This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Lagging Strand Replication Creates Evolutionary Hotspots Throughout The Genome

By

HARRIET KEMP

College of Medicine and Veterinary Medicine
UNIVERSITY OF EDINBURGH

A thesis submitted to the University of Edinburgh in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the College of Medicine and Veterinary Medicine.

SEPTEMBER 2015

Word count: twenty eight thousand
I declare that the work in this dissertation was carried out in accordance with the requirements of the University’s Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate’s own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .................................................... DATE: ..........................................
ACKNOWLEDGEMENTS

First and foremost I would like to acknowledge and thank Martin Taylor for his supervision during the course of my PhD. It cannot be stressed enough that his inspiration, enthusiasm and guidance have been vital to its instigation, progression and outcome. I would like to thank Grzegorz Kudla, Richard Baldock and James Prendergast for their input as my committee panel, their helpful feedback and discussions.

Joanna Pethick has been a significant and treasured ally throughout the course of this project, and has been a key support-network in many understated ways. I would also like to acknowledge the appreciated encouragement of Robert Young, Sara Perricone and Sarah Rennie.
The rate of DNA mutation is known to fluctuate across the genome but the patterns of mutation rate variation and molecular causes are poorly defined. It is important to understand these patterns of mutation as they influence where deleterious mutations are likely to arise and how rapidly sequences are likely to accumulate change between species, a measure often used as a proxy for functional constraint. In this work I investigate the relationship between DNA replication and apparent mutation hotspots adjacent to transcription factor binding sites. In eukaryotes both DNA strands are replicated simultaneously, the leading strand as a continuous stretch and the lagging strand as a series of discrete Okazaki fragments that are subsequently ligated together. Some transcription factors are able to bind the DNA lagging strand during replication and act as a partial barrier to DNA polymerase, resulting in the accumulation of Okazaki fragment junctions adjacent to these sites. I find that mutation rate is correlated genome wide with Okazaki junction frequency, suggesting that Okazaki junction processing may be error-prone. We present a mechanistic hypothesis to explain this locally elevated mutation rate and propose a role for lagging strand replication and its error-prone Pol α tract retention in the formation of these hotspots. I test this hypothesis using Okazaki fragment sequencing data from the yeast *Saccharomyces cerevisiae* to identify peaks in Okazaki junctions. When these peaks are aligned and orientated, so that the direction of lagging strand replication is uniform, I find a peak in substitution rate immediately downstream of Okazaki junctions, precisely where Pol α tract retention is predicted to occur. Novel binding motifs are identified within the underlying DNA of these junctions that can be assigned to known strong and fast-binding transcription factors, previously implicated in the phasing of nucleosomes, such as Reb1. I show that mutation hotspots adjacent to transcription factor binding sites are a conserved feature of eukaryotic genomes. In the human genome I predict sites of preferential Pol α retention using DNase I hypersensitivity footprint data. We observe that those footprints predicted as germline-specific manifest an elevated mutation signature. I propose that the rapid binding of some transcription factors to DNA following replication is required for nucleosome positioning or other important functions, however this incurs a cost in terms of locally elevated mutation rate adjacent to and within the sequence specific binding site. As a consequence these binding sites are biologically important mutational hotspots whose functional significance has been systematically underestimated by standard measures of sequence constraint.
# Table of Contents

| List of Tables | xi |
| List of Figures | xiii |

## 1 Introduction

1.1 The structure of the eukaryotic genome .......................... 3
   1.1.1 The genome and its building blocks ............................. 3
   1.1.2 The basics of cell structure, cell division and DNA replication ........ 5
   1.1.3 Protein coding DNA ........................................... 6
   1.1.4 Non-protein coding DNA ................................------- 8

1.2 Selection .................................................. 10
   1.2.1 Sites evolving approximately neutrally .......................... 10
   1.2.2 Purifying selection ........................................... 11
   1.2.3 Positive selection ........................................... 12
   1.2.4 Measuring selection ......................................... 13

1.3 Mutation rate ............................................. 15
   1.3.1 Genomic mutation rate ....................................... 15
   1.3.2 The effect of CpG dinucleotides and G + C content on mutation rate .... 15
   1.3.3 The effect of replication timing and replication strand on mutation rate ... 17
   1.3.4 Transcription associated mutagenesis .......................... 17
   1.3.5 Promoter mutation rate ..................................... 18
   1.3.6 Male biased mutation rate .................................... 19
   1.3.7 The nucleosome hump ........................................ 19

1.4 Substitution rate ........................................... 20
   1.4.1 Interpreting substitution rate ................................ 20
   1.4.2 From counting between-species changes to HKY .................. 22
   1.4.3 phastCons ................................................. 23
   1.4.4 GERP ..................................................... 25
   1.4.5 Polymorphism rate .......................................... 25

1.5 Research objectives ........................................ 27
TABLE OF CONTENTS

4.3.2 Okazaki junctions and polymorphism rate track neatly at Reb1 and Rap1 binding sites ............................................. 64
4.4 The lagging strand hypothesis: a discussion ............................................. 66
  4.4.1 Yeast transcription factor conservation ............................................. 66
  4.4.2 Intrinsic coupling of Okazaki junction and nucleotide substitution rates 67

5 Okazaki junction Super-peaks ......................................................... 69
  5.1 Super-peaks: data and methodologies ............................................. 70
    5.1.1 Defining Okazaki junction super-peaks ............................................. 70
    5.1.2 Orientating polymorphism rate at Okazaki junction super-peaks .......... 70
    5.1.3 Annotating genomic details to super-peaks ............................................. 71
    5.1.4 Motif discovery at super-peaks ............................................. 71
    5.1.5 Sub-dividing super-peaks by underlying motifs ............................................. 72
  5.2 Super-peaks: results ............................................................. 72
    5.2.1 Identifying Okazaki junction super-peaks ............................................. 72
    5.2.2 Polarisation by direction of replication creates conservation asymmetry . . 74
    5.2.3 De novo identification of underlying super-peak motifs ............................................. 74
    5.2.4 Assigning binding motifs to known transcription factors ............................................. 76
    5.2.5 Comparison of annotations between super-peak subcategories ................. 77
  5.3 Super-peaks: a discussion ......................................................... 82
    5.3.1 Super-peaks as a model for the Pol α tract retention hypothesis ................. 82
    5.3.2 Localisation of sequence motifs at super-peak sites ............................................. 83

6 DNase I Footprints ........................................................................ 87
  6.1 Footprints: an introduction ............................................................. 87
  6.2 Footprints: data and methodologies ............................................. 90
    6.2.1 Calculating DHS and footprint coverage at DHS footprint edges ................. 90
    6.2.2 Calculating between-species substitution rate at DHS footprint edges ................. 90
  6.3 Footprints: results ............................................................. 91
    6.3.1 A shoulder of substitution rate at human footprint edges ............................................. 91
  6.4 Footprints: a discussion ............................................................. 92
    6.4.1 Footprints as a model for the Pol α tract retention hypothesis ................. 92
    6.4.2 Considering evolutionary forces at shoulders ............................................. 93

7 Mutation versus selection: soma and the germline ............................................. 95
  7.1 Mutation versus selection: an introduction ............................................. 95
  7.2 Mutation versus selection: data and methodologies ............................................. 97
    7.2.1 Splitting footprints into cell type categories ............................................. 97
    7.2.2 Calculation of additional annotations ............................................. 97
TABLE OF CONTENTS

7.3 Mutation versus selection: results ........................................... 99  
  7.3.1 Increased substitution rate at germline-associated footprints .......... 99  
  7.3.2 Regional effects of substitution between cell type genres ............. 102  
  7.3.3 Germline-specific correlation between CpG dinucleotide frequency and footprint coverage .................................................. 104  
7.4 Mutation versus selection: a discussion ................................... 108  
  7.4.1 Housekeeping bias in elevated substitution rate ....................... 108  
  7.4.2 CpG effects are not responsible for substitution shoulders .......... 109  
  7.4.3 Germline DNA binding and the broader substitution rate landscape .... 109  
8 Conclusions ............................................................................. 111  
  8.1 Conclusions: key findings ...................................................... 111  
  8.2 Conclusions: implications ...................................................... 112  
  8.3 Conclusions: future direction .................................................. 116  
    8.3.1 Super-peaks ................................................................... 116  
    8.3.2 Human Okazaki fragment data ......................................... 116  
    8.3.3 Derived Allele Frequencies ............................................. 117  
    8.3.4 CTCF binding profiles .................................................... 117  
  8.4 Conclusions: supporting material ............................................ 120  
A Data sources ............................................................................ 125  
  A.1 Human and yeast transcription factor ChIP-seq data .................... 125  
  A.2 DNase hypersensitivity site coverage ...................................... 129  
  A.3 Methyl CpG Reduced Representation Bisulfite Seq tracks ............ 131  
B Manuscripts ............................................................................. 133  
Bibliography ................................................................................. 153
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Summary of data sources.</td>
<td>37</td>
</tr>
<tr>
<td>3.1 Transcription factors with known binding motif and available ChIP-seq alignment data. Class and family provided by Jaspar (<a href="http://jaspar.genereg.net/">http://jaspar.genereg.net/</a>) [75]</td>
<td>45</td>
</tr>
<tr>
<td>7.1 List of footprint cell types, their corresponding categories and genres. Annotations provided by Thurman et al. [117]</td>
<td>98</td>
</tr>
<tr>
<td>A.1 List of transcription factor ChIP-seq cell types and sources</td>
<td>125</td>
</tr>
<tr>
<td>A.2 List of DHS coverage data sources</td>
<td>129</td>
</tr>
<tr>
<td>A.3 List of Methyl CpG RRBS data sources</td>
<td>131</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 The structure of DNA.</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Basic mitosis and genome replication.</td>
<td>6</td>
</tr>
</tbody>
</table>

1.1 **The structure of DNA.** [a] In its simplest form DNA is made from two nucleotide strands tightly wound in a double stranded helix. Each nucleotide is made of a pentose carbon sugar (purple), an adenine (A; green), thymine (T; red), cytosine (C; blue) or guanine (G; yellow) nitrogenous base, and a phosphate group (orange), which connects each sugar to form the strand’s phosphate backbone. Each nucleotide and subsequent strand is polarised from 3’ to 5’, with each strand antiparallel to the other. Nucleotides are matched in complementary base pairs across strands: A to T and C to G. [b] DNA is wrapped twice around clusters of eight histone proteins to form nucleosomes. Nucleosome strings of DNA undergo further coiling to produce a chromatin fibre, which itself undergoes extensive packaging to form the chromosome.

1.2 **Basic mitosis and genome replication.** [a] During cell division the cell’s DNA must be replicated and split into separate nuclear compartments (pink) by mitosis. Each chromosome of DNA (purple and green) is first replicated (interphase), tightly packaged and condensed (prophase), and aligned along the cell’s centre by spindle fibres (metaphase). The spindle fibres separate the two copies of each chromosome by pulling in opposite directions towards the poles of the cell’s nucleus (anaphase). The cell’s singular nucleus then divides into two (telophase), followed by the remaining cell. [b] During DNA replication the cell’s replication machinery prises open multiple sections of single stranded DNA at origins of replication (orange). Replication continues bidirectionally along the length of each chromosome, replicating each strand of double-stranded DNA simultaneously (green and blue).
1.3 **Protein synthesis.** DNA is converted to pre-mRNA by RNA polymerase in the cell nucleus. The pre-mRNA molecule is complementary to its DNA template, in the same way two DNA strands are complementary to each other in the double stranded DNA helix. The exception of the RNA code is uracil (U), which pairs with A nucleotides (in place of T). The pre-mRNA molecule is processed, to remove untranslated regions, to produce mature mRNA. mRNA is exported from the nucleus to the cytoplasm of the cell where it is recruited by ribosomes. Ribosomes read the mRNA in three nucleotide codons, simultaneously pairing them with their corresponding amino acid to produce the protein encoded by the DNA template. ................................................................. 7

1.4 **Functional consequences of substitutions.** The functional consequence of an ‘A’ to ‘G’ substitution will vary within the genome, depending on the functionality of the DNA it is found within. Substitutions within non-functional DNA, such as introns and intergenic regions, are unlikely to be of functional importance, along with substitutions at synonymous sites within functionally active protein-coding DNA. Substitutions at non-synonymous sites within protein-coding DNA are likely to have a strong functional consequence, as well as substitutions within transcription factor binding sites at promoters and enhancers. The effect of purifying selection will reflect that of functional importance. Sites within regions of low functional importance, under a low influence of purifying selection, reflect genomic neutrality. ................................. 12

1.5 **The principles of Derived Allele Frequencies (DAFs).** Polymorphisms within non-functionally important sites drift within a population, increasing or decreasing in frequency purely by chance. There will therefore be a large number of polymorphisms of low frequency within the population, and a small number of polymorphisms of high frequency within the population. These sites are considered to be evolving approximately neutrally (brown). Polymorphisms within functionally important sites will also drift within a population, increasing or decreasing in frequency primarily by chance, but a small number will be removed or fixed within a population due to the fitness they confer to the organism. Sites undergoing positive selection will have more polymorphisms of high frequency within the population and less polymorphisms of low frequency within the population than a site evolving approximately neutrally (turquoise). Conversely, sites undergoing purifying selection will have less polymorphisms of high frequency within the population and more polymorphisms of low frequency within the population than a site evolving approximately neutrally (yellow). The DAFs of a test sample, such as transcription factor binding sites, can be compared to sites of known neutrality to infer the mechanism of selection acting upon them. .................................................................................................................. 14
1.6 **CpG deamination.** [a] deamination of an unmethylated C nucleotide in a CpG dinucleotide forms Uracil (U). U nucleotides, which are foreign to DNA, are recognised by uracil glycosylase, flagging the erroneous U nucleotides for removal and repair to C by base excision repair. [b] deamination of a methylated C nucleotide in a CpG dinucleotide forms T. As T is a native DNA component it will not be flagged as foreign. During replication the template T will be paired with A, resulting in a CG to TA substitution.

1.7 **Interpretation of substitution rate at the nucleosome hump.** Substitution rate (brown), increases where nucleosomes are bound to the DNA and is reduced in linker regions. This could be explained by purifying selection, whereby linker regions are of higher functional importance and so mutations are removed at higher efficiency from these sites than in nucleosome-bound regions where mutations are permitted to accumulate. This model assumes a uniform mutation rate. Alternatively, such a pattern in substitution could be explained by increased mutation rate at nucleosome-bound DNA. In this model it would be the nucleosome-bound DNA that would be of greatest interest for the identification of functional variants, as opposed to the linker regions. This demonstrates how the misinterpretation of the evolutionary forces acting at these sites will lead to the misinterpretation of those variants underlying them. (Figure adapted from Semple and Taylor [105]).

1.8 **Measures of nucleotide substitution rate probability.** [a] Jukes-Cantor assigns equal probability to all nucleotide changes ($\alpha$). [b] HKY accounts for an individual rate of change for each specific nucleotide. $\pi_A$, $\pi_T$, $\pi_C$ and $\pi_G$ denote the rate of change for A, T, C and G nucleotides respectively, with $\kappa$ weighting transversion substitutions.

1.9 **The phastCons model.** phastCons calculates substitution rate between species using fixed tree topologies. Each site is tested against its corresponding tree but where branch lengths are shortened, representing conservation (left), or lengthened, representing lack of conservation (centre). If the frequency of variation at a site matches the branch length in the conserved model it will be scored 1, while if it matches the branch length in the nonconserved model it will be scored 0. The majority of sites will fall between the two states and be scored accordingly. This is represented by the state-transition model (right), where each state, conserved (c), or non-conserved (n), is associated with a phylogenetic model ($\Psi_c$ and $\Psi_n$, respectively). These models are identical except for their scaling. $\mu$ and $\nu$ ($0 \leq \mu, \nu \leq 1$) define all state-transition probabilities. (Figure adapted from Siepel et al. [106]).
1.10 **The GERP model.** Substitution rates are calculated on a per-nucleotide basis between primates (green), placental mammals (purple) and marsupials (orange; left). Each column (top right), represents a single aligned nucleotide position between species. For each column an observed rate of variation is calculated according to known nucleotide variations and species divergence. An expected rate is calculated for each column by summing the remaining branches from a neutral tree model, after removing species with a gap character. Black, blue and red values in expected ratios correspond to the correlated neutral trees (bottom right). Each nucleotide position is scored as the sum of deviations from expectation at each site, thereby representing a value of rejected substitutions (R). (Figure adapted from Cooper et al. [17]). 26

1.11 **High-resolution analysis of transcription factor binding sites.** Substitution rate profiles at human transcription factor binding sites, represented by phastCons scores. [a] the substitution rate profile for combined human transcription factors (black), the average substitution rate across the genome (green), and the substitution rate pattern when centred purely on DNase I Hypersensitivity Sites (DHS; red). [b] the substitution rate profile for the single human transcription factor Nuclear Transcription Factor Y (NFYA). A site of low substitution rate lies directly at the transcription factor binding site, flanked by distinct peaks in substitution rate, (highlighted in yellow), ~ 10 nucleotides long. A further modest decrease in substitution before a gradual increase back to the genomic background level produces a “shoulder” in substitution. (Figure adapted from Boyle et al. [8]). 28

1.12 **Selective and mutational models for explaining the conservation shoulder.** Conservation “shoulders” can be explained by two alternative, non-mutually exclusive models: selection (left) and mutation (right). In both models substitution rate (brown) is the combined effect of mutation rate (red) and purifying selection (green). Blue spheres represent hypothetical factors bound to DNA. Weak to strong factor-DNA interactions are shown by pink bars (light to dark). Under the selective model substitution rate is explained by uniform mutation rate and selection which reflects the strength of factor-DNA binding interaction. In this example selection is reduced immediately adjacent to central binding interactions due to steric hindrance. Under the mutational model selection is strongest at core binding interactions and gradually returns to the genomic background rate as factor-DNA interactions reduce. In this model mutation rate is not uniform, but instead increases either side of the central binding interaction (red asterisks). 29
3.1 **Substitution rate profile at NFYA binding sites.** NFYA binding sites are centred on their motif mid-point and orientated by motif strand, so that for all sites motifs are aligned and in the same orientation (illustrated by the regional motif sequence strength by PWM, top plot). Sites are split into four groups of equal numbers according to their binding strength (ChIP-Seq coverage). Substitution rate (GERP) for each group is plotted in red, from light to dark (weakest to strongest binding strength). Genomic neutrality (y=0; based on genome-wide 4D sites, the implications of which are discussed in sections 1.4.4 and 3.4.2), is illustrated by the green dotted line. Bottom and top x-axes differ. The top plot represents the strongest ChIP-Seq coverage quartile at increased magnification, matching that of its above PWM.

3.2 (one of six.)

4.1 **DNA replication and its polymerases.** Replication extends from origins of replication continuously on the leading strand, mediated by Pol $\epsilon$. To comply with 5’ to 3’ replication the lagging strand is replicated discontinuously, forming short Okazaki fragments. Okazaki fragments are initiated by Pol $\alpha$, first synthesising an RNA primer (purple) followed by an error-prone stretch of DNA (blue). Pol $\alpha$ is replaced by Pol $\delta$ which continues DNA synthesis with increased fidelity. Pol $\alpha$ tracts are recognised and removed by the Pol $\alpha$ processing enzyme FEN-1, leaving a one nucleotide gap in their place. Neighbouring Okazaki fragments are then joined by DNA Ligase I.

4.2 **Strand-specific Okazaki fragment alignment.** [a] distribution of paired-end sequencing hits mapped to either the Watson (blue) or Crick (orange) strands across S. cerevisiae chromosome 10. Known origins of replication are indicated as grey vertical dashed lines [86]. [bd] Okazaki 5’ (red) and 3’ (blue) termini, at bound nucleosomes in [b] wild type, [c] reduced processivity Pol32 knockout strain, and [d] at the binding sites of the transcription factors Abd, Reb1 and Rap1. All sites are orientated by the direction of Okazaki synthesis. (Figures adapted from Smith and Whitehouse [108]).

4.3 **The lagging strand hypothesis: substitution rate directed by Okazaki polarisation.** Replication is illustrated in top schematic moving through double stranded DNA from right to left by Pol $\epsilon$ on the leading DNA strand and both Pol $\alpha$ (purple to blue) and Pol $\delta$ (black) on the lagging strand. Factors (cylinders) are bound at Okazaki junctions prior to ligation and involved in upstream Pol $\delta$ dissociation. If factors were to bind marginally upstream of the downstream Okazaki fragment’s Pol $\alpha$: Pol $\delta$ junction the remaining portion of the error-prone Pol $\alpha$ tract would be retained in the newly synthesised DNA (illustrated in central diagram, boxed). Over evolutionary time this process would accumulate mutations at such sites. Such a substitution rate would be asymmetrical when sites are orientated by direction of lagging strand replication, as Pol $\alpha$ tracts will only be retained downstream of Okazaki junctions (bottom plot; orange).
4.4 Coverage bias between Okazaki frequency and Okazaki junction rate. [a] forward (blue) and reverse strand (orange) Okazaki junction rate across chromosome 10. Autonomous Replicating Sequence (ARS) Consensus Sequence (ACS) origins of replication are shown in red. [b] average log2 ratio between Watson and Crick strand coverage is correlated to directionality of origins of replication, as represented by ACS [24]. [c] average log2 ratio between forward and reverse strand Okazaki junction rates diminishes replication origin directionality. All data unsmoothed. . . . . . . . . . 65

4.5 Polymorphism rate and Okazaki junction rate at Reb1 and Rap1 binding site peaks. [a] Reb1 and [b] Rap1 primary and monomer peaks with mean read coverage >30 are aligned to their binding motif mid-points. Orientation is arbitrary. Both polymorphism rate (red) and Okazaki junction rate (blue) are positioned and scaled so that the mean value over the viewed window matches the dashed grey line, and 10% deviations from the mean correspond to dotted grey lines. The central 11nt -5 to 5, were excluded from the calculation of distribution means as the curves clearly diverge over the sequence specific binding site nucleotides. As well as unsmoothed polymorphism values (points), values are plotted as smoothed (curve, spline 34 degrees of freedom). Okazaki junction rate is unsmoothed. . . . . . . . . . . . . . . . . . . . . . 66

5.1 Uni- and bi-directional super-peak distributions. [a] super-peaks (blue vertical line), are categorised as uni-directional if a peak surpasses the Okazaki junction rate threshold of 3.47% (grey shading), only on the forward strand (orange), or only on the reverse strand (blue), within a 20 nucleotide exclusion zone (vertical black lines). [b] super-peaks are categorised as bi-directional if peaks surpassing 3.47% significance are present on both the forward and reverse strand within 20 nucleotides. Both examples are super-peaks from chromosome 10. [c] strand ratios are calculated by dividing the number of Okazaki fragments on the forward strand by the number of Okazaki fragments on the reverse strand. The absolute log value is plotted for uni-directional and bi-directional super-peaks. Bean plots visualise the distribution of strand ratios (curved volume) as well as individual data points (thin horizontal bars). The dotted horizontal line represents the mean value of the whole data set, while the mean of each category is shown as a thick black line. As the maximum sample input for each category is limited to 5000 values, a random sample has been extracted from the uni-directional super-peaks category for visualisation. . . . . . . . . . . . . . . . . . 73
5.2 **Super-peaks aligned and orientated by consensus direction of Okazaki synthesis.** The most extreme nucleotide positions of Okazaki junctions across the genome, termed “super-peaks”, are aligned, and orientated so that the consensus direction of synthesis is from left to right. Pol α tract retention would be expected immediately downstream of aligned super-peaks (see schematic, top left). DNase I hypersensitivity (grey shading, top left) corresponds to Okazaki junction rate (blue points, top left). Polymorphism rate (black line, bottom left) is plotted with its 95% confidence interval of permuted values (grey). In the expanded view (right), DNase I hypersensitivity, nucleosomes and H2A.Z containing nucleosomes are plotted individually. Forward (red) and reverse (blue) strand transcripts combined.

5.3 **De novo motifs at super-peak junctions.** Sequences surrounding super-peaks were input into the de novo motif discovery software DREME [3]. Five motifs were identified as significantly enriched at super-peaks. [a] each motif, in descending significance, named dreme 1 to 5; their p values (green), frequencies at super-peaks (purple), and bit-weighted motifs. [b] localisation of dreme motifs at super-peaks. Regions are orientated so that the consensus direction of Okazaki synthesis is from left to right. Dreme matches represent frequency of mid-point coordinates, for each dreme motif.

5.4 **Top novel-binding-motif : known-binding-motif matches.** Positional weight matrix (PWM) logos for the two strongest known-motif matches, to each of the five novel dreme motifs, as assigned by Jaspar [75]. The alignment score percentages, as provided by Jaspar, are shown with each factor.

5.5 **Motif-divided subcategories of super-peaks aligned and orientated by consensus direction of Okazaki synthesis.** [a-e] subcategories of super-peaks containing de-novo motifs dreme1 to 5, respectively. [f] super-peaks containing non of the five dreme motifs. In all plots super-peaks belonging to each subcategory are aligned, and orientated so that their consensus direction of Okazaki synthesis is from left to right. Plots contain DNase I hypersensitivity (grey shading, top left), Okazaki junction rate (blue points, top left), polymorphism rate (black line, bottom left), the 95% confidence interval of permuted polymorphism rate (grey lines, bottom left), an expanded view of DNase I hypersensitivity (top right), bound nucleosome dyad positions (second from bottom, right), bound H2A.Z containing nucleosome dyad positions (second from bottom, right), and forward (red) and reverse (blue) strand transcripts (bottom right).
6.1 **DNase I hypersensitivity footprints.** A close-up of DNase I Hypersensitivity Sites (DHSs; full sites several kb wide), are shown, with DNase I cleavage frequency per nucleotide (grey and black). Gaps of low cleavage frequency, the DHS footprints, are visible within. These footprints are expanded further below (black). Footprints (red bars) are on average less than fifty bp. (Figure adapted from Hesselberth et al. [46]). 88

6.2 **Okazaki junctions and substitution rate at aligned yeast DHS footprints.** Okazaki junction rate (blue) and substitution rate (black) are aligned to DHS footprint left-hand edges (grey). Okazaki junctions on the forward strand (dark blue line) and Okazaki junctions on the reverse strand (dark blue points) are separated. Bound proteins are predicted to lie immediately downstream of aligned footprint left-hand edges (top schematic). Pol δ dissociation is predicted to occur when first encountering a bound protein. Pol α tract retention is predicted immediately downstream of Pol δ dissociation. Substitution rate peaks at footprint edges and reduces within the footprint itself, where purifying selection is predicted. 89

6.3 **Substitution rate at aligned human DHS footprints.** Between species GERP scores (red) are aligned at DHS footprint left-hand edges ($x = 0$). 95% confidence intervals of permuted GERP values are included (pink). DHS coverage (grey) and footprint coverage (blue) are plotted. Bound proteins are predicted to lie immediately downstream of aligned footprint left-hand edges (top schematic). 91

7.1 **Evolutionary consequences of cell type-specific selection and mutation.** A schematic to illustrate how elevated mutation rate and selection will effect DNA differently over evolutionary time when the selectional forces are the same but the cell type in which they act differ. Each of the two circles on the left of the diagram represent the same stretch of DNA, in two different cell types. The top circle represents this stretch of DNA in a heart cell. In this fictitious heart cell a factor binds on the left. Consequently the left-hand binding site undergoes purifying selection (green), based on its cell specific function, but also elevated mutation rate (red). The bottom circle represents this same stretch of DNA in a gamete. In this fictitious gamete a factor binds on the right. Consequently the right-hand binding site undergoes purifying selection (green), based on germ cell specific functions, and also elevated mutation rate (red). The bottom left plot represents how these signals will combine within a population over evolutionary time (dark brown). Selection will act at both binding sites and consequently both sites of constraint will be observed within the population. Elevated mutation rate will only be passed on to future generations if it occurs within the germline, therefore a peak in substitution rate will be seen at the right-hand binding site but not the left. 96

7.2 (one of one..) 101
7.2 Substitution rate at somatic-like and germline-like subsets of human footprints. DNase I Hypersensitivity Coverage (grey), footprint coverage (blue), non-CpG-associated human polymorphism rate (black) and between species GERP constraint scores (red) are plotted across footprint sites divided into four subsets, according to the cell types within which they are found: [a] footprints found in at least twenty of the thirty five somatic cell lines, and not in any other cell type category (‘strong somatic’; the most confidently called somatic-like footprints), [b] footprints found in at least ten of the thirty five somatic cell lines, and not in any other cell type category (‘weak somatic’; the less confidently called somatic-like footprints), [c] footprints only found in the pluripotent cell line, and not in other cell type categories (‘weak germline’; the less confidently called germline-like footprints) and [d] only those footprints found in all of the five cell type categories, implicated as “housekeeping” footprints (‘strong germline’; the most confidently called germline-like footprints). 95% confidence intervals for non-CpG-associated polymorphism rate (grey) and between species GERP constraint scores (pink) are calculated from permuted values across all footprints combined. Y axes for substitution rates are equal throughout a to d.

7.3 Footprint subset substitution rates within an extended genomic window. DNase I Hypersensitivity Coverage (grey), footprint coverage (blue), non-CpG-associated human polymorphism rate (black) and between species GERP constraint scores (red) are plotted across footprint sites divided into four subsets, according to the cell types within which they are found: [a] footprints found in at least twenty of the thirty five somatic cell lines, and not in any other cell type category (strong somatic), [b] footprints found in at least ten of the thirty five somatic cell lines, and not in any other cell type category (weak somatic), [c] footprints only found in the pluripotent cell line and not in other cell type categories (weak germline) and [d] only those footprints found in all five cell type categories (strong germline). 95% confidence intervals for non-CpG-associated polymorphism rate (grey) and between species GERP constraint scores (pink) are calculated from permuted values across all footprints combined. Y axes for substitution rates are equal throughout a to d.
7.4 The breakdown of polymorphism rate and CpG dinucleotide frequencies between somatic-like and germline-like subsets of human footprints. Close-up [a:d] and extended [e:h] genomic window views. Overall polymorphism rate (black), non-CpG-associated polymorphisms (blue), CpG-associated polymorphisms (magenta), CpG dinucleotide frequency (purple), methylated CpG dinucleotide rate (green) and footprint coverage (orange) are plotted across footprint sites divided into each of the four subsets: strong somatic ([a], [e]), weak somatic ([b], [f]), weak germline ([c], [g]) and strong germline ([d], [h]). With the exception of methylated CpG rate, where Y axes differ due to considerations for regional bias, Y axes are equal throughout a to h.

8.1 Evidence for the lagging strand hypothesis: elevated substitution rate at sites of predicted Pol α tract retention. A mechanistic model.

8.2 Classification of distinct CTCF binding motifs: Genome-wide CTCF-bound locations as identified by Rhee and Pugh. The left-hand panel shows a colour chart representation of each CTCF binding site sequence, centred by motif midpoint. Sites (rows) are divided into six categories based upon the presence or absence of sequence motifs within them. Four sub-motifs were identified, marked above the left-hand panel, (green, blue, yellow and pink), and shown above their underlying sequences PWMs, (bottom right). The top right-hand table is coloured to demarcate the combinatorial use of each CTCF sub-motif. (Figure adapted from Rhee and Pugh [102]).

8.3 High-resolution analysis of CTCF binding sites. Boyle et al. plot DNase I hypersensitivity at CTCF binding sites. Above is the cumulative footprinting signal at all CTCF motif predicted sites. Centre is just those sites with no upstream DNase I digestion spike. Bottom is just those CTCF sites including a large increase in DNase I digestion upstream of the CTCF motif. The light gray bar indicates the location of the known CTCF motif. The dark gray bar represents a novel binding motif, both of which are shown as PWM logos below. The novel binding motif is only detected in CTCF footprints that contain the small upstream region with a spike in DNase I hypersensitivity. (Figure adapted from Boyle et al. [8]).
8.4 Substitution rate at CTCF binding sites. [a] substitution rate patterns at CTCF binding footprints as reported by Boyle et al., represented by phastCons scores [8]. The authors report an apparent lack of conservation "shouldering" at CTCF sites. (Figure adapted from Boyle et al. [8]). [b] my analysis of substitution rate at CTCF binding sites. All sites are centred on their motif mid-point and orientated by motif strand, so that for all sites motifs are overlapping and in the same orientation (illustrated by the motif sequence strength by PWM, top). Sites are split into four groups of equal numbers according to their binding strength (ChIP-Seq coverage). Substitution rate (GERP) for each group is plotted in red, from light to dark (weakest to strongest binding strength).

8.5 Sliding window peak comparison. Peak A (green) and peak B (orange) are hypothetical sequence tag binding profiles of two CTCF-bound locations. Positive and negative values denote sequence tags aligned to either the forward or reverse strand, respectively. All possible overlapping combinations of the two peaks are considered, for both possible strand orientations. A score (right of each plot), is derived by summing the total area of both peaks that is not overlapping (overlapping areas are shown in yellow). As the sum of each peak is normalised to 1, the maximum score (no overlap) is 2, while the minimum score (complete overlap) is 0. The optimum overlap (0.54 in this example; red), represents the distance (similarity) between each peak pair.
The human body is made from trillions of cells [6], almost all of which contain copies of the same genetic DNA code; the entirety of which is known as the genome. The genome is what we inherit from our parents, and our offspring inherit from us. It contains blueprints for all the proteins the body will need to make in order to function, as well as the instructions controlling how each one of these is made. Each blueprint is under strict regulation, to maintain tight control of which proteins are made in each cell and when, and consequently how each cell acts and functions. In this way different cells of the body appear and act differently, but their underlying genetic code remains the same.

The core drive behind the study of genetics is to explain how sequences and interactions in our DNA manifest as characteristics and disease. The hope is to find the sequences and interactions that contribute to the morphology we all share or to pin-point variations that make us differ. Variation, within a population or between species, is a common occurrence; although only a small percentage of our genome differs between any two individuals, approximately 0.1% [80], this still amounts to approximately 6 million single nucleotide sites. The large majority of these variants will be of no functional consequence to the organism [60]. The complex challenge is in identifying those that are.

One approach for the identification of functional variants is to assess the biological activity of the DNA within which they are found. A proportion of our genome is believed to be non-functional; comprising ancient genomic relics or past events of duplication [9]. If a stretch of DNA is of no functional importance then mutations will be permitted to accumulate, as their presence will
have no affect on the viability of the organism [130]. A truly non-functional region of DNA is considered the genomic background level for variation. Conversely, functionally important DNA will be depleted of variation, as mutations in such regions will likely disadvantage the organism, reduce viability and suppress genomic propagation. In rare instances mutations at functionally important sites improve viability, and in these instances variation may rise above that of the genomic background level. Genomic fluctuation in variation across the genome as a consequence of biological fitness is governed by the process of selection.

Variation across the genome can be quantified as the rate of substitution. Substitution is far from uniform but peaks and troughs; both at single nucleotide resolution and regionally. When interpreting patterns of substitution it is often incorrectly assumed that mutation rate is uniformly random, in the sense that mutations are just as likely to occur at any point in the genome [60]. Consequently, under this assumption, patterns in substitution would purely be a result of selection: the selection against mutations deleterious to the organism or selection for mutations advantageous to the organism. Under this assumption the focus for functional variants would be in the regions of the genome where substitution is low: selection will act to preferentially remove new mutations in these regions from the population, therefore those that persist are likely to be of functional consequence.

It is now known that certain regions of the genome have undergone high substitution rate as a consequence of elevated mutation, as opposed to lack of selection. CpG dinucleotides, for example, due to their susceptibility to deamination, are well known sites for increase mutation rate at the single nucleotide level [74]. Both the DNA bound to nucleosomes [99], and nucleosome free regions of promoters [18], displays elevated substitution rate spanning comparatively large genomic distances, which is speculated also to be a consequence of elevated mutation. This contradiction to the assumption that mutation rate is random and uniform adds an added complexity to the understanding of functionally important genomic sites. If this mutational effect is taken into account, it is no longer correct to state that high substitution rate represents DNA of low functional importance. It must also be considered that high substitution rate has the potential to represent regions of the genome that are both functionally important and under elevated mutation rate. This is a consideration still largely ignored when mining the genome for variants of functional importance, but a consideration I wish to focus upon and contribute toward with this thesis.

To aid the understanding of how fluctuating mutation rate shapes our genome I investigate unusual evolution patterns surrounding the binding sites of DNA transcription factors. An elevated rate of DNA change has been identified immediately adjacent to these sites [8]. These poorly explored patterns are unusual and difficult to explain as they are in otherwise well
conserved functional regions of the genome. We speculate that these sites may hold a clue to mutation rate, and by proposing and testing a mechanistic hypothesis for how they form hope to significantly contribute towards our current understanding of the evolutionary forces that shape our genome.

1.1. THE STRUCTURE OF THE EUKARYOTIC GENOME

1.1.1 The genome and its building blocks

The genetic information of a cell is encoded by Deoxyribose Nucleic Acid (DNA). At its smallest resolution DNA is a long thin molecule of two strands coiled around a central axis to form a helix. Each strand is made from a string of nucleotide proteins, each composed of a deoxyribose sugar, a phosphate group, and a nitrogenous base (Figure 1.1a) [127]. Approximately ten nucleotides per strand make up one turn of the DNA helix.

The phosphate group of each nucleotide connects to the sugar of the subsequent nucleotide to create a chain of nucleotides connected by a phosphate backbone. Each sugar is made of five carbon molecules, numbered one to five. The fifth carbon of the sugar connects to the nucleotide’s phosphate group. This is referred to as the five prime (5') side of the nucleotide. The third carbon is connected to a hydroxyl group which interacts with the phosphate of the next nucleotide in the chain. This is referred to as the three prime (3') side of the nucleotide. The two strands of the DNA double helix are orientated in opposite directions, in an anti-parallel fashion, with one strand orientated 5' to 3', and the other 3' to 5' [127]. The base of each nucleotide, connected to carbon two of the sugar, extends into the centre of the helix where it is paired with a base from the opposite strand. Four types of nitrogenous bases are commonly found within DNA: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). It is these four bases that provide the “alphabet” of the genetic code. Pairing of bases from opposite strands is not random, but instead an A base will always be found opposite and paired with a T, and a C base will always be found opposite and paired with a G, and vice versa [127]. In this way each strand is essentially a template of the other. Sequences of paired strands from corresponding sections of DNA are referred to as each other’s reverse compliment.

Further packaging takes place to convert naked DNA into its chromosomal structure. First DNA is wrapped around nucleosomes to create chromatin. Nucleosomes are complexes of eight histone proteins, whose purpose is to package and order the DNA, as well as to regulate it. Approximately 147 base pairs of DNA wrap twice around each histone complex to form the nucleosome, with
Figure 1.1: **The structure of DNA.** [a] In its simplest form DNA is made from two nucleotide strands tightly wound in a double stranded helix. Each nucleotide is made of a pentose carbon sugar (purple), an adenine (A; green), thymine (T; red), cytosine (C; blue) or guanine (G; yellow) nitrogenous base, and a phosphate group (orange), which connects each sugar to form the strand's phosphate backbone. Each nucleotide and subsequent strand is polarised from 3’ to 5’, with each strand antiparallel to the other. Nucleotides are matched in complementary base pairs across strands: A to T and C to G. [b] DNA is wrapped twice around clusters of eight histone proteins to form nucleosomes. Nucleosome strings of DNA undergo further coiling to produce a chromatin fibre, which itself undergoes extensive packaging to form the chromosome.

Linker regions of unwrapped DNA spanning an average of 80 nucleotides between [52]. Still further coiling, super-coiling, looping and long-range interactions act upon the DNA to create its complex 3D structure [100] (Figure 1.1 b).

The genome is the term used to describe an organism’s complete DNA sequence. The human
1.1. THE STRUCTURE OF THE EUKARYOTIC GENOME

The eukaryotic genome is made up of twenty four chromosomes, ranging in length from ~ 48 million base pairs (chromosome 21), to 249 million base pairs (chromosome 1) [http://genome.ucsc.edu/]. Comparatively, the yeast *Saccharomyces cerevisiae* (*S.cerevisiae*) genome is made up of sixteen chromosomes, ranging in length from ~ 270 thousand base pairs (chromosome 6), to 12 million base pairs (chromosome 1) [http://www.yeastgenome.org/].

1.1.2 The basics of cell structure, cell division and DNA replication

Every living organism is made from either a single cell or multiple cells. Each cell contains DNA, encoding the proteins the organism requires to function and survive, as well as the protein machinery required for their production, cell maintenance, and specific cellular behavior. All living organisms can be described as either prokaryotes or eukaryotes, depending on the structure of the cells from which they are composed. Eukaryotes, such as animals, birds, plants and yeasts, are made from cells whose DNA is contained within a nucleus; a sub-compartment of the cell enclosed by a membrane barrier. Their DNA is normally organised as multiple linear chromosomes. Prokaryotes, such as bacteria, lack a nuclear sub-compartment, and their often singular circular DNA is instead within the body of the cell [111]. It is widely believed that eukaryotes are a descendant of a prokarotic organism, who over approximately 850 million years have developed distinct architecture, components and cellular functions [15].

Both prokaryotic and eukaryotic cells propagate by division: the cell will first replicate its DNA, to produce two exact copies, and then split in two, with one copy of the genome in each. In prokaryotes cell division is known as binary fission, in eukaryotes mitosis. Although the purpose of binary fission and mitosis are the same, the methods and mechanisms by which they are achieved differ significantly. Subsequent discussions within this thesis focuses only upon those of the eukaryotic cell.

Mitosis is divided into distinct phases: in interphase chromosomes replicate to form two identical copies, in prophase they condense and compact, in metaphase the two copies align at the centre of the cell, in anaphase they are separated and pulled to opposite ends of the cell, and in telophase the cell divides into two, with a single copy of each chromosome in each (Figure 1.2 a), reviewed in [122]. During interphase, DNA replication is initiated in concert from multiple points throughout the chromosome, termed origins of replication (Figure 1.2 b). From each origin of replication the DNA molecule is prised apart exposing short stretches of single stranded DNA. Replication by DNA polymerase moves bilaterally from each origin of replication until reaching the product of its neighbours, resulting in two double stranded chromosomal copies, ready to begin prophase. It has long been observed that DNA polymerases are unidirectional, traveling only in the 5' to
Figure 1.2: **Basic mitosis and genome replication.** [a] during cell division the cell’s DNA must be replicated and split into separate nuclear compartments (pink) by mitosis. Each chromosome of DNA (purple and green) is first replicated (interphase), tightly packaged and condensed (prophase), and aligned along the cell’s centre by spindle fibres (metaphase). The spindle fibres separate the two copies of each chromosome by pulling in opposite directions towards the poles of the cell’s nucleus (anaphase). The cell’s singular nucleus then divides into two (telophase), followed by the remaining cell. [b] during DNA replication the cell’s replication machinery prises open multiple sections of single stranded DNA at origins of replication (orange). Replication continues bidirectionally along the length of each chromosome, replicating each strand of double-stranded DNA simultaneously (green and blue).

3’ orientation [88]. As the DNA’s two strands are antiparallel they must therefore be treated differently: the leading strand extends from an origin of replication in a 5’ to 3’ manner and therefore is replicated continuously, while the lagging strand extends from the origin of replication in a 3’ to 5’ manner, and must therefore be replicated discontinuously as short stretches of DNA (termed Okazaki fragments) [88], later ligated together.

### 1.1.3 Protein coding DNA

Although the definition varies depending on context, for the purpose of genome annotation genes can be considered stretches of genome that encode a protein. The human genome contains around
1.1. THE STRUCTURE OF THE EUKARYOTIC GENOME

Figure 1.3: **Protein synthesis.** DNA is converted to pre-mRNA by RNA polymerase in the cell nucleus. The pre-mRNA molecule is complementary to its DNA template, in the same way two DNA strands are complementary to each other in the double stranded DNA helix. The exception of the RNA code is uracil (U), which pairs with A nucleotides (in place pf T). The pre-mRNA molecule is processed, to remove untranslated regions, to produce mature mRNA. mRNA is exported from the nucleus to the cytoplasm of the cell where it is recruited by ribosomes. Ribosomes read the mRNA in three nucleotide codons, simultaneously pairing them with their corresponding amino acid to produce the protein encoded by the DNA template.

19,000 such genes [27], amounting to approximately 1.8% of the genome’s total length [80]. The *S.cerevisiae* genome contains an estimate 6,091 genes [69], but their percentage of the total genome is predicted to be much higher.

Genes are translated to proteins in a roughly three step process (Figure 1.3): first the length of the gene’s DNA is copied by RNA polymerase to create an intermediate messenger molecule made from Ribo Nucleic Acid (RNA), known as pre-mRNA, by the process of transcription. The pre-mRNA is spliced to remove any portions within the gene not to be converted to protein (introns and untranslated regions), by RNA processing, producing mRNA (reviewed in [5]). mRNA is exported from the nucleus where it is bound by ribosomal complexes, which translate the genetic code from RNA nucleotides to protein amino acids. The ribosome reads the mRNA three nucleotides at a time. Different three nucleotide combinations (codons), correspond to different protein amino acids. For example the codon ‘ACG’ will be matched by the ribosome to a Threonine amino acid, while simultaneously adjoining it to the chain of amino acids preceeding it. The complete amino acid chain is then folded, according to the interactions of the amino acids it
contains, to produce the final encoded protein.

Other ‘genes’ are transcribed to RNA by RNA polymerases but instead of further translation to protein the RNA molecules themselves play a role in the cell (for reviews see [28, 43, 83]). Transfer RNAs for example (tRNAs; see review [132]), are recruited by ribosomes as recognition molecules to bridge the gap between mRNA codons and protein amino acids.

1.1.4 Non-protein coding DNA

The ∼ 98.2% of the human genome that is not protein coding is not purely redundant. Throughout eukaryotes there is great speculation over how much non-protein coding DNA has regulatory and structural roles. The authors of the Encyclopedia of DNA Elements (ENCODE) claimed that biochemical function could be assigned to 80% of the human genome [26]. However there is much criticism over the criteria used by the ENCODE project in their definition of “biochemically functional” [25]. Other studies have suggested that ∼ 5% of the human genome is functionally conserved with other species [98], with the potential of an additional 4% being functionally conserved in the human lineage alone [124], placing the figure of functional DNA more at ∼ 9%. In support of this figure statistical methods predict that, according to the rate of harmful mutations that accumulate per generation, a minimum of 90% of the genome is functionally inactive [58, 104].

Of this non-functional DNA it is suggested that between half and two thirds is made up of transposable elements [20, 38]. Transposable elements are stretches of DNA able to change position within the genome, either by transcription from DNA to RNA, reverse transcription back to DNA and reinsertion elsewhere into the genome, or by cutting and pasting sections of DNA directly [76]. Millions of copies of these elements exist within the genome, the majority of which are now inactive. Other examples of non-functional DNA are highly repetitive DNA elements, which expand and contract even between individuals by unequal crossing over in mitosis or replication slippage, as well as pseudogenes, which are redundant copies of an original functional gene. It has been estimated that between 12,600 and 19,700 pseudogenes exist within the human genome [92]. Introns, between the coding sections of genes, are believed to be largely non-functional, although their terminal sequences are known to be important in determining precisely where mRNA must be cut [36].

Many classes of functional non-coding DNA have also been characterised: Telomeres, which flag the ends of each chromosome, are highly repetitive DNA sequences known for their role in maintaining chromosomal integrity [45]. Centromeres, found at the mid-point of chromosomes,
are paramount for the correct alignment and separation of sister chromatids along the cell’s central axes during mitosis [118].

A considerable proportion of functional non-coding DNA is assigned to sites involved in the regulation of gene expression. Key players in this regulation are transcription factors: a varied family of function-specific proteins that bind in a precise sequence-specific manner to well defined transcription factor binding sites in DNA to modulate its activity [102]. These binding sites are short nucleotide sequences, typically between five and twelve nucleotides in length [75], which act as a recognition motif for their corresponding transcription factor to bind.

Transcription factors interact with four general classifications of gene regulators: promoters, enhancers, silencers and insulators. Promoters are regions of DNA upstream of genes, normally within ~ 500 base pairs of the gene’s transcription start site [18]. They contain multiple transcription factor binding sites, which mediate accurate initiation of transcription by evicting nucleosomes and modifying histone DNA markers [12]. Enhancers positively regulate promoters, irrespective of chromosomal orientation or distance [56]. They are often thousands of base pairs from their target promoters, or on separate chromosomes, and rely on looping of DNA to orchestrate close proximity for gene activation [128]. Active enhancers are typically devoid of nucleosomes, facilitating transcription factor accessibility. Enhancers often contribute additively and partly redundantly to the overall expression pattern of their target genes (reviewed [7]). Silencers have the opposite function to enhancers, using transcription factor binding to intentionally promote dense nucleosome binding called heterochromatin. Such heterochromatin formation prevents transcription factor-mediated promoter initiation and subsequent gene transcription [96]. The fourth category of regulators, insulators, are DNA elements that protect promoters located in one region of accessible chromatin from enhancers targeting a second region of accessible chromatin (enhancer-blockers), or block the spread of heterochromatin by silencers (barriers) (reviewed [34]).

Due to the sequence specific nature of many transcription factor binding sites it can be inferred that mutations at these sites will have significant consequences on the expression levels of the genes they regulate. In accordance, it has been calculated that transcription factor binding sites contributed an estimate 9,017 adaptive substitutions per hundred generations in the human lineage, roughly equal to the estimate for coding sequence [1].
CHAPTER 1. INTRODUCTION

1.2 Selection

1.2.1 Sites evolving approximately neutrally

Any change in an organism’s DNA sequence will be as a result of mutation; whereby nucleotide bases are either erroneously changed, or whole nucleotides inserted or deleted, or whole chromosomal regions rearranged. The degree to which such mutations affect the viability of their host organism, or the host organism’s offspring, will depend on whether or not these mutations occur in functional DNA, and consequently whether the cell’s essential functions are affected. If mutations occur in functional DNA there will be an increased likelihood that they will affect the organism’s viability, and consequently affect whether such mutations are passed on to future generations and spread throughout the subsequent population.

Selection is the process by which detrimental mutations are more likely to be filtered out of a population (purifying selection), and advantageous mutations more likely to be retained (positive selection), as a consequence of organism fitness. As the majority of mutations will occur in non-functional DNA the majority of sites will be subjected to neither purifying nor positive selection [60]. Such sites are said to be evolving neutrally. Neutrally evolving sites are subject merely to genetic drift, which governs that a mutation’s frequency within a population will fluctuate over time, purely as a result of random stochastic sampling and chance [60]. By genetic drift, the frequency of a mutation within a population may fluctuate for thousands of years before becoming either removed from the population pool completely or becoming fixed in each genome of the population permanently (until being subjected to further mutation).

Identifying sites of neutrally evolving DNA within the genome is important for functional analysis; the rate of evolution at these sites will represent a regional and genomic background rate of genetic drift, devoid of selective pressures, providing a base-rate value to which other sites can be compared. A challenge is the identification of sites truly evolving neutrally. Using the rate of evolution at non-functional sites provides a simplistic model, but due to the previously discussed lack of clarity regarding non-coding DNA (section 1.1.4), excluding non-coding regions with functional activity is problematic. Pseudogenes and ancient repeats, thought to be no longer functional, are sites where both single base changes and insertions and deletions are free to accumulate under neutrality [94]. However, active genes composed from such sites are known, such as the human gene syncytin, expressed in the placenta and derived from an extinct retrovirus [82]. Misincorporation of such regions would skew neutral estimates. Such sites excluded, using a genome-wide value from non-functional DNA as a proxy for neutrality still does not account for regional variability of mutation rate.
Often used as a proxy for neutrality are four-fold degenerative (4D) sites, also known as synonymous sites. These are positions within protein-coding sequence where the nucleotide can be exchanged for any other base, without affecting the amino acid encoded by the codon it is a part of. Thirty two of a possible sixty four codons contain a 4D site at their third position. As single nucleotide changes at these sites will not affect the protein they encode, they are in theory free to evolve neutrally under genetic drift [130]. In the majority of cases this theory will hold, but it cannot be ruled out that changes to the DNA and subsequent mRNA sequence they encode will not have structural effects to either mRNA packaging within the ribosome or incorrect processing during intron splicing, for example. The benefit of using 4D sites as a proxy for neutrality is that regional mutation bias is accounted for.

1.2.2 Purifying selection

Combined research between Kimura, Ohta, King and Jukes demonstrated with the “nearly neutral theory of molecular evolution” that mutations slightly beneficial or deleterious to an organism behave in the majority of instances like those at neutral sites [60, 61, 87]. They determined that in order for a selectional effect to be observed, both a) the degree to which a mutation affects organism fitness and b) the effective population size of that organism, must be significant. When fitness effect is low or the effective population size too small, the frequency within a population of a mutation at a functional site will again be governed purely by genetic drift. In this manner new slightly deleterious mutations may by chance become fixed within a population, or slightly advantageous mutations lost. When the population size is large, and/or mutational effect on fitness significant, selection will have the power to contribute to substitution rate frequency, alongside genetic drift. In this manner selection acts to bias the stochastic sampling of genetic drift.

Purifying selection is the removal of mutation causing unfavourable traits from a population, due to the reduced viability and reproductive fitness of those organisms they are found within. Our genomes have been shaped by purifying selection for many millions of years, meaning that our functionally active DNA is already extremely optimal. Any mutations in these regions are therefore most likely to be disruptive as opposed to advantageous, and consequently acted upon by purifying selection.

The most commonly used sites for quantifying strong signals of positive selection are non-synonymous sites. While synonymous sites can be exchanged for any nucleotide and their codon still encode the same protein amino acid, a nucleotide alteration at a non-synonymous site has the potential to change the codon’s encoded amino acid. New mutations at these sites are likely
CHAPTER 1. INTRODUCTION

Figure 1.4: **Functional consequences of substitutions.** The functional consequence of an ‘A’ to ‘G’ substitution will vary within the genome, depending on the functionality of the DNA it is found within. Substitutions within non-functional DNA, such as introns and intergenic regions, are unlikely to be of functional importance, along with substitutions at synonymous sites within functionally active protein-coding DNA. Substitutions at non-synonymous sites within protein-coding DNA are likely to have a strong functional consequence, as well as substitutions within transcription factor binding sites at promoters and enhancers. The effect of purifying selection will reflect that of functional importance. Sites within regions of low functional importance, under a low influence of purifying selection, reflect genomic neutrality.

\[ \text{Functional effect} \]
\[ \text{Purifying selection} \]

1.2.3 Positive selection

In rare instances DNA mutations within functionally important regions of the genome will confer a selective advantage to the organism. If such mutations have a significant enough effect on organism fitness, and the effective population size is large, positive selection will act to increase their frequency within the population, due to the increased viability of such organisms in comparison to their fellow population.

Examples of positively selected mutations within the human genome are mutations that influence skin hypo-pigmentation with migration to colder climates [79], and adaptation to low oxygen...
1.2. SELECTION

concentration at high altitude [131].

Positively selected mutations within regulatory DNA are believed to have the greatest contribution towards divergence between species [62]. They are therefore of great interest in evolutionary studies, as a means of explaining how we ourselves have diverged from the primate lineage. Pollard et al. used comparative genomics to identify what they believed to be sites that have undergone the most extensive positive selection within our genomes over the last ~ five million years, since our split from our most common ancestor the chimpanzee [97]. Although many of their findings have since been accounted for by other mechanisms (see section 1.4.1), their statistically strongest Human Accelerated Region (HAR), HAR1, has been assigned as a novel RNA gene expressed during neocortical development, and could consequently contribute towards the differences in brain development between human and chimpanzee populations [4].

Zhang and Li found no evidence of positive selection when looking at DNA changes within 182 housekeeping and 148 tissue-specific human genes [134]. This supports the theory that while sites undergoing positive selection have a significant effect on the organism, as demonstrated by the significance of their proceeding traits, such sites occur at an extremely rare frequency throughout protein-coding regions of the genome.

1.2.4 Measuring selection

Within protein-coding sequences, the number of non-synonymous substitutions per possible non-synonymous sites ($K_a$), is compared to the number of synonymous substitutions per possible synonymous sites ($K_S$), to calculate the net degree and direction of selection [50]. If $K_a/K_S = 1$, with the rate of non-synonymous and synonymous substitutions equal, the region is inferred as evolving neutrally. If the rate of non-synonymous substitutions is higher than the rate of synonymous substitutions, with $K_a/K_S > 1$, the region is inferred to be under positive selection, and inversely $K_a/K_S < 1$ infers purifying selection.

Comparison of such $K_a/K_S$ scores within and between species, by a McDonald-Kreitman test [77], adds a further complexity to the measurement of selection. If the ratio of nonsynonymous to synonymous variation between species is lower than the ratio of nonsynonymous to synonymous variation within species ($Dn/Ds < Pn/Ps$), negative selection is implied. Conversely, if the ratio of nonsynonymous to synonymous variation within species is lower than the ratio of nonsynonymous to synonymous variation between species ($Dn/Ds > Pn/Ps$), positive selection is implied.

A third means of deciphering the direction of selection is by Derived Allele Frequency (DAF)
Figure 1.5: The principles of Derived Allele Frequencies (DAFs). Polymorphisms within non-functionally important sites drift within a population, increasing or decreasing in frequency purely by chance. There will therefore be a large number of polymorphisms of low frequency within the population, and a small number of polymorphisms of high frequency within the population. These sites are considered to be evolving approximately neutrally (brown). Polymorphisms within functionally important sites will also drift within a population, increasing or decreasing in frequency primarily by chance, but a small number will be removed or fixed within a population due to the fitness they confer to the organism. Sites undergoing positive selection will have more polymorphisms of high frequency within the population and less polymorphisms of low frequency within the population than a site evolving approximately neutrally (turquoise). Conversely, sites undergoing purifying selection will have less polymorphisms of high frequency within the population and more polymorphisms of low frequency within the population than a site evolving approximately neutrally (yellow). The DAFs of a test sample, such as transcription factor binding sites, can be compared to sites of known neutrality to infer the mechanism of selection acting upon them.

Comparisons [57]. Allele frequencies are the rate at which different nucleotide variations are found within a population. DAFs have distinct profiles between sites undergoing purifying selection and sites undergoing positive selection: the number of alleles undergoing purifying selection which are at a high frequency within the population is low, while the number of alleles undergoing purifying selection which are at a low frequency within the population is high. Conversely, the number of alleles undergoing positive selection which are at a high frequency within the population is high, while the number of alleles undergoing positive selection which are at a low frequency within a population is low (Figure 1.5). By comparing DAFs from a test group, such as transcription factor binding sites, with a proxy neutral background sample, the degree of acting selection within the test group can be inferred.
1.3 Mutation rate

1.3.1 Genomic mutation rate

DNA mutations are the incorrect incorporation of nucleotide bases within the genomic sequence, in the form of either single nucleotide point mutations (the most common and the focus of this thesis), insertions, deletions or chromosomal rearrangements. For mutations to be observed throughout evolution they must occur in the germline; in humans either the egg of the female, or sperm of the male, or their direct cellular ancestors. Mutations occurring in other cells of the body (somatic cells), may affect an organism’s fitness, and thus still be subjected to selection, but will not be passed on to the next generation. Such somatic mutations are important in the development of cancer, for example.

Mutations can be caused by external mutagens, such as radiation, sunlight or tobacco smoke, as well as intracellular mutagens like water and oxygen. Under normal circumstances, however, the most common mutations are those caused spontaneously during DNA replication [112].

Faithful replication of the genome is crucial to organism and species survival. The human genome must replicate itself $2 \times 10^{14}$ times to produce the average adult human, as well as a further $10^{16}$ times to maintain it by cell regeneration. As each replication event must replicate $6 \times 10^9$ new nucleotides, this would amount to approximately $6 \times 10^{26}$ single nucleotide replication events in the average human life span. To maintain such a feat without error would require an extraordinary level of replicative fidelity. Although not absolute, considering it has been estimated that single nucleotide DNA mutations persist at a rate of $1.16 \times 10^{-8}$ to $2.17 \times 10^{-8}$ mutations per base pair per lifetime (reviewed in [14]), the fidelity of DNA replication is extremely high.

The most basic assumption of single nucleotide point mutation rate is that it is random, and therefore equally uniform across the genome [60]. It is now known that this is far from true. I detail examples of known variations in mutation rate below.

1.3.2 The effect of CpG dinucleotides and G + C content on mutation rate

Within DNA the combined presence of C and G nucleotides has several unique properties that subject them to a higher rate of mutation than other dinucleotide base combinations. Firstly, cytosine residues are prone to spontaneous deamination throughout the genome (the removal of their amine group by the addition of $H_2O$). In mammals C residues are unmethylated, except in
Figure 1.6: CpG deamination. **[a]** deamination of an unmethylated C nucleotide in a CpG dinucleotide forms Uracil (U). U nucleotides, which are foreign to DNA, are recognised by uracil glycosylase, flagging the erroneous U nucleotides for removal and repair to C by base excision repair. **[b]** deamination of a methylated C nucleotide in a CpG dinucleotide forms T. As T is a native DNA component it will not be flagged as foreign. During replication the template T will be paired with A, resulting in a CG to TA substitution.

the context of CpG dinucleotides, where the C can either be methylated or unmethylated. If a C residue has no methyl group attached to its base its deamination will form a Uracil (U) residue. U residues are foreign to DNA and consequently efficiently recognised by uracil glycosylase and removed, preventing mutation. If, however, the C residue of a CpG dinucleotide is methylated, by the addition of a methyl group to its base, its deamination will result in a T residue. The cell finds it harder to discriminate a T residue as it is a native base of DNA. Therefore, if such a T escapes recognition by mismatch repair, it will be paired with an A residue during DNA replication, resulting in a CG to AT conversion (Figure 1.6). This is known as the CpG effect.

CpG dinucleotides are consequently under-represented throughout the human genome; the exception being focal clusters of CpG dinucleotides called CpG islands. Even so, the rate of CpG dinucleotide associated mutations in humans is estimated to be between ten- and 18-fold the rate of non-CpG dinucleotide associated mutations [74].

CpG methylation is not a universal feature of all organisms but is restricted to certain species: methylated CpG dinucleotides have been identified in the genomes of birds and mammals, but not in the genomes of reptiles and yeast, for example.

A second property of C and G nucleotides is GC-biased gene conversion. During germ cell pro-
duction cross-over occurs between complementary chromosomes by the process of recombination. GC-biased gene conversion is the process by which cross-over between AT and GC nucleotide combinations will result in a greater number of germ cells carrying G and C nucleotides than A and T nucleotides, supposedly due to GC-biased repair of AC and GT mismatches in double stranded recombination [23]. GC-biased gene conversion is supported by observations that G and C nucleotide variants are elevated close to recombination hotspots [109], as is overall GC content [110]. The effect is reported to be restricted to highly recombining parts of the genome, irrespective of coding or functional potential [33].

1.3.3 The effect of replication timing and replication strand on mutation rate

Wolfe et al. compared the rate of DNA variation at 4D sites in mice and rats and observed differences in the rate and pattern of mutation over different regions of the genome [130]. Due to the similarity of mutation rate between genes physically close to each other the authors proposed the rate of mutation they observed was a consequence of replication timing, whereby the distance from an origin of replication will have a direct effect on the timing of the region’s DNA replication. The authors proposed depletion of the pool of nucleotide precursors as replication progresses as a possible explanation.

Jørgensen and Schierup used computationally predicted origins of replication in human germline cells to investigate this same theory, using germline patterns in variation as a measure of mutation [54]. They observed early replicating regions to have a lower mutation rate than late replicating regions, with a steady increase in variation over distance. The authors also investigated strand-specific mutation rate, a further signature of DNA replication, revealing mutation rate to be higher within genes on the lagging strand than the leading strand. They suggest that collisions between RNA and DNA polymerases may play a role in such a mutation pattern, speculating selection towards gene orientation and density at origins of replication as a consequence.

1.3.4 Transcription associated mutagenesis

Protein-coding DNA undergoing transcription to mRNA is found in an unwound open chromatin state, so as to facilitate the binding of transcriptional machinery. Once bound the machinery separates the two DNA strands to expose short stretches of single stranded DNA. Only one of these strands, the template strand, is then transcribed.
Transcription-Associated Mutagenesis (TAM) is the phenomenon of increased mutation rate within transcribed regions of the genome. It has been predicted to occur as a result of the transient exposure of single stranded DNA and this DNA’s consequent susceptibility to prolonged chemical damage [59]. In comparison to replication, which uniformly generates one copy of the genome per turn, transcription of DNA to RNA is non-uniform and clustered, occurring at highly variable rates, in a strand specific manner. Comparisons can therefore be made between the rate of transcription and rate of mutation, as well as transcribed and non-transcribed strands.

Transcription-associated mutagenesis has been observed in the yeast *S.cerevisiae* and human genome [91]. In yeast mutation accumulation models, sites that became mutated were those sites transcribed at a level ~ 50% higher than expected by random [91]. In the same report the authors used multiple linear regression analysis to estimated that doubling the expression level of an averagely expressed gene in the human genome would result in an increased mutation rate of 15% [91]. In both instances mutation rate is higher on the non-transcribed strand compared to the transcribed strand. This asymmetry is predicted to be as a result of the inherent proof reading ability of RNA polymerase: during transcription the polymerase will stall upon encountering DNA damage or nucleotide misincorporation and recruit Transcription Coupled Repair (TCR) to correct it (reviewed in [40]). Such a repair mechanism is absent from the non-transcribed strand.

These results infer transcription to be a mutagen within the genome, with more highly expressed genes consequently prone to a higher rate of mutation. As such mutations are most likely to be deleterious, selective pressures must have acted to prevent transcription at a high level, so as to keep mutation rate minimal [91].

### 1.3.5 Promoter mutation rate

Promoters are non-coding genomic regions of known functional importance, containing multiple transcription factor binding sites, conferring control to gene’s transcriptional activities [18]. It is therefore contrary to expectation that Taylor *et al.* found mutation rate at gene promoter regions to be elevated in primates [115]. The authors found 46% of primate promoters mutated at a higher rate than the expected background rate, as calculated by ancient repeat sequences. Mutations were more frequently non-CpG-associated than CpG-associated, excluding CpG dinucleotides as the cause. Promoters and comparison sites were within close proximity, thereby excluding chromosomal fluctuations in mutation rate as a consequence of replication timing.

The authors suggest such an elevated mutation rate could be linked to the unusual chromatin structure of nucleosomes at gene promoter regions. It has been observed that chromatin is
unusually open (unraveled) upstream of transcription start sites, exposing single stranded stretches of DNA for transcription factor binding and gene regulation [103]. Such an open chromatin structure has been proposed to associate with elevated mutation rate within the human genome [35]. The reason for such an elevated mutation rate remains unknown, but as with transcription associated mutagenesis, transient exposure of single stranded DNA without the presence of transcription coupled repair may subject such regions to prolonged exposure to chemical damage.

1.3.6 Male biased mutation rate

Recent studies have estimated that 85% of new single nucleotide point mutations originate in the paternal lineage (95% CI = 70.8-98.5%) [13]. This figure was calculated by tracing the parental origin of 26 single nucleotide variants uniquely mapped to an isolate population. It has been suggested the increased mutation rate is due to the larger number and continuous nature of cell divisions in spermatogenesis. Female eggs arise from a finite number of 23-33 cell divisions, whereas male sperm monotonically increases every 15-16 days as a result of mitotic maintenance of the spermatogonial pool [51], and will as a consequence be subjected to a significantly larger number of replication cycles. Consequently, due to the inevitably imperfect fidelity of DNA replication, it would be expected that mutation rate would increase with paternal age. This is supported by a second study looking at variants unique to an Icelandic population, which predicted approximately two new mutations per year of a father’s life, and assigned paternal lineage to 94% of new single nucleotide variants (90% CI = 80.1-100%) [63].

1.3.7 The nucleosome hump

Nucleosomes contain ∼ 147 base pairs of DNA and are found at regular intervals throughout the eukaryotic genome. In yeast S.cerevisiae, nucleosomes occupy 70-80% of the genome [126], irrespective of protein coding potential. Washietl and Warnecke separately reported an elevated substitution rate in correlation with nucleosome positioning throughout the S.cerevisiae genome [125, 126]; observing DNA variation over nucleosome-bound DNA to be 5 to 15% higher than that of unbound linker regions. This positional dependency was highly significant, with the null hypothesis that the pattern was produced by chance rejected with $P \sim 10^{-25}$ [126]. Substitution rate was highest at the nucleosome dyad (the binding mid-point), gradually decreasing with distance, to produce a “hump” of elevated substitution over bound nucleosome sites.

The explanation for such a hump remains illusive, with both authors arguing the opposite side
to whether (non-exclusively) selection or elevated mutation rate is responsible for the pattern observed. Warnecke et al. suggest that selection has acted upon linker regions, to maintain linker-based nucleotide-exclusion signals, for the facilitation of their correct positioning. They argue that linker regions exhibit distinct patterns of codon and amino acid usage, which reflect the linker’s needs to be rigid to prevent nucleosome formation [125]. Washietl et al., however, argue that if selection were responsible for shaping the nucleosome hump, then the correlation between dyad distance and substitution rate would be absent from synonymous sites. The authors observe positional dependencies at all three codon positions, as well as intronic sites, and consequently propose elevated mutation rate in nucleosome-bound DNA to be responsible for the observed “hump” [126].

Prendergast and Semple suggest the patterns at nucleosomes are influenced by both mutational bias and selection [99]. Furthermore, the hump appears to be the composite effect of varying base change patterns. While T→C, T→G, A→C and A→G changes form strong humps over nucleosomes, C→T, G→T, C→A and G→A changes display a corresponding dip. Chen et al. present evidence that reduced C→T substitutions within the hump are due to reduced cytosine deamination as a consequence of DNA-nucleosome binding, and subsequent protection against single stranded DNA exposure [16]. This is demonstrated by a negative correlation between the frequency by which a nucleosome is bound and C→T substitution rate.

1.4 Substitution rate

1.4.1 Interpreting substitution rate

Substitution rate is the quantification of genetic variation per nucleotide between a population or between species. It is therefore a measure of the combined effect of mutation rate and its proceeding selection, throughout the population, over time. Between-species substitution rate is calculated by nucleotide changes in relation to the phylogenetics of the species under comparison. Phylogenetic analysis produces evolutionary trees between species, assigning commonality between species and ancestors, and assigning distance lengths to each of the branches of the tree representing time since evolutionary divergence (see Figure 1.9, for example).

Substitution rate has been used to identify regions of functionally active DNA and regions of DNA predicted to have played a role in species divergence. Traditionally, under the assumption that purifying selection has removed mutations from sites of functional importance, regions of DNA with low substitution rate are predicted to be functional. Conversely, regions of DNA with
1.4. SUBSTITUTION RATE

Figure 1.7: **Interpretation of substitution rate at the nucleosome hump.** Substitution rate (brown), increases where nucleosomes are bound to the DNA and is reduced in linker regions. This could be explained by purifying selection, whereby linker regions are of higher functional importance and so mutations are removed at higher efficiency from these sites than in nucleosome-bound regions where mutations are permitted to accumulate. This model assumes a uniform mutation rate. Alternatively, such a pattern in substitution could be explained by increased mutation rate at nucleosome-bound DNA. In this model it would be the nucleosome-bound DNA that would be of greatest interest for the identification of functional variants, as opposed to the linker regions. This demonstrates how the misinterpretation of the evolutionary forces acting at these sites will lead to the misinterpretation of those variants underlying them. (Figure adapted from Semple and Taylor [105]).

Elevated or neutral substitution rate are predicted to be of low functional importance, under the assumption that a lack of purifying selection at these sites has allowed mutations to accumulate (Figure 1.7).

Under these assumptions regions of low between-species substitution rate with species-specific elevated substitution rate have been flagged as candidates for species-specific divergence by positive selection. This traditional assumption is based on a uniform mutation rate across the genome and consequently assigns only purifying, positive or neutral selective pressures to fluctuations in substitution rate. It does not take into account regional and nucleotide level fluctuations in mutation rate, which we now know to be prevalent across the genome. In such
a manner Haygood et al. identified human promoters undergoing apparent positive selection, without taking into account the possibility of species-specific elevated mutation rate at these sites [44]. Taylor et al. have since provided evidence that it is in fact the latter that is responsible for elevated substitution rate at these sites [116]. In a similar manner Pollard et al. [97] reported positive selection as the driving force behind their HAR elements, again ignoring the potential for genomic fluctuations in mutation rate. This has since been followed by reports that elevated mutation rate, by GC-biased gene conversion, is prevalent at a large number of HARs [33].

Substitution rate is an invaluable tool for deciphering the genome and for the identification of important deleterious and advantageous variants within it. However these signals will be misunderstood if substitution rate and mutation rate are not accounted for with care.

Several different measures of substitution rate have been formulated. Those of interest to this thesis are included below.

### 1.4.2 From counting between-species changes to HKY

The most simplistic method to quantify between-species substitution rate is the alignment of sequences from each species and subsequent counting of changes at each corresponding nucleotide position. In this manner a corresponding nucleotide position which is the same nucleotide base in all species studied would be identified as fully conserved, and a nucleotide position with all nucleotide bases present at equal frequencies identified as most divergent. Three major flaws render this method insufficient: 1) large quantities of nucleotide positions will not be aligned between species, so each of these positions will be ignored, 2) different nucleotide conversions are prone to different frequencies of occurrence. For example A→G substitutions are more likely to occur than A→T substitutions, and 3) no consideration is made for the ancestral relationship between species under comparison. There is therefore no means to distinguish original ancestral nucleotide bases from substituted nucleotide bases.

Simply counting the number of changes at corresponding nucleotide positions will also underestimate true divergence, since some positions may have undergone multiple changes, such as from A→G and then again from G→A. To overcome this problem Charles Cantor and Thomas Juke developed a probabilistic model in 1965 to derive nucleotide substitution rate [55] (Figure 1.8a). It is the most basic model for determining substitution rate, again assuming that the rate of substitution is the same between all nucleotides (which is well known not to be the case), but able to account for ancestral states. Jukes-Cantor considers all possible nucleotide scenarios to estimate multiple changes (a maximum likelihood framework), with each scenario weighted and
1.4. SUBSTITUTION RATE

Figure 1.8: Measures of nucleotide substitution rate probability. [a] Jukes-Cantor assigns equal probability to all nucleotide changes (α). [b] HKY accounts for an individual rate of change for each specific nucleotide. $\pi_A$, $\pi_T$, $\pi_C$ and $\pi_G$ denote the rate of change for A, T, C and G nucleotides respectively, with $\kappa$ weighting transversion substitutions.

Hasegawa, Kishino and Yano (HKY), built upon Jukes-Cantor and other preceding models to account for the observed differences in nucleotide conversion rates [42]. For example, across the genome nucleotides are more likely to convert to another nucleotide that is similar to it in molecular structure, such as A→G (purine nucleotides) and T→C (pyrimidine nucleotides). Such substitutions are termed transitions. The alternative scenario is transversion, where by nucleotides switch between purine and pyrimidine classes: A→T, A→C, G→T and G→C. These substitutions are rarer throughout the genome. As well as taking transition and transversion rates into account the HKY model takes into account the genomic rates of all nucleotide base substitutions individually (Figure 1.8b). By incorporating multiple parameters the HKY model creates a more realistic simulation of how nucleotide sequences evolve. Although the model has complexities it is still not overly complex, assuming an A→G is equally as likely as a A←G for example, thereby allowing efficient analysis of the large datasets composing the eukaryotic genome. Missing alignment data is still not accounted for in such a model, and consequently any such sites are still discarded or dealt with inconsistently between lineages, which may contribute to biasing against difficult to sequence regions of the genome.

1.4.3 phastCons

The phastCons programme is designed specifically to identify conserved elements within multiply aligned sequences, while also providing genome-wide per-nucleotide rates of substitution [106].
CHAPTER 1. INTRODUCTION

Figure 1.9: The **phastCons model.** phastCons calculates substitution rate between species using fixed tree topologies. Each site is tested against its corresponding tree but where branch lengths are shortened, representing conservation (left), or lengthened, representing lack of conservation (centre). If the frequency of variation at a site matches the branch length in the conserved model it will be scored 1, while if it matches the branch length in the nonconserved model it will be scored 0. The majority of sites will fall between the two states and be scored accordingly. This is represented by the state-transition model (right), where each state, conserved (c), or non-conserved (n), is associated with a phylogenetic model ($\Psi_c$ and $\Psi_n$, respectively). These models are identical except for their scaling. $\mu$ and $\nu$ ($0 \leq \mu, \nu \leq 1$) define all state-transition probabilities. (Figure adapted from Siepel *et al.* [106]).

It is based upon a statistical model that uses branch lengths between phylogenetic trees, for the species under investigation, to determine the evolutionary order of nucleotide changes. Scores are assigned per nucleotide from zero to one, with zero representing the lowest probability that the site is within a conserved element and one representing the highest probability that the site is within a conserved element. Zero is assumed as both nonconserved and neutral (Figure 1.9).

phastCons scores have been pre-calculated for four data sets: Human, the fruit fly *Drosophila melanogaster*, worm *Caenorhabditis elegans* and *S.cerevisiae*. The human data set is calculated with phylogenetic comparisons to the four vertebrate species mouse, rat, chicken and *Fugu rubripes*. The *S. cerevisiae* data set calculated with phylogenetic comparisons to six other yeast *Saccharomyces* strains.

phastCons does not require prior calculation of neutral substitution rate and is capable of dealing
with alignment gaps. It is not, however, able to detect substitution rate above that of the neutral estimate.

### 1.4.4 GERP

Genomic Evolutionary Rate Profiling (GERP) is another statistical measure providing a genome-wide per-nucleotide substitution rate, through the comparison of the rate of change at between-species orthologous sites [17]. As opposed to a rate of substitution, GERP is a measure of substitution deficit. GERP assigns zero as genomic neutrality, calculated from 4D sites over the genome. When calculating variation for the complete tree, a minimum of -12.3 describes sites with the lowest deficit in substitution rate (highest substitution rate, lowest sequence constraint), and a maximum of 6.17 describes sites with the highest deficit in substitution rate (lowest substitution rate, highest sequence constraint). Consequently, regions with negative GERP values represent sites with greater than neutral substitution rates, evolving faster than the genomic average, while sites with positive GERP values represent regions with lower than neutral substitution rates, evolving at a rate slower than the genomic average.

GERP is capable of intelligently dealing with alignment gaps, thereby reducing the assignment of erroneous scores. Pre-calculated genomic scores are readily available making extraction of values simple and fast. Importantly it is able to detect substitution rate above that of neutral, indicating either positive selection or elevated mutation rate, giving it an added advantage over phastCons. However, because GERP uses a prior assumption about the genomic rate of neutrality, it does not account for regional fluctuations throughout the genome.

Human GERP scores were calculated by phylogenetic comparisons to 28 other vertebrate species. These species include 12 further primates, 13 placental mammals and 3 marsupials (Figure 1.10).

It is interesting to note that in publishing the GERP algorithm its authors propose GERP as a measure of selection, and do themselves assume mutational uniformity in their analysis. However substitution rate, as measured by GERP, provides a useful model for the combined effects of mutation rate and selection.

### 1.4.5 Polymorphism rate

Polymorphism rate is a per-nucleotide measure of genetic variation within a population. Polymorphisms are essentially variations in a transient stage, yet to be fixed or removed from within a
Figure 1.10: The GERP model. Substitution rates are calculated on a per-nucleotide basis between primates (green), placental mammals (purple) and marsupials (orange; left). Each column (top right) represents a single aligned nucleotide position between species. For each column an observed rate of variation is calculated according to known nucleotide variations and species divergence. An expected rate is calculated for each column by summing the remaining branches from a neutral tree model, after removing species with a gap character. Black, blue and red values in expected ratios correspond to the correlated neutral trees (bottom right). Each nucleotide position is scored as the sum of deviations from expectation at each site, thereby representing a value of rejected substitutions (R). (Figure adapted from Cooper et al. [17]).

The major caveat with the inference of polymorphism rate is acquisition biases, whereby sequence coverage of the genome may vary, resulting in sites of poor confidence and consequently polymorphism rate misinterpretation. The study of human polymorphism rate is constantly developing and accuracy of its values improving; an increasing amount of human genomes are being sequenced and coverage of sequencing is increasing with decreased cost of protocol, providing greater confidence in polymorphism values presented. The 1000 Genomes Project set about to
sequence a thousand human genomes for the quantification of sequence variations within them [80]. This data is readily available and provides all Single Nucleotide Polymorphisms (SNPs) across the genome (all nucleotide positions with variation above a frequency of 1%). For each SNP the ancestral nucleotide, each varying nucleotide identified, and their calculated rate of prevalence within the population is provided.

The Saccharomyces Resequencing Project provides genome wide sequencing for 37 *S.cerevisiae* strains, providing data to calculate a comparable genome-wide single-nucleotide polymorphism rate across its genome [70].

1.5 Research objectives

1.5.1 Specific motivations

The study of substitution rate across the genome is one approach for identifying functional variants, for phenotypic and disease states, by assigning functional importance to the regions of DNA within which they are found. However, selection and mutation rate both shape substitution rate across the genome, with (in the majority of cases) both acting as opposing forces to the overall pattern that’s observed: elevated mutation drives up substitution rate while (purifying) selection drives it down. Such a consideration is largely ignored in variant studies due to the current immature understanding of mutation variation across the genome. It is this understanding that I hope to contribute towards with the work here in presented.

As a means of investigating variations in mutation rate across the genome our attention was drawn to the substitution pattern observed at the binding sites of the human transcription factor Nuclear Transcription Factor Y (NFYA) [8]. The substitution rate at various transcription factor binding sites have been calculated [8], revealing unusual patterns described as conservation “shoulders” (Figure 1.11). A strong signal of conservation, represented by low substitution rate, lies directly at transcription factor binding sites. This is predicted, due to the sequence specific nature of transcription factor binding motifs, and consequently the high functional importance of these sites. However, flanking either side of this low substitution is a sudden and marked increase in substitution rate, approximately ten nucleotides long. This was a novel observation, and although the authors acknowledge it as unusual and of interest, they are unable to explain its occurrence. Substitution rate then returns to the genomic background rate, as predicted.

I am interested in the pattern observed at the NFYA binding site as it has two possible mechanistic
CHAPTER 1. INTRODUCTION

Figure 1.11: High-resolution analysis of transcription factor binding sites. Substitution rate profiles at human transcription factor binding sites, represented by phastCons scores. [a] the substitution rate profile for combined human transcription factors (black), the average substitution rate across the genome (green), and the substitution rate pattern when centred purely on DNase I Hypersensitivity Sites (DHS; red). [b] the substitution rate profile for the single human transcription factor Nuclear Transcription Factor Y (NFYA). A site of low substitution rate lies directly at the transcription factor binding site, flanked by distinct peaks in substitution rate, (highlighted in yellow), ~10 nucleotides long. A further modest decrease in substitution before a gradual increase back to the genomic background level produces a “shoulder” in substitution. (Figure adapted from Boyle et al. [8]).

Explanations: the first is selection, whereby NFYA binding inhibits the binding of other factors in its immediate vicinity, rendering these regions of DNA functionally unimportant, (the effect of steric hindrance), and consequently permitting neighbouring mutations to accumulate. The second is elevated mutation rate, whereby a mechanistic or evolutionary force is occurring at the same position as NFYA binding which elevates the rate of mutation above that of the genomic background level (Figure 1.12).

The first explanation, selection, implies shoulders are of low functional importance. Consequently variations at these sites are unlikely to be of interest. The second explanation however, site specific elevated mutation, implies that although shoulders are in regions of functional importance, (demonstrated by low surrounding substitution rate), mutation rate is elevated, rendering them potential hotspots for new variants with high probability of functional consequence.

Conservation shoulders at NFYA binding sites have the potential to contradict the assumption that mutation rate is random and uniform across the genome, thereby contradicting the assumption that patterns in substitution rate are a consequence solely of selection. If so, they provide discrete sites on which to test mechanistic hypotheses explaining elevated mutation.
1.5. RESEARCH OBJECTIVES

Figure 1.12: Selective and mutational models for explaining the conservation shoulder. Conservation “shoulders” can be explained by two alternative, non-mutually exclusive models: selection (left) and mutation (right). In both models substitution rate (brown) is the combined effect of mutation rate (red) and purifying selection (green). Blue spheres represent hypothetical factors bound to DNA. Weak to strong factor-DNA interactions are shown by pink bars (light to dark). Under the selective model substitution rate is explained by uniform mutation rate and selection which reflects the strength of factor-DNA binding interaction. In this example selection is reduced immediately adjacent to central binding interactions due to steric hindrance. Under the mutational model selection is strongest at core binding interactions and gradually returns to the genomic background rate as factor-DNA interactions reduce. In this model mutation rate is not uniform, but instead increases either side of the central binding interaction (red asterisks).

rate occurrence. I therefore wish to investigate substitution rate at NFYA binding sites and the binding sites of other transcription factors further.

1.5.2 Research questions

By investigating substitution rate at transcription factor binding sites I wish to answer the following biological questions:

- How wide-spread are conservation shoulders? Are they found at the binding sites of other transcription factors? Are they found within other species?
• Are conservation shoulders a result of reduced purifying selection, or instead a new example of site specific elevated mutation rate? What is their mechanistic cause?

• Do NFYA's conservation shoulders have implications for our current understanding of functional DNA, and variants found within it? Could they help us identify new disease causing variants?

1.5.3 Approach

Throughout this thesis I use computational approaches to tackle the research questions I have proposed above, using extensive publicly available data sets to coordinate DNA replication, binding and substitution events across the human and *S.cerevisiae* genomes.

I propose to first assess the significance of conservation shoulders. A systematic investigation of substitution rate at transcription factor binding sites has yet to be published for any eukaryote, even though transcription factor binding data is readily available for many (see [102] for example). Identification of shoulders in both mammalian and yeast species would imply an evolutionarily conserved relationship, thereby adding to its significance, as well as providing an extensively larger pool of data from which to test a mechanistic hypothesis.

We hypothesise that error-prone Pol α tract retention during lagging strand replication is responsible for site specific elevated mutation rate across the eukaryotic genome, including at the binding sites of some but not all transcription factors. I will first test this hypothesis at the binding sites of the *S.cerevisiae* transcription factor Reb1, as well as at positions across the *S.cerevisiae* genome predicted to retain the highest rate of Pol α tracts.

As yet, accurate lagging strand Okazaki fragment mapping, as used to determine Pol α retention in the *S.cerevisiae* genome, is not available for the human genome. We further hypothesis that DNase I hypersensitivity footprints could act as a proxy for determining sites of Okazaki termination and consequently the potential for Pol α tract retention. I will therefore progress to investigate DNase I hypersensitivity footprints as a means of drawing comparisons between Pol α tract retention and substitution rate in the human genome.

The presence of elevated substitution rate at sites of error-prone Pol α tract retention would provide suggestive evidence that such a pattern is a consequence of site specific elevated mutation rate, as opposed to site specific reduction in purifying selection. I would like to add further support to this hypothesis, and therefore propose to identify additional mutation-specific lines of evidence. To do this I will compare the substitution profiles of human between-species conservation, which
is primarily shaped by selection, and human-specific polymorphism rate, which is primarily shaped by mutation. I also propose to compare these profiles between footprints not present in the germline, which can only be effected by selection, and footprints only present in the germline, where mutation rate will also be observed.

With this approach I will test a novel hypothesis to prove or disprove a significant caveat in the maintenance of the human genome.
2.1 Methods

Analysis methods and data sources common to multiple chapters are summarised in this chapter. Each results chapter also contains details of analysis specific to those results.

2.1.1 Reference genomes

All human analyses were performed on the hg19 (GRCh37) reference genome. All yeast analyses were performed on the sacCer3 (V64) *S. cerevisiae* reference genome. Yeast datasets originally obtained with coordinates on other assemblies were projected onto the sacCer3 assembly using liftOver (v261) [65], with the corresponding chain files obtained from http://www.yeastgenome.org.

2.1.2 Computational and statistical analysis

Analysis and all statistical calculations were performed in R (version 3.0.0). Analysis-specific software and algorithms were implemented in Perl (version 5.14.1). Lines of fit used the smooth.spline function. Sliding window averages used the rollapply function from the CRAN Zoo package with centre alignment and null padding.
2.1.3 Defining transcription factor binding sites

Binding sites were identified for transcription factors where a) ChIP-seq binding data was publicly available and b) a consensus binding motif is known. Both sets of data were available for thirty one human transcription factors, and the yeast transcription factor Reb1 [102]. ChIP-seq data was downloaded from the NCBI Sequence Read Archive (SRA) [http://www.ncbi.nlm.nih.gov/sra] (see summary Table 2.1 and Table A.1 for detailed date sources). Reads were aligned to the reference genome using Bowtie2 (version 2.0.0) [66], where a maximum of 10 best hits were reported per read. Alignment peaks were called with the Model-based Analysis of ChIP-seq (MACS) software, (version 2.0.10) [135]. MACS defined peaks were recorded as a binding site only if a binding motif for that transcription factor was also present within 50 nt of the ChIP-seq peak summit. The presence or absence of a motif was determined using the Motif Occurrence Detection Suite (MOODS), [64], (version 1.0.1; download date 04-03-2013), and motif positional weight matrices (PWMs) downloaded from JASPAR [75] (file source: http://jaspar.genereg.net/html/DOWNLOAD/JASPAR-R_CORE/pfm/nonred-undant/pfm_all.txt; file dated 19-11-2013). Motifs were required to match with a MOODS significance of 0.005. The most significant motif within 50 nt of the peak summit was recorded. “Secondary” sites were noted if another peak of higher ChIP-seq coverage was present within 400nt.

2.1.4 Yeast substitution measures

Yeast polymorphism data was obtained from the Saccharomyces Genome Resequencing project [70] (file source: ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/latest/misc.tgz; download date 07-05-2014). A polymorphic difference between any of the 37 sequenced S. cerevisiae strains was called as a polymorphic site. Sites with n > 2 alleles were counted as one polymorphic site. Only nucleotide point substitutions were considered; insertions and deletions were excluded. The polymorphism rate reported is the number of polymorphic sites divided by the number of sacCer3 sites with sequence coverage in at least two additionally sequenced strains. For each nucleotide a polymorphism count value (0-3/NA) represents a count of the number of non-reference alleles at a site, and a polymorphism call value (0-1/NA) represents the confidence by which the rate is called. Only polymorphism rates where the call value is 1 are included in analysis.
2.1.5 **Human substitution measures**

1000 Genomes data was used to define SNPs across the genome, representational of polymorphism rate [80]. Files were downloaded from the 1000 genomes database, release 20110521 (file source: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521; files dated 10-10-2012).

Genomic Evolutionary Rate Profiling (GERP) scores [17] were used as a measure of between species nucleotide diversity across 46 vertebrate species. Single nucleotide resolution bigWig files were obtained from the UCSC genome browser (hg19) (file source: http://hgdownload.cse.ucsc.edu-gdb/hg19/bbi/All_hg19_RS.bw; download date 11-04-2011). For consistency of presentation with plots of polymorphism rate the y-axes of plots showing GERP scores have been inverted so that greater constraint is low and greater diversity is high.

2.1.6 **Okazaki fragment sequence processing**

Okazaki fragment sequence data was obtained from Smith and Whitehouse [108]. Analysis primarily focussed on the larger “replicate” library (GEO:GSM835651), but results were confirmed in the “sample” library (GEO:GSM835650; file source: GSM835650_wt_sample.bed.gz; download date 14-15-2012). Results shown are for de-duplicated read data (identical start and end coordinates were considered duplicates). De-duplication minimises potential biases in PCR amplification. Qualitatively similar results were obtained with non-de-duplicated data and support identical conclusions. The Okazaki fragment strand ratio was calculated as the sum of per nucleotide read coverage on the forward strand divided by the same measure for the reverse strand. For each strand separately, the number of fragment termini at each nucleotide was compared to the average coverage of the six nucleotides upstream from it. Percentage fragment termination per nucleotide was calculated, providing a genome wide Okazaki junction rate.

2.1.7 **Yeast genomic annotations**

Annotated origins of replication were obtained from Eaton et al. [24]. DNase I hypersensitive sites and footprints were obtained from Hesselberth et al. [46] (file source: Hesselberth_2009_DNaseI_hypersensitive_sites_V64.bed; file date: 09-12-2011) and nucleosome position, occupancy and positional fuzzyness (positional heterogeneity) measures from Jiang and Pugh [52] (file source: jiangPugh2009dyadV64.ordered; file date 08.11.2012). H2A.Z containing nucleosome dyad positions were obtained from Guillemette et al. [39] (file source: Guillemette_2005_H2AZ_vs_H2-
CHAPTER 2. METHODS

*B_ChIP_chip_V64.bedgraph*; file date 16-12-2011), and forward and reverse stand transcripts downloaded from the UCSC sacCer3 database (table: *sgdGene*; download date 27-08-2012).

### 2.1.8 Human genomic annotations

DNase I hypersensitivity site footprints are provided from Hesselberth *et al.* [46], and downloaded from the ensemble ENCODE directory (file source: *ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011*; files dated 07-09-2012). Two files were used: *all.footprints.gz*, containing all footprints per cell type, and *combined.fps.gz*, containing adapted footprint boundaries, compensating for cell type specific variation of footprint coordinates at similar nucleotide positions. DNase I hypersensitivity sites were downloaded from UCSC hg19 database [53] (Table A.2). Nucleotide sequences downloaded from the UCSC hg19 database were used to calculate CpG dinucleotide rate. CpG methylation was downloaded from the UCSC database ENCODE HAIB Methyl Reduced Representation Bisulfite Sequence (RRBS) track (see supplementary table A.3 for file sources).

### 2.1.9 Novel nucleotide motif discovery

Nucleotide sequence FASTA files were either inputted into the “Discriminative Regulatory Expression Motif Elicitation” (DREME) software (version 4.9.1) [3], with default settings, to identify significantly re-occurring nucleotide motifs, or processed by the “seqLogo” package [89] (version 1.30.0) in R, to calculate the average frequency of each nucleotide base across all sites.
### Table 2.1: Summary of data sources.

<table>
<thead>
<tr>
<th>Data</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-exo (CTCF,Rap1,Reb1)</td>
<td>SRA:SRP008309</td>
</tr>
<tr>
<td>ChIP-seq (CREB1,ETS1,HNF4A</td>
<td>SRA:SRP008797</td>
</tr>
<tr>
<td>MEF2A,NFIC,SRF1,USF1)</td>
<td></td>
</tr>
<tr>
<td>ChIP-seq (CEBPA)</td>
<td>SRA:SRP008215</td>
</tr>
<tr>
<td>ChIP-seq (BRCA1,E2F1,ELK1,ELK4,GATA2,GATA3 MAX,MYC,IRF1,STAT1,YY1)</td>
<td>SRA:SRP007993, SRA:SRP003811, SRA:SRP007486, SRA:SRP006770</td>
</tr>
<tr>
<td>ChIP-seq (FOXA1,NKX31)</td>
<td>SRA:SRP010705</td>
</tr>
<tr>
<td>ChIP-seq (IRF2)</td>
<td>SRA:SRP012028</td>
</tr>
<tr>
<td>ChIP-seq (NFkB)</td>
<td>SRA:SRP002125</td>
</tr>
<tr>
<td>ChIP-seq (NFYA)</td>
<td>SRA:SRP005174</td>
</tr>
<tr>
<td>ChIP-seq (NR2F1,TFAP2A)</td>
<td>SRA:SRP006560</td>
</tr>
<tr>
<td>ChIP-seq (NR3C1)</td>
<td>SRA:SRP011408</td>
</tr>
<tr>
<td>ChIP-seq (PPARG)</td>
<td>SRA:SRP009091</td>
</tr>
<tr>
<td>ChIP-seq (REST)</td>
<td>SRA:SRP010385, SRA:SRP017090, SRA:SRP006944</td>
</tr>
<tr>
<td>TF binding motifs</td>
<td><a href="http://jaspar.genereg.net/html/DOWNLOAD/JASPAR_CORE/pfm/nonredundant/pfm_all.txt">http://jaspar.genereg.net/html/DOWNLOAD/JASPAR_CORE/pfm/nonredundant/pfm_all.txt</a></td>
</tr>
<tr>
<td>Yeast polymorphism data</td>
<td>ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/latest/misc.tgz</td>
</tr>
<tr>
<td>Human polymorphism data</td>
<td>ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521</td>
</tr>
<tr>
<td>Human between-species GERP</td>
<td><a href="http://hgdownload.cse.ucsc.edu/gbdb/hg19/bbi/All_hg19_RS.bw">http://hgdownload.cse.ucsc.edu/gbdb/hg19/bbi/All_hg19_RS.bw</a></td>
</tr>
<tr>
<td>Okazaki fragment sequencing</td>
<td>GEO:GSM835651, GEO:GSM835650</td>
</tr>
<tr>
<td>Yeast DHS footprints</td>
<td><a href="http://downloads.yeastgenome.org/published_datasets/Hesselberth_2009_PMID_19305407/track_files/Hesselberth_2009_DNaseI_hypersensitive_sites_V64.bed">http://downloads.yeastgenome.org/published_datasets/Hesselberth_2009_PMID_19305407/track_files/Hesselberth_2009_DNaseI_hypersensitive_sites_V64.bed</a></td>
</tr>
<tr>
<td>Yeast nucleosome dyad positions</td>
<td>jiangPugh2009dyadV64.ordered</td>
</tr>
<tr>
<td>Yeast H2A.Z dyad positions</td>
<td>Guillemette_2005_H2AZ_vs_H2-B_ChIP_chip_V64.bedgraph</td>
</tr>
<tr>
<td>Human DHS footprints</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011/all.footprints.gz</td>
</tr>
<tr>
<td></td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011/combined.fps.gz</td>
</tr>
<tr>
<td>Human Methyl CpG RRBS</td>
<td><a href="http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/wgEncodeHaibMethylRrbs*UwSitesRep1.txt.gz">http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/wgEncodeHaibMethylRrbs*UwSitesRep1.txt.gz</a> (*see Table A.3)</td>
</tr>
</tbody>
</table>
The wavering rate of substitution across the genome is the result of complex influences from selection and mutation. When undertaking the challenging task of identifying variants within our DNA affecting phenotypes or disease, having a better understanding of how selection and mutation shape a region of interest will assist in determining the importance of the underlying DNA sequence, and consequently the importance of any variants within it.

To develop a better understanding of these processes our attention was drawn to the substitution rate pattern at the binding sites of the transcription factor NFYA (Figure 1.11) [8]. At these sites an unusual conservation “shoulder” in observed; describing a rapid spike in substitution rate either side of the binding motif, in an region of otherwise low substitution. These sites could either be formed by purifying selection due to a lack of sequence constraint, or site-specific elevated mutation rate, as a consequence of a mechanism yet to be defined.

### 3.1 Transcription factor substitution profiles: an introduction

Transcription factors are cellular proteins which control gene-expression through direct DNA binding [8]. A cell's behavior and characteristics are determined by which proteins it produces and when. Consequently the correct coordination of protein production by transcription factors is critical for correct cellular function [114]. For this reason, alteration to transcription factor
regulation is believed to be key to species diversity [11], as well as highly associated with disease: it has been estimated that a third of human developmental disorders are as a result of dysfunctional transcription factors [19], and aberrant transcription factors are known to be common contributors towards cancer [31].

Depending on definition, transcription factors are only those proteins that bind and regulate DNA by sequence-specific interactions. Different transcription factors recognise and bind different DNA sequence motifs, normally in the region of 5 to 12 nucleotides in length [75]. This contrasts with nucleosomes, for example, which contain 147 nucleotides of DNA bound in a non-sequence-specific manner (although some trends in their underlying nucleotide sequence have been observed [125]).

Transcription factors are composed of a DNA-binding domain, as well as functional domains which instigate the recruitment of further factors or regulate DNA directly. It is the characteristics of their DNA-binding domains which differentiate families of transcription factors, such as zinc-finger (675 human TFs), homeodomain (257 human TFs) and helix-loop-helix (87 human TFs) transcription factors (reviewed in [73]). It is the specific 3D structures of these domains and their amino acid interactions which facilitate specificity of sequence binding, and determine their binding to one DNA sequence but not another.

It has been estimated that there are between 2,000 and 3,000 different types of human sequence-specific transcription factors [119]. While many transcription factors perform specific functions others display diverse regulatory roles. CTCF for example, has been shown to act as isolator, enhancer and repressor [95]. The CTCF protein is composed of eleven zinc fingers, orientated asymmetrically across its binding site in varying combinations, resulting in a heterogeneous binding footprint [29]. Rhee and Pugh have described seven different motif combinations at CTCF binding sites [102] (detailed in section 8.3.2). CTCF’s binding heterogeneity presents one possible way by which it perform diverse functions, depending on the combination by which it binds, and consequently the conformation it takes.

NFYA is one component of a protein-trimer, along with NFYB and NFYC. NFYA is the subunit providing capacity to bind avidly to a CCAAT motif in transcriptional regulatory regions [21]. Genes whose expression is controlled by NFYA are associated with pro-growth, anti-apoptotic roles [21], defining NFYA in cell-cycle control. Evidence exists that NFYA can act both as a gene activator and gene repressor [93], depending on its cooperation with transcriptional co-factors [10], or its own post-translational modification [133].

Understanding transcription factor binding is extremely important in the understanding of the genome, its disease states and its diversity. Because transcription factors bind in a sequence specific manner their binding sites are frequently conserved [8]. Consequently mutation of these
3.2 Transcription factor substitution profiles: data and methodologies

3.2.1 Defining human transcription factor binding sites

To determine the extent of conservation “shouldering” around transcription factor binding sites, data was compiled for those transcription factors where a) ChIP-seq binding data is publicly available and b) a consensus binding motif is known. Both sets of data were available for thirty one human transcription factors. ChIP-seq data was downloaded from the NCBI Sequence Read Archive (SRA) [http://www.ncbi.nlm.nih.gov/sra]. Date sources provided in Table 2.1 and Supplementary Table A.1.

ChIP-seq SRA replicate files were converted to FASTQ files using fastq-dump, part of the NCBI SRA Toolkit software (version 2.1.9; download date 19-01-2012). FASTQ reads were aligned to the hg19 reference genome using Bowtie2 (version 2.0.0) [66], where a maximum of 10 best hits were reported per read. Variables: [bowtie -S -k 10 –best -m 10]

Resulting SAM files were converted to BAM files and sorted using SAMtools (version 0.1.16) [67]. Binding peaks were defined using the Model-based Analysis of ChIP-Seq (MACS) software (version 2.0.10) [135], for all replicates.

MACS defined peaks were recorded as a binding site only if a binding motif for that transcription factor was also present within 50 nt of the ChIP-seq peak summit. The presence or absence of a motif was determined using the Motif Occurrence Detection Suite (MOODS; version 1.0.1; download date 04-03-2013) [64], using a motif significance threshold of 0.005. Consensus binding
motifs were downloaded as positional weight matrices (PWM) from JASPAR [75], (file source: http://jaspar.genereg.net/html/DOWNLOAD/JASPAR_CORE/pfm/nonredundant/pfm_all.txt; file dated 19-11-2013).

Peaks between replicates were combined using corresponding motif mid-points. For each binding site the motif mid-point, motif strand orientation (provided by MOODS), and ChIP-seq coverage (provided by MACS), was recorded. “Secondary” sites were noted if another peak of higher ChIP-seq coverage was present within 400nt.

3.2.2 Calculating substitution rate at transcription factor binding sites

Genomic Evolutionary Rate Profiling (GERP) scores [17], were used as a measure of substitution rate. Scores were downloaded as bigWig files from the UCSC hg19 database (file source: All_hg19_RS.bw; file dated 15-04-2011). Binding peak coordinates for each transcription factor were used to define a focal position about which GERP scores were extracted.

For each transcription factor the following algorithm was implemented in R (version 3.0.0):

1. GERP scores (extracted for each non-secondary, non-telomeric binding site, centred on motif mid-point), were orientated according to their binding motif strand (i.e. when binding motifs were on the reverse strand GERP arrays were inverted).

2. Binding sites were ranked according to their ChIP-Seq coverage at the binding peak summit and divided into four groups of equal size; quartiling transcription factor binding sites according to strength of binding.

3. Orientated, motif mid-point-centred, per nucleotide GERP arrays for each quartile were summed, averaged and plotted.

3.2.3 Producing Positional Weight Matrix logos at binding sites

Positional Weight Matrices (PWMs) were calculated from all nucleotide sequences at each binding site, for each transcription factor. Nucleotide sequences were downloaded as FASTA files, per chromosome, from the UCSC hg19 database. Refined sequence arrays for each binding region were then also saved as FASTA files. FASTA files were processed by the “seqLogo” package [89]
3.3. TRANSCRIPTION FACTOR SUBSTITUTION PROFILES: RESULTS

Figure 3.1: **Substitution rate profile at NFYA binding sites.** NFYA binding sites are centred on their motif mid-point and orientated by motif strand, so that for all sites motifs are aligned and in the same orientation (illustrated by the regional motif sequence strength by PWM, top plot). Sites are split into four groups of equal numbers according to their binding strength (ChIP-Seq coverage). Substitution rate (GERP) for each group is plotted in red, from light to dark (weakest to strongest binding strength). Genomic neutrality (y=0; based on genome-wide 4D sites, the implications of which are discussed in sections 1.4.4 and 3.4.2), is illustrated by the green dotted line. Bottom and top x-axes differ. The top plot represents the strongest ChIP-Seq coverage quartile at increased magnification, matching that of its above PWM.

(Version 1.30.0) in R (version 3.0.0), to calculate the average frequency of each nucleotide base, across each nucleotide position, for all binding sites.

3.3 Transcription factor substitution profiles: results

3.3.1 Motif dependent substitution asymmetry at NFYA binding sites

To begin to elucidate evolutionary forces shaping the NFYA conservation “shoulder”, I further investigate the substitution profile at NFYA binding sites. ChIP-Seq data was used to determine NFYA binding interaction peaks across the genome. GERP scores [17], as a measure of substitution rate, were centred on each NFYA motif mid-point and orientated according to binding motif strand. Binding sites were ranked according to their binding site strength, determined by ChIP-Seq coverage, to distinguish between frequency of occupation. GERP scores within each coverage quartile were averaged and plotted (Figure 3.1).
A detailed pattern in substitution rate around NFYA sites arises. GERP’s assumed genomic neutrality, calculated by genome-wide 4D sites, is centred on zero (green dotted line). Consequently values below the green line on the plot are considered conserved, values above considered diverged. The most striking profile is for sites with highest ChIP-Seq coverage (dark red), which are the binding sites I call with the strongest confidence and represent sites of strongest DNA-binding interaction. Substitution at these sites is extremely low at the binding motif, as expected by purifying selection at functionally important sequence-dependent sites. Either side of the binding motif are peaks of strong divergence, representing high substitution rate, matching the conservation “shoulder” reported by Boyle et al. [8]. Notably there is significant asymmetry surrounding the binding motif. This asymmetry is a signal not observe by Boyle et al., presumably as their sites were not orientated by motif strand. The nucleotide base Positional Weight Matrices (PWMs) identifies the nucleotide strength of the known motif only, with no notable asymmetry, although subtleties are debatable.

Decreased substitution rate at NFYA binding sites is consistent with purifying selection. However the pronounced effect of the sites’ proximately elevated substitution rate, particularly at sites with strongest binding signal is not predicted by purifying selection. If the genome-wide neutrality provided by GERP is correct for the region, such elevated substitution rate must be indicative of either site-specific mutation rate or positive selection.

3.3.2 Varied factor-dependent substitution rate profiles

To further investigate the extent of conservation shoulders throughout the genome I look at all available transcription factor binding sites. I compile data for those transcription factors where a) ChIP-seq binding data has been previously published, to confirm binding interactions and b) a consensus binding motif is known, to identify precise sites of binding for accurate alignment. Both sets of data were available for thirty one human transcription factors (Table 3.1).

Again binding sites for each transcription factor were divided into quartiles according to their ChIP-Seq coverage, as a means of ranking according to strength of binding interaction and frequency of occupation. Substitution rates, defined by GERP, were again centred on motif mid-points and orientated according to motif strand (Figure 3.2).

Between transcription factors I see varied and specific profiles. Common between all factors is extremity of profiles at sites of highest binding confidence (dark red), as well as asymmetry when orientated by motif strand. Also common is a spike in substitution rate adjacent to and/or within motifs. Spikes often surpass the genomic background level, indicating positive selection
### Table 3.1: Transcription factors with known binding motif and available ChIP-seq alignment data

Class and family provided by Jaspar ([http://jaspar.genereg.net/](http://jaspar.genereg.net/)) [75]

<table>
<thead>
<tr>
<th>name</th>
<th>class</th>
<th>family</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>other</td>
<td>other</td>
</tr>
<tr>
<td>CEBPA</td>
<td>zipper-type</td>
<td>leucine zipper</td>
</tr>
<tr>
<td>CREB1</td>
<td>zipper-type</td>
<td>leucine zipper</td>
</tr>
<tr>
<td>CTCF</td>
<td>zinc-coordinating</td>
<td>beta-beta-alpha zinc finger</td>
</tr>
<tr>
<td>E2F1</td>
<td>winged helix-turn-helix</td>
<td>E2F</td>
</tr>
<tr>
<td>ELK1</td>
<td>winged helix-turn-helix</td>
<td>Ets</td>
</tr>
<tr>
<td>ELK4</td>
<td>winged helix-turn-helix</td>
<td>Ets</td>
</tr>
<tr>
<td>ETS1</td>
<td>winged helix-turn-helix</td>
<td>Ets</td>
</tr>
<tr>
<td>FOXA1</td>
<td>winged helix-turn-helix</td>
<td>Forkhead</td>
</tr>
<tr>
<td>FOXO3</td>
<td>winged helix-turn-helix</td>
<td>Forkhead</td>
</tr>
<tr>
<td>GATA2</td>
<td>zinc-coordinating</td>
<td>GATA</td>
</tr>
<tr>
<td>GATA3</td>
<td>zinc-coordinating</td>
<td>GATA</td>
</tr>
<tr>
<td>HNF4A</td>
<td>zinc-coordinator</td>
<td>hormone-nuclear receptor</td>
</tr>
<tr>
<td>IRF1</td>
<td>winged helix-turn-helix</td>
<td>IRF</td>
</tr>
<tr>
<td>IRF2</td>
<td>winged helix-turn-helix</td>
<td>IRF</td>
</tr>
<tr>
<td>MAX</td>
<td>zipper-type</td>
<td>helix-loop-helix</td>
</tr>
<tr>
<td>MEF2A</td>
<td>other alpha-helix</td>
<td>MADS</td>
</tr>
<tr>
<td>Myc</td>
<td>zipper-type</td>
<td>helix-loop-helix</td>
</tr>
<tr>
<td>NFIC</td>
<td>other</td>
<td>NFI CCAAT-binding</td>
</tr>
<tr>
<td>NKxB</td>
<td>Ig-fold</td>
<td>Rel</td>
</tr>
<tr>
<td>NFYA</td>
<td>other alpha-helix</td>
<td>NFY CCAAT-binding</td>
</tr>
<tr>
<td>NKX3-1</td>
<td>helix-turn-helix</td>
<td>Homeo</td>
</tr>
<tr>
<td>NR2F1</td>
<td>zinc-coordinating</td>
<td>hormone-nuclear receptor</td>
</tr>
<tr>
<td>NR3C1</td>
<td>zinc-coordinating</td>
<td>hormone-nuclear receptor</td>
</tr>
<tr>
<td>PPARG</td>
<td>zinc-coordinating</td>
<td>hormone-nuclear receptor</td>
</tr>
<tr>
<td>REST</td>
<td>zinc-coordinating</td>
<td>beta-beta-alpha zinc finger</td>
</tr>
<tr>
<td>SRF</td>
<td>other</td>
<td>alpha-helix</td>
</tr>
<tr>
<td>STAT1</td>
<td>Ig-fold</td>
<td>Stat</td>
</tr>
<tr>
<td>TFAP2A</td>
<td>zipper-type</td>
<td>helix-loop-helix</td>
</tr>
<tr>
<td>YY1</td>
<td>zinc-coordinating</td>
<td>beta-beta-alpha zinc finger</td>
</tr>
<tr>
<td>USF1</td>
<td>zipper-type</td>
<td>helix-loop-helix</td>
</tr>
</tbody>
</table>
CHAPTER 3. SUBSTITUTION PROFILES AT HUMAN TRANSCRIPTION FACTOR BINDING SITES

assumed neutrality
first ChIP−Seq coverage quartile (lowest)
second ChIP−Seq coverage quartile
third ChIP−Seq coverage quartile
forth ChIP−Seq coverage quartile (highest)

Figure 3.2: (one of six..)
3.3. TRANSCRIPTION FACTOR SUBSTITUTION PROFILES: RESULTS

![Graphs showing transcription factor substitution profiles](image)

Figure 3.2: (..two of six..)
Figure 3.2: (...)three of six..
3.3. TRANSCRIPTION FACTOR SUBSTITUTION PROFILES: RESULTS

Figure 3.2: (..four of six..)
CHAPTER 3. SUBSTITUTION PROFILES AT HUMAN TRANSCRIPTION FACTOR BINDING SITES

Figure 3.2: (five of six.)
Figure 3.2: Substitution rate profiles at transcription factor binding sites. Transcription factors (top left), are labeled along with their total number of binding sites (top right), and ordered alphabetically. GERP scores are centred on binding motif mid-points ($x = 0$; light blue), and orientated according to motif strand. Binding sites are divided into quartiles according to strength of ChIP-seq coverage. Each quartile is averaged and plotted separately, from light pink (weakest ChIP-seq coverage) to dark red (strongest ChIP-seq coverage). The green dotted line marks a GERP score of zero, i.e. assumed genomic background neutrality. Y-axes are inverted for consistency with further plots, therefore GERP scores above zero (below the green dotted line) indicate constraint and reduced substitution rate, while a score below zero (above the green dotted line) indicating divergence and increased substitution rate. Top and bottom X-axes differ. Top plots represent the strongest ChIP-Seq coverage quartile at increased magnification, matching that of their above PWM.

or elevated mutation rate at binding sites of multiple transcription factors across the genome. Conservation shoulders are therefore not purely a characteristic of NFYA binding.

In some instances patterns in substitution appear explicable by underlying sequence, demonstrated by PWM logos. For example, IRF1, IRF2, NKxB, REST, STAT1 and TFAP2A motifs are made of poorly conserved mid-sequences flanked by strongly conserved outer sequences. In reflection, substitution rate forms a bi-modal “\$\sqrt[4]{\text{}}\$”. Such consistently low bit-score positions within motifs could be considered proxies to neutrality within non-coding sequence, akin to 4D sites in coding sequence.
Another contributor to substitution patterns may be CpGs, which are prone to mutation by deamination (chapter 1.3.2). For example, MAX and USF1 both display spikes in substitution rate at the same position as a CpG within their motif. In contrast, the pronounced spike in GERP score at the centre of the TFAP2A motif is not obviously associated with a bias for CpG dinucleotides.

### 3.4 Transcription factor substitution profiles: a discussion

#### 3.4.1 Defining transcription factor binding by both ChIP-Seq coverage and motif orientation

Boyle et al.’s NFYA figure is a key observation as the conservation shoulders they observe have the potential to further explain the poorly understood forces of selection and mutation which shape our genome. Here are sites within well conserved regions that display a marked elevation in substitution rate, for which there is no known explanation. To begin to understand the underlying forces responsible, we first wished to assess the extent of the phenomenon, by plotting substitution rate at all available human transcription factor binding sites.

In the same manner as Boyle et al. [8], transcription factor binding sites were defined by both ChIP-Seq and binding motifs. Two novel approaches were then taken to refine binding site definition, which extended my research further than that of Boyle’s:

- **Tiering by strength of binding**: binding sites were quartiled according to their ChIP-Seq coverage, representing strength and occupation of binding. By this means I compare patterns in substitution rate between sites where transcription factor binding is strong and frequent with sites where transcription factor binding is weak and infrequent.

- **Orientation**: when binding motifs were identified within a ChIP-Seq peak the motif strand was also recorded. This gave me the power to orientate binding sites so that in all instances motifs are aligned in the same orientation.

It is evident that both binding strength tiering and orientation provide additional information regarding substitution rate at these sites. While conservation directly at binding sites and divergence at shoulders are modest at weakly and infrequently bound sites, both observations are exaggerated at sites of strongest binding. In other words, the stronger the binding site conservation the greater the shoulder divergence. This is remarkable: even though purifying
selection increases with strength of binding, shoulders are not likewise reduced, as expected under a model of uniform mutation rate, but are instead further elevated.

Division by quartiles thus provides both improved substitution rate resolution, when binding is strongest, and also the direction of resolution trend, by comparing plots as strength and frequency of occupation increases. At NFYA sites, as well as for CTCF, ELK4, REST, TRAP2A and USF1 for example, the resolution trend extends in opposite directions at binding sites and shoulders; decreasing and increasing respectively. It is notable that NFYA, CTCF and REST, which show the strongest signals, are also factors known to bind DNA with strong affinity, with extensive DNA contact [22, 113]. For other factors, CEBPA, CREB1, FOXO3, SRF, YY1, GATA2, GATA3 and NFIC, for example, the resolution trend for the whole region, as binding strength increases, is towards a decrease in substitution rate.

Orientation by motif strand reveals substantial asymmetry in substitution rate profiles. At sites of strongest binding, substitution rate is greater within the downstream shoulder of NFYA binding sites, while the upstream shoulder differs only a small degree between quartiles and from the genomic background rate. The orientation of CTCF binding sites reveals a highly asymmetrical profile, including a discrete “shoulder” on the upstream side of the binding site only. The CTCF protein is made up of 11 zinc fingers, able to bind in different asymmetrical combinations and conformations. Rhee and Pugh have reported seven different CTCF binding motifs [102]. While six of these motifs contain the core motif identified in our plots, three also contain an additional, less prominent motif, approximately five nucleotides upstream, fitting neatly with the upstream asymmetry in reduced substitution we observe. Orientation thus provides increased substitution rate detail, providing evidence that motif orientation is not arbitrary regarding substitution rate pattern.

### 3.4.2 GERP as a measure of both conservation and divergence

There are several measures available that quantify substitution rate across the genome (section 1.4). Each is based on a different algorithm, taking into account different considerations. One such measure is phastCons; used to measure substitution rate by Boyle et al. for transcription factor binding site profiles. phastCons incorporates only two alternate models: 1) neutral evolution and 2) purifying selection [106]. It is unable to identify regions evolving faster than neutral expectation. Consequently all phastCons values are below that or equal to that of the genomic background rate, acting as a useful measure for purifying selection. An alternative measure of substitution rate is GERP constraint scores [17]. GERP is able to measure both rates below the genomic background rate (purifying selection), and those above the genomic background rate.
(positive selection or elevated mutation). This gives it an added advantage over phastCons when investigating the possibility of locally elevated mutation rate.

We observe this advantage by comparing substitution rate profiles of NFYA binding sites between both measures. In Boyle’s figure, conservation shoulders spike above the neighbouring trend in substitution, but do not reach that of the genomic background rate. This indicates, according to phasCons, that the highest rate of substitution within the shoulders is still lower than the genomic average, and therefore under stronger purifying selection than the average genomic site. Comparatively, conservation shoulders extend well above the genomic background rate in my figures, where GERP scores are used as a measure of substitution rate. Any rates below zero, according to GERP, indicate DNA evolving at a rate faster than that of the genomic average. This cannot be explained by just neutral evolution and purifying selection alone, and thus challenges the hypothesis that steric hindrance is responsible for observed shoulders. Rates past zero are instead an indication of either positive selection (a rare occurrence across the genome, and difficult to explain considering strong purifying selection in the vicinity), or elevated mutation rate. It is as a consequence of using GERP constraint scores, rather than phastCons, that we are able to observe these processes, not previously recognised by Boyle et al. However, it must also be taken into consideration that GERP and phastCons algorithms assign genomic neutrality by different means: while phastCons adapts neutrality for regional variability (explained in section 1.4.3 [106]), GERP assigns a single rate of neutrality across the genome (explained in section 1.4.4 [17]). That rates surpass neutral under GERP could therefore be an algorithm-specific artifact, identifying an underestimation of GERP neutrality in such regions. If this is the case, it must then be asked why substitution rate as a whole is higher in such regions than the average site across the genome.

3.4.3 Disentangling compositional bias

All binding sites are aligned and orientated by the binding motif identified within them. As the specific sequence within a binding motif is the identifier for corresponding specific factors to bind, mutations within them are likely to have deleterious functional consequences. Consequently binding motifs are under strong purifying selection and exhibit low substitution rate. By aligning to such sites we intrinsically align to sites of low substitution rate, the resulting patterns of which must be taken into account when drawing conclusions about evolutionary effects such as elevated mutation rate at such regions.

At NFYA binding sites, nucleotides central to its motif are under strongest purifying selection, with outer nucleotides tapering to background level. This creates a single strong and narrow
3.4. TRANSCRIPTION FACTOR SUBSTITUTION PROFILES: A DISCUSSION

(motif length-dependent) dip in substitution rate. Comparatively IRF1, IRF2, NFκB, REST, STAT1, TFAP2A and USF1 display two strongly conserved nucleotide sequences flanking several nucleotides of reduced conservation. The result is an approximate “\%/\%” bi-modal pattern with low substitution at conserved flanks and increased substitution within.

CpG dinucleotides within binding motifs should also be considered. Due to their predisposition to deamination, CpG dinucleotides are a known cause of site-specific elevated mutation rate (see section 1.3.2). Consequently CpG dinucleotides are rare within binding motifs, or remain unmethylated to prevent subsequent deamination. However both MAX and USF1 contain a central CpG dinucleotide. In reflection these binding sites exhibit a centrally elevated substitution rate. At USF1 sites the CpG dinucleotide occurs in the interim nucleotides of a bi-modal motif, resulting in a potentially combinatory effect of reduced conservation and a CpG site-specific elevated mutation rate. Consequentially its central peak in substitution rate is high.

There are exceptions: CEBPA, for example, has the bi-modal motif from which a “\%/\%” substitution pattern would be expected, but has a simple dip in substitution rate at its binding site, similar to that of NFYA. HNF4A forms a “\%/\%” substitution pattern, albeit mild, but so too does the conservation of its motif, predicting a “\%/\%” substitution form. CTCF sites are beautifully intricate, but well defined sites of low substitution are not obviously accounted for by their underlying sequence conservation (namely the second dip upstream of the core binding motif).

An explanation for such discrepancies is multiple binding conformations by single transcription factors. CTCF is a good example for this, with its multiple binding configurations and corresponding varied motifs [102]. The presence of a discreet upstream motif, perhaps just visible in the sequence PWM logo, exactly aligns to the second upstream drop in conservation. If CTCF binds DNA in multiple conformations the current plot will combine their substitution rate profiles, blurring interesting evolutionary signals. It would therefore be of interest to separate profiles according to their binding interactions, to assess patterns in substitution rate devoid of confounding overlap.

Strength or occupancy of binding is useful when deciphering evolutionary forces; some substitution profiles are explained purely by underlying sequence and purifying selection. However other signals are not as conveniently accounted for, and there are clearly additional signals at such sites. Shoulders are formed outside binding motifs, such as at NFYA sites, for example. At other sites, such as at the binding sites of TFAP2A, a central dip in sequence conservation (peak in substitution), extends far past the genomic background level, which cannot be explained purely by lack of conservation. Compared to GERP’s neutral proxy these patterns would confirm either diversifying (positive) selection or elevated mutation rate.

55
In Chapter 3 I demonstrate that “shoulders” of locally elevated substitution rate are a common, but not universal, feature of human transcription factor binding site evolution. These peaks in substitution rate extend well above that of the local genomic background rate and genome wide estimate of neutral evolution, implicating a role for either positive selection or elevated mutation rate in their formation. As may be expected, it appears that the stronger or more stable the binding of a transcription factor to the DNA, the greater the apparent purifying selection on the sequence specific binding site. However, strikingly, I find that stronger binding also corresponds to greater elevation of nucleotide substitution rate in shoulder regions. This observation seems difficult to reconcile with a model of positive selection, and seems most consistent with a model where stable binding of proteins to DNA in some way leads to local elevation of nucleotide substitution mutations.

Intriguingly, Smith and Whitehouse recently showed that three specific transcription factors, (and nucleosomes), can act as partial barriers to DNA replication in yeast [108]. This suggested the interesting possibility that protein binding to DNA during replication could perturb normal replication processes and may have an impact on replication fidelity. I explore this possibility in this chapter and introduce a mechanistic model that could explain locally elevated mutation rate at protein binding sites. This model was first proposed by Martin Taylor and developed jointly with myself throughout the course of this thesis.
CHAPTER 4. THE LAGGING STRAND HYPOTHESIS

4.1 The lagging strand hypothesis: an introduction

During every cell cycle each cell must make a complete replica of its genome, with minimum error; both the entirety of its DNA as well as the epigenetic marks that accompany it. Replication of DNA does not occur as a single molecule, but starts at multiple ‘origins of replication’ across the genome (Figure 4.1). Replication fork protein complexes pry open the neighbouring DNA double helix in both directions, to expose short stretches of single stranded DNA [88]. The exposed single stranded DNA is simultaneously replicated by DNA polymerases.

DNA is only replicated in the 5’ to 3’ direction [88]. Consequently only one of the DNA strands, the leading strand, can be replicated in a continuous manner, a process performed by DNA polymerase \( \epsilon \) (Pol \( \epsilon \); reviewed in [123]). As the other strand, the lagging strand, is orientated in the opposite direction, continuous 5’ to 3’ replication is not an option. The lagging strand is therefore replicated discontinuously in many short 5’ to 3’ fragments, as new stretches of single stranded DNA are exposed, termed “Okazaki Fragments” [88]. Okazaki synthesis is initiated by the error-prone DNA primase/polymerase \( \alpha \) (Pol \( \alpha \)) [85], prior to high fidelity elongation by
4.1. THE LAGGING STRAND HYPOTHESIS: AN INTRODUCTION

polymerase δ (Pol δ) [121]. As the new upstream Pol δ fragment is made it displaces the Pol α tract of the fragment ahead of it. The resulting Pol α flap is removed and degraded, and the two neighbouring fragments ligated together to form a continuous strand.

Smith and Whitehouse have published sequence alignment data of Okazaki fragments across the yeast *Saccharomyces cerevisiae* genome [108]. Fragments were produced in a DNA ligase I depleted yeast background; the enzyme required for the ligation of neighbouring Okazaki fragments into one. As DNA ligase I is an essential gene in yeast this was implemented as a temperature sensitive mutation grown at the non-permissive temperature prior to DNA extraction and sequence library preparation. In this system, complete Okazaki fragments were synthesised by Pol α and Pol δ, Pol α flaps still removed and degraded, but ligation of neighbouring Okazaki fragments inhibited. Smith and Whitehouse's fragments were not present in cells until DNA replication. When the purified DNA was treated with a recombinant DNA ligase the number of fragments diminished, illustrating the presence of nicked DNA as opposed to single-stranded gaps or flaps, confirming the desired properties of Okazaki fragments. Fragments were aligned to the *S. cerevisiae* genome using deep sequencing, providing genome wide strand-specific depth of coverage (Figure 4.2 a). From this data the frequency at which lagging strand replication occurs on each strand at each nucleotide position was inferred. Despite the elegance of this model, it must be remembered that it represents a severely (lethally) perturbed system that may not recapitulate normal lagging strand replication patterns with complete fidelity.

Smith and Whitehouse plot the 5' and 3' ends of Okazaki fragments. They show that Okazaki termini are preferentially located at nucleosome dyads (Figure 4.2 b). To distinguish between cause and effect, between Okazaki junction formation and nucleosome binding, they plotted the termini of Okazaki fragments from a Pol δ Pol32 subunit knockout strain, the subunit associated with increased processivity. In the reduced Pol δ processivity strain Okazaki junctions are shifted upstream of nucleosome dyads (Figure 4.2 c). Smith and Whitehouse conclude that bound nucleosomes are acting as partial barriers to Pol δ to cause their dissociation and subsequent termini. When processivity is reduced, so too is their momentum to track around each nucleosome, and as a consequence dissociation is premature.

Smith and Whitehouse report that Okazaki termini are also enriched at known transcription factor binding sites, specifically those of Afb1, Reb1 and Rap1 (Figure 4.2 d). These are three transcription factors known to be essential in chromatin association [2, 90]. The authors propose that these factors bind more rapidly and tightly to newly replicated DNA than other factors. Alignment of fragment termini to their binding sites, specifically to the side first to encounter Pol δ, shows that along with nucleosomes the binding of Afb1, Reb1, and Rap1 to DNA has the potential to influence Pol δ dissociation and Okazaki termination by blocking the upstream Pol δ
Our attention was drawn to Smith and Whitehouse's publication, and the inspiration for our developing hypothesis was drawn from it, due to the striking similarity between the distribution of Okazaki termini at nucleosome dyads and the 'nucleosome hump'; the observed hump in elevated mutation rate surrounding nucleosome dyads. If the two observations are correlated one could infer a link between the frequency of Okazaki junctions and elevated mutation rate, and a potential link between Okazaki junctions at transcription factor binding sites, elevated mutation rate, and the formation of substitution rate shoulders at these sites.

The involvement of nucleosomes and transcription factors in Okazaki termination predicts their
4.2. THE LAGGING STRAND HYPOTHESIS: DATA AND METHODOLOGIES

rapid binding to replicating DNA and its newly synthesised Okazaki fragments. If their binding acts as a barrier and causes Pol δ dissociation, then the location that they bind will influence how far the Pol δ molecule of the new upstream fragment can proceed, and consequently how much Pol α and possibly Pol δ tract of the downstream fragment is displaced, removed and degraded. As Pol α is error-prone any retention of its tract is likely to harbour an elevated mutation rate. We therefore hypothesis that substitution rate would be elevated immediately downstream of Okazaki fragment termini if nucleosomes or transcription factors were ever, in their haste, to bind anywhere upstream of a fragment’s Pol α-δ junction (Figure 4.3).

Based on Smith and Whitehouse’s evidence we propose a candidate mechanism for the elevated substitution rate shoulders observed at the binding sites of a subset of transcription factors. We propose such transcription factors may possess binding properties lending them as partial barriers to Pol δ in lagging strand replication. In such instances their rapid binding to the lagging strand prior to Okazaki fragment ligation will have the potential to inhibit complete displacement of the Pol α tract, and in rare instances result in error-prone Pol α tract retention.

We consequently predict conservation shoulders to be a consequence of elevated mutation rate, as opposed to positive selection. For this hypothesis to hold substitution rate must correlate with Pol α retention, itself correlated to Okazaki termini. I begin to explore this hypothesis by correlating the frequency and pattern of Okazaki junctions with substitution rate at experimentally determined binding sites for the yeast transcription factors Reb1 and Rap1; both implicated by Smith and Whitehouse to be strong and fast binding during replication [108], capable of Pol δ dissociation and subsequently potentially responsible for Okazaki junction formation.

4.2 The lagging strand hypothesis: data and methodologies

4.2.1 Calculating Okazaki junction rate

Genome wide Okazaki junction rates were defined in the group by Martin Taylor: Okazaki fragment sequence data was obtained from Smith and Whitehouse [108]. Analysis primarily focussed on the larger “replicate” library (GEO:GSM835651) but results were confirmed in the “sample” library (GEO:GSM835650). Rather than using separate Okazaki 5’ and 3’ end counts we produced a normalised Okazaki junction rate measure. This is the average of (1) the fraction of upstream Okazaki fragments that terminate with a 3’ end at a focal nucleotide, and (2) the fraction of downstream Okazaki fragments whose 5’ end is at the focal nucleotide. The upstream and downstream coverage measures were based on mean Okazaki read coverage for
CHAPTER 4. THE LAGGING STRAND HYPOTHESIS

Figure 4.3: **The lagging strand hypothesis: substitution rate directed by Okazaki polarisation.** Replication is illustrated in top schematic moving through double stranded DNA from right to left by Pol ε on the leading DNA strand and both Pol α (purple to blue) and Pol δ (black) on the lagging strand. Factors (cylinders) are bound at Okazaki junctions prior to ligation and involved in upstream Pol δ dissociation. If factors were to bind marginally upstream of the downstream Okazaki fragment’s Pol α:δ junction the remaining portion of the error-prone Pol α tract would be retained in the newly synthesised DNA (illustrated in central diagram, boxed). Over evolutionary time this process would accumulate mutations at such sites. Such a substitution rate would be asymmetrical when sites are orientated by direction of lagging strand replication, as Pol α tracts will only be retained downstream of Okazaki junctions (bottom plot; orange).

the nucleotides located between 5 and 12 nucleotides upstream (downstream) of the focal 3’ (5’) end. This Okazaki Junction rate was calculated at single nucleotide resolution over both strands of the sacCer3 genome. Original data was aligned to the May 2008 *S.cerevisiae* genome assembly, deduct as V59. LiftOver was performed to convert coordinates to the current genome release, V64 (sacCer3). Read coverage and fragment termini were calculated on a per nucleotide basis.

62
4.2.2 Defining transcription factor binding sites

Reb1 and Rap1 ChIP-exo data (Table A.1), published by Rhee and Pugh [102], was downloaded from the NCBI Sequence Read Archive (SRA) [http://www.ncbi.nlm.nih.gov/sra]. SRA files were converted to colourspace FASTQ file format, using the fastq-dump software (download date 19-01-2012). Colourspace FASTQ reads were sorted using a purposefully written Perl algorithm according to corresponding transcription factor barcodes, and barcodes removed from each read.

Colourspace FASTQ reads were aligned by Bowtie [66] to the V64 (sacCer3) reference genome (files dated 13-11-2009), to produce SAM file output. Following the protocol by Rhee and Pugh [102], up to three mismatches across the length of each tag sequence were allowed, and the 3' most 6bp removed, (reportedly higher error rates); variables: [bowtie -C -S -3 6 –chunkmbs 50000 -n 3 -y -k 10 –best -m 10].

Conversion of files from SAM to BAM, peak definition with MACS and motif confirmation through MOOIDS followed the protocol as described in 3.2.1 for human transcription factor binding site definition.

Following the protocol defined by Rhee and Pugh, sites were defined as monomer if no other peaks were present within 100 bp. Where two or more peaks were present within 100 bp the peak with the highest occupancy was labeled as the primary peak, and additional peaks secondary. Telomeric sites were excluded using annotations within the sacCer3 sgdOther UCSC table [http://www.yeastgenome.org/]. Primary and monomer non-telomeric sites were used for further analysis, (Reb1 \( n = 834 \); Rap1 \( n = 486 \)).

4.2.3 Calculating polymorphism rate at Reb1 binding sites

Polymorphism rate and call frequency was previously calculated across the \( S.cerevisiae \) genome within the group by Martin Taylor (unpublished). Polymorphism data was downloaded from the sacCer3 genome project [ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryot-es/fungi/Saccharomyces_cerevisiae/SacCer_Apr2011/; download date: 07-05-2014]. For each nucleotide a polymorphism count value (0-3/NA) represents polymorphism rate, and a polymorphism call value (0-1/NA) represents the confidence by which the rate is called. Only polymorphism rates where the call value is 1 are included in analysis.

Transcription factor binding sites were orientated arbitrarily, their neighbouring polymorphism count and call values extracted, and rates averaged per nucleotide.
CHAPTER 4. THE LAGGING STRAND HYPOTHESIS

4.3 The lagging strand hypothesis: results

4.3.1 Okazaki junction rate is independent of Okazaki coverage

Okazaki coverage and read ends are provided by Smith and Whitehouse across the yeast *S. cerevisiae* genome [108]. This data was downloaded and used for our extended analysis. In Smith and Whitehouse’s figures it is evident that 5’ (red) and 3’ (blue) Okazaki ends do not entirely match, as would be expected for correlating Okazaki junctions (Figure 4.2 b:d). Various acquisition biases such as size selection of fragments could be responsible for this. Consequently, instead of quantifying fragment ends at each nucleotide position, as investigated by Smith and Whitehouse, we quantify the rate of Pol δ drop-off, by quantifying the rate at which Okazaki fragments start and terminate at each nucleotide position. The number of fragment ends at each nucleotide was compared to neighbouring sequence coverage, to produce a genome wide Okazaki junction rate (Figure 4.4 a).

Smith and Whitehouse’s original strand specific Okazaki coverage reflects the rate of lagging strand replication over each region of the genome. Coverage is therefore higher upstream of origins of replication (Autonomously replicating sequence (ARS) consensus Sequences (ACS)) on the ‘Crick’ strand and downstream of origins of replication on the ‘Watson’ strand (Figure 4.4 b). Our calculation of junction rate, however, displays a reduced relationship to neighbouring origins of replication (Figure 4.4 c). This is illustrated by plotting junction rate across chromosome 10, where no regional patterns are observed (Figure 4.4 a). By calculating junction rate we therefore increase independence from the influence of Okazaki synthesis frequency.

4.3.2 Okazaki junctions and polymorphism rate track neatly at Reb1 and Rap1 binding sites

Smith and Whitehouse report the association of Okazaki termini with nucleosome dyads and the specific strong and fast binding transcription factors Abf1, Reb1 and Rap1 [108]. I note the similarity in distribution of Okazaki termini around the nucleosome dyad and the mutation rate “hump” previously observed at nucleosome dyads by Prendergast and Semple [99], and speculate a relationship between Okazaki termini and rate of mutation. To investigate whether a correlation exists I plot polymorphism rate and Okazaki junction rate at the binding sites of Reb1 and Rap1 (high resolution binding data is not currently available for Abf1) (Figure 4.5).
4.3. THE LAGGING STRAND HYPOTHESIS: RESULTS

Figure 4.4: **Coverage bias between Okazaki frequency and Okazaki junction rate.** [a] forward (blue) and reverse strand (orange) Okazaki junction rate across chromosome 10. Autonomously Replicating Sequence (ARS) Consensus Sequence (ACS) origins of replication are shown in red. [b] average log2 ratio between Watson and Crick strand coverage is correlated to directionality of origins of replication, as represented by ACS [24]. [c] average log2 ratio between forward and reverse strand Okazaki junction rates diminishes replication origin directionality. All data unsmoothed.

Okazaki junction rate and polymorphism rate were plotted so that the mean and 10% deviations over the viewed window of each were aligned along the horizontal plane. The resulting plot illustrates a striking similarity between the distribution of Okazaki junctions and polymorphism rate at these sites, both in profile and magnitude. Both data sets begin to dramatically spike approximately twenty nucleotides from binding site mid-points, with the highest rate for both values at approximately ten nucleotides up and downstream. The two data sets peak at almost
4.4 The lagging strand hypothesis: a discussion

4.4.1 Yeast transcription factor conservation

Conservation shoulders were reported at human transcription factor binding sites by Boyle et al. [8], and further supported by my own analysis (Figure 3.2). Here I plot substitution rate at Reb1 and Rap1 transcription factor binding sites in yeast S.cerevisiae (Figure 4.5). At human binding...
sites, between-species GERP constraint scores were used as a measure of substitution rate. In the case of *S. cerevisiae*, polymorphism data is used.

A shoulder of conservation is seen both at Reb1 and Rap1 binding sites, as with human sites: the binding site is low in substitution rate and well conserved, an expected consequence of purifying selection, but a sharp and unexpected increase in substitution rate follows immediately flanking on both sides, before gradually returning to an approximate genomic background level.

Okazaki fragment data is only currently available for the yeast *S. cerevisiae* genome. Based on similarities between conservation shoulders at yeast-bound transcription factors and the human transcription factor conservation shoulder, we propose the same driving mechanism is present in both humans and yeast, and therefore that it is an evolutionary conserved process. It would be desirable to directly measure Okazaki junction rates in human cells, to compare them with protein binding sites and nucleotide substitution patterns, but for the moment at least this technically demanding feat has not been achieved. However, I later introduce a possible proxy for peaks of Okazaki junction rates that utilises existing human genomic data.

### 4.4.2 Intrinsic coupling of Okazaki junction and nucleotide substitution rates

Interesting patterns in substitution rate are found in the immediate vicinity of transcription factor binding sites in humans. The causes of these patterns are unknown. We observe a similarity between the published distribution of Okazaki termini and potential elevated mutation rate at nucleosome dyads [108, 125, 126]. As Okazaki termini also align to the boundaries of certain strong and fast binding transcription factor binding sites we speculate potential for further correlation between the distribution of Okazaki termini and mutation rate at such binding sites. We hypothesis a role for lagging strand replication in elevated mutation rate at such sites, by retention of error-prone Pol α tracts immediately downstream of induced Okazaki junctions.

To test this hypothesis I plot polymorphism rate and Okazaki junction rate at binding sites of the yeast transcription factors Reb1 and Rap1; the purpose being to observe whether similarities exist in profiles between the two data sets. We see a striking match between polymorphism and junction rate distributions, both in profile and magnitude. This provides strong evidence for a link between Okazaki termini and elevated mutation rate, in support of our model. It does not however provide a directional causality. Our hypothesis proposes the rare retention of error-prone Pol α tracts at these sites to be responsible for the patterns we observe in polymorphism rate. Further analysis will be required to conclude such a hypothesis.
have previously shown that some sequence specific DNA binding proteins are associated with regions of locally elevated nucleotide substitution rate, and that these regions correspond precisely to regions with high Okazaki junction rate in yeast. This led us to propose a simple mechanistic model, implying that stretches of DNA synthesised by error-prone Pol α can be preferentially trapped in the fully replicated genome at sites where proteins bind rapidly and strongly to the lagging strand, prior to ligation of component Okazaki fragments.

My preceding analyses have depended on genomic annotations to define the location of specific binding sites. However, with single-nucleotide, strand-specific measures of Okazaki junction rate, it is possible to define local peaks of Okazaki junctions independently of external annotations. This potentially has three advantages: 1) there may be other stronger barriers to Pol δ processivity than those currently identifiable by known transcription factor binding sites, 2) defining peaks of Okazaki junctions will define a heterogenous population of Pol δ barriers, their collective analysis less confounded by nucleotide compositional distortions; an inherent property when aligning a homogeneous population of sequence specific transcription factor binding sites, 3) defining Okazaki junction peaks inherently involves setting an arbitrary threshold. Tuning of this threshold will dynamic selection of specific subsets of extremities, ensuring sufficiently large numbers of sites suitable for statistical analysis; a major limitation for the majority of yeast transcription factor binding sites.

In this chapter I define Okazaki junction super-peaks, explore their underlying biology and use them to further test the Pol α retention hypothesis.
5.1 Super-peaks: data and methodologies

5.1.1 Defining Okazaki junction super-peaks

Okazaki junction super-peaks were previously defined within the group by Martin Taylor (unpublished). Okazaki reads were aligned to the genome and the rate of Okazaki junctions calculated on a per nucleotide basis, (as previously described in 4.2.1). The single nucleotide position within the genome with the highest rate of Okazaki fragment junctions was defined as the first super-peak. Super-peaks were then defined in descending order until reaching a peak threshold, defined as:

\[
\text{peakThresh} = \text{genomeMean} + (\text{genomeSd} \times 5) = 0.0346
\]

Where:

- \(\text{genomeMean}\) = mean genome wide junction rate (0.0062; i.e. 0.62% of neighbouring Okazaki reads form a junction at a specific nucleotide position)
- \(\text{genomeSd}\) = standard deviation of genome wide junction rate

Consequently, peaks are sites where at least 3.47% of neighbouring Okazaki fragments form a junction at that specific nucleotide.

Only sites where the sum of forward and reverse strand reads were greater than thirty were considered, to filter for sites of poor sequence coverage, which through higher stochastic variance give false peak signals. Exclusion zones of twenty nucleotides were placed around each defined super-peak so that no two peaks were called on the same strand within this range. If two peaks were present within twenty nucleotides on different strands the two peaks were combined as a “bi-directional” peak (as opposed to a “uni-directional” peak). In the case of multiple candidates for bi-directional peaks the closest pairs were used.

5.1.2 Orientating polymorphism rate at Okazaki junction super-peaks

Polymorphism rate was calculated as previously described in 4.2.3 from the sacCer3 genome project [http://www.yeastgenome.org/] (file source: ftp://ftp.ncbi.nlm.nih.gov/genbank/ genomes/Eukaryotes/fungi/Saccharomyces_cerevisiae/SacCer_Apr2011; download date: 07-05-2014).
Super-peaks were orientated according to their Okazaki fragment strand ratios: forward strand Okazaki coverage was divided by reverse strand Okazaki coverage, thereby determining the consensus direction of Okazaki synthesis. For uni-directional peaks, if reverse strand coverage is greater than forward strand coverage at the super-peak position, all subsequent data was rotated. For bi-directional peaks, strand coverage was calculated at the site of the strongest super-peak. If the strongest peak was on the reverse strand, and coverage consequently greater on the reverse strand, all subsequent data would be rotated, and aligned to the reverse strand peak. If the strongest peak was on the forward strand, and coverage consequently greater on the forward strand, data would not be rotated. In this way all super-peaks are orientated so that the consensus direction of Okazaki fragment synthesis is from left to right.

5.1.3 Annotating genomic details to super-peaks

Yeast DNase I hypersensitivity coverage, nucleosomes dyad positions, the dyad positions of the specific H2A.Z containing nucleosome, and forward and reverse strand transcripts were annotated at super peak sites:

- DNase I hypersensitivity sites downloaded from Hesselberth et al. [46] (file source: Hesselberth_2009_DNaseI_hypersensitive_sites_V64.bed; file date: 09-12-2011).
- Nucleosome dyad positions downloaded from Jiang and Pugh [52] (file source: jiang-Pugh2009dyadV64.ordered; file date 08.11.2012)
- H2A.Z containing nucleosome dyad positions from Guillemette et al. [39] (file source: Guillemette_2005_H2AZ_vs_H2-B_Chip_chip_V64.bedgraph; file date 16-12-2011)
- Forward and reverse stand transcripts downloaded from the UCSC sacCer3 database, [http://www.yeastgenome.org/] (table: sgdGene; download date: 27-08-2012).

5.1.4 Motif discovery at super-peaks

Nucleotide sequences surrounding each super-peak, within a 2000 nucleotide window, were downloaded as FASTA files from the sacCer3 genome project [http://www.yeastgenome.org/]. FASTA files were inputted into the “Discriminative Regulatory Expression Motif Elicitation” (DREME) software [3], with default settings, to identify significantly re-occurring nucleotide motifs within super-peak regions. Positional Weight Matrices (PWMs) of each significant motif identified were saved for further analysis.
The Motif Occurrence Detection Suite (MOODS) software [64] (version: 1.0.1; download date: 04-03-2013) was used to re-identify significant motifs from DREME, using PWMs, and save mid-point coordinates of each novel motif within super-peak regions.

5.1.5 Sub-dividing super-peaks by underlying motifs

Super-peaks were assigned subcategories according to novel motifs identified by DREME within them. Super-peaks were matched to DREME motif categories if the mid-point of that motif was present within ten nucleotides of the super-peak. Super-peaks may therefore be assigned to multiple DREME motif categories if multiple motif mid-points are present within ten nucleotides of the super-peak. A separate subcategory contains those super-peaks without any significant DREME motif within ten nucleotides.

5.2 Super-peaks: results

5.2.1 Identifying Okazaki junction super-peaks

To test the Pol $\alpha$ retention hypothesis, we are interested to see if polymorphism rate increases at sites in the genome where error-prone Pol $\alpha$ tract retention is expected to occur: immediately downstream of Okazaki fragment junctions. I focus upon sites where Okazaki junctions most frequently align, termed “super-peaks”. Super-peaks were defined as described in 5.1.1, whereby all peaks throughout the genome above a defined junction-to-neighbouring-read-coverage threshold (3.47%) were selected ($n = 96,951$). Peaks are either “uni-directional”, if a defined peak is only present on one strand ($n = 94,538$; Figure 5.1a), or “bi-directional”, if defined peaks are present on both the forward and reverse strands, within a twenty nucleotide span ($n = 2,413$; Figure 5.1b).

Uni- and bi-directional peaks were correlated with their strand bias ratios. Ratios represent the division of reads between the forward and reverse strands at each site, indicative of the frequency of Okazaki fragment synthesis in each genomic direction (Figure 5.1c). Distribution of strand bias ratios suggests that a peak is either uni- or bi-directional as a consequence of the quantity of Okazaki synthesis on each strand: sites where the strand ratio bias is highest, where synthesis is most likely to be on one strand compared to the other (all sites where abs(log(ratio)) > 1.84; $n = 57,477$) are uni-directional, whereas sites with a lower strand bias ratio, where Okazaki
Figure 5.1: **Uni- and bi-directional super-peak distributions.** [a] super-peaks (blue vertical line), are categorised as uni-directional if a peak surpasses the Okazaki junction rate threshold of 3.47% (grey shading), only on the forward strand (orange), or only on the reverse strand (blue), within a 20 nucleotide exclusion zone (vertical black lines). [b] super-peaks are categorised as bi-directional if peaks surpassing 3.47% significance are present on both the forward and reverse strand within 20 nucleotides. Both examples are super-peaks from chromosome 10. [c] strand ratios are calculated by dividing the number of Okazaki fragments on the forward strand by the number of Okazaki fragments on the reverse strand. The absolute log value is plotted for uni-directional and bi-directional super-peaks. Bean plots visualise the distribution of strand ratios (curved volume) as well as individual data points (thin horizontal bars). The dotted horizontal line represents the mean value of the whole data set, while the mean of each category is shown as a thick black line. As the maximum sample input for each category is limited to 5000 values, a random sample has been extracted from the uni-directional super-peaks category for visualisation.

Fragment synthesis will be occurring on both strands more equally, will be more likely to be bi-directional. In other words, barriers to lagging strand replication are efficient barriers, no matter which strand is replicated as the lagging strand, albeit the exact site of the Okazaki junctions enrichment can be offset by up to 20 or 30 nucleotides between strands. Our ability to reliably detect the peak on both strands is dependent on read coverage of the strand that is less often the lagging strand.
5.2.2 Polarisation by direction of replication creates conservation asymmetry

For subsequent analyses, all super-peaks are orientated by their consensus direction of Okazaki synthesis, using each peak’s strand ratio bias. Additional annotations for each peak region are rotated accordingly, so that in all instances the consensus direction of Okazaki synthesis is from left to right, and aligned to super-peaks (Figure 5.2). In this orientation, any Pol α tract retention, if present, would be preferentially located immediately to the right of aligned super-peaks.

I observe a peak in polymorphism rate immediately to the right of aligned super-peaks, spanning approximately four nucleotides, exactly where error-prone Pol α tract retention is predicted to occur. The peak in polymorphism rate clearly surpasses the 95% confidence interval created by permutations of randomly sampled nucleotide values.

Additional annotations illustrate well defined nucleosome phasing, both at super-peaks, and at further approximate 180 nucleotide intervals; a pattern subtly followed by DNase I hypersensitivity. H2A.Z containing nucleosomes are mildly depleted at super-peaks, with an approximate but poorly defined inverse correlation to nucleosome occupancy as a whole. Forward and reverse strand transcripts show a slight asymmetry around super-peaks, indicating a moderate increase in both transcripts upstream of junctions.

5.2.3 De novo identification of underlying super-peak motifs

Super-peaks are identified independently of known binding interactions. However, part of our hypothesis implies a role for DNA-bound factors in the displacement of Pol δ and subsequent Okazaki junction formation. We therefore are interested to see if specific binding interactions produce Okazaki junction super-peaks. I run novel motif discovery analysis on sequences underlying super-peaks to investigate whether specific factors may repeatedly be associated with them.

Five significant motifs were identified within super-peak sequences. Of these, two motifs can be found as prominent spikes: “dreme1” immediately downstream of super-peak junctions, and “dreme4” directly at junctions (Figure 5.3). “dreme2”, “dreme3” and “dreme5” are found at a frequency and motif strength passing significance, but do not form defined spikes; “dreme2” and “dreme5” displaying a mild depletion directly at junctions, compared to neighbouring nucleotides.
Figure 5.2: **Super-peaks aligned and orientated by consensus direction of Okazaki synthesis.** The most extreme nucleotide positions of Okazaki junctions across the genome, termed “super-peaks”, are aligned, and orientated so that the consensus direction of synthesis is from left to right. Pol α tract retention would be expected immediately downstream of aligned super-peaks (see schematic, top left). DNase I hypersensitivity (grey shading, top left) corresponds to Okazaki junction rate (blue points, top left). Polymorphism rate (black line, bottom left) is plotted with its 95% confidence interval of permuted values (grey). In the expanded view (right), DNase I hypersensitivity, nucleosomes and H2A.Z containing nucleosomes are plotted individually. Forward (red) and reverse (blue) strand transcripts combined.
Figure 5.3: De novo motifs at super-peak junctions. Sequences surrounding super-peaks were input into the de novo motif discovery software DREME [3]. Five motifs were identified as significantly enriched at super-peaks. [a] each motif, in descending significance, named dreme 1 to 5; their p values (green), frequencies at super-peaks (purple), and bit-weighted motifs. [b] localisation of dreme motifs at super-peaks. Regions are orientated so that the consensus direction of Okazaki synthesis is from left to right. Dreme matches represent frequency of mid-point coordinates, for each dreme motif.

5.2.4 Assigning binding motifs to known transcription factors

Five significant motifs underlie super-peak junctions. To assign these motifs to binding motifs of known transcription factors I entered their PWMs into Japsar’s custom matrix motif alignment search tool (Figure 5.4).

Without ChIP-seq data for all factors of interest it is not possible to assign with confidence the binding of specific factors to super-peaks. Potential factor binding can only be inferred by the motifs present. In some instances, such as with dreme1 and dreme2, certain factors stand out as significantly enriched Sfp1 and Reb1, respectively). In other instances, as with dreme3, dreme4 and dreme5, no one factor represents a strong candidate.
5.2. SUPER-PEAKS: RESULTS

5.2.5 Comparison of annotations between super-peak subcategories

When all super-peaks are aligned and orientated by their consensus direction of Okazaki synthesis we see a clear asymmetrical increase in substitution rate immediately downstream of Okazaki junctions, where error-prone Pol α tract retention is predicted to occur. To investigate which factors bind DNA at these sites I identify sequence motifs with statistical significance within super-peak regions. Super-peaks were divided into six subcategories: five for those super-peaks containing one of the five novel motifs, and a sixth category for those super-peaks containing none of the five motifs. For each of these six subcategories all super-peaks within them were aligned and orientated. Additional annotations were plotted accordingly, for those genomic locations assigned to super-peaks within the corresponding subcategory only (Figure 7.4).
CHAPTER 5. OKAZAKI JUNCTION SUPER-PEAKS

Figure 5.5: (one of three..)
5.2. SUPER-PEAKS: RESULTS

Figure 5.5: (..two of three..)
Figure 5.5: (three of three.)
5.2. SUPER-PEAKS: RESULTS

Figure 5.5: Motif-divided subcategories of super-peaks aligned and orientated by consensus direction of Okazaki synthesis. [a:e] subcategories of super-peaks containing de-novo motifs dreme1 to 5, respectively. [f] super-peaks containing non of the five dreme motifs. In all plots super-peaks belonging to each subcategory are aligned, and orientated so that their consensus direction of Okazaki synthesis is from left to right. Plots contain DNase I hypersensitivity (grey shading, top left), Okazaki junction rate (blue points, top left), polymorphism rate (black line, bottom left), the 95% confidence interval of permuted polymorphism rate (grey lines, bottom left), an expanded view of DNase I hypersensitivity (top right), bound nucleosome dyad positions (second from top, right), bound H2A.Z containing nucleosome dyad positions (second from bottom, right), and forward (red) and reverse (blue) strand transcripts (bottom right).

I investigate whether formation of super-peaks by individual binding interactions results in differences in annotation:

- DNase I hypersensitivity profiles remain peaked at super-peaks in all six subcategories. This suggests that DNase I hypersensitivity is a common feature of heterogeneous DNA replication barriers. Taking into account increased noise from reduced sample size, most evident for dreme3 super-peaks, a subtle variation is evident between the shape of DHS peaks within subcategories. For example, between dreme1 and dreme2 subcategories the DHS profiles have a defined peak, and are broader, respectively.

- In nearly all subcategories a clear asymmetrical increase in substitution rate is visible immediately downstream of Okazaki junctions, where error-prone Pol α tract retention is predicted to occur. The profile of these peaks is noisier than when plotted for all super-peaks combined, due to reduced sample size, but still evident. Defined peak patterns in polymorphism rate are however distinguishable between super-peak subcategories. For example, although polymorphism rate peaks immediately downstream of aligned super-peak junctions in both dreme1 and dreme4 subcategories, the peak is broad in the former, and acute in the latter. It could also be inferred that two additional peaks in polymorphism rate flank that at the super-peak approximately 20 nucleotides up and downstream in the dreme2 subcategory, a pattern mildly visible also at dreme1 super-peaks. At dreme3 and dreme5 super-peaks the spike in polymorphism rate is almost entirely diminished.

- In the expanded view, DNase I hypersensitivity is similar between subcategories. What differs is the exaggerated phasing of DHS at dreme3 super-peaks, compared to all other subcategories, as well as the higher percentage coverage of DHS at dreme2 super-peaks.

- The location of DNA-bound nucleosomes at super-peaks displays notable variation between super-peak subcategories. At non-dreme super-peaks nucleosomes are clearly phased, at ap-
approximately 180 nucleotide intervals, both at super-peaks and up and downstream. dreme3, dreme 4 and dreme5 super-peaks display a similar pattern, although less distinguished. dreme2 super-peaks display a weakened pattern of phasing, arguably with reduced phasing at the super-peak itself, while in dreme1 super-peaks nucleosome phasing is completely ablated, both at super-peaks, and up and downstream.

- Although binding of H2A.Z containing nucleosomes exhibited no clear pattern when all super-peaks were combined, a modest relationship between bound H2A.Z containing nucleosomes and transcripts is visible when super-peaks are divided into subcategories. This relationship is either direct, in the case of dreme1, dreme2 and dreme5, or inverse, in the case of dreme2 and dreme4.

### 5.3 Super-peaks: a discussion

#### 5.3.1 Super-peaks as a model for the Pol α tract retention hypothesis

We hypothesise that retention of error-prone Pol α tracts during lagging strand replication has the potential to increase substitution rate and shape polymorphism rate across the genome. We test this hypothesis in the yeast *S.cerevisiae* genome due to the availability of lagging strand Okazaki fragment read alignment data. We approach the hypothesis by identifying sites across the genome with the highest rate of Okazaki junctions, as we predict these “super-peaks” to be the best candidates for observing substitution rate effects produced by such Pol α tract retention.

A strand bias ratio is calculated for each super-peak, by dividing forward and reverse strand Okazaki coverage. Strand bias ratios therefore act as indicators both for consensus direction of Okazaki synthesis, (the strand with the highest coverage), and also strength of the consensus direction of synthesis, (the higher the ratio, the higher the confidence).

In regards to calling the strength of the consensus direction of Okazaki synthesis, success of strand bias ratios is supported by the breakdown of super-peaks into both uni- and bi-directional categories: sites with lower strand bias ratios are more likely to produce super-peaks on both the forward and reverse strand, compared to sites with high strand bias ratios which are more likely to contain a super-peak on one strand only. Therefore, barriers themselves do not appear to be directional.

Super-peaks are aligned and rotated according to their consensus direction of Okazaki synthesis,
as determined by their strand bias ratio. A clear asymmetrical peak in polymorphism rate is observed immediately downstream of super-peak junctions, where Pol $\alpha$ tract retention is predicted to occur. This observation supports both the use of super-peaks as targets for identifying Pol $\alpha$ tract retention and subsequent increased substitution rate, as well as the use of strand bias ratios in determining the correct consensus direction of Okazaki synthesis. At all sites, however, it must be considered that a proportion of Okazaki synthesis will still occur in the non-consensus direction: the lower the strand bias ratio, the higher the proportion. All plots consequently contain a degree of mixed signal.

5.3.2 Localisation of sequence motifs at super-peak sites

Okazaki super-peaks were identified directly from Okazaki junction rates without a prerequisite for DNA binding interactions. However, our hypothesis implies a role for DNA bound factors in the displacement of Pol $\delta$ and subsequent Okazaki junction formation. In support of this prediction, DNase I hypersensitivity coverage indicates increased binding at super-peak sites and notably this is consistently enriched downstream of the Okazaki junction peak. This suggests a simple model where protein binding to DNA presents a road-block to Pol $\delta$, promoting polymerase displacement as it approaches the barrier. The peak of elevated substitution rate is consistently located between the Okazaki junction peak and the peak of maximal DNase I hypersensitivity. We are interested in whether specific factors are involved in displacement of Pol $\delta$ during Okazaki fragment synthesis, and what these factors are.

By running de novo motif discovery analysis I identify five motifs at a significant frequency at these sites. When dividing all super-peaks by the presence or absence of such motifs, however, it is apparent that the large majority are void of such enriched sequences ($n = 90,975; 93.84\%$). A smaller percentage are enriched, with dreme4 the most common motif within super-peaks ($n = 2,237; 2.31\%$), followed by dreme1 ($n = 2,171; 2.24\%$). This indicates that while there is clearly a subset of factors that represent common polymerase barriers (e.g. Reb1:dreme2), there may be either a) a much larger, heterogeneous population of binding factors that are individually too rare to manifest a significant motif enrichment, and/or b) a significant as yet unidentified polymerase dissociating factor binding in a non-sequence specific manner.

The first, most significant motif, “dreme1”, localises a few nucleotides downstream of Okazaki super-peaks, at high frequency. Comparison of the motif to known transcription factor binding motifs highlights the transcription factor Sfp1 as a potential candidate. Visual comparison of the novel motif confirms Sfp1 as a strong match (taking into account possible truncation of the novel motif by the software’s poor handling of palindromes). Sfp1 is known to interact with
the transcription factor Rap1 [37], one of the three transcription factors shown by Smith and Whitehouse to associate with Okazaki termini [108], further supporting Sfp1’s legitimacy as a candidate for super-peak formation. That Sfp1 binds DNA both directly in a sequence specific manner and indirectly through other factors such as Rap1 causes acknowledged difficulties in defining SPF1 binding sites experimentally.

The second motif, “dreme2”, is found recurrently at statistical significance but at reduced frequency, and with reduced peak definition. It co-localises at high frequency with the binding motif of transcription factor Reb1. Reb1 was another of the three transcription factors shown by Smith and Whitehouse to associate with Okazaki termini [108]. These findings convincingly support Smith and Whitehouse’s results, and confirm Reb1’s involvement in super-peak junction formation.

“dreme4” localises as a substantial peak directly at super-peak junctions. The highest matching known transcription factor motif is that of DAL81. It’s alignment score is high (94.81%), but comparison by eye is not convincing, and no known DAL81 interactions implicate its involvement. I am unable to suggest a potential transcription factor candidate for this motif with confidence at this time.

Motif assignment to known transcription factors is limited. Binding interaction data such as ChIP-seq is not available for all factors. Potential assignments can therefore only be speculated. The motifs we have identified may belong to a separate entity, separate to the known transcription factor binding motifs available to us. I did not detect significant enrichment of Rap1 or Abf1 binding site motifs at the Okazaki junction super-peaks, although the results of Smith and Whitehouse [108] and my own work in previous chapters show that these specific factors are associated with both local peaks in Okazaki junction and nucleotide substitution rates. I also find a large population of Okazaki junction super-peak sites that do not appear to harbour any of the identified novel motifs. This argues there may be a large population of heterogeneous factors that can elicit Pol δ disassociation and locally elevated mutation rate, but the frequency that they bind the genome is not sufficient to reveal a significant motif enrichment in our bulk analysis.

Although the analysis performed in this chapter has rediscovered Reb1 as a major barrier to Pol δ processivity, the analysis has several important distinctions from that performed by Smith and Whitehouse [108]. Smith and Whitehouse’s work considered the enrichment of Okazaki fragment ends over known transcription factor binding motifs. From that, they highlighted only Reb1, Rap1 and Abf1 (a paralogue of Reb1) amongst transcription factors as barriers to Pol δ. In contrast, while I find that Reb1 is one of (but not the most) abundant, strong barriers to Pol δ, there appear to be many heterogeneous barriers. I can reconcile this apparent conflict by considering the
different style of analyses. Inherent in transcription factor binding site prediction by PWMs, as performed by Smith and Whitehouse, is a high rate of false-positive binding site prediction. Their analysis is also limited to transcription factors with known binding sites. Here I have turned the analysis around, using experimentally defined peaks of Okazaki junctions to look for enriched motifs, an approach that is not constrained by a pre-determined reported binding motif and is largely unaffected by false-positive motif based binding site prediction. This has allowed me to identify several highly significant Okazaki junction peak associated motifs, such as that of Reb1, but also a large heterogeneous population of sites. This suggest many factors may promote Pol δ dissociation, but that individually each may occur at a low frequency genome wide. In contrast Reb1 has over 800 distinct, strong and experimentally confirmed binding sites, and non-Reb1 sites may also represent less efficient barriers, as already seen for Rap1 (see Chapter 4).
The yeast *S. cerevisiae* genome has proved a useful model organism for the study of Pol α retention, due to the availability of its genome wide mapped Okazaki fragment reads. Upon investigation of Reb1 and Rap1 sites I see peaks in substitution rate exactly where Pol α tract retention is predicted to occur. To strengthen this evidence I wish to show support of the Pol α tract retention hypothesis in the human genome. The benefit of investigating sites of Okazaki junctions, as opposed to known transcription factor binding sites, is independence from a prerequisite bias towards sites undergoing purifying motif selection, capable of confounding the signal of interest. No Okazaki alignment data is available for the human genome. I therefore turn to DNase I Hypersensitivity Site (DHS) footprints.

### 6.1 Footprints: an introduction

Deoxyribonuclease I (DNase I) is an enzyme in the cell nucleus that cleaves phosphodiester linkages of the DNA backbone, resulting in both single and double stranded DNA breaks. DNase I is only able to access exposed sites of the genome that are unbound by proteins. DNA wrapped around nucleosomes, for example, cannot be accessed by DNase I, as well as sites of transcription factor-bound DNA. The frequency of DNA strand breaks in a DNase I environment, termed DNase I hypersensitivity [32], is therefore a measure of a nucleotide’s protein-bound state. DNase I hypersensitivity is useful not just as a measure of where proteins are or are not bound to DNA, but also of how these binding interactions differ over time and between cell types.
CHAPTER 6. DNASE I FOOTPRINTS

Figure 6.1: **DNase I hypersensitivity footprints.** A close-up of DNase I Hypersensitivity Sites (DHSs; full sites several kb wide), are shown, with DNase I cleavage frequency per nucleotide (grey and black). Gaps of low cleavage frequency, the DHS footprints, are visible within. These footprints are expanded further below (black). Footprints (red bars) are on average less than fifty bp. (Figure adapted from Hesselberth et al. [46]).

The majority of the genome is wrapped around nucleosomes, in a tightly closed chromatin state [107]. Regions of the genome undergoing protein interactions are found in an open chromatin state, so as to facilitate binding. Regions of open chromatin undergoing DNA-protein interactions will therefore be evident as broad regions (several kb), of high DNase I hypersensitivity (Figure 6.1). These peaks of high DNase I cleavage are DNase I Hypersensitivity Sites (DHSs). Within DHSs, specific protein interactions will again inhibit DNase I cleavage. These small DNase I hypersensitivity-free pockets, often less than fifty bp, within DHSs, are DHS footprints.

I have previously identified a total of 116,676 transcription factor binding interactions across the human genome (Chapter 3), from 31 different motif dependent transcription factors. A caveat when using such sites to investigate the Pol α retention hypothesis is that the substitution rate signal of interest is likely to be confounded by the sequence specificity and consequent purifying selection at each binding motif. Super-peaks of Okazaki junctions in the yeast genome are one way of excluding selection bias, by focusing on sites predictive of Pol α tract retention, as opposed to specific binding interactions. DHS footprints present another such means. We predict that at each footprint a protein is bound to DNA. If a protein is bound during lagging strand replication it may act as a partial Pol δ barrier, promote Pol δ dissociation and consequently result in Okazaki junction formation. We therefore predict the alignment of Okazaki junctions immediately upstream of the left-hand edge of such footprints, and subsequent Pol α tract retention immediately downstream of such junctions.

This theory has been tested in the yeast *S.cerevisiae* genome, where the alignment of Okazaki
Figure 6.2: Okazaki junctions and substitution rate at aligned yeast DHS footprints. Okazaki junction rate (blue) and substitution rate (black) are aligned to DHS footprint left-hand edges (grey). Okazaki junctions on the forward strand (dark blue line) and Okazaki junctions on the reverse strand (dark blue points) are separated. Bound proteins are predicted to lie immediately downstream of aligned footprint left-hand edges (top schematic). Pol δ dissociation is predicted to occur when first encountering a bound protein. Pol α tract retention is predicted immediately downstream of Pol δ dissociation. Substitution rate peaks at footprint edges and reduces within the footprint itself, where purifying selection is predicted.

junctons to footprint edges was visualized (Martin Taylor, unpublished; Figure 6.2). Compared to footprint mid-points (data not shown), Okazaki junctions align neatly to aligned left-hand footprint edges. Again substitution rate peaks at Okazaki junctions, coinciding with the location of predicted Pol α tract retention. We use this evidence as a basis to investigate substitution rate at DHS footprint edges in the human genome.
6.2 Footprints: data and methodologies

6.2.1 Calculating DHS and footprint coverage at DHS footprint edges

DNase I hypersensitivity site footprints are provided by Hesselberth et al. [46], and downloaded from the ensemble ENCODE directory (file source: ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011; files dated 07-09-2012). Two files were used: all.footprints.gz, containing all footprints per cell type, and combined.fps.gz, containing adapted footprint boundaries, compensating for cell type specific variation of footprint coordinates. The two files were combined to allocate cell types to each pair of adapted footprint boundaries, (the only coordinates used for all further DHS footprint analysis). For simplicity only left-hand footprint boundaries are used for analysis.

DNase I hypersensitivity sites were downloaded from UCSC hg19 database for 37 of the 41 cell types of interest [53] (Table A.2).

DHS and footprint percentage coverage arrays were saved for 1000 nucleotides either side of each footprint left-hand edge. DHSs and footprints were only marked on their corresponding arrays if the cell type that the DHS or footprint was found in also contained the footprint in question. Coverage was summed and averaged for all footprints.

6.2.2 Calculating between-species substitution rate at DHS footprint edges

GERP scores were downloaded as bigWig files from the UCSC hg19 database [17] (file source: All_hg19_RS.bw; file dated 15-04-2011). An array of GERP scores, 1000 nucleotides either-side of each left-hand edge coordinate, was saved for every DHS footprint. In this way all footprint arrays are analogous for left-hand edge alignment. Array scores were summed and averaged.
Figure 6.3: **Substitution rate at aligned human DHS footprints.** Between species GERP scores (red) are aligned at DHS footprint left-hand edges \((x = 0)\). 95% confidence intervals of permuted GERP values are included (pink). DHS coverage (grey) and footprint coverage (blue) are plotted. Bound proteins are predicted to lie immediately downstream of aligned footprint left-hand edges (top schematic).

### 6.3 Footprints: results

#### 6.3.1 A shoulder of substitution rate at human footprint edges

A total of 8,374,967 DHS footprints have been previously identified across the human genome [84]. All footprints were aligned by their left-hand edges, (the side first encountered by the approaching upstream Pol δ), and associated substitution rates averaged over all sites. Between species Genomic Evolutionary Rate Profiling (GERP) constraint scores [17], were plotted (Figure 6.3). GERP scores spike at footprint left-hand edges, as predicted by factor-binding Okazaki junction formation and subsequent alignment of error-prone Pol α tract retention.
It is interesting to note a dearth of footprint coverage upstream of aligned edges, potentially as a result of steric hindrance. It is at this dearth that GERP scores spike. GERP scores remain below the genomic background average at these sites. These results neither rule out Pol $\alpha$-induced elevated mutation rate or reduced purifying selection by steric hindrance, but instead identify the potential for a mutual effect between the two at corresponding genomic sites.

### 6.4 Footprints: a discussion

#### 6.4.1 Footprints as a model for the Pol $\alpha$ tract retention hypothesis

To validate or disprove the Pol $\alpha$ hypothesis within the human genome I focus my attention upon DHS footprints. I hypothesise these footprints to be a means of identifying the most probable sites of aligned Pol $\alpha$ tract retention, devoid of pre-requisite confounding features of binding motif selectional constraint. Compared to Okazaki junction super-peaks, however, where junctions themselves are aligned, the exact location of junctions in relation to footprints can only be inferred. Interaction between protein structures external to DNA binding interactions could mean Pol $\delta$ is prone to dissociation prior to reaching footprint boundaries (the “barrier hypothesis”). In addition, footprint coordinates have been combined over cell lines. The use of aligned footprint edges as a proxy for aligned Okazaki junctions is however well supported by Okazaki junction and footprint edge correlations in yeast *S.cerevisiae* (Figure 6.2).

At aligned footprints in the human genome GERP displays a dip in substitution within footprints, consistent with their selective constraint. In addition it significantly peaks directly upstream of footprint left hand edges, similar to peaks in yeast, and consistent with predicted sites of Pol $\alpha$ retention. I therefore present DHS footprints as a heterogeneous binding site model for the investigation of the lagging strand hypothesis.

We are not able to draw conclusions about the evolutionary mechanism shaping such peaks. Taken alone they could indicate either elevated mutation rate or a lack of purifying selection, or contributions from the two. Consequently, while identification of these sites has been successful, the challenge of explaining them has yet to be achieved.
6.4. FOOTPRINTS: A DISCUSSION

6.4.2 Considering evolutionary forces at shoulders

An absence of footprint coverage is visible immediately upstream of aligned footprint left-hand edges, for a stretch of approximately ten nucleotides. A simple explanation is steric hindrance. A spike in GERP at this site would correspond well with such an observation. Another consideration for the dearth of coverage is the footprint definition procedure. Footprints were identified as short regions of reduced DNase I hypersensitivity, flanked by DNase I hypersensitivity passing a threshold of significance [46]. No two footprints could therefore be identified as immediately neighbouring, with this criteria being met. Adjacent footprints will therefore be rejected or combined. Substitution rate as a consequence of selection at such sites would not be effected under these circumstances.

If error-prone Pol α tract retention is occurring at footprints, and this process results in elevated mutation rate, the difficulty lies in identifying the exact signal. Without Okazaki alignment data we cannot align direct sites of significant Okazaki junction rate, and therefore must infer association between footprint edges and Okazaki termination. Due to this fact Okazaki junctions, and subsequent Pol α tract retention, may be too dispersed and spread out within the vicinity of footprint edges for their effects to be distinctively observed. Although in the yeast genome a clear signal of correlation was observed, additional evolutionary contributions in the human genome, such as CpG deamination, or different factor binding interactions, could further distort the Pol α tract signal in the human genome.

When all footprints are aligned GERP peaks at an approximate score of 0; the proposed value of genomic background substitution rate. The fact that GERP scores do not extend beyond 0, and are significantly below the regional background rate calculated by permuted values, means there is no direct evidence of positive selection or elevated mutation rate at these sites. When transcription factor binding sites were aligned, however, and their substitution rate plotted, GERP scores frequently extended well beyond zero, ruling out purifying selection as the only contributor to the pattern observed.
In principal, the shoulders of elevated substitution rate that I have found proximal to some transcription factor binding sites and DNase I footprint edges can be explained by either selection or locally elevated mutation rate. The mechanistic model I propose provides support for and argues that mutation rate is primarily responsible. I therefore predict that binding sites occupied in germline cells exhibit the mutation shoulder effect but that binding sites only occupied in a restricted set of non-germline cells only show the influence of selection and not the mutational shoulder (I illustrate the underlying theory below). In this chapter I test this prediction, accepting the limitation that we currently have very few measures of protein-DNA interaction for the cell types of the human germline. Building on the observations of previous chapters I use DNase I footprint edges that have been consistently measured for a diversity of cell types, including differentiated somatic cells and embryonic stem cells; the latter representing one of the cell types of the germline.

### 7.1 Mutation versus selection: an introduction

Selection is capable of exerting an observable evolutionary effect regardless of the cell type in which it acts. A mutation rate effect, on the other hand, will only be seen throughout evolution if it occurs within the germline (Figure 7.1). For example, if an advantageous or deleterious substitution occurs at a heart-specific binding site within a heart cell, this mutation may confer a selective advantage or disadvantage, respectively. Such a substitution will consequently effect the
Figure 7.1: **Evolutionary consequences of cell type-specific selection and mutation.** A schematic to illustrate how elevated mutation rate and selection will effect DNA differently over evolutionary time when the selectional forces are the same but the cell type in which they act differ. Each of the two circles on the left of the diagram represent the same stretch of DNA, in two different cell types. The top circle represents this stretch of DNA in a heart cell. In this fictitious heart cell a factor binds on the left. Consequently the left-hand binding site undergoes purifying selection (green), based on its cell specific function, but also elevated mutation rate (red). The bottom circle represents this same stretch of DNA in a gamete. In this fictitious gamete a factor binds on the right. Consequently the right-hand binding site undergoes purifying selection (green), based on germ cell specific functions, and also elevated mutation rate (red). The bottom left plot represents how these signals will combine within a population over evolutionary time (dark brown). Selection will act at both binding sites and consequently both sites of constraint will be observed within the population. Elevated mutation rate will only be passed on to future generations if it occurs within the germline, therefore a peak in substitution rate will be seen at the right-hand binding site but not the left.

fitness of the organism, and so effect its frequency within a population. In contrast, if a factor only binds DNA in cells of the heart, and as a result of such an interaction this binding site undergoes elevated mutation rate within heart cells only, these mutations would not be passed on to future generations. If however a factor binds DNA in the germline, and consequently the binding site undergoes elevated mutation rate, these mutations can be passed on to future generations.
7.2 Mutation versus selection: data and methodologies

7.2.1 Splitting footprints into cell type categories

DNase I hypersensitivity footprints were defined as previously described (6.2.1). Five cell type categories, (immortalized, malignant, multipotent, normal and pluripotent), are described by Thurman et al. for 97 cell types [117]. 41 of these categorised cell types match the cell types of the DHS footprint data (Table 7.1). Each cell type is further assigned a cell type ‘genre’ (somatic, germline-like, or undefined). Each individual footprint is assigned one or more cell type categories and genres, according to the categories and genres of the cell type/s it is present within.

Footprints are divided into four footprint subsets for further analysis, according to the cell type categories and genres within which they are present:

- **Strong somatic**: only footprints in > 20 somatic cell lines (i.e. those footprints present within at least twenty somatic-like cell types and absent from all other germlike-like or undefined cell types)

- **Weak somatic**: only footprints in > 10 somatic cell lines (i.e. those footprints present within at least ten somatic-like cell types and absent from all other germlike-like or undefined cell types)

- **Weak germline**: footprints only in pluripotent cell lines (i.e. those footprints present within a pluripotent cell line and absent from all other cell type categories)

- **Strong germline**: only footprints in all five cell type categories (i.e. those footprints present within immortalized, malignant, multipotent, normal and pluripotent cell lines)

7.2.2 Calculation of additional annotations

- GERP scores were calculated at each individual footprint site as previously described in 6.2.2.

- 1000 Genomes data was used to derive SNPs across the genome, representational of polymorphism rate [80]. Files were downloaded from the 1000 genomes database, release 20110521 (file source: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521; files dated 10-10-2012). All SNPs within 1000 nucleotides of a DHS footprint left edge were saved, with the coordinate, reference allele, alternative allele and ancestral allele recorded.
Table 7.1: **List of footprint cell types, their corresponding categories and genres.** Annotations provided by Thurman et al. [117]

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>Cell Type Category</th>
<th>Cell Type Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG10803</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>AoAF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>CD20+</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>fBrain</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>fHeart</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>fLung</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HAEpiC</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HA-h</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HCF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HCM</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HCPEpiC</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HEEpiC</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HFF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HIP_epiC</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HMF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HMVEC-dBl-Ad</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HMVEC-dBl-Neo</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HMVEC-dLy-Neo</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HMVEC-LLy</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HPAF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HPdLF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HPF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HRCEpiC</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HSMM</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>HVMF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>IMR90</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>NH-A</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>NHDF-Ad</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>NHDF-neo</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>NHLF</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>SAEC</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>SKMC</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>Th1</td>
<td>Normal</td>
<td>Somatic</td>
</tr>
<tr>
<td>GM12865</td>
<td>Immortalized</td>
<td>Somatic</td>
</tr>
<tr>
<td>GM06990</td>
<td>Immortalized</td>
<td>Somatic</td>
</tr>
<tr>
<td>H7-hESC</td>
<td>Pluripotent</td>
<td>Germline-like</td>
</tr>
<tr>
<td>CD34+_Mobilized</td>
<td>Multipotent</td>
<td>Un-defined</td>
</tr>
<tr>
<td>SK-N-SH_RA</td>
<td>Malignant</td>
<td>Un-defined</td>
</tr>
<tr>
<td>NB4</td>
<td>Malignant</td>
<td>Un-defined</td>
</tr>
<tr>
<td>K562</td>
<td>Malignant</td>
<td>Un-defined</td>
</tr>
<tr>
<td>HepG2</td>
<td>Malignant</td>
<td>Un-defined</td>
</tr>
</tbody>
</table>
• C→T, G→A transversion and non-C→T, G→A polymorphisms were calculated from 1000 genomes data as above.

• CpG frequency was calculated by downloading nucleotide sequences as FASTA files, per chromosome, from the UCSC hg19 database (files dated 20-03-2009). For 1000 nucleotides either side of each footprint left-hand edge, both C and G nucleotides of each CpG dinucleotides were scored as 1 on a binary array.

• CpG methylation was downloaded from the ENCODE database as Reduced Representation Bisulfite Sequence [81]. For 1000 nucleotides either side of each footprint left edge, if a methylated CpG dinucleotide was present, the percentage methylation, (as provided by ENCODE), was saved at the position of both C and G nucleotides. Percentage methylation was only scored if the cell type the methylated CpG dinucleotide is found in also contained the footprint in question.

• DHS and footprint coverage were calculated at each individual footprint site as previously described in 6.2.1.

Values were divided into their corresponding footprint subsets before being summed and averaged for each.

### 7.3 Mutation versus selection: results

#### 7.3.1 Increased substitution rate at germline-associated footprints

I wished to investigate the presence or absence of CpG-independent elevated mutation within the vicinity of DHS footprints. Substitution rate is the combined effect of both selection and germline-specific mutation rate over evolutionary time. One way of distinguishing selection from mutation is by looking at substitution rate at sites only found in somatic cell lines, where mutational effect will be absent, and comparing them to substitution rate at sites only found in the germline, where mutation rate will be included.

As a proxy for somatic-only footprints, I filter for those footprints present in at least twenty of the thirty five somatic cell lines, which aren’t present in any other cell type category (‘strong somatic’; \( n = 96,198 \)), and compare them to those footprints present in at least ten of the thirty five somatic cell lines, which also aren’t present in any other cell type category (‘weak somatic’; \( n = 425,413 \)) (Figure 7.2 a:b). I would predict footprints of the former category to be assigned as somatic-
only with higher confidence, and so would predict their profile to be a greater representation of somatic-like substitution rate than the profile of the latter.

Very limited footprint data is available for germline cell types. As a proxy for germline-only footprints, I filter for those footprints only present in the pluripotent cell line (‘weak germline’; \( n = 514,872 \)), and compare them to those footprints present in all five cell type categories (‘strong germline’; \( n = 178,842 \)) (Figure 7.2 c:d). I predict footprints of the latter category to be essentially “housekeeping” footprints. If they are present within normal, immortalised, malignant, multipotent and pluripotent cell lines, I predict with high probability, (but by no means with absolute certainty), that they will also be present within the germline. I therefore predict footprints of the latter category to be assigned as germline-only with higher confidence, and so would predict their profile to be a greater representation of germline-like substitution rate than the profile of the former.

Footprint coordinates were adjusted to account for variations between footprints from different cell lines. Consequently footprint coverage profiles vary depending on the number of cell lines in each subset. Footprint coordinates are only available for one pluripotent cell line. Consequently footprint coverage at pluripotent-only footprints (weak germline), provides the clearest profile of footprint coverage. At these sites a clear increase of coverage at aligned footprints is observable, as well as a dearth immediately upstream of their left-hand edges. In other subsets, footprint coordinates are adjusted across many cell lines. This is evident by reduced clarity of footprint coverage. However in all subsets coverage is highest at aligned footprints and depleted immediately upstream of their left-hand edges. With the exception of footprints found in all five cell type categories (strong germline), polymorphism rate is erratic and difficult to convincingly separate from the local background estimate (Figure 7.2 a:c). However, clear patterns are seen for all GERP score profiles.

GERP constraint scores demonstrate equal profiles in both somatic-like categories, peaking immediately upstream of aligned footprint left-hand edges, (at the site of depleted footprint coverage), and troughing most significantly within the binding region of aligned footprints, (at the site of highest footprint coverage). This is an indication of purifying selection within footprint bound sites, and reduced purifying selection within sites of probable steric hindrance (the height of which is more pronounced for strong somatic footprints). It is interesting to note that while the footprint profile of GERP is lower on the y-axis for strong somatic footprints (indicating reduced overall substitution; Figure 7.2 a), compared to weak somatic footprints (Figure 7.2 b), so too is the visible neighbouring rate for the former compared to the latter (~ 0.4 and ~ 0.3 respectively). To further explore these patterns, in the next section, I extend my analyses to broader windows around footprint edges.
7.3. Mutation versus Selection: Results

![Graphs](image)

Figure 7.2: (one of one.)
CHAPTER 7. MUTATION VERSUS SELECTION: SOMA AND THE GERMLINE

Figure 7.2: **Substitution rate at somatic-like and germline-like subsets of human footprints.** DNase I Hypersensitivity Coverage (grey), footprint coverage (blue), non-CpG-associated human polymorphism rate (black) and between species GERP constraint scores (red) are plotted across footprint sites divided into four subsets, according to the cell types within which they are found: [a] footprints found in at least twenty of the thirty five somatic cell lines, and not in any other cell type category (‘strong somatic’; the most confidently called somatic-like footprints), [b] footprints found in at least ten of the thirty five somatic cell lines, and not in any other cell type category (‘weak somatic’; the less confidently called somatic-like footprints), [c] footprints only found in the pluripotent cell line, and not in other cell type categories (‘weak germline’; the less confidently called germline-like footprints) and [d] only those footprints found in all of the five cell type categories, implicated as “housekeeping” footprints (‘strong germline’; the most confidently called germline-like footprints). 95% confidence intervals for non-CpG-associated polymorphism rate (grey) and between species GERP constraint scores (pink) are calculated from permuted values across all footprints combined. Y axes for substitution rates are equal throughout a to d.

Between each germline-like category, the profile of GERP constraint scores varies dramatically. For weak germline footprints (Figure 7.2 c), GERP matches that of somatic-like footprints, albeit with a further reduction in the spike of substitution at depleted footprint coverage. At strong germline footprints (Figure 7.2 d), a different profile is observed: from approximately fifteen nucleotides either side of aligned footprint left-hand edges substitution rate increases, until peaking immediately upstream of aligned footprint left-hand edges, at a rate substantially in excess of the genome-wide neutral estimate and local background substitution rate. As with weak germline footprints, the observable neighbouring substitution rate is elevated compared to somatic-like footprints (GERP \( \sim 0 \)). Non-CpG-associated polymorphism rate at strong germline footprints approximately follows that of GERP (Figure 7.2 d). Its profile peaks immediately upstream of aligned footprint left-hand edges, at the same nucleotide as the peak in GERP.

### 7.3.2 Regional effects of substitution between cell type genres

To further understand the broader context of differences in substitution rates in the 50 nt flanking DHS footprints (Figure 7.2), I consider a wider window encompassing 1 kb in either direction of aligned footprint left-hand edges (Figure 7.3 a:d).

GERP scores mirror DHS coverage, in both somatic-like subsets of binding sites and the weak germline-like subset (Figure 7.3 a:c). Substitution rate is lowest at strong somatic footprints, and
7.3. MUTATION VERSUS SELECTION: RESULTS

![Graphs showing mutation versus selection results.]

Figure 7.3: (one of one.)
Figure 7.3: **Footprint subset substitution rates within an extended genomic window.**

DNase I Hypersensitivity Coverage (grey), footprint coverage (blue), non-CpG-associated human polymorphism rate (black) and between species GERP constraint scores (red) are plotted across footprint sites divided into four subsets, according to the cell types within which they are found: [a] footprints found in at least twenty of the thirty five somatic cell lines, and not in any other cell type category (strong somatic), [b] footprints found in at least ten of the thirty five somatic cell lines, and not in any other cell type category (weak somatic), [c] footprints only found in the pluripotent cell line and not in any other cell type category (weak germline) and [d] only those footprints found in all five cell type categories (strong germline). 95% confidence intervals for non-CpG-associated polymorphism rate (grey) and between species GERP constraint scores (pink) are calculated from permuted values across all footprints combined. Y axes for substitution rates are equal throughout a to d.

nearly completely absent in the weak germline subset. This implies the profile is a consequence of purifying selection within DHSs. The GERP profile at strong germline footprints, however, is distinct (Figure 7.3 d); it is not an inversion of DHS coverage, but instead intricately peaks and troughs.

In the first three footprint subsets, non-CpG-associated polymorphism is approximately uniform, with a marginal dip mirroring that of DHS coverage (Figure 7.3 a:c). The fourth subset, strong germline, again displays a distinct profile (Figure 7.3 d): Non-CpG-associated polymorphism rate increases significantly over sites of greatest footprint coverage, to peak at the site of aligned footprints.

When extending the genomic window to one thousand nucleotides, I observe a return of all four GERP profiles to the same approximate rate (~ -0.1). With the exception of strong germline footprints, non-CpG-associated polymorphism also returns to a uniform rate at the extended window’s outer edges (~ 0.5).

### 7.3.3 Germline-specific correlation between CpG dinucleotide frequency and footprint coverage

I observe distinct patterns in non-CpG-associated polymorphism rate between somatic-like footprints and strong germline-like footprints. By considering only non-CpG associated polymorphisms, the elevated polymorphism rate observed (Figure 7.3 d) should not be a consequence of methyl-CpG deamination mutation. However, to more generally understand the substitution rate...
patterns at these sites I investigate CpG dinucleotide frequency, CpG-associated polymorphisms and CpG-independent polymorphisms, to determine whether these patterns too differ between footprint subsets (Figure 7.4 a-h).

CpG dinucleotides are locally depleted adjacent to DHS footprint edges relative to proximal sequences both outside and inside the footprint site (Figure 7.4). The fraction of methylated CpGs in this footprint edge region is also depleted relative to the flanking 10 to 20 nucleotides. Both of these patterns hold true for all four cell-type categories, but are most pronounced for strong germline footprints (Figure 7.4 d). Consistent with these observations, CpG-associated polymorphism rate is locally depleted in these footprint edge regions.

Combined polymorphism rate drops over aligned footprint left-hand edges at the single nucleotide level. Differentiating between CpG-associated and non-CpG-associated polymorphisms reveals this drop to be the consequence of two distinguishable profiles. 1) a drop in CpG-associated polymorphism rate immediately upstream of aligned footprint left-hand edges, and 2) a drop in non-CpG-associated polymorphism rate immediately downstream of aligned footprint left-hand edges. By dividing footprints into somatic-like and germline-like subsets I observe a correlation between somatic-like footprint confidence and the degree of these two profiles: The drop in CpG-associated polymorphisms is least distinguishable in strong germline footprints and the non-CpG-associated drop is no longer apparent. Instead a peak in non-CpG-associated polymorphisms is observable immediately upstream of aligned footprint left-hand edges in this subset, which is not visible at somatic footprints.

Over the wider genomic window, non-CpG-associated polymorphism rate is indistinguishable from CpG-associated polymorphism rate, and consequently neither are distinguishable from the rate of all polymorphisms combined. At weak somatic and weak germline footprints these profiles are approximately uniform across the region. It is only strong somatic footprints that display a marginal dip in polymorphism rate within the broader region of aligned sites. At strong germline footprints this profile is inverted and amplified.

Methylated CpG values do not cover the whole genome but are enriched for high CpG sites, such as promoters. Y axes are therefore not comparable in this instance due to regional bias, and profiles cannot be taken as absolute. What can be inferred, is that for the majority of DHSs, where methylation data is available, at the single nucleotide level a general correlation exists between footprint coverage and CpG methylation. Over the winder genomic window however, a general inverse relationship exists between footprint coverage and CpG methylation. The exception is for strong germline footprints. At these sites methylation levels remain significantly higher than for all other footprint subsets (∼ 0.12 compared to ∼ 0.03, respectively).
CHAPTER 7. MUTATION VERSUS SELECTION: SOMA AND THE GERMLINE

Figure 7.4: (one of two.)
7.3. MUTATION VERSUS SELECTION: RESULTS

Figure 7.4: (..two of two..)
Figure 7.4: The breakdown of polymorphism rate and CpG dinucleotide frequencies between somatic-like and germline-like subsets of human footprints. Close-up [a:ad] and extended [e:h] genomic window views. Overall polymorphism rate (black), non-CpG-associated polymorphisms (blue), CpG-associated polymorphisms (magenta), CpG dinucleotide frequency (purple), methylated CpG dinucleotide rate (green) and footprint coverage (orange) are plotted across footprint sites divided into each of the four subsets: strong somatic ([a], [e]), weak somatic ([b], [f]), weak germline ([c], [g]) and strong germline ([d], [h]). With the exception of methylated CpG rate, where Y axes differ due to considerations for regional bias, Y axes are equal throughout a to h.

CpG frequency profiles are similar for somatic-like footprints and weak germline germline-like footprints, both at the single nucleotide level and over a wider genomic window. At these sites CpG frequency is equal to background level (~ 0.02) for the entire region, with the exception of its profile at aligned footprints where frequency peaks at highest footprint coverage and dips at the adjacent upstream dearth. Again strong germline-like footprints are an exception; at these sites CpG frequency rises gradually across the wider region, in accordance with footprint coverage, eventually peaking at aligned footprint sites. While the peak of CpG rate in the first three subsets of footprints is approximately 0.03, CpG frequency in the latter is four times this value, at approximately 0.12. A clear dip is CpG frequency is still present immediately upstream of aligned footprint left-hand edges at these sites.

7.4 Mutation versus selection: a discussion

7.4.1 Housekeeping bias in elevated substitution rate

“Housekeeping” footprints, present within all cell type categories, are used as a proxy for those footprints predicted with the highest confidence to also be present within the germline. Such sites display a striking between-species and polymorphism based elevation in substitution rate, clearly distinct to somatic-only footprints. Due to the inherent properties of selection and mutation, patterns at strong somatic footprints represent purely the effect of selection, while patterns at strong germline footprints, the “housekeepers”, represent effects of both selection and mutation combined (the fact that footprint subsets are used merely as proxies must be considered). That the most pronounced elevation in substitution is observed at strong germline footprints, for both measures, is as predicted by elevated mutation, in keeping with the ‘lagging strand hypothesis’.
While the elevation in substitution at strong germline footprints is distinct, all footprint subcategories exhibit a modest ‘peak’ in between-species substitution. Such a peak could be due to the inclusion of unidentified germline footprints, and consequently result from elevated mutation, or could be due to reduced constraint through steric hindrance, and consequently result from (a lack of) purifying selection. Such reduced constraint through steric hindrance would be akin to 4D sites in protein coding sequence, providing a local estimate of neutrality, about which nucleotide level changes in substitution can be compared. That substitution profiles directly at footprints remain similar between categories (with the exception of strong germline footprints), while overall constraint of shoulders varies, suggests the possibility of elevated mutation rates not just in shoulder regions but also in sequence specific binding site regions.

The peak of increased substitution rate at strong germline footprints is broader than at somatic-like footprints, suggesting footprint-associated elevated mutation rate occurs over a larger window than that predicted by steric hindrance alone. At aligned footprint edges exact Okazaki junctions are not defined, but are predicted to occur within the vicinity, thereby fitting with such an observation. The pattern of strong germline-like GERP substitution rate consequently fits with a model of elevated mutation rate as a result of error-prone Pol α tract retention, also excepting a potential combined effect of steric hindrance and elevated mutation rate at strong germline-like footprint shoulders.

### 7.4.2 CpG effects are not responsible for substitution shoulders

Methylated CpG dinucleotides are a known contributor towards site-specific elevated mutation rate. I observe that CpG dinucleotides are locally depleted within shoulder regions, inconsistent with shoulders being a consequence of the CpG effect. CpG methylation is also depleted in shoulder regions, providing further support. Consistent with these two points, CpG-associated substitutions are locally depleted in shoulder regions, demonstrating CpG effects are not responsible for elevated substitution rate at shoulders. Furthermore, I observe a local elevation of non-CpG-associated substitution specifically in the shoulder region of strong germline footprints, consistent with a non-CpG dependent mutational process.

### 7.4.3 Germline DNA binding and the broader substitution rate landscape

Over the wider DHS region, GERP scores mirror footprint coverage, most dramatically at strong somatic sites, indicative of selection. Both DHS coverage and footprint coverage support the
notion of footprint clustering, within an approximately 500 bp window. GERP scores are therefore consistent with the cumulative effect of purifying selection at footprint sites over DHSs.

At the single nucleotide level somatic-like footprints exhibited both increased substitution, at sites of probable steric-hindrance, as well as decreased substitution as a consequence of purifying selection at aligned binding sites. As substitution rate over the wider DHS region is extremely low, if it is a cumulative effect of many footprint sites, it suggests that purifying selection at sites has a stronger selectional evolutionary consequence than that of steric hindrance.

Substitution rate increases over the DHS region as confidence in calling somatic-only footprints is reduced. As the potential inclusion of germline footprints increases, so too does the potential inclusion of mutation rate into the substitution rate profile. What I observe suggests that mutation is working against purifying selection over entire DHSs. This is supported by evidence of elevated mutation rate at the single nucleotide level at aligned footprints.

Non-CpG associated polymorphism rate remains noisy across somatic-like footprints. As polymorphism rate is a more accurate measure of mutation than selection, due to its shorter evolutionary age, we expect a less defined polymorphism rate signal at somatic-like sites. At strong germline sites non-CpG-associated polymorphism rate increases along with footprint coverage. The complex GERP profile at these sites can be interpreted as a combinatory effect of underlying footprint-specific purifying selection, as shown most clearly by GERP scores in Figure 7.3 a, combined with footprint-specific elevated mutation rate, as shown most clearly by non-CpG-associated polymorphism rate in Figure 7.3 d.

Alternative mechanisms must still be considered regarding germline-specific elevated substitution across the DHS region; such as regional increase in CpG content (Figure 7.4 h), an open chromatin environment, or proximal transcriptional activity. The fact that the CpG profile does not precisely adhere to the same profile as footprint coverage and polymorphism rate, that there is a regional depletion in methylation rate rather than enrichment, and that non-CpG-associated polymorphisms are also enriched, is inconsistent with elevated CpG mutagenesis.
8.1 Conclusions: key findings

Due to a) the similarity between mutation rate and Okazaki termini at nucleosomes, and b) evidence that strong and fast binding transcription factors are involved in Okazaki termination, I investigate elevated substitution rate at transcription factor binding sites and provide a mechanistic model of error-prone Pol α tract retention during lagging strand replication. I demonstrate that “shoulders” of locally elevated substitution rate are a common, but not universal, feature of human transcription factor binding site evolution. As may be expected, it appears that the stronger or more stable the binding of a transcription factor to the DNA, the greater the apparent purifying selection of its sequence specific binding site. However, strikingly, I find that stronger binding also corresponds to greater elevation of nucleotide substitution rate in shoulder regions.

I observe in yeast S.cerevisiae that substitution rate is elevated immediately downstream of aligned and orientated Okazaki junctions, precisely where error-prone Pol α tract retention is predicted to occur, and that junction-specific elevated substitution rate occurs devoid of nucleosome phasing in the presence of predicted transcription factor binding. I provide further support for Smith and Whitehouse’s observation [108] that Reb1 binding acts as a partial barrier to Pol δ dissociation during lagging strand replication.

In the human genome I move to DNase I Hypersensitivity footprints and their left-hand edge
alignment as a proxy for Okazaki junction localisation to test our hypothesis. My results indicate site specific non-CpG-associated elevated substitution, restricted to germline-like footprints, reflective of elevated mutation rate and in support of error-prone Pol α tract retention. That substitution rate patterns exist in a non-CpG-associated manner is indicative of a mutational signature inexplicable purely by mutagenic sequences and nucleotide sequence bias. Furthermore, while no Okazaki fragment data is available for the human genome, shoulders of substitution rate at human footprint edges are consistent with those observed in yeast S.cerevisiae, implicating a widespread phenomenon in the genome at protein binding sites in both yeast and humans.

In principal, the shoulders of elevated substitution rate I observe proximal to some transcription factor binding sites and DHS footprint edges can be explained by locally elevated mutation rate or reduced purifying selection by steric hindrance. The mechanistic model I have proposed and provided support for (Figure 8.1) argues that site-specific elevated mutation rate is the prominent cause.

8.2 Conclusions: implications

The common assumption that mutation rate is random across the genome confounds functional analysis in the search for important genetic variants. By providing evidence that transcription factors are involved in Okazaki termination, and that substitution rate is elevated immediately downstream of aligned Okazaki junctions, I provide strong evidence that error-prone Pol α tract retention during lagging strand replication is responsible for conservation “shoulders” observed at transcription factor binding sites.

It must be considered that due to the nature of DNA replication (Figure 4.1), even at error-prone Pol α retention hotspots, Pol α synthesis occurs at low frequency, with the majority of DNA synthesised by Pol δ and Pol ε. Furthermore only rarely will Pol α introduce errors, and replication fidelity processes, including efficient mismatch repair at the 5’ ends of Okazaki fragments [68, 78], will in the majority of case mitigate such Pol α replication errors. These factors, combined with the likely rare occurrence of incomplete Pol α displacement, predict the signal of elevated substitution rate as a consequence of error-prone Pol α tract retention to be minute. It is an impressive result, therefore, that at the population level error-prone Pol α retention appears visible and mutagenic, implying molecular processes remain insufficient to compensate for Pol α retention and its lack of proofreading activity, and that error-prone Pol α retention is greatly underestimated when considering mutagenic forces across the genome and throughout evolution.
Figure 8.1: Evidence for the lagging strand hypothesis: elevated substitution rate at sites of predicted Pol α tract retention. A mechanistic model.
Transcription factors Reb1 and CTCF both display such substitution rate shoulders. Both factors are strong and fast binders, known to position nucleosomes [30, 41]. Nucleosomes bind replicating DNA extremely quickly, prior to Okazaki ligation [108]. Strong and fast binding factors such as Reb1 and CTCF may therefore bind replicating DNA even faster, to facilitate nucleosome phasing. Such fast binding may on occasion result in binding upstream of Okazaki Pol α-Pol δ transitions, premature dissociation of the upstream Pol δ molecule, and consequently downstream error-prone Pol α tract retention.

The fact that binding of factors in the germline creates site-specific elevated mutation rate means it comes at a genomic cost. In order for this process to prevail, selection must act to retain it, suggesting that immediate transcription factor binding and subsequent nucleosome phasing during lagging strand replication is essential for organism viability.

Pol α associated mutagenesis could prove to have great significance for human genetics, due to its elevated mutation rate at transcription factor and regulatory protein binding sites. Such increased mutagenesis will have been substantially obscured by strong purifying selection at these sites, necessary to maintain functionality. I therefore predict conservation shoulders to be hotspots for regulatory mutations, due to their presence within sites of functional importance, combined with their site-specific elevated mutation rate. These hotspots may explain the rapid evolutionary turnover of transcription factor sites [120], and the difficulty in non-coding functional site prediction by inter-species sequence conservation comparisons. Furthermore, as hyper-mutable loci, transcription factor binding sites may be frequently mutated in inherited disease and neoplastic tumors. The ability to link replication to mutation rate at transcription factor binding sites will shed important light on divergence and conservation among functional regulatory elements.

In section 1.3 I summarise known examples of non-uniform elevated mutation rate within the genome. The work I provide within this thesis fits well with these previously reported observations:

- **The effect of replication strand on mutation rate**: previous investigation of strand-specific mutation rate identified a higher rate of mutation within genes on the lagging strand compared to genes on the leading strand (1.3.3) [54]. The authors propose collision between RNA and DNA polymerases as an explanation. My work sheds new light on this observation; error-prone Pol α retention is a phenomenon restricted to the lagging strand and capable of elevating site-specific mutation rate. I therefore provide this work as a new or additional explanation for replication strand biased mutation rate.

- **Promoter mutation rate**: promoters are reported to show elevated mutation rate within
the primate lineage, contrary to the effect of purifying selection predicted at such regions of functional importance (1.3.5) [115]. The authors were unable to explain this observation but acknowledged the unusually unravelled chromatin structure and transcription factor accessibility of primate-specific promoters. From my work I now know that transcription factor binding within such regions could lead to early Pol δ dissociation and error-prone Pol α retention, consequently providing potential for elevated mutation rate.

• **Male biased mutation rate:** it has been observed that between 71 and 100% of new single nucleotide point mutations arise in the paternal lineage (1.3.6) [14, 63]. This is explained by the infidelity of DNA replication and the significantly increased number of cell divisions, and consequently DNA replications, of the male spermatogonial pool compared to the female egg. Here I provide an explanation for this infidelity by DNA replication-specific error-prone Pol α retention. The work by Campbell *et al.* [14], Kong *et al.* [63] and the work I provide within this thesis are therefore mutually supportive.

• **The nucleosome hump:** substitution rate is elevated at sites of nucleosome binding throughout the genome (1.3.7). It is under debate whether this phenomenon is as a result of purifying selection [125], elevated mutation rate [126], or both [99]. Our work provides direct evidence that substitution rate is elevated at nucleosomes as a result of site-specific elevated mutation rate by error-prone Pol α retention during lagging strand replication. This does not rule out the contribution of purifying selection at these sites, but provides a previously unavailable mechanistic model for the explanation of nucleosome hump formation.

Such sites may prove good candidate locations to focus GWAS functional variant searches. The vast majority of variants identified by GWAS studies remain unexplained; it may be that such variants fall within error-prone Pol α hotspots. Variants within such hotspots should be considered to be of particular interest, due to their location within regions of functional constraint and functional importance, along with their potentially abnormally increased rate of mutation. Along with protein coding potential, for example, identifying the locality of variants to DHS footprint frequency could be a (much needed) means of prioritising GWAS hits.

Other factors, along with error-prone Pol α retention, are not ruled out within this work. In addition to impaired displacement of Pol α DNA tracts by Pol δ, protein binding may also impair access of replication-related repair factors, such as Exo1, to correct errors in Pol α synthesised DNA [68], for example. However it does not appear that the mismatch repair machinery is generally obstructed at such sites, as mismatch repair efficiency at nucleosomes is uniform with respect to dyad position [71].
In summary, I hereby contribute to the field of genomics and molecular evolution by demonstrating that DNA mutations synthesised by error-prone Pol \( \alpha \) contribute to the genome, likely elevating mutation rate at specific regulatory sites of relevance to both human genetics and the shaping of the genome during evolution.

### 8.3 Conclusions: future direction

#### 8.3.1 Super-peaks

90,975 peaks in Okazaki junctions passing super-peak criteria remain un-assigned to DREME-identified motifs. This is the large majority of super-peaks (93.84%), and implies that in the majority of instances super-peaks are devoid of an identifiable homogeneous sequence. This implies either a) the binding of a large number of heterogeneous factors, b) the binding of one or several proteins in a non-sequence specific manner or c) the formation of Okazaki junctions in the absence of protein binding. In either case, peaks in Okazaki junctions are being formed across the genome through a means I have yet to identify, thereby highlighting this category as of particular interest, warranting further investigation.

#### 8.3.2 Human Okazaki fragment data

The most direct means of identifying the strongest sites of error-prone Pol \( \alpha \) tract retention is by quantifying the rate of Okazaki fragment junctions across the genome, as we show successfully for the yeast \( S.cerevisiae \) genome (Chapter 5). This was made possible by the availability of publicly-accessible Okazaki fragment deep-sequencing data (4.2.1) [108]. To definitively show the correlation between sites of highest error-prone Pol \( \alpha \) retention (immediately downstream of Okazaki junction super-peaks) and substitution rate in the human genome, human Okazaki fragment deep-sequencing alignment data is required. As yet, presumably due to technical limitations, no Okazaki fragment alignment data is available for the human genome. Once it becomes available its super-peak analysis is sure to produce exciting and important results.
8.3. CONCLUSIONS: FUTURE DIRECTION

8.3.3 Derived Allele Frequencies

DAF test do not measure mutation rate. If however substitution rate is high, and yet DAFs indicate purifying selection, this would diminish the potential for high substitution rate as a consequence of positive selection. It could then be inferred that high substitution rate is a consequence of elevated mutation.

Beneficial to the investigation of mutation rate at transcription factor shoulders would be the calculation of DAFs over combined DHS footprint sites, comparing rates between somatic-like and germline-like footprint subsets. To do so, SNPs should be extracted at each nucleotide position either side of footprint left edges. At each position the number of rare (<=0.015) and common (>0.015) SNP frequencies would be compared (as provided within the 1000 Genomes data [80]), to provide DAF values.

DAF values would be calculated for three additional nucleotide subsets, against which DAF values from DHS footprints can be compared:

- **Four-fold degenerate (4FD) sites**: the DAF at 4FD sites will be higher than that of the background rate.

- **Non-synonymous (NS) DAF ratio**: their DAF will be lower than that of the background rate.

- **Stop-gained DAF ratio**: nucleotides where a base change will not only have a non-synonymous effect, but will alter the codon so that it creates a premature stop codon, therefore truncating their encoded protein. Their DAF will therefore be even lower than that of other non-synonymous sites.

For DAFs to support the Pol α retention hypothesis and site-specific elevated mutation rate, we would expect values equal to or above neutral at sites of elevated substitution rate in the vicinity of aligned footprint left-hand edges.

8.3.4 CTCF binding profiles

By uniformly aligning and orientating human transcription factor binding sites I observe a subset that display a peak in substitution rate, termed a conservation “shoulder”. We believe that in certain instances an even more detailed substitution profile could be teased out of these plots.
CTCF, for example, is able to bind in multiple conformations, and is able to recognise variable binding motifs. By plotting substitution rate for all CTCF binding sites combined, I combine and therefore blur profiles of individual interactions. The profile I observe is consequently the result of multiple profiles super-imposed. It would be more informative to observe individual substitution rate profiles when binding interactions are separated.

Rhee and Pugh have described seven different motif combinations at CTCF binding sites [102] (Figure 8.2). They identify a smaller and less defined upstream motif, in addition to that of the well known nineteen base pair core motif, present in approximately 28% of binding sites. Boyle et al. have investigated in detail DNase I hypersensitivity around CTCF binding sites, as a means of deciphering its binding footprints [8] (Figure 8.3). All DNase I hypersensitivity peaks genome-wide were assigned to transcription factors by motif identification. The common nineteen base pair core motif of CTCF was used. When plotting all CTCF sites, centred on their binding footprints, Boyle et al. observed a spike in DNase I hypersensitivity immediately upstream of the core motif. They then split the binding sites according to the presence or absence of this site, (~ 80% and 20% respectively), and again DNase I hypersensitivity around the two groups was observed. They report an additional upstream dip in DNase I hypersensitivity, and corresponding novel motif identification, only in the group of binding sites where the additional DNase I hypersensitivity spike is present.

When Boyle et al. plotted substitution rate at CTCF binding sites they reported no apparent “shoulder” of conservation. They show a simple negative spike in substitution rate at the binding site (Figure 8.4 a). By plotting substitution rate at motif-aligned and motif-orientated CTCF binding sites, I see an asymmetrical unequal bi-modal distribution in conservation sites (Figure 8.4 b). Although Boyle et al. orientated their DNase I hypersensitivity plots by the underlying motif, they did not do so when plotting substitution rate. As CTCF is known to bind asymmetri-cally, this could explain why they observe such a simplified substitution rate pattern spike, as any strand-specific patterns would be merged.

From the evidence presented, of seven distinct motif combinations, and two distinct DNase I hypersensitivity profiles, we hypothesise that if divided correctly then different groups of CTCF binding sites will also display a number of distinct substitution rate profiles. A striking correlation can already be seen between the DNase I hypersensitivity profiles in the lower plot of Figure 8.3 and our detailed substitution rate profile for all CTCF binding sites combined (Figure 8.4 b). Already we see that the correct alignment and orientation of motifs reveals clarity at these sites. We predict the acuteness of CTCF shoulders to increase when profiles are separated according to binding interactions.
One method to separate CTCF binding sites would be by the identification of different motif module combinations, as described by Boyle et al. However, a problem arises that while some CTCF binding sites display a strong consensus binding motif, at a significant number of other CTCF sites this consensus sequence is vague and motif assignment is not accurate. To bypass this dependency I propose a novel means of grouping and orientating CTCF binding sites through raw sequence data peak alignment and hierarchical clustering of all CTCF-DNA interactions. This algorithm would cluster CTCF binding sites according to the similarity of their DNA-binding interactions (Figure 8.5).

CTCF binding site peaks have been calculated as described in 3.2.1 using MACS [135] and MOODS [64] software, to provide peak coverage (number of tags) and motif strength values, respectively. For further analysis, a final set of CTCF primary bound locations have been determined following criteria adapted from Rhee and Pugh [102], whereby:

- A peak (> 5 normalised tags) must be identifiable in at least 2 of 3 replicates, where replicate pairs are no more than 125 bp apart.

- The mean tag score of those replicates included must be > 25 normalised tags.

Binding site coverage for CTCF is in the form of ChIP-exo. The 5’ end of each tag therefore represents nucleotide specific factor-DNA interactions. It is the profiles of these binding interactions, for individual CTCF binding peaks that we wish to compare between all CTCF binding sites.

Strand separated SAM files have been converted to WIG files using a custom written algorithm implemented in Perl (version 5.14.1), whereby WIG scores denote the number of tag 5’ ends aligned per nucleotide. A second algorithm, combining CTCF bound locations with WIG files of 5’ tag positions, compares all peaks to each other, by means of a “sliding window” (Figure 8.5), to create a distance matrix of best possible overlap scores per peak pair. The distance matrix would be clustered using hierarchical clustering in R to define distinct CTCF binding clusters.

The aim would be to calculate substitution rate profiles for defined interaction-dependent subsets of CTCF peaks. I would hypothesise both increased clarity and increased magnitude of substitution rate profiles between subsets.
CHAPTER 8. CONCLUSIONS

8.4 Conclusions: supporting material

The computational work detailed in this thesis is supported by additional laboratory experimental validation. Laboratory work has been carried out by Martin Reijns and James Ding of the Andrew Jackson Laboratory, Edinburgh University IGMM Human Genetics Unit, in collaboration with Martin Taylor, Edinburgh University IGMM Human Genetics Unit, Principal Investigator and direct supervisor to this project. At the initial time of writing, our combined research was under publication review and all work included in this thesis unpublished. Since my original thesis submission our paper has been accepted and published in Nature [101] (attached as Appendix B). Identification and mapping of substitution rates at \textit{S. cerevisiae} and human transcription factors, \textit{S. cerevisiae} Okazaki junction super-peak identification and orientation, super-peak motif identification and sub-division and human DHS footprint substitution mapping are included within the publication, having been created as a result of this thesis.

Laboratory validation focuses upon mapping of ribonucleotide incorporation within the yeast \textit{S. cerevisiae} genome. Ribonucleotide incorporation is a known feature of Pol \(\alpha\). Ribonucleotide frequency is therefore a proxy for Pol \(\alpha\) tract retention. In a Pol \(\alpha\) mutant strain, where an amino acid substitution near the active site results in ribonucleotide incorporation eight times higher than wild type, ribonucleotide incorporation increases specifically on the lagging strand, thereby showing direct evidence of Pol \(\alpha\) tract retention within the yeast \textit{S. cerevisiae} genome, in support of our hypothesis.

When ribonucleotide incorporation is plotted at DHS footprints and Okazaki junction super-peaks, ribonucleotide incorporation aligns to aligned footprint left-hand edges and immediately downstream of aligned and orientated super-peaks, directly at sites of increased substitution rate. This shows a clear correlation between error-prone Pol \(\alpha\) tract retention and increased substitution rate at sites predicted within this work, providing experimental support for the results I report within this thesis.
Figure 8.2: **Classification of distinct CTCF binding motifs: Genome-wide CTCF-bound locations as identified by Rhee and Pugh.** The left-hand panel shows a colour chart representation of each CTCF binding site sequence, centred by motif midpoint. Sites (rows) are divided into six categories based upon the presence or absence of sequence motifs within them. Four sub-motifs were identified, marked above the left-hand panel, (green, blue, yellow and pink), and shown above their underlying sequences PWMs, (bottom right). The top right-hand table is coloured to demarcate the combinatorial use of each CTCF sub-motif. (Figure adapted from Rhee and Pugh [102]).
Figure 8.3: **High-resolution analysis of CTCF binding sites.** Boyle *et al.* plot DNase I hypersensitivity at CTCF binding sites. Above is the cumulative footprinting signal at all CTCF motif predicted sites. Centre is just those sites with no upstream DNase I digestion spike. Bottom is just those CTCF sites including a large increase in DNase I digestion upstream of the CTCF motif. The light gray bar indicates the location of the known CTCF motif. The dark gray bar represents a novel binding motif, both of which are shown as PWM logos below. The novel binding motif is only detected in CTCF footprints that contain the small upstream region with a spike in DNase I hypersensitivity. (Figure adapted from Boyle *et al.* [8]).
Figure 8.4: **Substitution rate at CTCF binding sites.** [a] substitution rate patterns at CTCF binding footprints as reported by Boyle et al., represented by phastCons scores [8]. The authors report an apparent lack of conservation "shouldering" at CTCF sites. (Figure adapted from Boyle et al. [8]). [b] my analysis of substitution rate at CTCF binding sites. All sites are centred on their motif mid-point and orientated by motif strand, so that for all sites motifs are overlapping and in the same orientation (illustrated by the motif sequence strength by PWM, top). Sites are split into four groups of equal numbers according to their sequence strength (ChIP-Seq coverage). Substitution rate (GERP) for each group is plotted in red, from light to dark (weakest to strongest binding strength).
Figure 8.5: Sliding window peak comparison. Peak A (green) and peak B (orange) are hypothetical sequence tag binding profiles of two CTCF-bound locations. Positive and negative values denote sequence tags aligned to either the forward or reverse strand, respectively. All possible overlapping combinations of the two peaks are considered, for both possible strand orientations. A score (right of each plot), is derived by summing the total area of both peaks that is not overlapping (overlapping areas are shown in yellow). As the sum of each peak is normalised to 1, the maximum score (no overlap) is 2, while the minimum score (complete overlap) is 0. The optimum overlap (0.54 in this example; red), represents the distance (similarity) between each peak pair.
### A.1 Human and yeast transcription factor ChIP-seq data

Table A.1: List of transcription factor ChIP-seq cell types and sources

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>ChIP-seq SRA file</th>
<th>Download date</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>SRR502137.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>BRCA1</td>
<td>SRR502138.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>BRCA1</td>
<td>SRR502423.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>BRCA1</td>
<td>SRR502424.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>BRCA1</td>
<td>SRR502507.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>BRCA1</td>
<td>SRR502508.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>BRCA1</td>
<td>SRR502624.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>BRCA1</td>
<td>SRR502625.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>CEBPA</td>
<td>SRR343314.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>CEBPA</td>
<td>SRR343315.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>CEBPA</td>
<td>SRR343324.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>CEBPA</td>
<td>SRR343329.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>CREB1</td>
<td>SRR343314.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>CREB1</td>
<td>SRR343315.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>CREB1</td>
<td>SRR343324.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>CREB1</td>
<td>SRR343329.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>Gene</td>
<td>SRA ID</td>
<td>Date</td>
</tr>
<tr>
<td>--------</td>
<td>---------------</td>
<td>------------</td>
</tr>
<tr>
<td>CTCF</td>
<td>SRR346401.sra</td>
<td>23-01-2012</td>
</tr>
<tr>
<td>CTCF</td>
<td>SRR346401.sra</td>
<td>23-01-2012</td>
</tr>
<tr>
<td>E2F1</td>
<td>SRR502355.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>E2F1</td>
<td>SRR502356.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>ELK4</td>
<td>SRR502086.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>ELK4</td>
<td>SRR502087.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>ELK4</td>
<td>SRR502583.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>ELK4</td>
<td>SRR502584.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>ETS1</td>
<td>SRR351712.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>ETS1</td>
<td>SRR351713.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>ETS1</td>
<td>SRR351837.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>ETS1</td>
<td>SRR351838.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>ETS1</td>
<td>SRR577867.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>ETS1</td>
<td>SRR577868.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR069066.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR069067.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR203393.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR203394.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR203395.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR309193.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR309194.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR309195.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXA1</td>
<td>SRR309196.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXO3</td>
<td>SRR403973.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXO3</td>
<td>SRR403974.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>FOXO3</td>
<td>SRR617227.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>GATA2</td>
<td>SRR502078.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA2</td>
<td>SRR502079.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA2</td>
<td>SRR502129.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA2</td>
<td>SRR502130.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA2</td>
<td>SRR502581.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA2</td>
<td>SRR502582.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA3</td>
<td>SRR361464.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA3</td>
<td>SRR361465.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA3</td>
<td>SRR502275.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA3</td>
<td>SRR502276.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA3</td>
<td>SRR502277.sra</td>
<td>05-03-2013</td>
</tr>
</tbody>
</table>
## A.1. Human and Yeast Transcription Factor Chip-seq Data

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>SRA Accession</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATA3</td>
<td>SRR502538.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>GATA3</td>
<td>SRR502539.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>HNF4A</td>
<td>SRR351747.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>HNF4A</td>
<td>SRR351748.sra</td>
<td>05-03-2013</td>
</tr>
<tr>
<td>IRF1</td>
<td>SRR502396.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>IRF1</td>
<td>SRR502397.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>IRF1</td>
<td>SRR502398.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>IRF1</td>
<td>SRR502399.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>IRF1</td>
<td>SRR502496.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>IRF1</td>
<td>SRR502497.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>IRF1</td>
<td>SRR502502.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>IRF1</td>
<td>SRR502503.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>IRF2</td>
<td>SRR524938.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>MAX</td>
<td>SRR502021.sra</td>
<td>26-02-2013</td>
</tr>
<tr>
<td>MAX</td>
<td>SRR502022.sra</td>
<td>26-02-2013</td>
</tr>
<tr>
<td>MAX</td>
<td>SRR502482.sra</td>
<td>26-02-2013</td>
</tr>
<tr>
<td>MAX</td>
<td>SRR502483.sra</td>
<td>26-02-2013</td>
</tr>
<tr>
<td>MAX</td>
<td>SRR502548.sra</td>
<td>26-02-2013</td>
</tr>
<tr>
<td>MAX</td>
<td>SRR502549.sra</td>
<td>26-02-2013</td>
</tr>
<tr>
<td>MAX</td>
<td>SRR502696.sra</td>
<td>26-02-2013</td>
</tr>
<tr>
<td>MAX</td>
<td>SRR502697.sra</td>
<td>26-02-2013</td>
</tr>
<tr>
<td>MEF2</td>
<td>ASRR351594.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>MEF2</td>
<td>SRR351595.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>MEF2</td>
<td>SRR351839.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>MEF2</td>
<td>SRR351840.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>MEF2</td>
<td>SRR577877.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>MEF2</td>
<td>SRR577878.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>MYC</td>
<td>SRR502238.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>MYC</td>
<td>SRR502239.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>MYC</td>
<td>SRR502317.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>MYC</td>
<td>SRR502318.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>MYC</td>
<td>SRR502500.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>MYC</td>
<td>SRR502501.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>MYC</td>
<td>SRR502602.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>MYC</td>
<td>SRR502603.sra</td>
<td>20-02-2013</td>
</tr>
<tr>
<td>NFIC</td>
<td>SRR577692.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFIC</td>
<td>SRR577693.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFIC</td>
<td>SRR577920.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>Gene</td>
<td>SRA ID</td>
<td>Date</td>
</tr>
<tr>
<td>----------</td>
<td>----------------------</td>
<td>------------</td>
</tr>
<tr>
<td>NFIC</td>
<td>SRR577921.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFIC</td>
<td>SRR577972.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFIC</td>
<td>SRR577973.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFIC</td>
<td>SRR577974.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFIC</td>
<td>SRR577975.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFKB</td>
<td>SRR038524.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFKB</td>
<td>SRR038525.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFKB</td>
<td>SRR038526.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFKB</td>
<td>SRR038527.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFKB</td>
<td>SRR038528.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NFYA</td>
<td>SRR090231.sra</td>
<td>24-01-2013</td>
</tr>
<tr>
<td>NKKX31</td>
<td>SRR203391.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NKKX31</td>
<td>SRR203392.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NR2F1</td>
<td>SRR191359.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NR3C1</td>
<td>SRR446023.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>NR3C1</td>
<td>SRR446024.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>PPARG</td>
<td>SRR358615.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>PPARG</td>
<td>SRR358616.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>PPARG</td>
<td>SRR358617.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>REB1</td>
<td>SRR346368.sra</td>
<td>04-01-2012</td>
</tr>
<tr>
<td>REB1</td>
<td>SRR346373.sra</td>
<td>04-01-2012</td>
</tr>
<tr>
<td>REB1</td>
<td>SRR346400.sra</td>
<td>04-01-2012</td>
</tr>
<tr>
<td>REST</td>
<td>SRR400395.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>REST</td>
<td>SRR400396.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>REST</td>
<td>SRR400397.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>REST</td>
<td>SRR400398.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>REST</td>
<td>SRR611594.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR351777.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR351778.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR351821.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR351822.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR351857.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR351858.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR577734.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR577735.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR577887.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR577888.sra</td>
<td>06-03-2013</td>
</tr>
<tr>
<td>SRF</td>
<td>SRR577911.sra</td>
<td>06-03-2013</td>
</tr>
</tbody>
</table>
### A.2 DNase hypersensitivity site coverage

Table A.2: List of DHS coverage data sources

<table>
<thead>
<tr>
<th>Data</th>
<th>Download date</th>
</tr>
</thead>
<tbody>
<tr>
<td>wgEncodeUwDnaseA549Pk</td>
<td>04-08-2011</td>
</tr>
<tr>
<td>wgEncodeUwDnaseAg10803Pk</td>
<td>04-08-2011</td>
</tr>
<tr>
<td>wgEncodeUwDnaseAoafPk</td>
<td>04-08-2011</td>
</tr>
</tbody>
</table>
APPENDIX A. DATA SOURCES

wgEncodeUwDnaseGm12865Pk 04-08-2011
wgEncodeUwDnaseH7esPk 04-08-2011
wgEncodeUwDnaseAg10803Pk 04-08-2011
wgEncodeUwDnaseAoafPk 04-08-2011
wgEncodeUwDnaseGm12865Pk 04-08-2011
wgEncodeUwDnaseH7esPk 04-08-2011
wgEncodeUwDnaseHaePk 04-08-2011
wgEncodeUwDnaseHahPk 04-08-2011
wgEncodeUwDnaseHaspPk 04-08-2011
wgEncodeUwDnaseHcfPk 04-08-2011
wgEncodeUwDnaseHcfaaPk 04-08-2011
wgEncodeUwDnaseHcmPk 04-08-2011
wgEncodeUwDnaseHcpePk 04-08-2011
wgEncodeUwDnaseHeePk 04-08-2011
wgEncodeUwDnaseHffPk 04-08-2011
wgEncodeUwDnaseHgfPk 04-08-2011
wgEncodeUwDnaseHipePk 04-08-2011
wgEncodeUwDnaseHmvecdbladPk 04-08-2011
wgEncodeUwDnaseHpafPk 04-08-2011
wgEncodeUwDnaseHpdlfPk 04-08-2011
wgEncodeUwDnaseHpfPk 04-08-2011
wgEncodeUwDnaseHrcePk 04-08-2011
wgEncodeUwDnaseHsmmPk 04-08-2011
wgEncodeUwDnaseHuvecPk 04-08-2011
wgEncodeUwDnaseHvmfPk 04-08-2011
wgEncodeUwDnaseNb4Pk 04-08-2011
wgEncodeUwDnaseNhaPk 04-08-2011
wgEncodeUwDnaseNhdfadPk 04-08-2011
wgEncodeUwDnaseNhdfneoPk 04-08-2011
wgEncodeUwDnaseNhlfPk 04-08-2011
wgEncodeUwDnaseSkmcPk 04-08-2011
wgEncodeUwDnaseTh1Pk 04-08-2011
### A.3 Methyl CpG Reduced Representation Bisulfite Seq tracks

#### Table A.3: List of Methyl CpG RRBS data sources

<table>
<thead>
<tr>
<th>Data</th>
<th>Download date</th>
</tr>
</thead>
<tbody>
<tr>
<td>wgEncodeHaibMethylRrbsAg10803UwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsAg10803UwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsGm06990UwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsGm06990UwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHaeUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHaeUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHcfUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHcfUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHcmUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHcmUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHcepUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHcepUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHeeUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHeeUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHepg2DukeSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHepg2DukeSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHipeUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHipeUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHrceUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHrceUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHsmmDukeSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsHsmmDukeSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsImr90UwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsImr90UwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsK562HaibSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsK562HaibSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsNb4UwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsNb4UwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsNhaUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsNhaUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsNhdfneoUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsNhdfneoUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
</tbody>
</table>
### APPENDIX A. DATA SOURCES

<table>
<thead>
<tr>
<th>Dataset Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>wgEncodeHaibMethylRrbsSaecUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsSaecUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsSkmcUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsSkmcUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsSknshraUwSitesRep1</td>
<td>29-07-2012</td>
</tr>
<tr>
<td>wgEncodeHaibMethylRrbsSknshraUwSitesRep2</td>
<td>29-07-2012</td>
</tr>
</tbody>
</table>
APPENDIX

MANUSCRIPTS
Lagging-strand replication shapes the mutational landscape of the genome

Martin A. M. Reijns1*, Harriet Kemp2*, James Ding1, Sophie Marion de Proe3, Andrew P. Jackson1 & Martin S. Taylor2

The origin of mutations is central to understanding evolution and of key relevance to health. Variation occurs non-randomly across the genome, and mechanisms for this remain to be defined. Here we report that the 5′ ends of Okazaki fragments have significantly increased levels of nucleotide substitution, indicating a replicative origin for such mutations. Using a novel method, emRiboSeq, we map the genome-wide contribution of polymerases, and show that despite Okazaki fragment processing, DNA synthesized by error-prone polymerase-α (Pol-α) is retained in vivo, comprising approximately 1.5% of the mature genome. We propose that DNA-binding proteins that rapidly re-associate post-replication act as partial barriers to Pol–δ-mediated displacement of Pol–α-synthesized DNA, resulting in incorporation of such Pol–α tracts and increased mutation rates at specific sites. We observe a mutational cost to chromatin and regulatory protein binding, resulting in mutation hotspots at regulatory elements, with signatures of this process detectable in both yeast and humans.

Selection pressure to maintain TF binding, and obscuring any mutational signal at these nucleotides.

Given that both classes of sites (nucleosomes and TFs) are present genome-wide and represent different biological processes, this association was probably the direct consequence of protein binding at these sites. However, to rule out site-specific biases in sequence as a confounding explanation for the observed distributions, we randomly sampled the rest of the genome for trinucleotides of identical sequence compositions and calculated the substitution rate at these sites, on a nucleotide-by-nucleotide position basis (Extended Data Fig. 1h–j). This resulted in loss of the observed patterns, establishing that nucleotide composition bias was not a contributing factor. Furthermore, the observed association was not restricted to polymorphism rates, as yeast inter-species nucleotide substitution patterns at both nucleosome and Reb1 TF binding sites were identical (Extended Data Fig. 1k, l).

Substitutions correlate with OJs

We were struck by the similarity of the distribution of Saccharomyces cerevisiae OJ sites at nucleosomes17 to that previously reported for nucleotide substitutions7,8,10–12, and set out to investigate the potential reasons for this. We established that nucleotide substitution and OJ distributions are highly correlated (Pearson’s correlation coefficient $r = 0.76$, $P = 2.2 \times 10^{-15}$) and essentially identical in pattern (Fig. 1a). Furthermore, differences in OJ distribution by nucleosome type (genic versus non-genic), spacing or consistency of binding were mirrored by the substitution rate distribution (Extended Data Fig. 1a–f). We found similar strong correlation in the regions directly surrounding TF binding sites of Reb1 (Fig. 1b; Pearson’s correlation $r = 0.57$, $P = 5.6 \times 10^{-12}$) and Rap1 (Extended Data Fig. 1g), providing further evidence for a direct association. At the sequence-specific binding sites themselves, substitution rates were depressed relative to the OJ, resulting from strong selection pressure to maintain TF binding, and obscuring any mutational signal at these nucleotides.

Figure 1 | Increased substitution rates at OJs. a, Nucleotide (nt) substitution rates (red) closely correlate with increased OJ site frequency (blue) at nucleosome (a) and Reb1 (b) binding sites. S. cerevisiae polymerase rates per nucleotide computed using sequences from nucleosome (n = 27,586) and Reb1 binding sites (n = 881). Individual data points, open circles. Solid curves, best-fit splines. Mean, dashed grey line; ±10% dotted grey lines.
We therefore concluded that OJ frequency and nucleotide substitution rates could be causally related, and set out to investigate the potential mechanism for this association.

Mutations at 5′ ends of OFs

The synthesis and processing of OFs is directional. Therefore, substitution rates would be expected to be asymmetrical relative to the direction of synthesis, if a component of this process was the cause. As most of the genome is preferentially replicated with either the forward or reverse strand as the lagging strand, we orientated regions by their dominant direction of lagging-strand synthesis. This revealed substantially increased nucleotide substitution rates immediately downstream of OJs (Fig. 2a), the level of mutational signal correlating with OJ site frequency. Quantification of substitution rates for the five nucleotides immediately upstream and downstream of the OJ (Fig. 2b) demonstrated that high frequency OJ sites (11-fold increased OJ rate relative to baseline; top 99.9th centile of sites) displayed the highest substitution rate \((P < 2.2 \times 10^{-14})\), with significant increases \((P < 2.2 \times 10^{-14})\) for medium frequency sites, \((6.1\text{-fold}, 99–99.9\text{th centile})\) but not low frequency sites \(\left(P = 0.3\right)\). OJ sites \(<99\text{th centile}\). This was not due to site-specific sequence biases, as the increase in substitution rate was lost after a trinucleotide preserving genome shuffle. Therefore, point mutations are enriched at the 5′ ends of mature OFs of frequently occurring OJ sites, sites that correspond to protein barriers to Pol-δ processivity\(^2\).

Pol-α DNA retention hypothesis

We next considered which aspect of lagging-strand synthesis might be responsible. OFs are generated by the consecutive actions of Pol-α and Pol-δ (Fig. 2c). When the previously synthesized, downstream OF is encountered, OF processing occurs\(^18\), involving the coordinated action of FEN1 and DNA2 nucleases\(^15,16\) in conjunction with continuing DNA strand displacement of lagging-strand synthesis. This revealed substantially increased mutation rates immediately downstream of OJs as the lagging strand, we orientated regions by their dominant direction of OF synthesis by Pol-α. These act as partial barriers to Pol-δ synthesis by Pol-α (Fig. 2d). Our model would predict (1) that Pol-α tracts are retained at a considerable level within the mature genome post-replication, and (2) that mutational signatures arising from such Pol-α-synthesized DNA will be increased at many DNA-binding protein sites in eukaryotes.

EmRiboSeq

To address where error-prone Pol-α DNA is retained in vivo, we used the incorporation of ribonucleotides into genomic DNA to track the activity of specific DNA polymerases. Ribonucleotides are covalently incorporated into genomic DNA by replicative polymerases\(^27,28\), although they are normally efficiently removed by ribonucleotide excision repair, a process initiated by the type 2 RNase H enzyme (RNase H2)\(^29\). In RNase-H2-deficient budding yeast, such ribonucleotides are generally well tolerated: Δrnh201 yeast has proliferation rates identical to wild type under normal growth conditions\(^30\), and therefore this genetic background ribonucleotides can be used as a ‘label’ to track polymerase activity. Furthermore, the contribution of specific polymerases can be studied using polymerases with catalytic site point mutations (such as Pol-δ-\((\text{Leu612Met})\), Pol-δ-\((\text{Leu612Met})\) and Pol-ε (Met644Gly)) that incorporate ribonucleotides at higher rates than their wild-type counterparts (refs 21, 26, 27, 30 and J. S. Williams, A. R. Clausen & T. A. P. Pol-α-synthesized DNA at 5′ends will be increased at many DNA-binding protein sites in eukaryotes. This would be desirable, as unlike other replicative DNA polymerases, Pol-α lacks 3′-to-5′ proofreading exonuclease activity, limiting its intrinsic fidelity\(^23\). On the other hand, studies on the mutagenesis pattern of reduced fidelity polymerase mutants in yeast demonstrate that Pol-α-synthesized DNA does contribute to the genome\(^21,24–26\). How comprehensive the removal or retention of such DNA is in vivo is unknown, but notably the retention of error-prone Pol-α-synthesized DNA at the 5′ end of OFs would provide a straightforward explanation for the increased mutation rates we observed. Given that protein barriers have been shown to influence OF processing\(^17\), we therefore propose that Pol-α-synthesized DNA is preferentially retained at sites where proteins bind shortly after initial OF DNA synthesis (Fig. 2c). Our model would predict (1) that Pol-α tracts are retained at a considerable level within the mature genome post-replication, and (2) that mutational signatures arising from such Pol-α-synthesized DNA will be increased at many DNA-binding protein sites in eukaryotes.

![Figure 2](image-url)  
**Figure 2** | **Frequent nucleotide substitutions at OF 5′ ends.** a. Mutation rates are increased downstream of OJs. Substitution polymorphisms (red) and OJ rate (blue) in regions surrounding high frequency OJs (top 0.1%). \(n = 5,660\) sequences orientated for dominant direction of OF synthesis. b. Mutation rates correlate with OJ peak size. Mutations are significantly enriched downstream of the junction (pink), compared to genome shuffle controls (light green/pink). Sites grouped by OJ frequency. Points denote mean and error bars denote s.d. from 100 bootstrap samples or genome shuffles (controls); statistics by paired two-sided t-test. c. Hypothesis: DNA synthesized by non-proofreading Pol-α is preferentially trapped in regions rapidly bound by proteins post-replication. These act as partial barriers to Pol-δ displacement of Pol-α-synthesized DNA, resulting in locally increased mutations.

![Figure 3](image-url)  
**Figure 3** | **Mapping DNA synthesis in vivo using emRiboSeq.** a. Replicative polymerases can be tracked using point mutants with increased ribonucleotide incorporation. Schematic of replication fork with Pol-ε (asterisk denotes Met644Gly mutant) and ribonucleotide incorporation rates for each polymerase. b. Schematic of emRiboSeq methodology. c. Schematic of replication. d, e. Mapping of leading/lagging-strand synthesis and replication origins using emRiboSeq. Ratio of OF reads\(^27\) between forward and reverse strands of chromosome 10 (Chr10); d. Corresponds to the ratio of their respective ribonucleotide content (e) for Pol-δ (orange), whereas Pol-ε (cyan) shows negative correlation. Intersections with x axes correspond to replication origins and termination regions (c–e). Experimentally validated origins (dotted pink lines). f. Pol-α* DNA is detected genome-wide by emRiboSeq as a component of the lagging strand. Strand ratios are shown as best-fit splines, x axes denote log of ratios (d–f).
Kunkel, personal communication; Fig. 3a). Yeast strains expressing these mutant polymerases have previously been used to demonstrate that Pol-α and Pol-δ are the major leading- and lagging-strand polymerases, respectively, by measuring strand-specific alkaline sensitivity of particular genomic loci30–32.

To track directly the genome-wide contribution of polymerases, we developed a next-generation sequencing approach, which we term emRiboSeq (for embedded ribonucleotide sequencing), that determines the strand-specific, genome-wide distribution of embedded ribonucleotides. This is achieved by treatment of genomic DNA with recombinant RNase H2 to generate nicks 5′ of embedded ribonucleotides, followed by ligation of a sequencing adaptor to the 3′-hydroxyl group of the deoxynucleotide immediately upstream of the ribonucleotide (Fig. 3b and Extended Data Fig. 2a). Subsequent ion-sensor semiconductor sequencing permits strand-specific mapping of ribonucleotide incorporation sites.

Control experiments using endonucleases of known sequence specificity demonstrated 99.9% strand specificity and 99.9% site specificity for the technique (Extended Data Fig. 2b–d). Using RNase H2-deficient Pol-ε(Met644Gly) and Pol-δ(Leu612Met) yeast strains, we then mapped the relative contributions of these respective polymerases genome-wide (Fig. 3c–e and Extended Data Figs 3 and 4). We found that ribonucleotide incorporation in the Pol-δ(Leu612Met) strain was substantially enriched on the DNA strand that is preferentially synthesized by lagging-strand synthesis37, in keeping with its function as the major lagging-strand polymerase30,33,34, while ribonucleotide incorporation in the Pol-ε (Met644Gly) strain exhibited an entirely reciprocal pattern consistent with its function as the leading-strand polymerase30,33,34 (Fig. 3e). Furthermore, points at which neither enzyme showed strand preference (intersection of both Pol-α and Pol-ε plots with the x-axis) corresponded precisely with annotated origins of replication. Other intersection points were also evident that correspond to replication termination regions, as well as putative, non-annotated origins. The latter overlapped with early replicating regions36 (Extended Data Fig. 3b, c). Therefore, we concluded that emRiboSeq can be used to determine the distribution of polymerase activity genome-wide, and has utility for the identification of replication origin and termination sites.

Pol-α-synthesized DNA ~1.5% of genome

Having demonstrated the validity of our technique through detailed mapping of the major replicative polymerases, we next examined the contribution of Pol-α-synthesized DNA to the budding yeast genome. Significantly, the Pol-α(Leu688Met) Arnh201 strain had a strand ratio distribution identical to that seen for Pol-δ(Leu612Met) Arnh201, consistent with the expected role for Pol-α in lagging-strand replication (Fig. 3f). Furthermore, the Pol-α(Leu688Met) pattern of strand incorporation was reciprocal to that of a wild-type polymerase strain (POL-ε), which displayed leading-strand bias, in keeping with a strong propensity for ribonucleotide incorporation by leading-strand polymerase Pol-ε compared to Pol-δ (ref. 37). Increased ribonucleotide retention on the lagging strand was also present in DNA from stationary phase Pol-α(Leu688Met) Arnh201 yeast (Extended Data Fig. 3d), demonstrating that Pol-α-derived DNA is retained in the mature genome post-replication and that this signal was not due to the transient presence of Pol-α DNA during S-phase.

To provide biochemical validation, we performed alkaline gel electrophoresis on genomic DNA extracted from Pol-α(Leu688Met), Pol-δ (Leu612Met) and Pol-ε(Met644Gly) Arnh201 yeast. Ribonucleotide incorporation was detected in all three strains (Extended Data Fig. 4a–c) and increased ribonucleotide incorporation was also detected in genomic DNA from stationary phase Pol-α(Leu686Met) yeast (Fig. 4a–c), consistent with Pol-α tract retention in mature genomic DNA. To quantify the contribution of Pol-α DNA to the genome, we used densitometry measurements from the alkaline gels to calculate ribonucleotide incorporation rates38. We detected 1,500 embedded ribonucleotides per genome in Arnh201 genomic DNA, which increased to 2,400 sites per genome for Pol-α(Leu688Met) (Fig. 4c). Observed ribonucleotide incorporation rates correspond to the product of the incorporation frequency of each polymerase and the amount of DNA it contributes to the genome. Using the in vitro ribonucleotide incorporation rates of wild-type and mutant polymerases and the number of embedded ribonucleotides embedded in vivo (Extended Data Figs 3a and 4a–c), we estimated the relative contributions of each of the replicative polymerases to the genome (Fig. 4d), calculating the contribution of Pol-α to be 1.5 ± 0.3% (mean ± s.d.).

RNase H enzymes may contribute to the removal of OF RNA primers39,40 and consequently Arnh201 strains could have altered levels of Pol-α-synthesized DNA to that seen in wild-type strains. This confounding factor was excluded using an RNH201 separation-of-function mutant41, which established that retention of Pol-α DNA was independent of a role for RNase H2 in RNA primer removal (Extended Data Fig. 5).

In conclusion, Pol-α-synthesized DNA makes a small but significant contribution to the genome, relative to the major replicative polymerases, confirming the first prediction of our model.

Mutational cost of TF binding in humans

As OF processing is a conserved process in eukaryotes, we next considered whether an OF-related mutational signature was also present in humans. Substitution rates are also increased at nucleosome cores in humans30 with an identical distribution to yeast. Furthermore, the TF NYFA has an unexplained ‘shoulder’ of increased substitution proximal to its binding sites40, reminiscent of the Reb1 pattern (Fig. 1b). We therefore investigated whether similar mutational patterns are present at other experimentally defined human TF and chromatin protein binding sites. Increased inter-species nucleotide substitution rates were detected flanking essential binding site residues, for many, but not all TFs, as well as CTCF binding sites (Fig. 5a, b and Extended Data Fig. 6). Therefore, we considered whether a OF-related mutational signature was also present in humans. Finally, to extend our analysis beyond common TF binding sites, we investigated whether the same mutational signature could be found for a broad range of regions at which regulatory proteins bind, regions we identified by the presence of DNase I footprints. Our preceding analysis of TFs suggested that nucleotide substitutions would be increased immediately adjacent to the protein binding region defined by such footprints. In yeast we found that DNase I footprint edges served as a good proxy for increased O) rate with significantly elevated substitution rates

Figure 4| Pol-α DNA synthesis contributes ~1.5% of the mature genome. a, Increased ribonucleotide incorporation in Pol-α+ stationary phase yeast is detected by alkaline gel electrophoresis. kb, kilobases; WT, wild type. b, Quantification confirms significantly increased rates in the Pol-α+ genome (n = 6 independent experiments; error bars denote s.e.m.; statistics by paired two-sided t-test). c, Estimate of relative contribution of polymerases to the genome (n = 4 independent experiments; error bars denote s.e.m.).
processes, including efficient mismatch repair at the 5’ end of OEs, will mitigate Pol-α replication errors. Additionally, Pol-α DNA will be incorporated at relatively low frequency (Extended Data Fig. 8), with most DNA at such sites still synthesized by Pol-δ and Pol-ε. However, over evolutionary timescales, it seems that these processes are insufficient to compensate fully for the lack of Pol-α proofreading activity. An alternative possibility is that protein binding may impede access of replication-related repair factors, such as Exol to correct errors in Pol-α-synthesized DNA45. However, it does not appear that the mismatch repair machinery is generally obstructed at such sites, as mismatch repair efficiency at nucleosomes is reported to be uniform with respect to dyad position46.

Nucleosome formation has a key role in ensuring genome stability46, and consequently there is an imperative for the rapid repackaging of the genome post-replication. However, we now show that this comes at the cost of increased mutation at specific sites, detectable on an evolutionary timescale. OEs-associated mutagenesis could also have importance for human genetics, as it increases mutation rates at TF and regulatory protein binding sites. Such increased mutagenesis has been substantially obscured by strong purifying selection at these sites necessary to maintain functionality. Notably, increased mutation suggests that they will be evolutionary hotspots, and may help to explain the rapid evolutionary turnover of TF sites47 and the difficulty in non-coding functional site prediction by interspecies sequence conservation comparisons. Furthermore, as hyper-mutable loci, TF binding sites may be frequently mutated in inherited disease and neoplasia.

In summary, we demonstrate that DNA synthesized by Pol-α contributes to the eukaryotic genome, probably increasing mutations at specific regulatory sites of relevance to both human genetics and the shaping of the genome during evolution.

Note added in proof: Three studies, published concurrently with this paper, have independently developed similar methods to determine the genome-wide distribution of embedded ribonucleotides48,49,50, demonstrating the utility of ribonucleotides as markers of replication enzymology in budding yeast.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 October 2014; accepted 5 January 2015.

Published online 26 January 2015.

Figure 5 | OF mutational signatures are conserved in humans. a, Nucleotide substitutions (plotted as GERP scores) are increased immediately adjacent to TF NFYA binding sites (n = 5,110). Pink to brown: lower to higher quartiles of ChIP-exo peak height (reflecting strength of binding/occupancy). Stronger binding correlates with substitution rate in the ‘shoulder’ region (asterisk). b, Increased substitution rates are not a consequence of local sequence composition effects. Strongest binding sites (brown) compared to trinucleotide preserving shuffle (black). c, Model showing nucleotide substitution profiles are the sum of mutation rate and selective pressure. d, Interspecies substitution rates are also increased adjacent to DNase I footprint edges (asterisk) (n = 33,350). Sequences aligned to left footprint edges as indicated in schematic. Right footprint edge is indistinct owing to heterogeneity in footprint length. Substitution rates are no longer increased after trinucleotide preserving shuffle from local flanking sequences (black). Brown dashes and grey shading denote 95% confidence intervals (b, d).

Discussion

Here we establish a mutational signature at protein binding sites that we suggest could result from the activity of the replicative polymerase Pol-α. We use a novel technique, emRiboSeq, to demonstrate that error-prone DNA synthesized by Pol-α is retained in the mature lagging strand. EmRiboSeq tracks genome-wide in vivo polymerase activity using ribonucleotides as a ‘non-invasive’ label, and will have significant future use for the in vivo study of DNA polymerases in replication and repair. Further optimization of emRiboSeq should permit high-resolution examination of the role of polymerases at specific sites, such as Pol-α tract retention at protein binding sites. It will also be a useful method for defining replication origin and termination sites, and furthermore will facilitate the investigation of physiological roles of genome-embedded ribonucleotides.40-44

A direct relationship between OF junctions and mutation frequency is indicated by the significant correlations between substitution rate and OF junction sites at diverse protein binding sites, although future experimental validation will be needed to establish causality formally. We find that substitution rates are specifically increased downstream of such junction sites, suggesting a replicative origin for such mutations. As Pol-α DNA tracts occur genome-wide, and Pol-δ processing of OEs is impaired by DNA-bound proteins47, we propose that retention of Pol-α DNA is increased at these functionally important sites, and is responsible for the increased mutation rate (Fig. 2c). Replication fidelity

METHODS
Yeast reference genome and annotation. All analyses were performed on the sacCer3 (V64) S. cerevisiae reference genome assembly. Data sets originally obtained with coordinates on other assemblies were projected into the sacCer3 assembly using liftOver (v261) with the corresponding chain files obtained from http://www.yeastgenome.org. All regions of the sacCer3 genome were used for read alignment but analyses including strand ratios and all rate estimates excluded the following multi-copy regions: the mitochondrial genome, rDNA locus chrXII:459153–461153 and any 100-nucleotide segment with mappability score of <0.9 (gem-mappability52 with k = 100). In total, this masked 951,532 nucleotides (7.8%) of the reference genome. Gene structure annotations were the Saccharomyces Genome Database (SGD) consensus annotations extracted from the University of California, Santa Cruz (UCSC) genome browser in November 2013. Annotated origins of replication were obtained from ref. 53. DNase I hypersensitive sites and footprints were obtained from ref. 54, and nucleosome position, occupancy and positional fuzziness (positional heterogeneity) measures were from ref. 55. Yeast replication timing data was obtained from ref. 36, where we have plotted the percentage of heavy-light (replicated) DNA (pooled samples data set), Higher percentage indicates earlier average replication time. Yeast polymorphisms and between species substitution rates. Yeast polymorphism data was obtained from the Saccharomyces Genome Resequencing project44. A polymorphic difference between any of the 37 sequenced S. cerevisiae strains was called as a polymorphic site. Sites with n > 2 alleles were only counted once as a polymorphic site. Only nucleotide position substitutions were considered, insertions and deletions were excluded. The polymorphism rate reported is the number of polymorphic sites divided by the number of sacCer3 sites with sequence coverage in at least 20k reads and at least 100 sequenced strains.

Yeast between species substitution rates were calculated from MultiZ stacked pairwise alignments obtained from the UCSC genome browser (Supplementary Table 1). Alignments for five sensu stricto yeast species (S. cerevisiae, S. paradoxus, S. mikatue, S. kudriavzevii and S. bayanus) were extracted from the original seven species alignment. The reference assembly names and phylogenetic relationship are represented by the tree (((sacCer3, sacPar), sacMik), sacKud, sacBay). Substitution rates were calculated over whole chromosomes using baseml from the paml57 package (version 4.6) under the HKY85 substitution model with ncatG = 5 categorical gamma. Per-nucleotide relative rate estimates (branch length multipliers) were obtained over the sacCer3 genome.

Human conservation measures. GERP scores53 were used as a measure of between species nucleotide diversity across 46 vertebrate species. Single-nucleotide resolution bigWig files were obtained from UCSC genome browser (hg19). For consistency of presentation with plots of polymorphism rate and yeast between species nucleotide substitution rate, the y axes in plots showing GERP scores have been inverted so that greater constraint is low and greater diversity is high.

OF sequence processing. OF sequence data was obtained from ref. 17 (GEO accession GSM385651). Analysis primarily focused on the larger ‘replicate’ library but results were confirmed in the ‘sample’ library (GEO accession GSM385650). The OF strand ratio was calculated as the sum of OF nucleotide read coverage on the forward strand divided by the same measure for reverse strand reads. OF strand ratios were calculated in windows of 2,001 nucleotides. A pseudo count of 1 read-covered nucleotide was added to both strands in each window to avoid divisions by zero. Results shown are for de-duplicated read data (identical start and end coordinates were considered duplicates). De-duplication minimises potential biases in PCR amplification, qualitatively similar results were obtained with non-de-duplicated data and support identical conclusions.

Rather than using separate Okazaki 5′ and 3′ end counts that did not always correlate well, probably due to amplification biases, sequencing and size selection biases; we produced a normalized OF rate measure. This is the average of (1) the fraction of upstream OFs that terminate with a 3′ end at a focal nucleotide, and (2) the fraction of downstream OFs whose 5′ end is at the focal nucleotide. The upstream and downstream coverage measures were based on mean Okazaki read coverage for the nucleotides located between 5 and 12 nucleotides upstream (downstream) of the focal 3′ (5′) end. This OF rate was calculated at single nucleotide resolution over both strands of the sacCer3 genome.

EmRiboSeq alignment and processing. Sequence reads (see Supplementary Table 2 for runs and read numbers) were aligned to the unmasked sacCer3 genome with bowtie2 (version 2.0.0). Subsequent filtering and format conversion were performed using Samtools (version 0.1.18) and BEDTools (version 2.16.2). Only reads with a mapping quality score >30 were kept for analysis. As there had been substantial duplication, de-duplication, downstream mapping was not performed. Read 5′-end counts were summed per strand at single nucleotide resolution over the yeast reference genome. Note that under the emRiboSeq protocol, the ribonucleotide incorporation site would be one nucleotide upstream and on the opposite strand to the mapped read 5′ end. To facilitate comparison between libraries of differing read depth, read counts were normalized to sequence tags per million mapped into the non-masked portion of the genome.

Defining TF binding sites. Reb1 and Rap1 ChiP-exo data was obtained from ref. 58 (Sequence Read Archive accession SRA044886). Sequence bar codes were clipped and sequences sorted using Perl (version 5.18.2). Reads were aligned using bowtie2 (version 2.0.0). Following the previously published protocol59 up to three mismatches across the length of each tag sequence were allowed, and the 3′ most 6 base pairs (bp) removed. Peaks were called with MACS (version 2.0.10). Following ref. 58, sites were defined as monomer if no other peaks were present within 100 bp. Where two or more peaks were present within 100 bp the peak with the highest occupancy was labelled as the primary peak. Telomeric sites were excluded using annotations within the sacCer3 sgdOther UCSC table (http://www.yeastgenome.org). The presence or absence of a motif was determined using the Motif Occurrence Detection Suite (MOODS)60 (version 1.0.1). Consensus binding motifs positional weight matrices were obtained from JASPAR61 (http://jaspar.genereg.net/). The matching motif significance threshold was set at 0.005. Multiple peaks were aligned (x = 0) to the midpoint of the JASPAR defined motif. Human TF binding sites were defined using ChiP-seq data (Supplementary Table 1) as for yeast, except that the peak clustering threshold was reduced to 50 nucleotides.

Statistical analyses. All statistical calculations were performed in R (version 3.0.0). Lines of fit used the smooth.spline function with degrees of freedom: Fig. 1a, 18 degrees; Fig. 1b, 34 degrees; Fig. 3d–f, 80 degrees of freedom (strand ratio calculated in 2,001-nucleotide consecutive windows). Sliding window averages used the rollapply function from the Zoo package with centre alignment and null padding. Pearson’s correlation was performed with the cor.test function in R. Paired Student’s t-test was performed on the raw data. Mann–Whitney tests with the wilcox.test function and lowess (locally weighted scatterplot smoothing) with the lowess function and default parameters. No statistical methods were used to predetermine sample size.

Rate estimates with compositional correction. Polymorphism and OF rates were calculated separately for each nucleotide (A, T, C or G) and the average of these for rates used as the reported or plotted measure for a nucleotide site or group of sites. This corrects for mononucleotide compositional biases that are abundant when sampling specific features of a genome. The between-species relative substitution rate calculation incorporates a compositional correction. The rate estimates shown are the number of observations divided by the number of sites with non-missing data.

Trinucleotide preserving shuffles. Every nucleotide of the sacCer3 genome was assigned to one of 64 categories based on the identity of that nucleotide and its flanking nucleotides. A vector of transformations was produced by swapping the genomic coordinate of a nucleotide for one with an identical category chosen at random. Swaps between masked and unmasked sites (see above) were prevented. 100 such vectors were produced. For a set of stacked coordinates (for example, Fig. 1a comprising 27,586 sequences, each of 251 nucleotides), every nucleotide of every sequence was substituted through the transformation vector, for a randomly selected proxy, matched for the same trinucleotide context and their corresponding rate or annotation used. This provides a compositionally well-matched null expectation. With 100 independent transformation vectors we provide empirically derived 95% confidence bounds and standard deviations on those null expectations. For human sites, shuffles were confined to shuffling the region of interest (100–300 nucleotides distant from the binding site for TF analysis and 1,000–2,000 nucleotides distant for DNase I footprint analysis). Human genomic coordinates in the ENCODE ‘Duke Excluded Regions’62 and those positions with a uniqueness score of <0.9 (gem-mappability52 with k = 100) were excluded from shuffles.

Sites selected for analysis. Thresholds were applied to define specific subsets of sites to be evaluated. For the presented data (Fig. 1a) nucleosomes with an occupancy of >80%, positional fuzziness63 of <30, with at least 30 OF reads over them, and located more than 200 nucleotides from transcription start sites were used. Other combinations (Extended Data Fig. 1) of these parameters gave qualitatively similar results and support the same conclusions. Reb1 (and Rap1) sites were defined as the primary ChiP-exo peak at a site, with sequences aligned (x = 0) to the centre of the highest scoring Reb1/Rap1 position weight matrix match within 50 nucleotides of the ChiP-exo peak summit. DNase I footprints from 41 human cell types were previously combined64 into consensus footprints (combined.fps.gz). We intersected the combined footprints with those found in each cell type using BEDTools (version 2.17.0) to identify the subset (n = 33,530) that were detected in all 41 cell types. The left-edge coordinate as defined in the combined footprint file was used as the focal nucleotide (x = 0) for analysis.

Comparison of polymorphism rates. The five nucleotide positions downstream and the five upstream of the focal OJ position (excluding x = 0 in both cases) were scored for their polymorphism rate (Fig. 2b). Rate deltas were calculated as upstream minus downstream in 100 bootstrap replicates and a paired two-sided t-test.
performed against the same calculation performed on 100 trinucleotide preserving genome shuffles of the same sites. This tests whether the difference in rate between upstream and downstream positions is greater in the observed data than the shuffled data.

**DNA purification.** Yeast strains were grown in 30 °C in YPDA to mid-log phase (see Supplementary Table 3 for a list of strains) or to saturation for stationary phase. Per 5 A_{260nm} units, cell pellets were resuspended in 200 μl lysozyme buffer (2% Triton X-100, 1% SDS, 0.5 M NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA). An equal volume of TE-equilibrated phenol and glass beads (0.40–0.60 mm diameter, Sartorius) were added, and cells lysed by vortexing for 2 min; 200 μl TE buffer was then added, followed by an additional 1 min of vortexing. After centrifugation, the aqueous phase was further extracted with equal volumes of phenol/chloroform isoamylalcohol (25:24:1) and chloroform. Total nucleic acids were precipitated with 1 ml of 100% ethanol, and dissolved in 0.5 M NaCl. RNA was degraded by treatment with 10 μg RNase A (Roche) for 1 h at room temperature. DNA was finally purified with an equal volume of Ampure XP beads (Beckman Coulter) and eluted in nuclease-free water. For library preparations DNA was isolated from up to 40 A_{260nm} units.

**Alkaline gel electrophoresis.** Isolated genomic DNA (0.5 μg) was treated with recombinant RNase H2, purified as previously described and stained with SYBR Gold (Life Technologies). Densitometry measurements and derivation of ribonuclease incorporation rates as previously described. Percentage genome contribution for each replicative polymerase (x) was calculated using the following formula: N_{polar}F_{polar}/(N_{polar}F_{polar} + N_{apolar}F_{apolar} + N_{apolar}F_{polar}), with N_{polar}, the number of ribonucleic acids incorporated in one yeast genome for the mutant polymerase, above that detected in the Δrnh201 POL strain, measured on the same alkaline gel, and F_{polar} the frequency of incorporation by that polymerase (see Fig. 3a).

**EmRiboSeq library preparation and sequencing.** DNA was sonicated using a Bioruptor Plus (Diagenode) to achieve an average fragment length of approximately 400 bp. Fragmented DNA was concentrated by ethanol precipitation and size selected using 1.2 volumes of Ampure XP DNA. DNA was quantified by nanodrop and up to 5 μg was used for NEBNext End Repair and dA-tailing (New England BioLabs) following the manufacturer’s guidelines. After the end-repair reaction, DNA was purified using 1.2 volumes of Ampure XP. Subsequent steps were performed in the presence of Ampure XP beads, capturing the DNA by adding NaCl and PEG8000, to final concentrations of 1.25 M and 10%, respectively. The trP1 adaptor (see below) was attached using NEBNext Quick Ligation with 120 pmol of adaptor per microgram of DNA for 14–18 h at 16 °C.

Terminal transferase (NEB) was then used to block any free 3′ ends with ddATP. Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific) was then used to amplify the library (Thermo Scientific) and up to 5 μg was used for NEBNext End Repair and dA-tailing (New England BioLabs) following the manufacturer’s guidelines. After the end-repair reaction, DNA was purified using 1.2 volumes of Ampure XP. Subsequent steps were performed in the presence of Ampure XP beads, capturing the DNA by adding NaCl and PEG8000, to final concentrations of 1.25 M and 10%, respectively. The trP1 adaptor (see below) was attached using NEBNext Quick Ligation with 120 pmol of adaptor per microgram of DNA for 14–18 h at 16 °C. Fragments with biotinylated A adaptor were captured on streptavidin-coupled M-280 Dynabeads (Life Technologies) following the manufacturer’s guidelines, and non-biotinylated strands were released in 0.15 M NaOH. Single-stranded fragments were concentrated by ethanol precipitation.

**Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific) was then used for second strand synthesis with primer A to produce a double stranded library.** Size selection of fragments between 200 and 300 bp in size was performed using 2% E-Gel EX (Life Technologies). Finally, this library was quality checked and quantified using a 2100 Bioanalyzer (Agilent Technologies) before emulsion PCR, using the Ion Torrent One Touch, and next generation sequencing on the Ion Torrent PGM or Proton platform (Life Technologies).

**Oligonucleotides and adaptor design.** Custom oligonucleotides were synthesized by Eurogentec. Adaptor primer pairs were annealed by heating at 95 °C for 5 min and cooling gradually. Sequences of the adaptor primer pairs were as follows. Adaptor 1 (trP1): trP1-top, 5′-CCTCTCTATGGGCAGTCGGTGAT-phosphorothioate-T′-3′; trP1-bottom, 5′-phosphate-ATCACCGACTGCCCCAT AGAGAGGC-dideoxy-3′. Adaptor 2 (A): A-top, 5′-phosphate-CTGAGTCGGA GACACCGAGGGATGAGTG-phosphorothioate-3′; A-bottom, 5′-biotin-CCATCCTC ATCCCTGCGTGCTCTGGACTCGAGNNNNNNN3-phosphorothioate-3′. The sequence for primer A used in second strand synthesis was 5′-CCATCTCATC CCFGCGGTGCCTCGCGAC-3′.

**Data sources, sequencing data and S. cerevisiae strains.** Documented in Supplementary Tables 1–3.
Extended Data Figure 1 | Increased OJ and polymorphism rates correlate at binding sites of different nucleosome classes and at Rap1 binding sites.

a–f, OJ and polymorphism rates are strongly correlated for different classes of S. cerevisiae nucleosomes. Data presented as in Fig. 1a, for different sub-classes of S. cerevisiae nucleosomes, demonstrating that OJ and polymorphism rates co-vary in all cases. Transcription start site proximal nucleosomes (d) are probably subject to strong and asymmetrically distributed selective constraints, which is likely to explain the modestly reduced correlation for this subset. Such transcription start site proximal nucleosomes were excluded from analyses of other categories presented (b, c, e, f), except ‘all nucleosomes’ (a).

g, OJ and polymorphism rates are correlated for the S. cerevisiae TF, Rap1. Data presented, as for Reb1 in Fig. 1b, show increased OJ and polymorphism rates around its binding site, with a dip corresponding to its central recognition sequence. h–j, Increased polymorphism and OJ rates at Rap1 (h), nucleosome (i) and Reb1 (j) binding sites are not caused by biases in nucleotide content. Distributions calculated as for g, Fig. 1a and b, respectively, using a trinucleotide preserving genome shuffle. Pink shaded areas denote 95% confidence intervals for nucleotide substitution rates (100 shuffles).

k, l, Polymorphism (red) and between-species (black) substitution rates are highly correlated for nucleosome (k) and Reb1 (l) binding sites. Best fit splines shown only. y axes scaled to demonstrate similar shape distribution. Values plotted as percentage relative to the mean rate for all data points (central 11 nucleotides excluded for calculation of mean in g, l).
Extended Data Figure 2 | EmRiboSeq methodology and validation.

**a.** Schematic of emRiboSeq library preparation. rN, ribonucleotide.

**b–d.** Validation of strand-specific detection of enzymatically generated nicks through linker-ligation. Nb.BtsI nicking endonuclease cleaves the bottom strand of its recognition site releasing a 5' fragment (cyan) with a free 3'-OH group after denaturation, to which the sequencing adaptor (pink) is ligated, allowing sequencing and mapping of this site to the genome (b). Nb.BtsI libraries have high reproducibility between Δnh201 POL and Δnh201 Pol-α* (pol1-L868M) strains after normalizing read counts to sequence tags per million (TPM). Bona fide Nb.BtsI sites were equally represented, at maximal frequency, in both libraries (c). Those with lower frequencies represented sites in close proximity to other Nb.BtsI sites, causing their partial loss during size selection. Additionally, Nb.BtsI-like sites were detected as the result of star activity. Libraries were also prepared using BciVI restriction enzyme digestion, that did not show such star activity (data not shown), allowing calculation of the site specificity for the method (>99.9%). Summed signal at Nb.BtsI sites shows >99.9% strand specificity (blue, correct strand; grey, opposite strand) and >99% single nucleotide resolution (d).
Extended Data Figure 3 | Mapping replicative polymerase DNA synthesis using emRiboSeq.  

*a* Point mutations in replicative polymerases elevate ribonucleotide incorporation rates, permitting their contribution to genome synthesis to be tracked. Schematic of replication fork with polymerases and their ribonucleotide incorporation rates (refs 27, 30 and J. S. Williams, A. R. Clausen & T. A. Kunkel, personal communication) as indicated (POL denotes wild-type polymerases; asterisk denotes point mutants). Embedded ribonucleotides indicated by ‘R’; additional incorporation events due to polymerase mutations highlighted by shaded circles.

*b*, *c*, Mapping of leading/ lagging-strand synthesis by Pol-δ* and Pol-ε* yeast strain using emRiboSeq (as in Fig. 3) highlights both experimentally validated (pink dotted lines) and putative (grey dotted lines) replication origins. These often correspond to regions of early replicating DNA

*d*, Pol-α* DNA is detected genome-wide by emRiboSeq as a component of the lagging strand in stationary phase yeast, as shown by the opposite pattern for a polymerase wild-type strain. Strand ratios are shown as best-fit splines with 80 degrees of freedom, y-axes show log2 of the strand ratio calculated in 2,001-nucleotide windows (*b*–*d*).
Extended Data Figure 4 | Quantification of in vivo ribonucleotide incorporation by replicative polymerases. a, b, Representative alkaline gel electrophoresis of genomic DNA from yeast strains with mutant replicative DNA polymerases (a), with accompanying densitometry plots (b). Embedded ribonucleotides are detected by increased fragmentation of genomic DNA following alkaline treatment in an RNase H2-deficient (Arnh201) background. Increased rates are seen with all three mutant polymerases (indicated by asterisk, as defined in Extended Data Fig. 3a), and are reduced in Pol-e* which contains the point mutation Met644Leu, a mutation that increases selectivity for dNTPs over rNTPs27. c, Quantification of average ribonucleotide incorporation in polymerase mutants from four independent experiments. DNA isolated from mid-log phase cultures; error bars denote s.e.m. Overall ribonucleotide content is the product of incorporation frequency and the total contribution of each polymerase, resulting in the total ribonucleotide content detected to be highest for Pol-e* (14,200 per genome), followed by Pol-δ* (4,300 per genome), Pol-α* (2,700 per genome), POL (1,900 per genome) and Pol-e’ (860 per genome). d, Most of the yeast genome exhibits directional asymmetry in replication (median 4:1 strand ratio). Count of genomic segments calculated for consecutive 2,001-nucleotide windows over the yeast genome based on reanalysis of OF sequencing data17 denoted as ‘Okazaki-seq’. The strand asymmetry ratio was calculated after re-orienting all regions such that the predominant lagging strand was the forward strand. e–g, Genome-wide quantification of strand-specific incorporation of wild-type and mutant replicative DNA polymerases determined by emRiboSeq reflects their roles in leading- and lagging-strand replication. A close to linear correlation with Okazaki-seq strand ratios is observed. The strand ratio preference for lagging-strand ribonucleotide incorporation for independent libraries (including stationary phase libraries for POL and Pol-α*, marked by diamonds) was plotted against the lagging:leading-strand ratio determined using Okazaki-seq data (only ratios $\leq 1:1$ for the latter are shown for clarity). There was high reproducibility between experiments in strand ratio preferences. Lines are lowess smoothed (see Methods) representations of the full data sets (representative examples given in f and g). f, g, Scatter plots illustrating the individual strand ratio data points for 2,001-nucleotide windows, for stationary phase POL (f) and Pol-α* (g) yeast. Pearson’s correlation = 0.49, $P < 2.2 \times 10^{-16}$ for POL (f); correlation = 0.75, $P < 2.2 \times 10^{-16}$ for Pol-α* (g).
Extended Data Figure 5 | Pol-α-synthesized DNA retention is independent of RNase H2 processing of RNA primers. a, b, The ribonucleotide content of genomic DNA is unchanged between Δrnh201 strains transformed with empty vector (−) or vector expressing Rnh201 separation-of-function mutant (sf), that retains the ability to cleave RNA:DNA hybrids, including RNA primers, but cannot cleave single embedded ribonucleotides. In contrast, the same vector expressing wild-type Rnh201 (wt) fully rescues alkaline sensitivity of the DNA. As complementation with the separation-of-function mutant had no detectable effect on the ribonucleotide content seen in the Pol-α(Leu868Met) Δrnh201 strain, retention of Pol-α-synthesized DNA appears to be independent of a putative role for RNase H2 in RNA primer removal. Representative result shown for $n=3$ independent experiments. c, Wild-type and mutant Rnh201 are expressed at equal levels, as shown by immunodetection of the C-terminal FLAG tag. Loading control, actin.
Extended Data Figure 6 | Elevated substitution rates are observed adjacent to many human TF binding sites. a–d, Nucleotide substitution rates (plotted as GERP scores) are elevated immediately adjacent to REST (a, b) and CTCF binding sites (c, d). Colour intensity shows quartiles of ChIP-seq peak height (pink to brown: lower to higher), reflecting strength of binding/occupancy. Stronger binding correlates with greater increases of proximal substitution rate in the ‘shoulder’ region (asterisk). Increased substitution rates are not a consequence of local sequence composition effects (b, d). Strongest binding quartile of sites (brown) is shown compared to a trinucleotide preserving shuffle (black) based on the flanking sequence (100–300 nucleotides from motif midpoint) of the same genomic locations. Brown dashed line and grey shading denote 95% confidence intervals. e, Substitution rates plotted as GERP scores for human TF binding sites identified in ChIP-seq data sets (in conjunction with binding site motif). Sites aligned (x = 0) on the midpoint of the TF binding site within the ChIP-seq peak (colours as for a–d). Dashed black line shows y = 0, the genome wide expectation for neutral evolution.
Extended Data Figure 7 | OJ and polymorphism rates are increased at yeast DNase I footprints. a, b, DNase I footprint edges correspond, genome-wide, to increased OJ rates and locally elevated polymorphism rates in S. cerevisiae (a), a pattern that is maintained when footprints associated with Reb1 and Rap1 binding sites are excluded (b). Genome-wide DNase I footprints ($n = 6,063$) and excluding those within 50 nucleotides of a Reb1 or Rap1 binding site ($n = 5,136$) were aligned to their midpoint. c, d, Aligning DNase I footprints on their left edge rather than midpoint (to compensate for substantial heterogeneity in footprint size) demonstrates a distinct shoulder of elevated polymorphism rate at the aligned edge (c), with a significant elevation compared to nearby sequence upstream from the footprint (d). DNase I footprints from a were aligned to their left edge ($x = 0$) with corresponding polymorphism rates shown (c). The increased polymorphism rate cannot be explained by local sequence compositional distortions (d). Nucleotide substitution rates in the 11 nucleotides centred on the DNase footprint edge (pink line), and another 11 nucleotides encompassing positions $-35$ to $-25$ relative to the footprint edge (green line) were quantified. Darker pink and green filled circles denote the mean of observed substitution rates and lighter shades denote the mean for the same sites after trinucleotide preserving genomic shuffles. Error bars denote s.d.; statistics by Mann–Whitney test. e, Model shows that correlation of increased nucleotide substitution and OJ rates are consistent with increased mutation frequency across heterogeneous DNase I footprints. Polymorphism is reduced at sequence-specific binding sites within the footprints, owing to functional constraint. Therefore, the effect of OF-related mutagenesis in these regions is most sensitively detected in the region immediately adjacent to the binding site (left of vertical dashed blue line, representing footprints aligned to their left edge). This ‘shoulder’ of increased nucleotide substitutions represents sites with increased, OJ-associated mutation is followed by a region of depressed substitution rates, owing to selective effects of the functional binding sites within the footprints (to the right of the dashed blue line). Signals further to the right are not interpretable given the heterogeneity in DNase I footprint sizes. Given strong selection at TF and DNase I footprint sites, this ‘shoulder’ of elevated nucleotide substitutions could represent a measure for the local mutation rate for such regions, analogous to that measured by the fourfold degenerate sites in protein coding sequence.
Extended Data Figure 8 | Model to show Pol-α DNA tract retention downstream of protein binding sites. a, OF priming occurs stochastically, with the 5’ end of each OF initially synthesized by Pol-α and the remainder of the OF synthesized by Pol-δ. b, c, OF processing: when Pol-δ encounters the previously synthesized OF, Pol-δ continues to synthesize DNA displacing the 5’ end of the downstream OF, which is removed by nucleases to result in mature OFs which are then ligated. The OJs of such mature OFs before ligation were detected previously after depletion of temperature-sensitive DNA ligase I. They demonstrated that if a protein barrier is encountered (grey circle), Pol-δ progression is impaired, leading to reduced removal of the downstream OF (b). Given that ~1.5% of the mature genome is synthesized by Pol-α, a proportion of lagging strands will retain Pol-α-synthesized DNA (red). When Pol-δ progression is impaired by protein binding, this will lead to an increased fraction of fragments containing Pol-α-synthesized DNA downstream of such sites (c).
BIBLIOGRAPHY


   High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells.

   Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms.

    Switching polo-like kinase-1 on and off in time and space.

    Natural selection on protein-coding genes in the human genome.

    The logic of chromatin architecture and remodelling at promoters.

    Estimating the human mutation rate using autozygosity in a founder population.

    Properties and rates of germline mutations in humans.

    Cell evolution and earth history: stasis and revolution.
    *Philosophical Transactions of the Royal Society of London*, 361:969–1006.

    Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes.
[17] Cooper, G. M., Stone, E. A., Asimenos, G., Program, N. C. S., Green, E. D., Batzoglou, S., ,
    Distribution and intensity of constraint in mammalian genomic sequence.
    Genome, 15:901–913.

    Comprehensive analysis of transcriptional promoter structure and function in 1% of the human
    genome.
    Genome Research, 16:1–10.

    Patterns of evolutionary constraints on genes in humans.
    BMC Evolutionary Biology, 8.

    Repetitive elements may comprise over two-thirds of the human genome.
    PLoS Genetics, 7:1–12.

    A perspective of promoter architecture from the ccaat box.
    Cell Cycle, 8:4127–4137.

    A multiplicity of ccaat box-binding proteins.
    Cell, 50:863–872.

    A new perspective on isochore evolution.

    Conserved nucleosome positioning defines replication origins.
    Genes & Development, 24:748–753.

    Genomics: Encode explained.

    An integrated encyclopedia of dna elements in the human genome.
Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes.

Long non-coding rnas: new players in cell differentiation and development.

An exceptionally conserved transcriptional repressor, ctcf, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes.
*Molecular and Cellular Biology*, 16:2802–2813.

The insulator binding protein ctcf positions 20 nucleosomes around its binding sites across the human genome.

Structural and functional properties of genes involved in human cancer.
*BMC Genomics*, 7.

Dnase footprinting: a simple method for the detection of protein-dna binding specificity.

Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution.

Insulators: exploiting transcriptional and epigenetic mechanisms.
*Nature Reviews Genetics*, 7:703–713.

Does radial nuclear organisation influence dna damage?

Patterns and rates of intron divergence between humans and chimpanzees.
Genome Research, 19:2090–2100.


Cell, 137:445–458.


Nature Genetics, 39:1140–1144.


[http://genome.ucsc.edu/](http://genome.ucsc.edu/).


Rates and fitness consequences of new mutations in humans.
*Genetics*, 190:295–304.

Transcription as a source of genome instability.

Evolutionary rate at the molecular level.

Non-darwinian evolution.

Evolution at two levels in humans and chimpanzees.

Rate of de novo mutations and the importance of father's age to disease risk.

Moods: fast search for position weight matrix matches in dna sequences.

The ucsf genome browser and associated tools.

Ultrafast and memory-efficient alignment of short dna sequences to the human genome.
*Genome Biology*, 10:R25.

The sequence alignment/map (sam) format and samtools.
Exonuclease 1 preferentially repairs mismatches generated by dna polymerase alpha.

Re-annotation of protein-coding genes in the genome of saccharomyces cerevisiae based on
support vector machines.

[70] Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P.,
Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O’Kelly,
M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen, A. N., Jones, M., Quail,
Population genomics of domestic and wild yeasts.

[71] Lujan, S. A., Clausen, A. R., Clark, A. B., MacAlpine, H. K., MacAlpine, D. M., Malec, E. P.,
Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composi-
tion.

A jurassic eutherian mammal and divergence of marsupials and placentals.

An overview of the structures of protein-dna complexes.
*Genome biology*, 1.

Rate, molecular spectrum, and consequences of human mutation.

[75] Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman,
Jaspar 2014: an extensively expanded and updated open-access database of transcription
factor binding profiles.

Induction of instability at selected loci in maize.  
*Genetics*, 38:579–599.

Adaptive protein evolution at the adh locus in drosophila.  

Differential correction of lagging-strand replication errors made by dna polymerases alpha and delta.  

The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model.  

An integrated map of genetic variation from 1,092 human genomes.  

Genome-scale dna methylation maps of pluripotent and differentiated cells.  

Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis.  

The rise of regulatory rna.  

An expansive human regulatory lexicon encoded in transcription factor footprints.

Two dna polymerases may be required for synthesis of the lagging dna strand of simian virus 40.

Oridb: a dna replication origin database.

Slightly deleterious mutant substitutions in evolution.

Mechanism of dna chain growth. i. possible discontinuity and unusual secondary structure of newly synthesized chains.

Sequence logos for dna sequence alignments.
*Bioconductor*.

Multiple sequence-specific factors generate the nucleosome-depleted region on cln2 promoter.

Genomic evidence for elevated mutation rates in highly expressed genes.

The gencode pseudogene resource.
*Genome Biology*, 13:R51.

The nfy transcription factor functions as a repressor and activator of the von willebrand factor promoter.
High rate of dna loss in the drosophila melanogaster and drosophila virilis species groups.

Ctcf: master weaver of the genome.

Epigenetic silencing mechanisms in budding yeast and fruit fly: different paths, same destinations.

Pedersen, J. S., Bejerano, G., Baertsch, R., Rosenbloom, K. R., Kent, J., and Haussler, D.
Forces shaping the fastest evolving regions in the human genome.

What fraction of the human genome is functional?

Widespread signatures of recent selection linked to nucleosome positioning in the human lineage.

[100] Rao10, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson,
A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping.

[101] Reijns, M. A. M., Kemp, H., Ding, J., de Proce, S. M., Jackson, A. P., and Taylor, M. S.
Lagging-strand replication shapes the mutational landscape of the genome.

Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution.
Genome-wide identification of dnasei hypersensitive sites using active chromatin sequence libraries.

Revising the human mutation rate: implications for understanding human evolution.

The structure of change.

Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.

Mysteries of heterochromatic sequences unravelled.

Intrinsic coupling of lagging-strand synthesis to chromatin assembly.

Human polymorphisms around recombination hotspots.

The influence of recombination on human genetic diversity.

Some aspects of the biology of cells and their possible evolutionary significance.
*Cambridge University Press*.

*Human molecular genetics.*

Dnase footprint signatures are dictated by factor dynamics and dna sequence.

Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.
_Cell_, 126:663–676.

Heterotachy in mammalian promoter evolution.

Rapidly evolving human promoter regions.

The accessible chromatin landscape of the human genome.
_Nature_, 489:75–82.

Focus on the centre: the role of chromatin on the regulation of centromere identity and function.
_EMBO Journal_, 28:2337–2348.


The sequence of the human genome.


Evolution of transcription factor binding in metazoans - mechanisms and functional implications.


