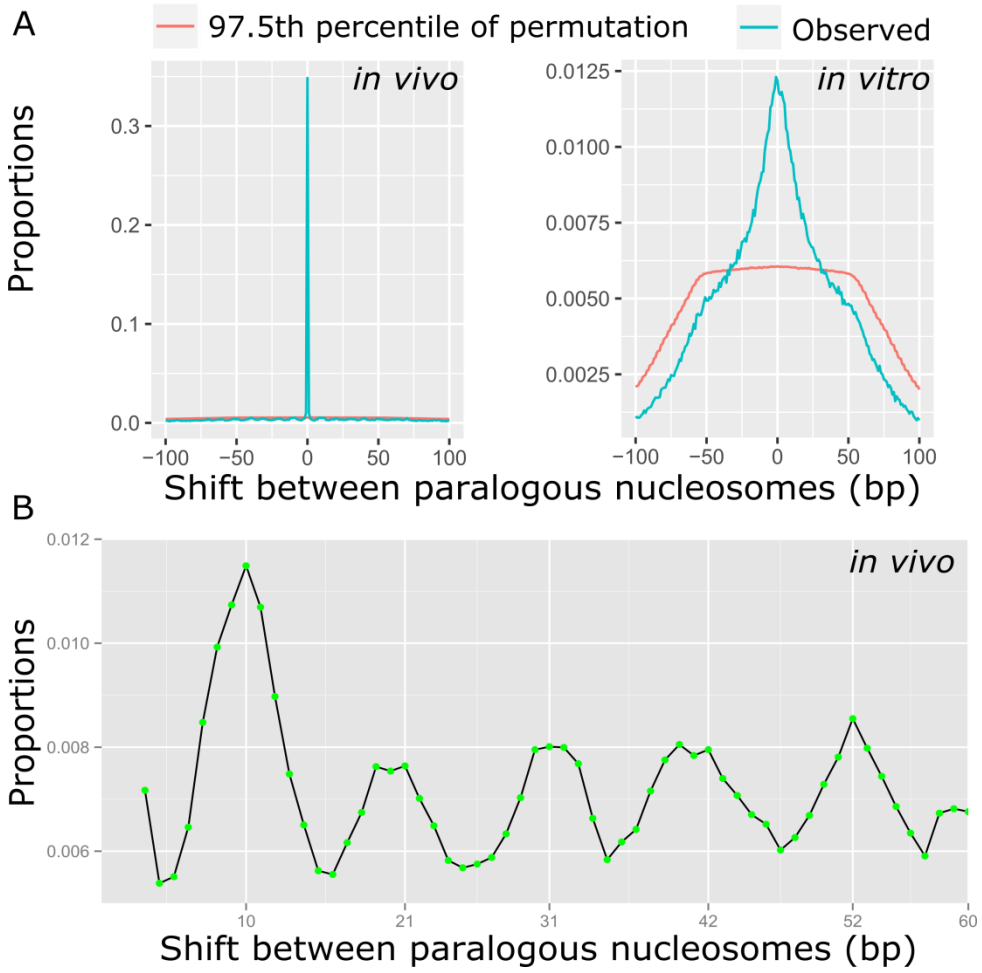


Figure 2.9. Nucleosome positioning is generally conserved between paralogous regions. (A). Distribution of differences in positions (shift) between paralogous nucleosomes *in vivo* and *in vitro* respectively. Negative numbers in the X axis correspond to a shift 5 prime relative to the randomly selected reference dyad. Blue lines are observed shifts between paralogous nucleosomes while red lines are permutated shifts (based on 1000 permutations) assuming independence between the positions of nucleosomes across duplicons. (B). Observed periodicity in nucleosome positioning divergence of ~10 bp in *in vivo* dataset. A complete helical turn is approximately 10.4 bp, so the corresponding distances of 1 to 5 helical turns of 10, 21, 31, 42, and 52 are indicated.

2.3.2.3 Correlated evolution among neighbouring nucleosomes

Nucleosome positioning has been postulated to follow the "statistical principle", in which the positioning of one nucleosome acts as a barrier and impacts the location of other neighbouring nucleosomes (Kornberg and Stryer 1988). I sought evidence for the influence of this phenomenon in nucleosome positioning evolution by investigating the correlations seen in shifts among proximal paralogous nucleosome pairs within a maximum distance of 1000 bp. To assess the significance of any observations I permuted the positions of nucleosome pairs across the genome, disrupting the link between proximal nucleosome pairs

but maintaining the relationship between each nucleosome and its paralog (Figure 2.10A). I was then able to determine whether there were dependencies between the shifts in the positioning of neighbouring nucleosome pairs above what would be expected by chance, given the distribution of shifts between paralogous nucleosomes observed genome-wide.

I found a significant enrichment of nucleosomes within close proximity that display approximately the same shifts relative to their corresponding paralogous nucleosome partners in both the *in vivo* and *in vitro* samples, indicated by the significantly higher frequency of zero differences in shift between paralogous nucleosomes in observed data than in the permutations (Figure 2.10B). I conclude that there are correlations between the extent of nucleosome repositioning following a duplication event and nearby nucleosomes (up to 1 kb) are observed to often shift by approximately the same distance supporting a role for long range effects of nucleosome positioning evolution even in the absence of *trans*-acting factors.

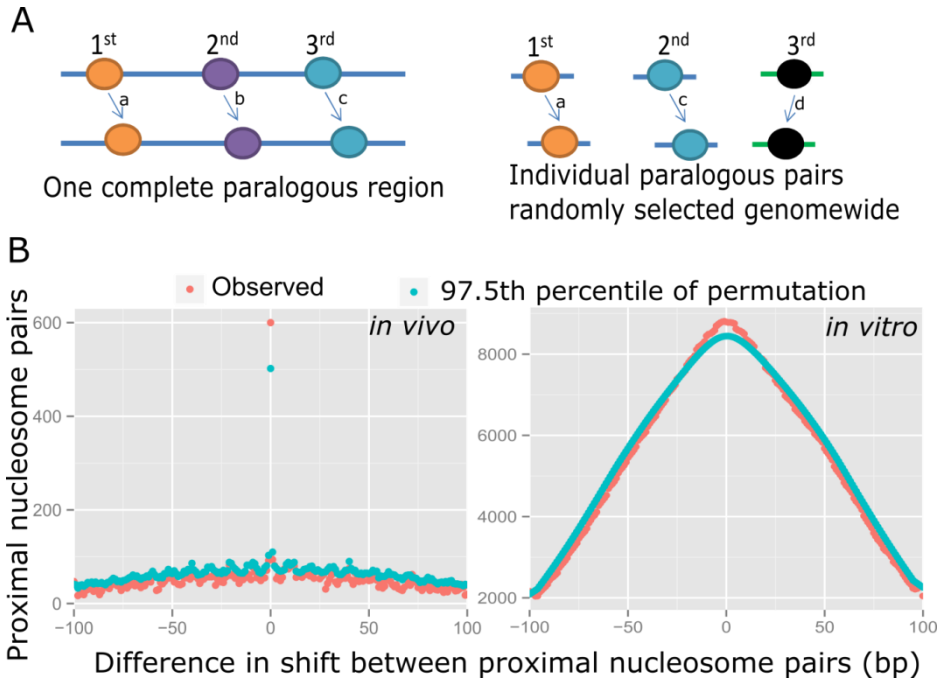


Figure 2.10. Shifts among proximal paralogous nucleosome pairs are correlated. (A). Schematic description of permutations. For any given paralogous nucleosome pair, the shift in their relative positioning was compared with that observed at pairs of paralogous nucleosomes within 1000 bp. To assess whether the observed shifts at neighbouring nucleosome pairs were correlated, pairs of nucleosomes were randomly shuffled 1000 times disrupting the link between neighbouring nucleosome pairs but maintaining the relationship between a nucleosome and its paralog. The difference in the shifts observed between neighbouring nucleosome pairs in the real data and

permuted data was then compared. For example in the left panel, to test whether a≈b and/or a ≈ c the difference in these shifts was compared to that observed at randomly selected pairs of paralogous nucleosomes (permutation; right panel). I repeated this process 1000 times. a and b were observed to more often be approximately the same in the real data than the permuted data highlighting that neighbouring pairs of nucleosomes display a correlation in their positioning divergence. (B). *in vivo* and *in vitro* samples. Because the resolution for *in vitro* nucleosome positioning is 10bp, data points were smoothed by a 10 bp moving average.

2.3.2.4 Sequence level determinants of nucleosome positioning

The conserved positioning of paralogous nucleosomes supports an important role for DNA sequence in nucleosome positioning (Segal et al. 2006; Kaplan et al. 2008; Valouev et al. 2011; Brogaard et al. 2012). To explore this relationship further, I investigated how various sequence features of duplicated regions are related to nucleosome positioning stability. It is thought that differential AT and GC base composition in core and linker regions is linked to nucleosome positioning, such that nucleosomes preferentially assemble at high GC content regions with relatively high AT content at their linker regions (Reynolds et al. 2010; Valouev et al. 2011). However it has been difficult to disentangle cause and effect for these associations, which may for example be a result of biased mutational spectra at core and linker regions, as hinted to be the case in yeast species by (Xing and He 2015). To investigate this further I compared the GC content in core and AT content in linker regions between nucleosomes showing conserved positions between paralogous regions with those showing larger relative shifts. To balance sample sizes, I divided the nucleosomes in both the *in vivo* and *in vitro* samples into two, approximately equal sized "high shift" and "low shift" groups based on the degree of their observed divergence in positioning. More stably positioned nucleosomes were observed to be associated with both a significantly higher average GC content in core regions and significantly elevated AT content in linker regions in both the *in vivo* and *in vitro* samples (Figure 2.11A). In addition, nucleosome positioning divergence was found to be significantly correlated with DNA composition by Spearman correlation tests (GC content in core, rho=-0.07 and p<2.2e-16 for *in vivo*, and rho=-0.07 and p<2.2e-16 for *in vitro*; AT content in linker, rho=-0.04 and p=1.3e-07 for *in vivo*, and rho=-0.03 and p<2.2e-16 for *in vitro)*. I then tested whether changes in DNA composition between duplicons is associated with divergence in nucleosome positioning. As can be seen in Figure 2.11B, divergence in both the GC content in core regions and AT content in linker regions is significantly correlated with changes in positioning between paralogous nucleosomes in both the *in vivo* and *in vitro* samples, supported by correlation

tests (Divergence in DNA composition in core, rho=0.07 and p=8.9e-15 for *in vivo,* and rho=0.05 and p<2.2e-16 for *in vitro;* Divergence in DNA composition in linker: rho=0.09 and p<2.2e-16 for *in vivo,* and rho=0.06 and p<2.2e-16 for *in vitro*). This suggests the strength of compositional bias generally associates with the stability of a nucleosome's position.

In addition to the role of DNA composition, I also examined the extent to which total DNA sequence divergence is correlated with nucleosome repositioning. Divergence in nucleosome positioning was found to be positively correlated with the amount of sequence divergence across the duplicated segments (Spearman test: *in vivo*, rho=0.15, p< 2.2e-16; *in vitro*, rho=0.09, p<2.2e-16). The divergence in nucleosome positioning between paralogous regions being significantly correlated with DNA sequence divergence in the nucleosome core, linker and whole nucleosome regions (core+linker) both *in vivo* and *in vitro* (Figure 2.12, top panel). In addition, the strength of the correlation between sequence and nucleosome positioning divergence was observed to be linked to whether the sequence divergence changed DNA composition. Both *in vivo* and *in vitro,* the correlation between sequence divergence and nucleosome repositioning was higher when examining sequence divergence that changes DNA composition (AT <=> GC) relative to changes that do not (AT <=> TA or GC <=> CG) (Figure 2.12).

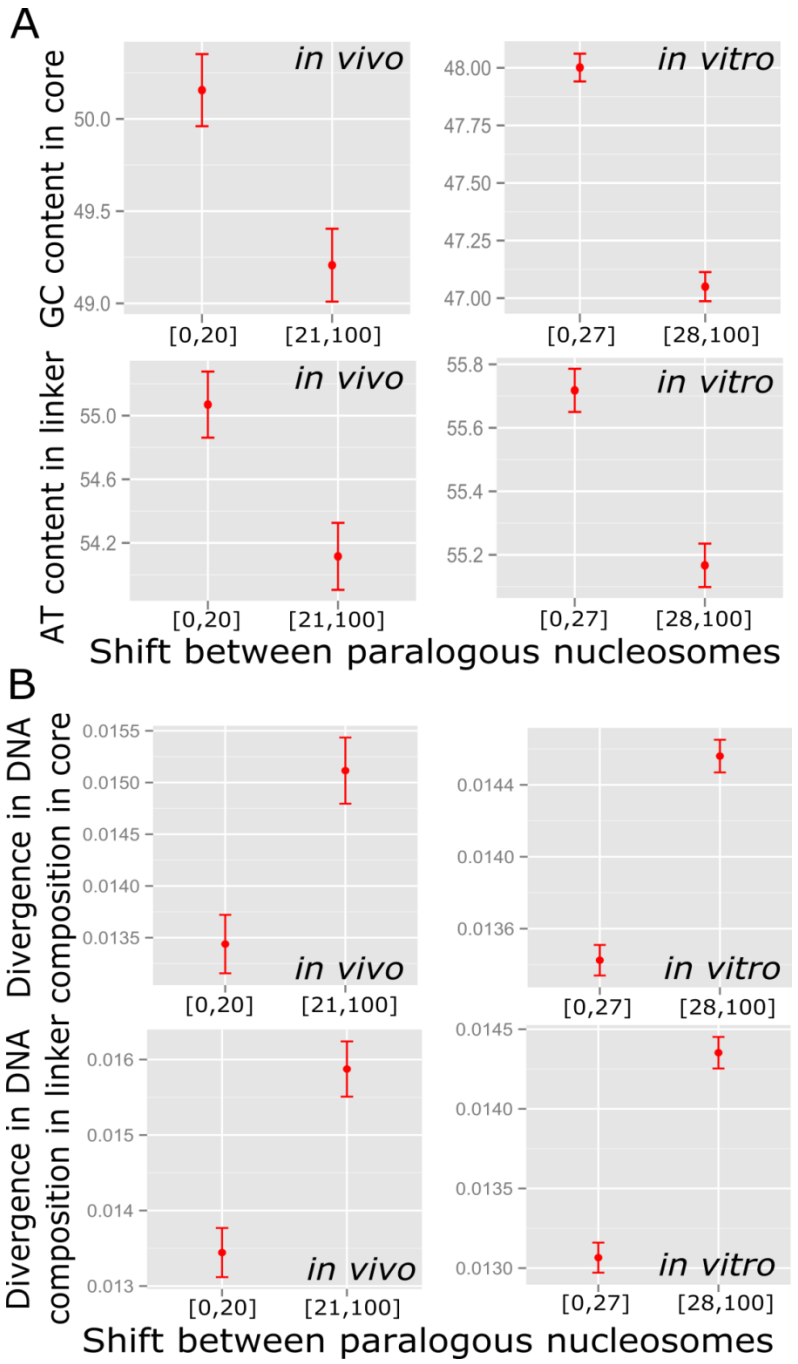
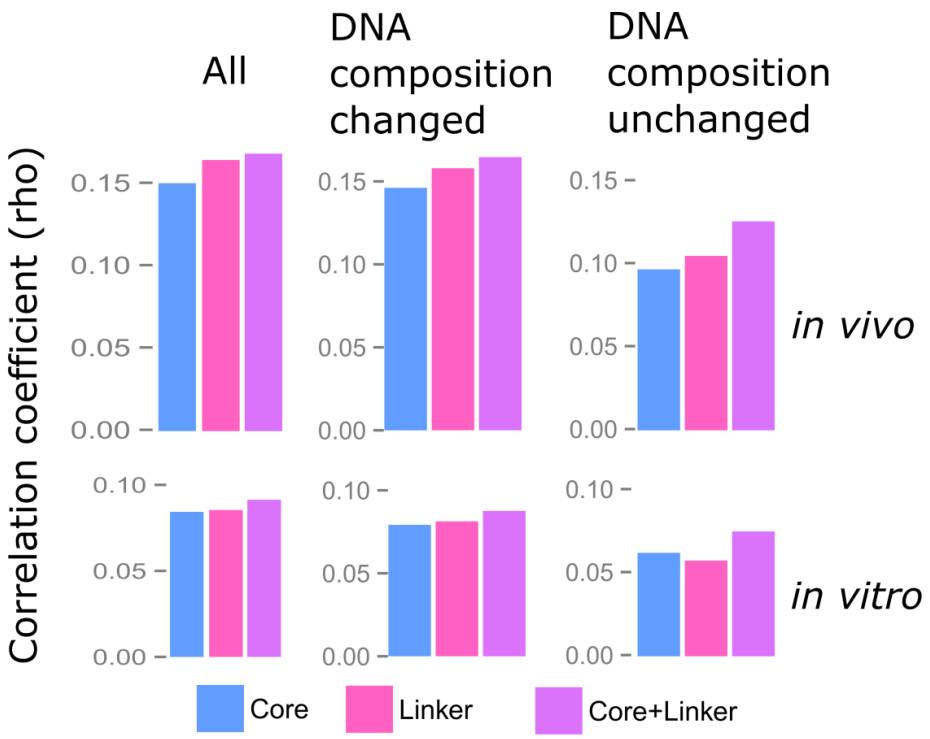


Figure 2.11. Nucleosome positioning divergence is associated with underlying sequence composition. Nucleosome pairs were classified into two approximately equal sized groups based on their observed shift. (A). Core GC and linker AT compositions associated with the "low" and "high" shift groups. GC content in core Mann-Whitney p=1.0e-11 for *in vivo* and p=1.5e-99 for *in vitro* samples; AT content in linker p=5.3e-09 for *in vivo* and p=7.8e-30 for *in vitro*. (B). Divergence in DNA composition associated with the "low" and "high" shift groups. Divergence in DNA composition in core Mann-Whitney p=8.8e-13 for *in vivo* and p=1.8e-71 for *in vitro* samples; Divergence in DNA composition in linker p=2.0e-23 for *in vivo* and p=1.9e-97 for *in vitro.* In addition, to address the potential biases introduced by the arbitrary grouping of paralogous nucleosomes based on the median size of

unsigned shift, Correlation between the unsigned shift and DNA sequence composition was assessed by Spearman correlation tests. (A): GC content in core, rho=-0.07 and p<2.2e-16 for *in vivo*, and rho=-0.07 and p<2.2e-16 for *in vitro*; AT content in linker, rho=-0.04 and p=1.3e-07 for *in vivo*, and rho=-0.03 and p<2.2e-16 for *in vitro*. (B): Divergence in DNA composition in core, rho=0.07 and p=8.9e-15 for *in vivo*, and rho=0.05 and p<2.2e-16 for *in vitro;* Divergence in DNA composition in linker: rho=0.09 and p<2.2e-16 for *in vivo*, and rho=0.06 and p<2.2e-16 for *in vitro*.



		p value	
Region	Type	*in vivo*	*in vitro*
Core	All	3.99e-78	2.17e-275
Linker	All	2.69e-93	2.68e-282
Core+Linker	All	1.09e-97	4.00e-323
Core	DNA composition changed	8.77e-74	1.04e-243
Linker	DNA composition changed	5.04e-86	1.24e-256
Core+Linker	DNA composition changed	2.07e-93	4.31e-298
Core	DNA composition unchanged	2.51e-33	1.62e-151
Linker	DNA composition unchanged	7.53e-39	4.69e-130
Core+Linker	DNA composition unchanged	4.19e-55	3.62e-220

Figure 2.12. Sequence divergence (number of base changes) is correlated with divergence in positioning between paralogous nucleosomes (unsinged shift) *in vivo* and *in vitro*. Sequence divergence was categorized into two types based on whether it changes DNA composition or not: DNA composition changed (AT <=> GC) and DNA composition unchanged (AT <=> TA or GC <=> CG). Correlation was then assessed in nucleosome core, linker and whole nucleosome regions (core+linker). Divergence in positioning between paralogous nucleosomes was tested using the Spearman non-parametric correlation. Correlation coefficients (rho) were plotted in the top panel and corresponding p values are shown in the bottom panel for both *in vivo* and *in vitro* samples.

2.3.2.5 The local genomic environment affects positioning between paralogous nucleosomes

Though DNA composition and sequence divergence appear to be important determinants of nucleosome positioning, the different local chromosomal environments that duplicons occupy may also play important roles in divergence in positioning between paralogous nucleosomes. I first assessed whether the divergence in the chromatin states of nucleosomes in each paralogous pair is associated with the divergence in positioning. I found that the association only existed across *in vivo* paralogous nucleosomes: paralogous nucleosomes show a more conserved positioning if the chromatin state is the same (Figure 2.13). However, the absence of the effect of chromatin state on *in vitro* paralogous nucleosomes is not surprising, since the chromatin state is established by epigenetic marks and chromatin profiles, none of which would be expected to affect the positions of nucleosomes reconstructed *in vitro* that only involves the DNA sequence and canonical core histones from another species.

In addition, the divergence in positioning between paralogous nucleosomes in duplications which have occurred between different chromosomes (inter-chromosomal) is significantly higher than those on the same chromosome (intra-chromosomal) *in vivo* and *in vitro* (Figure 2.14). Although inter-chromosomal duplications are generally older and therefore more divergent at the sequence level (Kimura 1983), changes in nucleosome positioning remain significantly larger between human inter-chromosomal duplicons having accounted for levels of sequence divergence (*in vivo*, p=7.3e-10; *in vitro*, p=3.4e-79).

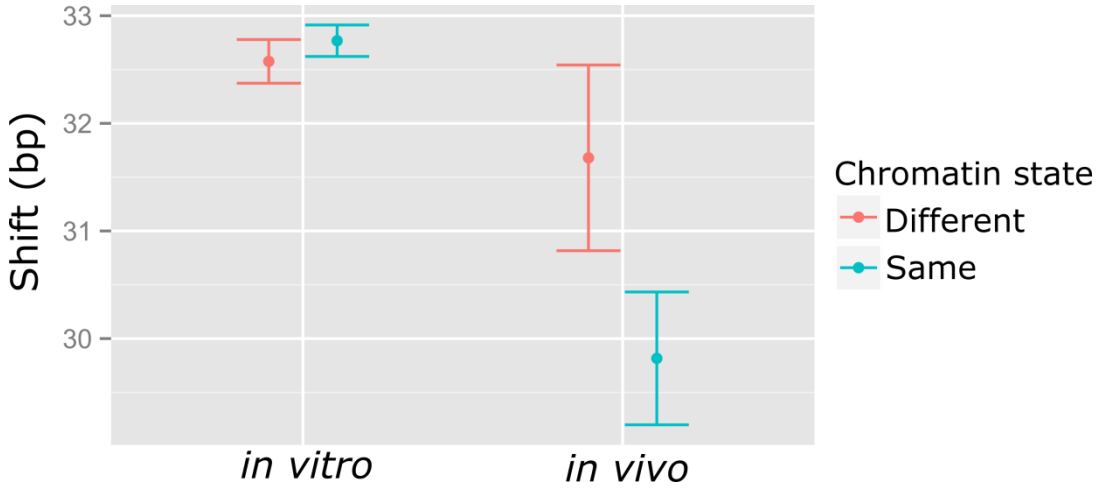


Figure 2.13. Local chromatin states affect divergence in positioning between paralogous

nucleosomes *in vivo* but not *in vitro*. The mean and 95% confidence interval were plotted for the shifts between paralogous nucleosomes that are either in the same chromatin (red) or different chromatin states (blue) for both *in vitro* and *in vivo*. P values obtained from the comparison of the unsigned shifts between paralogous nucleosomes by the Mann-Whitney test: *in vitro*, p=0.22; *in vivo*, p=1.3e-4.

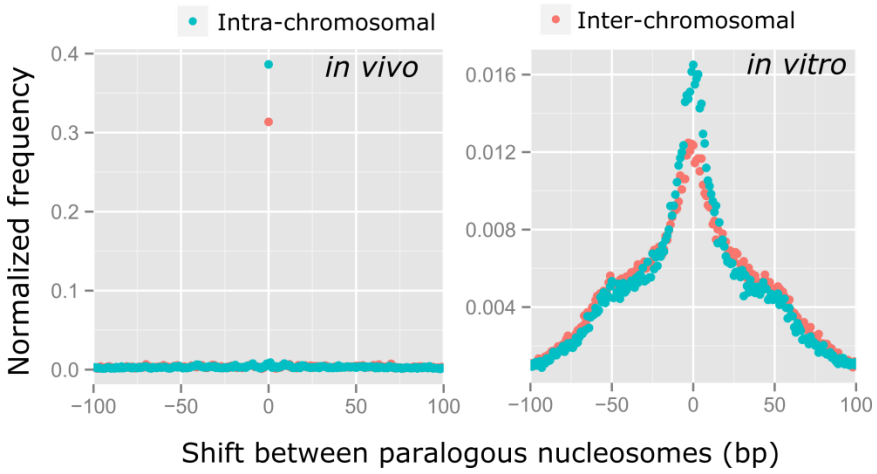


Figure 2.14. Local chromosomal environments affect divergence in positioning between paralogous nucleosomes *in vivo* and *in vitro*. Distribution of differences in positions (shift) between paralogous nucleosomes is plotted for both intra-chromosomal (blue points) and inter-chromosomal duplications (red points). Negative numbers in the X axis correspond to a shift 5 prime relative to the randomly selected reference dyad. P values obtained from the comparison of the unsigned shifts between paralogous nucleosomes by the Mann-Whitney test: *in vivo*, p=1.3e-24; *in vitro*, p=5.4e-160.

2.3.2.6 Strong agreement between role of sequence features in vivo and in vitro

Thus I can discern the influence of many, doubtless inter-correlated, variables upon the divergence of nucleosome positioning, from sequence composition to various aspects of genomic function, including promoters and TSSs. Human promoters are GC rich and intrinsically nucleosome favouring but the nucleosome occupancy and positioning at 5' prime of genes are under the dynamic regulation of transcription activity (Schones et al. 2008; Tillo et al. 2010; Vavouri and Lehner 2012). To disentangle the relative contributions of these variables to nucleosome repositioning in both *in vivo* and *in vitro* samples, I included eight different variables in multiple linear regression models to assess their relative importance to nucleosome positioning *in vivo* and *in vitro*. The features linked to divergence in nucleosome positioning between paralogous regions were observed to be highly correlated

between *in vivo* and *in vitro* (Table 2.4, Figure 2.15). Not only was there a large overlap in the factors significantly linked to nucleosome position divergence in both species, but also the size and direction of their coefficients were generally similar, suggesting they have a similar impact on nucleosome positioning. Consequently the relative contribution of DNA sequence and chromosomal environment in nucleosome positioning appears to be largely consistent whether in the presence or absence of *trans*-acting chromatin binding factors.

Relative importance analysis revealed that GC content at the nucleosome core appears to have a considerably stronger relative association with nucleosome positioning *in vitro* than *in vivo* (Figure 2.15). On the other hand sequence divergence in both linker and core regions appears more important *in vivo.* I hypothesised that this may be due to a role for DNA binding proteins in nucleosome positioning whose motifs may be disrupted through sequence divergence. The lack of these proteins in the *in vitro* sample would mean that sequence changes at their motifs would be irrelevant to nucleosome positioning in the resulting *in vitro* dataset.

To explore this further I compared nucleosome occupancies between corresponding paralogous regions where a given motif is maintained in one duplicon but is lost in the other, encompassing 242 motifs (see Appendix). The loss of particular protein binding motifs was notably associated with changes in nucleosome occupancy between duplicons *in vivo* but not *in vitro.* For example disruption of CDX2 and TEAD2 binding motifs was associated with increased occupancy *in vivo,* but not *in vitro,* consistent with abrogated binding of this protein at these locations through the disruption of its motif, with associated effects on nucleosome occupancy (Figure 2.16). A role for CDX2 in nucleosome positioning agrees with previous observations of nucleosome occupancy increasing at CDX2 binding sites in CDX2 knockout cells (Verzi et al. 2010) but to my knowledge this is the first time that binding of TEAD2, involved in the hippo signalling pathway (Je et al. 2015: 2), has been linked to nucleosome positioning. For the vast majority of motifs their gain or loss had no detectable effect on nucleosome occupancy. It is unclear how representative paralogous regions are since paralogous regions comprise only 5% of the human genome sequence (Marques-Bonet et al. 2009) and thus our focus on the paralogous regions might lack the power to detect the effects on nucleosome occupancy of rare motifs. In addition, a number of proteins are expressed in a cell type restricted manner, and further study of other cell types using this approach may highlight further motifs whose loss is linked to nucleosome positioning evolution in particular cell types.

Table 2.4. Relative effects of DNA local sequences and chromosomal environments on positioning between paralogous nucleosomes by multiple linear regression.

	in vivo		***in vitro***	
	Coefficient	**p value**	**Coefficient**	**p value**
Intra-chromosomal duplication[1]	-3.55e+00	2.28e-10	-2.31e+00	<2.0e-16
Average duplication length (kb)	-2.43e-02	2.12e-03	-1.54e-02	<2.0e-16
Average distance to TSS (log2)	-5.13e-01	1.28e-02	-1.15e-01	0.229
Difference in distance to TSS (log2)	2.79e-01	2.7e4-02	6.90e-01	4.66e-12
GC content in core (percentage)	-7.42e-01	<2.0e-16	-5.85e-01	<2.0e-16
AT content in linker (percentage)	-5.80e-01	<2.0e-16	-4.25e-01	<2.0e-16
Sequence divergence in core (number of changes)	3.02e-01	3.61e-06	1.48e-01	<2.0e-16
Sequence divergence in linker (number of changes)	5.56e-01	1.39e-10	1.32e-01	4.67e-10
Interaction between average distance and difference in distance to TSS	NA	NA	-2.61e-02	7.53e-05

[1]The reference level being inter-chromosomal duplications.
"NA" means variable was dropped following model selection according to AIC.

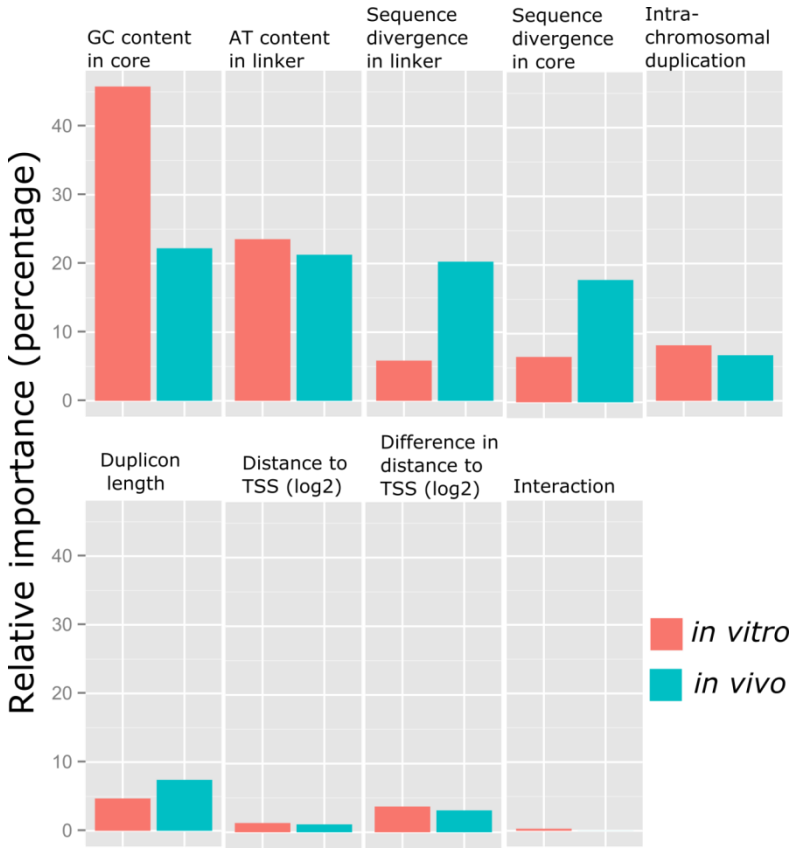


Figure 2.15. Relative importance analysis of local sequence and chromosomal environment features on positioning divergence *in vivo* and *in vitro*.

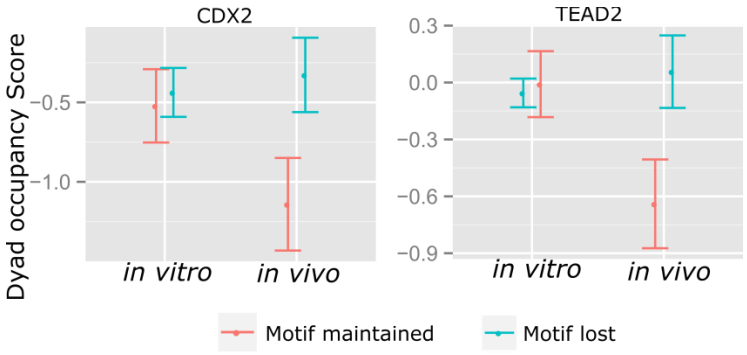


Figure 2.16. CDX2 and TEAD2 binding motifs affect nucleosome positioning between paralogous regions *in vivo* but not *in vitro*. The mean and 95% confidence interval were plotted for the dyad occupancy scores at the 41 positions from -20 bp to +20bp relative to the midpoint of the motif and compared between paralogous duplicons with (Motif maintained) and without the motif (Motif lost) both *in vivo* and *in vitro*. Bonferroni corrected p values following a Mann-Whitney test were: CDX2 p= 1 for *in vitro* and p= 0.034 for *in vivo*; TEAD2 p=1 for *in vitro* and p=0.004 for *in vivo*.

2.3.3 No obvious effect of the difference in mapping stringency and mapping software on the inference of nucleosome positions and the pattern of paralogous nucleosome positioning

Due to the high sequence similarity between segmental duplicates, I used the strictest parameters when re-mapping paired-end reads (stringent mapping). To assess the implication of the mapping stringency, we also obtained the genome wide midpoint profile in the same *in vivo* sample from Gaffney et al. (2012) where default parameters were applied to map paired-end reads back to reference genome by BWA read mapper (Li and Durbin 2009) and midpoints were inferred from fragments of exact 147 bp (http://eqtl.uchicago.edu/nucleosomes/midpoints/mnase_mids_combined_147.wig.gz). Nucleosome positions were inferred exactly the same way as for *in vivo* nucleosomes from stringent mapping (no mismatch and mapped by Bowtie2 read mapper). As shown in Figure 2.17, the positions of 93% and 94% of nucleosomes genome-wide and within segmental duplicates were identical between two different mapping approaches for *in vivo* human sample. In addition, the pattern of relative positioning between paralogous nucleosomes derived from this default mapping approach was similar as that derived from the strictest mapped approach discussed above (Figure 2.18). Thus I conclude that mapping stringency

and choice of read mappers are not an issue in comparing nucleosome positioning between paralogous regions at least if nucleosome positions are inferred from paired-end reads of exact 147 bp apart.

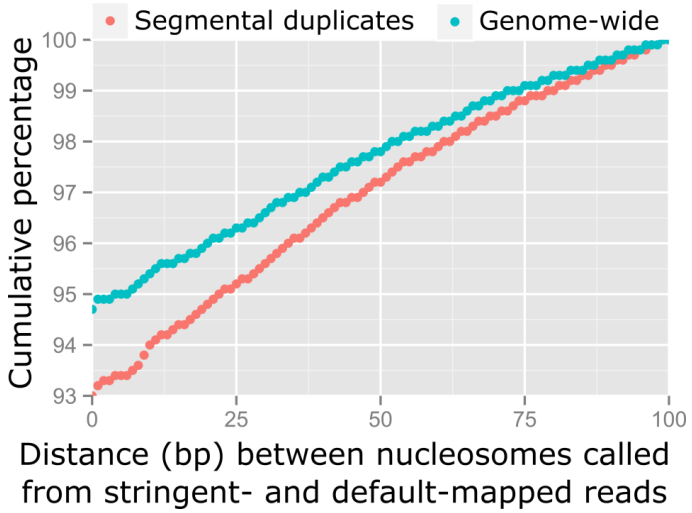


Figure 2.17. Comparison of *in vivo* nucleosomes derived from differently mapped pair-end reads. The cumulative distribution was plotted from a total of 2,864,947 and 94126 reciprocally closest nucleosomes genome-wide (blue) and within duplicated segments (red) within a distance up to 100 bp.

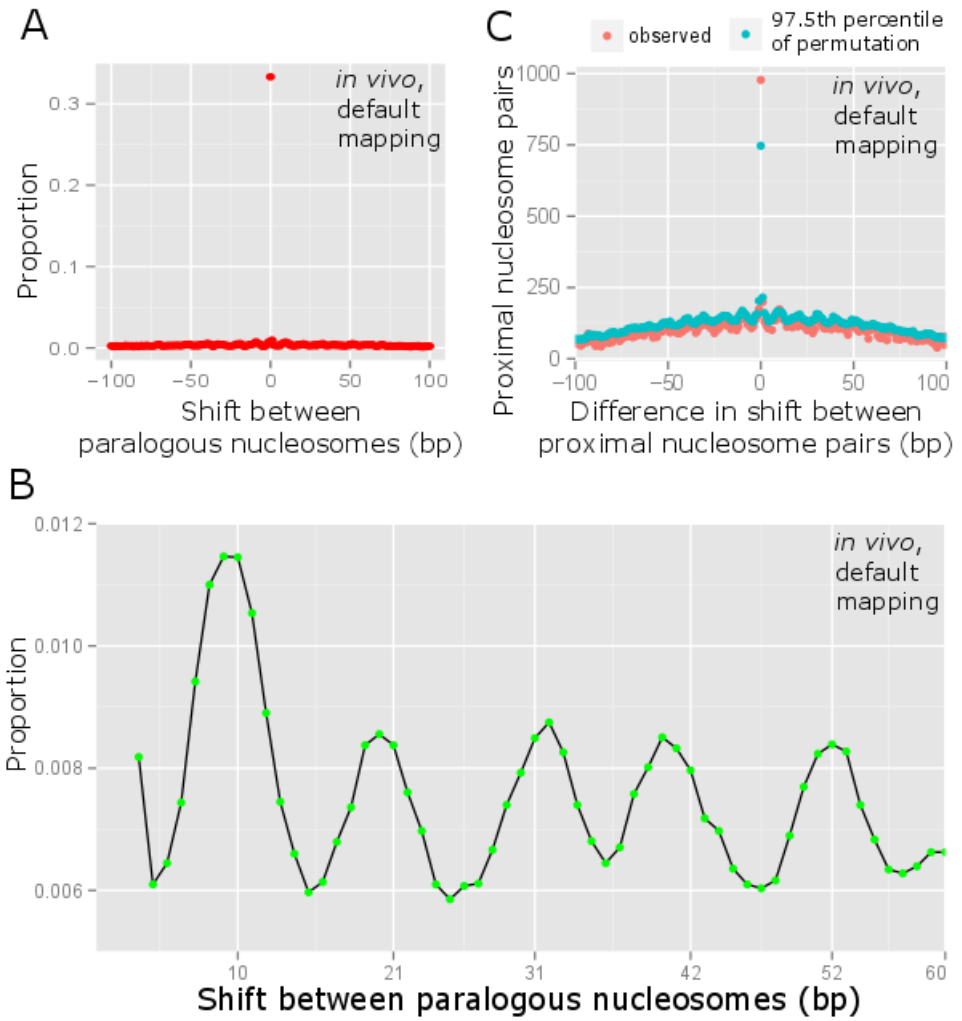


Figure 2.18. Pattern of nucleosome positioning between paralogous regions derived from paired-end reads that mapped to the human reference genome with default parameter and by BWA. (A). Distribution of signed differences in positions (shift) between paralogous nucleosomes. (B). Observed periodicity in nucleosome positioning divergence of ~10 bp. The corresponding distances of 1 to 5 helical turns of 10, 21, 31, 42, and 52 are indicated, based on the assumption that a complete helical turn is approximately 10.4 bp. (C). Shifts among proximal paralogous nucleosome pairs are correlated.

2.4 Discussion

In the study by Valouev et al. (2011), the involvement of the DNA sequence in placing nucleosomes at many genomic sites has been deduced by the observation of the consistent peaks of ~147 bp in the distograms and of well positioned nucleosomes at many genomic positions with the underlying DNA sequence features that they called "container sites" both *in vivo* and *in vitro*. In this chapter, I first directly compared the distance (up to 100 bp) from nucleosomes *in vitro* to their closest nucleosomes *in vivo* by limiting our focus on the reciprocally closest nucleosome pairs, as similarly done in Kaplan et al. (2008). I observed a mild impact of DNA sequence in the nucleosome positioning, as indicated by a median size of 39 bp in the divergence in translational positioning among ~2.5 million pairs of *in vivo* and *in vitro* nucleosomes. I speculate that the comparison of nucleosome translational positioning between *in vivo* and *in vitro* samples might introduce biases in detecting the effects of DNA sequence on nucleosome positioning, due to the variations in sample handling, sequencing strategies (paired-end for *in vivo* and single-end sequencing for *in vitro* samples respectively), and bioinformatics procedures to infer nucleosome positions *in vivo* and *in vitro*. Nonetheless, I have successfully demonstrated that both DNA sequence features and chromatin states were associated with nucleosome positioning evolution, echoing the claims by Schones et al. (2008) that *trans*-acting factors are more important in regulating nucleosome positioning at *cis*-regulatory genomic sites while DNA sequences are the primary determinants of nucleosome positioning at non-regulatory sites.

To mitigate the biologically irrelevant biases, and control for the confounding effects from *trans*-acting factors, we then compared the translational positioning of nucleosomes between paralogous regions within the same *in vivo* and *in vitro* human samples side by side. By doing so, the technical biases and variations in the broad cell environments, such as the abundances of *trans*-acting binding factors, should be consistently controlled for. These variations may be expected to confound other studies comparing species, replicates or cell types (Tirosh et al. 2010; Hughes et al. 2012; Struhl and Segal 2013).

I showed that nucleosome positioning is generally well conserved between paralogous regions in both datasets. Following the duplication of a region and its insertion into a new genomic location nucleosomes generally reassemble at the same, or similar, locations.

The base-pair resolution of the nucleosome positioning calls in the *in vivo* sample enabled me to explore how nucleosomes relocate. As far as we are aware, it is the first time that the rotational positioning has been directly tested by comparing the translational positioning of

nucleosomes from paralogous regions. Before this study, the rotational positioning has usually been deduced from the ~10 bp periodicity in the dinucleotide frequencies along the nucleosomal DNA (Kaplan et al. 2008; Zhang et al. 2009; Valouev et al. 2011; Gaffney et al. 2012), or in the distribution of the DNase I cleavages and the MNase fragment midpoints with respective to the called nucleosome dyads (Gaffney et al. 2012). I observed that nucleosomes do not simply drift from their location during evolution; but rather particular sizes of shifts are favoured, in particular multiples of around 10 bp. The DNA helix makes one complete turn every approximately 10 bp and matching 10 bp periodicities in AT and GC rich sequences have been observed to underlie nucleosomes in various studies. This has often been attributed to AT/AA/TT dinucleotides being periodically favoured at the minor grooves of the DNA double helix that face histone octamer, such that nucleosomes with this periodicity are stable (Albert et al. 2007; Cui and Zhurkin 2009). This observed preferential shift of 10 base pairs is expected to maintain these periodicities relative to the nucleosome dyad, suggesting DNA structural constraints restrict the evolution of nucleosome positioning.

In addition to the strong rotational preference where nucleosome positioning has diverged, another feature in the nucleosome positioning evolution between paralogous regions is that neighbouring proximal nucleosomes (up to 1 kb) often shift by approximately the same distance and thus evolve as an array both *in vivo* and *in vitro*, supporting a barrier model and the "statistical principle" in nucleosome positioning (Kornberg and Stryer 1988; Mavrich et al. 2008). It has been previous shown that well positioned nucleosomes could act as the barrier but *trans*-acting factors, such as chromatin remodelling complex, are required to created regularly spaced nucleosome arrays (Valouev et al. 2011; Zhang et al. 2011; Gaffney et al. 2012); however, the observed correlation in nucleosome positioning *in vitro* suggests a role for long range effects of intrinsic histone-DNA interaction in nucleosome positioning evolution.

I go on to show that the comparison of duplicons *in vitro* and *in vivo* provide natural controls for the effect of DNA sequence changes and the disruption of DNA binding motifs on nucleosome occupancy. To further support the role of DNA sequence in the nucleosome positioning, features of DNA sequence including DNA composition, divergence in DNA composition and DNA sequence all showed significant links with divergence in nucleosome positioning between paralogous regions in both *in vivo* and *in vitro* samples. Although inter-chromosomal duplications were linked to slightly higher divergences in nucleosome positioning than between intra-chromosomal duplicons, strong conservation in nucleosome positioning was still observed. This suggests that the chromosomal neighbourhood or broader environment a sequence is inserted into plays a more limited role in nucleosome

positioning in humans than the sequence itself. Not surprisingly, the relative importance analysis confirmed that the most important variables in nucleosome positioning between paralogous regions are features of DNA sequence.

Similar to previous observations that nucleosome positioning could be regulated by the interaction of transcription factors with DNA sequence, I also found that the birth or death of certain protein binding motifs in human genome, such as CDX2 and TEAD2, are associated with nucleosome positioning evolution between paralogous regions (Valouev et al. 2011). However, it is not a universal feature of all transcription factors except a subset of proteins called "pioneer transcription factors" (Zaret and Carroll 2011), with the majority of motifs having no observable role in nucleosome repositioning between paralogous regions. Previous studies have shown that when genes get duplicated, one copy is constrained to keep the original function, and the other is free to diverge, acquiring new function or losing function which results in the copy that loses function evolving to pseudogene. One possible mechanism for acquiring or losing functions might be the birth and/or loss of functional cis-regulatory motifs (Force et al. 1999; Stoltzfus 1999; Rastogi and Liberles 2005; Rastogi and Liberles 2005; Bailey and Eichler 2006; Zheng et al. 2007). However, it has been found that both copies of segmental duplicates experience elevated evolution (Kostka et al. 2010; Lorente-Galdos et al. 2013).

Controlling for confounding factors is a common problem when trying to determine the role of DNA sequence or local environment in the evolution of chromatin and epigenetic states. Here I have shown that the analysis of paralogous regions allows for *trans*-factors to be controlled for and can potentially provide a clearer picture of the role of DNA sequence and other factors in nucleosome position evolution. The results presented here and in particular the strong agreement between *in vivo* and *in vitro* datasets, point towards DNA sequence playing a primary role in determining nucleosome positioning evolution.

Chapter 3: Conserved Determinants of Nucleosome Positioning Evolution across Eukaryotes

3.1 Introduction

In Chapter 2, I have compared the nucleosome positioning evolution between paralogous regions in the human genome using *in vivo* and *in vitro* datasets. In this chapter, I aimed to directly contrast the nucleosome positioning between paralogous regions in human and yeast genomes simultaneously and deduce potentially conserved or divergent determinants of nucleosome positioning evolution across eukaryotes, by analysing another two independent *in vivo* human (10 base pair resolution, provided by James Prendergast) and yeast (base pair resolution, Brogaard et al. 2012) nucleosome positions maps. The side by side comparison of the relative nucleosome positioning between paralogous regions in both human and yeast genomes respectively should resolve the issue that is inevitable in the direct inter-species comparison of nucleosome positioning, namely that the levels of *trans*-acting chromatin binding factors are expected to vary widely between cell types, replicates and organisms.

Particularly, I addressed the following questions: 1) To what extent is the positioning between paralogous nucleosome pairs conserved in human and yeast genomes following a duplication event? 2) Does the repositioning of proximal paralogous nucleosome pairs over time follow the statistical positioning principle? 3) To what extent is the underlying DNA sequence important in determining how nucleosome positioning evolves between paralogous regions? 4) Is the local chromosomal environment an important determinant of the positioning between paralogous nucleosomes? 5) To what extent are the determinants of nucleosome positioning conserved across eukaryotes?

3.2 Materials and Methods:

3.2.1 Summary of datasets and software used

3.2.1.1 Inference of paralogous regions in the yeast genome by LASTZ

The list of paralogous segments (sequence identity ≥ 90%; length ≥ 1 kb) was obtained from http://humanparalogy.gs.washington.edu/, as described in Chapter 2.

The list of duplicated regions in the yeast genome was obtained by whole genome self-alignment: the yeast genome sequence (SGD/sacCer2 version) was downloaded from the UCSC genome browser (Kent et al. 2002) and LASTZ was used to align the whole genome sequence to itself to retrieve paralogous regions (Harris 2008). The general workflow of LASTZ includes: 1) build a position table of DNA oligomers from the target sequences (the whole yeast genome) which is loaded into memory; 2) scan each query sequence (each chromosome for example) against the position table to find matches in the target sequence, which are called seeds; 3) infer high-scoring segment pairs (HSPs) from matched seeds and furthered chained into anchors using syntenic information; and 4) individual anchors were further extended by local alignment and back-end filtering to obtain duplicated segments. The power of LASTZ lies in the fact that, instead of requiring parameters from end users, it automatically infers scoring parameters from input sequences. The command used to infer paralogous segments in the yeast genome is: *lastz yeast.fa[multiple] yeast.fa –chain –notrivial format=general:name1,start1,end1,name2,start2,end2,length1,length2,strand1, strand2,identity,text1,text2 > yeast_duplicates.txt.*

Due to much smaller the size of the yeast genome than that of the human genome (Chapter 2), paralogous segments with sequence identity ≥ 80% and length ≥ 200 bp were selected to obtain enough data for meaningful analysis (Harris 2008). I identified a total of 459 pairs of paralogous regions in the yeast genome, with a median average size of 746 bp. The size distribution is plotted in Figure 3.1.

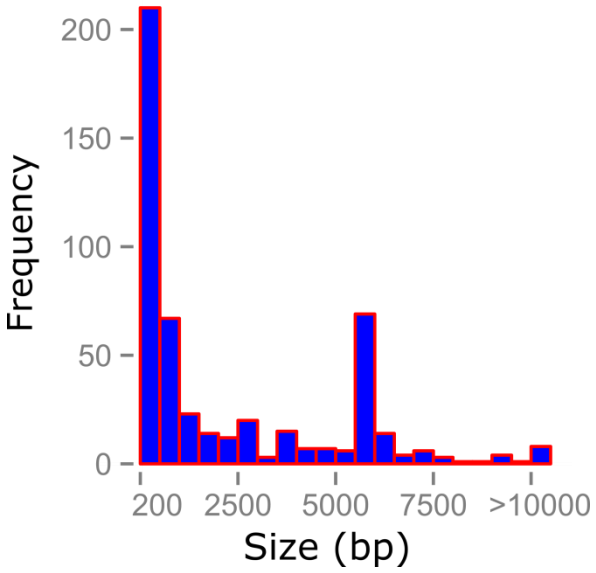


Figure 3.1. Size distributions of average paralogous duplicons in the human genome.

3.2.1.2 Nucleosome positioning map in the yeast and human genomes

The nucleosome positioning map in the yeast genome was obtained from (Brogaard et al. 2012). In contrast to the progressive digestion of Linker DNA towards nucleosome cores by MNase, the authors used a chemical approach in which the cleavage selectively occurs at -1 and +6 positions relative to the centre of engineered nucleosomes which carry a cytosine at serine residue of H4 core histone (H4S47C). Thus positions of nucleosomes by this method can be determined with base pair accuracy. DNA fragments with ends marking centres of neighbouring nucleosomes were selected and sequenced from six replicate experiments (four single-end and two paired-end replicates). The list of nucleosomes used in this analysis was based on the unique map in which 67,543 nucleosomes were called from the combined dataset of six replicate experiments such that the maximum overlap of two neighbouring nucleosomes is 40 base pairs, corresponding to a centre-to-centre distance of at least 107 bp.

The *in vivo* human nucleosome positions data were provided by James Prendergast (bench supervisor). Briefly reads used to infer nucleosome positions were obtained from histone modification data in human H1 cells that were collated from a combination of both the ENCODE (T.E.P. Consortium 2012) and NIH Epi-genomics Roadmap (Bernstein et al. 2010) projects. Reads were mapped with Bowtie allowing no mismatches and only a single best hit (Langmead and Salzberg 2012). Mapped reads less than 35 bp long were also

discarded. Nucleosome positions were determined using the NPS software by default parameters (Zhang et al. 2008). The reasons we used another independent *in vivo* human nucleosome positions dataset in this chapter are that we would like to use another dataset as a validation for what we observed in Chapter 2, and we would like sample sizes to be approximately balanced for yeast and human paralogous nucleosome pairs (we retrieved 4,479 and 4,462 paralogous nucleosome pairs in the yeast and human genomes respectively).

3.2.2 Analysis procedure

Linker regions have been estimated at approximately 20bp in yeast (Brogaard et al. 2012), and so 20bp at each side of the core regions was annotated as linker region in yeast dataset. In addition, we set the maximum shift in yeast as 80bp, as done in Tirosh et al. (2010). The length of linker region and the maximum shift in humans were defined as described in Chapter 2. Unless stated explicitly, the analytic processes were similar to that described in Chapter 2, including defining paralogous nucleosome, permutation analysis to detect the correlation in positioning between proximal paralogous nucleosome pairs, and multiple linear regression and relative importance test.

3.3 Results

3.3.1 Nucleosome positioning is well conserved following yeast and human duplications

I retrieved 4,462 and 4,479 paralogous nucleosome pairs in the human and yeast genomes respectively. As shown in Figure 3.2, 33% (1,496 pairs) of paralogous pairs of nucleosomes displayed no evidence of changing position at all following a duplication event in yeast; whereas in human 25% (1,121 pairs) of paralogous pairs of nucleosomes were shifted by less than 10 bp, the approximate resolution of the algorithm used to call nucleosome positioning (Zhang et al. 2008). Beyond this there are heavy biases in both species to shifts of less than 40 bp, such that the cores of paralogous nucleosome pairs occupy broadly equivalent positions. I conclude that following the duplication of a region, nucleosomes in general assemble at similar locations on the original and duplicated copies in both yeast and human genomes, suggesting an important relative contribution of local DNA sequence to nucleosome positioning in both yeast and human genomes.

3.3.2 Yeast nucleosome positioning divergence shows strong periodicity

Similarly to the *in vivo* human nucleosome dataset in Chapter 2, the unusually high, base-pair resolution of the yeast nucleosome dataset used in this study (Brogaard et al. 2012) allowed me to discover periodicities in shifts following a duplication event, indicating that particular sizes of shifts are favoured. As shown in Figure 3.2C, following a segmental duplication event yeast nucleosomes were observed to preferentially shift by a multiple of complete helical turns as observed previously in the human *in vivo* analysis (Figure 2.10B). The strong rotational preference in translational positioning between paralogous regions in both human (Figure 2.10B) and yeast (Figure 3.2C) *in vivo* datasets of base pair resolution suggests that this is a fundamental principle of nucleosome positioning evolution across eukaryotes.

Nucleosomes were not only observed to preferentially shift a multiple of 10 bp however, intriguingly an additional preference for shifts of 6 bp was also observed (Figure 3.2C). This corresponds to the known phase shift between the maxima of AA and TT dinucleotides at nucleosome core regions (Ioshikhes et al. 1996) and suggests a more complex interplay between underlying dinucleotide frequencies and nucleosome positioning evolution.

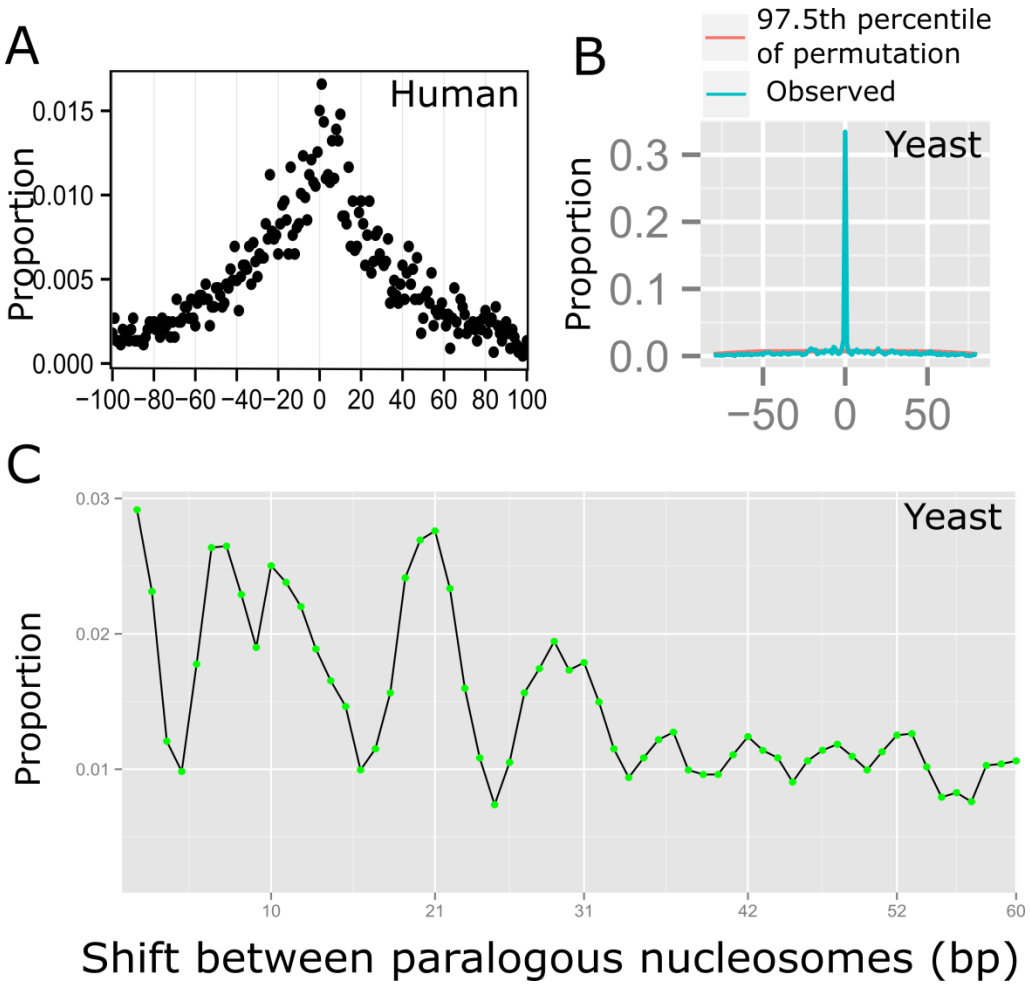


Figure 3.2. Nucleosome positioning is generally conserved between paralogous regions in human and yeast. (A) and (B) Distribution of signed differences in positions between paralogous nucleosomes in human and yeast datasets respectively. Negative numbers in the X axis correspond to a shift 5 prime relative to the reference strand. Permutation was also carried out in yeast species (B) to assess whether there was conservation in positioning between paralogous nucleosomes. Blue lines are observed shifts between paralogous nucleosomes while red lines are permutated shifts (based on 1000 permutations) assuming independence between the positions of nucleosomes between duplicons. (C) Observed periodicity in nucleosome positioning divergence of ~10 bp in yeast. The corresponding distances of 1 to 5 helical turns of 10, 21, 31, 42, and 52 are indicated, based on the assumption that a complete helical turn is approximately 10.4 bp.

3.3.3 Long range correlations in nucleosome positioning evolution in human and yeast genomes

To decide whether the role of "statistical principle" in nucleosome positioning between paralogous regions was evolutionarily conserved, I investigated the correlations seen in shifts among proximal paralogous nucleosome pairs within a maximum distance of 1000 bp

in both human and yeast species by comparing correlations observed against that generated from permutations, which was carried out similarly as described in Figure 2.11A. I found a significant enrichment of nucleosomes within close proximity that display approximately the same size shifts relative to their corresponding paralogous nucleosome partners in both the yeast and human genomes (Figure 3.3). Thus I concluded that nucleosome positioning by the constraints from neighbouring nucleosomes is a general mechanism that is conserved across eukaryotes.

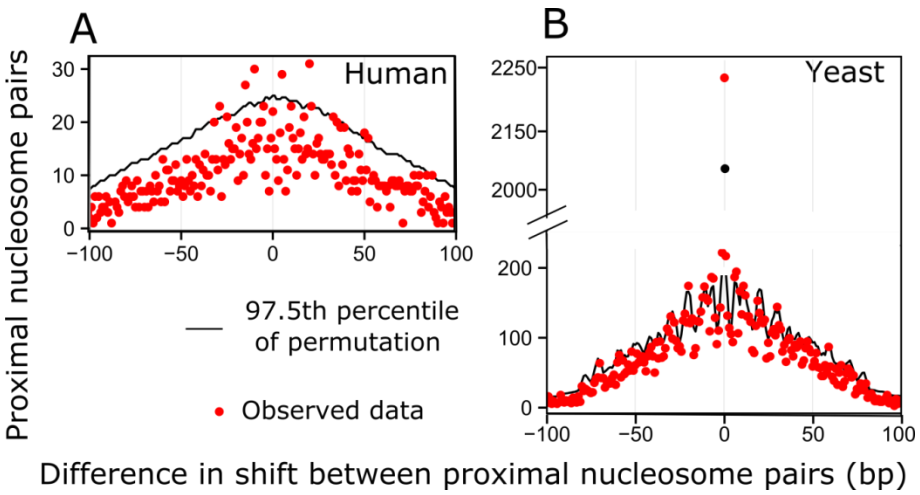


Figure 3.3. Shifts among proximal paralogous nucleosome pairs are correlated. Detailed schematic description on the permutation procedure could be found in Figure 2.11A. (A) Human dataset. (B) Yeast dataset.

3.3.4 The local DNA composition bias is associated with nucleosome positioning evolution

The "container site" mechanism for nucleosome translational positioning has been found and only was directly tested in human cells; studies in *Schizosaccharomyces pombe* have found different rules for DNA sequence in nucleosome positioning and nucleosome cores are actually coincident with AT rich sequences, directly contradicting with that found in human nucleosomes which feature GC rich sequences (Lantermann et al. 2010; Valouev et al. 2011; Gaffney et al. 2012; Moyle-Heyrman et al. 2013**)**. In addition, a recent study from Xing and He (2015) suggested that the GC content bias in core in *Saccharomyces cerevisiae* is a by-product of mutation bias.

To extend our analyses of the impact of DNA composition to yeast species, while controlling for *trans*-acting factors, I investigated how various sequence features of duplicated regions are related to nucleosome positioning stability in both human and yeast genomes. I compared the GC content in core and AT content in linker regions between nucleosomes showing conserved positions between paralogous regions with those showing larger relative shifts. To balance sample sizes, I divided the nucleosomes in both human and yeast datasets into three, approximately equal sized groups and tested the differences among the three groups by Kruskall-Wallis test. The three groups in yeast being: 1) Shift: 0 bp (1496 nucleosome pairs); 2) Shift: 1-27 bp (1513 pairs); 3) Shift: 28-80 bp (1470 pairs) and in the human dataset: 1) Shift: 0-15 bp (1499 pairs); 2) Shift: 16-40 bp (1493 pairs); 3) Shift: 41-100 bp (1470 pairs). To address the potential biases introduced by the arbitrary grouping of unsigned shift sizes between paralogous nucleosomes, I also performed Spearman correlation test between unsigned shift and DNA sequence composition. As shown in Figure 3.4A and B, the highest GC content in core regions is associated with nucleosomes of the most conserved positioning between paralogous regions in both human and yeast, though the difference does not reach significance in humans (human, $p = 0.673$; yeast, $p = 1.53e-11$), which is supported by the Spearman correlation test (human, rho=-0.01, p=0.34; yeast, rho=-0.11, p=3.72e-13). The observation is supported by further evidence in yeast. Firstly, paralogous nucleosomes with identical positioning are generally more stable, indicated by the nucleosome centre positioning score to noise ratio (Pearson's Chi-squared, p-value < 2.2e-16; Brogaard et al. 2012). In addition, I compared the average binding and occupancy scores across nucleosome cores based on the model proposed by Kaplan et al. (2008), and in Figure 3.4E and F observed that paralogous nucleosomes with identical positioning are associated with the highest average binding and occupancy scores (Kruskal-Wallis test: binding score, $p < 2.2e-16$; occupancy score, p=3.386e-16. Spearman correlation test: binding score, rho= -0.10, p=3.28e-12; occupancy score, rho=-0.10, p=1.04e-10). This suggests that the relative positioning between paralogous nucleosomes is associated with the stability of nucleosome formation, thus confirming that the compositional bias plays an important role in nucleosome positioning.

However, AT content in linker shows an intriguing contrast between species. I observed the expected bias in higher AT content in linkers for nucleosomes of the more conserved positioning between paralogous regions in human data but the opposite in yeast, a significant decline in AT content in linker for nucleosomes with conserved positioning (Figure 3.4C and D. Kruskal-Wallis test: human, p=0.044; yeast, p=7.863e-09. Spearman corelation test: human, rho=-0.10, p=3.28e-12; yeast, rho=-0.10, p=1.04e-10.). A plausible explanation would be that this

may be due to the differences in linker histone biology between species. Firstly, the extremely low abundance of linker histone Hho1p in yeast (1 molecule per 4 ~ 40 nucleosomes) suggests that the role of linker region realised through linker histone is quite limited in nucleosome positioning (Bates and Thomas 1981; Freidkin and Katcoff 2001). Secondly, the effect of Hho1p is directed by the contact of "wing" domains with TT dimers situated at ±75 and ±76 bp relative to the dyad and bends the linker DNA around the histone core (Cui and Zhurkin 2009). I thus assessed the relative enrichment of the TT dimers at ±75 and ±76 bp relative to the dyad and found that nucleosomes displaying no shift between paralogous nucleosomes (Shift: 0 bp) are generally more enriched with TT dimers at these locations (Pearson's Chi-squared Test, p=4.6e-4).

Collectively this highlights that more conserved positioning between paralogous nucleosomes is associated with stronger local DNA compositional biases in both human and yeast genomes.

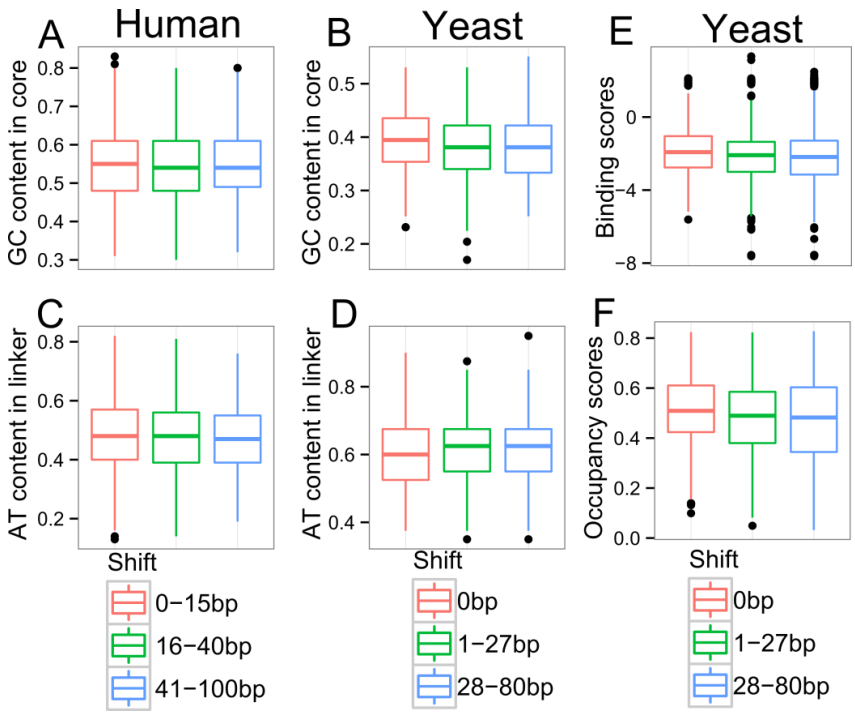


Figure 3.4. Nucleosome positioning divergence is associated with underlying sequence composition. The distribution of GC content in core and AT content in linker were compared among three groups of nucleosome with different degrees of divergence in positioning between paralogous nucleosomes in human (A and C) and yeast (B and D). Nucleosome binding and occupancy scores at each site of nucleosome cores were predicted and averaged. The distribution of average scores for individual nucleosomes in yeast was then compared between the three groups (E and F). P values were obtained using the Kruskal-Wallis rank sum test as appropriate for each plot shown (A: p = 0.673; B: p = 1.53e-11; C: p = 0.044; D: p = 7.863e-09; E: p < 2.2e-16; F: p =

3.386e-16). In addition, the correlation coefficients (rho) and p values from Spearman correlation tests between positioning divergence (unsigned shift) and DNA sequence composition are: A, rho=-0.01, p=0.34; B, rho=-0.11, p=3.72e-13; C, rho=-0.04, p= 0.007; D, rho=0.09, p=9.30e-10; E, rho= -0.10, p=3.28e-12; F, rho=-0.10, p=1.04e-10.p=1.04e-10.

3.3.5 Sequence divergence and the nucleosome repositioning

In addition to the role of DNA composition, I also examined whether DNA sequence divergence is associated with nucleosome repositioning in both human and yeast genomes. Overall the relative positioning of paralogous nucleosomes (unsigned distance between paralogous nucleosomes) is correlated with total sequence divergence (nucleosome core plus linker regions; Spearman's rho = 0.10, p = 1.94e-10 in humans; rho = 0.17, p < 2.2e-16 in yeast). In addition, to assess whether different choices in the definition of reference dyads have an impact on the interpretation, I compared the divergence at the DNA sequence level between paralogous regions showing different degrees of nucleosome positioning conservation in more detail using two approaches. In the first approach, the reference dyad locations were defined as the dyad of a randomly selected nucleosome from each paralogous pair, and in the second the mid-point between dyad locations was taken as the reference dyad in both yeast and human datasets. The results obtained from these two approaches were largely consistent, and one of the possible reasons might be that the majority of nucleosomes have shifted relatively small distances. Sequence substitutions were separated into two groups: AT <=> TA or GC <=> CG, i.e. substitutions that do not alter DNA composition, and AT <=> GC changes that do alter sequence composition between paralogous regions (Figure 3.5).

In yeast the number of both types of change was observed to be lower at pairs of nucleosomes conserved in their positioning. However in humans only the number of changes that altered DNA composition (AT <=> GC) was significantly higher across nucleosome cores at divergently positioned human nucleosomes (left two panels in Figure 3.5). This suggests that, at least in humans, only mutations that alter DNA composition are specifically linked to nucleosome positioning divergence. These results provide evidence of a link between DNA sequence divergence and nucleosome repositioning after eukaryotic duplication events.

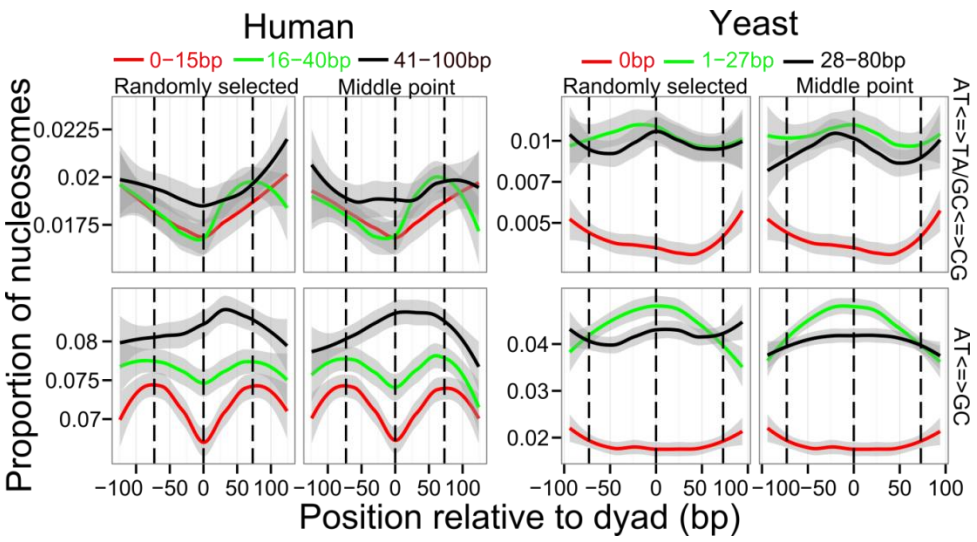


Figure 3.5. Sequence divergence is associated with divergence in nucleosome positioning between paralogous regions in human and yeast. Panels are split into AT <=> GC sequence changes that affect local DNA composition and AT<=>TA/CG<=>GC changes that do not. In the 'Randomly selected' analysis, the reference dyad is set to the dyad of a randomly selected nucleosome in each paralogous pair. In the 'Middle point' analysis the reference dyad is set to the mid-point between paralogous nucleosomes. The means and associated 95% confidence intervals are displayed as red, green and black lines and associated grey regions.

3.3.6 The local genomic environment affects positioning between paralogous nucleosomes

Though DNA composition and sequence divergence appear to be key determinants of nucleosome repositioning, the different local chromosomal environments that duplicons occupy may also play important roles in divergence in positioning between paralogous nucleosomes. For example, the divergence in positioning between paralogous nucleosomes in duplications which have occurred between different chromosomes (inter-chromosome) is significantly higher than those on the same chromosome (intra-chromosome) in both human and yeast (Figure 3.6A and B). Although inter-chromosomal duplications are generally older and more divergent at the sequence level (Kimura 1983), changes in nucleosome positioning remain larger between human inter-chromosomal duplicons having accounted for any sequence divergence affects (Spearman partial correlation test, human adjusted rho = 0.07, p=1.9e-7; yeast adjusted rho = 0.03, p = 0.06).

For most eukaryotic genes, nucleosomes are well positioned around the TSSs with the strength of nucleosome positioning decaying with increasing distance from the TSSs

(Schones et al. 2008; Valouev et al. 2011; Hughes et al. 2012). This suggests that the TSS acts as an influential local chromosomal environment and could potentially affect the positioning of nucleosomes in opposition to their broader sequence preferences. To explore this I examined the positioning of paralogous nucleosomes relative to their proximity to the closest TSS. I divided nucleosome pairs into three categories as follows. 1) Both nucleosomes being within 200 bp of their closest TSSs (Both proximal). 2) Both nucleosomes at least 1000 bp from their closest TSS (Neither proximal). 3) One nucleosome within 200 bp of a TSS and the other at least 1000 bp away (One proximal). As one might expect, the positioning of paralogous nucleosomes in which both are proximal to and under the strong effect of their TSSs (Both proximal) are the most conserved in both human and yeast datasets (Figure 3.6C and D). The positioning of paralogous nucleosomes in which only one is under the strong influence of a TSS is most divergent in yeast, which could reflect sub-functionalisation of TSS-associated nucleosomes following duplication. However, in humans the average divergence of positioning between paralogous nucleosomes was similar whether only one nucleosome is proximal to a TSS or neither. Again, as expected given the functional constraints of occupying an exon, nucleosome positioning was also observed to be significantly more conserved following a duplication event where both nucleosomes were in exons in both yeast and humans (Figure 3.6E and F).

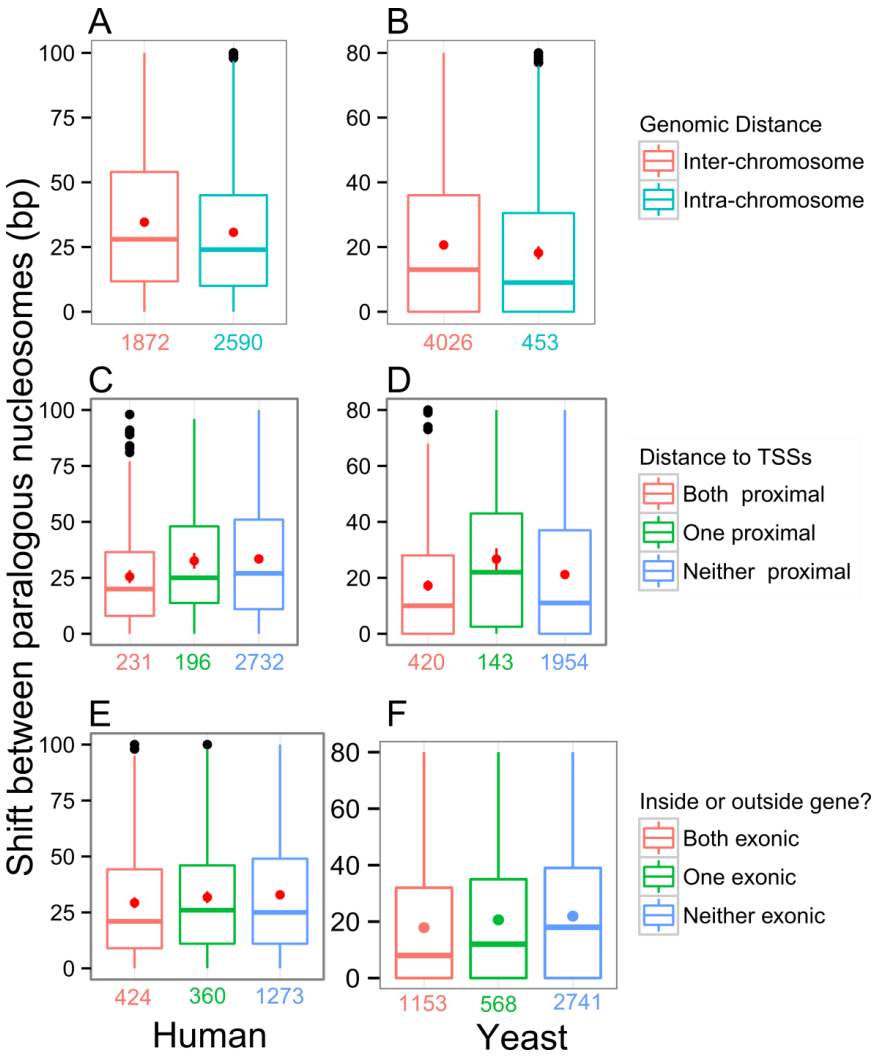


Figure 3.6. Local chromosomal environments affect divergence in positioning between paralogous nucleosomes. Paralogous nucleosomes residing on the same chromosome are compared with nucleosomes located on different chromosomes in human (A) and yeast (B). Comparison of paralogous nucleosome pairs (C: human; D: yeast) where both nucleosomes are within 200 bp of a TSS (Both proximal), where one nucleosome is within 200 bp of a TSS but the other is at least 1000 bp away (One proximal), and where both nucleosomes are at least 1000 bp away from their nearest TSS (Neither proximal). The effect on positioning is also compared (E: human; F: yeast) for pairs where both nucleosomes are in exons (Both exonic), where one nucleosome is in an exon but the other is intergenic (One exonic), and where both nucleosomes are in intergenic regions (Neither exonic). In each case the mean shift between paralogous nucleosomes is plotted in each boxplot in red (mean ± 95% confidence interval). The numbers at the bottom of each boxplot reflect the numbers of paralogous nucleosome pairs in each group. P values were obtained using the Mann-Whitney test (two groups) and Kruskal-Wallis rank sum test (more than two groups) as appropriate for each plot shown (A: p = 3.3e-06; B: p = 0.015; C: p = 8.6e-05; D: p = 6.08e-05; E: p = 0.03; F: p = 4.916e-09).

3.3.7 Integrative analysis of paralogous nucleosome divergence covariates by multiple linear regression and relative importance test

Finally, to disentangle the relative contributions of several variables that might affect each other to nucleosome repositioning in the human and yeast lineages, I included 7 distinct variables in multiple linear regression models and analysed the normalized relative importance of each independent variable using the "lmg" method from the "relaimpo" R package (Table 3.1,Table 3.2; Lindeman and Merenda 1980; Ulrike 2006). Despite being separated by up to 1 billion years of evolution the features controlling the divergence in nucleosome positioning between paralogous regions were observed to be generally correlated between human and yeast. Not only was there a large overlap in the factors significantly linked to nucleosome position divergence in both species, but also the size and direction of their coefficients were generally similar, suggesting they have a similar impact on nucleosome positioning evolution in these two substantially different species. For example the AT<=>GC base changes that alter the DNA composition was observed to be significantly associated with the divergence in nucleosome positioning between paralogous regions in both species when accounting for these other factors, with 10 such base changes (in nucleosome core plus linker regions) associated with a 3bp larger shift on average in both humans and yeast (Table 3.1). The relative importance analysis revealed that such base changes were ranked the most important factor contributing to nucleosome repositioning in both species with the distance to TSS also ranked highly in both human and yeast (Table 3.2).

However, one factor was found to be inconsistent between species, the relative contribution of AT content in linker regions. A general higher AT content in linker regions is linked to a significantly decreased probability of the corresponding nucleosome shifting following a segmental duplication event in human but an increased probability of shifting in yeast. Examining the correlation between nucleosome shifts and linker AT content in isolation from other factors captures this observed difference between species. Whereas on average larger AT linker contents are associated with greater nucleosome shifts in yeast, nucleosomes in regions of intermediate AT content show the largest shifts in humans (Figure 3.7). A plausible explanation would be that the different sequence characteristics at these regions are linked to the known divergence in structure and distribution of the linker histone protein between these eukaryotes. Although the other core histone proteins are comparatively well conserved between these species the linker histone protein has shown an elevated rate of divergence and is not found associated with all nucleosomes in yeast unlike

in humans (Bates and Thomas 1981; Freidkin and Katcoff 2001; Downs et al. 2003; Cui and Zhurkin 2009; Osmotherly 2010).

Table 3.1. Relative effects of DNA local sequences and chromosomal environments on positioning between paralogous nucleosomes according to optimised multiple linear regression models.

	Human		**Yeast**	
	Coefficient	**p value**	**Coefficient**	**p value**
Intra-chromosomal duplication[1]	-3.4	2.6e-5	-1.9	0.08
One proximal to TSS[2]	7.4	2.8e-3	8	2.7e-4
Neither proximal to TSS[2]	10.2	5.1e-8	4.5	1.8e-4
Average duplication length (kb)	-0.04	0.02	-0.22	2.0e-3
GC content in core (percentage)	-0.289	1.5e-5	-0.259	2.3e-5
AT content in linker (percentage)	-0.357	7.3e-12	0.099	0.02
AT <=> TA or GC <=> CG (number of changes)[3]	0.0	0.95	-0.4	0.06
AT <=> GC (number of changes)[4]	0.3	1.9e-8	0.3	7.2e-5

[1]The reference level for the Intra-chromosomal duplication is the Inter-chromosomal duplication
[2]The reference level in the distance to TSS is the Both proximal to TSS
[3]AT <=> TA or GC <=> CG: i.e. sequence divergence that does not change the local DNA composition
[4]AT <=> GC: i.e. sequence divergence that changes the local DNA composition

Table 3.2: Relative importance analysis of DNA local sequences and chromosomal environments on positioning between paralogous nucleosomes.

	Human	**Yeast**
AT <=> GC	0.26	0.27
Distance to TSSs	0.20	0.17
AT content in linker region	0.19	0.12
Genomic distance	0.15	0.02
Average duplication length	0.11	0.10
GC content in core region	0.07	0.22
AT <=> TA or GC <=> CG	0.03	0.11

Note: the analysis has been done by the "lmg" method in R package relaimpo (Lindeman and Merenda 1980; Ulrike 2006)

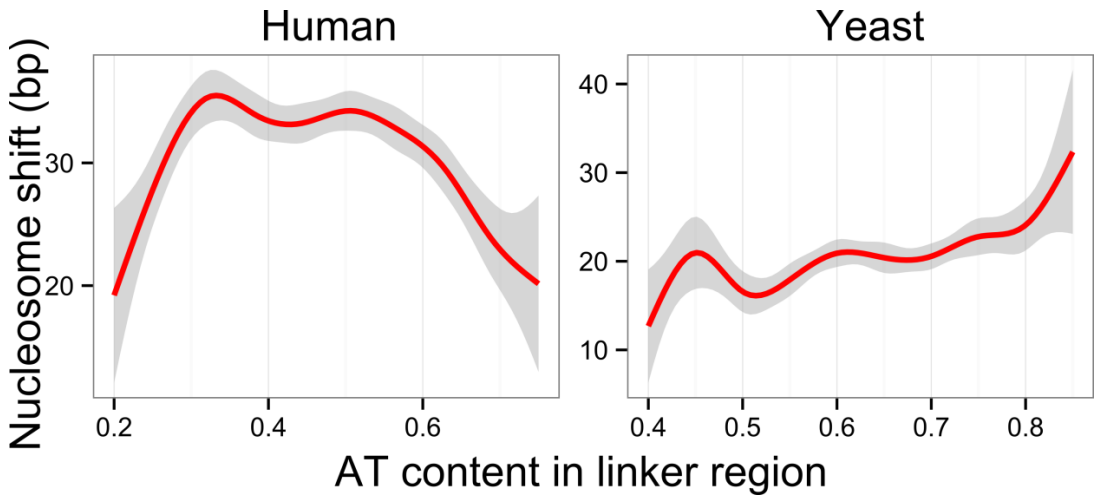


Figure 3.7. The effects of linker region AT content on nucleosome repositioning in human and yeast genomes. The mean and 95% confidence interval of nucleosome shifts are displayed as red lines and grey regions after applying loess smoothing. AT content was rounded to the nearest percentage and percentage values with less than 10 regions were excluded.

3.4 Discussion

In this chapter, I show that nucleosome positioning is generally well conserved between paralogous regions in both human and yeast genomes. I observed two features in nucleosome positioning evolution between paralogous regions. Firstly, nucleosome positioning by the constraints from neighbouring nucleosomes is a general mechanism that is conserved across eukaryotes. Secondly, the strong rotational preference in translational positioning appears to not only be the case in humans (Chapter 2) but also in yeast, suggesting that this is a fundamental principle of nucleosome positioning evolution between paralogous regions across eukaryotes (Hughes and Rando 2014). Intriguingly a peak at around 6 bp was also observed in nucleosome shifts in yeast, which corresponds to the observed phase shift between the AA and TT dinucleotide 10 bp periodicities underlying nucleosomes (Ioshikhes et al. 1996). However no similar peaks were observed at other multiples of 6 bp suggesting this preferential shift of 6 bp may be unrelated to such underlying, periodic patterns.

In both yeast and humans, a greater level of sequence divergence between paralogous regions was associated with larger changes in the locations of corresponding nucleosomes. However separating changes into those that change GC content and those that do not highlighted that the number of changes that affect base composition (AT<=>GC) are linked to nucleosome repositioning when controlling for other factors, and was found to be the most important factor of those analysed. The total number of AT<=>TA or GC<=>CG changes was uncorrelated to the size of nucleosome shifts, highlighting that it is not simply higher mutation rates, indicative of a longer time since duplication, that is linked to nucleosome repositioning. This suggests certain sets of nucleosomes are more primed to evolve following a duplication event due to being less constrained by their underlying sequence patterns. The nucleosome positioning was conserved between paralogous regions irrespective of whether the duplicons were in close proximity or on different chromosomes, though nucleosome positioning was marginally less well conserved between inter-chromosomal than intra-chromosomal duplicons in humans. This was the case even having controlled for sequence divergence, suggesting that the chromosomal neighbourhood or broader environment a sequence is inserted into plays a limited role in nucleosome positioning in humans, confirming what we found in Chapter 2.

In spite of these species being separated by approximately 1 billion years of evolution, accounting for all other factors the contribution of individual factors to nucleosome positioning, as measured by their coefficients, was surprisingly well conserved between yeast and humans. This is especially surprising given previous reports that distinct

nucleosome positioning mechanisms have evolved between yeast species (Lantermann et al. 2010). Despite the high level of conservation of core histone proteins, across eukaryotes linker histones have evolved comparatively rapidly (Cui and Zhurkin 2009; Osmotherly 2010). Thus, the observed difference in the association of linker AT content with the positioning between paralogous nucleosomes might be a reflection of the evolution of linker histones and DNA sequence features between human and yeast species. In multicellular, eukaryotic organisms, the abundance of linker histone is approximately 1 molecule per nucleosome; however the abundance of linker histone Hho1p in yeast is about 1 molecule per 4 ~ 40 nucleosomes (Bates and Thomas 1981; Freidkin and Katcoff 2001; Downs et al. 2003; Cui and Zhurkin 2009; Osmotherly 2010). The low abundance of the yeast linker histone Hho1p would suggest the stabilisation of nucleosomes by the linker histone is restricted to a subset of the total pool of nucleosomes. In addition, while the nucleosome core is extremely conserved the linker region is much shorter in yeast than human (~ 20 bp in yeast but ~ 50 bp in human), potentially a reflection of the linker region evolving to be more important in human than yeast, as supported by the relative importance analysis.

Once again in this chapter, I have shown that the analysis of paralogous regions allows for *trans* factors to be controlled for and can potentially provide a clearer picture of the role of DNA sequence and other factors in nucleosome position evolution.

Chapter 4: Nucleosome Positioning Dynamics and Interplay with Mutational Spectra in Disease

4.1 Introduction

In Chapter 2 and 3, I have investigated the nucleosome positioning dynamics in evolution. In this chapter, I focused on directly comparing nucleosome positioning between multiple cancer and non-cancerous cell lines. As discussed in Chapter1, nucleosome positioning has been linked to the mutational spectra and several clues suggest that the positioning of nucleosomes and the patterns of histone modifications or variants they carry may differ between cancer and normal cell lines. Firstly, various histone variants (e.g. H2.A.Z) and modifications (e.g. H3K4me3) are enriched at promoter regions and promoter activities are globally altered in cancer (Barski et al. 2007; Zhang et al. 2008; Jin et al. 2009a). Secondly, aberrant DNA methylation, histone modification, and nucleosome positioning are intertwined in cancers, leading to the dysregulation of many genes including oncogenes and tumour suppressor genes and may also cause genome instability (Portela and Esteller 2010; Brait and Sidransky 2011). DNA methylation has been shown to increase nucleosome stability and be associated with nucleosome positioning and occupancy (Lin et al. 2007; Chodavarapu et al. 2010; Collings et al. 2013). Thirdly mutations in chromatin remodelling complexes including SWI/SNF have also been observed to be involved in cancer (Fraga et al. 2005; Esteller 2007; Wilson and Roberts 2011).

Previous work from different groups have shown that sequence divergence is linked with nucleosome positioning (Chen et al. 2012; Xing and He 2015; Prendergast and Semple 2011), suggesting that nucleosome positioning might be a potential factor to affect mutational spectra. For example, Prendergast and Semple (2011) have shown that both mutation rates and patterns of selection observed in the human lineage are correlated with nucleosome positioning. Furthermore the direction and strength of selection observed was predicted to maintain the optimal variation in local GC content for nucleosome positioning. Firstly to detect whether the alterations in nucleosome positioning between cell types is a

possible contributor to the difference in the mutation spectra between germline and somatic mutations, I compared the patterns of nucleosome positioning in human cancer lines.I have also compared the difference between somatic and germline mutational spectra and their compositional biases. I was especially interested in the following questions:

1. Are the positioning at TSSs and phasing of nucleosomes different or generally conserved among different cell lines?

2. How do somatic and germline mutational spectra differ between classes of genomic annotations and chromatin states?

3. Are the rates of particular types of mutation, for example transitions or transversions, different among germline and somatic mutations?

4.2 Methods and materials

4.2.1 ENCODE source data to call nucleosome positions

Nucleosome positioning data from 11 cell lines (Table 4.1) derived using ChIP-Seq were all downloaded from the Encyclopaedia of DNA Elements (ENCODE) Project. The ENCODE project was established in 2003 and aims to identify and annotate the cell line specific functional elements in the human genome, including but not limited to RNA transcripts, transcription factor binding sites, and chromatin structures that are both three-dimensional, based on the chromatin interaction capturing techniques like Hi-C, and one-dimensional, based on histone modification data by ChIP-Seq protocol (Consortium 2004; Ecker et al. 2012; Dekker et al. 2013; Kellis et al. 2014). I analysed and compared the positioning and phasing of three types of nucleosomes (Valouev et al. 2011). They were: 1) nucleosomes carrying a specific histone modification from a range of 11 types (Table 4.2), 2) all nucleosomes carrying any examined modification, and 3) bulk nucleosomes not selected for any specific histone modification in cell lines GM12878 and K562.

4.2.2 Reads mapping and nucleosome positions calling

Reads, either from the two replicates (Replicate number 1 and 2) of each histone modification in each cell line or derived from bulk nucleosomes in cell lines GM12878 (replicate number 8) and K562 (replicate number 4), were mapped to the current human genome assembly (hg19, NCBI37) using Bowtie2 (Langmead et al. 2009) with default parameters (See workflow in Figure 4.1). Nucleosome dyad positions were estimated from positioned nucleosomes derived using NPS (Zhang et al. 2008) by taking the midpoint of called nucleosomes as in the method of (Reynolds et al. 2010). Each core detected by NPS is given a p-value which is defined as the probability of observing that number of reads by chance, given the current dataset of reads and reference genome (Zhang et al. 2008). Only nucleosome cores predicted by NPS with p value not greater than a conservative threshold of $1x10^{-5}$ were accepted. The numbers of called nucleosomes carrying a specific histone modification and all nucleosomes carrying any histone modification is summarized in Table 4.2. For bulk nucleosomes not selected for any specific histone modification, we called 179,783 and 224,275 nucleosomes in GM12878 and K562 respectively.

Table 4.1. The list of cell lines used for detecting nucleosome positioning

Cell line	Description
Hela-S3	Immortalized cell line; cervical cancer
HepG2	Liver carcinoma
K562	Immortalized cell line; chronic myelogenous leukemia
Dnd41	T cell leukemia with Notch mutation
HUVEC	Human umbilical vein endothelial cells
GM12878	Lymphoblastoid cell line
H1-hESC	Embryonic stem cell
HMEC	Mammary epithelial cells
HSMM	Skeletal muscle myoblasts
NHEK	Epidermal keratinocytes
NHLF	Lung fibroblasts

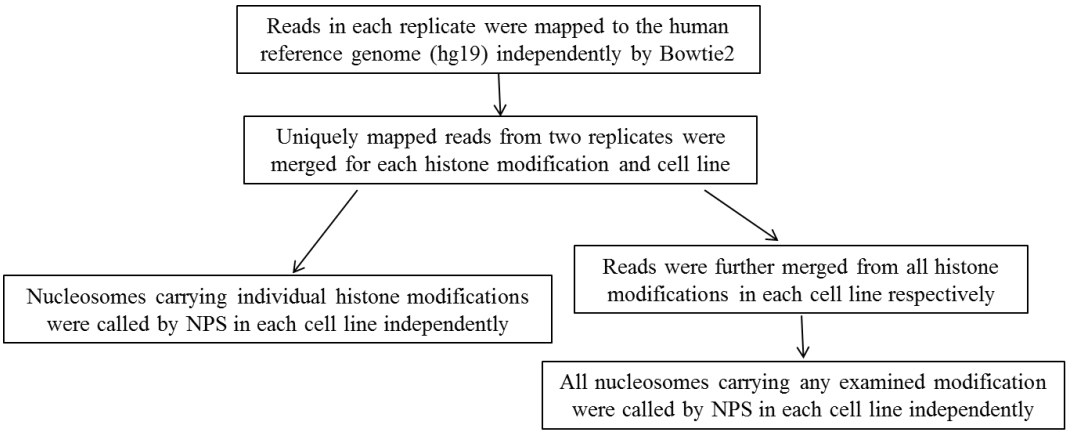


Figure 4.1. Workflow to derive positions of nucleosomes carrying a specific histone modification and all nucleosomes carrying any examined histone modifications.

Table 4.2. Summary of number of nucleosomes called from histone modifications and variants data in this study.

	H2A.Z	**H3K9ac**	**H3K9me3**	**H3K27ac**	**H3K27me3**	**H3K36me3**	**H3K4me1**	**H3K4me2**	**H3K4me3**	**H3K79me2**	**H4k20me1**	**Any mark**
Hela-S3	98490	106088	29535	152399	49847	114389	244838	194056	127712	252456	84814	346793
HepG2	77918	137471	NA	139850	102621	102583	471965	285286	163771	298446	59529	510912
K562	251118	141543	33272	149685	60063	93091	226150	191195	150696	271351	42364	437751
Dnd41	123799	169173	98574	213895	NA	425016	302866	200360	138951	394592	152557	486488
HUVEC	225098	125004	NA	189566	80659	52750	294139	209961	113637	427365	10771	501188
GM12878	159392	115199	31399	169387	8698	71209	177360	225491	151106	309531	6136	411645
H1-hESC	110318	65386	85385	48585	83188	49247	104573	198450	107143	296984	25645	341312
HMEC	197494	98930	35135	178828	20508	53714	297577	252510	120435	457641	9091	508627
HSMM	171338	112939	29673	187184	21099	140546	165159	237004	116769	364421	17723	473599
NHEK	239323	125916	77122	191153	54999	41278	257842	235285	123242	492364	12611	537944
NHLF	240826	80436	32942	NA	36314	88555	141116	188174	117787	468371	5132	472510

"NA": the specific histone modification data was not available in a particular cell line
"Any mark": all nucleosomes carrying any histone modification and variant examined

4.2.3 TCGA dataset and overview of variant calling

Exome sequencing data were retrieved from The Cancer Genomic Atlas (TCGA) project by Alison Meynert. TCGA (http://cancergenome.nih.gov/) provides comprehensive exome sequencing data for tumour and matched normal tissues from the same patient. Exome sequencing targets the protein-coding portion of the genome (Teer and Mullikin 2010). Data for 997 patients covering 17 cancer types (Table 4.3) was remapped and variants (single nucleotide and insertions/deletions) were called using Genome Analysis Toolkit (GATK, McKenna et al. 2010; DePristo et al. 2011) and stored in a local MyQSL database by Alison Meynert. Briefly, pre-aligned BAM files were downloaded from TCGA, sequencing reads were extracted into FASTQ format by Picard (version 1.43, http://broadinstitute.github.io/picard/), and subsequently aligned to the reference human genome (hg19) using BWA pre-alignment (version 0.5.9, Li and Durbin 2009) followed by the Stampy read mapper (version 1.0.12, Lunter and Goodson 2011). Duplicated reads were marked by Picard (version 1.43).

4.2.4 Simultaneous per-patient variant calling and mutation rate calculation

Genotypes in both normal and tumour tissues were determined simultaneously in each patient. This approach calls all single nucleotide variants in tumour-normal sample pairs for each patient by the joint calling mode by the unified genotyper of GATK and was stored in a MySQL database. This was done by Alison Meynert. I limited the analysis to positions that were covered by at least 10 reads in both matched tumour and normal tissue. The coverage cut-off excluded sites without enough read depth to call variants accurately.

For a given position meeting these criteria, variants were categorised based on the following sequential steps: 1) if the ratio of alternative read depth to total read depth (alternative plus reference read depth) in the cancer sample was ≥ 0.1, and in the matching normal sample it was ≤ 0.01, we called it a potential cancer specific mutation; 2) if the ratio in the cancer sample was ≤ 0.01 and in the matching normal sample was ≥ 0.1, we called it a potential somatic mutation; and 3) if both of the ratios were ≥ 0.1, we called it a potential germline variant.

Conservative methods were employed to identify mutations likely to be specific to cancer, somatic, or germline cells. Potential cancer or normal tissue specific mutations were filtered against dbSNP132 variants, 1000 genome variants, and the germline specific variants called. In addition, for mutations that were present only in the cancer tissues (cancer specific

mutations), the genotype of the same position in normal matching tissue must be reference homozygous (0/0) and in the cancer tissue must be variant heterozygous or homozygous (0/1 or 1/1). Correspondingly for somatic mutations, the genotype in cancer tissue must be 0/0, and in normal tissue must be 0/1 or 1/1. For potential germline variants that were present in both normal and cancer tissues, only those observed in 1000 genomes and different to the ancestral allele were kept.

Due to the design of exome sequencing, regions upstream of TSS are not as well covered and as deeply sequenced as downstream exonic regions. To account for such sequencing coverage bias, we calculated the total number of patients with enough coverage at each position in a 2kb window across Fantom5 TSSs (1kb at each side, The FANTOM Consortium and the RIKEN PMI and CLST (dgt) 2014). Consequently, the mutation rate at each position from a given TSS was measured by dividing the observed number of patients carrying a variant by the total number of patients that had enough coverage at the position to call a variant if it existed.

Table 4.3. List of disease in the study.

Cancer type	**Patient number**
Bladder Urothelial Carcinoma (BLCA)	15
Breast invasive carcinoma (BRCA)	110
Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC)	14
Colon adenocarcinoma (COAD)	9
Glioblastoma multiforme (GBM)	208
Head and Neck squamous cell carcinoma (HNSC)	85
Kidney renal clear cell carcinoma (KIRC)	168
Kidney renal papillary cell carcinoma (KIRP)	16
Acute Myeloid Leukemia (LAML)	54
Brain Lower Grade Glioma (LGG)	50
Lung adenocarcinoma (LUAD)	26
Lung squamous cell carcinoma (LUSC)	53
Ovarian serous cystadenocarcinoma (OV)	73
Prostate adenocarcinoma (PRAD)	40
Stomach adenocarcinoma (STAD)	19
Thyroid carcinoma (THCA)	19
Uterine Corpus Endometrioid Carcinoma (UCEC)	38

4.3 Results

4.3.1 Nucleosome phasing is well conserved among different human cell lines

To investigate whether nucleosome positioning is relatively consistent among different human cell lines as described in Table 4.1, I first compared the inter-nucleosome distance of all nucleosomes carrying any histone modification for 11 cell lines. Most nucleosomes are distant to the nearest other called nucleosome (data not shown) so I confined our analysis only to nucleosomes within 500bp of their nearest neighbours (Figure 4.2). The pattern of nucleosome spacing (inter-nucleosome distance) was found to be consistent among different datasets (cell lines). All cell lines show a similar peak of inter-nucleosomes distances at ~190bp, corresponding to an inter-nucleosome linker length of ~43 bp given that the length of nucleosome core is 147 bp.

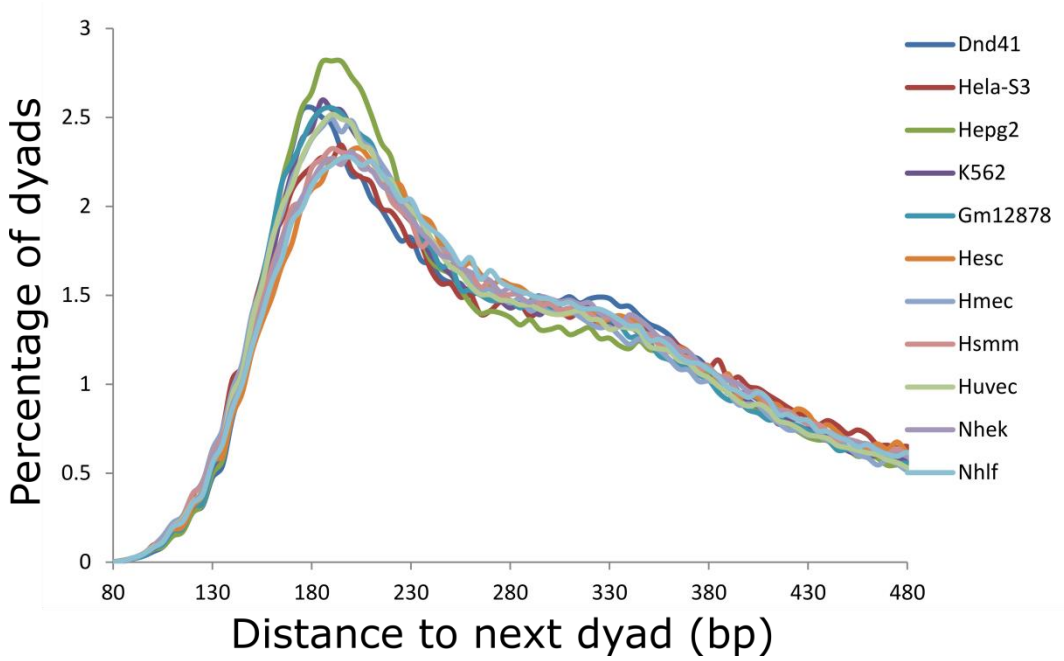


Figure 4.2. Nucleosome phasing is conserved among different cell lines. Inter-nucleosome distances were limited to 500bp. Frequency is normalised to account for the different numbers of nucleosomes covered by each dataset (cell line), by calculating the percentage of nucleosomes with each inter-nucleosome distance up to 500 bp (the frequency at each distance divided by the total number of nucleosomes for each dataset). Normalised frequency (percentage of dyads) is plotted against the distance between two successive nucleosome dyads.

I then calculated the average frequency for each inter-nucleosome distance across 7 normal cell lines and 4 cancer cells respectively and the distribution of the cell line averaged normal and cancer inter-nucleosome distance confirmed that the nucleosome spacing was generally conserved between normal and cancer genomes, suggesting that the nucleosome organization in the cancer genome was not globally altered compared to that in the normal genome (Figure 4.3). The overall consistency in the global nucleosome organizations was also observed in a recent study (West et al. 2014) that nucleosome positioning only changes locally at key regulatory regions during cell differentiation and reprogramming. It might be the case that the cancer genome features the local alteration in nucleosome organization, while global patterns are generally maintained.

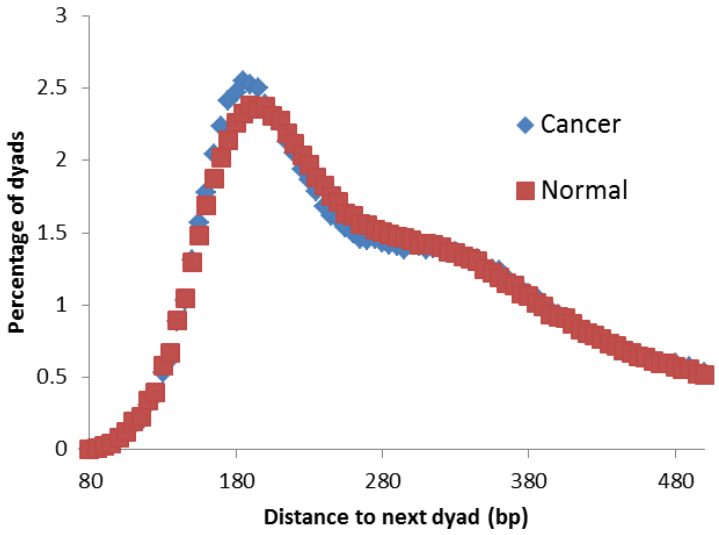


Figure 4.3. Nucleosome phasing is globally conserved between normal and cancer genomes. The normalized frequency of each inter-nucleosome distance was averaged across 7 normal and 4 cancer cell lines respectively. The averaged frequency was further normalized, such that the percentages of dyads with individual inter-nucleosome distances up to 500 bp in both normal and cancer genomes sum up to 100% respectively, and plotted for normal and cancer genomes.

I lastly directly compared the pattern of phasing of GM12878 to K562 based on bulk nucleosomes not selected for any specific histone modification (Figure 4.4). I observed two peaks in the inter-nucleosome distance distributions for both cell lines: a main peak at ~190

bp and a secondary at ~380 bp (equivalent to the distance of two regularly spaced nucleosomes. The phasing is also relatively conserved between GM12878 and K562 cell lines, except that the secondary peak in GM12878 is higher, suggesting that nucleosomes are slightly more regularly spaced.

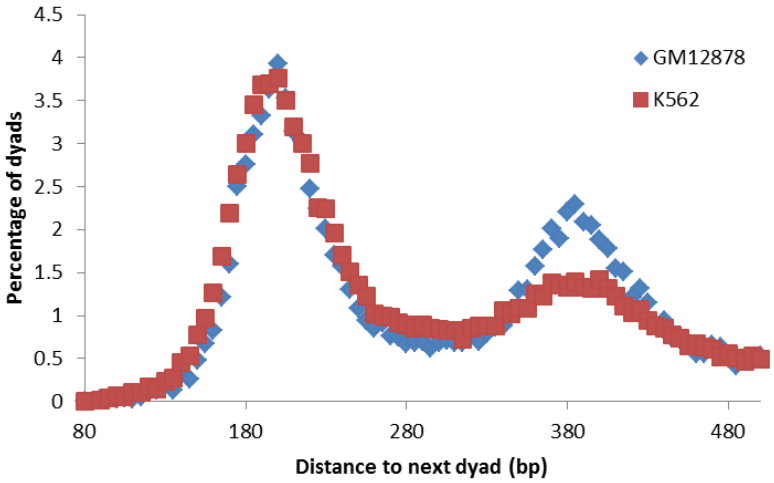


Figure 4.4. Global nucleosome phasing is relatively conserved between GM12878 normal and K562 cancer cell lines. Inter-nucleosome distances are limited up to 500bp. Normalised frequency (percentage) is plotted against the distance between two successive nucleosome dyads.

4.3.2 Nucleosome positioning around FANTOM5 TSSs is moderately conserved among different human cell lines

To further investigate whether the nucleosome organizations are conserved across different cell lines, I also compared the positioning of nucleosomes around FANTOM5 TSSs among 11 cell lines based on all nucleosomes carrying any histone modification (Figure 4.5). Annotation of TSSs in FANTOM5 data is based on the Cap Analysis of Gene Expression (CAGE) approach and provides the most precise annotation of TSSs which are used ubiquitously across almost all the cell types in the human body, without being confounded by the alternative promoter usage, due to the usage of the robust threshold to define the 5' end of the full transcripts which in turn were supported by other evidence to ensure the identified peaks are genuine TSSs (Shiraki et al. 2003; Carninci et al. 2006; The FANTOM Consortium and the RIKEN PMI and CLST (dgt) 2014). The -2 and +1 nucleosomes

immediately flanking the TSSs are strongly positioned in each cell line, as seen at most eukaryotic genes (Jiang and Pugh 2009). The first nucleosome immediately upstream of a TSS is called the -1 nucleosome and usually lost, generating nucleosome depleted region. Correspondingly, the first nucleosome immediately downstream of a TSS is called the +1 nucleosome. While the nucleosomes upstream and downstream of the -1 and +1 ones are called -2 and +2 nucleosomes and so on. Variable positioning to different extents can be observed for nucleosomes downstream of the +1 nucleosome among different cell lines (comparing peaks at +2, +3, and +4 nucleosome positions in Figure 4.5). It was further exemplified by the comparison of nucleosome positioning in Dnd41 cell line to that in H1-hESC and K562 cell lines respectively (Figure 4.6). In Figure 4.6A, the overall positioning between Dnd41 and H1-hESC is correlated (rho = 0.7970706; $p < 2.2e-16$). When zoomed in to the region from -800bp to +800bp relative to the TSS, it clearly shows that the +2, +3, and +4 nucleosomes show shifts of different degrees while the +1 nucleosome is strongly positioned. Comparison between Dnd41 and K562 in Figure 4.6B shows a similar pattern.

In addition, the nucleosome positioning was also consistent between normal (averaged across 7 cell lines) and cancer (averaged across 4 cell lines) genomes, based on the combined dataset of nucleosomes carrying any histone modification examined (Figure 4.7).

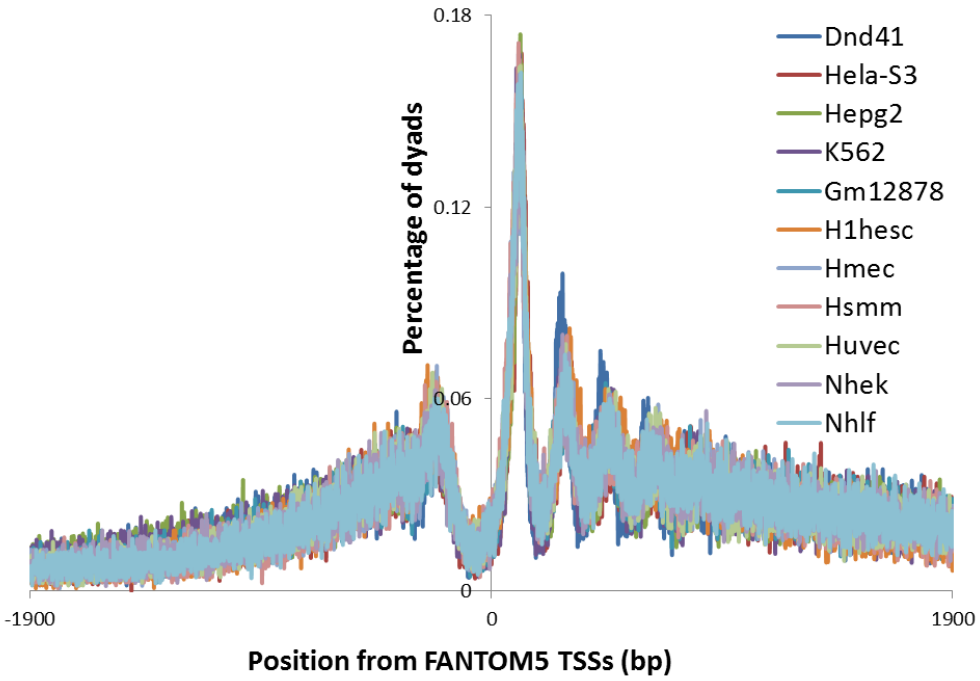


Figure 4.5. Nucleosome positioning around FANTOM5 TSSs. Frequency is normalised to account for the different numbers of nucleosomes covered by each dataset (cell line), by calculating the percentage of nucleosomes at each position relative to TSS (the frequency at each distance divided by the total number of nucleosomes for each dataset). Normalised frequency (percentage) is plotted against relative position from FANTOM5 TSS for 11 cell line.

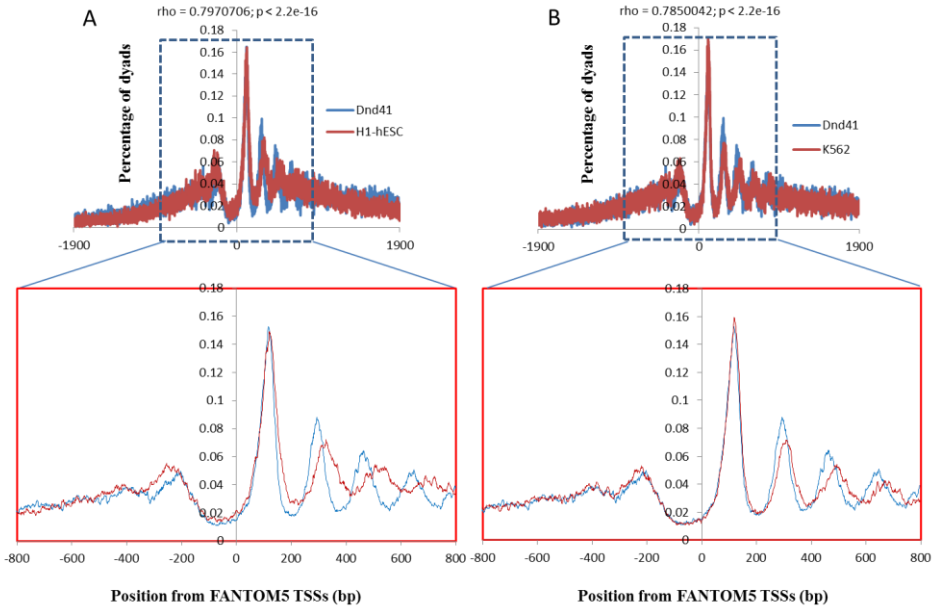


Figure 4.6. Variable positioning of nucleosomes downstream of TSSs in different cell types. Although the +1 nucleosome appears tightly controlled and is consistently positioned between cell types, nucleosomes further downstream display variable positioning peaks in different cell lines.

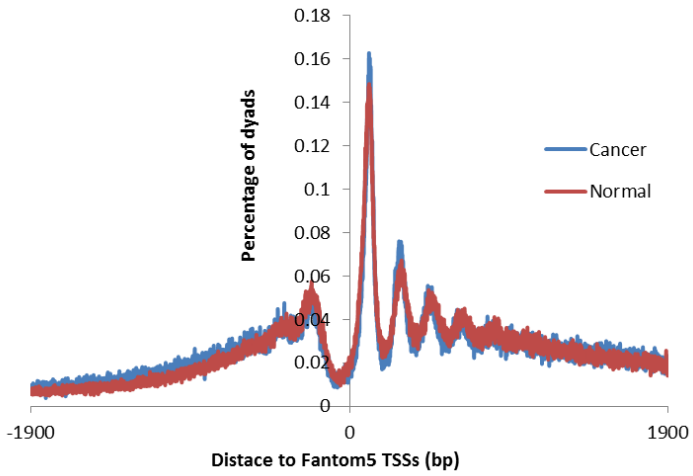


Figure 4.7. Nucleosome positioning around FANTOM5 TSSs in normal and cancer cell lines.

4.3.3 Mutation spectra around FANTOM5 TSSs

To assess whether particular genomic regions around TSSs are more prone to certain types of mutations, I focused on the region from -500 to +500bp across TSSs and compared the substitution density among germline, cancer, and somatic substitutions called from TCGA. Figure 4.8 shows that the number of total valid sites (≥10X coverage) at regions upstream of FANTOM5 TSSs is lower than that at regions downstream, confirming that regions upstream were sparsely covered and inadequately sequenced as expected.

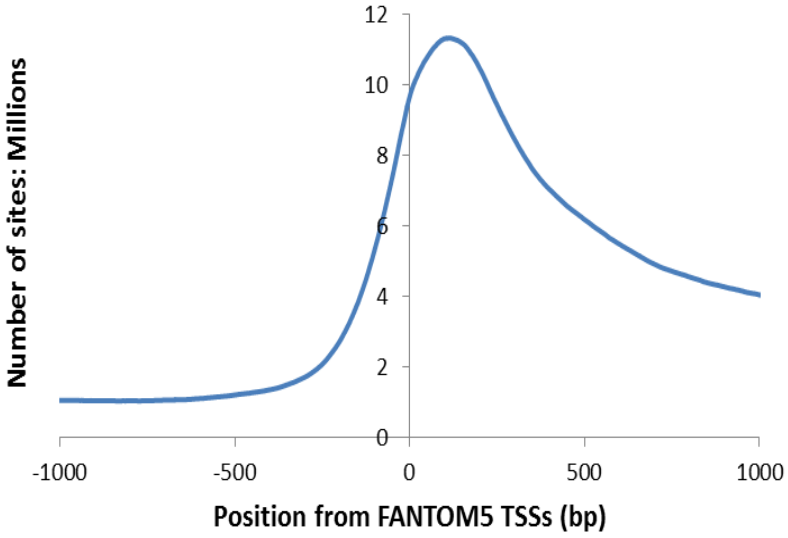


Figure 4.8. Number of valid sites at each position relative to FANTOM5 TSSs. The total number of valid sites for each position relative to TSS was calculated by adding up, across 997 patients, the number of valid sites (covered by ≥ 10 reads in both normal and cancer tissues) observed in the same relative position from each of the 39445 FANTOM5 TSSs.

The position dependent substitution density distribution (20bp sliding window) in Figure 4.9 shows that substitution density is higher in germline than somatic mutation across the region analysed. The substitution density, in both germline and cancer substitutions, is relatively even across the whole region. The substitution density in somatic substitutions is relatively higher at regions upstream of -100bp relative to TSS than at regions downstream. I also observed that the substitution density is higher in transitions than transversions in germline while it appears similar in somatic transitions. Thus the difference in the

substitution density seen across TSSs is mainly between the germline and somatic mutations, which is further supported by the investigation into the composition of substitutions of germline to somatic substitutions in this region as a whole, and results are presented in next section.

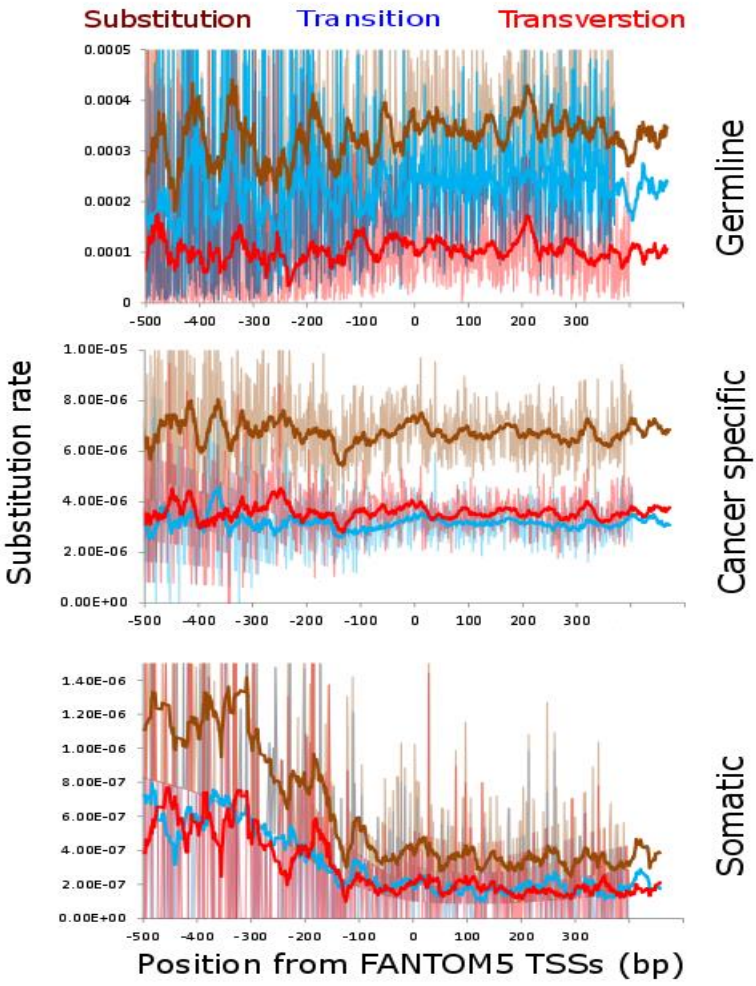


Figure 4.9. Substitution density (original and 20bp sliding window) from -500bp to +500bp relative to FANTOM5 TSSs between germline and somatic mutations. Germline mutations whose different ancestral allele could be ascertained were included. Purple lines represent total substitutions. Blue lines transitions and red lines transversions.

4.3.4 Composition is distinct between germline and somatic substitutions around FANTOM5 TSSs

Pearson's Chi-squared tests were applied to investigate whether the composition of somatic substitutions is different from that of germline substitutions across the region from -500bp to +500bp relative to FANTOM5 TSSs as a whole. The raw data are presented in Table **4.4**. Figure 4.10A shows that, compared to germline substitutions, transitions are relatively under represented (observed to expected ratios: germline, 1.01; cancer, 0.69; somatic, 0.74) while transversions are relatively enriched (observed to expected ratios: germline, 0.99; cancer, 1.67; somatic, 1.55) in somatic substitutions (chi-squared test: $p < 2.2e-16$). The ratio of the observed frequency of transitions to transversions is 2.2, 0.9, and 1.0 in germline, cancer, and normal datasets respectively. The transition/transversion bias observed for germline mutations was consistent with previous findings, and was mainly attributable to the elevated rate in C->T base change due to the deamination of methylated cytosine to thymine in the CpG context (Gojobori et al. 1982; Zhang and Gerstein 2003; Arnheim and Calabrese 2009). However, I also acknowledge the potential confounding from sequencing error that might contribute to the reduced transition/transversion ratio observed for somatic mutations, in the sense that mutation calls as a result of sequencing error would be in theory enriched in what appear to be transversions because possible transversions (C:G <=> A:T and A:T <=> T:A) are twice as many as possible transitions (C:G <=> T:A).

I then investigated the relative contribution of different types of base changes to the transition and transversion biases between germline and somatic substitutions (Figure 4.10B and Table **4.4**). The relative enrichment of transitions in germline substitutions and/or deprivation in somatic substitutions is mainly determined by G:C -> A:T. The enrichment index of G:C -> A:T transitions in germline substitutions is 1.02 while that in somatic transitions is much lower (cancer: 0.28; somatic: 0.41). It strongly supports a well-accepted observation that higher rate of transitions than transversions is a general feature of vertebrate evolution which is at least partly due to the relatively high rate of mutation of methylated cytosine to thymine due to deamination (Colot and Rossignol 1999). In contrast, A:T -> G:C transitions are relatively enriched in somatic rather than germline substitutions, with an observed to expected ratio of 0.99, 1.57, and 1.47 in germline, cancer, and somatic respectively. A:T -> C:G transversions are enriched in somatic substitutions, with the enrichment index of 0.94, 3.78, and 3.03 in germline, cancer, and somatic substitutions respectively. Also, A:T -> T:A transversions are enriched in somatic mutations (germline: 0.97; cancer: 2.22; somatic: 2.37) while the difference in enrichment index in G:C -> T:A

and G:C -> C:G transversions is relatively smaller between germline and somatic substitutions.

Collectively these analyses suggest that mutational compositions are different between germline and somatic mutations at regions across TSSs.

Table 4.4. Observed frequency of substitutions in germline and somatic substitutions in the region from -500bp to +500bp relative to FANTOM5 TSSs as a whole.

	Germline	Cancer	Somatic	Total
AT>TA	103482	4709	260	108451
AT>GC	448197	14184	689	463070
AT>CG	113377	9062	377	122816
GC>TA	218656	3224	184	222064
GC>CG	218356	5093	240	223689
GC>AT	984093	5338	407	989838
Total	2086161	41610	2157	2129928

	Germline	Cancer	Somatic	Total
Transition	1432290	19522	1096	1452908
Transversion	653871	22088	1061	677020
Total	2086161	41610	2157	2129928

Note: Pearson's chi-squared test was applied to detect whether the mutational composition (six types of base specific base changes, top part of table; transition vs. transversion, bottom part of table) is independent of mutational classes (germline, cancer and somatic). The null model assumes that the mutational composition is independent of the mutational classes. For example, the expected number of germline mutation that is transition under the null model is derived as: 1) the proportion of mutations that are germline is $\frac{2086161}{2129928}$; 2) the proportion of mutations that are transition is $\frac{1452908}{2129928}$; and 3) the expected number is $\frac{2086161}{2129928} * \frac{1452908}{2129928} * 2129928 = 1423053$. In test between transition and transversion, X-squared = 9194.314, df = 2, and p-value < 2.2e-16. Post-hoc tests to look for significant differences in terms of transversion vs. transition substitutions was done by the "chisqPostHoc" function in the R package "NCStats" (https://rforge.net/NCStats/), and the FDR-adjusted p values for contrasts Germline vs. Cancer, Germline vs. Somatic, and Cancer vs. Somatic were all equal 0. In tests among different types of base changes, X-squared = 37070.32, df = 10, and p-value < 2.2e-16. The post-hoc tests for pair-wise comparisons between different base substitution types (15 comparisons in total) and between germline and somatic classes (3 comparisons) were all significant.

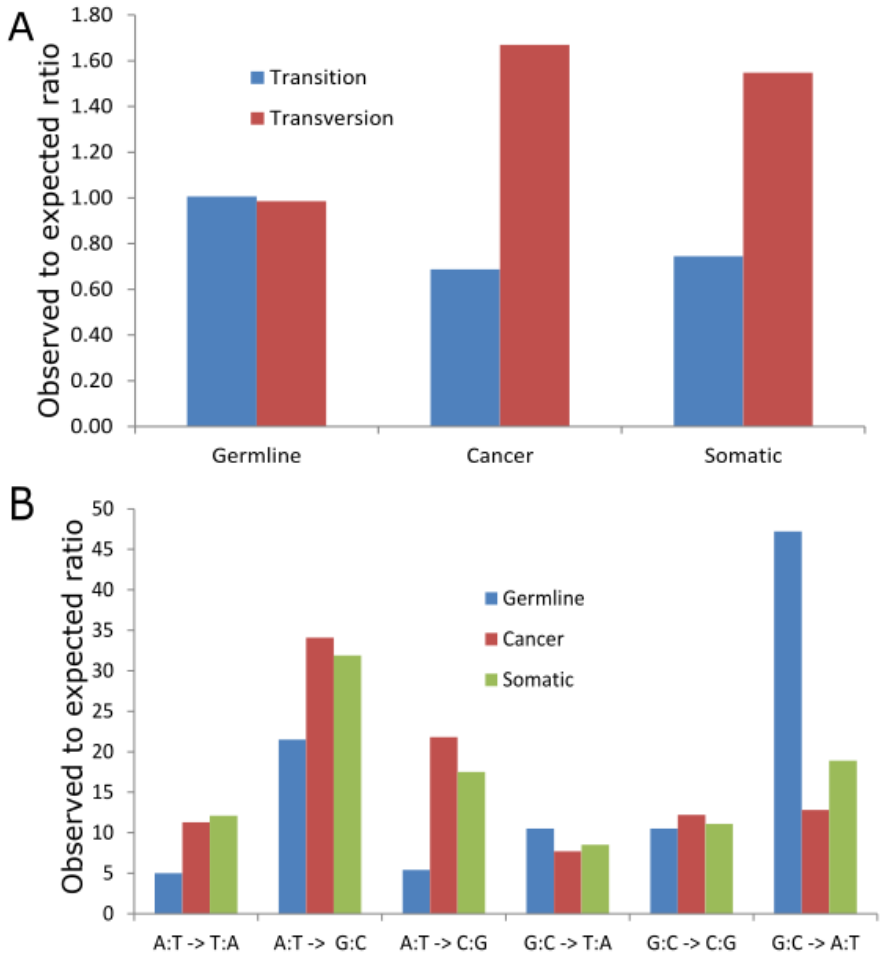


Figure 4.10. Transition and transversion biases between germline and somatic substitutions in the region from -500bp to +500bp relative to FANTOM5 TSSs as a whole. A: transition and transversion biases between somatic and germline substitutions. B: enrichment biases in different types of base changes between somatic and germline substitutions. Though theoretically there are 12 types of individual base changes, they are collapsed into 6 types. The logic behind this is that if we observe a T -> A change at a given genomic position, we cannot tell whether it comes from the + strand or – strand.

4.3.5 Mutation spectra around nucleosome dyads

To test whether the nucleosome structure is associated with the biases between germline and somatic substitutions, I compared the spectra of both germline substitutions against that of somatic mutations around nucleosome dyads. The nucleosome dyad dataset used in this analysis, covering 817,774 autosomal nucleosomes, was produced by Schones et al. (2008) and used in Prendergast and Semple (2011) where the inter-species sequence divergence and intra-species sequence diversity (SNPs) were found to be higher in nucleosome cores. In this

analysis, unlike the calculation of the position dependent substitution density across FANTOM5 TSSs, I directly compared the position dependent frequency (counts) of mutations across dyads without normalizing the frequency at each position by the total number of valid sites meeting the criterion of 10X coverage. Theoretically not correcting for the total number of valid sites at each position relative to the nucleosome dyad might introduce bias in the calculation of the position-dependent mutational density; however, we speculate that it is less of a problem since the variation in sequencing coverages across the nucleosomes should not be as substantial as that observed at regions across TSSs.

Figure 4.11 shows that substitution frequency (20bp sliding window) is generally lower in somatic than germline across the region from -500bp to +500bp relative to the nucleosome dyad. Consistent with results comparing transitions and transversions across FANTOM5 TSSs, the frequency of transitions is relatively higher than that of transversions in germline substitutions but the difference is not obvious in somatic substitutions.

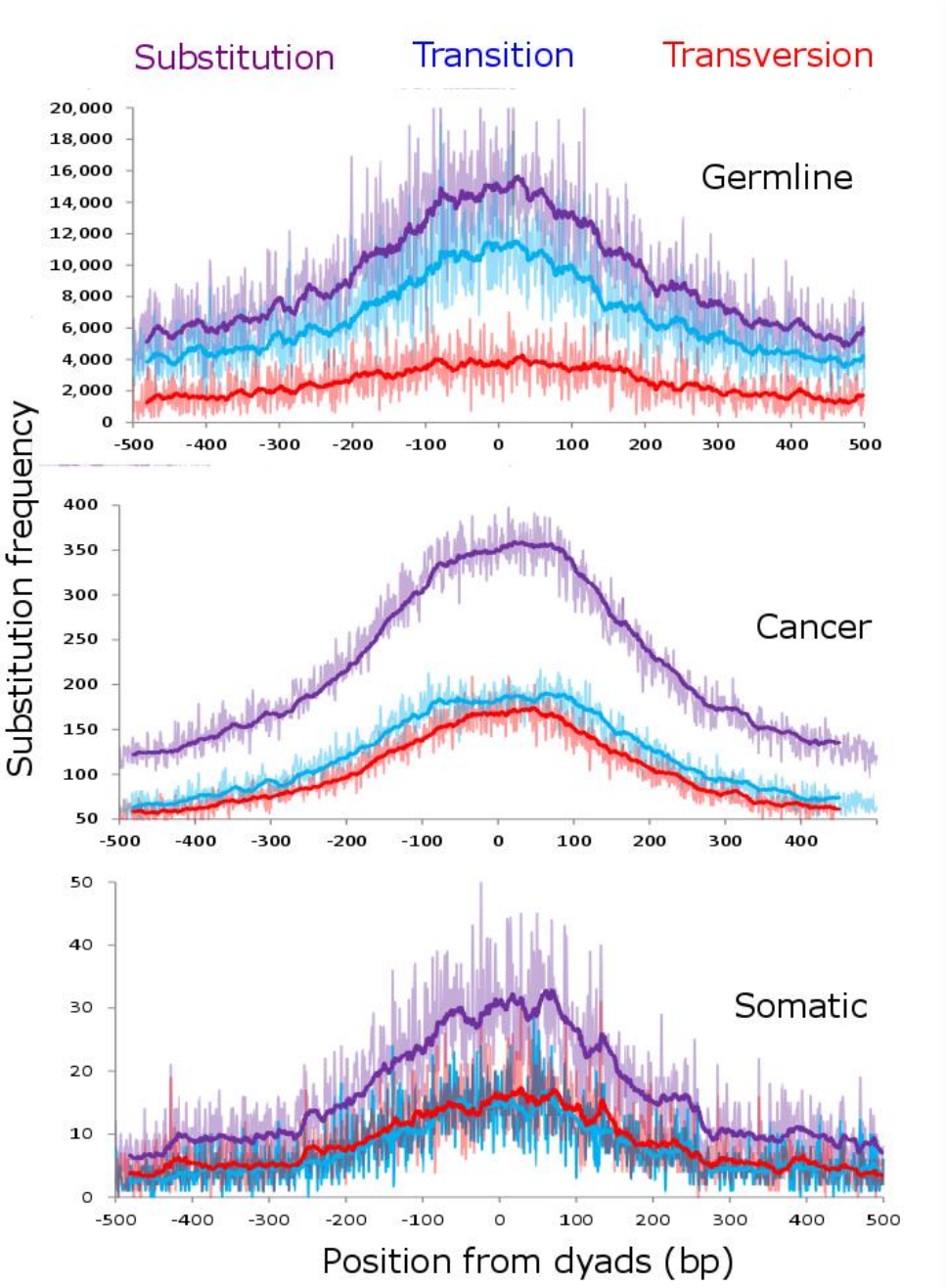


Figure 4.11. Substitution frequency (original and 20bp sliding window) from -500bp to +500bp relative to nucleosome dyad between germline and somatic mutations. Only germline substitutions whose different ancestral allele could be ascertained were included. Purple lines represent total substitutions; blue lines indicate transitions and red transversions.

4.3.6 Composition of substitutions is distinct between germline and somatic substitutions around nucleosome dyads

To statistically test the difference in the substitution frequency around the nucleosome dyads between germline and somatic substitutions, I adopted a similar strategy as in the comparison of substitutions around TSSs by comparing the observed frequency to the expected frequency under the null model using a Pearson's Chi-squared test. The raw data is presented in Table **4.5**. In addition, instead of focusing on the region from -500bp to +500bp relative to nucleosome dyads, we limited the analysis to the region from -125bp to +125bp relative to dyads as it more accurately reflects the real length of a nucleosome (150bp of nucleosome core and 50bp of linker region at both sides).

Figure 4.12A shows that, compared to germline substitutions, transitions are relatively under represented (observed/expected ratio: germline, 1.01; cancer, 1.97; somatic, 1.93) while transversions are relatively enriched (observed/expected ratio: germline, 0.98; cancer, 1.67; somatic, 1.55) among somatic substitutions ($p < 2.2e-16$). The ratio of the observed frequency of transitions to transversions (Ti/Tv ratio) is 2.81, 0.89, and 0.92 in germline, cancer, and somatic substitutions respectively.

Figure 4.12B shows that the enrichment of transitions in germline substitutions and/or deprivation of transversions is mainly by G:C -> A:T. The enrichment index of G:C -> A:T transitions in germline substitutions is 1.02 while that in somatic transitions is much lower (cancer: 0.29; somatic: 0.40). In contrast, A:T -> G:C transitions are relatively enriched in somatic rather than germline substitutions, with an enrichment index of 0.99, 1.38, and 1.18 observed in germline, cancer, and somatic substitutions respectively. Results are quite consistent in the contradictory observations of G:C -> A:T and A:T -> G:C transitions in the transitions enrichment between germline and somatic substitutions, across both FANTOM TSSs and nucleosome dyads. A:T -> C:G transversions are enriched in somatic substitutions, with the enrichment index of 0.93, 3.84, and 3.10 in germline, cancer, and somatic substitutions respectively. Also, A:T -> T:A transversions are enriched in somatic mutations (germline: 0.95; cancer: 3.05; somatic: 3.97) while the difference in enrichment index in G:C -> T:A and G:C -> C:G transversions is relatively smaller between germline and somatic substitutions.

Table 4.5. Observed frequency of substitutions in germline and somatic substitutions on the region from -125bp to +125bp relative to nucleosome dyads.

	Germline	Cancer	Somatic	Total
AT>TA	133673	10208	1125	145006
AT>GC	823621	27063	1973	852657
AT>CG	171616	16778	1147	189541
GC>TA	295333	6893	637	302863
GC>CG	320027	10021	745	330793
GC>AT	1760593	11887	1400	1773880
Total	3504863	82850	7027	3594740

	Germline	Cancer	Somatic	Total
Transition	2584214	38950	3373	2626537
Transversion	920649	43900	3654	968203
Total	3504863	82850	7027	3594740

Note: Pearson's chi-squared test was applied to detect whether the mutational composition (six types of base-specific base changes, top part of table; transition vs. transversion, bottom part of table) is independent of mutational classes (germline, cancer and somatic). The null model assumes that the mutational composition is independent of the mutational classes. For example, the expected number of germline mutation that is transition under the null model is derived as: 1) the proportion of mutations that are germline is $\frac{3504863}{3594740}$; 2) the proportion of mutations that are transition is $\frac{2626537}{3594740}$; and 3) the expected number is $\frac{3504863}{3594740} * \frac{2626537}{3594740} * 3594740 = 2560867$. In test between transitions and transversion, X-squared = 31610.08, df = 2, and p-value < 2.2e-16. The FDR adjusted p values in post-hoc tests for Germline vs. Cancer and Germline vs. Somatic equal 0 while that for Cancer vs. Somatic equals 0.114. In test among different types of base changes, X-squared = 80916.25, df = 10, and p-value < 2.2e-16. The post-hoc tests for pair-wise comparisons between different base substation types (15 comparisons in total) and between germline and somatic classes (3 comparisons) were all significant.

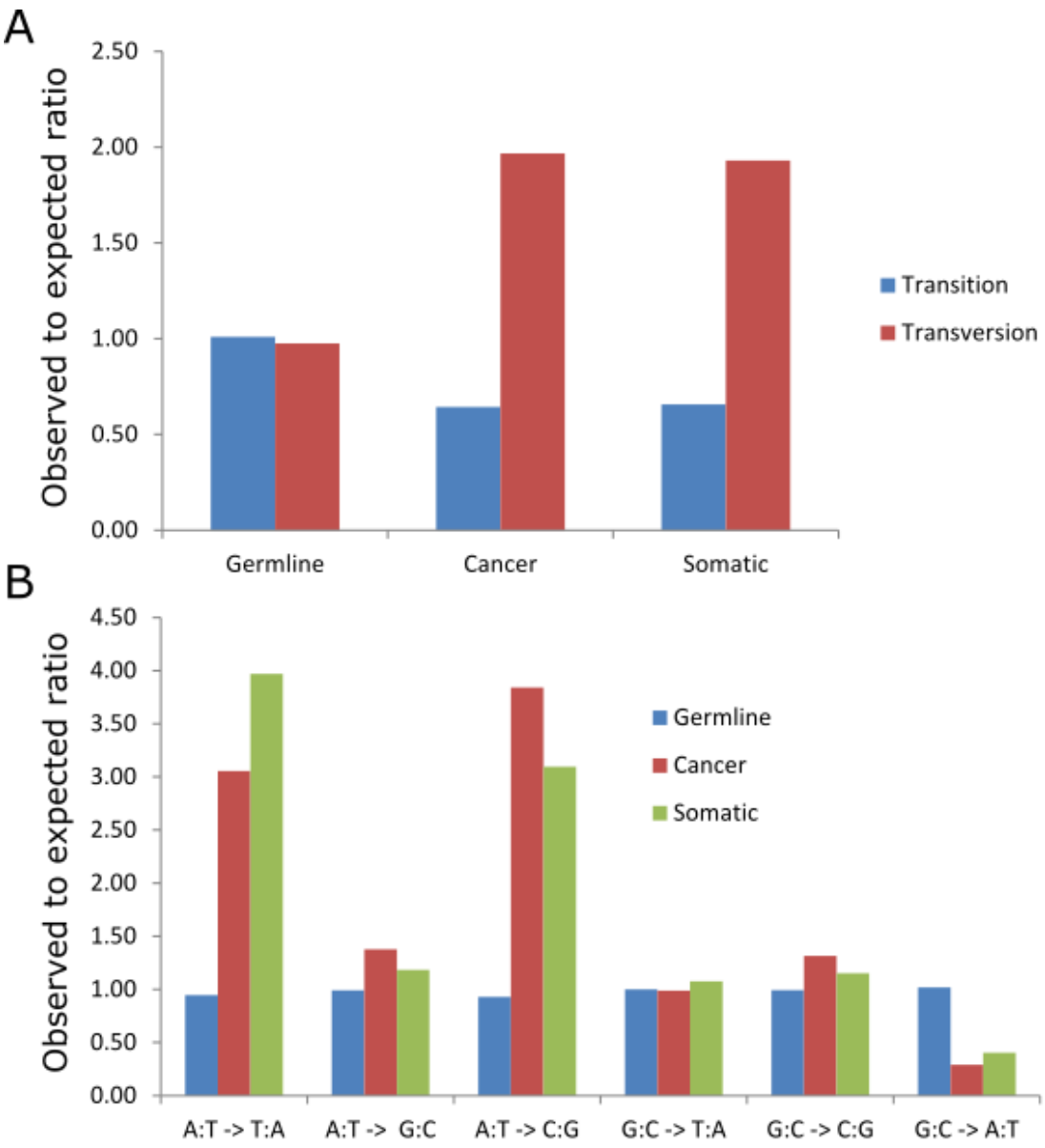


Figure 4.12. Transition and transversion biases between germline and somatic substitutions on the region from -125bp to +125bp relative to the nucleosome dyad. A: transition and transversion biases between somatic and germline substitutions. B: enrichment biases in different types of base changes between somatic and germline substitutions.

4.4 Discussion

In this chapter, I have shown that nucleosome spacing among different human cancer and normal cell lines is well conserved. Overall nucleosome positioning around FANTOM5 TSSs is consistent among cell lines. In addition, the positioning of -1 and +1 nucleosomes is well conserved while that of further downstream nucleosomes shows shifts of different degrees among different cell lines. Although the nucleosome positioning has been observed to alter at specific genomic loci associated with a subset of genes during cell differentiation and reprogramming and in cancer cells (Schones et al. 2008; Portela and Esteller 2010; Brait and Sidransky 2011; Hassler and Egger 2012; Plass et al. 2013; West et al. 2014), the global patterns of the positions of bulk nucleosomes or nucleosomes carrying specific histone modifications were not noticeably globally altered in cancer cells with respect to normal cell lines, suggesting the contrasts seen between somatic and germline mutational spectra do not appear to be attributable to alterations in nucleosome positioning between cell types.

The germline and somatic substitution spectra across FANTOM5 TSSs and nucleosome dyads are different. Using the Pearson's Chi-squared test, we found that the composition of somatic substitutions is quite distinct from that of germline substitutions. Also, the transition bias has been shown to be exclusively determined by the deprivation of G:C -> A:T transitions in somatic substitutions, although we could not exclude the possible confounding from the errors in variant calling. In addition, it should not be surprising to observe the expected transition/transversion ratio in germline variants since I only included the germline variants if also seen in 1000 genomes (filtering against 1000 genomes should be expected to reduce the false positive germline variants detected). The biased enrichment of A:T -> C:G and A:T -> T:A changes are responsible for the general transversion enrichment in somatic mutations.

One limitation in the mutation spectra analysis is that when we defined somatic mutations at a non-polymorphic genomic site (non-SNP site), we did not take into consideration the effect of the loss of heterozygosity (LOH) in tumour cells and the LOH rate was substantial in some cancer genomes (data not shown); thus a called somatic mutation might be a false positive in the sense that it might be a de novo germline mutation but the alternate base has been lost in cancer cells due to LOH in cancer. Future study will be focused on comparing mutational spectra at genomic regions where no LOH was detected in tumour cells.

Chapter 5: Genetic Determinants of Somatic Mutation Rates in Blood Cells

5.1 Introduction

As discussed in Chapter 1, cancer can be regarded as a disease of the genome and is the result of an evolutionary process within populations of cells (Crespi and Summers 2005; Jones et al. 2008; Ye et al. 2009; Heng et al. 2010; Heng et al. 2011). A crucial requirement for the initiation and progression of cancer is the accumulation of somatically acquired DNA mutations in normal cells. The sources of DNA mutations are diverse, including but not limited to, exogenous and endogenous DNA mutagens, decreased fidelity of DNA replication and defects in DNA repair pathways (Salk et al. 2010).

Genomic DNA suffers constant damage by both endogenous and exogenous mutagens (Salk et al. 2010). Examples of endogenous DNA damage are the oxidation of DNA by reactive oxygen species (ROS) such as the 8−oxo-deoxyguanosine (8−oxo-dG) (Nishimura 2006), and the deamination of methylated cytosine at CpG sites (Helleday et al. 2014). The environmental mutagens can be either chemical or physical, and the DNA damage caused is predominantly a stochastic process and shows characteristic patterns associated with different sources. The non-ionizing radiation by UV light causes the covalent modification between neighbouring pyrimidines, resulting in CC:GG ->TT:AA mutations which are often predominant in skin cancers (Salk et al. 2010; Helleday et al. 2014). Another well studied example is that the lung cancer caused by tobacco smoke features elevated G->T mutations (Hecht 1999; Lee et al. 2010; Pleasance et al. 2010; Helleday et al. 2014).

Though not perfect, DNA replication is a process of high fidelity and is associated with an extremely low frequency of spontaneous mutations (Salk et al. 2010). The bulk replication of DNA is carried out by DNA polymerases Pol δ and Pol ε (Garg and Burgers 2005). The proofreading domains in POLD1 (the catalytic subunit of Pol δ) and POLE (the catalytic subunit of Pol ε) which work as exonuclease guarantee the low replication error rate. Thus defects in the proofreading functions can increase the mutation rates, generating mutator phenotypes (Lawrence A Loeb 2011; Heitzer and Tomlinson 2014). For example,

both germline and somatic mutations in POLD1 and POLE were discovered in the colorectal cancers (Briggs and Tomlinson 2013; Palles et al. 2013; Heitzer and Tomlinson 2014).

A crucial mechanism against mutation accumulation is the DNA repair system, such as the involvement of the p53 in DNA repair by arresting cells at G1 phase, and mutations in p53 can lead to increased DNA damage (Calvert and Frucht 2002). Distinct DNA repair pathways involving more than 100 repair genes have been found to repair both single-strand breaks (SSBs) and double strand breaks (DSBs) and thus the defects in DNA repair pathways are often associated with increased spontaneous mutation rates (Salk et al. 2010; Lawrence A Loeb 2011; Dietlein et al. 2014). For example, the DNA mismatch repair (MMR) pathway targets and repairs the mis-incorporated bases during DNA replication, as well as some insertions and deletions (indels) (Dietlein et al. 2014; Helleday et al. 2014). Mutations in MMR genes, including MSH2 and MSH6, have been found to be associated with microsatellite instability (MSI) characterised by a high frequency of indels at simple tandem sequence repeats, and in particular linked to increased mutation rates in colorectal cancer (Network 2012). In addition, by assessing the single nucleotide variants in 652 tumours, Supek and Lehner (2015) found that differential DNA mismatch repair underlies variations in the mutation rates of intervals measured at mega-base scale across the genome. Both the base excision repair (BER) and nucleotide excision repair (NER) pathways recognise and repair SSBs (Helleday et al. 2008; Dietlein et al. 2014). While the BER pathway targets and excises the single damaged base, the NER pathway usually excises a short stretch of ~30 nucleotides (Dietlein et al. 2014). In BER pathway, the single damaged base is excised by a specific DNA glycosylase, the resulting abasic site is cleaved by APEX1 nuclease, and finally the SSB is repaired by a DNA repair complex. The defects in the BER pathway have been shown to have a causative role in colorectal cancer (Farrington et al. 2005). The NER pathway is a major DNA repair pathway involved in maintaining genome integrity and can be either global genome repair (GGR) or transcription-coupled repair (TCR) (Kamileri et al. 2012; Dietlein et al. 2014). TCR can cause strand bias in the repair efficiency such that the DNA damage on the transcribed strand can be repaired more efficiently compared to that on non-transcribed strand (Nouspikel 2009; Kamileri et al. 2012). The DSBs can be repaired by homologous recombination (HR) and non-homologous end-joining (NHEJ). The HR pathway uses the sister chromatid as the template and generally happens in S and G2 phases (Chapman et al. 2012); the NHEJ pathway does not need a sister chromatid but directly ligates broken DNA ends, and preferentially happens in the G1 phase (Hartlerode and Scully 2009). Heterozygous mutations in HR genes, including BRCA1, BRCA2 and RAD51C are associated with increased cancer risks, and homozygous mutations

(loss-of-function biallelic mutations) in those genes have been discovered in different cancers (Dietlein et al. 2014). However, NHEJ is usually error prone and has been shown to be involved in the formation of translocations, representing a source of genome instability and a variant in NHEJ gene Ligase IV has been found to be associated with a decrease in breast cancer risk (Kuschel et al. 2002; Dietlein et al. 2014).

The rate of the accumulation of somatic mutations is therefore affected by many genetic and genomic factors (Salk et al. 2010; Hodgkinson and Eyre-Walker 2011; Dietlein et al. 2014; Helleday et al. 2014). There are likely many further genes that directly or indirectly affect somatic mutation rates and potentially play an important role in the emergence of diseases such as cancer (Lawrence A Loeb 2011; Network 2012; Supek and Lehner 2015).

In addition to the discovery of germline mutations that predispose individuals to diseases, genome-wide association study (GWAS) has also been successful in understanding the genetic architecture of quantitative trait phenotypes (like blood pressure and height) affected by multiple loci where each locus has only a moderate effect (Cho et al. 2009; Stranger et al. 2011; Visscher et al. 2012). For example, GWAS has been applied to successfully identify the loci that affect blood pressure (ICBP 2011), lipid levels (Global Lipids Genetics Consortium 2013), adult human height (Wood et al. 2014), and body mass index (Speliotes et al. 2010).

In this chapter, I focused on the somatic mutation rate of single nucleotide variants (SNVs) in blood derived normal cells. I tested whether an individual's age is associated with the number of somatic mutations they carry in their normal tissue and searched for potential genetic determinants of somatic mutation rate using GWAS.

5.2 Methods

5.2.1 Patient selection and raw variant calling

Data used in this chapter are somatic mutations from normal blood derived cells, a subset of the TCGA dataset used in Chapter 4. The rates of loss of heterozygosity (LOH) in cancers can be substantial, and the LOH events in cancer tissues can confound the somatic mutations called in blood derived normal cells, as a called normal cells specific SNV might be a false positive in the sense that it might be a germline mutation but the alternate base has been lost in the cancer cells due to the LOH. We thus only analysed SNVs in normal blood samples where the overall LOH rates in corresponding tumour tissue were low (arbitrarily chosen as ≤ 0.11 to maximize the number of cases). The list of overall LOH rate and LOH regions in cancer tissues was detected by ExomeCNV (Sathirapongsasuti et al. 2011) and provided by Alison Meynert. The ExomeCNV detects the LOH in cancer tissues, based on contrasts in depth-of-coverage and minor allele frequency between cancer and paired normal tissues at known SNP sites (1000 genomes) across the genome (Sathirapongsasuti et al. 2011). In total, 372 out of 997 patients were selected (the list of included patient numbers per cancer is summarised in Table 5.1) where the control (non-cancer) tissue sample sequenced was blood. Varscan (version 2.3.5, Koboldt et al. 2012) was used to detect SNVs specific to normal somatic cells and also germline SNVs based on the normal-cancer sample pairs.

To call somatic and germline SNVs in normal blood cells using Varscan, the analysis protocol was reversed from that traditionally used to call somatic mutations in tumour samples, i.e. the pre-aligned BAM files from primary tumour tissues were used as the control files. The general workflow is depicted in Figure 5.1. For each patient, pileup inputs that summarize the base calls at the reads mapping to each genomic position were generated respectively for normal blood cells and primary tumour tissue by Samtools (version 0.1.19, Li et al. 2009) from Stampy realigned BAM files (by Alison Meynert, Lunter and Goodson 2011). At a given position, bases were only included if the base quality was $\geqslant 20$ (-Q 20) and the minimum mapping quality of the containing read was $\geqslant 1$ (-q 1), to exclude reads that mapped to multiple genomic regions. Normal cells specific somatic mutations and germline variants were called by Varscan with default parameters.

Table 5.1: List of patient numbers whose control tissues are from blood derived normal cells.

Cancer type	Patient number
Bladder Urothelial Carcinoma (BLCA)	4
Breast invasive carcinoma (BRCA)	45
Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC)	6
Colon adenocarcinoma (COAD)	7
Glioblastoma multiforme (GBM)	89
Head and Neck squamous cell carcinoma (HNSC)	37
Kidney renal clear cell carcinoma (KIRC)	35
Kidney renal papillary cell carcinoma (KIRP)	4
Brain Lower Grade Glioma (LGG)	43
Lung adenocarcinoma (LUAD)	10
Lung squamous cell carcinoma (LUSC)	7
Ovarian serous cystadenocarcinoma (OV)	1
Prostate adenocarcinoma (PRAD)	28
Stomach adenocarcinoma (STAD)	4
Thyroid carcinoma (THCA)	19
Uterine Corpus Endometrioid Carcinoma (UCEC)	33

5.2.2 Variants filtering and mutation rate calculation

To exclude the possibility that a given genomic position is well sequenced in some patients but not others, only genomic positions that were covered by at least 10 reads in both cancer and normal blood cells for all 372 patients were included and analysed. This was done to reduce the bias, such as hypermutated regions represented in one sample but not another. To exclude the impact of LOH in cancer, any genomic positions which were located in the regions defined as LOH in the cancer were discarded. Somatic mutations are rare events and are unlikely to occur multiple times at the same position. Therefore any normal cells specific mutations that were seen in another patient were excluded. In addition, potential normal cells specific mutations were further filtered against dbSNP135 variants, 1000 genome variants, and germline variants from 372 patients.

The overall somatic mutation rate in each patient was calculated as the ratio of the number of normal cells specific somatic mutations to the total number of valid sites passing the above filters.

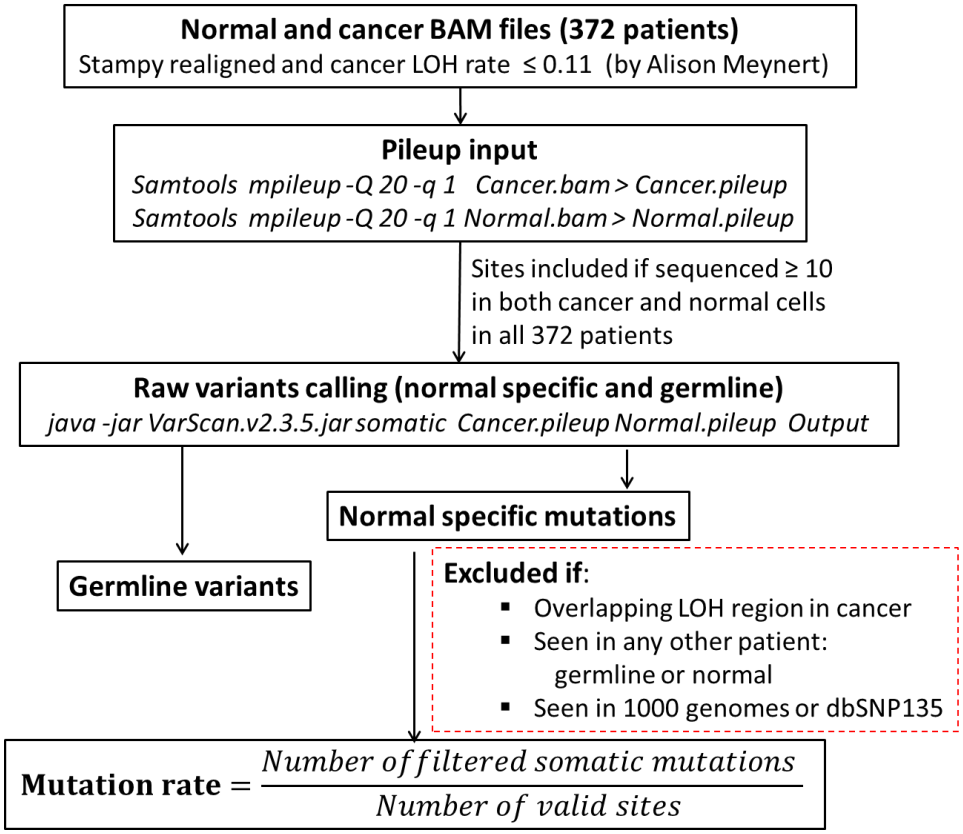


Figure 5.1. Work flow for calling normal cells specific somatic variants and germline variants.

5.2.3 Genome-wide association study (GAWS) for genetic determinants of normal cells specific somatic mutation rates

The quantitative trait association between the overall normal cells specific somatic mutation rates and germline variants (mean and median number of called variants per individual are 91438 and 86068 respectively) called by Varscan was tested by PLINK (Purcell et al. 2007) with adaptive permutations on a per-SNP basis with default parameters except that the maximum number of permutations per SNP was set as 100000000 (--aperm

10 100000000 0.0001 0.01 5 0.001) instead of the default 1000000, where the phenotype is the overall normal cells specific somatic mutation rates (for more information please see "Quantitative trait association" section in the PLINK manual: http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml#qt)). The command used was: plink --lfile germline --allow-no-sex --assoc --aperm 10 100000000 0.0001 0.01 5 0.001 --qt-means –out germline_aperm.

To account potential confounders, I also did a post-GWAS quality control analysis. To only keep SNPs with high quality, a given SNP must meet the following criteria: the proportion of patients with missing genotypes was <0.05, minor allele frequency (MAF) was >0.01, and Hardy-Weinberg Equilibrium (p>0.001) must be met. The command used was: *plink --lfile germline --noweb --allow-no-sex --maf 0.01 --geno 0.05 --hwe 0.001 --out germline.filtered --make-bed.*

5.3 Results and discussions

5.3.1 The overall somatic mutation rate and age

The variation in the somatic mutation rates in blood cells was considerable across 372 individuals, ranging from ~1.3 per 10^6 to ~6.3 per 10^4 bases with a median rate of ~3.8 per 10^6 bases (Figure 5.2). Age has been shown to have a linear relationship with the germline mutation rates (Crow 2000; Hodgkinson and Eyre-Walker 2011; Ségurel et al. 2014), such that Kong et al. (2012) noted an increase of ~2 *de novo* germline mutations per year while Michaelson et al. (2012) noted a rate of ~1.02 *de novo* mutations per year. In order to investigate whether the somatic mutations accumulate in the normal blood cells as a function of age, I performed a correlation test but failed to find a significant relationship across the 372 individuals (Pearson's product-moment correlation test: p=0.1546 and rho=0.07). However, examination of Figure 5.2 highlights that 7 individuals showed evidence of a mutator phenotype, i.e. they exhibited a substantially higher somatic mutation rate than the other individuals in this dataset. Intriguingly 6 out of 7 of these individuals were colon adenocarcinoma patients. Deficient DNA repair is a common feature of colon cancer, germline mutations in mismatch repair genes being linked to several of the major heritable colorectal cancer syndromes (Peltomäki 2001). Age and mutation numbers in these "hyper-mutated" individuals did show a significant positive correlation with an average rate of ~4.8 per 10^6 bases per year. The higher mutational load in these individuals potentially leading to the relationship between age and mutation number being easier to detect, however the numbers are small and further investigation is required to follow up this result.

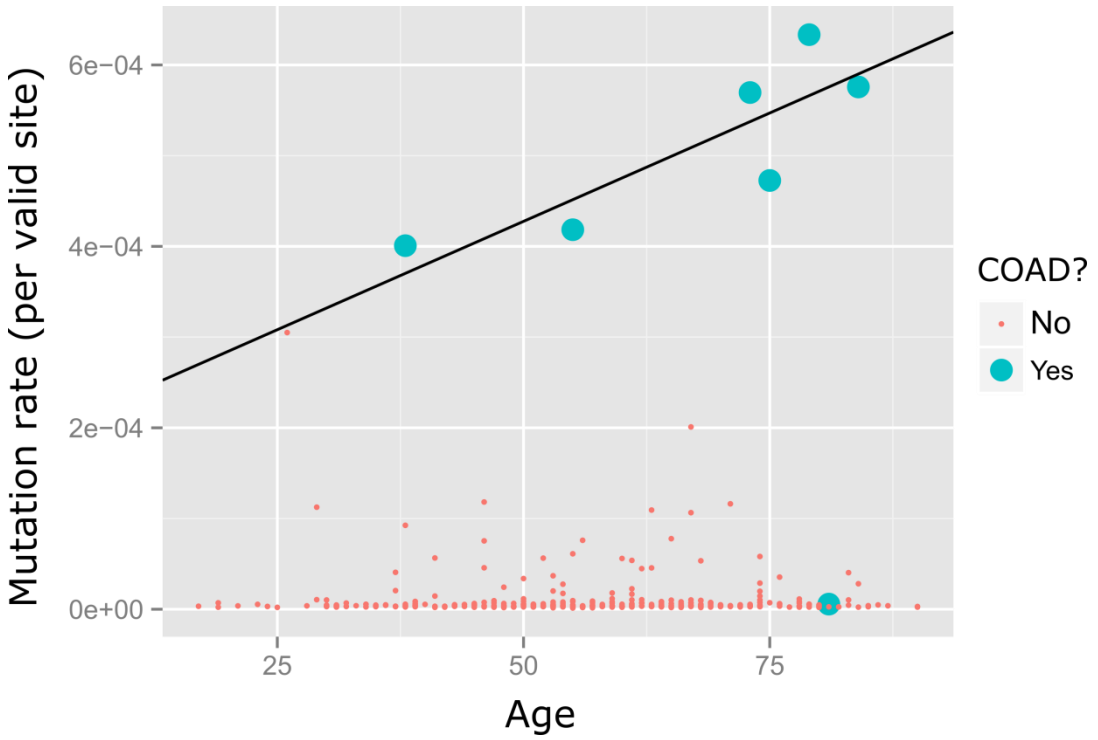


Figure 5.2. Relationship between age and the somatic mutation rates in blood derived normal cells in 372 TCGA patients. Individuals with colon adenocarcinoma (COAD) were displayed as blue while others as red points. The linear relationship between age and mutation rate (represented by black line) was based on a subset of 7 individuals with substantially higher somatic mutation rate: p = 0.004 and adjusted R-squared = 0.80.

5.3.2 Genetic determinants of somatic mutation rate

To search for genetic determinants of these human somatic mutation rates, I performed a quantitative trait association using PLINK with per-SNP based adaptive permutations (see Methods). I obtained a total of 601 germline variants whose genotype was significantly correlated with mutation rate at a p value of ~1.0e-08 which is approximately the level for a genome-wide significant result. The effects of 601 germline variants were annotated using the ANNOVAR program (Wang et al. 2010). Out of 601 significant germline variants, 108 were either nonsynonymous or stop gaining mutations affecting 41 genes in total, and 44 were *de novo* in the sense that they were not seen in 1000 genomes (T. 1000 G.P. Consortium 2012), esp6500 (https://esp.gs.washington.edu/drupal/) and dbSNP135 databases. The summary of genes harbouring significant germline variants is in Table 5.2. Some of these genes are already associated to cancer associated pathways. CDC27 encodes a component of the anaphase-promoting complex and is associated with cell cycle regulation, and has been linked to lung and prostate adenocarcinoma (Gonzalez-Perez et al. 2013; Ahn

et al. 2014; Rubio-Perez et al. 2015). BCLAF1 encodes a transcriptional repressor and can induce apoptosis, and has been shown to regulate the tumorigenic potential of colon cancer cells (Lee et al. 2012; Zhou et al. 2014). CTDSP2 involves in the metabolism pathway which is the major source of reactive oxidative species and associates with diseases including glioblastoma (GBM) (Ping et al. 2015). However, many of the genes in the list seem unlikely to affect the accumulation of mutations in somatic tissues, including OR4C3 and MUC6 (Table 5.2). Although it is possible these variants are tagging causative variants nearby, MUC6 has several similar paralogs and therefore variants in these genes are at greater risk of being false positives (Treangen and Salzberg 2012).

No evidence of population structure was found. However after applying quality control to filter putative false positive variants and linkage disequilibrium pruning, the number of significant germline variants dropped from 601 to 44, mainly due to the genotype information missing. Out of 44 significant germline variants, 13 were nonsynonymous mutations affecting seven genes in total: AK2 (1), CNN2 (3), MAP1B (1), OR4C3 (4), PLIN4 (1), PRB4 (1), and ZNF141 (2).

Though number of significant germline variants dropped markedly due to the 5% threshold applied to missing genotypes, this approach so far shows promise for detecting candidate variants and genes linked to somatic mutation rates. Expanding this preliminary analysis to more individuals represented in the TCGA and other cohorts has the potential to enable the first comprehensive study of the genetic determinants of somatic mutation rates.

Table 5.2. Summary of genes harbouring germline variants associated with variation in somatic mutation rates.

Genes	Germline variants (count)		Summary
	All	***de novo***	
AK2	1	0	Adenylate kinase 2
AKAP13	1	0	A kinase (PRKA) anchor protein 13
ANKRD36	10	3	Ankyrin Repeat Domain-Containing Protein 36
BCLAF1	2	0	BCL2-associated transcription factor 1
CDC27	13	4	Cell Division Cycle 27
CNN2	5	5	Calponin H2, Smooth Muscle
CPEB3	1	1	Cytoplasmic polyadenylation element binding protein3
CTBP2	4	2	C-terminal binding protein 2
CTDSP2	6	6	CTD (carboxy-terminal domain) small phosphatase 2
DUX2	1	0	Double homeobox 2
GPR89B	1	1	G protein-coupled receptor 89B
HNRNPCL1	4	0	Heterogeneous nuclear ribonucleoprotein C-like 1
ITPR1	1	1	Inositol 1,4,5-Trisphosphate Receptor, Type 1
KRT18	4	2	Keratin 18
KRT8	1	1	Keratin 8
LOC440563	6	0	Heterogeneous nuclear ribonucleoprotein C-like
LOC649330	4	0	Heterogeneous nuclear ribonucleoprotein C-like
LST1	1	1	Leukocyte specific transcript 1
MAP1B	1	1	Microtubule-associated protein 1B
MTCH2	1	0	Mitochondrial carrier 2
MUC16	2	2	Mucin 16, cell surface associated
MUC2	3	1	Mucin 2, oligomeric mucus/gel-forming
MUC6	6	1	Mucin 6, oligomeric mucus/gel-forming
MYH7B	1	1	Myosin, heavy chain 7B, cardiac muscle, beta
OR4C3	5	0	Olfactory receptor, family 4, subfamily C, member 3
OR8U1	1	0	Olfactory receptor, family 8, subfamily U, member 1
OTOP1	2	1	Otopetrin 1
PCMTD1	3	0	Protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 1
PLIN4	1	0	Perilipin 4
POTED	1	1	POTE ankyrin domain family, member D
PRB4	1	0	Proline-rich protein BstNI subfamily 4
PRSS3	4	4	Protease, serine, 3
PSMB10	1	1	Proteasome subunit, beta type, 10
RPS20	1	1	Ribosomal protein S20
SHANK3	1	1	SH3 and multiple ankyrin repeat domains 3
TNC	1	0	Tenascin C
TRPM6	1	0	Transient Receptor Potential Cation Channel, Subfamily M, Member 6
ZNF141	2	0	zinc finger protein 141
ZNF585A	1	1	Zinc finger protein 585A
ZNF585B	1	1	Zinc finger protein 585B
ZNF717	1	0	Zinc Finger Protein 717

Note: only significant nonsynonymous or stop-gaining germline variants (p value ~1.0e-08) were included.
"De novo" means germline variants not exist in 1000 genomes, ESP6500 and dbSNP135.

Chapter 6: Discussion

In this thesis, I have investigated nucleosome positioning dynamics in evolution and cancer, the interplay between mutational spectra and nucleosome structure, and also the role of age and genetic determinants in the rate of accumulation of somatic mutations in blood derived normal cells.

In Chapter 2 and 3, I found that nucleosome positioning is generally well conserved between paralogous regions in both the human and yeast genomes: following the duplication of a region and its insertion into a new genomic location nucleosomes generally reassemble at the same, or similar, locations, supporting the importance of the *cis*-acting DNA sequence in nucleosome positioning (Kaplan et al. 2008). I have also observed strong rotational preference and constraints from neighbouring nucleosomes in nucleosome positioning evolution where nucleosome positioning has diverged in both species (Chapter 2 and 3), consistent with previous studies (Albert et al. 2007; Zhang et al. 2009; Struhl and Segal 2013; Hughes and Rando 2014). As far as I am aware, it is the first time that the rotational positioning has been directly tested by comparing the translational positioning of nucleosomes from paralogous regions. Although there are clues suggesting that the positioning of nucleosomes might differ between cancer and normal cell lines (Fraga et al. 2005; Barski et al. 2007; Esteller 2007; Zhang et al. 2008; Jin et al. 2009b; Portela and Esteller 2010; Brait and Sidransky 2011; Wilson and Roberts 2011; Collings et al. 2013), I have shown that global nucleosome organization is broadly conserved across cancer and non-cancerous cell lines, as suggested by the conserved positioning across FANTOM5 TSSs and strongly conserved global phasing (Chapter 4). The local change in nucleosome positioning has been also observed during cell differentiation and reprogramming (West et al. 2014).

As well as the chromosomal environment, both the DNA composition and sequence divergence have been found to be associated with shifts in nucleosome positioning during evolution in both yeast and human genomes. The GC content at the nucleosome core is inversely related to the nucleosome positioning divergence in both species; while showing an inverse relationship in the human genome, the AT content in linker has been observed to be positively correlated with nucleosome positioning divergence in the yeast genome (Chapter 2

and 3), suggesting that the observed difference might be a reflection of the evolution of linker histones and DNA sequence features between human and yeast species (Bates and Thomas 1981; Freidkin and Katcoff 2001; Downs et al. 2003; Cui and Zhurkin 2009; Osmotherly 2010). I have also found, in both yeast and human genomes, that DNA sequence features appear to be more important than local chromosomal environments in nucleosome positioning evolution, while controlling for *trans*-acting factors that can potentially confound inter-species comparisons.

The observed correlation between nucleosome positioning evolution between paralogous regions and sequence divergence (Chapter 2 and 3), suggests that nucleosome structure might be associated with mutational spectra (Prendergast and Semple 2011; Chen et al. 2012). I also investigated the interplay between chromatin structure and DNA sequence variation, with a particular focus on the spectra of (germline and somatic) substitutions in Chapter 4. Based upon a set of 997 patients covering 17 cancer types, I observed that both somatic and germline substitutions are enriched at sequences coinciding with nucleosome cores. This is consistent with a previous study where nucleosome positioning has been shown to influence the types and rates of substitutions across the genome during human evolution (Prendergast and Semple 2011). In addition, transitions appear to be enriched in germline relative to somatic substitutions at nucleosome core regions; this difference in transition to transversion ratio is also seen at transcription start sites (TSSs) genome wide. The contrasts seen between somatic and germline mutational spectra do not appear to be attributable to alterations in nucleosome positioning between cell types, since I have shown that the global nucleosome organizations have been broadly conserved across cancer and non-cancerous cell lines. Instead the particular mutational profiles seen for somatic and germline cells occur upon a common landscape of conserved chromatin structure. However there are some potential drawbacks in the analysis in this chapter: 1) somatic substitutions called were filtered based on the number of reads in paired cancer and normal tissues but variants were not called by programs specifically designed for somatic mutation detection (e.g. Varscan), possibly compromising accuracy. 2) I was not able to take into consideration co-founding factors such as LOH in cancer tissues, and thus the substitutions called may contain false positives. Future work will be focused on addressing those potential drawbacks.

In Chapter 5, I also found that the somatic mutation rates in blood derived normal cells varied considerably across 372 individuals. Although no correlation was detected between age and overall somatic mutation rates, the linear dependency of the rate of somatic mutation on age was observed in a subset of individuals that carry colon adenocarcinoma and show elevated mutation rates (mutator phenotype). The mutator phenotype is considered to be an

important initiating event during the initiation and progression of most of the sporadic cancers (Lawrence A. Loeb 2011). In addition, I identified a list of candidate germline variants that potentially predispose to increased somatic mutation rates. However, I failed to discover any gene that has been previously shown to be directly involved in the DNA repair pathways and most of these germline variants are expected to be false positives. Since this analysis was quite coarse and preliminary, future improvements, such as the inclusion of covariates and the detection of germline variants based on pooled data of 372 normal blood samples by GATK, which has been designed specifically for the detection of germline variants, should improve the power.

The observed mutational landscapes in different tumours are thought to be the result of the differential exposures to individual mutational processes operating in different cellular lineages (Nik-Zainal et al. 2012; Stephens et al. 2012; Alexandrov et al. 2013; Helleday et al. 2014), and the exposure rates of certain (but not all) mutational processes appear to show a linear association with age in cancer genomes (Alexandrov et al. 2013). Future work will focus on defining the particular mutational processes operating in the blood derived normal cells, their association with age, and further searches for genetic determinants of somatic mutation rates.

Together these analyses contribute to an integrated view of genome evolution, encompassing the divergence of DNA sequence and chromatin structure, and initial explorations of how they may interact in human disease.

Bibliography

Aarts E, Verhage M, Veenvliet JV, Dolan CV, van der Sluis S. 2014. A solution to dependency: using multilevel analysis to accommodate nested data. Nat. Neurosci. 17:491–496.

Ahn JW, Kim HS, Yoon J-K, Jang H, Han SM, Eun S, Shim HS, Kim H-J, Kim DJ, Lee JG, et al. 2014. Identification of somatic mutations in EGFR/KRAS/ALK-negative lung adenocarcinoma in never-smokers. Genome Med. 6:18.

Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. 2007. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature 446:572–576.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. 2013. Signatures of mutational processes in human cancer. Nature [Internet]. Available from: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12477.html

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. 3:246–259.

Allan J, Fraser RM, Owen-Hughes T, Keszenman-Pereyra D. 2012. Micrococcal nuclease does not substantially bias nucleosome mapping. J. Mol. Biol. 417:152–164.

Ananda G, Chiaromonte F, Makova KD. 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. Genome Biol. 12:R27.

Anderson JD, Widom J. 2001. Poly(dA-dT) Promoter Elements Increase the Equilibrium Accessibility of Nucleosomal DNA Target Sites. Mol. Cell. Biol. 21:3830–3839.

Anon. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9:e1001046.

Arnheim N, Calabrese P. 2009. Understanding what determines the frequency and pattern of human germline mutations. Nat. Rev. Genet. 10:478–488.

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat. Rev. Genet. 8:619–631.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat. Rev. Genet. 7:552–564.

Bai L, Ondracka A, Cross FR. 2011. Multiple Sequence-Specific Factors Generate the Nucleosome-Depleted Region on CLN2 Promoter. Mol. Cell 42:465–476.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129:823–837.

Bates DL, Thomas JO. 1981. Histones H1 and H5: one or two molecules per nucleosome? Nucleic Acids Res. 9:5883–5894.

Becker PB ed. 1999. Equilibrium and Dynamic Nucleosome Stability - Springer. In: Methods in Molecular Biology™. Humana Press. Available from: http://link.springer.com/protocol/10.1385%2F1-59259-681-9%3A61#page-2

Bell O, Tiwari VK, Thomä NH, Schübeler D. 2011. Determinants and dynamics of genome accessibility. Nat. Rev. Genet. 12:554–564.

Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL. 2004. Global nucleosome occupancy in yeast. Genome Biol. 5:R62.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. 28:1045–1048.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

Bondarenko VA, Steele LM, Újvári A, Gaykalova DA, Kulaeva OI, Polikanov YS, Luse DS, Studitsky VM. 2006. Nucleosomes Can Form a Polar Barrier to Transcript Elongation by RNA Polymerase II. Mol. Cell 24:469–479.

Brait M, Sidransky D. 2011. Cancer epigenetics: above and beyond. Toxicol. Mech. Methods 21:275–288.

Bram S, Ris H. 1971. On the structure of nucleohistone. J. Mol. Biol. 55:325–336.

Briggs S, Tomlinson I. 2013. Germline and somatic polymerase ϵ and δ mutations define a new class of hypermutated colorectal and endometrial cancers. J. Pathol. 230:148–153.

Brogaard K, Xi L, Wang J-P, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. Nature 486:496–501.

Calvert PM, Frucht H. 2002. The Genetics of Colorectal Cancer. Ann. Intern. Med. 137:603–612.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38:626–635.

Chaffer CL, Weinberg RA. 2011. A Perspective on Cancer Cell Metastasis. Science 331:1559–1564.

Chapman JR, Taylor MRG, Boulton SJ. 2012. Playing the end game: DNA double-strand break repair pathway choice. Mol. Cell 47:497–510.

Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet J-P, Ahmann GJ, Adli M, et al. 2011. Initial genome sequencing and analysis of multiple myeloma. Nature 471:467–472.

Chen W, Liu Y, Zhu S, Green CD, Wei G, Han J-DJ. 2014. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. Nat. Commun. 5:4909.

Chen X, Chen Z, Chen H, Su Z, Yang J, Lin F, Shi S, He X. 2012. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science 335:1235–1238.

Chodavarapu RK, Feng S, Bernatavichute YV, Chen P-Y, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, et al. 2010. Relationship between nucleosome positioning and DNA methylation. Nature 466:388–392.

Choi JK, Kim Y-J. 2009. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat. Genet. 41:498–503.

Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban H-J, Yoon D, Lee MH, Kim D-J, Park M, et al. 2009. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat. Genet. 41:527–534.

Chung H-R, Dunkel I, Heise F, Linke C, Krobitsch S, Ehrenhofer-Murray AE, Sperling SR, Vingron M. 2010. The Effect of Micrococcal Nuclease Digestion on Nucleosome Positioning Data. PLoS ONE 5:e15754.

Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS. 2002. Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4. Mol. Cell 9:279–289.

Clapier CR, Cairns BR. 2009. The Biology of Chromatin Remodeling Complexes. Annu. Rev. Biochem. 78:273–304.

Clark KL, Halay ED, Lai E, Burley SK. 1993. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature 364:412–420.

Cohen NM, Kenigsberg E, Tanay A. 2011. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell 145:773–786.

Collings CK, Waddell PJ, Anderson JN. 2013. Effects of DNA methylation on nucleosome stability. Nucleic Acids Res. 41:2918–2931.

Colot V, Rossignol JL. 1999. Eukaryotic DNA methylation as an evolutionary device. BioEssays News Rev. Mol. Cell. Dev. Biol. 21:402–411.

Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al. 2011. Variation in genome-wide mutation rates within and between human families. Nat. Genet. 43:712–714.

Consortium T 1000 GP. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Consortium TEP. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640.

Consortium TEP. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Crespi B, Summers K. 2005. Evolutionary biology of cancer. Trends Ecol. Evol. 20:545–552.

Crow JF. 2000. The origins, patterns and implications of human spontaneous mutation. Nat. Rev. Genet. 1:40–47.

Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. 19:24–32.

Cui F, Zhurkin VB. 2009. Distinctive sequence patterns in metazoan and yeast nucleosomes: implications for linker histone binding to AT-rich and methylated DNA. Nucleic Acids Res. 37:2818–2829.

Cutter AR, Hayes JJ. 2015. A brief review of nucleosome structure. FEBS Lett.

Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ. 2002. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 Å Resolution†. J. Mol. Biol. 319:1097–1113.

Davuluri RV, Grosse I, Zhang MQ. 2001. Computational identification of promoters and first exons in the human genome. Nat. Genet. 29:412–417.

Dehal P, Boore JL. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLoS Biol 3:e314.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. 14:390–403.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast Computation and Applications of Genome Mappability. PLoS ONE 7:e30377.

Dietlein F, Thelen L, Reinhardt HC. 2014. Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches. Trends Genet. 30:326–339.

Don PK, Ananda G, Chiaromonte F, Makova KD. 2013. Segmenting the human genome based on states of neutral genetic divergence. Proc. Natl. Acad. Sci. 110:14699–14704.

Downs JA, Kosmidou E, Morgan A, Jackson SP. 2003. Suppression of homologous recombination by the Saccharomyces cerevisiae linker histone. Mol. Cell 11:1685–1692.

Drew HR, Travers AA. 1985. DNA bending and its relation to nucleosome positioning. J. Mol. Biol. 186:773–790.

Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E. 2012. Genomics: ENCODE explained. Nature 489:52–55.

Eltsov M, Maclellan KM, Maeshima K, Frangakis AS, Dubochet J. 2008. Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. Proc. Natl. Acad. Sci. U. S. A. 105:19732–19737.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. 28:817–825.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473:43–49.

Esteller M. 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat. Rev. Genet. 8:286–298.

Fan X, Moqtaderi Z, Jin Y, Zhang Y, Liu XS, Struhl K. 2010. Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. Proc. Natl. Acad. Sci. 107:17945–17950.

Fan Y, Nikitina T, Zhao J, Fleury TJ, Bhattacharyya R, Bouhassira EE, Stein A, Woodcock CL, Skoultchi AI. 2005. Histone H1 Depletion in Mammals Alters Global Chromatin Structure but Causes Specific Changes in Gene Regulation. Cell 123:1199–1212.

Farrington SM, Tenesa A, Barnetson R, Wiltshire A, Prendergast J, Porteous M, Campbell H, Dunlop MG. 2005. Germline susceptibility to colorectal cancer due to base-excision repair gene defects. Am. J. Hum. Genet. 77:112–119.

Finch JT, Lutter LC, Rhodes D, Brown RS, Rushton B, Levitt M, Klug A. 1977. Structure of nucleosome core particles of chromatin. Nature 269:29–36.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Fraga MF, Ballestar E, Villar-Garea A, Boix-Chornet M, Espada J, Schotta G, Bonaldi T, Haydon C, Ropero S, Petrie K, et al. 2005. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nat. Genet. 37:391–400.

Freidkin I, Katcoff DJ. 2001. Specific distribution of the Saccharomyces cerevisiae linker histone homolog HHO1p in the chromatin. Nucleic Acids Res. 29:4043–4051.

Fudenberg G, Mirny LA. 2012. Higher-order chromatin structure: bridging physics and biology. Curr. Opin. Genet. Dev. 22:115–124.

Fussner E, Ching RW, Bazett-Jones DP. 2011. Living without 30nm chromatin fibers. Trends Biochem. Sci. 36:1–6.

Fu Y, Sinha M, Peterson CL, Weng Z. 2008. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. 4:e1000138.

Gaffney DJ, Keightley PD. 2005. The scale of mutational variation in the murid genome. Genome Res. 15:1086–1094.

Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK. 2012. Controls of nucleosome positioning in the human genome. PLoS Genet. 8:e1003036.

Garg P, Burgers PMJ. 2005. DNA Polymerases that Propagate the Eukaryotic DNA Replication Fork. Crit. Rev. Biochem. Mol. Biol. 40:115–128.

Gazave E, Marqués-Bonet T, Fernando O, Charlesworth B, Navarro A. 2007. Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol. 8:R21.

Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. 366:883–892.

Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA. 2004. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell 118:555–566.

Global Lipids Genetics Consortium. 2013. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45:1274–1283.

Gojobori T, Li WH, Graur D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18:360–369.

Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. 2013. IntOGen-mutations identifies cancer drivers across tumor types. Nat. Methods 10:1081–1082.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al. 2007. Patterns of somatic mutation in human cancer genomes. Nature 446:153–158.

Groemping U, Matthias L. 2013. relaimpo: Relative importance of regressors in linear models. Available from: https://cran.r-project.org/web/packages/relaimpo/index.html

Harris RS. 2008. Improved pairwise alignment of genomic DNA. Available from: http://gradworks.umi.com/32/99/3299002.html

Hartlerode AJ, Scully R. 2009. Mechanisms of double-strand break repair in somatic mammalian cells. Biochem. J. 423:157–168.

Hassan AH, Prochasson P, Neely KE, Galasinski SC, Chandy M, Carrozza MJ, Workman JL. 2002. Function and Selectivity of Bromodomains in Anchoring Chromatin-Modifying Complexes to Promoter Nucleosomes. Cell 111:369–379.

Hassler MR, Egger G. 2012. Epigenomics of cancer – emerging new concepts. Biochimie 94:2219–2230.

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat. Genet. 39:1140–1144.

Hecht SS. 1999. Tobacco Smoke Carcinogens and Lung Cancer. J. Natl. Cancer Inst. 91:1194–1210.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. 39:311–318.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. HOMER program. Mol. Cell 38:576–589.

Heitzer E, Tomlinson I. 2014. Replicative DNA polymerase mutations in cancer. Curr. Opin. Genet. Dev. 24:107–113.

Helleday T, Eshtad S, Nik-Zainal S. 2014. Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. [Internet]. Available from: http://www.nature.com.ezproxy.is.ed.ac.uk/nrg/journal/vaop/ncurrent/full/nrg3729.html

Helleday T, Petermann E, Lundin C, Hodgson B, Sharma RA. 2008. DNA repair pathways as targets for cancer therapy. Nat. Rev. Cancer 8:193–204.

Hellmann I, Prüfer K, Ji H, Zody MC, Pääbo S, Ptak SE. 2005. Why do human diversity levels vary at a megabase scale? Genome Res. 15:1222–1231.

Heng HHQ, Stevens JB, Bremer SW, Liu G, Abdallah BY, Ye CJ. 2011. Evolutionary mechanisms and diversity in cancer. Adv. Cancer Res. 112:217–253.

Heng HHQ, Stevens JB, Bremer SW, Ye KJ, Liu G, Ye CJ. 2010. The evolutionary mechanism of cancer. J. Cell. Biochem. 109:1072–1084.

Higasa K, Hayashi K. 2006. Periodicity of SNP distribution around transcription start sites. BMC Genomics 7:66.

Hodgkinson A, Chen Y, Eyre-Walker A. 2012. The large-scale distribution of somatic mutations in cancer genomes. Hum. Mutat. 33:136–143.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12:756–766.

Hong L, Schroth GP, Matthews HR, Yau P, Bradbury EM. 1993. Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 "tail" to DNA. J. Biol. Chem. 268:305–314.

Hughes AL, Jin Y, Rando OJ, Struhl K. 2012. A Functional Evolutionary Approach to Identify Determinants of Nucleosome Positioning: A Unifying Model for Establishing the Genome-wide Pattern. Mol. Cell [Internet]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22885008

Hughes AL, Rando OJ. 2014. Mechanisms Underlying Nucleosome Positioning in vivo. Annu. Rev. Biophys. 43:null.

Hughes A, Rando OJ. 2009. Chromatin "programming" by sequence - is there more to the nucleosome code than %GC? J. Biol. 8:96.

Hu G, Schones DE, Cui K, Ybarra R, Northrup D, Tang Q, Gattinoni L, Restifo NP, Huang S, Zhao K. 2011. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. Genome Res. 21:1650–1658.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. U. S. A. 101:13994–14001.

ICBP TIC for BP. 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478:103–109.

Igo-Kemenes T, Horz W, Zachau HG. 1982. Chromatin. Annu. Rev. Biochem. 51:89–121.

Ioshikhes I, Bolshoy A, Derenshteyn K, Borodovsky M, Trifonov EN. 1996. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J. Mol. Biol. 262:129–139.

Jansen A, Verstrepen KJ. 2011. Nucleosome Positioning in Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev. 75:301–320.

Je EM, Choi YJ, Chung YJ, Yoo NJ, Lee SH. 2015. TEAD2, a Hippo pathway gene, is somatically mutated in gastric and colorectal cancers with high microsatellite instability. APMIS 123:359–360.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. 10:161–172.

Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G. 2009a. H3.3/H2A.Z double variant-containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions. Nat. Genet. 41:941–945.

Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G. 2009b. H3.3/H2A.Z double variant–containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions. Nat. Genet. 41:941–945.

Johnson PLF, Hellmann I. 2011. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. Genome Biol. Evol. 3:842–850.

Jones S, Chen W-D, Parmigiani G, Diehl F, Beerenwinkel N, Antal T, Traulsen A, Nowak MA, Siegel C, Velculescu VE, et al. 2008. Comparative lesion sequencing provides insights into tumor evolution. Proc. Natl. Acad. Sci. U. S. A. 105:4283–4288.

Kamileri I, Karakasilioti I, Garinis GA. 2012. Nucleotide excision repair: new tricks with old bricks. Trends Genet. 28:566–573.

Kaplan N, Moore I, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, Hughes TR, Lieb JD, Widom J, Segal E. 2010. Nucleosome sequence preferences influence in vivo nucleosome organization. Nat. Struct. Mol. Biol. 17:918–920.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2008. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458:362–366.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458:362–366.

Kassabov SR, Zhang B, Persinger J, Bartholomew B. 2003. SWI/SNF Unwraps, Slides, and Rewraps the Nucleosome. Mol. Cell 11:391–403.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428:617–624.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. 111:6131–6138.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D. 2002. The Human Genome Browser at UCSC. Genome Res. 12:996–1006.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge [Cambridgeshire]; New York: Cambridge University Press

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell 155:27–38.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22:568–576.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. Nature 488:471–475.

Kornberg R. 1981. The location of nucleosomes in chromatin: specific or statistical. Nature 292:579–580.

Kornberg RD. 1974. Chromatin structure: a repeating unit of histones and DNA. Science 184:868–871.

Kornberg RD. 1977. Structure of Chromatin. Annu. Rev. Biochem. 46:931–954.

Kornberg RD, Stryer L. 1988. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. Nucleic Acids Res. 16:6677–6690.

Kornberg RD, Thomas JO. 1974. Chromatin structure; oligomers of the histones. Science 184:865–868.

Kostka D, Hahn MW, Pollard KS. 2010. Noncoding Sequences Near Duplicated Genes Evolve Rapidly. Genome Biol. Evol. 2:518–533.

Koszul R, Caburet S, Dujon B, Fischer G. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J. 23:234–243.

Kulaeva OI, Hsieh F-K, Studitsky VM. 2010. RNA polymerase complexes cooperate to relieve the nucleosomal barrier and evict histones. Proc. Natl. Acad. Sci. 107:11325–11330.

Kuschel B, Auranen A, McBride S, Novik KL, Antoniou A, Lipscombe JM, Day NE, Easton DF, Ponder BAJ, Pharoah PDP, et al. 2002. Variants in DNA double-strand break repair genes and breast cancer susceptibility. Hum. Mol. Genet. 11:1399–1407.

LangkjAEr RB, Cliften PF, Johnston M, Piškur J. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature 421:848–852.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

Längst G, Manelyte L. 2015. Chromatin Remodelers: From Function to Dysfunction. Genes 6:299–324.

Lantermann AB, Straub T, Strålfors A, Yuan G-C, Ekwall K, Korber P. 2010. Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. Nat. Struct. Mol. Biol. 17:251–257.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214–218.

Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat. Genet. 36:900–905.

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, et al. 2010. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 465:473–477.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. 2007. A high-resolution atlas of nucleosome occupancy in yeast. Nat. Genet. 39:1235–1244.

Lee YY, Yu YB, Gunawardena HP, Xie L, Chen X. 2012. BCLAF1 is a radiation-induced H2AX-interacting partner involved in γH2AX-mediated regulation of apoptosis and DNA repair. Cell Death Dis. 3:e359.

Li B, Carey M, Workman JL. 2007. The Role of Chromatin during Transcription. Cell 128:707–719.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Lindeman R, Merenda PF. 1980. Introduction to Bivariate and Multivariate Analysis. Glenview, Ill.: Longman Higher Education

Lin JC, Jeong S, Liang G, Takai D, Fatemi M, Tsai YC, Egger G, Gal-Yam EN, Jones PA. 2007. Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island. Cancer Cell 12:432–444.

Liu L, De S, Michor F. 2013. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nat. Commun. 4:1502.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of Next-Generation Sequencing Systems. BioMed Res. Int. 2012:e251364.

Loeb LA. 2011. Human cancers express mutator phenotypes: origin, consequences and targeting. Nat. Rev. Cancer 11:450–457.

Loeb LA. 2011. Human cancers express mutator phenotypes: origin, consequences and targeting. Nat. Rev. Cancer 11:450–457.

Lorente-Galdos B, Bleyhl J, Santpere G, Vives L, Ramírez O, Hernandez J, Anglada R, Cooper GM, Navarro A, Eichler EE, et al. 2013. Accelerated exon evolution within primate segmental duplications. Genome Biol. 14:R9.

Lowary PT, Widom J. 1997. Nucleosome packaging and nucleosome positioning of genomic DNA. Proc. Natl. Acad. Sci. U. S. A. 94:1183–1188.

Lowary PT, Widom J. 1998. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J. Mol. Biol. 276:19–42.

Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389:251–260.

Lunter G, Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21:936–939.

Luzzati V, Nicolaieff A, Masson F. 1961. [Structure of desoxyribonucleic acid in solution. Study by the diffusion of x-rays at small angles]. J. Mol. Biol. 3:185–201.

Maeshima K, Eltsov M. 2008. Packaging the genome: the structure of mitotic chromosomes. J. Biochem. (Tokyo) 143:145–153.

Maeshima K, Hihara S, Eltsov M. 2010. Chromatin structure: does the 30-nm fibre exist in vivo? Curr. Opin. Cell Biol. 22:291–297.

Makova KD, Hardison RC. 2015. The effects of chromatin organization on variation in mutation rates in the genome. Nat. Rev. Genet. 16:213–223.

Mardis ER. 2008. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9:387–402.

Marfella CGA, Imbalzano AN. 2007. The Chd family of chromatin remodelers. Mutat. Res. Mol. Mech. Mutagen. 618:30–40.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457:877–881.

Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. 18:1073–1083.

McCall M, Brown T, Kennard O. 1985. The crystal structure of d(G-G-G-G-C-C-C-C) a model for poly(dG) · poly(dC). J. Mol. Biol. 183:385–396.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. 2012. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. Cell 151:1431–1442.

Moyle-Heyrman G, Zaichuk T, Xi L, Zhang Q, Uhlenbeck OC, Holmgren R, Widom J, Wang J-P. 2013. Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning. Proc. Natl. Acad. Sci.:201315809.

Musselman CA, Lalonde M-E, Côté J, Kutateladze TG. 2012. Perceiving the epigenetic landscape through histone readers. Nat. Struct. Mol. Biol. 19:1218–1227.

Nelson HC, Finch JT, Luisi BF, Klug A. 1987. The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature 330:221–226.

Network TCGA. 2012. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487:330–337.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. 2012. Mutational processes molding the genomes of 21 breast cancers. Cell 149:979–993.

Nishimura S. 2006. 8-Hydroxyguanine: From its discovery in 1983 to the present status. Proc. Jpn. Acad. Ser. B Phys. Biol. Sci. 82:127–141.

Nouspikel T. 2009. DNA repair in mammalian cells : Nucleotide excision repair: variations on versatility. Cell. Mol. Life Sci. CMLS 66:994–1009.

Olins AL, Olins DE. 1974. Spheroid chromatin units (v bodies). Science 183:330–332.

Osmotherly LM. 2010. Structural and biochemical studies of the yeast linker histone, Hho1p. Available from: http://www.repository.cam.ac.uk/handle/1810/228702

Ozsolak F, Song JS, Liu XS, Fisher DE. 2007. High-throughput mapping of the chromatin structure of human promoters. Nat. Biotechnol. 25:244–248.

Palles C, Cazier J-B, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Guarino Almeida E, et al. 2013. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat. Genet. 45:136–144.

Papamichos-Chronakis M, Watanabe S, Rando OJ, Peterson CL. 2011. Global Regulation of H2A.Z Localization by the INO80 Chromatin-Remodeling Enzyme Is Essential for Genome Integrity. Cell 144:200–213.

Papp B, Pál C, Hurst LD. 2003. Evolution of cis-regulatory elements in duplicated genes of yeast. Trends Genet. 19:417–422.

Pardon JF, Wilkins MH. 1972. A super-coil model for nucleohistone. J. Mol. Biol. 68:115–124.

Peltomäki P. 2001. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. Hum. Mol. Genet. 10:735–740.

Ping Y, Deng Y, Wang L, Zhang H, Zhang Y, Xu C, Zhao H, Fan H, Yu F, Xiao Y, et al. 2015. Identifying core gene modules in glioblastoma based on multilayer factor-mediated dysfunctional regulatory networks through integrating multi-dimensional genomic data. Nucleic Acids Res.:gkv074.

Plass C, Pfister SM, Lindroth AM, Bogatyrova O, Claus R, Lichter P. 2013. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. Nat. Rev. Genet. 14:765–780.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin M-L, Beare D, Lau KW, Greenman C, et al. 2010. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463:184–190.

Pohlert T. 2015. PMCMR: Calculate Pairwise Multiple Comparisons of Mean Rank Sums. Available from: https://cran.r-project.org/web/packages/PMCMR/index.html

Polach KJ, Widom J. 1995. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. J. Mol. Biol. 254:130–149.

Polak P, Arndt PF. 2008. Transcription induces strand-specific mutations at the 5' end of human genes. Genome Res. 18:1216–1223.

Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA, et al. 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518:360–364.

Pon JR, Marra MA. 2015. Driver and Passenger Mutations in Cancer. Annu. Rev. Pathol. Mech. Dis. 10:25–50.

Portela A, Esteller M. 2010. Epigenetic modifications and human disease. Nat. Biotechnol. 28:1057–1068.

Pray-Grant MG, Daniel JA, Schieltz D, Yates JR, Grant PA. 2005. Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature 433:434–438.

Prendergast JGD, Campbell H, Gilbert N, Dunlop MG, Bickmore WA, Semple CAM. 2007. Chromatin structure and evolution in the human genome. BMC Evol. Biol. 7:72.

Prendergast JGD, Chambers EV, Semple CAM. 2014. Sequence level mechanisms of human epigenome evolution. Genome Biol. Evol.

Prendergast JGD, Semple CAM. 2011. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. Genome Res. 21:1777–1787.

Pugh BF. 2010. A preoccupied position on nucleosomes. Nat. Struct. Mol. Biol. 17:923–923.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Radman-Livaja M, Rando OJ. 2010. Nucleosome positioning: How is it established, and why does it matter? Dev. Biol. 339:258–266.

Ramakrishnan V, Finch JT, Graziano V, Lee PL, Sweet RM. 1993. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. Nature 362:219–223.

Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M, Smale ST. 2009. A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling. Cell 138:114–128.

Rando OJ, Winston F. 2012. Chromatin and Transcription in Yeast. Genetics 190:351–387.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. 5:28.

Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, Zeevi D, Sharon E, Weinberger A, Segal E. 2012. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat. Genet. 44:743–750.

R Development Core Team. 2009. {R: A language and environment for statistical computing}. Vienna, Austria: R Foundation for Statistical Computing

Reynolds SM, Bilmes JA, Noble WS. 2010. Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in Saccharomyces cerevisiae and Homo sapiens. PLoS Comput. Biol. 6:e1000834.

Richmond TJ, Davey CA. 2003. The structure of DNA in the nucleosome core. Nature 423:145–150.

Richmond TJ, Finch JT, Rushton B, Rhodes D, Klug A. 1984. Structure of the nucleosome core particle at 7 |[angst]| resolution. Nature 311:532–537.

Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez A, Lopez-Bigas N. 2015. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. Cancer Cell 27:382–396.

Salk JJ, Fox EJ, Loeb LA. 2010. Mutational heterogeneity in human cancers: origin and consequences. Annu. Rev. Pathol. 5:51–75.

Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S-I, Ogawa M, Matsushima K, Gu SG, Kasahara M, Ahsan B, et al. 2009. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science 323:401–404.

Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF. 2011. Exome Sequencing-Based Copy-Number Variation and Loss of Heterozygosity Detection: ExomeCNV. Bioinformatics [Internet]. Available from: http://bioinformatics.oxfordjournals.org/content/early/2011/08/09/bioinformatics.btr462

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U. S. A. 103:1412–1417.

Schlegel RA, Haye KR, Litwack AH, Phelps BM. 1980. Nucleosome repeat lengths in the definitive erythroid series of the adult chicken. Biochim. Biophys. Acta 606:316–330.

Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell 132:887–898.

Schuster-Böckler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature [Internet]. Available from: http://www.nature.com.ezproxy.webfeat.lib.ed.ac.uk/nature/journal/vaop/ncurrent/full/nature11273.html

Scipioni A, Pisano S, Anselmi C, Savino M, De Santis P. 2004. Dual role of sequence-dependent DNA curvature in nucleosome stability: the critical test of highly bent Crithidia fasciculata DNA tract. Biophys. Chem. 107:7–17.

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J. 2006. A genomic code for nucleosome positioning. Nature 442:772–778.

Segal E, Widom J. 2009. Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr. Opin. Struct. Biol. 19:65–71.

Ségurel L, Wyman MJ, Przeworski M. 2014. Determinants of Mutation Rate Variation in the Human Germline. Annu. Rev. Genomics Hum. Genet. 15:47–70.

Sekinger EA, Moqtaderi Z, Struhl K. 2005. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. Mol. Cell 18:735–748.

Semple CAM, Taylor MS. 2009. Molecular biology. The structure of change. Science 323:347–348.

She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark R, Graves T, Gulden CL, Alkan C, et al. 2004. The structure and evolution of centromeric transition regions within the human genome. Nature 430:857–864.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. U. S. A. 100:15776–15781.

Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. 2008. Dynamic Remodeling of Individual Nucleosomes Across a Eukaryotic Genome in Response to Transcriptional Perturbation. PLOS Biol 6:e65.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J 'an, Mägi R, et al. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. 42:937–948.

Spencer CCA, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The Influence of Recombination on Human Genetic Diversity. PLoS Genet 2:e148.

van Steensel B. 2011. Chromatin: constructing the big picture. EMBO J. 30:1885–1895.

Stein A, Takasuka TE, Collings CK. 2010. Are nucleosome positions in vivo primarily determined by histone–DNA sequence preferences? Nucleic Acids Res. 38:709–719.

Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR, et al. 2012. The landscape of cancer genes and mutational processes in breast cancer. Nature 486:400–404.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. J. Mol. Evol. 49:169–181.

Stormo GD. 2000. DNA binding sites: representation and discovery. Bioinformatics 16:16–23.

Stranger BE, Stahl EA, Raj T. 2011. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait. Genetics 187:367–383.

Strohner R, Németh A, Nightingale KP, Grummt I, Becker PB, Längst G. 2004. Recruitment of the Nucleolar Remodeling Complex NoRC Establishes Ribosomal DNA Silencing in Chromatin. Mol. Cell. Biol. 24:1791–1798.

Struhl K, Segal E. 2013. Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. 20:267–273.

Studitsky VM, Clark DJ, Felsenfeld G. 1994. A histone octamer can step around a transcribing polymerase without leaving the template. Cell 76:371–382.

Studitsky VM, Kassavetis GA, Geiduschek EP, Felsenfeld G. 1997. Mechanism of Transcription Through the Nucleosome by Eukaryotic RNA Polymerase. Science 278:1960–1963.

Supek F, Lehner B. 2015. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature [Internet]. Available from: http://www.nature.com.ezproxy.is.ed.ac.uk/nature/journal/vaop/ncurrent/full/nature14173.html?WT.ec_id=NATURE-20150226

Supek F, Lehner B, Hajkova P, Warnecke T. 2014. Hydroxymethylated Cytosines Are Associated with Elevated C to G Transversion Rates. PLoS Genet 10:e1004585.

Suter B, Schnappauf G, Thoma F. 2000. Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res. 28:4083–4089.

Talavera D, Taylor MS, Thornton JM. 2010. The (non)malignancy of cancerous amino acidic substitutions. Proteins 78:518–529.

Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ. 2007. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. Nat. Struct. Mol. Biol. 14:1025–1040.

Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CAM. 2006. Heterotachy in mammalian promoter evolution. PLoS Genet. 2:e30.

Taylor MS, Massingham T, Hayashizaki Y, Carninci P, Goldman N, Semple CAM. 2008. Rapidly evolving human promoter regions. Nat. Genet. 40:1262–1263.

Teer JK, Mullikin JC. 2010. Exome sequencing: the sweet spot before whole genomes. Hum. Mol. Genet. 19:R145–R151.

Tessarz P, Kouzarides T. 2014. Histone core modifications regulating nucleosome structure and dynamics. Nat. Rev. Mol. Cell Biol. 15:703–708.

Thåström A, Bingham LM, Widom J. 2004. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. J. Mol. Biol. 338:695–709.

Thåström A, Lowary PT, Widlund HR, Cao H, Kubista M, Widom J. 1999. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences1. J. Mol. Biol. 288:213–229.

The FANTOM Consortium and the RIKEN PMI and CLST (dgt). 2014. A promoter-level mammalian expression atlas. Nature 507:462–470.

Thoma F, Koller T, Klug A. 1979. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. J. Cell Biol. 83:403–427.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. Nature 489:75–82.

Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen J-Q. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455:105–108.

Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 10:442.

Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Field Y, Lieb JD, Widom J, Segal E, Hughes TR. 2010. High Nucleosome Occupancy Is Encoded at Human Regulatory Sequences. PLoS ONE 5:e9129.

Tims HS, Gurunathan K, Levitus M, Widom J. 2011. Dynamics of Nucleosome Invasion by DNA Binding Proteins. J. Mol. Biol. 411:430–448.

Tirosh I, Barkai N. 2008. Two strategies for gene regulation by promoter nucleosomes. Genome Res. 18:1084–1091.

Tirosh I, Sigal N, Barkai N. 2010. Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences. Mol. Syst. Biol. 6:365.

Tollefsbol TO ed. 2011. Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications - Springer. In: Methods in Molecular Biology. Humana Press. Available from: http://link.springer.com.ezproxy.is.ed.ac.uk/protocol/10.1007%2F978-1-61779-316-5_20

Tolstorukov MY, Volfovsky N, Stephens RM, Park PJ. 2011. Impact of chromatin structure on sequence variability in the human genome. Nat. Struct. Mol. Biol. 18:510–515.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13:146–146.

Tremethick DJ. 2007. Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber. Cell 128:651–654.

Tuszynski J. 2014. caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc. Available from: https://cran.r-project.org/web/packages/caTools/index.html

Ulrike G. 2006. Relative Importance for Linear Regression in R: The Package relaimpo. J. Stat. Softw.

Vaillant C, Palmeira L, Chevereau G, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2010. A novel strategy of transcription regulation by intragenic nucleosome ordering. Genome Res. 20:59–67.

Valastyan S, Weinberg RA. 2011. Tumor Metastasis: Molecular Insights and Evolving Paradigms. Cell 147:275–292.

Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. Nature 474:516–520.

Van Bortle K, Corces VG. 2012. Nuclear Organization and Genome Function. Annu. Rev. Cell Dev. Biol. 28:163–187.

Vavouri T, Lehner B. 2012. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. Genome Biol. 13:R110.

Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. New York, NY: Springer New York Available from: http://link.springer.com/10.1007/978-0-387-21706-2

Verzi MP, Shin H, He HH, Sulahian R, Meyer CA, Montgomery RK, Fleet JC, Brown M, Liu XS, Shivdasani RA. 2010. Differentiation-specific histone modifications reveal dynamic chromatin interactions and alternative partners for the intestinal transcription factor CDX2. Dev. Cell 19:713–726.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five Years of GWAS Discovery. Am. J. Hum. Genet. 90:7–24.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38:e164–e164.

Washietl S, Machné R, Goldman N. 2008. Evolutionary footprints of nucleosome positions in yeast. Trends Genet. TIG 24:583–587.

Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N. 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res. 20:90–100.

West JA, Cook A, Alver BH, Stadtfeld M, Deaton AM, Hochedlinger K, Park PJ, Tolstorukov MY, Kingston RE. 2014. Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. Nat. Commun. [Internet] 5. Available from: http://www.nature.com/ncomms/2014/140827/ncomms5719/full/ncomms5719.html

Wickham H. 2014. reshape: Flexibly reshape data. Available from: https://cran.r-project.org/web/packages/reshape/index.html

Wickham H. 2015. plyr: Tools for Splitting, Applying and Combining Data. Available from: https://cran.r-project.org/web/packages/plyr/index.html

Wickham H, Chang W. 2015. ggplot2: An Implementation of the Grammar of Graphics. Available from: https://cran.r-project.org/web/packages/ggplot2/index.html

Widom J. 2001. Role of DNA sequence in nucleosome stability and dynamics. Q. Rev. Biophys. 34:269–324.

Wilson BG, Roberts CWM. 2011. SWI/SNF nucleosome remodellers and cancer. Nat. Rev. Cancer 11:481–492.

Wilson Sayres MA, Makova KD. 2011. Genome analyses substantiate male mutation bias in many species. BioEssays 33:938–945.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713.

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J 'an, Kutalik Z, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 46:1173–1186.

Xing K, He X. 2015. Mutation bias, rather than binding preference, underlies the nucleosome-associated G+C% variation in eukaryotes. Genome Biol. Evol.:evv053.

Xu F, Zhang K, Grunstein M. 2005. Acetylation in Histone H3 Globular Domain Regulates Gene Expression in Yeast. Cell 121:375–385.

Ye CJ, Stevens JB, Liu G, Bremer SW, Jaiswal AS, Ye KJ, Lin M-F, Lawrenson L, Lancaster WD, Kurkinen M, et al. 2009. Genome based cell population heterogeneity promotes tumorigenicity: The evolutionary mechanism of cancer. J. Cell. Physiol. 219:288–300.

Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. 2005. Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309:626–630.

Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. Genes Dev. 25:2227–2241.

Zentner GE, Henikoff S. 2013. Regulation of nucleosome dynamics by histone modifications. Nat. Struct. Mol. Biol. 20:259–266.

Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K. 2009. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat. Struct. Mol. Biol. 16:847–852.

Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K. 2010. Evidence against a genomic code for nucleosome positioning. Nat. Struct. Mol. Biol. 17:920–923.

Zhang Y, Shin H, Song JS, Lei Y, Liu XS. 2008. Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq. BMC Genomics 9:537.

Zhang Z, Gerstein M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. 31:5338–5348.

Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, Pugh BF. 2011. A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome. Science 332:977–980.

Zheng D. 2008. Asymmetric histone modifications between the original and derived loci of human segmental duplications. Genome Biol. 9:R105.

Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, et al. 2007. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. Genome Res. 17:839–851.

Zhou X, Li X, Cheng Y, Wu W, Xie Z, Xi Q, Han J, Wu G, Fang J, Feng Y. 2014. BCLAF1 and its splicing regulator SRSF10 regulate the tumorigenic potential of colon cancer cells. Nat. Commun. 5:4581.

Zhurkin VB. 1985. Sequence-Dependent Bending of DNA and Phasing of Nucleosomes. J. Biomol. Struct. Dyn. 2:785–804.

Appendix

The divergence in the occupancy of dyads around 242 motifs in human genome

Motif	Dataset	Motif maintained		Motif lost		Student's *t*-test (p value)		Mann-Whitney test (p value)	
		Mean	95% CI	Mean	95% CI	Raw	Adjusted	Raw	Adjusted
AMYB (HTH)	*in vitro*	0.049	[-0.113, 0.21]	0.071	[-0.017, 0.159]	0.82152	1.00000	0.89670	1.00000
AMYB (HTH)	*in vivo*	-0.359	[-0.684, -0.034]	-0.148	[-0.294, -0.002]	0.24442	1.00000	0.14094	1.00000
AP-1 (bZIP)	*in vitro*	-0.369	[-0.648, -0.09]	-0.074	[-0.189, 0.042]	0.04454	1.00000	0.01454	1.00000
AP-1 (bZIP)	*in vivo*	-0.333	[-0.668, 0.003]	-0.21	[-0.443, 0.023]	0.56150	1.00000	0.27542	1.00000
AP-2alpha (AP2)	*in vitro*	0.291	[0.042, 0.539]	0.252	[0.125, 0.379]	0.78916	1.00000	0.50042	1.00000
AP-2alpha (AP2)	*in vivo*	-0.15	[-0.497, 0.198]	0.161	[-0.061, 0.383]	0.11871	1.00000	0.17668	1.00000
AP-2gamma (AP2)	*in vitro*	0.256	[0.033, 0.48]	0.213	[0.112, 0.315]	0.70431	1.00000	0.56097	1.00000
AP-2gamma (AP2)	*in vivo*	-0.3	[-0.645, 0.045]	0.15	[-0.107, 0.408]	0.01227	1.00000	0.01936	1.00000
Ap4 (bHLH)	*in vitro*	-0.243	[-0.563, 0.077]	0.075	[-0.113, 0.264]	0.10440	1.00000	0.14094	1.00000
Ap4 (bHLH)	*in vivo*	-0.259	[-0.546, 0.028]	-0.022	[-0.269, 0.224]	0.22201	1.00000	0.21568	1.00000
AR-halfsite (NR)	*in vitro*	0.164	[0.023, 0.304]	0.151	[0.033, 0.27]	0.89868	1.00000	0.91919	1.00000
AR-halfsite (NR)	*in vivo*	-0.522	[-0.872, -0.171]	-0.09	[-0.284, 0.104]	0.03135	1.00000	0.02965	1.00000
ARE (NR)	*in vitro*	0.012	[-0.374, 0.399]	0.24	[0.088, 0.391]	0.27213	1.00000	0.06840	1.00000
ARE (NR)	*in vivo*	0.42	[0.046, 0.793]	0.452	[0.201, 0.704]	0.88149	1.00000	0.44418	1.00000
Arnt:Ahr (bHLH)	*in vitro*	0.242	[0.094, 0.39]	0.105	[0.008, 0.203]	0.14617	1.00000	0.03165	1.00000
Arnt:Ahr (bHLH)	*in vivo*	0.047	[-0.221, 0.316]	-0.015	[-0.203, 0.174]	0.70088	1.00000	0.76786	1.00000
Atf1 (bZIP)	*in vitro*	-0.064	[-0.305, 0.177]	-0.08	[-0.242, 0.082]	0.91861	1.00000	0.93385	1.00000

Atf1 (bZIP)	*in vivo*	-0.194	[-0.588, 0.199]	-0.176	[-0.399, 0.047]	0.92122	1.00000	0.81763	1.00000
Atf3 (bZIP)	*in vitro*	0.573	[0.259, 0.887]	0.609	[0.376, 0.843]	0.84802	1.00000	0.49540	1.00000
Atf3 (bZIP)	*in vivo*	1.637	[1.474, 1.8]	1.739	[1.557, 1.92]	0.38986	1.00000	0.35599	1.00000
Atf4 (bZIP)	*in vitro*	-0.142	[-0.342, 0.058]	-0.106	[-0.228, 0.015]	0.77484	1.00000	0.41116	1.00000
Atf4 (bZIP)	*in vivo*	-0.338	[-0.637, -0.04]	-0.188	[-0.434, 0.057]	0.38865	1.00000	0.16503	1.00000
Atoh1 (bHLH)	*in vitro*	0.03	[-0.199, 0.258]	0.287	[0.17, 0.403]	0.05916	1.00000	0.03568	1.00000
Atoh1 (bHLH)	*in vivo*	0.102	[-0.159, 0.362]	0.034	[-0.173, 0.24]	0.66798	1.00000	0.77728	1.00000
Bach1 (bZIP)	*in vitro*	0.338	[0.069, 0.607]	0.303	[0.106, 0.5]	0.85245	1.00000	0.92979	1.00000
Bach1 (bZIP)	*in vivo*	0.946	[0.635, 1.257]	0.964	[0.744, 1.184]	0.91649	1.00000	0.51021	1.00000
Bach2 (bZIP)	*in vitro*	-0.11	[-0.349, 0.128]	0	[-0.157, 0.157]	0.44459	1.00000	0.19257	1.00000
Bach2 (bZIP)	*in vivo*	-0.033	[-0.386, 0.32]	0.432	[0.172, 0.692]	0.05666	1.00000	0.01165	1.00000
BATF (bZIP)	*in vitro*	-0.044	[-0.35, 0.261]	-0.054	[-0.295, 0.187]	0.96198	1.00000	0.63623	1.00000
BATF (bZIP)	*in vivo*	0.239	[-0.02, 0.499]	0.38	[0.129, 0.632]	0.48220	1.00000	0.64951	1.00000
Bcl6 (Zf)	*in vitro*	-0.418	[-0.618, -0.219]	-0.181	[-0.288, -0.074]	0.03169	1.00000	0.03397	1.00000
Bcl6 (Zf)	*in vivo*	-0.6	[-0.87, -0.33]	-0.387	[-0.624, -0.15]	0.23560	1.00000	0.16702	1.00000
bHLHE40 (bHLH)	*in vitro*	0.319	[0.049, 0.589]	0.438	[0.261, 0.615]	0.47067	1.00000	0.31208	1.00000
bHLHE40 (bHLH)	*in vivo*	0.693	[0.295, 1.092]	0.851	[0.59, 1.111]	0.50268	1.00000	0.54354	1.00000
BMAL1 (bHLH)	*in vitro*	0.108	[-0.186, 0.402]	0.015	[-0.182, 0.211]	0.60107	1.00000	0.62963	1.00000
BMAL1 (bHLH)	*in vivo*	0.106	[-0.226, 0.437]	0.591	[0.377, 0.805]	0.02043	1.00000	0.00435	1.00000
BMYB (HTH)	*in vitro*	-0.003	[-0.156, 0.15]	0.023	[-0.042, 0.088]	0.76260	1.00000	0.80333	1.00000
BMYB (HTH)	*in vivo*	-0.345	[-0.616, -0.073]	-0.208	[-0.37, -0.047]	0.38823	1.00000	0.14348	1.00000
BORIS (Zf)	*in vitro*	0.083	[-0.214, 0.38]	0.363	[0.149, 0.577]	0.15898	1.00000	0.22218	1.00000
BORIS (Zf)	*in vivo*	-0.038	[-0.427, 0.351]	0.273	[-0.025, 0.572]	0.14986	1.00000	0.31386	1.00000
Brachyury (T-box)	*in vitro*	-0.101	[-0.395, 0.193]	0.218	[0.093, 0.343]	0.05249	1.00000	0.01105	1.00000
Brachyury (T-box)	*in vivo*	-0.064	[-0.403, 0.276]	0.138	[-0.133, 0.41]	0.42587	1.00000	0.22572	1.00000
bZIP:IRF (bZIP,IRF)	*in vitro*	-0.665	[-0.859, -0.471]	-0.27	[-0.381, -0.159]	0.00133	0.64429	0.00007	0.03591
bZIP:IRF (bZIP,IRF)	*in vivo*	-0.816	[-1.178, -0.455]	-0.468	[-0.625, -0.312]	0.09921	1.00000	0.04278	1.00000
c-Jun-CRE (bZIP)	*in vitro*	-0.014	[-0.312, 0.284]	0.087	[-0.093, 0.266]	0.54191	1.00000	0.61181	1.00000

c-Jun-CRE (bZIP)	*in vivo*	-0.081	[-0.51, 0.348]	-0.055	[-0.369, 0.259]	0.91721	1.00000	0.70033	1.00000
c-Myc (bHLH)	*in vitro*	0.064	[-0.164, 0.292]	0.042	[-0.064, 0.148]	0.84510	1.00000	0.91187	1.00000
c-Myc (bHLH)	*in vivo*	0.105	[-0.223, 0.432]	0.15	[-0.052, 0.351]	0.82233	1.00000	0.56216	1.00000
CArG (MADS)	*in vitro*	-0.384	[-0.658, -0.11]	0.014	[-0.139, 0.168]	0.02435	1.00000	0.00416	1.00000
CArG (MADS)	*in vivo*	-0.364	[-0.71, -0.018]	0.104	[-0.124, 0.332]	0.04127	1.00000	0.00488	1.00000
Cdx2 (Homeobox)	*in vitro*	-0.522	[-0.753, -0.291]	-0.437	[-0.591, -0.283]	0.51742	1.00000	0.62483	1.00000
Cdx2 (Homeobox)	*in vivo*	-1.141	[-1.433, -0.85]	-0.327	[-0.561, -0.093]	0.00006	0.02755	0.00007	0.03442
CEBP:AP1 (bZIP)	*in vitro*	-0.244	[-0.427, -0.062]	-0.072	[-0.191, 0.047]	0.13748	1.00000	0.07960	1.00000
CEBP:AP1 (bZIP)	*in vivo*	-0.603	[-0.946, -0.261]	-0.13	[-0.306, 0.045]	0.00873	1.00000	0.03165	1.00000
CEBP:CEBP (bZIP)	*in vitro*	-0.253	[-0.52, 0.014]	-0.235	[-0.367, -0.102]	0.89122	1.00000	0.19737	1.00000
CEBP:CEBP (bZIP)	*in vivo*	-0.569	[-0.934, -0.205]	-0.128	[-0.397, 0.142]	0.03895	1.00000	0.05728	1.00000
CEBP (bZIP)	*in vitro*	-0.315	[-0.508, -0.121]	-0.084	[-0.196, 0.028]	0.06099	1.00000	0.12643	1.00000
CEBP (bZIP)	*in vivo*	-0.718	[-1.095, -0.341]	-0.227	[-0.424, -0.03]	0.01397	1.00000	0.01240	1.00000
Chop (bZIP)	*in vitro*	-0.379	[-0.615, -0.143]	-0.141	[-0.261, -0.022]	0.06842	1.00000	0.12186	1.00000
Chop (bZIP)	*in vivo*	-0.386	[-0.785, 0.013]	0.059	[-0.235, 0.352]	0.03674	1.00000	0.04041	1.00000
CLOCK (bHLH)	*in vitro*	-0.241	[-0.499, 0.018]	0.176	[0.032, 0.32]	0.00575	1.00000	0.00282	1.00000
CLOCK (bHLH)	*in vivo*	0.244	[-0.105, 0.592]	0.251	[-0.024, 0.525]	0.97297	1.00000	0.94455	1.00000
CRE (bZIP)	*in vitro*	-0.16	[-0.509, 0.189]	0.183	[-0.008, 0.374]	0.02980	1.00000	0.13662	1.00000
CRE (bZIP)	*in vivo*	0.28	[-0.083, 0.642]	0.329	[0.048, 0.609]	0.82677	1.00000	0.86742	1.00000
CRX (Homeobox)	*in vitro*	-0.089	[-0.307, 0.128]	-0.164	[-0.297, -0.032]	0.54821	1.00000	0.36047	1.00000
CRX (Homeobox)	*in vivo*	-0.123	[-0.53, 0.285]	-0.15	[-0.373, 0.073]	0.89800	1.00000	0.95587	1.00000
CTCF-SatelliteElement	*in vitro*	0.383	[0.073, 0.693]	0.873	[0.637, 1.108]	0.00695	1.00000	0.00004	0.02040
CTCF-SatelliteElement	*in vivo*	1.405	[1.295, 1.515]	1.821	[1.631, 2.012]	0.00022	0.10550	0.00000	0.00000
CTCF (Zf)	*in vitro*	-0.099	[-0.496, 0.298]	0.208	[-0.083, 0.499]	0.23466	1.00000	0.23887	1.00000
CTCF (Zf)	*in vivo*	0.011	[-0.409, 0.43]	0.175	[-0.186, 0.536]	0.50680	1.00000	0.61000	1.00000
E-box (bHLH)	*in vitro*	0.516	[0.182, 0.849]	0.424	[0.198, 0.65]	0.66408	1.00000	0.50430	1.00000
E-box (bHLH)	*in vivo*	0.67	[0.43, 0.91]	0.904	[0.667, 1.141]	0.18125	1.00000	0.26376	1.00000
E2A (bHLH)	*in vitro*	0.24	[-0.006, 0.487]	0.25	[0.12, 0.38]	0.94722	1.00000	0.84642	1.00000

E2A (bHLH)	*in vivo*	0.181	[-0.14, 0.501]	0.346	[0.074, 0.618]	0.47217	1.00000	0.30766	1.00000
E2A (bHLH), near_PU.1	*in vitro*	0.247	[0.123, 0.37]	0.188	[0.098, 0.277]	0.47432	1.00000	0.61000	1.00000
E2A (bHLH), near_PU.1	*in vivo*	-0.162	[-0.481, 0.157]	0.379	[0.187, 0.57]	0.00242	1.00000	0.00540	1.00000
E2F1 (E2F)	*in vitro*	-0.134	[-0.375, 0.108]	0.287	[0.158, 0.416]	0.00184	0.88995	0.00156	0.75715
E2F1 (E2F)	*in vivo*	-0.041	[-0.414, 0.332]	0.388	[0.154, 0.623]	0.06215	1.00000	0.00993	1.00000
E2F4 (E2F)	*in vitro*	-0.364	[-0.604, -0.124]	-0.031	[-0.165, 0.102]	0.02040	1.00000	0.04576	1.00000
E2F4 (E2F)	*in vivo*	-0.497	[-0.852, -0.142]	-0.072	[-0.359, 0.215]	0.06355	1.00000	0.05001	1.00000
E2F6 (E2F)	*in vitro*	-0.147	[-0.369, 0.075]	0.222	[0.123, 0.321]	0.00245	1.00000	0.00172	0.83276
E2F6 (E2F)	*in vivo*	-0.018	[-0.384, 0.348]	0.177	[-0.031, 0.384]	0.39337	1.00000	0.08453	1.00000
E2F7 (E2F)	*in vitro*	-0.423	[-0.687, -0.159]	-0.035	[-0.248, 0.179]	0.01045	1.00000	0.02541	1.00000
E2F7 (E2F)	*in vivo*	0.001	[-0.286, 0.287]	0.435	[0.176, 0.694]	0.05183	1.00000	0.00113	0.54827
E2F (E2F)	*in vitro*	-0.257	[-0.592, 0.079]	0.02	[-0.213, 0.253]	0.18753	1.00000	0.03369	1.00000
E2F (E2F)	*in vivo*	0.569	[0.365, 0.773]	0.897	[0.695, 1.099]	0.02480	1.00000	0.00142	0.68757
EBF1 (EBF)	*in vitro*	-0.118	[-0.349, 0.112]	0.235	[0.078, 0.392]	0.01507	1.00000	0.02364	1.00000
EBF1 (EBF)	*in vivo*	-0.639	[-0.963, -0.316]	0.075	[-0.17, 0.32]	0.00047	0.22956	0.00134	0.65092
EBF (EBF)	*in vitro*	0.13	[-0.227, 0.487]	0.43	[0.18, 0.679]	0.17407	1.00000	0.09784	1.00000
EBF (EBF)	*in vivo*	0.091	[-0.25, 0.433]	0.329	[0.074, 0.584]	0.17524	1.00000	0.12144	1.00000
EBNA1 (EBV_virus)	*in vitro*	1.487	[1.256, 1.718]	1.79	[1.594, 1.986]	0.02962	1.00000	0.00002	0.00974
EBNA1 (EBV_virus)	*in vivo*	2.826	[2.752, 2.901]	3.038	[2.932, 3.144]	0.00277	1.00000	0.00000	0.00003
Egr1 (Zf)	*in vitro*	0.219	[0.066, 0.373]	0.246	[0.141, 0.351]	0.76190	1.00000	0.70526	1.00000
Egr1 (Zf)	*in vivo*	0.013	[-0.295, 0.321]	0.104	[-0.108, 0.316]	0.57855	1.00000	0.75380	1.00000
Egr2 (Zf)	*in vitro*	0.21	[-0.008, 0.427]	0.311	[0.15, 0.471]	0.39318	1.00000	0.22796	1.00000
Egr2 (Zf)	*in vivo*	0.085	[-0.297, 0.466]	0.251	[-0.033, 0.535]	0.44527	1.00000	0.16702	1.00000
EHF (ETS)	*in vitro*	-0.162	[-0.28, -0.044]	-0.083	[-0.166, 0.001]	0.29888	1.00000	0.17668	1.00000
EHF (ETS)	*in vivo*	-0.525	[-0.806, -0.245]	-0.352	[-0.541, -0.163]	0.32325	1.00000	0.28367	1.00000
EKLF (Zf)	*in vitro*	-0.194	[-0.442, 0.053]	0.085	[-0.06, 0.23]	0.04512	1.00000	0.07126	1.00000
EKLF (Zf)	*in vivo*	-0.137	[-0.46, 0.185]	0.195	[-0.02, 0.41]	0.06592	1.00000	0.08793	1.00000
ELF1 (ETS)	*in vitro*	-0.161	[-0.401, 0.079]	0.046	[-0.067, 0.16]	0.11434	1.00000	0.06753	1.00000

ELF1 (ETS)	*in vivo*	-0.519	[-0.918, -0.121]	0.063	[-0.215, 0.342]	0.00768	1.00000	0.00942	1.00000
ELF5 (ETS)	*in vitro*	-0.356	[-0.562, -0.15]	-0.119	[-0.217, -0.02]	0.03559	1.00000	0.06613	1.00000
ELF5 (ETS)	*in vivo*	-0.382	[-0.682, -0.083]	-0.273	[-0.444, -0.103]	0.45162	1.00000	0.37526	1.00000
Elk1 (ETS)	*in vitro*	-0.219	[-0.506, 0.069]	0.009	[-0.109, 0.127]	0.12408	1.00000	0.37830	1.00000
Elk1 (ETS)	*in vivo*	-0.608	[-0.983, -0.232]	-0.028	[-0.276, 0.219]	0.01136	1.00000	0.01112	1.00000
Elk4 (ETS)	*in vitro*	-0.308	[-0.546, -0.071]	-0.037	[-0.171, 0.098]	0.04601	1.00000	0.11524	1.00000
Elk4 (ETS)	*in vivo*	-0.469	[-0.808, -0.13]	-0.054	[-0.279, 0.172]	0.02598	1.00000	0.01616	1.00000
Eomes (T-box)	*in vitro*	-0.077	[-0.232, 0.078]	-0.041	[-0.115, 0.032]	0.63429	1.00000	0.75604	1.00000
Eomes (T-box)	*in vivo*	-0.002	[-0.315, 0.312]	0.014	[-0.148, 0.175]	0.92781	1.00000	0.92243	1.00000
ERE (NR), IR3	*in vitro*	0.161	[-0.113, 0.435]	0.429	[0.235, 0.623]	0.09658	1.00000	0.10686	1.00000
ERE (NR), IR3	*in vivo*	-0.101	[-0.481, 0.279]	0.074	[-0.154, 0.301]	0.46430	1.00000	0.39040	1.00000
ERG (ETS)	*in vitro*	0.013	[-0.143, 0.169]	-0.09	[-0.2, 0.019]	0.25922	1.00000	0.18272	1.00000
ERG (ETS)	*in vivo*	-0.529	[-0.805, -0.253]	-0.231	[-0.417, -0.046]	0.08823	1.00000	0.10660	1.00000
Erra (NR)	*in vitro*	0.099	[-0.06, 0.258]	0.107	[0.009, 0.205]	0.93282	1.00000	0.64286	1.00000
Erra (NR)	*in vivo*	-0.007	[-0.334, 0.321]	0.11	[-0.124, 0.344]	0.52505	1.00000	0.60352	1.00000
Esrrb (NR)	*in vitro*	0.153	[-0.015, 0.32]	0.061	[-0.041, 0.163]	0.30231	1.00000	0.40391	1.00000
Esrrb (NR)	*in vivo*	-0.05	[-0.407, 0.308]	0.233	[-0.032, 0.497]	0.16943	1.00000	0.17371	1.00000
ETS:E-box (ETS, bHLH)	*in vitro*	-0.128	[-0.443, 0.187]	0.117	[-0.108, 0.342]	0.24408	1.00000	0.14033	1.00000
ETS:E-box (ETS, bHLH)	*in vivo*	-0.028	[-0.328, 0.272]	0.232	[-0.053, 0.518]	0.19101	1.00000	0.12942	1.00000
ETS:RUNX (ETS, Runt)	*in vitro*	0.183	[-0.145, 0.511]	0.302	[0.082, 0.523]	0.50447	1.00000	0.38842	1.00000
ETS:RUNX (ETS, Runt)	*in vivo*	0.09	[-0.208, 0.388]	0.681	[0.417, 0.946]	0.00532	1.00000	0.00370	1.00000
Ets1-distal (ETS)	*in vitro*	-0.179	[-0.423, 0.065]	-0.175	[-0.339, -0.01]	0.97538	1.00000	0.88268	1.00000
Ets1-distal (ETS)	*in vivo*	-0.356	[-0.651, -0.06]	-0.141	[-0.421, 0.14]	0.30514	1.00000	0.40591	1.00000
ETS1 (ETS)	*in vitro*	0.061	[-0.118, 0.241]	-0.008	[-0.111, 0.095]	0.53927	1.00000	0.19207	1.00000
ETS1 (ETS)	*in vivo*	-0.421	[-0.632, -0.209]	-0.18	[-0.34, -0.019]	0.05883	1.00000	0.12413	1.00000
ETS (ETS)	*in vitro*	-0.119	[-0.407, 0.169]	0.055	[-0.085, 0.194]	0.25976	1.00000	0.07960	1.00000
ETS (ETS)	*in vivo*	-0.521	[-0.975, -0.067]	0.169	[-0.096, 0.435]	0.00506	1.00000	0.01020	1.00000
ETV1 (ETS)	*in vitro*	0.022	[-0.117, 0.161]	-0.04	[-0.133, 0.053]	0.40099	1.00000	0.44901	1.00000

ETV1 (ETS)	*in vivo*	-0.531	[-0.864, -0.199]	-0.217	[-0.43, -0.004]	0.08745	1.00000	0.02548	1.00000
EWS:ERG-fusion (ETS)	*in vitro*	-0.457	[-0.631, -0.284]	-0.249	[-0.357, -0.142]	0.05815	1.00000	0.05458	1.00000
EWS:ERG-fusion (ETS)	*in vivo*	-0.81	[-1.164, -0.457]	-0.336	[-0.564, -0.108]	0.03013	1.00000	0.03018	1.00000
EWS:FLI1-fusion (ETS)	*in vitro*	-0.247	[-0.447, -0.046]	-0.163	[-0.308, -0.019]	0.51885	1.00000	0.27135	1.00000
EWS:FLI1-fusion (ETS)	*in vivo*	-0.563	[-0.918, -0.207]	-0.122	[-0.372, 0.128]	0.04610	1.00000	0.03319	1.00000
Fli1 (ETS)	*in vitro*	-0.121	[-0.309, 0.067]	-0.013	[-0.116, 0.089]	0.26347	1.00000	0.30942	1.00000
Fli1 (ETS)	*in vivo*	-0.389	[-0.677, -0.1]	-0.116	[-0.311, 0.079]	0.11924	1.00000	0.05341	1.00000
Fox:Ebox (Forkhead, bHLH)	*in vitro*	-0.001	[-0.13, 0.127]	0.086	[-0.013, 0.185]	0.19297	1.00000	0.35613	1.00000
Fox:Ebox (Forkhead, bHLH)	*in vivo*	-0.317	[-0.579, -0.056]	-0.376	[-0.591, -0.161]	0.69734	1.00000	0.85649	1.00000
FOXA1:AR (Forkhead,NR)	*in vitro*	-0.961	[-1.281, -0.64]	-0.669	[-0.871, -0.467]	0.10316	1.00000	0.01840	1.00000
FOXA1:AR (Forkhead,NR)	*in vivo*	-0.744	[-1.178, -0.31]	-0.397	[-0.641, -0.152]	0.20515	1.00000	0.00250	1.00000
FOXA1 (Forkhead)	*in vitro*	-0.279	[-0.706, 0.148]	-0.184	[-0.384, 0.016]	0.68783	1.00000	0.87472	1.00000
FOXA1 (Forkhead)	*in vivo*	0.083	[-0.245, 0.411]	0.157	[-0.087, 0.401]	0.72906	1.00000	0.38586	1.00000
Foxa2 (Forkhead)	*in vitro*	-0.264	[-0.444, -0.083]	-0.053	[-0.18, 0.073]	0.05788	1.00000	0.01494	1.00000
Foxa2 (Forkhead)	*in vivo*	-0.464	[-0.73, -0.199]	-0.562	[-0.771, -0.353]	0.55200	1.00000	0.35563	1.00000
Foxh1 (Forkhead)	*in vitro*	-0.073	[-0.262, 0.116]	-0.097	[-0.2, 0.007]	0.81633	1.00000	0.81763	1.00000
Foxh1 (Forkhead)	*in vivo*	-0.608	[-0.907, -0.31]	-0.472	[-0.671, -0.273]	0.38317	1.00000	0.41178	1.00000
Foxo1 (Forkhead)	*in vitro*	0.027	[-0.143, 0.196]	0.045	[-0.028, 0.118]	0.83821	1.00000	0.40070	1.00000
Foxo1 (Forkhead)	*in vivo*	-0.17	[-0.477, 0.137]	-0.159	[-0.317, -0.002]	0.95214	1.00000	0.94852	1.00000
FOXP1 (Forkhead)	*in vitro*	-0.149	[-0.353, 0.055]	-0.152	[-0.244, -0.06]	0.98113	1.00000	0.80946	1.00000
FOXP1 (Forkhead)	*in vivo*	-0.602	[-0.892, -0.311]	-0.503	[-0.749, -0.257]	0.52577	1.00000	0.40914	1.00000
FXR (NR),IR1	*in vitro*	0.351	[0.102, 0.599]	0.356	[0.172, 0.541]	0.97348	1.00000	0.44145	1.00000
FXR (NR),IR1	*in vivo*	0.17	[-0.193, 0.532]	0.302	[0.04, 0.565]	0.52262	1.00000	0.21056	1.00000
GABPA (ETS)	*in vitro*	0.068	[-0.135, 0.272]	0.006	[-0.094, 0.105]	0.56104	1.00000	0.87541	1.00000
GABPA (ETS)	*in vivo*	-0.411	[-0.756, -0.065]	0.002	[-0.171, 0.175]	0.02497	1.00000	0.02808	1.00000
GATA:SCL (Zf, bHLH)	*in vitro*	-0.077	[-0.4, 0.247]	0.054	[-0.099, 0.207]	0.46227	1.00000	0.20841	1.00000
GATA:SCL (Zf, bHLH)	*in vivo*	-0.244	[-0.679, 0.192]	0.194	[-0.048, 0.435]	0.05727	1.00000	0.01260	1.00000
Gata1 (Zf)	*in vitro*	-0.288	[-0.437, -0.139]	0.037	[-0.058, 0.131]	0.00010	0.04740	0.00021	0.10001

Gata1 (Zf)	*in vivo*	-0.615	[-0.955, -0.274]	-0.221	[-0.445, 0.002]	0.03790	1.00000	0.03692	1.00000
Gata2 (Zf)	*in vitro*	-0.368	[-0.517, -0.219]	0.005	[-0.094, 0.104]	0.00003	0.01441	0.00007	0.03442
Gata2 (Zf)	*in vivo*	-0.72	[-1.068, -0.371]	-0.215	[-0.367, -0.062]	0.00631	1.00000	0.01174	1.00000
GATA3 (Zf)	*in vitro*	-0.415	[-0.611, -0.219]	-0.185	[-0.277, -0.092]	0.03759	1.00000	0.00670	1.00000
GATA3 (Zf)	*in vivo*	-0.442	[-0.81, -0.073]	-0.22	[-0.401, -0.04]	0.28866	1.00000	0.17871	1.00000
GATA3 (Zf), DR4	*in vitro*	-0.723	[-1.014, -0.433]	-0.387	[-0.616, -0.158]	0.09280	1.00000	0.09153	1.00000
GATA3 (Zf), DR4	*in vivo*	-0.358	[-0.683, -0.033]	0.303	[0.064, 0.542]	0.00044	0.21183	0.00022	0.10420
GATA3 (Zf), DR8	*in vitro*	-0.737	[-1.033, -0.44]	-0.011	[-0.17, 0.148]	0.00017	0.08220	0.00001	0.00431
GATA3 (Zf), DR8	*in vivo*	-0.157	[-0.523, 0.209]	-0.163	[-0.422, 0.096]	0.97957	1.00000	0.68323	1.00000
Gata4 (Zf)	*in vitro*	-0.421	[-0.603, -0.239]	-0.103	[-0.192, -0.015]	0.00403	1.00000	0.00019	0.09239
Gata4 (Zf)	*in vivo*	-0.623	[-0.928, -0.317]	-0.149	[-0.326, 0.028]	0.00214	1.00000	0.01112	1.00000
GATA (Zf), IR3	*in vitro*	-0.452	[-0.74, -0.163]	-0.005	[-0.24, 0.231]	0.01768	1.00000	0.01244	1.00000
GATA (Zf), IR3	*in vivo*	0.272	[-0.036, 0.58]	0.16	[-0.111, 0.43]	0.64031	1.00000	0.91506	1.00000
GATA (Zf), IR3	*in vitro*	0.034	[-0.261, 0.33]	0.341	[0.105, 0.576]	0.11561	1.00000	0.04319	1.00000
GATA (Zf), IR3	*in vivo*	0.82	[0.639, 1]	1.102	[0.904, 1.3]	0.04030	1.00000	0.00867	1.00000
Gfi1b (Zf)	*in vitro*	-0.006	[-0.172, 0.16]	0.196	[0.097, 0.295]	0.05314	1.00000	0.04132	1.00000
Gfi1b (Zf)	*in vivo*	-0.203	[-0.543, 0.136]	-0.261	[-0.465, -0.058]	0.74648	1.00000	0.71902	1.00000
GFY-Staf (?, Zf)	*in vitro*	-0.156	[-0.433, 0.12]	0.203	[-0.071, 0.477]	0.08798	1.00000	0.03733	1.00000
GFY-Staf (?, Zf)	*in vivo*	0.682	[0.512, 0.852]	1.16	[0.952, 1.368]	0.00042	0.20210	0.00015	0.07325
GLI3 (Zf)	*in vitro*	-0.072	[-0.364, 0.221]	0.189	[0.034, 0.344]	0.12019	1.00000	0.12368	1.00000
GLI3 (Zf)	*in vivo*	-0.143	[-0.514, 0.229]	0.196	[-0.038, 0.43]	0.12498	1.00000	0.07199	1.00000
GRE (NR), IR3	*in vitro*	-0.256	[-0.54, 0.029]	0.125	[-0.04, 0.29]	0.03633	1.00000	0.01694	1.00000
GRE (NR), IR3	*in vivo*	-0.036	[-0.397, 0.325]	0.212	[-0.056, 0.481]	0.24659	1.00000	0.08621	1.00000
GRHL2 (CP2)	*in vitro*	-0.224	[-0.468, 0.02]	-0.04	[-0.19, 0.111]	0.17676	1.00000	0.06429	1.00000
GRHL2 (CP2)	*in vivo*	-0.302	[-0.639, 0.035]	-0.069	[-0.278, 0.14]	0.27308	1.00000	0.13844	1.00000
HIF-1a (bHLH)	*in vitro*	0.21	[-0.047, 0.467]	0.214	[0.101, 0.327]	0.97240	1.00000	0.75380	1.00000
HIF-1a (bHLH)	*in vivo*	0.119	[-0.121, 0.358]	-0.04	[-0.248, 0.167]	0.31185	1.00000	0.30503	1.00000
HIF2a (bHLH)	*in vitro*	0.097	[-0.109, 0.304]	0.175	[0.07, 0.28]	0.49485	1.00000	0.43256	1.00000

HIF2a (bHLH)	*in vivo*	-0.199	[-0.517, 0.118]	-0.02	[-0.201, 0.162]	0.35242	1.00000	0.39040	1.00000
Hnf1 (Homeobox)	*in vitro*	-0.993	[-1.265, -0.721]	-0.566	[-0.765, -0.367]	0.01469	1.00000	0.00778	1.00000
Hnf1 (Homeobox)	*in vivo*	-0.83	[-1.192, -0.467]	-0.511	[-0.785, -0.236]	0.16979	1.00000	0.02965	1.00000
HNF4a (NR), DR1	*in vitro*	-0.034	[-0.237, 0.169]	-0.054	[-0.17, 0.062]	0.87253	1.00000	0.57791	1.00000
HNF4a (NR), DR1	*in vivo*	-0.602	[-0.939, -0.265]	-0.093	[-0.305, 0.119]	0.03387	1.00000	0.00277	1.00000
HNF6 (Homeobox)	*in vitro*	-0.547	[-0.72, -0.373]	-0.307	[-0.418, -0.196]	0.03673	1.00000	0.02451	1.00000
HNF6 (Homeobox)	*in vivo*	-0.741	[-1.006, -0.477]	-0.26	[-0.488, -0.032]	0.00809	1.00000	0.00605	1.00000
HOXA2 (Homeobox)	*in vitro*	-0.575	[-0.839, -0.311]	-0.113	[-0.281, 0.056]	0.01098	1.00000	0.01120	1.00000
HOXA2 (Homeobox)	*in vivo*	-0.332	[-0.671, 0.006]	0.319	[0.07, 0.569]	0.00143	0.69103	0.00319	1.00000
HOXA9 (Homeobox)	*in vitro*	-0.38	[-0.603, -0.157]	-0.102	[-0.183, -0.022]	0.01318	1.00000	0.02364	1.00000
HOXA9 (Homeobox)	*in vivo*	-0.775	[-1.015, -0.535]	-0.271	[-0.477, -0.064]	0.00247	1.00000	0.00170	0.82388
Hoxb4 (Homeobox)	*in vitro*	-0.341	[-0.582, -0.1]	0.073	[-0.056, 0.201]	0.00632	1.00000	0.00034	0.16561
Hoxb4 (Homeobox)	*in vivo*	-0.748	[-1.082, -0.414]	0.047	[-0.173, 0.267]	0.00015	0.07466	0.00010	0.04660
Hoxc9 (Homeobox)	*in vitro*	-0.493	[-0.652, -0.334]	-0.323	[-0.444, -0.202]	0.09782	1.00000	0.05823	1.00000
Hoxc9 (Homeobox)	*in vivo*	-0.973	[-1.266, -0.68]	-0.404	[-0.62, -0.188]	0.00077	0.37418	0.00221	1.00000
HOXD13 (Homeobox)	*in vitro*	-0.66	[-0.884, -0.436]	-0.462	[-0.602, -0.322]	0.11352	1.00000	0.03478	1.00000
HOXD13 (Homeobox)	*in vivo*	-0.978	[-1.337, -0.619]	-0.346	[-0.649, -0.043]	0.00685	1.00000	0.00441	1.00000
HRE (HSF)	*in vitro*	-0.77	[-1.119, -0.422]	-0.207	[-0.383, -0.032]	0.00532	1.00000	0.01020	1.00000
HRE (HSF)	*in vivo*	-0.28	[-0.627, 0.067]	0.014	[-0.249, 0.277]	0.18658	1.00000	0.05974	1.00000
IRF1 (IRF)	*in vitro*	-0.773	[-1.016, -0.531]	-0.304	[-0.469, -0.139]	0.00425	1.00000	0.00046	0.22050
IRF1 (IRF)	*in vivo*	-0.958	[-1.233, -0.683]	-0.435	[-0.703, -0.168]	0.00735	1.00000	0.00640	1.00000
IRF2 (IRF)	*in vitro*	-0.786	[-1.086, -0.486]	-0.056	[-0.234, 0.121]	0.00023	0.11187	0.00004	0.02139
IRF2 (IRF)	*in vivo*	-0.428	[-0.709, -0.146]	-0.076	[-0.339, 0.188]	0.03645	1.00000	0.03369	1.00000
IRF4 (IRF)	*in vitro*	-0.313	[-0.521, -0.105]	-0.151	[-0.272, -0.029]	0.20399	1.00000	0.23292	1.00000
IRF4 (IRF)	*in vivo*	-0.417	[-0.803, -0.031]	-0.153	[-0.398, 0.091]	0.17239	1.00000	0.28619	1.00000
Isl1 (Homeobox)	*in vitro*	-0.251	[-0.417, -0.085]	-0.089	[-0.211, 0.033]	0.12856	1.00000	0.07040	1.00000
Isl1 (Homeobox)	*in vivo*	-0.475	[-0.764, -0.187]	-0.124	[-0.332, 0.084]	0.02561	1.00000	0.06840	1.00000
ISRE (IRF)	*in vitro*	-0.528	[-0.801, -0.256]	-0.219	[-0.458, 0.019]	0.07510	1.00000	0.07199	1.00000

ISRE (IRF)	*in vivo*	-0.032	[-0.248, 0.185]	0.178	[-0.049, 0.404]	0.13342	1.00000	0.11487	1.00000
Jun-AP1 (bZIP)	*in vitro*	0.592	[0.279, 0.905]	0.653	[0.415, 0.892]	0.73291	1.00000	0.38076	1.00000
Jun-AP1 (bZIP)	*in vivo*	1.628	[1.383, 1.872]	1.767	[1.547, 1.987]	0.36983	1.00000	0.14529	1.00000
JunD (bZIP)	*in vitro*	-0.182	[-0.517, 0.152]	0.429	[0.178, 0.681]	0.01427	1.00000	0.00404	1.00000
JunD (bZIP)	*in vivo*	0.997	[0.709, 1.286]	1.088	[0.872, 1.304]	0.61654	1.00000	0.19096	1.00000
Klf4 (Zf)	*in vitro*	-0.171	[-0.41, 0.067]	0.142	[0.024, 0.26]	0.02545	1.00000	0.00844	1.00000
Klf4 (Zf)	*in vivo*	-0.099	[-0.375, 0.177]	-0.055	[-0.335, 0.226]	0.79688	1.00000	0.56844	1.00000
KLF5 (Zf)	*in vitro*	0.265	[0.09, 0.44]	0.201	[0.097, 0.305]	0.47658	1.00000	0.85365	1.00000
KLF5 (Zf)	*in vivo*	-0.213	[-0.539, 0.113]	-0.136	[-0.36, 0.088]	0.64945	1.00000	0.53026	1.00000
Lhx2 (Homeobox)	*in vitro*	-0.38	[-0.603, -0.157]	-0.267	[-0.396, -0.138]	0.37169	1.00000	0.42178	1.00000
Lhx2 (Homeobox)	*in vivo*	-0.574	[-0.918, -0.231]	-0.421	[-0.646, -0.196]	0.44719	1.00000	0.23520	1.00000
Lhx3 (Homeobox)	*in vitro*	-0.518	[-0.731, -0.305]	-0.202	[-0.362, -0.042]	0.03648	1.00000	0.01747	1.00000
Lhx3 (Homeobox)	*in vivo*	-0.406	[-0.707, -0.105]	-0.117	[-0.373, 0.139]	0.15287	1.00000	0.12596	1.00000
LXRE (NR), DR4	*in vitro*	0.135	[-0.214, 0.485]	0.063	[-0.149, 0.276]	0.71704	1.00000	0.61000	1.00000
LXRE (NR), DR4	*in vivo*	0.241	[-0.079, 0.561]	0.215	[-0.038, 0.469]	0.88984	1.00000	0.62305	1.00000
MafA (bZIP)	*in vitro*	0.179	[0.019, 0.34]	0.164	[0.057, 0.27]	0.86051	1.00000	0.91187	1.00000
MafA (bZIP)	*in vivo*	0.071	[-0.181, 0.324]	0.443	[0.215, 0.672]	0.04503	1.00000	0.05458	1.00000
MafF (bZIP)	*in vitro*	-0.585	[-0.808, -0.361]	-0.03	[-0.171, 0.111]	0.00041	0.19851	0.00018	0.08833
MafF (bZIP)	*in vivo*	-0.721	[-1.07, -0.372]	-0.308	[-0.596, -0.02]	0.05614	1.00000	0.06910	1.00000
MafK (bZIP)	*in vitro*	-0.069	[-0.34, 0.202]	-0.041	[-0.218, 0.137]	0.86305	1.00000	0.69690	1.00000
MafK (bZIP)	*in vivo*	0.05	[-0.256, 0.355]	0.327	[0.074, 0.58]	0.10987	1.00000	0.10262	1.00000
Max (bHLH)	*in vitro*	-0.013	[-0.283, 0.256]	0.066	[-0.07, 0.201]	0.61250	1.00000	0.21226	1.00000
Max (bHLH)	*in vivo*	0.077	[-0.241, 0.394]	0.186	[-0.08, 0.451]	0.61818	1.00000	0.35131	1.00000
Maz (Zf)	*in vitro*	0.184	[0.011, 0.356]	0.108	[0, 0.215]	0.46168	1.00000	0.47149	1.00000
Maz (Zf)	*in vivo*	-0.267	[-0.577, 0.044]	0.156	[-0.05, 0.361]	0.02025	1.00000	0.05112	1.00000
Mef2a (MADS)	*in vitro*	-0.426	[-0.631, -0.221]	-0.411	[-0.511, -0.31]	0.89401	1.00000	0.82481	1.00000
Mef2a (MADS)	*in vivo*	-0.91	[-1.216, -0.604]	-0.635	[-0.838, -0.432]	0.15501	1.00000	0.13844	1.00000
Mef2c (MADS)	*in vitro*	-0.823	[-1.005, -0.642]	-0.645	[-0.796, -0.493]	0.13934	1.00000	0.08707	1.00000

Mef2c (MADS)	*in vivo*	-1.036	[-1.325, -0.746]	-0.951	[-1.197, -0.706]	0.66670	1.00000	0.95587	1.00000
Meis1 (Homeobox)	*in vitro*	-0.054	[-0.222, 0.114]	0.129	[0.02, 0.238]	0.08125	1.00000	0.02451	1.00000
Meis1 (Homeobox)	*in vivo*	-0.077	[-0.406, 0.252]	0.006	[-0.211, 0.222]	0.66317	1.00000	0.97041	1.00000
MITF (bHLH)	*in vitro*	0.105	[-0.067, 0.277]	0.172	[0.063, 0.28]	0.52448	1.00000	0.32285	1.00000
MITF (bHLH)	*in vivo*	-0.331	[-0.582, -0.08]	0.062	[-0.137, 0.261]	0.01634	1.00000	0.02145	1.00000
MYB (HTH)	*in vitro*	-0.085	[-0.256, 0.086]	0.065	[-0.034, 0.164]	0.13151	1.00000	0.11068	1.00000
MYB (HTH)	*in vivo*	-0.28	[-0.526, -0.034]	-0.321	[-0.464, -0.179]	0.77170	1.00000	0.90456	1.00000
Myf5 (bHLH)	*in vitro*	0.379	[0.15, 0.609]	0.311	[0.194, 0.428]	0.62664	1.00000	0.25940	1.00000
Myf5 (bHLH)	*in vivo*	0.15	[-0.237, 0.537]	0.203	[-0.041, 0.447]	0.80153	1.00000	0.64457	1.00000
MyoD (bHLH)	*in vitro*	0.563	[0.347, 0.778]	0.346	[0.166, 0.525]	0.11559	1.00000	0.14094	1.00000
MyoD (bHLH)	*in vivo*	0.064	[-0.364, 0.492]	0.241	[-0.073, 0.554]	0.41915	1.00000	0.39870	1.00000
MyoG (bHLH)	*in vitro*	0.45	[0.271, 0.628]	0.321	[0.192, 0.451]	0.28859	1.00000	0.36536	1.00000
MyoG (bHLH)	*in vivo*	0.062	[-0.211, 0.335]	0.087	[-0.117, 0.29]	0.88820	1.00000	0.66966	1.00000
n-Myc (bHLH)	*in vitro*	-0.2	[-0.483, 0.082]	0.013	[-0.135, 0.16]	0.16125	1.00000	0.07336	1.00000
n-Myc (bHLH)	*in vivo*	0.186	[-0.149, 0.52]	0.214	[-0.013, 0.442]	0.88921	1.00000	0.89303	1.00000
Nanog (Homeobox)	*in vitro*	0.076	[-0.141, 0.293]	0.16	[-0.031, 0.351]	0.53457	1.00000	0.56843	1.00000
Nanog (Homeobox)	*in vivo*	0.298	[0.023, 0.573]	0.127	[-0.088, 0.343]	0.33046	1.00000	0.51023	1.00000
NeuroD1 (bHLH)	*in vitro*	0.059	[-0.105, 0.223]	0.126	[0.036, 0.216]	0.47883	1.00000	0.18478	1.00000
NeuroD1 (bHLH)	*in vivo*	0.018	[-0.282, 0.318]	0.199	[0.021, 0.377]	0.31113	1.00000	0.12368	1.00000
NF-E2 (bZIP)	*in vitro*	0.121	[-0.21, 0.451]	0.292	[0.064, 0.521]	0.43137	1.00000	0.16139	1.00000
NF-E2 (bZIP)	*in vivo*	0.959	[0.697, 1.222]	1.057	[0.795, 1.319]	0.60769	1.00000	0.55592	1.00000
NF1-halfsite (CTF)	*in vitro*	0.085	[-0.099, 0.269]	0.159	[0.054, 0.265]	0.46672	1.00000	0.97793	1.00000
NF1-halfsite (CTF)	*in vivo*	0.277	[0.01, 0.545]	0.164	[-0.024, 0.351]	0.46441	1.00000	0.54859	1.00000
NF1:FOXA1 (CTF,Forkhead)	*in vitro*	-0.582	[-1.116, -0.048]	0.078	[-0.248, 0.404]	0.04598	1.00000	0.02511	1.00000
NF1:FOXA1 (CTF,Forkhead)	*in vivo*	-0.253	[-0.649, 0.143]	-0.004	[-0.359, 0.35]	0.35949	1.00000	0.41178	1.00000
NF1 (CTF)	*in vitro*	0.153	[-0.179, 0.484]	0.435	[0.278, 0.593]	0.11268	1.00000	0.05975	1.00000
NF1 (CTF)	*in vivo*	-0.019	[-0.343, 0.305]	0.453	[0.23, 0.677]	0.02123	1.00000	0.04320	1.00000
NFAT:AP1 (RHD,bZIP)	*in vitro*	-0.968	[-1.3, -0.636]	-0.446	[-0.595, -0.296]	0.00535	1.00000	0.00274	1.00000

NFAT:AP1 (RHD,bZIP)	*in vivo*	-0.809	[-1.139, -0.48]	-0.372	[-0.621, -0.122]	0.04045	1.00000	0.00716	1.00000
NFAT (RHD)	*in vitro*	-0.665	[-0.875, -0.455]	-0.415	[-0.517, -0.314]	0.02738	1.00000	0.00677	1.00000
NFAT (RHD)	*in vivo*	-0.69	[-1.01, -0.37]	-0.135	[-0.35, 0.081]	0.00345	1.00000	0.00613	1.00000
NFkB-p50,p52 (RHD)	*in vitro*	-0.411	[-0.765, -0.056]	-0.177	[-0.362, 0.008]	0.21984	1.00000	0.46377	1.00000
NFkB-p50,p52 (RHD)	*in vivo*	-0.457	[-0.802, -0.111]	-0.228	[-0.47, 0.013]	0.18187	1.00000	0.12254	1.00000
NFkB-p65-Rel (RHD)	*in vitro*	-0.388	[-0.772, -0.003]	0.064	[-0.18, 0.309]	0.06598	1.00000	0.01590	1.00000
NFkB-p65-Rel (RHD)	*in vivo*	0.115	[-0.126, 0.356]	0.521	[0.264, 0.777]	0.02394	1.00000	0.09052	1.00000
NFkB-p65 (RHD)	*in vitro*	-0.013	[-0.253, 0.227]	-0.009	[-0.165, 0.148]	0.97819	1.00000	0.84642	1.00000
NFkB-p65 (RHD)	*in vivo*	-0.756	[-1.124, -0.388]	-0.093	[-0.398, 0.213]	0.01204	1.00000	0.01455	1.00000
NFY (CCAAT)	*in vitro*	-0.017	[-0.211, 0.177]	0.187	[0.09, 0.284]	0.06957	1.00000	0.03018	1.00000
NFY (CCAAT)	*in vivo*	-0.532	[-0.812, -0.252]	-0.163	[-0.348, 0.022]	0.03196	1.00000	0.04667	1.00000
Nkx2.1 (Homeobox)	*in vitro*	0.058	[-0.107, 0.223]	0.045	[-0.068, 0.157]	0.89872	1.00000	0.88570	1.00000
Nkx2.1 (Homeobox)	*in vivo*	-0.47	[-0.753, -0.187]	-0.022	[-0.219, 0.175]	0.01347	1.00000	0.00710	1.00000
Nkx2.5 (Homeobox)	*in vitro*	0.144	[-0.009, 0.297]	0.172	[0.077, 0.267]	0.77708	1.00000	0.82026	1.00000
Nkx2.5 (Homeobox)	*in vivo*	-0.336	[-0.622, -0.05]	-0.085	[-0.26, 0.09]	0.15222	1.00000	0.16419	1.00000
Nkx3.1 (Homeobox)	*in vitro*	-0.006	[-0.216, 0.204]	0.176	[0.088, 0.264]	0.13742	1.00000	0.04929	1.00000
Nkx3.1 (Homeobox)	*in vivo*	-0.335	[-0.605, -0.065]	-0.117	[-0.302, 0.067]	0.18726	1.00000	0.12368	1.00000
Nkx6.1 (Homeobox)	*in vitro*	-0.671	[-0.953, -0.388]	-0.23	[-0.429, -0.03]	0.00720	1.00000	0.01550	1.00000
Nkx6.1 (Homeobox)	*in vivo*	-0.355	[-0.663, -0.048]	-0.107	[-0.418, 0.203]	0.27218	1.00000	0.21914	1.00000
NPAS2 (bHLH)	*in vitro*	0.029	[-0.12, 0.178]	0.101	[0.027, 0.175]	0.41174	1.00000	0.23156	1.00000
NPAS2 (bHLH)	*in vivo*	-0.051	[-0.319, 0.218]	0.025	[-0.146, 0.196]	0.60662	1.00000	0.78199	1.00000
Nr5a2 (NR)	*in vitro*	-0.011	[-0.265, 0.243]	-0.02	[-0.17, 0.13]	0.95676	1.00000	0.78084	1.00000
Nr5a2 (NR)	*in vivo*	0.273	[-0.12, 0.666]	0.058	[-0.202, 0.319]	0.38289	1.00000	0.44697	1.00000
NRF1 (NRF)	*in vitro*	0.523	[0.194, 0.852]	0.862	[0.582, 1.142]	0.06852	1.00000	0.20887	1.00000
NRF1 (NRF)	*in vivo*	1.134	[0.957, 1.311]	1.534	[1.247, 1.822]	0.01648	1.00000	0.77724	1.00000
Nrf2 (bZIP)	*in vitro*	0.039	[-0.329, 0.407]	0.41	[0.167, 0.654]	0.10407	1.00000	0.02541	1.00000
Nrf2 (bZIP)	*in vivo*	0.944	[0.68, 1.208]	1.143	[0.932, 1.353]	0.17175	1.00000	0.06294	1.00000
NRF (NRF)	*in vitro*	0.068	[-0.253, 0.389]	0.479	[0.247, 0.71]	0.02881	1.00000	0.04566	1.00000

NRF (NRF)	*in vivo*	0.64	[0.313, 0.967]	0.788	[0.492, 1.083]	0.50298	1.00000	0.29464	1.00000
Nur77 (NR)	*in vitro*	-0.491	[-0.735, -0.247]	-0.297	[-0.446, -0.149]	0.20031	1.00000	0.08124	1.00000
Nur77 (NR)	*in vivo*	-0.648	[-1.014, -0.283]	-0.103	[-0.33, 0.123]	0.02743	1.00000	0.00056	0.27137
Oct2 (POU,Homeobox)	*in vitro*	-0.491	[-0.69, -0.293]	-0.43	[-0.552, -0.308]	0.61269	1.00000	0.41178	1.00000
Oct2 (POU,Homeobox)	*in vivo*	-0.617	[-0.945, -0.29]	-0.359	[-0.61, -0.107]	0.21300	1.00000	0.11813	1.00000
OCT4-SOX2-TCF-NANOG (POU,Homeobox,HMG)	*in vitro*	-0.732	[-0.949, -0.516]	-0.458	[-0.616, -0.3]	0.01841	1.00000	0.00942	1.00000
OCT4-SOX2-TCF-NANOG (POU,Homeobox,HMG)	*in vivo*	-1.216	[-1.564, -0.869]	-0.715	[-0.968, -0.461]	0.01530	1.00000	0.00156	0.75719
Oct4:Sox17 (POU,Homeobox,HMG)	*in vitro*	-0.322	[-0.609, -0.035]	-0.142	[-0.271, -0.013]	0.27151	1.00000	0.20389	1.00000
Oct4:Sox17 (POU,Homeobox,HMG)	*in vivo*	-0.488	[-0.845, -0.131]	-0.388	[-0.666, -0.111]	0.66326	1.00000	0.48874	1.00000
Oct4 (POU,Homeobox)	*in vitro*	-0.533	[-0.733, -0.333]	-0.305	[-0.437, -0.173]	0.04269	1.00000	0.02252	1.00000
Oct4 (POU,Homeobox)	*in vivo*	-0.71	[-1.088, -0.332]	-0.418	[-0.661, -0.176]	0.11107	1.00000	0.24403	1.00000
Olig2 (bHLH)	*in vitro*	-0.059	[-0.227, 0.11]	-0.008	[-0.108, 0.092]	0.59276	1.00000	0.38027	1.00000
Olig2 (bHLH)	*in vivo*	-0.343	[-0.629, -0.056]	-0.119	[-0.317, 0.078]	0.16862	1.00000	0.07126	1.00000
p53 (p53)	*in vitro*	0.339	[-0.161, 0.839]	0.66	[0.399, 0.922]	0.23023	1.00000	0.01260	1.00000
p53 (p53)	*in vivo*	1.364	[1.208, 1.521]	1.309	[1.146, 1.472]	0.59800	1.00000	0.21562	1.00000
p63 (p53)	*in vitro*	-0.08	[-0.347, 0.188]	0.287	[0.084, 0.49]	0.02654	1.00000	0.04929	1.00000
p63 (p53)	*in vivo*	0.566	[0.251, 0.88]	0.62	[0.375, 0.864]	0.77513	1.00000	0.42781	1.00000
PAX3:FKHR-fusion (Paired,Homeobox)	*in vitro*	-0.195	[-0.429, 0.04]	-0.15	[-0.356, 0.056]	0.79548	1.00000	0.99264	1.00000
PAX3:FKHR-fusion (Paired,Homeobox)	*in vivo*	-0.673	[-1.056, -0.291]	-0.111	[-0.412, 0.19]	0.02000	1.00000	0.00891	1.00000
PAX5 (Paired,Homeobox)	*in vitro*	0.325	[0.105, 0.544]	0.291	[0.188, 0.394]	0.77283	1.00000	0.56097	1.00000
PAX5 (Paired,Homeobox)	*in vivo*	-0.146	[-0.521, 0.229]	0.12	[-0.085, 0.324]	0.15349	1.00000	0.37830	1.00000
PAX5 (Paired, Homeobox), condensed	*in vitro*	0.22	[-0.1, 0.541]	0.455	[0.316, 0.594]	0.20294	1.00000	0.10361	1.00000
PAX5 (Paired, Homeobox), condensed	*in vivo*	0.195	[-0.175, 0.566]	0.604	[0.355, 0.852]	0.04133	1.00000	0.03369	1.00000

Pax7 (Paired,Homeobox)	*in vitro*	-0.467	[-0.808, -0.125]	-0.122	[-0.401, 0.158]	0.15941	1.00000	0.11962	1.00000
Pax7 (Paired,Homeobox)	*in vivo*	-1.119	[-1.47, -0.768]	-0.271	[-0.614, 0.073]	0.00224	1.00000	0.00066	0.32186
Pax7 (Paired, Homeobox), long	*in vitro*	0.009	[-0.242, 0.259]	0.266	[0.015, 0.516]	0.15104	1.00000	0.15181	1.00000
Pax7 (Paired, Homeobox), long	*in vivo*	1.144	[1, 1.289]	1.193	[1.051, 1.335]	0.65497	1.00000	0.50387	1.00000
Pax7 (Paired, Homeobox), longest	*in vitro*	-0.267	[-0.595, 0.061]	-0.063	[-0.276, 0.15]	0.20725	1.00000	0.02171	1.00000
Pax7 (Paired, Homeobox), longest	*in vivo*	0.919	[0.752, 1.086]	0.936	[0.791, 1.082]	0.88358	1.00000	0.05086	1.00000
Pax8 (Paired, Homeobox)	*in vitro*	0.593	[0.406, 0.779]	0.533	[0.447, 0.62]	0.58047	1.00000	0.52422	1.00000
Pax8 (Paired, Homeobox)	*in vivo*	0.39	[-0.009, 0.788]	0.409	[0.23, 0.589]	0.92498	1.00000	0.86742	1.00000
PBX1 (Homeobox)	*in vitro*	0.057	[-0.274, 0.387]	0.277	[0.009, 0.544]	0.33128	1.00000	0.10760	1.00000
PBX1 (Homeobox)	*in vivo*	0.589	[0.261, 0.918]	0.648	[0.361, 0.934]	0.77305	1.00000	0.53741	1.00000
Pbx3 (Homeobox)	*in vitro*	-0.018	[-0.285, 0.248]	0.077	[-0.083, 0.236]	0.51900	1.00000	0.29895	1.00000
Pbx3 (Homeobox)	*in vivo*	-0.4	[-0.771, -0.03]	0.151	[-0.157, 0.46]	0.01149	1.00000	0.03995	1.00000
Pdx1 (Homeobox)	*in vitro*	-0.431	[-0.607, -0.255]	-0.135	[-0.237, -0.033]	0.00907	1.00000	0.00221	1.00000
Pdx1 (Homeobox)	*in vivo*	-0.649	[-0.89, -0.408]	-0.116	[-0.29, 0.058]	0.00037	0.17905	0.00026	0.12652
Phox2a (Homeobox)	*in vitro*	-0.483	[-0.736, -0.231]	-0.508	[-0.712, -0.304]	0.88476	1.00000	0.97057	1.00000
Phox2a (Homeobox)	*in vivo*	-0.612	[-0.991, -0.232]	-0.63	[-0.86, -0.399]	0.93751	1.00000	0.39040	1.00000
Pitx1 (Homeobox)	*in vitro*	0.115	[-0.197, 0.426]	0.115	[-0.102, 0.332]	0.99904	1.00000	0.86742	1.00000
Pitx1 (Homeobox)	*in vivo*	0.089	[-0.225, 0.402]	0.672	[0.426, 0.917]	0.00326	1.00000	0.00024	0.11831
PPARE (NR), DR1	*in vitro*	0.143	[-0.029, 0.314]	0.084	[-0.003, 0.17]	0.54311	1.00000	0.25163	1.00000
PPARE (NR), DR1	*in vivo*	-0.261	[-0.519, -0.004]	0.028	[-0.12, 0.176]	0.05921	1.00000	0.00736	1.00000
PRDM14 (Zf)	*in vitro*	0.126	[-0.123, 0.374]	0.092	[-0.013, 0.198]	0.80661	1.00000	0.73287	1.00000
PRDM14 (Zf)	*in vivo*	-0.622	[-0.847, -0.397]	-0.257	[-0.523, 0.01]	0.01874	1.00000	0.02336	1.00000
PRDM1 (Zf)	*in vitro*	-0.366	[-0.518, -0.214]	-0.257	[-0.36, -0.154]	0.20208	1.00000	0.17078	1.00000
PRDM1 (Zf)	*in vivo*	-0.75	[-1.065, -0.436]	-0.41	[-0.647, -0.173]	0.10161	1.00000	0.11922	1.00000
PRDM9 (Zf)	*in vitro*	-0.028	[-0.271, 0.214]	0.228	[0.11, 0.346]	0.06155	1.00000	0.01381	1.00000
PRDM9 (Zf)	*in vivo*	0.083	[-0.26, 0.426]	0.261	[0.041, 0.481]	0.41153	1.00000	0.20555	1.00000

PR (NR)	*in vitro*	-0.042	[-0.191, 0.108]	-0.01	[-0.14, 0.121]	0.75398	1.00000	0.40591	1.00000
PR (NR)	*in vivo*	-0.422	[-0.745, -0.1]	-0.124	[-0.348, 0.101]	0.13319	1.00000	0.27135	1.00000
Ptf1a (bHLH)	*in vitro*	0.339	[0.055, 0.624]	0.227	[0.048, 0.406]	0.50650	1.00000	0.45813	1.00000
Ptf1a (bHLH)	*in vivo*	0.313	[-0.028, 0.654]	0.427	[0.187, 0.667]	0.58229	1.00000	0.49839	1.00000
PU.1-IRF (ETS:IRF)	*in vitro*	-0.381	[-0.533, -0.229]	-0.218	[-0.298, -0.138]	0.06816	1.00000	0.00892	1.00000
PU.1-IRF (ETS:IRF)	*in vivo*	-0.457	[-0.737, -0.177]	-0.311	[-0.491, -0.131]	0.40213	1.00000	0.40591	1.00000
PU.1 (ETS)	*in vitro*	-0.216	[-0.402, -0.03]	-0.108	[-0.232, 0.015]	0.37055	1.00000	0.46581	1.00000
PU.1 (ETS)	*in vivo*	-0.345	[-0.64, -0.05]	-0.165	[-0.339, 0.008]	0.33459	1.00000	0.11922	1.00000
RARg (NR)	*in vitro*	0.324	[0.026, 0.621]	0.486	[0.224, 0.748]	0.43376	1.00000	0.20060	1.00000
RARg (NR)	*in vivo*	1.316	[1.034, 1.598]	1.26	[1.042, 1.478]	0.74977	1.00000	0.98150	1.00000
REST-NRSF (Zf)	*in vitro*	1.484	[1.295, 1.673]	2.026	[1.805, 2.247]	0.00039	0.18771	0.00000	0.00184
REST-NRSF (Zf)	*in vivo*	3.222	[3.036, 3.407]	2.997	[2.914, 3.08]	0.02660	1.00000	0.21880	1.00000
Reverb (NR), DR2	*in vitro*	0.282	[0.054, 0.511]	0.226	[0.022, 0.43]	0.68777	1.00000	0.94118	1.00000
Reverb (NR), DR2	*in vivo*	0.378	[0.056, 0.7]	0.443	[0.211, 0.674]	0.72867	1.00000	0.94455	1.00000
Rfx1 (HTH)	*in vitro*	0.05	[-0.231, 0.331]	0.093	[-0.097, 0.283]	0.80527	1.00000	0.78797	1.00000
Rfx1 (HTH)	*in vivo*	-0.328	[-0.772, 0.116]	-0.064	[-0.317, 0.19]	0.24105	1.00000	0.08287	1.00000
Rfx2 (HTH)	*in vitro*	0.649	[0.286, 1.011]	0.645	[0.391, 0.898]	0.98383	1.00000	0.35609	1.00000
Rfx2 (HTH)	*in vivo*	1.33	[1.145, 1.515]	1.49	[1.3, 1.681]	0.18616	1.00000	0.04035	1.00000
Rfx5 (HTH)	*in vitro*	0.199	[-0.034, 0.432]	0.389	[0.233, 0.546]	0.23739	1.00000	0.35372	1.00000
Rfx5 (HTH)	*in vivo*	-0.345	[-0.704, 0.014]	-0.041	[-0.334, 0.252]	0.19625	1.00000	0.02741	1.00000
RFX (HTH)	*in vitro*	0.016	[-0.371, 0.403]	0.648	[0.405, 0.891]	0.00654	1.00000	0.00218	1.00000
RFX (HTH)	*in vivo*	0.721	[0.429, 1.012]	0.908	[0.654, 1.162]	0.32994	1.00000	0.28408	1.00000
RORgt (NR)	*in vitro*	-0.262	[-0.496, -0.028]	-0.032	[-0.251, 0.186]	0.14385	1.00000	0.07054	1.00000
RORgt (NR)	*in vivo*	-0.332	[-0.64, -0.024]	0.177	[-0.129, 0.483]	0.00646	1.00000	0.01472	1.00000
RUNX-AML (Runt)	*in vitro*	0.068	[-0.123, 0.259]	0.09	[0.001, 0.179]	0.84005	1.00000	0.40591	1.00000
RUNX-AML (Runt)	*in vivo*	-0.047	[-0.315, 0.222]	0.063	[-0.144, 0.269]	0.47630	1.00000	0.42178	1.00000
RUNX1 (Runt)	*in vitro*	0.009	[-0.178, 0.196]	-0.076	[-0.17, 0.017]	0.40526	1.00000	0.56097	1.00000
RUNX1 (Runt)	*in vivo*	-0.335	[-0.642, -0.028]	-0.03	[-0.189, 0.128]	0.07861	1.00000	0.06632	1.00000

RUNX2 (Runt)	*in vitro*	0.037	[-0.149, 0.224]	0.032	[-0.053, 0.117]	0.95866	1.00000	0.73982	1.00000
RUNX2 (Runt)	*in vivo*	-0.276	[-0.521, -0.03]	0.03	[-0.188, 0.248]	0.06028	1.00000	0.05578	1.00000
RUNX (Runt)	*in vitro*	-0.067	[-0.301, 0.166]	0.048	[-0.067, 0.163]	0.36354	1.00000	0.29636	1.00000
RUNX (Runt)	*in vivo*	-0.317	[-0.644, 0.009]	0.156	[-0.058, 0.369]	0.00700	1.00000	0.01659	1.00000
RXR (NR), DR1	*in vitro*	0.236	[0.074, 0.398]	0.188	[0.11, 0.267]	0.61042	1.00000	0.43256	1.00000
RXR (NR), DR1	*in vivo*	-0.317	[-0.624, -0.01]	0.013	[-0.185, 0.21]	0.07016	1.00000	0.04516	1.00000
SCL (bHLH)	*in vitro*	0.128	[-0.119, 0.375]	0.182	[-0.031, 0.395]	0.74517	1.00000	0.83833	1.00000
SCL (bHLH)	*in vivo*	-0.043	[-0.321, 0.235]	0.246	[-0.086, 0.577]	0.14405	1.00000	0.25398	1.00000
Six1 (Homeobox)	*in vitro*	-0.47	[-0.726, -0.214]	-0.165	[-0.269, -0.061]	0.05660	1.00000	0.14284	1.00000
Six1 (Homeobox)	*in vivo*	-0.476	[-0.863, -0.089]	-0.353	[-0.613, -0.092]	0.59593	1.00000	0.57159	1.00000
Smad2 (MAD)	*in vitro*	0.104	[-0.061, 0.27]	0.125	[0.027, 0.222]	0.83537	1.00000	0.78909	1.00000
Smad2 (MAD)	*in vivo*	-0.131	[-0.368, 0.107]	0.088	[-0.083, 0.259]	0.14762	1.00000	0.15130	1.00000
Smad3 (MAD)	*in vitro*	0.13	[0.009, 0.252]	0.006	[-0.096, 0.109]	0.11036	1.00000	0.12413	1.00000
Smad3 (MAD)	*in vivo*	-0.252	[-0.462, -0.043]	0.025	[-0.165, 0.214]	0.05154	1.00000	0.03820	1.00000
Smad4 (MAD)	*in vitro*	0.221	[0.08, 0.362]	0.142	[0.068, 0.216]	0.32581	1.00000	0.37029	1.00000
Smad4 (MAD)	*in vivo*	-0.022	[-0.225, 0.182]	0.058	[-0.11, 0.226]	0.47833	1.00000	0.64457	1.00000
Sox2 (HMG)	*in vitro*	-0.195	[-0.369, -0.022]	-0.245	[-0.332, -0.157]	0.62044	1.00000	0.88268	1.00000
Sox2 (HMG)	*in vivo*	-0.703	[-1.018, -0.389]	-0.267	[-0.449, -0.085]	0.01126	1.00000	0.00484	1.00000
Sox3 (HMG)	*in vitro*	-0.073	[-0.237, 0.09]	-0.195	[-0.31, -0.08]	0.18879	1.00000	0.21179	1.00000
Sox3 (HMG)	*in vivo*	-0.553	[-0.838, -0.267]	-0.14	[-0.298, 0.019]	0.01232	1.00000	0.01986	1.00000
Sox6 (HMG)	*in vitro*	-0.339	[-0.498, -0.179]	-0.097	[-0.2, 0.006]	0.00941	1.00000	0.01534	1.00000
Sox6 (HMG)	*in vivo*	-0.334	[-0.695, 0.028]	-0.162	[-0.328, 0.003]	0.40469	1.00000	0.19099	1.00000
Sp1 (Zf)	*in vitro*	0.146	[-0.24, 0.532]	0.246	[0.012, 0.479]	0.68028	1.00000	0.44421	1.00000
Sp1 (Zf)	*in vivo*	0.037	[-0.279, 0.353]	0.38	[0.098, 0.663]	0.08756	1.00000	0.09596	1.00000
SPDEF (ETS)	*in vitro*	-0.104	[-0.296, 0.087]	-0.022	[-0.097, 0.054]	0.40740	1.00000	0.38334	1.00000
SPDEF (ETS)	*in vivo*	-0.128	[-0.403, 0.148]	0.068	[-0.099, 0.236]	0.23306	1.00000	0.07187	1.00000
Srebp1a (bHLH)	*in vitro*	0.285	[0.044, 0.526]	0.359	[0.223, 0.495]	0.57864	1.00000	0.57349	1.00000
Srebp1a (bHLH)	*in vivo*	-0.099	[-0.52, 0.323]	0.033	[-0.215, 0.28]	0.59241	1.00000	0.96671	1.00000

Srebp2 (bHLH)	*in vitro*	0.378	[0.132, 0.623]	0.418	[0.266, 0.57]	0.78733	1.00000	0.56844	1.00000
Srebp2 (bHLH)	*in vivo*	-0.182	[-0.522, 0.158]	0.237	[-0.019, 0.494]	0.03673	1.00000	0.04929	1.00000
STAT1 (Stat)	*in vitro*	-1.093	[-1.404, -0.781]	-0.367	[-0.693, -0.04]	0.00186	0.89803	0.00833	1.00000
STAT1 (Stat)	*in vivo*	-0.172	[-0.533, 0.189]	-0.066	[-0.366, 0.234]	0.62385	1.00000	0.97411	1.00000
Stat3+il21 (Stat)	*in vitro*	-0.174	[-0.46, 0.112]	-0.107	[-0.291, 0.077]	0.67140	1.00000	0.61830	1.00000
Stat3+il21 (Stat)	*in vivo*	-0.31	[-0.617, -0.002]	-0.275	[-0.545, -0.006]	0.86472	1.00000	0.95562	1.00000
Stat3 (Stat)	*in vitro*	-0.429	[-0.722, -0.137]	-0.114	[-0.281, 0.053]	0.08833	1.00000	0.04929	1.00000
Stat3 (Stat)	*in vivo*	-0.404	[-0.782, -0.026]	-0.276	[-0.566, 0.013]	0.56847	1.00000	0.44144	1.00000
STAT4 (Stat)	*in vitro*	-0.391	[-0.695, -0.086]	-0.419	[-0.636, -0.201]	0.86645	1.00000	0.81665	1.00000
STAT4 (Stat)	*in vivo*	-0.201	[-0.504, 0.101]	-0.103	[-0.343, 0.136]	0.59445	1.00000	0.20388	1.00000
STAT5 (Stat)	*in vitro*	-0.789	[-1.176, -0.402]	-0.442	[-0.781, -0.103]	0.18954	1.00000	0.09505	1.00000
STAT5 (Stat)	*in vivo*	0.211	[-0.161, 0.584]	0.303	[-0.072, 0.679]	0.72376	1.00000	0.38585	1.00000
STAT6 (Stat)	*in vitro*	-0.802	[-1.099, -0.506]	-0.64	[-0.88, -0.4]	0.38470	1.00000	0.33946	1.00000
STAT6 (Stat)	*in vivo*	-0.384	[-0.755, -0.013]	-0.466	[-0.866, -0.065]	0.77267	1.00000	0.33479	1.00000
T1ISRE (IRF)	*in vitro*	-0.016	[-0.209, 0.177]	0.359	[0.111, 0.608]	0.03204	1.00000	0.01844	1.00000
T1ISRE (IRF)	*in vivo*	1.45	[1.36, 1.539]	1.666	[1.502, 1.831]	0.02561	1.00000	0.02917	1.00000
TATA-Box (TBP)	*in vitro*	-0.287	[-0.463, -0.111]	-0.207	[-0.334, -0.08]	0.48532	1.00000	0.56844	1.00000
TATA-Box (TBP)	*in vivo*	-0.915	[-1.256, -0.573]	-0.601	[-0.799, -0.403]	0.12006	1.00000	0.11596	1.00000
Tbet (T-box)	*in vitro*	-0.144	[-0.299, 0.012]	-0.033	[-0.113, 0.047]	0.25038	1.00000	0.05975	1.00000
Tbet (T-box)	*in vivo*	-0.319	[-0.577, -0.06]	-0.201	[-0.414, 0.013]	0.48184	1.00000	0.44901	1.00000
Tbox:Smad (T-box, MAD)	*in vitro*	0.084	[-0.179, 0.347]	0.343	[0.212, 0.473]	0.09166	1.00000	0.06341	1.00000
Tbox:Smad (T-box, MAD)	*in vivo*	-0.052	[-0.368, 0.264]	0.378	[0.135, 0.622]	0.01302	1.00000	0.01363	1.00000
Tbx20 (T-box)	*in vitro*	0.135	[-0.137, 0.407]	0.126	[-0.027, 0.279]	0.95066	1.00000	0.47720	1.00000
Tbx20 (T-box)	*in vivo*	-0.177	[-0.541, 0.187]	0.344	[0.133, 0.556]	0.02853	1.00000	0.08453	1.00000
Tbx5 (T-box)	*in vitro*	0.196	[0.009, 0.383]	0.062	[-0.084, 0.209]	0.32002	1.00000	0.22975	1.00000
Tbx5 (T-box)	*in vivo*	-0.12	[-0.426, 0.186]	0.039	[-0.181, 0.26]	0.41434	1.00000	0.42512	1.00000
Tcf12 (bHLH)	*in vitro*	0.349	[0.125, 0.572]	0.233	[0.088, 0.377]	0.42003	1.00000	0.31833	1.00000
Tcf12 (bHLH)	*in vivo*	0.295	[-0.057, 0.647]	0.41	[0.164, 0.657]	0.60550	1.00000	0.51622	1.00000

Tcf3 (HMG)	*in vitro*	-0.345	[-0.571, -0.12]	-0.153	[-0.285, -0.021]	0.13234	1.00000	0.19737	1.00000
Tcf3 (HMG)	*in vivo*	-0.495	[-0.909, -0.08]	-0.221	[-0.517, 0.075]	0.18196	1.00000	0.18580	1.00000
Tcf4 (HMG)	*in vitro*	-0.197	[-0.389, -0.005]	-0.143	[-0.254, -0.032]	0.62701	1.00000	0.28410	1.00000
Tcf4 (HMG)	*in vivo*	-0.503	[-0.845, -0.161]	-0.062	[-0.299, 0.175]	0.02120	1.00000	0.04875	1.00000
Tcfcp2l1 (CP2)	*in vitro*	0.19	[-0.136, 0.517]	0.422	[0.162, 0.683]	0.29213	1.00000	0.25940	1.00000
Tcfcp2l1 (CP2)	*in vivo*	0.2	[-0.142, 0.542]	0.616	[0.336, 0.897]	0.10186	1.00000	0.00510	1.00000
TCFL2 (HMG)	*in vitro*	-0.331	[-0.646, -0.016]	-0.186	[-0.406, 0.035]	0.39463	1.00000	0.14283	1.00000
TCFL2 (HMG)	*in vivo*	0.022	[-0.338, 0.381]	0.164	[-0.086, 0.414]	0.54297	1.00000	0.11276	1.00000
TEAD2 (TEA)	*in vitro*	-0.009	[-0.182, 0.165]	-0.055	[-0.131, 0.021]	0.61807	1.00000	0.35613	1.00000
TEAD2 (TEA)	*in vivo*	-0.64	[-0.874, -0.406]	0.057	[-0.134, 0.248]	0.00000	0.00052	0.00001	0.00426
TEAD4 (TEA)	*in vitro*	-0.102	[-0.269, 0.066]	-0.131	[-0.228, -0.034]	0.73158	1.00000	0.59891	1.00000
TEAD4 (TEA)	*in vivo*	-0.541	[-0.778, -0.304]	-0.133	[-0.329, 0.062]	0.01978	1.00000	0.00300	1.00000
TEAD (TEA)	*in vitro*	-0.05	[-0.223, 0.123]	-0.071	[-0.182, 0.04]	0.82600	1.00000	0.93385	1.00000
TEAD (TEA)	*in vivo*	-0.754	[-1.114, -0.394]	-0.239	[-0.52, 0.042]	0.02057	1.00000	0.01082	1.00000
THRa (NR)	*in vitro*	0.182	[-0.002, 0.367]	0.27	[0.153, 0.388]	0.42632	1.00000	0.12643	1.00000
THRa (NR)	*in vivo*	-0.166	[-0.573, 0.242]	0.122	[-0.082, 0.326]	0.19743	1.00000	0.43800	1.00000
Tlx (NR)	*in vitro*	0.037	[-0.185, 0.26]	0.364	[0.255, 0.473]	0.00687	1.00000	0.00126	0.60798
Tlx (NR)	*in vivo*	0.125	[-0.264, 0.513]	0.29	[0.071, 0.509]	0.51833	1.00000	0.24825	1.00000
TR4 (NR), DR1	*in vitro*	-0.296	[-0.661, 0.069]	-0.012	[-0.235, 0.21]	0.19267	1.00000	0.17426	1.00000
TR4 (NR), DR1	*in vivo*	-0.286	[-0.66, 0.088]	0.263	[-0.059, 0.584]	0.03094	1.00000	0.00640	1.00000
USF1 (bHLH)	*in vitro*	0.181	[-0.027, 0.389]	0.145	[0.001, 0.288]	0.76862	1.00000	0.91875	1.00000
USF1 (bHLH)	*in vivo*	0.098	[-0.257, 0.452]	0.277	[0.046, 0.509]	0.38053	1.00000	0.34134	1.00000
Usf2 (bHLH)	*in vitro*	0.183	[-0.057, 0.423]	0.171	[0.059, 0.283]	0.93624	1.00000	0.97780	1.00000
Usf2 (bHLH)	*in vivo*	0.039	[-0.245, 0.323]	0.308	[0.096, 0.519]	0.09742	1.00000	0.09505	1.00000
VDR (NR), DR3	*in vitro*	-0.078	[-0.332, 0.177]	0.189	[0.053, 0.325]	0.04284	1.00000	0.06166	1.00000
VDR (NR), DR3	*in vivo*	-0.173	[-0.53, 0.184]	0.268	[0.055, 0.48]	0.03122	1.00000	0.01793	1.00000
X-box (HTH)	*in vitro*	-0.201	[-0.612, 0.209]	0.354	[0.178, 0.531]	0.02559	1.00000	0.01590	1.00000
X-box (HTH)	*in vivo*	0.079	[-0.353, 0.511]	0.282	[-0.078, 0.642]	0.25354	1.00000	0.19736	1.00000

YY1 (Zf)	*in vitro*	-0.921	[-1.341, -0.502]	-0.595	[-0.973, -0.216]	0.23605	1.00000	0.20506	1.00000
YY1 (Zf)	*in vivo*	-0.196	[-0.683, 0.291]	-0.476	[-0.745, -0.207]	0.28838	1.00000	0.35854	1.00000
ZBTB33 (Zf)	*in vitro*	0.997	[0.75, 1.244]	1.068	[0.881, 1.255]	0.65650	1.00000	0.07382	1.00000
ZBTB33 (Zf)	*in vivo*	2.266	[2.181, 2.352]	2.554	[2.387, 2.721]	0.00100	0.48545	0.00030	0.14449
ZFX (Zf)	*in vitro*	0.194	[-0.012, 0.4]	0.459	[0.331, 0.588]	0.04352	1.00000	0.03907	1.00000
ZFX (Zf)	*in vivo*	-0.322	[-0.617, -0.028]	0.271	[-0.004, 0.545]	0.00244	1.00000	0.01703	1.00000
ZNF143|STAF (Zf)	*in vitro*	0.195	[0.004, 0.386]	0.168	[0.023, 0.312]	0.82206	1.00000	0.73149	1.00000
ZNF143|STAF (Zf)	*in vivo*	-0.163	[-0.507, 0.181]	-0.015	[-0.265, 0.235]	0.45913	1.00000	0.64952	1.00000
Znf263 (Zf)	*in vitro*	0.322	[0.169, 0.476]	0.304	[0.212, 0.396]	0.78664	1.00000	0.92652	1.00000
Znf263 (Zf)	*in vivo*	-0.148	[-0.4, 0.103]	-0.011	[-0.229, 0.207]	0.40457	1.00000	0.33201	1.00000
ZNF711 (Zf)	*in vitro*	0.334	[0.136, 0.532]	0.434	[0.269, 0.599]	0.46335	1.00000	0.63138	1.00000
ZNF711 (Zf)	*in vivo*	-0.227	[-0.573, 0.12]	-0.148	[-0.406, 0.109]	0.67821	1.00000	0.84558	1.00000