Patterns of selective constraint across the mouse and human genomes





George Powell

St Anne's College, University of Oxford

Supervisors: Professor Cecilia Lindgren, Dr Ann-Marie Malon, and Dr Michelle Simon

Submitted for the degree of *Doctor of Philosophy*

May 2022

Table of Contents

Abstract 4

Chapter 1 5

Background 5

1.1 Survivorship bias and selective constraint 5
1.2 Factors that affect the distribution and frequency of variants across genomes 8
1.2.1 The mutation rate spectrum 8
1.2.2 Dominance and selection 14
1.2.3 Linked selection 15
1.2.4 Genetic drift and effective population size 19
1.3 Selective constraint across the human genome 20
1.3.1 Selective constraint across human protein-coding genes 22
1.3.2 Selective constraint across human non-coding sequences 26
1.3.3 Functional and clinical correlations of constraint metrics 27
1.3.4 Limitations of constraint metrics 30
1.4 Selective constraint across the mouse genome 31
1.4.1 In-vivo mammalian gene-phenotype associations from knockout experiments 33
1.4.2 Characterisation of the mouse regulatory landscape from biochemical assays 34
1.4.3 Human-mouse homology and mouse model limitations 35
1.4.4 Selective constraint across the mouse genome 37
1.4.5 A source of mouse genetic diversity 39
1.5 Thesis aims and outline 41

Chapter 2 44

Quantifying the effect of local sequence context on substitution rates across the mouse genome 44

2.1 Abstract 44
2.2 Introduction 45
2.3 Results 47
2.3.1 The effect of local sequence context on substitution rates. 47
2.3.2 Power simulations with increasing k-mer sizes 52
2.3.4 GC content and mutability 54
2.3.5 The effect of methylation on CG dinucleotide substitution rates 56
2.3.6 Evidence mutations are less frequent in nucleotide contexts where they are deleterious 58
2.4 Discussion 62
2.5 Methods 65
2.5.1 Data download and analysis software 65
2.5.2 Filtering variants and genomic regions 66
2.5.3 Local sequence context models and probabilities of substitution 66
2.5.4 Methylation states and probabilities of substitution 68
2.5.5 Testing the prediction accuracy of substitution models 69

2.5.6 Power testing 70
2.5.7 Mutational effect and probability of occurrence 71

Chapter 3 74

Selective constraint against nonsynonymous variation in the mouse lineage 74
3.1 Abstract 74
3.2 Introduction 75
3.3 Results 77
3.3.1 Quantifying mouse gene constraint against nonsynonymous variation. 77
3.3.2 Mouse gene constraint and knockout phenotype 78
3.3.3 Human pathogenic variant site enrichment 84
3.4 Discussion 86
3.5 Methods 88
3.5.1 Raw data and dependencies 88
3.5.2 Defining genes, quality variants, and coding consequences 89
3.5.3 Transcript specific probabilities of nonsynonymous and synonymous substitution 90
3.5.4 Calculating gene-wide intolerance of nonsynonymous variation 92
3.5.5 Assessing mouse gene constraint and knockout phenotype 93
3.5.6 Defining orthologues and correlation with other measures of constraint 96
3.5.7 Assessing pathogenic variant enrichment in the human orthologues 96

Chapter 4 98

Human-mouse conservation of non-coding features and disease-associated loci 98
4.1 Abstract 98
4.2 Introduction 99
4.3 Results and Discussion 101
4.4 Conclusions 107
4.5 Methods 107
4.5.1 Calling topologically associated domains (TADs) 107
4.5.2 Compiling an annotation model of genomic features across the human and mouse genomes 108
4.5.3 Human and mouse pairwise alignment 110
4.5.4 Tissue-specific enhancers 110
4.5.5 Defining human variant sites associated with Mendelian and complex disease 111
4.5.6 Assessing the alignment of human variant sites associated with Mendelian and complex disease to the mouse genome 112

Chapter 5 114

Characterising the mouse non-coding genome by genetic diversity 114
5.1 Abstract 114
5.2 Introduction 115
5.3 Results 117
5.3.1 Ranking constrained regions across the mouse genome 117

5.3.2 Feature composition of constrained regions 120
5.3.4 Human pathogenic variant site enrichment across the most constrained regions of the mouse genome 124
5.4 Discussion 128
5.5 Methods 130
5.5.1 Raw data and computational dependencies 130
5.5.2 Defining high-quality variants and genomic loci 131
5.5.3 Determining sequence-specific probabilities of mutation 132
5.5.4 Ranking constrained regions across the genome 133
5.5.5 Genome-wide annotations 134
5.5.6 Feature composition of constrained regions 136
5.5.7 Human variant enrichment 136

Chapter 6 138

Summary and Conclusions 138
6.1 Summary of findings 138
6.2 Limitations and future research 141
6.2.1 Correlation between mutation rate and fitness consequences 141
6.2.3 Methodological limitations and comparisons with human constraint scores 142
6.2.4 Implications of linked selection 146
6.3 Concluding remarks 147

Supplementary Information 149
Chapter 1 149
Chapter 2 157
Chapter 3 165
Chapter 4 173
Chapter 5 181

References 186

Acknowledgments 208

Abstract

Understanding how genetic variation affects organism phenotype and fitness is a fundamental question in biology with implications for biomedical research. Constrained genomic loci (i.e. loci under selective constraint) are depleted of variation due to negative selection and are inferred to be functionally important to organism fitness. Recent sequencing in human populations has allowed researchers to make contemporary estimates of selective constraint across the genome. These estimates provide a basis for identifying functionally important elements and interpreting the fitness consequences of mutations. Mouse models are critical to our understanding of mammalian biology and are used to experimentally assess the function of genomic elements and genetic variation *in-vivo*. However, how selective constraint shapes patterns of variation across the mouse genome remains relatively understudied.

In this thesis, I identify and characterise constrained genes and genomic regions in mice using single nucleotide variation from sequencing in wild populations (*Mus musculus* and *Mus spretus*). I quantify how local sequence context affects substitution probabilities across the mouse genome in order to control for differences in mutability when estimating selective constraint. I show a nucleotide's 7-mer context (i.e. the three bases flanking the nucleotide on each side) explains more variance in substitution rates than either the 3-mer or 5-mer contexts. I then use these estimates of sequence mutability to identify constrained genes and genomic regions. The results show that many highly constrained genes are essential for embryonic development, and that mouse gene constraint is positively correlated with pleiotropy (i.e. a greater degree of constraint is correlated with an increased number of phenotype associations). Importantly, I also observe that the most constrained genes and genomic regions are enriched for variant sites associated with human disease, suggesting pathogenicity may have been conserved between the lineages.

I conclude by discussing how these results may help to interpret the functional importance of genomic elements, prioritise disease-associated loci for functional assessment, and their importance for facilitating more reliable inferences from mouse models.

Chapter 1

Background

1.1 Survivorship bias and selective constraint

In the middle of the Second World War, Abraham Wald, a Hungarian Jew who escaped Nazi Germany and fled to America, published “A method for estimating plane vulnerability based on damage of survivors”. Many US planes were being lost to enemy fire, but flying weight restrictions and finite wartime resources meant only limited armour could be added to the planes to provide further protection. Wald’s thesis aimed to determine which areas on planes were most critical to function and vulnerable to enemy fire and therefore should be prioritised for reinforcement, based on the distribution of bullet holes across planes that returned home (Figure 1). In contradiction to the US military's recommendation of adding armour to the areas that showed the most damage on surviving planes, Wald suggested that areas showing the least damage should be reinforced. Wald pointed out that any damage to surviving planes did not critically compromise their ability to fly as it did not prevent them from returning home. By extension, you would expect to see a lack of damage on surviving planes in areas where damage would severely compromise their ability to fly. Planes that were critically damaged in these vulnerable regions would have been brought down and made unavailable for assessment. Wald reasoned, therefore, that due to survivorship bias, areas with the least damage on returning (surviving)

planes could be inferred as being the most critical to function and vulnerable to enemy fire (Wald 1980)[1].

This same reasoning can be used to infer regions/sequences across genomes that are functionally important to the fitness of organisms and intolerant of mutation. When mutations occur across the genome and create variants, these variants are more likely to be lost from the population if they are deleterious to fitness. For example, if a variant reduces the reproductive success of the individual carrying it (i.e. it is deleterious to fitness), it is less likely that the allele will be passed to the next generation and spread, and more likely it will become extinct within the population. Variants that are lost from a population as a result of this mechanism are described as being purged by negative selection (also known as purifying selection). By assessing the distribution of genetic variation across genomes, it can be inferred that regions that are relatively depleted of variation have had variants purged by negative selection and have functional importance, specifically with regard to fitness (Figure 1). This intolerance of variation at the genomic level, due to negative selection, is described as selective constraint.

Selective constraint is quantitative and broadly defined in this thesis as intolerance of genetic variation at a given locus, relative to neutrally evolving loci, as a consequence of purifying selection. Loci that are highly constrained are more likely to have variants purged by negative selection and, therefore, to be depleted of variation and conserved over evolutionary time. The motivation for identifying constrained genomic regions is to infer loci that are critically or functionally important to the fitness of organisms. This has particular relevance for medical genetics, as estimates

[1] This analogy for selective constraint is used by (Havrilla et al. 2018).

of selective constraint can be used to prioritise candidate mutations that are likely to have a causative role in disease phenotypes (Bartha et al. 2018; Karczewski et al. 2020).

Throughout this thesis, I maintain a conceptual distinction between sequence conservation and sequence constraint. I use constraint to describe the likelihood that mutations across a given sequence will be deleterious and therefore purged. For example, the more constrained a sequence is the more likely mutations are to be deleterious and therefore purged by negative selection. By comparison, I use conservation to describe the degree of observed similarity between orthologous sequences. For example, if two orthologous sequences are identical they are described as conserved. These two terms are related as seqeucnece constraint will result in conservation over evolutionary time.

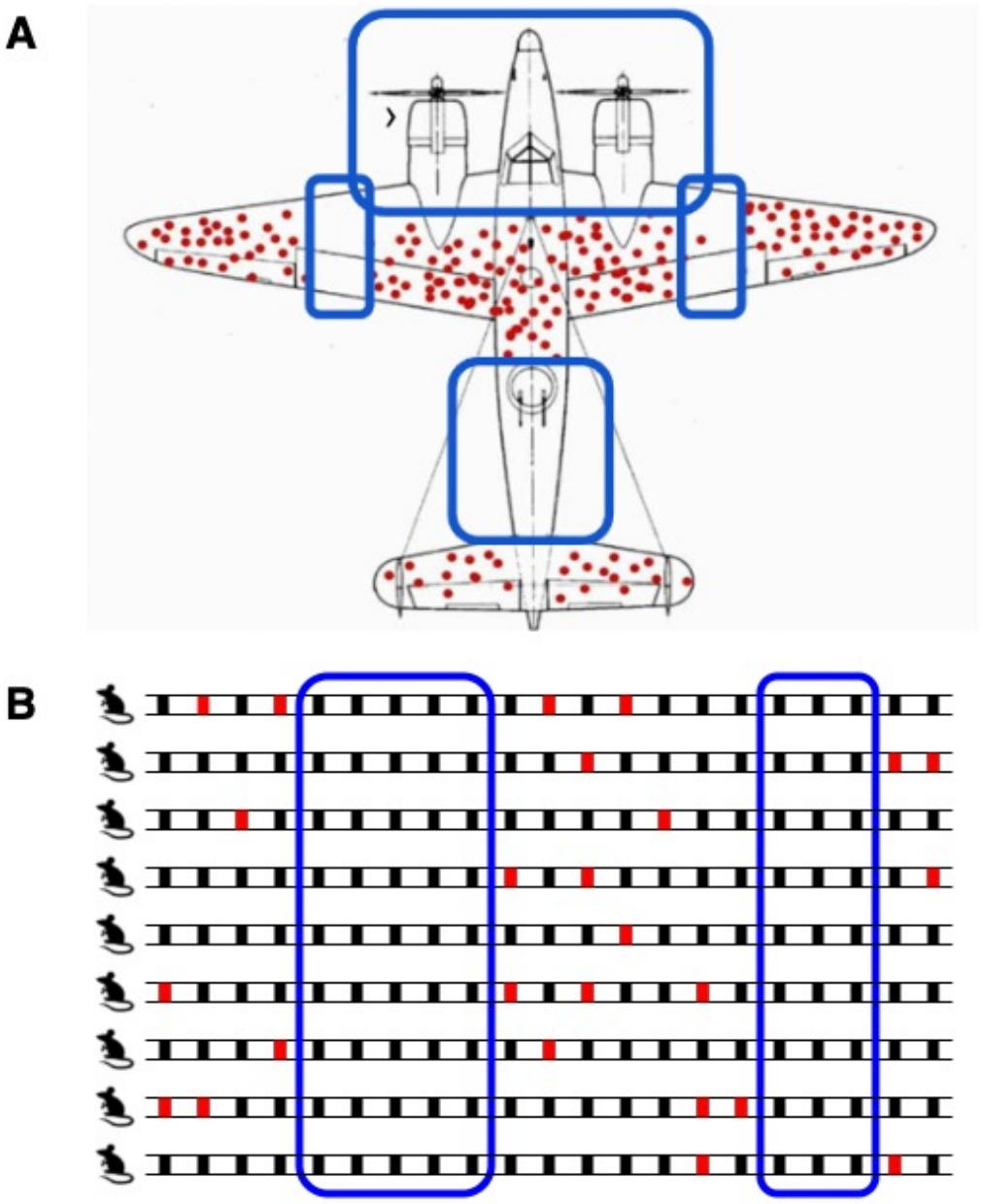


Figure 1.1 – Schematic illustration of survivorship bias and its implications for inferring regions that are intolerant of damage and variation across World War Two planes (A[2]) and species genomes (B), respectively. A) Areas on returning planes that have not taken any damage can be inferred as critical to the plane's ability to function and vulnerable to enemy fire (image credited to (Moll 2005)). Red dots indicate damage from enemy fire and blue rectangles highlight regions vulnerable to damage. B) Regions of a species genome that are relatively depleted of variation within a population can be inferred as critical to organism fitness and intolerant of mutation (i.e. constrained by negative selection). Red bars indicate variant sites along aligned DNA and blue rectangles highlight constrained loci).

[2] Image credited to (Moll 2005).

1.2 Factors that affect the distribution and frequency of variants across genomes

There are several factors in addition to negative selection that affects the distribution and frequency of variants across genomes, including the mutation rate spectrum, dominance, genetic drift, and linked selection. These factors need to be understood and/or accounted for when estimating selective constraint or interpreting selective constraint metrics. In the following subsections, I describe each of these factors, how they affect mutation rates across the genome and allele frequency changes within a population, and how they have the potential to impact or bias estimates of selective constraint.

1.2.1 The mutation rate spectrum

The mutation rate spectrum describes the frequency at which different mutation types occur at different loci across the genome. Differences in mutability are observed at different scales: from the local sequence scale (i.e. between nucleotides and their flanking contexts) to the megabase scale and between chromosomes (Gonzalez-Perez, Sabarinathan, and Lopez-Bigas 2019; Hodgkinson and Eyre-Walker 2011; Makova and Hardison 2015; Hardison et al. 2003; Matassi, Sharp, and Gautier 1999). Furthermore, the mutation rate spectrum has been shown to diverge between lineages and species over evolutionary time (K. Harris and Pritchard 2017; K. Harris 2015; Lynch 2010b). It is important to account for regional differences in mutability when estimating selective constraint, as neutrally evolving

regions (i.e. regions that are not under selection) that are less mutable are less likely to be variable. Therefore, if mutability is not accounted for, regions with relatively fewer variants could be erroneously inferred as constrained when the actual cause of this depletion is lower mutability.

Mutability is dependent on the interplay between the probability of DNA damage that causes a change to the nucleotide sequence (hereafter DNA damage) and the probability of its repair and faithful replication (Figure 2). DNA damage can have endogenous causes including replication error, spontaneous deamination, or oxidative damage; and exogenous causes such as harmful chemicals, or ionising radiation (Gonzalez-Perez, Sabarinathan, and Lopez-Bigas 2019; Poetsch, Boulton, and Luscombe 2018). Organisms have evolved complex mechanisms, involving hundreds of genes in eukaryotes (Lindahl and Wood 1999; Ségurel, Wyman, and Przeworski 2014), for repairing DNA damage to maintain its fidelity (Friedberg 2003; Chatterjee and Walker 2017). These repair mechanisms are active throughout different stages of the cell cycle and include base excision repair, nucleotide excision repair, mismatch repair, homologous recombination, and non-homologous end-joining (Friedberg 2003; Scully et al. 2019). Furthermore, translesion synthesis polymerases enable the replication of unrepaired DNA by bypassing the damaged section (Friedberg 2003; Wyrick and Roberts 2015). The probability of DNA damage, its repair and faithful replication vary across the genome, and the mutation rate spectrum results from the interplay between these probabilities.

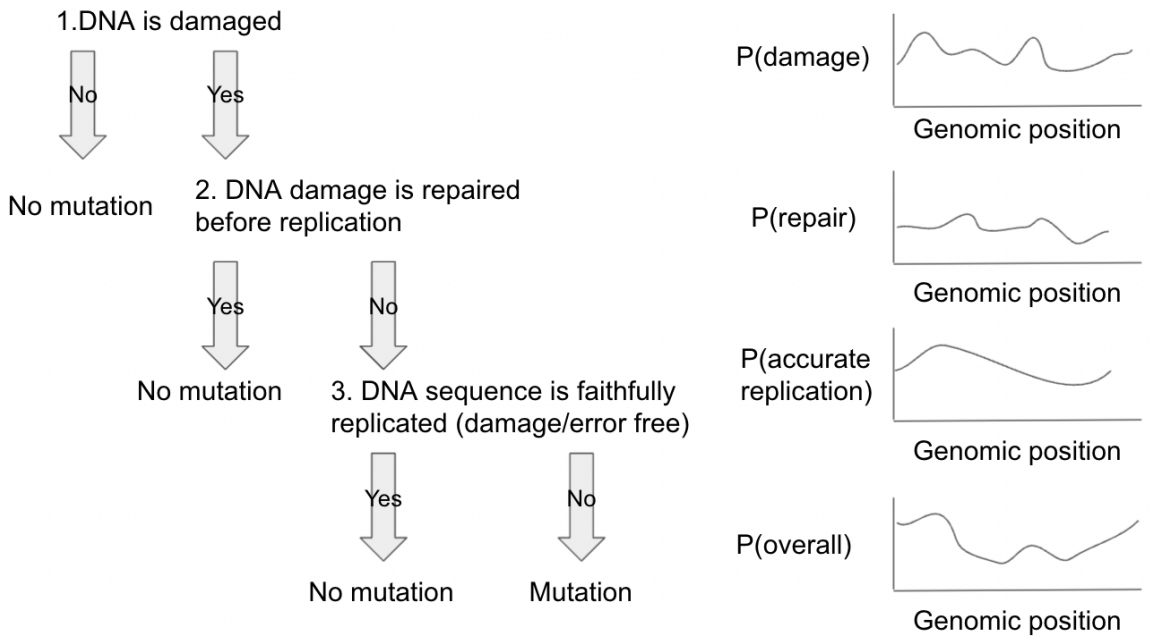


Figure 1.2 – There are three steps involved in mutation occurrence, and the probability of occurrence for each step affects the overall probability of mutation. First, DNA is damaged, causing a change to the nucleotide sequence, by either an exogenous or endogenous agent, or a replication error. Second, DNA damage repair mechanisms such as base excision repair, nucleotide excision repair, mismatch repair, homologous recombination, and non-homologous end joining, either return the DNA to its original state (leaving no mutation) or fail to correct the nucleotide sequence alteration. Finally, if the damage is not repaired, during DNA replication the damaged region is either bypassed by translesion synthesis polymerases or faithfully replicated by error-prone polymerases leaving no mutation, or incorrectly replicated resulting in a mutation. The probability of mutation at a given locus is therefore a function of its probabilities of damage, repair, and faithful replication.

Mutation rates correlate with several factors across the genome from the local to megabase scale, and even between chromosomes (Table 1.1). At the local scale (i.e. < 100 bp), these include the chemical properties of bases and their flanking nucleotides (i.e. the local sequence context), epigenetic modifications, and local chromatin features (Hodgkinson and Eyre-Walker 2011; Gonzalez-Perez, Sabarinathan, and Lopez-Bigas 2019). At the megabase scale, mutability is correlated with chromatin structure, replication time, distance to telomeres, and recombination rate, among

others (Makova and Hardison 2015; Schuster-Böckler and Lehner 2012). The greatest variation in mutability is at the local scale. For example, C and G nucleotides are approximately two to six times more mutable than A and T nucleotides (Lynch 2010a; Bulmer 1986; Gojobori, Li, and Graur 1982) and the probability of substitution for any given base differs up to 65-fold depending on the composition of its flanking nucleotides (Hwang and Green 2004; Aggarwala and Voight 2016). This is independent of CG dinucleotides which are approximately 10-fold enriched for C>T substitutions as a result of methylation and spontaneous cytosine deamination (J. C. Shen, Rideout, and Jones 1994; Duncan and Miller 1980; Coulondre et al. 1978). In comparison, across the human genome, most regions at the megabase scale have a mutation rate within two-fold of each other (Chimpanzee Sequencing and Analysis Consortium 2005). The causes of variation in mutability are difficult to infer due to collinearity between explanatory variables (Table 1.1). However, the majority of germline mutations occur during DNA replication during spermatogenesis, and multivariate analysis suggests replication time is the best predictor of germline mutability at the megabase scale (C.-L. Chen et al. 2010).

In summary, regional differences in mutability can confound estimates of selective constraint if not controlled for. Mutability varies across genomes as a product of the probability of damage and the probability of repair and faithful replication. These probabilities are dynamic and hard to dissect because they are affected by multiple correlated factors at the local and megabase scale (Table 1.1). Furthermore, they are liable to change over evolutionary time as repair mechanisms and the genomic landscape diverge.

Table 1.1 – Factors correlated with mutation rates.

Factors	Regional-scale	Mutation rate correlation	Description
GC content	Local	Positive	Cytosine (C) and guanine (G) are approximately two to six times more mutable than adenine (A) and thymine (T) (Lynch 2010a; Bulmer 1986; Gojobori, Li, and Graur 1982). This is hypothesised to result from decreased efficiency of exonuclease activity in separating C-G base pairs relative to A-T in order to excise erroneous nucleotides because A-T base pairs are bound by two hydrogen bonds whereas C-G base pairs are bound by three (Petruska and Goodman 1985; Hardison et al. 2003).
Flanking nucleotides	Local	NA	The probability of substitution for any given base differs up to 65-fold depending on the composition of its flanking nucleotides (Hwang and Green 2004; Aggarwala and Voight 2016). In human populations, the 7-mer context (i.e. the three flanking bases on either side of the base of interest) explains ~81% of the variability in substitution occurrence, more than either the 3-mer or 5-mer contexts (Aggarwala and Voight 2016).
CG methylation	Local	Positive	Methylated CG dinucleotides are ~10-fold enriched for C>T substitutions as a result of spontaneous cytosine deamination (J. C. Shen, Rideout, and Jones 1994; Duncan and Miller 1980).
Indels	Local	Positive	The substitution rate is increased ~6-fold within 50bp of an indel in the human lineage, with larger indels having greater effects (Tian et al. 2008; McDonald et al. 2011). It remains unclear whether this association is due to a shared cause, or whether one increases the probability of the other.
Replication time	Mb/kb	Positive	There is a positive correlation between replication time (i.e. when in the S-phase DNA is replicated) and mutability, with regions that are last to replicate having ~28% higher substitution rates than the earliest replicating regions across the human genome (C.-L. Chen et al. 2010), and ~10% higher across the mouse genome (Pink and Hurst 2010)). The cause of this correlation remains unclear. However, it is suggested to result from the relative decrease in time for mismatch repair of late replicating regions (C.-L. Chen et al. 2010), and the relative increase in time late replicating regions are

			single-stranded, as replication slows due to the depletion of free nucleotides (Stamatoyannopoulos et al. 2009).
Recombination rate	Mb/kb	Positive	There is a positive correlation between recombination rate and mutability, with regions that have the highest recombination rates having substitution rates that are ~50% higher than regions with the lowest recombination rates (Duret and Arndt 2008). This positive correlation is hypothesised to result from the mutagenic nature of recombination (Hellmann et al. 2005; Martin J. Lercher and Hurst 2002), and GC biased gene conversion (where GC mismatches are more likely to be repaired than AT mismatches during meiotic recombination (Kostka et al. 2012; Duret and Arndt 2008)). However, enrichment of variation at recombination hotspots could also result from linked selection (discussed in the following section).
Distance to telomere	Mb/kb	Negative	Distance to telomeres is negatively correlated with mutability (i.e. mutation rate increases closer to telomeres) (Makova and Hardison 2015). However, distance to telomeres is also correlated with recombination rate and GC content. This collinearity makes it difficult to infer the underlying cause of the correlation with mutability.
Sex chromosomes vs autosomes	Inter-chromosomal	X < autosomes < Y	The male germline accumulates more mutations from DNA replication errors compared with the female germline, as there are more cell divisions in spermatogenesis compared with oogenesis (this is known as male mutation bias) (M. J. Lercher, Williams, and Hurst 2001). The mutation rate of chromosomes, therefore, reflects the proportion of each chromosome found in males across the population. Y chromosomes have the highest mutation rate as they only occur in males. In comparison, ~50% of the autosomes are found in males, and ~33% of X chromosomes are found in males, resulting in proportionally lower mutation rates (Chimpanzee Sequencing and Analysis Consortium 2005; Ebersberger et al. 2002).
Chromatin structure	Local to Mb	Positive and negative	Mutability is correlated with chromatin structures from the megabase scale (open and closed chromatin) to the local scale (nucleosomes and linkers) (Makova and Hardison 2015; Gonzalez-Perez, Sabarinathan, and Lopez-Bigas 2019; Polak et al. 2015). This is hypothesised to result from open chromatin having increased exposure to mutagens, but conversely increased accessibility for repair mechanisms.

			However, open and closed chromatin correlates with replication time and distance to telomeres and were not significant in a multivariate model assessing mutation rates (C.-L. Chen et al. 2010).
Transcription/ exon density	Mb/kb	Minimal	At the Mb/kb scale probabilities of damage and repair (and therefore mutation) are affected by transcription and therefore associated with exon density. This is a consequence of two conflicting phenomena. The probability of damage is predicted to increase when DNA becomes single-stranded during transcription as single-stranded DNA is less stable and more accessible to mutagens (Hodgkinson and Eyre-Walker 2011; Makova and Hardison 2015) than double-stranded DNA. However, the probability of repair is also increased by transcription-coupled repair pathways (Hanawalt and Spivak 2008).

1.2.2 Dominance and selection

In diploid organisms, the impact an allele has on fitness is described by its selection and dominance coefficients (Fuller et al. 2019). For example, consider the fitness of the three diploid genotypes with wild-type allele (A) and deleterious allele (D), where h is the dominance coefficient and s is the selection coefficient:

Genotype:	AA	AD	DD
Fitness:	1	1-hs	1-s

The selection coefficient (s) describes the difference in fitness between the homozygous deleterious allele (DD) and the homozygous wild-type allele (AA). Deleterious selection coefficients range $0 < s <= 1$, where a selection coefficient of 0 would be neutral (i.e. equal to the wild-type and therefore would not affect fitness)

and a selection coefficient of 1 would be maximally deleterious (i.e. it would prevent reproduction in all homozygous carriers). For genotype AD, the effect on fitness is 1 minus the product of the dominance and selection coefficients. Dominance coefficients range 0 <= h <= 1, where a dominance coefficient of 0 represents complete dominance of allele A, a dominance coefficient of 1 represents complete dominance of allele D, and a dominance coefficient greater than 0 but less than 1 represents incomplete dominance (e.g. a dominance coefficient of 0.5 would have an effect on fitness halfway between the two homozygotes).

Dominance has important implications for estimates of selective constraint as loci can be constrained against homozygous variants, but completely unconstrained against heterozygous variants that are recessive (i.e. have a dominance coefficient of 0) (Fuller et al. 2019). While deleterious recessive variants are unlikely to reach fixation, they are more likely to persist within populations than deleterious dominant variants.

1.2.3 Linked selection

Natural selection affects the patterns of variation at sites in genetic linkage with alleles under direct selection (Figure 1.3). This phenomenon, termed linked selection, includes both selective sweeps that are a consequence of positive selection (J. M. Smith and Haigh 1974), and background selection that results from negative selection (B. Charlesworth, Morgan, and Charlesworth 1993). Both selective sweeps and background selection (described in more detail below) reduce genetic diversity at loci that are in genetic linkage with the allele under selection and therefore have the

potential to bias estimates of selective constraint if the reduction in variation is erroneously inferred to result from direct negative selection.

A selective sweep describes the rapid increase in the frequency of an allele that is under positive selection (J. M. Smith and Haigh 1974; Stephan 2019). Selective sweeps reduce variation across loci that are linked to the allele that is under positive selection, because neutral alleles that are in genetic linkage with the allele under selection are more likely to 'hitchhike' along and therefore proliferate, in a phenomenon known as genetic hitchhiking (J. M. Smith and Haigh 1974). Selective sweeps can be categorised as either hard or soft depending on the number of haplotypes the beneficial allele is on when the sweep begins (Pritchard, Pickrell, and Coop 2010; Cutter and Payseur 2013) (Figure 1.3). Hard or 'classic' sweeps describe when a *de novo* mutation creates a favourable allele that quickly spreads to fixation as a result of positive selection. Hard sweeps result in the greatest reduction in variation across linked sites because selection on the new allele started when it was in one individual, and therefore on one haplotype (Figure 1.3). In comparison, soft sweeps describe scenarios where the allele or alleles under selection are on different genomic backgrounds and therefore there is a smaller reduction in variation when they reach fixation (Figure 1.3). Soft sweeps can occur when an environmental change results in positive selection on a previously neutral allele (i.e. standing variation) that is present in the population across multiple haplotypes, or when multiple mutations creating favourable alleles occur simultaneously in different individuals and therefore on different haplotypes (which is more likely to occur in larger populations). Both of these soft sweep scenarios can involve more than one allele under positive selection if the alleles are similarly advantageous. The soft sweep completes when the sum of the advantageous allele frequencies equals one

(Pritchard, Pickrell, and Coop 2010). It should be noted that selective sweeps are not the only consequence of positive selection on a genome. Most traits are polygenic, comprising tens, if not hundreds or more loci that each make a small contribution towards a phenotype. Selection on polygenic traits results in polygenic adaptation, where there is weaker selection across individual alleles and subsequently across linked sites (Barghi, Hermisson, and Schlötterer 2020; Pritchard and Di Rienzo 2010; Cutter and Payseur 2013).

On the other hand, background selection describes the reduction in variation across regions that are in genetic linkage with loci constrained by negative selection (B. Charlesworth, Morgan, and Charlesworth 1993) (Figure 3). For example, a neutral variant that is not linked with any loci under selection will increase or decrease in frequency as a result of genetic drift (described in the subsection below). By comparison, a neutral variant that is in genetic linkage with a locus constrained by negative selection is less likely to increase in frequency. This is because if a deleterious mutation subsequently or simultaneously occurs in the constrained locus, the individuals' haplotype is more likely to be purged from the population. Background selection, therefore, skews the frequency spectrum of variants that are linked to loci constrained by negative selection towards rarity (Tajima 1989; Comeron 2017).

The relative importance and impact of selective sweeps and background selection on patterns of variation under different circumstances remain unclear and difficult to determine (Cutter and Payseur 2013; Comeron 2017; Jensen 2014; Barghi, Hermisson, and Schlötterer 2020; Pritchard and Di Rienzo 2010). Individual instances of background selection are expected to have a smaller effect on patterns of

variation than selective sweeps, as they only remove haplotypes on which deleterious alleles reside (Stephan 2010). However, *de novo* mutations are more likely to be deleterious than beneficial (Hodgkinson et al. 2013); therefore the increased prominence of background selection may result in a greater overall effect on the genome compared with selective sweeps (Rettelbach, Nater, and Ellegren 2019). Both types of linked selection (selective sweeps and background selection) are expected to contribute to the well established positive correlation between recombination rate and genetic diversity (Begun and Aquadro 1992; Cutter and Payseur 2013; Pouyet et al. 2018). This is because the diversity reducing effects of linked selection are greater in regions with low recombination rates, as the loci under selection are in genetic linkage with physically larger regions.

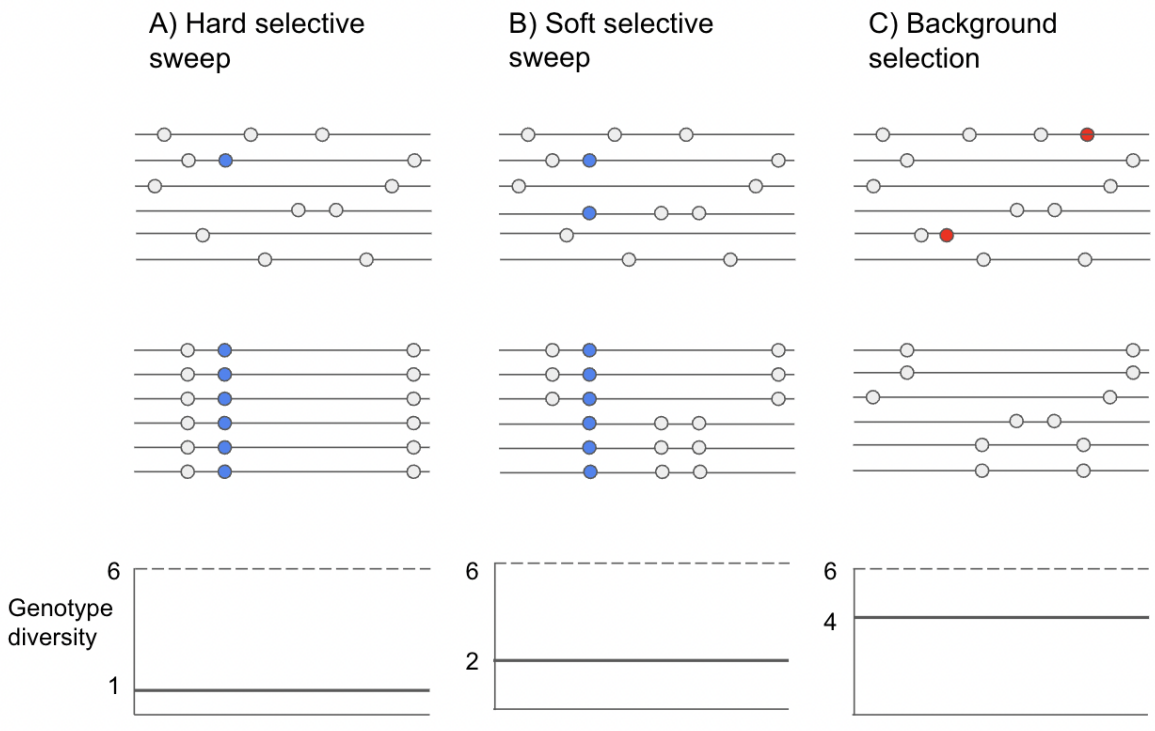


Figure 1.3 – Schematic highlighting the reduction in variation caused by hard (A) and soft (B) selective sweeps, and background selection (C). In a population of six individuals with six different haplotypes, a beneficial mutation (blue dot) that quickly spreads to fixation replacing all other haplotypes will reduce all variation at linked

sites (A). If a beneficial mutation spontaneously appears on two haplotypes and quickly spreads to fixation, variation will be reduced to two haplotypes (B). If deleterious mutations (red dots) appear the individual and haplotype will be removed from the population thereby reducing variation.

1.2.4 Genetic drift and effective population size

In addition to selection, changes in allele frequencies within populations are caused by genetic drift (Brian Charlesworth 2009). Genetic drift describes stochastic changes in allele frequencies that result from the random sampling of variants (i.e. sampling error) from one generation to the next in a finite population. In the absence of selection (i.e. neutral evolution), the frequency of alleles fluctuates across generations as a result of drift, with some alleles eventually removed from the population and others reaching fixation. The relative effectiveness of selection and drift are determined by the effective population size, a term that refers to the size of an idealised population that would lose heterozygosity at the same rate as the census population it describes (Husemann et al. 2016). The smaller the effective population size, the greater the effect of genetic drift as the sampling error from generation to generation increases, and the lower the relative efficiency of selection (Gravel 2016; Henn et al. 2015). Therefore, populations with small effective sample sizes are more likely to have mildly deleterious alleles fixed by genetic drift. This is particularly relevant for populations that have experienced a bottleneck or been recently founded (founder effect).

1.3 Selective constraint across the human genome

Several methods and corresponding metrics have been developed to quantify selective constraint across the human genome, including for protein-coding genes and non-coding sequences. These metrics have been used to distinguish genes and loci that are essential to human fitness, profile variant sets, and prioritise candidate mutations that are likely to have a causative role in human disease (Bartha et al. 2018; Karczewski et al. 2020). The premise of each method is to quantify the difference between the observed amount of variation across a locus with the expected amount of variation given neutral drift (i.e. no selection). Loci with relatively fewer variants than expected given neutrality are considered more constrained.

There are methodological differences between metrics that result in differences in the information captured. Methodological considerations include the type and zygosity of variants, and how to account for differences in sequence mutability. For example, constraint can be estimated across protein-coding genes for different variant classes (e.g. missense, or complete loss-of-function variants) and different zygosities (i.e. selection against heterozygotes or homozygotes). These differences are often dictated by the source of variation used to estimate constraint (Table 1.2). Early methods assess patterns of fixed variation between humans and other species, with the resulting metrics quantifying constraint against homozygous variation. In comparison, continually increasing sample sizes of whole-exome and whole-genome sequences within human populations have enabled increasingly powerful estimates of constraint against rare heterozygous variation, and predicted complete loss-of-function variation across and within protein-coding genes. In this section, I

outline and discuss metrics quantifying selective constraint across the human genome including protein-coding genes and non-coding sequences, the functional and clinical correlations of these metrics, and their limitations.

Table 1.2 – Sources of variation and their methodological implications for constraint metrics.

Source of variation	**Inter-species**	**Intra-species population sample**
Number of variants	Increases as a function of divergence time. More variation is expected the more distantly diverged the species are.	Increases as a function of divergence time and sample size (i.e. the number of sequenced individuals). More variation is expected from greater sample sizes that include more distantly diverged individuals.
Zygosity of variants	Variants are expected to have reached fixation in each lineage; therefore homozygous.	Variants fall on the allele frequency spectrum. 'Common' alleles with a greater frequency are more likely to appear as homozygous in the population, and 'rare' alleles with a lower frequency are more likely to be heterozygous. The greater the allele frequency the fewer the total number of alleles with that frequency within the population; therefore, there is more power to estimate selective constraint by considering rare, heterozygous variants.
Type of variation	In translated sequences, very few complete loss-of-function variants reach fixation.	A greater number of complete loss-of-function variants in translated sequences occur heterozygously due to

		dominance. This can enable the assessment of constraint against heterozygous complete loss-of-function variants.
Limitations	The more distantly diverged the species are, the more likely it is that constrained regions may have changed over evolutionary time. This is particularly relevant for regulatory features as they have a high evolutionary turnover rate compared with protein-coding sequences.	Requires a large number of individuals to be sequenced in order to be sufficiently powered – particularly with regards to constraint against homozygous variants. Recent population bottlenecks or founder effects that reduce the effective population size increase the probability that deleterious variants will be fixed by genetic drift. With large sample sizes substitutions at highly mutable loci, such as methylated CG dinucleotides, can reach saturation.

1.3.1 Selective constraint across human protein-coding genes

There are a number of metrics quantifying selective constraint across and within human protein-coding genes (Supplementary Table 1.1). Each method estimates selective constraint by quantifying the relative depletion of a given functional class of variants (e.g. predicted complete loss-of-function, missense) from the expected number given no selection (i.e. neutrality). The expected number of variants given neutrality is generally predicted based on the observed number of synonymous variants, which are synonyms for the same amino-acid, and therefore assumed to be

selectively neutral. However, it should be noted that different methods aim to capture different information. Notably, some quantify constraint against missense variation whilst others quantify selection against variants predicted to cause complete loss-of function (i.e. where the altered protein completely lacks the molecular function of the wild-type protein).

The first method to quantify selective constraint across human protein-coding genes was the dN/dS ratio (Subramanian 2011; Chimpanzee Sequencing and Analysis Consortium 2005; Kryazhimskiy and Plotkin 2008; Kimura 1977). The dN/dS ratio is frequently used in evolutionary and population genetics and quantifies selection on nonsynonymous variation as the gene-specific rate of nonsynonymous substitutions (dN) relative to the gene-specific rate of synonymous substitutions (dS) (Kryazhimskiy and Plotkin 2008). Genes with a lower rate of nonsynonymous substitutions than synonymous substitutions are considered constrained by negative selection. The dN/dS ratio was developed to quantify selection on substitutions that had reached fixation between divergent lineages (i.e. homozygous variants) (Kimura 1977; Goldman and Yang 1994; Muse and Gaut 1994) and was calculated for human genes using the human and chimpanzee reference genomes (Subramanian 2011; Chimpanzee Sequencing and Analysis Consortium 2005; Kryazhimskiy and Plotkin 2008; Kimura 1977).

Advances in genomic sequencing technologies subsequently enabled the assessment of selective constraint by using standing variation within human populations. The first metric to leverage standing variation to assess genic constraint in human populations was the Residual Variance Intolerance Score (RVIS) (Petrovski et al. 2013). RVIS is calculated by regressing the observed number of nonsynonymous

variants against the total observed variants in each gene, and then using the studentised residuals to rank genes in terms of how far they deviate from the linear prediction. Allele frequency can serve as a proxy for zygosity, as more common variants are more likely to occur as homozygotes in the population. RVIS considers 'common' variants (MAF > 0.1%), increasing the probability of identifying genes that are intolerant to homozygous variation. Following RVIS, Samocha et al (2014) developed a metric for assessing selective constraint against heterozygous missense variation, named the missense *Z*-score. They used a linear model to predict the expected number of missense variants in each gene given neutrality. The model was trained on the number of synonymous variants (presumed to be selectively neutral) as a function of sequence-specific probabilities of mutation determined by local sequence context, regional (kb) mutation rate, and sequencing depth. The difference between the observed and expected number of missense variants in each gene was then quantified as a *Z*-score.

The power to quantify selective constraint against protein-truncating variants predicted to cause complete loss-of-function has increased over time, in line with the the number of sequenced exomes and therefore the number of observed variants (Karczewski et al. 2020). Complete loss-of-function variants occur at very low frequencies in healthy adult cohorts as they tend to be deleterious, affecting fitness via haploinsufficiency (i.e. when the inactivation of one copy of a gene in a diploid organism results in a discernible change to the wild-type phenotype) (Karczewski et al. 2020; Fuller et al. 2019). The most recent and highly powered measure of selective constraint against heterozygous loss-of-function variation for human protein-coding genes is the loss-of-function observed expected upper bound fraction (LOEUF)(Karczewski et al. 2020). LOEUF expands the model developed by Samocha

et al (2014) (missense *Z*-score), used to predict the expected number of variants given no selection, by accounting for methylation state when calculating sequence-specific probabilities of mutation for CG dinucleotides. LOEUF is then calculated for each gene by dividing the observed number of loss-of-function variants by the expected number predicted with the neutral mutational model (upper limit of the 90% confidence interval). LOEUF is taken as the upper limit of the 90% confidence interval around this ratio to help control for gene length, as there is less power to distinguish between the observed and expected number of variants in shorter genes.

Larger sample sizes also enabled scores that use standing variation to identify regions within translated sequences that are constrained against amino-acid sequence altering variation, such as the MPC (for Missense badness, PolyPhen-2, and constraint)(Samocha et al. 2017), the constrained coding region (CCR) score (Havrilla et al. 2018), and gene variation intolerance rank (GeVIR) (Abramovs, Brass, and Tassabehji 2019). There is regional variation in constraint within protein-coding genes, with some regions being more constrained as they are more critical to the function of the protein (Havrilla et al. 2018). Havrilla et al (2018) defined constrained coding regions (CCR) within protein-coding genes by the distance between amino-acid sequence altering variation, controlling for sequencing depth, with longer regions defined as being more constrained. GeVIR is a composite score that combines CCRs with between species conservation to quantify selective constraint (Abramovs, Brass, and Tassabehji 2019).

In summary, early methods quantifying selective constraint across protein-coding genes capture selection against homozygous nonsynonymous variants by considering

fixed variation between species. Whole-exome sequencing then enabled the assessment of selective constraint using standing variation within human populations. This started with the assessment of selection against missense variation and progressed to selection against heterozygous predicted loss-of-function variation as the number of sequenced individuals, and therefore the power to detect constraint, increased. Metrics are highly positively correlated (Bartha et al. 2018); however, they capture subtly different information which is important to consider when deciding which metric to use and report.

1.3.2 Selective constraint across human non-coding sequences

Genomic sequences that do not include translated exons (hereafter described as non-coding) constitute approximately 98% of the total human genome and play an important role in determining human traits and disease phenotypes (di Iulio et al. 2018). The non-coding genome is enriched with complex trait and disease-associated variants identified via genome-wide association studies (GWAS) (Zhang and Lupski 2015) in addition to variants associated with Mendelian disorders (Zhang and Lupski 2015; Makrythanasis and Antonarakis 2013). Non-coding variants are expected to affect phenotypes via their effect on gene expression through transcriptional or post-transcriptional regulation (Edwards et al. 2013; Watanabe et al. 2019; Gallagher and Chen-Plotkin 2018). Quantifying selective constraint across the non-coding genome, or for specific non-coding features, can help uncover critically important non-coding elements, and differentiate between candidate disease-causing mutations.

There are several metrics that quantify selective constraint across the non-coding portion of the human genome, or for specific non-coding features (Supplementary Table 1.1). Genomic Evolutionary Rate Profiling (GERP) scores the strength of selection on single nucleotides between species using multiple sequence alignments from 33 mammalian species (Davydov et al. 2010). Nucleotides are considered constrained if they have fewer substitutions than expected given neutrality. Constrained sequences, covering a total of ~7% of the human genome, are defined using a permutation test to identify significant stretches of highly constrained (high GERP score) positions (Davydov et al. 2010). The Context-Dependent Tolerance Score (CDTS) quantifies selective constraint in 10 base-pair blocks across the human genome (di Iulio et al. 2018). CDTS uses a sliding window method, leveraging SNVs from 11,257 whole-genome sequenced individuals, and controls for window mutability by adjusting for heptonucleotide context[3]. More recently, Whiffin et al (2020) assessed constraint against variants in 5`UTRs that disrupt upstream open reading frames by considering variation from 15,708 whole-genome sequences. The authors demonstrate that variants that create new upstream start codons or remove stop codons in existing upstream open reading frames are strongly constrained by negative selection and that the strength of negative selection in the 5'UTR is positively correlated with the degree of selective constraint against loss-of-function variation (measured by LOEUF) across the corresponding gene (Whiffin et al. 2020). Other metrics, such as MutationTaster2 (Schwarz et al. 2014), and CADD (Kircher et al. 2014), integrate functional annotations defined from experimental biochemical data with patterns of genomic variation, with the specific aim of predicting

[3] A recent pre-print released on BioRxiv leverages the increased sample size of gnomAD (76,156 genomes) to quantify constrant across the genome in 1kb non-overlapping windows (S. Chen et al. 2022).

pathogenicity as opposed to quantifying selection. These metrics are not discussed further here but are reviewed by Elibeck et al (2017).

1.3.3 Functional and clinical correlations of constraint metrics

The motivation for quantifying selective constraint across the human genome is to identify loci and genes that are critical to fitness. This is an important step towards understanding the functional role of different genes and genetic elements and has implications for biomedical research and clinical medicine. Constrained genes are characterised as having fewer paralogues, being more highly expressed across a greater number of tissues, and being more central in protein-protein interaction networks (Supplementary Table 1.1). Furthermore, the most constrained genes are enriched for *in-vitro* cellular lethal knockouts and the orthologues of *in-vivo* developmental lethal knockouts in mice (Supplementary Table 1.1).

In the clinical setting, the utility of constraint metrics lies in their ability to predict disease-associated genes and/or variants. Metrics that are able to accurately predict disease association are useful for prioritising variants that are more likely to be pathogenic for further research. Predictive accuracy is generally demonstrated using gold-standard sets of Mendelian disease-associated genes and/or variants. Comparing the predictive accuracy of metrics using their original publications is challenging, as metrics are generally assessed against different truth sets. However, most metrics have the greatest area under the receiver operating characteristic (ROC) curve (70-90%) for haploinsufficient genes defined in the Online Mendelian Inheritance in Man (OMIM) database; and the most constrained genes, genic and

genomic regions are enriched for pathogenic variants defined by ClinVar (Bartha et al. 2018; di Iulio et al. 2018).

Metrics are also assessed on their capacity to predict essential genes whose homozygous deletion results in a complete loss of fitness. Essential genes have been identified across the mammalian genome using knockout experiments, where a target gene is rendered inoperative and the resulting phenotype assessed (Dickinson et al. 2016). Knockout experiments have been carried out *in-vitro* in cellular assays and *in-vivo* in model organisms, particularly mice (Cacheiro et al. 2020). Selective constraint is correlated with knockout lethality in both human cell lines and mice, with the strongest correlation for metrics that assess constraint against heterozygous loss-of-function variation (Bartha et al. 2018). However, this correlation is not perfect and there are several non-mutually exclusive explanations for this. First, knockout experiments assess biallelic inactivation, whereas loss-of-function metrics quantify constraint against heterozygous inactivation. For example, if homozygous inactivation causes lethality, heterozygous inactivation could be unconstrained if the wild-type allele is dominant (Fuller et al. 2019). Second, cellular lethality may not be replicated *in-vivo*, and/or there may be divergence in knockout lethality between humans and mice (Cacheiro et al. 2020). This is supported by the lack of complete overlap in lethality between orthologous knockouts in human cell lines and mouse models (Bartha et al. 2018).

Loss-of-function constraint metrics provide an indication as to the consequences of heterozygous loss-of-function variation *in-vivo* and have the potential to inform drug development (Plenge, Scolnick, and Altshuler 2013). If a gene is tolerant of loss-of-function variation it may indicate that reduced or eliminated expression of its

protein does not result in adverse fitness consequences. Therefore, its suppression as a drug target could be safer than for highly constrained genes. For example, naturally occurring protein-truncating variation in the *CCR5* gene was used to support the development of antiviral drugs that inhibit human immunodeficiency virus entry into cells by targeting the *CCR5* protein (a chemokine receptor on white blood cells) (Esté and Telenti 2007); and drugs that target genes in which naturally occurring homozygous loss-of-function variation has been observed in the population have a greater approval rate from phase 1 clinical trials than for other targets (Narasimhan et al. 2016). Recent research, however, shows that approved drug targets appear across the constraint spectrum and are, on average, slightly more constrained than all genes (Minikel et al. 2020). For example, there are highly constrained genes, such as HMGCR and PTGS2, that are the targets of approved inhibitors (Minikel et al. 2020). The greater average constraint of drug targets relative to all genes is expected to be confounded by other variables that correlate with drug target approval and constraint such as protein family, disease association, and tissue expression (Minikel et al. 2020). Therefore, constraint status alone is insufficient to nominate or exclude potential drug targets.

1.3.4 Limitations of constraint metrics

There are methodological and clinical limitations of constraint metrics. Methodologically, the power to detect selective constraint is limited by the sample size (i.e. the number of variants) and is intrinsically linked to the length and mutability of the genetic sequences under assessment (Karczewski et al. 2020). The greater the number of variants in the sample, the greater the power to estimate the

expected number of variants given neutrality, and therefore constraint. The number of variants increases as a function of the divergence time between lineages, with more variation expected the more distantly diverged individuals or species are. This creates a trade-off, however, because the more diverged the species or individuals are, the more likely it is that selective constraint may have changed over evolutionary time due to changing environments and/or epistasis (Huber, Kim, and Lohmueller 2020). This is particularly true for scores that quantify constraint across the non-coding genome using between-species variation, as non-coding functional elements have a high evolutionary turnover rate compared with protein-coding genes (Huber, Kim, and Lohmueller 2020). Furthermore, there is more power to determine constraint in longer, more mutable sequences. When selective constraint is determined based on the difference between the observed and expected number of variants given neutrality, shorter, less mutable sequences are more likely to be erroneously classified as unconstrained (Bartha et al. 2018). This is because longer, more mutable sequences are expected to contain a greater number of neutral variants and therefore they have more scope to be depleted of variation than shorter, less mutable sequences. Current sample sizes used to estimate constraint (LOEUF, using variation from gnomAD) are still insufficiently powered to robustly detect constraint against loss-of-function variation in approximately 30% of protein-coding genes (Karczewski et al. 2020).

There are additional methodological limitations that affect estimates of constraint across the non-coding genome. First, metrics that use human variation to quantify constraint across the whole genome in sliding windows have limited resolution given current sample sizes. For example, the Context Dependent Tolerance Score (CDTS), using variation from 11,257 whole genomes, measures constraint across windows

that are 550 bp in length (di Iulio et al. 2018). Second, models for quantifying constraint across the non-coding genome are unable to distinguish between variants based on effect, as regulatory sequences do not have a code that enables variant effect prediction such as in translated sequences (e.g. synonymous or nonsynonymous).

In a clinical setting, constraint metrics are biased towards the identification of pathogenic variants that cause disease prior to reproduction, and against those that cause disease later in life. This is because selection is likely weaker or non-existent against variants that cause disease after reproduction and therefore do not affect fecundity. These late-affecting pathogenic variants can persist in populations and the 'healthy' adult cohorts used to estimate constraint (Karczewski et al. 2020). For example, pathogenic variants in *BRCA1* are causally linked to adult-onset hereditary cancer (Hall et al. 1990; Findlay et al. 2018); however, *BRCA1* is relatively unconstrained with a pLI score of 0 (ranking among the 2% of least constrained genes) and a LOEUF score of 0.92 (less constrained than the median value of 0.91).

1.4 Selective constraint across the mouse genome

Model organisms play a critical role in genetic research: elucidating genotype function and contributing toward the development of human healthcare interventions by providing a level of *in-vivo* experimental control that is unachievable in human populations (Nadeau and Auwerx 2019; Brown 2021). The mouse (*Mus musculus*) is the most commonly used mammalian model organism

(Brown 2021), favourable relative to other species because it has a preferable balance between homology with humans and practicality (e.g. ease of husbandry and captive breeding, handleability, short generation time, and large litter sizes (Phifer-Rixey and Nachman 2015)). Mice were the first mammal used to demonstrate Mendel's laws in a series of papers by Lucien Cuenot between 1902 and 1905 (Hickman and Cairns 2003). Subsequently, mice were inbred to produce genetically homologous lineages to control for genetic background in experiments, with the first inbred strain (DBA) developed in 1909 (Phifer-Rixey and Nachman 2015; Peters et al. 2007). In 2002 the mouse genome of the most commonly used inbred strain (C57BL/6) was sequenced (Mouse Genome Sequencing Consortium et al. 2002), making it the second vertebrate reference genome after the human genome was sequenced in 2001 (Lander et al. 2001). Currently, mouse models are, amongst other things, being used to elucidate mammalian gene function (Brown and Moore 2012; Dickinson et al. 2016) and gene expression regulation (ENCODE Project Consortium et al. 2020)(described in more detail in the subsections below). Technological developments such as CRISPR/Cas9, a genetic engineering method for the targeted modification of DNA sequences in living organisms (Ran et al. 2013; Cong et al. 2013), have expanded the potential of mice to model the mechanistic function of human disease-associated loci through the humanisation of the mouse genome (Zhu et al. 2019; Devoy et al. 2011) and the generation of mouse knockouts. Mouse models have been used to decipher many facets of genetics and human disease including the mechanisms underlying Mendelian disease (Brown 2021; Cacheiro et al. 2020; Fairfield et al. 2015) and complex traits (Flint and Eskin 2012; Gonzales et al. 2018; Peters et al. 2007; Churchill et al. 2004), cancer (Zitvogel et al. 2016), disease caused by pathogens including COVID-19 (Muñoz-Fontela et al. 2020), and for therapeutic drug development (Nadeau and Auwerx 2019). In this section, I outline how mouse

models are currently being used to characterise mammalian genotype function, and the motivation for estimating selective constraint across the mouse genome.

1.4.1 In-vivo mammalian gene-phenotype associations from knockout experiments

The International Mouse Phenotyping Consortium (IMPC) is an internationally coordinated organisation that aims to characterise mammalian gene function by knocking out every protein-coding gene in the mouse genome (Dickinson et al. 2016). A gene knockout describes when a target gene is made inoperative (i.e. it has complete loss of function). The IMPC generates gene knockouts using CRISPR-Cas9 to create a deletion mutation in a critical exon of the target gene. IMPC knockouts are generated with an inbred strain of mice (C57BL/6N) that are genetically homogenous to control for genetic background. Seven male and seven female homozygous mouse knockouts are created for each gene and subject to a set of standardised phenotyping tests at set timepoints postweaning and postmortem (Brown 2021; Birling et al. 2021). These phenotyping tests are detailed by the International Mouse Phenotyping Resource of Standardises Screens (IMPReSS; http://www.mousephenotype.org/impress). Homozygous knockouts that result in preweaning lethality are subject to embryonic phenotyping tests to characterise the time of lethality during gestation and the morphological development of the embryo (Dickinson et al. 2016). Phenotypes determined throughout the pipelines are annotated with standardised Mammalian Phenotype Ontology terms (C. L. Smith and Eppig 2009).

Data from the IMPC has uncovered novel mammalian genotype-phenotype associations and provided insights into human disease. As of 2022, the IMPC has knocked out and phenotyped more than a third of mouse protein-coding genes (~7,000) (Brown 2021). Analysis of the resulting data has provided novel insights into developmental phenotypes (Dickinson et al. 2016) and sexual dimorphism in mammals (Karp et al. 2017). Furthermore, it has resulted in the discovery of mouse models for human diseases, such as auditory (Bowl et al. 2017) and metabolic dysfunction (Rozman et al. 2018).

The IMPC has produced a number of holistic studies outlining mouse associations to phenotypes and human disease. Analysis from the first 3,328 knockouts showed that approximately 90% of the resulting phenotype annotations were novel (Meehan et al. 2017). Furthermore, 40.5% of knockouts that have a human orthologue associated with disease share phenotypic annotations with the human disease (Meehan et al. 2017; Brown 2021).

1.4.2 Characterisation of the mouse regulatory landscape from biochemical assays

Mouse models are being used to understand the function of the non-protein-coding (hereafter non-coding) genome. Non-coding sequences comprise the majority (~98%) of the mouse and human genomes and include elements that regulate gene expression (Cheng et al. 2014). Differences in patterns of gene expression (i.e. differences in the time, magnitude, and cellular context at which translation occurs) dictate the extreme diversity of cell types that comprise complex organisms and the

phenotypic variation between individuals and species (ENCODE Project Consortium et al. 2020). In 2009, the Mouse Encyclopedia of DNA Elements (ENCODE) project was launched with the aim of comprehensively cataloguing the functional elements across the mouse non-coding genome (Mouse ENCODE Consortium et al. 2012). The Mouse ENCODE project complements the Human ENCODE project which started in 2003 by providing a comparative genomic reference that has yielded insight into the evolution of mammalian gene expression regulation (Cheng et al. 2014; Yue et al. 2014) and enabled the characterisation of regulatory elements active during fetal development that is unachievable with human samples (Gorkin et al. 2020; He et al. 2020). Functional genomic characteristics and elements, including open chromatin states, histone marks, transcription factor binding sites, and the 3D chromatin structure of DNA, are identified via standardised experimental assays across a variety of cell and tissue types (ENCODE Project Consortium et al. 2020). This data is being used to annotate candidate regulatory elements across the mouse genome and their activity status across different cell/tissue types (ENCODE Project Consortium et al. 2020; Zerbino et al. 2015).

1.4.3 Human-mouse homology and mouse model limitations

Despite their undoubted importance as models for answering fundamental questions about mammalian genotype function, questions remain about the mouse's capacity to approximate human genotype-to-phenotype pathways due to evolutionary divergence between the two species (Perlman 2016). Humans and mice have adapted to different environmental niches since their lineages diverged from a common ancestor approximately 85 million years ago (Springer and Murphy 2007). The

resulting phenotypic differences between the species are apparent and include size, metabolic rate, life-history traits such as lifespan, diet, microbiome, and susceptibility to pathogens (Perlman 2016). These phenotypic differences are reflected in the genomes of both species (Mouse Genome Sequencing Consortium et al. 2002; Yue et al. 2014).

The human and mouse genomes are both close to 3 billion base pairs in length and contain approximately 22,000 protein-coding genes, of which about 18,000 are orthologous (i.e. genes in different species that evolved from a common ancestral gene) (Frankish et al. 2019; Zerbino et al. 2018; Mouse Genome Sequencing Consortium et al. 2002). Orthologous protein-coding genes are highly conserved between the two species with an average of 84% amino-acid sequence identity (Herrero et al. 2016). Protein-coding sequences, however, make up approximately 2% of both species' genomes and the remaining 98% that is non-coding is less conserved (Yue et al. 2014). The majority of genetic diversity within and between species is located in non-coding genomic regions. Genetic variants in non-coding regions that impact organismal phenotype are expected to do so via their effect on gene expression regulation. As of 2020, candidate cis-regulatory element (cCRE) annotations curated by ENCODE cover approximately 7.9% and 3.4% of the human and mouse genomes respectively (ENCODE Project Consortium et al. 2020). The lower coverage of mouse annotations is a result of the smaller number of cell types and tissues sampled in mice. Over two-thirds (72%) of mouse cCREs have an orthologous sequence in the human genome, classified using LiftOver (Hinrichs et al. 2006) with a minimum matched score of 0.5. This is substantially more than the null expectation of 31% calculated using randomly selected sequences of equal lengths (ENCODE Project Consortium et al. 2020). However, only about half (47%) of mouse

cCREs have an orthologous sequence that is also annotated as a cCRE in the human genome (ENCODE Project Consortium et al. 2020), and it is unreported whether these cCREs have the same classification (e.g. predicted promoter matching predicted promoter) (ENCODE Project Consortium et al. 2020). Mouse cCREs with no human orthologue are enriched in close proximity to genes associated with stress response, immunity and certain metabolic processes (ENCODE Project Consortium et al. 2020). By comparison, mouse cCREs with human orthologues are enriched in close proximity to housekeeping genes required for the maintenance of basal cellular functions (ENCODE Project Consortium et al. 2020). The genomic similarities and dissimilarities between humans and mice suggest while some genotype-to-phenotype pathways may have been largely conserved between the species, others have substantially deviated, and others still have arisen or been lost in one lineage but not the other. Furthermore, they highlight the importance of carefully interpreting research using mouse models when making inferences about human biology.

1.4.4 Selective constraint across the mouse genome

Estimating selective constraint across the mouse genome can help identify functional loci and elements and elucidate the relationships between purifying selection and gene function. Furthermore, it can help characterise functional divergence between humans and mice and build a more complete picture of how selection has shaped the genetic differences between the species. This has the potential to help interpret research using mouse models and prioritise when the mouse may be a suitable model for a given human disease variant.

Current estimates of selective constraint across the mouse genome include dN/dS ratios (Herrero et al. 2016), and Genomic Evolutionary Rate Profiling (GERP) scores (Davydov et al. 2010; Huber, Kim, and Lohmueller 2020). Both of these estimates are also calculated in the human lineage and are described in the above subsection, but in brief: dN/dS estimates selective constraint across protein-coding genes by measuring the ratio of fixed nonsynonymous to synonymous polymorphisms between two species (Kryazhimskiy and Plotkin 2008). A dN/dS score of 1 indicates neutral evolution of a gene with regards to nonsynonymous variation, whereas a score of less than 1 indicates selective constraint: the closer dN/dS is to 0 the more constrained the gene. GERP estimates selective constraint against point mutations at single nucleotide sites across a genome by comparing the distribution of polymorphisms between multiple species alignments (Davydov et al. 2010). A GERP score for a given site indicates the difference between the observed number of substitutions at the site in the phylogeny and the expected number given neutrality. A GERP score of 0 indicates there have been no more substitutions than expected given neutrality at the given site, whereas a GERP score of 3 indicates there have been 3 fewer substitutions than expected and the site is constrained (Huber, Kim, and Lohmueller 2020).

Both dN/dS and GERP scores have limitations. GERP, and usually dN/dS, are calculated from fixed variation between distantly related species. This limits their ability to accurately characterise the current degree of constraint at given loci in the species of interest, as selection coefficients can change over evolutionary time as lineages diverge (Kryazhimskiy and Plotkin 2008; Huber, Kim, and Lohmueller 2020). GERP and dN/dS will capture constrained loci that have been under continuous purifying selection between all species in an alignment; however, the

power to detect constraint is substantially reduced at sites where selection has changed between lineages (Huber, Kim, and Lohmueller 2020; Kryazhimskiy and Plotkin 2008). This is particularly relevant for GERP scores at sites within non-coding regulatory elements, as these have a high functional turnover rate due to changes in regulatory architecture (Ludwig 2002) and epistatic effects on different genetic backgrounds (Bullaughey 2011) that occur over evolutionary time. Functional turnover describes the change in function of given genomic loci over evolutionary time and can result in changes in selection. For example, GERP scores are predicted to miss approximately half of the noncoding sites across the human genome that are currently constrained by purifying selection (Huber, Kim, and Lohmueller 2020).

The limitations that arise from estimating constraint using genetic variation from distantly diverged species can be overcome by using variation from a single population or metapopulation. However, simulations of dN/dS calculated from samples drawn from a single population of conspecifics show it is relatively insensitive to the selection coefficient over shorter evolutionary timescales (Kryazhimskiy and Plotkin 2008). Furthermore, dN/dS does not account for stochastic differences in the observed number of synonymous mutations between genes which can bias estimates.

1.4.5 A source of mouse genetic diversity

Selective constraint can be estimated from patterns of genetic diversity between mice. There is a trade-off when considering sources of genetic diversity between having a high number of variants (more power to identify constrained loci) and

minimising the phylogenetic distance between individuals in the sample (less time for functional elements and selection coefficients to diverge). In the human lineage, this has been overcome by extensive sampling and deep sequencing in human populations (Karczewski et al. 2020). However, comparable levels of sequencing have not been achieved in mouse populations.

Sources of mouse genetic diversity include inbred mouse strains and wild mouse population samples (Keane et al. 2011; Harr et al. 2016). Classical inbred strains of mice contain less genetic variation than their wild ancestors due to inbreeding (Salcedo, Geraldes, and Nachman 2007; Goios et al. 2007). Inbreeding removes genetic variation, increases homozygosity, and increases the likelihood of fixation of deleterious alleles (see the previous subsection for more detail). By comparison, wild mice exhibit patterns of genetic diversity that have not been affected by selective breeding. This makes wild mice preferable to their inbred strain counterparts as a source of genetic variation with which to calculate selective constraint across the mouse genome.

There are several species in the *Mus* genus, but the most commonly used for genetic research are the house mouse (*Mus musculus*) and the Algerian mouse (*Mus spretus*). The house mouse (*Mus musculus*) has three major lineages classified into three subspecies: the western house mouse (*Mus musculus domesticus*), the eastern house mouse (*Mus musculus musculus*), and the southeast-Asian house mouse (*Mus musculus casteneous*) (Harr et al. 2016). The Algerian mouse (*Mus spretus*) is a close relative of the house mouse (*Mus musculus*) (Lundrigan, Jansa, and Tucker 2002). The *Mus musculus* and *Mus spretus* lineages diverged ~2 million years ago (Galtier et al. 2004; Suzuki et al. 2004), with the *Mus musculus* lineage diverging into

subspecies ~1.5 million years later (~0.5 million years ago) (Bonhomme et al. 2007). *Mus musculus* and *Mus spretus* are morphologically and behaviourally distinct; however, they have been observed to hybridise in the wild in regions surrounding the Mederteranin Sea that are occupied by both species (Orth et al. 2002). Furthermore, both species are represented amongst (Adams et al. 2015) – or contributed genetic material to (Dejager, Libert, and Montagutelli 2009) – classical laboratory and wild-derived inbred mouse strains (Keane et al. 2011).

Throughout this thesis, I only consider single nucleotide variants (SNVs) from whole-genome sequencing of 67 wild mice, made publicly available by (Harr et al. 2016). These include 27 *Mus m. doimesticus* from 4 natural populations, 22 *Mus m. musculus* from 3 natural populations, 8 *Mus spretus* from one natural population, and 10 wild *Mus m. castaneous* for which genomic data is publicly available (Harr et al. 2016).

1.5 Thesis aims and outline

The overarching aim of this thesis is to evaluate selective constraint across the mouse genome using patterns of variation from populations of wild mice. My thesis is split into four analysis chapters in which I aim to:

1. Assess the effect of local sequence context and methylation state on substitution rates across the mouse genome. Specifically, I aim to determine which sequence context model best explains substitution probabilities in the

mouse lineage, in order to parameterise models for estimating mouse-specific selective constraint in later chapters. I show that the heptonucleotide context (i.e. the composition of the three bases flanking either side) of a given base increases the predictive accuracy of substitution occurrence between chromosomes across the mouse genome. I then consider potential drivers of the variation in substitution rates at the local sequence scale, highlighting substitutions in contexts where they are more likely to be deleterious to fitness occur at a lower rate. This is consistent with the hypothesis that the mutation rate spectrum at the local sequence scale has been shaped by natural selection.

2. Develop a method for estimating selective constraint against nonsynonymous variation for mouse protein-coding genes. I use this method to identify the most constrained genes across the mouse genome and compare them with their human orthologues. I show genes that cause preweaning lethality when knocked out in mice are enriched amongst the most constrained mouse genes, and mouse gene constraint is positively correlated with pleiotropy (i.e. a greater degree of constraint is correlated with an increased number of phenotype associations). Additionally, there is a positive correlation between mouse gene constraint and the number of Mendelian pathogenic variant sites in their human orthologue.

3. Quantify the degree of sequence conservation between humans and mice for twelve categories of genomic annotation, including protein-coding genes and regulatory features. I assess the alignment of human Mendelian and complex disease-associated variant sites with the mouse genome across these

annotation categories and show greater alignment of disease-associated variant sites than expected by chance.

4. Characterise patterns of selective constraint across the entire mouse genome, including non-coding regions. I show that the most constrained regions across the mouse genome are enriched for exonic and regulatory features, as expected. Furthermore, I highlight a positive correlation between regional constraint and the number of conserved human pathogenic variant sites across the mouse genome.

Chapter 2

Quantifying the effect of local sequence context on substitution rates across the mouse genome

Results from this chapter have been submitted to BioRxiv (Powell et al. 2021). I confirm that I produced all the material with no more input/advice than would be considered appropriate or acceptable for a thesis.

2.1 Abstract

Variation in sequence mutability has important implications for evolutionary models and predicting disease occurrence, and is driven in part by evolutionary divergence in mechanisms of DNA repair. The aim of this chapter is twofold: first, to assess the effect of local sequence context on substitution rates across the mouse genome; second, to investigate the relationship between sequence mutability and selection. I show that the 7-mer context (i.e. three bases on either side of the base of interest) explains more variation in substitution rates between chromosomes across the mouse genome than either the 3-mer or 5-mer contexts. Furthermore, I show that 7-mer substitutions with the potential to cause nonsense mutations when they occur in translated sequences occur at a lower rate across the genome than 7-mer

substitutions that lack this potential. I propose that mechanisms of DNA repair have evolved to prioritise substitutions at the local-sequence scale that are more likely to be deleterious to fitness.

2.2 Introduction

Germline mutations are the basis of evolution and the cause of inherited Mendelian and complex diseases. Most mutations are either neutral or deleterious to fitness, and therefore organisms have evolved mechanisms of repairing DNA to maintain sequence fidelity (K. Harris and Pritchard 2017). These repair mechanisms are complex, involving hundreds of genes in eukaryotes (Lindahl and Wood 1999; Ségurel, Wyman, and Przeworski 2014), and have diverged between lineages over evolutionary time, resulting in variation in sequence mutability (K. Harris 2015; Kumar and Subramanian 2002; Lindsay et al. 2019). Quantifying this variation in sequence mutability is important for parameterising evolutionary models, interpreting patterns of variation, and identifying both mutational signatures and pathogenic variants (Aggarwala and Voight 2016; Samocha et al. 2014; Karczewski et al. 2020; Helleday et al. 2008).

Variation in sequence mutability is influenced by local sequence context. For example, substitution rates are affected by a base's k-mer context, which describes the composition of its flanking nucleotides (K. Harris and Pritchard 2017; Aikens, Johnson, and Voight 2019; Schroeder et al. 2016). In human populations, the 7-mer context (i.e. the three bases on either side of the base of interest) has been shown to

account for more variation in substitution rates than either the 3-mer or 5-mer contexts (Aggarwala and Voight 2016). Variation in mutability between k-mers is partly driven by the chemical properties of bases and their epigenetic state (Ségurel, Wyman, and Przeworski 2014; Gonzalez-Perez, Sabarinathan, and Lopez-Bigas 2019; Makova and Hardison 2015). For example, the higher mutability of C-G relative to A-T base pairs is partly attributed to the difference in their number of connecting hydrogen bonds (Petruska and Goodman 1985; Hardison et al. 2003; Ségurel, Wyman, and Przeworski 2014), and C > T substitutions are enriched amongst CG dinucleotides due to methylation and spontaneous cytosine deamination (J. C. Shen, Rideout, and Jones 1994; Duncan and Miller 1980). However, variation in k-mer mutability is also affected by mechanisms of DNA repair, which can diverge over evolutionary time. For example, the TCC > TTC substitution rate has increased by 50% in the human European lineage since Europeans began diverging from Asians as a result of changes to DNA repair mechanisms (K. Harris 2015). The cause of these changes remains unclear, but they are tentatively linked to differences in UV exposure. TCC > TTC is the substitution most frequently induced by UV and present in melanoma skin cancers, and it is speculated that decreased exposure to UV in the European lineage may have reduced the selection pressure on mechanisms for repairing TCC > TTC substitutions, resulting in their relative increase in the rate of occurrence (K. Harris 2015). Nevertheless, determining how selection may have affected sequence mutability through the evolution of mechanisms of DNA repair remains poorly understood and is difficult to test empirically.

As mutation rates diverge over evolutionary time, it is important to quantify sequence mutability in the mouse lineage in order to parametrise mouse-specific models for inferring genomic regions under negative selection and identify

mutational signatures and pathogenic variants (Karczewski et al. 2020; Helleday et al. 2008; di Iulio et al. 2018). In this chapter, I assess the effect of local sequence context on substitution rates across the mouse genome using genetic variation between mice sampled from wild populations (Harr et al. 2016). This chapter has two main aims. First, to assess the contribution of flanking nucleotides to the ability to predict substitution mutations across the mouse genome by examining whether there are significant differences in the substitution occurrence variation explained by models which incorporate the 3-mer, 5-mer, and 7-mer sequence context. Second, to assess the relationship between sequence mutability and selection. Specifically, I hypothesise that substitutions in sequence contexts that are more likely to be deleterious to fitness have lower base rates of occurrence. In support of this hypothesis, I show that 7-mer substitutions with the potential to cause nonsense mutations occur at a lower rate across the genome than 7-mer substitutions that lack this potential. I propose that this variation in mutability is the result of selection on mechanisms of DNA repair.

2.3 Results

2.3.1 The effect of local sequence context on substitution rates.

I assessed the effect of local sequence context on the ability to predict substitution mutations across the mouse genome. I considered five different k-mer models to account for local sequence context (Table 2.1): the 1-mer, 3-mer, 5-mer, 7-mer, and A-CG models. The 1-mer model is the simplest model and does not consider the local sequence context of the base of interest. It describes 12 substitution rates, as each of

the four bases has three possible substitutions. The 3-mer model considers one flanking nucleotide on either side of the base of interest, making a 3-mer. It describes 192 substitution rates as there are 64 possible 3-mers and the middle base in each 3-mer has three possible substitutions. The 5-mer and 7-mer models consider 2 and 3 flanking nucleotides on either side of the base of interest, respectively. The number of substitution rates that a model describes increases exponentially as the number of flanking nucleotides expands, with the 7-mer model describing 49,152 substitution rates as there are 16,384 possible 7-mers and the middle base in each 7-mer has three possible substitutions. Substitution probabilities are known to be largely driven by the presence of CG dinucleotides, which have a high mutation rate from cytosine to thymine (M. Li and Chen 2011). I, therefore, expanded the simple 1-mer model to account for CG context and named this the A-CG model. The A-CG model describes 18 substitution rates as there are 6 possible contexts (A, T, C (non-CG), C (CG), G (non-CG), and G (CG)), each with three possible substitutions. For each k-mer model, I additionally calculated the probability of any mutation as the sum of each potential substitution for each k-mer.

To estimate substitution rates I considered autosomal single nucleotide variants (SNVs) from 8 populations of wild mice (*Mus musculus* sp. and *Mus spretus*) (Harr et al. 2016) and inferred the ancestral state of each allele from alignment with the *Mus caroli* reference genome. I filtered out 75.3% of the mouse genome, including genomic regions with low sequence coverage, repetitive elements, and regions with a functional annotation such as coding and regulatory features that are likely to be under selection (see Methods for more detail). I assessed the accuracy of each k-mer model to predict substitution occurrence using train/test resampling by chromosome. In brief, fifteen autosomes were randomly selected to calculate k-mer

probabilities of substitution and the remaining four autosomes were used to test predictive accuracy as the absolute error between the observed substitution rates (test autosomes) and predicted substitution rates (training autosomes), controlling for sequence composition. This was repeated 100 times to provide the mean absolute error (MAE) and 95% confidence intervals. Each ancestral k-mer and k-mer change has a complement on the reverse strand. For example, the ancestral 3-mer TAG on the forward strand must have the complementary k-mer CTA on the reverse strand, and a TAG > TTG substitution must result in a complementary CTA > CAA substitution. I, therefore, totalled the counts of each forward and reverse k-mer and k-mer change to calculate substitution probabilities, and report the substitution probabilities for adenine and cytosine only. This results in six substitution classes under consideration (noted as A>C, A>G, A>T, C>A, C>G, and C>T), plus the occurrence of any adenine or cytosine substitution (noted as A>* and C>*).

Table 2.1 – K-mer models. K-mers are defined by their flanking nucleotides. Note that each k-mer has a complementary k-mer on the reverse strand, effectively halving the total number of k-mers and k-mer changes.

K-mer model	N possible ancestral k-mers	N possible mutant k-mers	Ancestral k-mer example		Mutant k-mer example
1-mer	4	12	**C**	>	**A**
A-CG	12	18	**C** G	>	**A** G
3-mer	64	192	T **C** G	>	T **A** G

5-mer	1,024	3,072	A T **C** G T	>	A T **A** G T
7-mer	16,384	49,152	A A T **C** G T G	>	A A T **A** G T G

Of the k-mer models considered, the 7-mer context provides the greatest increase in predictive accuracy relative to the 1-mer and A-CG models for each substitution class, with predictive accuracy significantly increasing as the number of flanking bases is expanded (Figure 2.1) ($p < 0.001$, Supplementary Figure 2.1). The 7-mer model explains 84.6% ± 0.1% (mean ± 95% confidence interval) of the total variance for cytosine substitutions (C > *), and reduces mean absolute error (MAE) by 73.7% ± 0.3% (mean ± 95% confidence interval) relative to the 1-mer model. This is largely driven by the presence of CG dinucleotides and the frequency of CG > TG substitutions, with the A-CG model reducing MAE by 67.0% ± 0.3% for cytosine substitutions relative to the 1-mer model (Figure 2.1). The 7-mer model, however, reduces MAE for cytosine substitutions by 20.5% ±0.9% relative to the A-CG model, indicating the additional nucleotides on either side of a base of interest affect its substitution rate. This pattern is consistent for adenine substitutions (A > *), with the 7-mer model explaining 77.8% ± 0.2% (mean ± 95% confidence interval) of the variance, and reducing MAE by 34.5% ±0.7% relative to the 1-mer model. The absolute variance explained for each substitution model is provided in Supplementary Table 2.2.

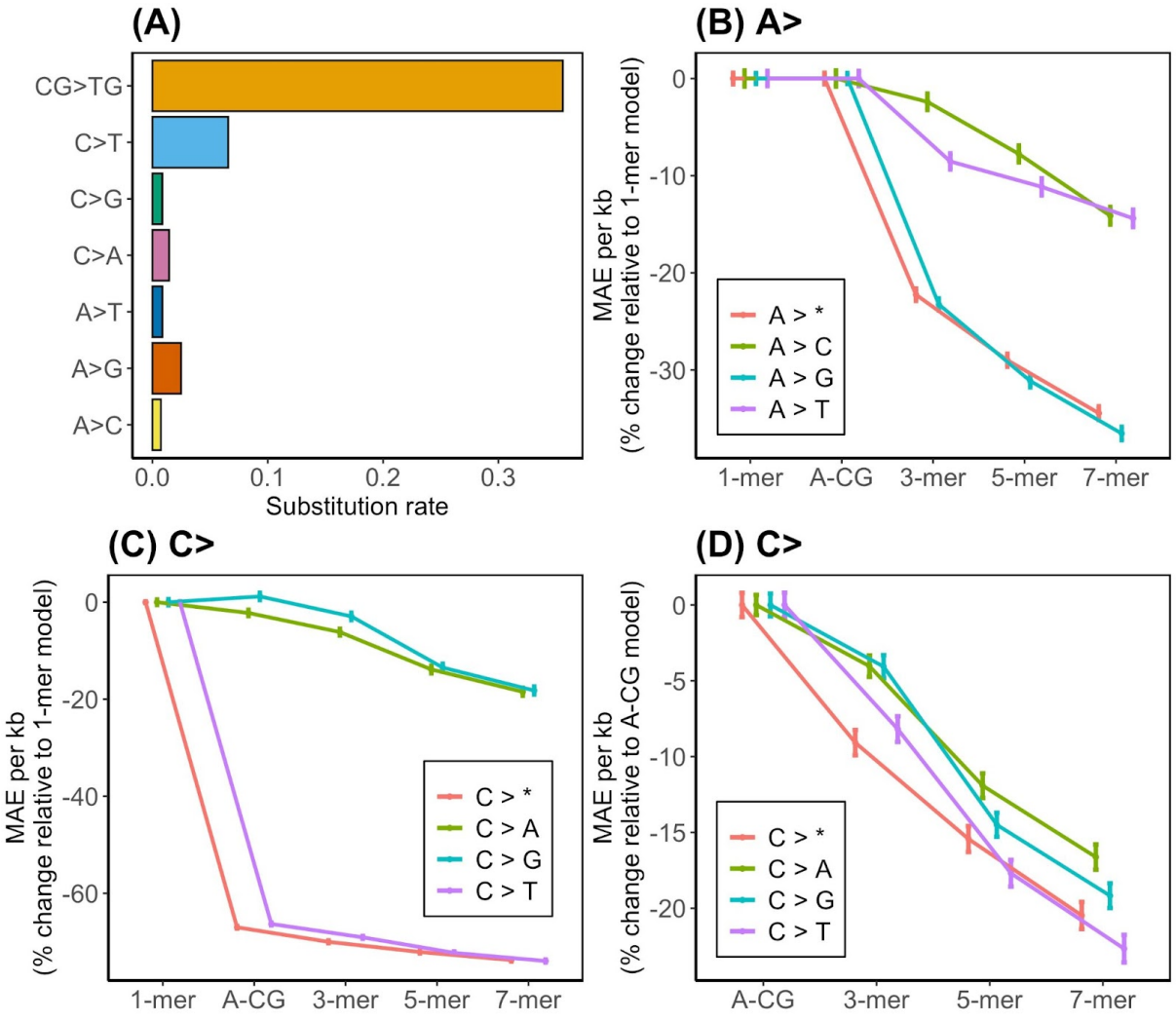


Figure 2.1 – A) 1-mer substitution rates by class. C > T substitutions occur at the greatest rate, largely driven by the mutability of CG > TG dinucleotides. B) Mean absolute error (MAE) per kb (95% confidence intervals) relative to the 1-mer model for predicting adenine substitutions. The 7-mer model provides the greatest improvement in MAE relative to the 1-mer model for all substitution classes, with the greatest improvement for A > G substitutions (36.5% ±0.7%) (mean ± 95% confidence interval). C) Mean absolute error (MAE) per kb (95% confidence intervals) relative to the 1-mer model for predicting cytosine substitutions. The A-CG model decreases MAE by 67.0% ±0.3% for C > T substitutions relative to the 1-mer model by considering the CG context. D) Mean absolute error (MAE) per kb (95% confidence intervals) relative to the A-CG model for predicting cytosine substitutions. The 7-mer model provides the greatest improvement in MAE relative to the A-CG model for all substitution classes, with the greatest improvement (22.7% ±0.9%) shown for C > T substitutions. Any potential substitution is denoted with a *.

2.3.2 Power simulations with increasing k-mer sizes

The number of k-mers (i.e. the potential combinations of A, C, G, and T) increases exponentially with k-mer length (Table 2.1). This reduces the average count for each k-mer and k-mer substitution in the sample, and therefore the statistical power to accurately determine probabilities of substitution. I, therefore, tested whether there is sufficient power to detect improvements in predictive accuracy between chromosomes, given the observed number of polymorphic sites and expanding k-mer sizes. I created two simulated datasets of substitution counts: one where the true probability of substitution for each k-mer is defined by its 3-mer context, and the other where it is defined by its 7-mer context. I then tested prediction accuracy for the different k-mer models using these simulated datasets to determine whether I would detect changes in prediction accuracy (see methods).

The simulations show that when probabilities of substitution are defined by the 3-mer context, MAE is significantly lower for each substitution class for the 3-mer model compared with the A-CG model (Figure 2.2)($p < 0.001$, paired t-tests, Supplementary Figure 2.2). MAE does not significantly differ between the 3-mer, 5-mer and 7-mer models as the additional bases do not aid prediction ($p > 0.05$, Supplementary Figure 2.2). When probabilities of substitution are defined by their 7-mer context, MAE of the 7-mer model is significantly lower than the MAE of the 3-mer and 5-mer models for each substitution class (Figure 2.2)($p < 0.05$, paired t-tests, Supplementary Figure 2.2). This indicates that there is sufficient power to detect improvements in MAE up to the 7-mer model for each substitution class, given the simulated effect size. In this chapter I do not extend the k-mer model beyond the

7-mer as there is insufficient power to distinguish improvements in MAE between the 7-mer and 9-mer models (Supplementary Figure 2.2).

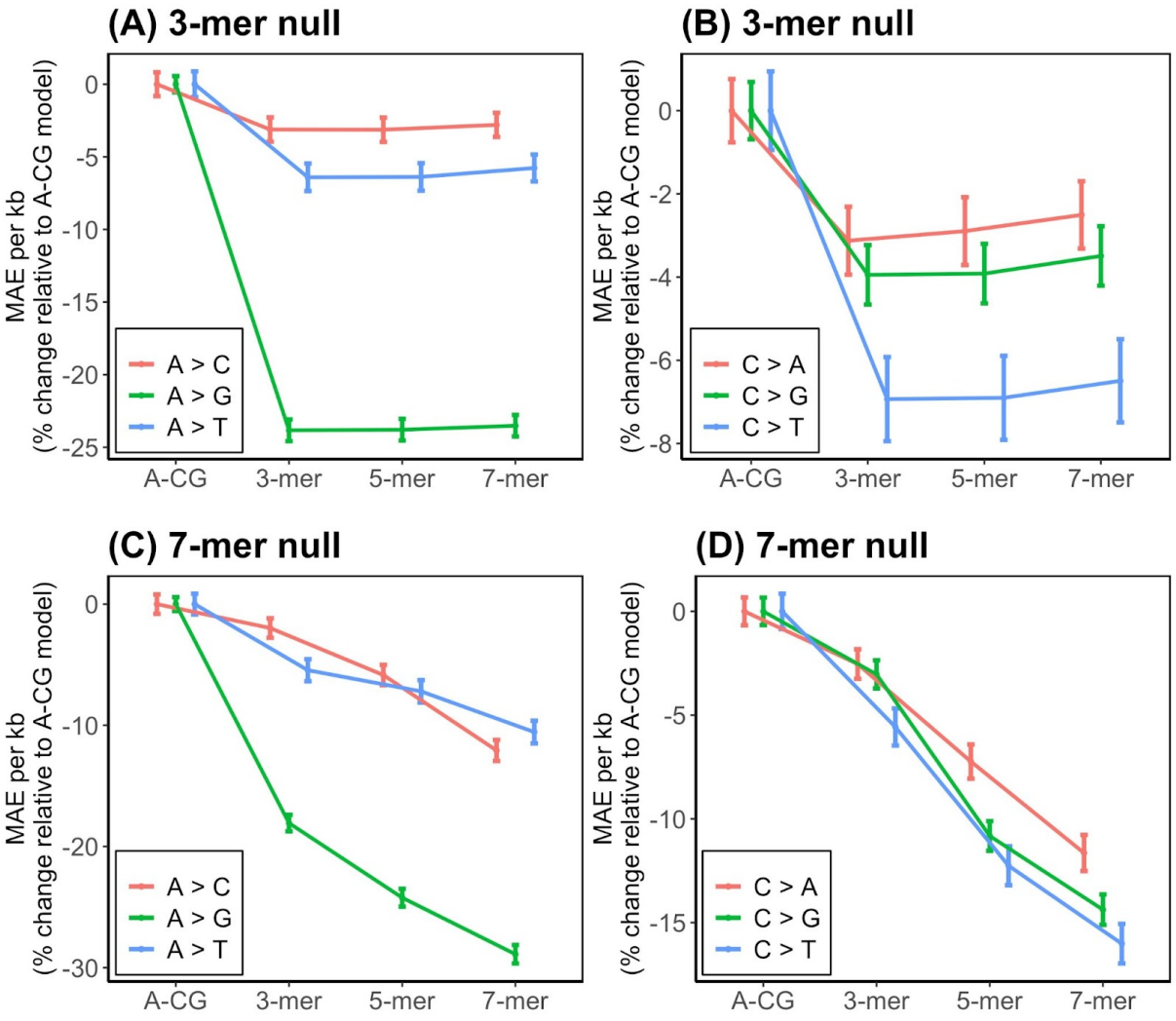


Figure 2.2 – Mean absolute error (MAE) per kb (95% confidence intervals) relative to the A-CG model given simulated probabilities of substitution determined by the 3-mer (A and B) and 7-mer (C and D) context for adenine and cytosine. When probabilities of substitution are defined by the 3-mer context, considering additional flanking bases does not improve prediction accuracy. However, when probabilities of substitution are defined by the 7-mer context there is sufficient power to detect significant increases in prediction accuracy.

2.3.4 GC content and mutability

Genomic regions with high GC content have higher mutation rates (Ségurel, Wyman, and Przeworski 2014). This is hypothesised to result from decreased efficiency of exonuclease activity in separating C-G base pairs relative to A-T in order to excise erroneous nucleotides because A-T base pairs are bound by two hydrogen bonds whereas C-G base pairs are bound by three (Petruska and Goodman 1985; Hardison et al. 2003). In order to determine the effect of GC content on substitution rates at the local sequence scale, I fit a generalised linear model to predict the number of cytosine or guanine bases in the 7-mer as a function of its substitution rate and found a significant positive relationship (estimate = 0.82±0.03, z = 25.7, p = 2.66e-167) (Figure 2.3). To test whether this relationship is inflated by CG deamination I fit another model which excluded all CG>TG substitutions and found a significant positive relationship remained (estimate = 1.90±0.12, z = 15.7, p = 1.18e-55). These results highlight that the relationship between GC content and mutability persists at a finer scale, and are consistent with the hypothesis that higher substitution rates are caused in part by a decreased efficiency of exonuclease activity in separating C-G base pairs relative to A-T in order to excise erroneous nucleotides.

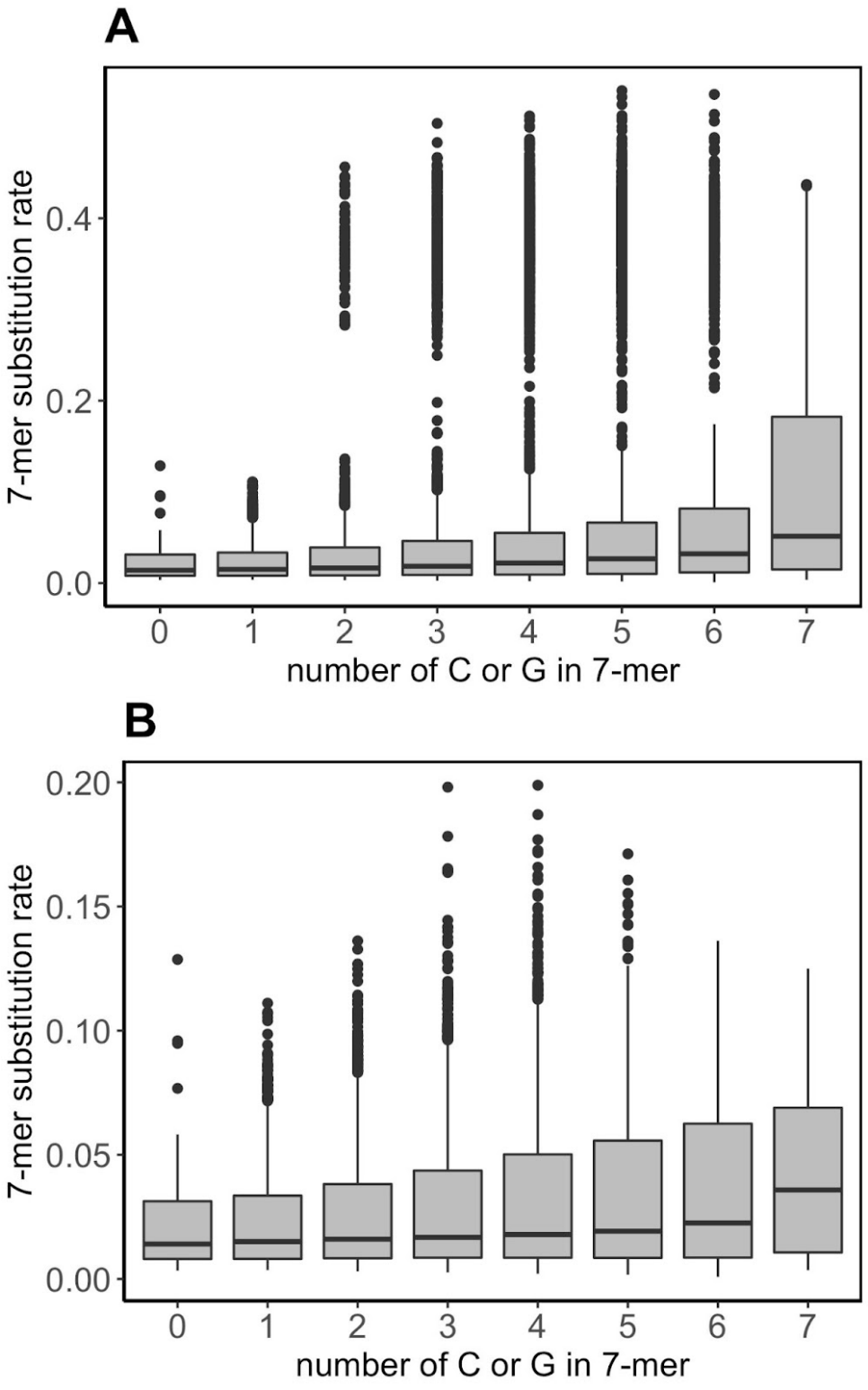


Figure 2.3 – Substitution rates of 7-mers defined by their composition of cytosine or guanine bases (A). 7-mers with more cytosine or guanine bases have on average higher substitution rates. This relationship persists when CG > TG substitutions are removed from the analysis (B).

2.3.5 The effect of methylation on CG dinucleotide substitution rates

CG dinucleotides are highly mutable across species genomes, particularly C > T substitutions resulting from deamination at methylated sites (J. C. Shen, Rideout, and Jones 1994; Duncan and Miller 1980). To assess how methylation status improves the prediction accuracy of substitution occurrence within CG dinucleotides across the mouse genome, I quantified the effect of methylation state on the CG mutability. To do this, I considered all CG dinucleotides across unmasked regions of the genome and inferred their methylation status in the germline from bisulfite sequencing in embryonic stem cells, which show a similar pattern of methylation to sperm cells in mice (Popp et al. 2010). In total, 72.9% of CGs are classified as methylated, 6.9% as unmethylated, and 20.2% were unclassified due to insufficient coverage (Supplementary Figure 2.5). The C > T substitution rate in methylated CG dinucleotides is 0.37 compared with 0.04 in unmethylated CGs, an 8.3-fold increase. By comparison, the C > T substitution rate is 0.35 across all CGs independent of methylation status. This is closer to the substitution rate for methylated CGs due to a higher proportion of methylated CGs relative to unmethylated CGs in the sample. Accounting for methylation state improves prediction accuracy for all CG substitution classes, with the effect greatest for C > T substitutions (Figure 2.4). MAE is reduced by 82.5% ±2.5% when predicting the occurrence of C > T substitutions across methylated CGs, and 99.7% ±0.1% across unmethylated CGs, relative to a null model where methylation status is not considered (see methods). The improvement in MAE is greater for unmethylated CGs relative to methylated CGs because the null substitution rate (calculated across all CGs independent of methylation status) has a

higher proportion of methylated CGs relative to unmethylated CGs. The improvement in MAE by considering methylation status is greater than using the 7-mer context alone, which improves the prediction accuracy of C > T substitutions by 22.7% ±0.9% relative to the A-CG model, further highlighting the importance of methylation state to predicting C > T substitution occurrence.

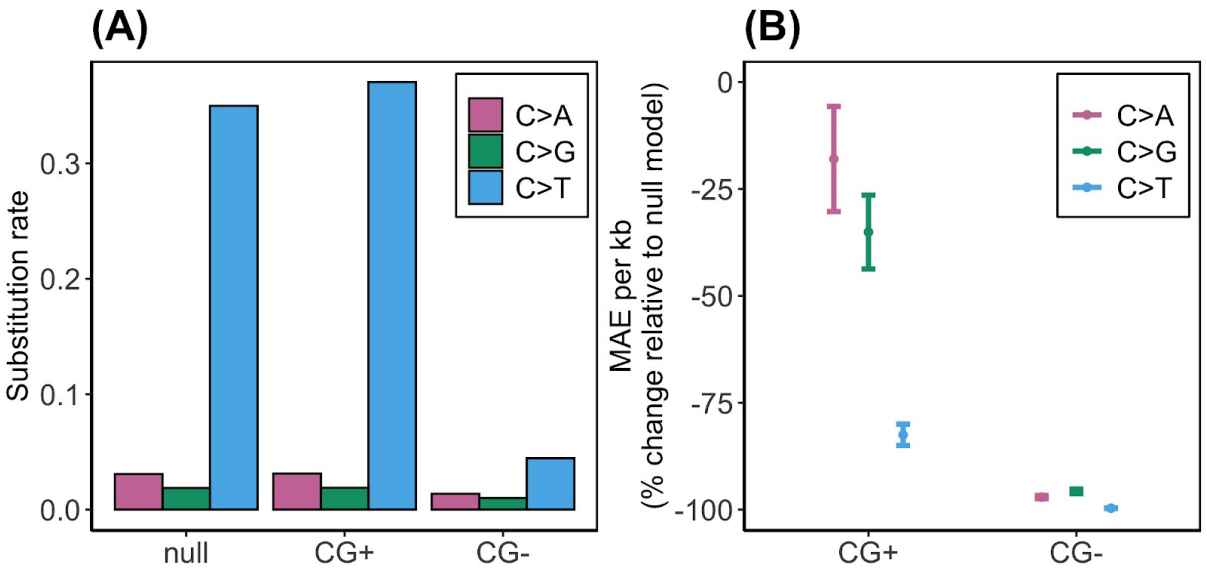


Figure 2.4 – A) Substitution rates of methylated (CG+), unmethylated (CG-), and all (null) dinucleotides by class. Methylated CG dinucleotides have a C > T substitution rate of 0.37 compared with 0.04 for unmethylated CG dinucleotides. In total, I classified 72.9% of CG dinucleotides across the unmasked genome as methylated, 6.9% are unmethylated, and 20.2% are unclassified because of low coverage. B) Mean absolute error (MAE) per kb (95% confidence intervals) relative to the null model for methylated (CG+) and unmethylated (CG-) dinucleotides. The null model determines substitution probabilities from all CG dinucleotides independent of methylation state, whereas the CG+ and CG- models use methylation state to determine substitution probabilities. Considering methylation status reduces MEA for each substitution class. The greatest decreases are for C > T substitutions, with a MAE decrease of -82.5%±0.2.5% and -99.7%±0.1% for methylated and unmethylated CG dinucleotides, respectively.

2.3.6 Evidence mutations are less frequent in nucleotide contexts where they are deleterious

If the mutation spectrum has been shaped by adaptive evolution at the local sequence scale, we would expect mutations to occur less frequently in nucleotide contexts where they are more likely to be deleterious. 7-mer context determines the protein-coding consequences of substitutions in translated sequences, thereby providing a basis for selection on mutation rate at the local sequence scale (Figure 2.5). Substitutions in some contexts but not others have the potential to cause highly deleterious nonsense mutations (Gorlov, Kimmel, and Amos 2006), so the average fitness effects of substitutions is likely to differ between 7-mers. I categorised substitutions into those that can cause nonsense mutations (hereafter "potential nonsense substitutions" (PNS)) and those that cannot (hereafter "no-nonsense substitutions" (NNS)) based on their 7-mer context alone and blind to where they occur in the genome. I found that potential nonsense substitutions have, on average, a substitution rate that is 31% lower than no-nonsense substitutions, and that this difference is statistically significant (Wilcoxen Signed-Rank Test, p = 1.52e-86)(Figure 2.6). With the exception of A>C substitutions (discussed below), this pattern remained consistent when I analysed each substitution class individually (Figure 2.6). The pattern was also maintained when I controlled for GC content, which correlates positively with mutability (Figure 2.3, Supplementary Figure 2.3).

If, as I hypothesise, mutation rates at the local sequence scale have undergone adaptive evolution, selection must have been able to overcome genetic drift to shape mutability (Hodgkinson and Eyre-Walker 2011). I tested this assertion using

simulations. Population genetics theory states that selection on mutational modifiers can overcome drift to shape mutation rates when the product of the following exceeds one: the effective population size (Ne), the per-nucleotide mutation rate (u), the magnitude of change in the mutation rate (du), the number of sites where a mutation would be under selection (Nsites), and the average selection coefficient of a mutation in one of these sites (s) (Lynch 2010b). Note that, in my simulations, Nsites equates to the number of sites in translated sequences where a substitution would cause a nonsense mutation, and s the average strength of selection against a nonsense mutation. I found that Nsites is greater than the minimum necessary for selection to overcome drift in all substitution classes other than A>C (Figure 2.6; Supplementary Figure 2.4). This is consistent with my results, as A>C substitutions are the only class in which no-nonsense and potential nonsense substitutions do not differ in their mutation rates (Figure 2.6).

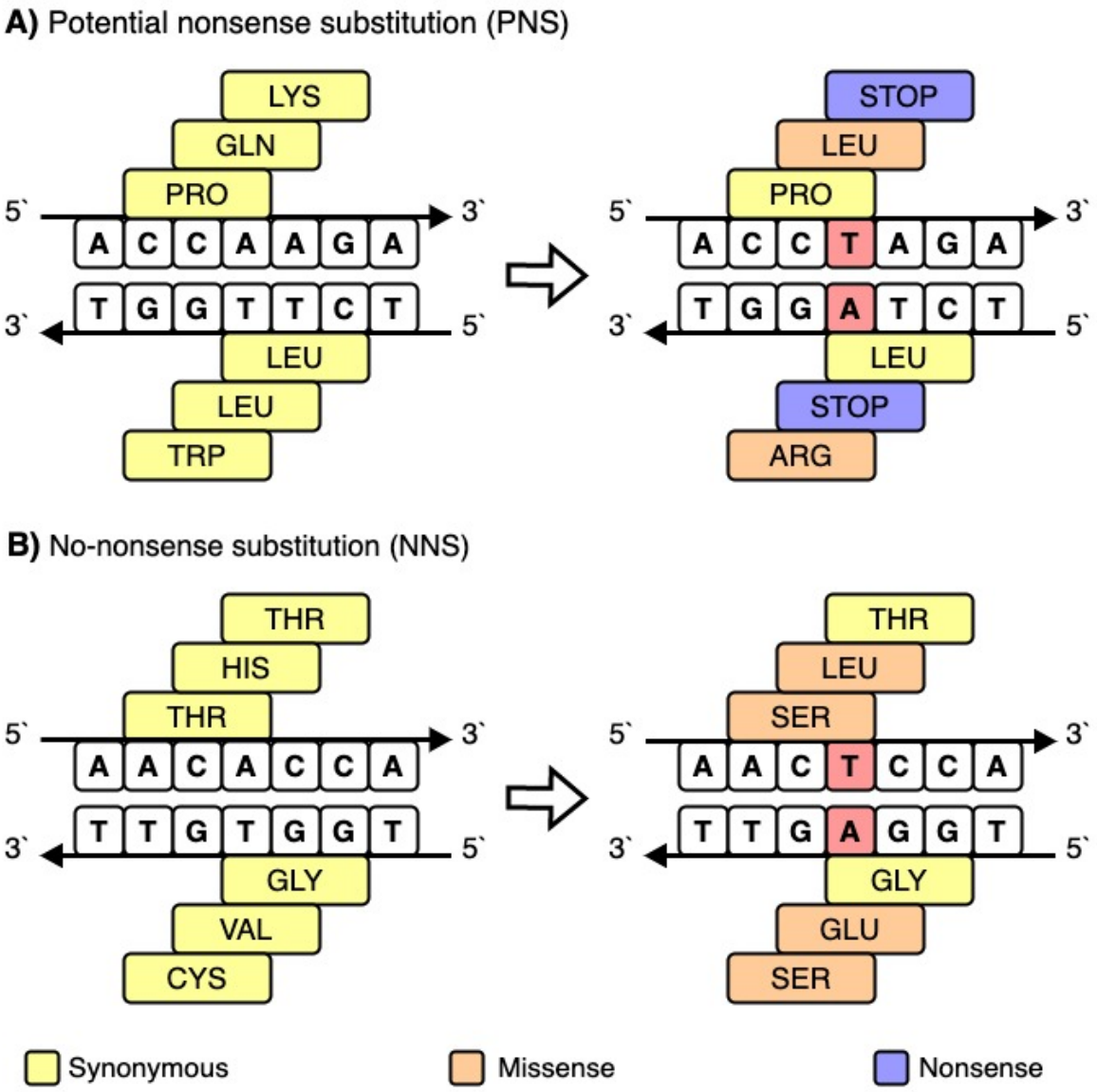


Figure 2.5 – Classification of 7-mer substitutions by their capacity to cause nonsense mutations in translated sequences. Substitution of the middle base in a given 7-mer has six potential coding consequences in translated sequences in the 5` to 3` direction: three on the forward strand and three on the complementary reverse strand. Substitutions in 7-mers that have the potential to cause one or more nonsense mutations when in translated sequences are classified as potential nonsense substitutions (PNS), otherwise, they are classified as no-nonsense substitutions (NNS). For example, an A>T substitution in the 7-mer ACCAAGA is classified as PNS because it has the potential to cause a nonsense mutation if it changes the codon AAG to TAG or the codon TTG to TAG on the reverse strand (A). By comparison, an A>T substitution in the 7-mer AACACCA is classified as NNS because it does not have the potential to cause a nonsense mutation (B).

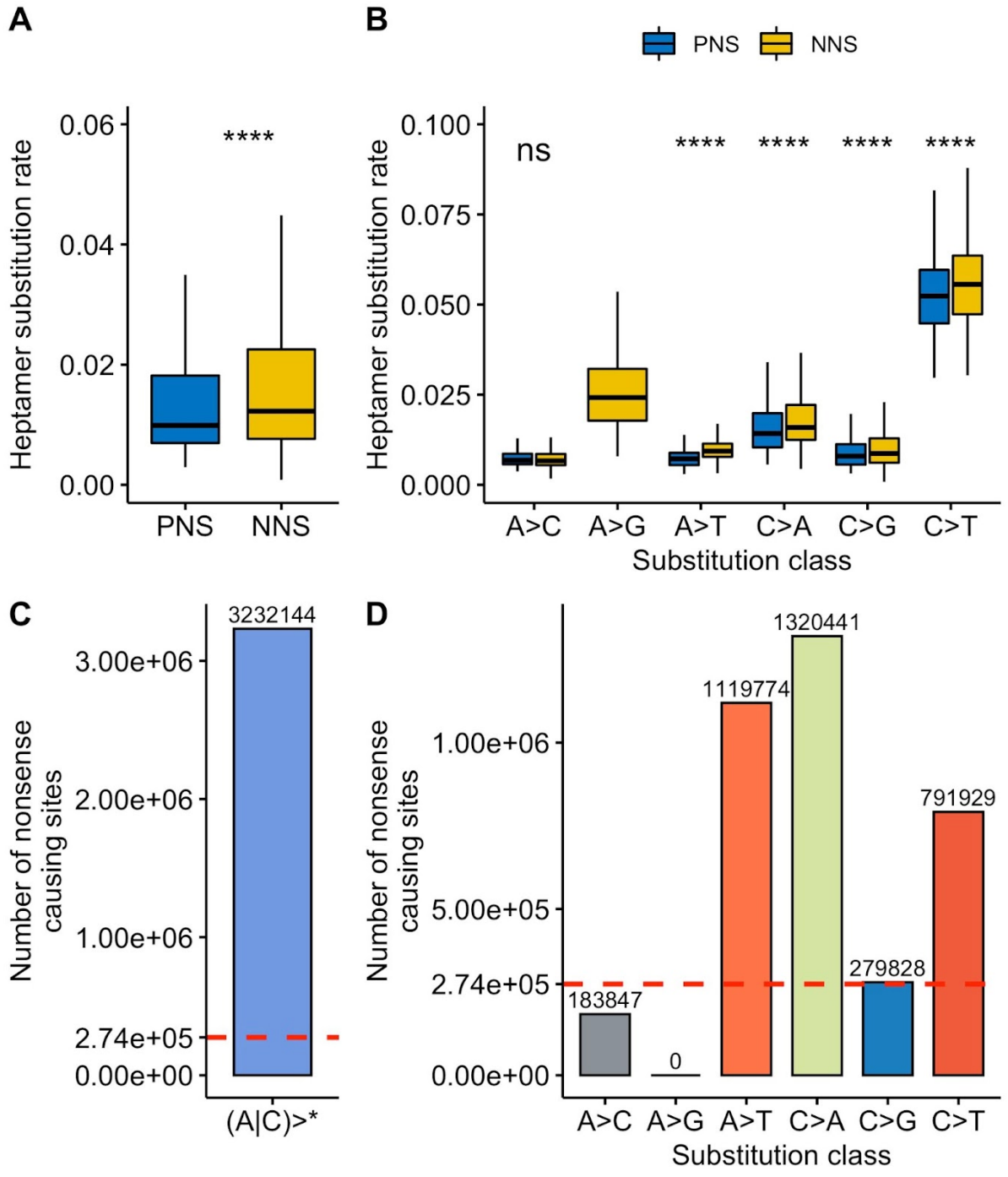


Figure 2.6 – (A) Potential nonsense substitutions (PNS) occur at a significantly lower rate than those that do not have the potential to cause nonsense mutations (i.e. no-nonsense substitutions, NNS) across the mouse genome. (B) This pattern holds across each substitution class except for A>C and could be explained by selection on mechanisms of DNA repair. (C and D) The number of sites in translated sequences in which a nonsense mutation can occur. (A|C)>* denotes any substitution class. The red line indicates the simulated minimum number of nonsense-causing sites required for selection to overcome genetic drift given a conservative effective

population size of 30,000 (Phifer-Rixey et al. 2012), a per-nucleotide substitution rate of 5.4e−9 (Uchimura et al. 2015), an order of magnitude reduction in mutation rate (Hodgkinson and Eyre-Walker 2011), and an average selection coefficient against nonsense substitutions of 0.025 (Gorlov, Kimmel, and Amos 2006). The number of sites where an A>C substitution will cause a nonsense mutation is below the simulated threshold to overcome drift (D), and this could explain the non-significant difference in substitution rates between PNC and NNC for this substitution class (B). Differences between groups were assessed with Wilcoxon tests and **** denotes a p-value <= 0.0001. Outliers, defined by being outside 1.5 times the interquartile range above the upper quartile and below the lower quartile, are not shown.

2.4 Discussion

In this chapter, I assessed the effect of local sequence context on substitution rates across the mouse genome, and the relationship between sequence mutability and selection. I have two main results: First, the three flanking nucleotides on either side of the base of interest (its 7-mer context) explain 84.6% and 77.8% of the variance in cytosine and adenine substitution rates in the mouse lineage, respectively, compared with a simple model that does not consider nucleotide context. This is 5.4% and 2.3% more than the 3-mer or 5-mer contexts for cytosine substitutions, and 54.8% and 18.8% more than the 3-mer or 5-mer contexts for adenine substitutions, respectively. Second, 7-mer substitutions that are more likely to be highly deleterious to fitness because they have the potential to cause nonsense mutations when they occur in translated sequences are less likely to occur across the genome than substitutions in contexts which lack this potential.

The importance of 7-mer context to substitution occurrence has important implications for evolutionary models. In particular, models that aim to quantify the

selective constraint from the relative enrichment or depletion of variation must control for regional differences in mutability (Samocha et al. 2014; di Iulio et al. 2018). Failure to do this may lead to false inferences of selection. For example, regions that are less mutable are likely to contain fewer variants, independent of selection. Our finding that 7-mer context improves the predictive accuracy of substitution occurrence in mice is consistent with human studies reporting similar findings (Aggarwala and Voight 2016). This suggests that the causes of 7-mer context importance have been conserved between the species. These causes remain unclear, but they could involve both fundamental properties of DNA, such as the chemical properties of different nucleotides, and DNA repair mechanisms. For example, sequence context may affect the ability of DNA polymerase to separate DNA in order to remove and replace an erroneous base. Erroneous bases that are surrounded by C or G nucleotides may be less likely to be corrected, as C-G nucleotides are joined by three hydrogen bonds and therefore require more energy to separate than A-T nucleotides, which are bound by two (Petruska and Goodman 1985; Hardison et al. 2003). The 7-mer context could account for the distance from an erroneous base by which DNA must be separated in order for repair. It remains to be determined whether the 7-mer context holds the same importance for non-mammalian species that are more distantly related to humans, such as drosophila.

Methodologically, this study assesses the accuracy of different context models to predict substitution occurrence by averaging across the full spectrum of potential contexts. There may be differences in the relative importance of different 7-mers to predict substitution occurrence that would be missed by this method. For example, if there was a difference in the likelihood of an A>C substitution in the 7-mer AAAAAAA compared with the 5-mer AAAAA, but not in the 7-mer GAAAAAG

compared with the 5-mer AAAAA, this nuance would be missed by the general model applied here. Determining which 7-mer contexts have the greatest predictive accuracy for substitution occurrence is an interesting question for future research.

Variation in sequence mutability between lineages captured by local sequence context suggests evolutionary divergence in mechanisms of DNA repair, with some lineages repairing specific substitutions more or less effectively than others (Ségurel, Wyman, and Przeworski 2014; K. Harris 2015; Kumar and Subramanian 2002). Evolutionary theory suggests the complexity of DNA proofreading and repair mechanisms in eukaryotes may limit the impact of selection on shaping the efficiency of individual genes in performing these processes (Sung et al. 2012; Lynch 2011). For example, if a variant changes the mutation rate in a given sequence context by a small amount then the effect on fitness may fall below the threshold at which selection is effective. Therefore, the fixation of the variant in the population would be dependent on genetic drift. This line of reasoning has led to the hypothesis that drift is the predominant cause of variation in mutability between species (Sung et al. 2012). I have shown, however, that 7-mer substitutions that are more likely to be deleterious to fitness (by having the potential to cause nonsense mutations in translated sequences) are less likely to occur across the genome. The genetic code that determines nonsense and synonymous mutations is universal to almost all species (Knight, Freeland, and Landweber 2001), but understanding whether selection drives further variation in mutability between lineages requires further research. For example, it would be interesting to identify k-mers known to be under lineage-specific negative selection and determine whether this is reflected in their substitution rates.

2.5 Methods

2.5.1 Data download and analysis software

I considered genetic variation between 67 mice (*Mus musculus* sp. and *Mus spretus*) sampled from 8 wild populations, described in detail by (Harr et al. 2016). Variant (vcf files) and coverage (bam files) information were downloaded from http://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/. The soft-masked mouse reference genome (GRCm38), *Mus musculus* to *Mus caroli* pairwise alignment (Herrero et al. 2016), GENCODE exonic annotations (Harrow et al. 2012), and the Ensembl Regulatory Build (Zerbino et al. 2015) were all downloaded from Ensembl ftp (ftp://ftp.ensembl.org/pub/release-101/) (v101). GERP conserved regions were calculated by Ensembl across 111 mammalian species alignments and were downloaded from Ensembl ftp (v101) (Herrero et al. 2016). Genome-wide methylation of CG dinucleotides in mouse embryonic stem cells was quantified by Stadler et al (2011) using bisulfite sequencing and was downloaded from Ensembl ftp (http://ftp.ensembl.org/pub/data_files/mus_musculus/GRCm38/dna_methylation_feature/NPC_5mC_Publication_Stadler2011_PMID22170606/) (Stadler et al. 2011). Human 7-mer substitution rates were downloaded from the supplementary information of Aggarwala and Voight (2016).

All analyses were conducted using custom scripts written in R (v3.6.2) (R Core Team 2017) and bash.

2.5.2 Filtering variants and genomic regions

I built a mask for the mouse genome (GRCm38) to filter low coverage regions, repetitive elements, and regions with functional annotation and/or are likely to be under selection. I filtered low coverage loci defined as having less than 10X coverage in more than 10% of individual mice in the sample. Read depth at each genomic position for each mouse in the sample was calculated from the bam coverage files using samtools (H. Li et al. 2009). I filtered repetitive elements and ambiguous (N) bases by excluding all soft-masked regions classified using RepeatMasker by Ensembl (Zerbino et al. 2018). I also filtered regions likely to be under selection. These included functional regions annotated as either “promoter”, “promoter flanking”, “enhancer”, “CTCF binding site”, “TF binding site”, or “open chromatin” by the Ensembl Regulatory Build (v101) (Zerbino et al. 2015); all exonic regions annotated by GenCode (Harrow et al. 2012); and GERP evolutionarily conserved regions (Herrero et al. 2016). In total, I masked 75.3% of the mouse genome.

I considered all 41,000,862 single nucleotide variants (SNVs) with “PASS” filter status across unmasked regions of the genome, amounting to approximately 67 SNVs per kb.

2.5.3 Local sequence context models and probabilities of substitution

I considered six k-mer models for determining the effect of local sequence context on substitution probabilities (Table 2.1): the 1-mer, 3-mer, 5-mer, 7-mer, 9-mer and A-CG models. The 1-mer model is the simplest model and does not consider the local sequence context of the base of interest. It has 12 parameters as each of the four bases has three possible substitutions. The 3-mer model considers one flanking nucleotide on either side of the base of interest, making a 3-mer. It has 192 parameters as there are 64 possible 3-mers (4^3) and the middle base in each 3-mer has three possible substitutions. The 5-mer, 7-mer, and 9-mer models consider 2, 3, and 4 flanking nucleotides on either side of the base of interest, respectively. The number of parameters increases exponentially as the number of flanking nucleotides considered expands, with the 9-mer model having 786,432 parameters as there are 262,144 possible 9-mers (4^9) and the middle base in each 9-mer has three possible substitutions. Substitution probabilities are largely driven by the presence of CG dinucleotides which have a high mutation rate from cytosine to thymine (M. Li and Chen 2011). I, therefore, expanded the simple 1-mer model to consider CG context and named this the A-CG model. The A-CG model has 18 parameters as there are 6 possible contexts (A, T, C (non-CG), C (CG), G (non-CG), and G (CG)), each with three possible substitutions.

I inferred the mutational direction of each SNV (i.e. the ancestral and mutant states) using Ensembl's alignment of the *Mus musculus* (GRCm38) and *Mus caroli* reference genomes. Following previously published protocols (K. Harris 2015) the ancestral state for each SNV was assumed to be the reference allele unless the alternate allele was shared with *Mus caroli*.

I calculated the substitution probability for each k-mer by dividing the counts of each k-mer change by the counts of each ancestral k-mer across the unmasked genome. Each ancestral k-mer and k-mer change has a complementary k-mer on the reverse strand. For example, the ancestral 3-mer TAG on the forward strand must have the complementary k-mer CTA on the reverse strand, and a TAG > TTG substitution must result in a complementary CTA > CAA substitution. I, therefore, totalled the counts of each forward and reverse k-mer and k-mer change to calculate substitution probabilities, effectively halving the number of parameters for each model, and report the substitution probabilities for adenine and cytosine only.

2.5.4 Methylation states and probabilities of substitution

Genome-wide methylation of CG dinucleotides in mouse embryonic stem cells was quantified by Stadler et al (2011) using bisulfite sequencing and were downloaded from Ensembl (Stadler et al. 2011; Zerbino et al. 2018) (v101). Genome-wide methylation data is not available for mouse germline cells. However, similar patterns of methylation have been observed between mouse embryonic stem cells and sperm cells (Popp et al. 2010), and I, therefore, used mouse embryonic stem cells as a proxy for the germline.

I classified all CG dinucleotides across the unmasked mouse genome as methylated if they have coverage greater than or equal to 5 and a percentage methylated greater than or equal to 60, and unmethylated if they have coverage greater than or equal to 5 and a percentage methylated less than or equal to 20. I excluded CG dinucleotides with coverage less than 5, or a percentage methylated more than 20 and less than 60.

The distribution of methylation percentages is shown in Supplementary Figure 2.5. In total, I classified 72.9% of CG dinucleotides as methylated, 6.9% as unmethylated, and excluded 20.2% from our analysis.

I calculated the probability for cytosine substitutions across three CG dinucleotide groups: all CG dinucleotides, unmethylated CG dinucleotides, and methylated CG dinucleotides. For each unmasked CG dinucleotide, I determined the substitution class by the alternate allele (i.e. C>A, C>G, C>T, G>A, G>C, G>T). I calculated the probabilities of substitution for each CG group by dividing the counts of each substitution class in each CG group, by the number of CGs in each CG group. I then combined counts between complementary strands to leave three substitution probabilities: C>A, C>G, and C>T.

2.5.5 Testing the prediction accuracy of substitution models

I used resampling by chromosome to test the prediction accuracy of local sequence context models. For each iteration, I split the data into training and testing datasets. Four autosomes were randomly selected to be the testing data, and the remaining fifteen autosomes were used to train the models. Due to differences in chromosome size, this meant that the testing data could range between 28.0% and 14.1% of the total dataset. Probabilities of substitution were calculated for each k-mer model using the training data, following the methodology described above. To minimise bias introduced by the composition of test chromosomes (i.e. some k-mers appearing at a much higher frequency than others) each 9-mer was given an equal weighting in calculating substitution occurrence. This means the observed number of

substitutions in the test data was calculated by totalling the observed substitution rates for each 9-mer. The absolute error was then calculated for each local sequence context model and substitution class (i.e. A>C, A>G, A>T, A>*, C>A, C>G, C>T, C>*) as the sum of the absolute difference between the observed (testing data) and predicted (training data) number of each substitution class for each 9-mer. Resampling was conducted 100 times and the mean absolute error and 95% confidence interval for each k-mer model and substitution class were calculated across the iterations. Differences in absolute error between models are assessed using t-tests.

I tested the accuracy to predict CG dinucleotide substitutions given different methylation states using resampling by chromosome, with four chromosomes randomly sampled for the test dataset in each iteration. Probabilities of substitution were calculated for each CG dinucleotide group in the training dataset following the methodology described above. Probabilities of substitution for methylated and unmethylated CG dinucleotides were calculated with the test dataset. The prediction accuracy of each CG group was determined as the difference between the observed (testing data) and predicted (training data) fraction for each substitution class.

2.5.6 Power testing

The power to capture the effect of expanding sequence context models on the prediction accuracy of substitution occurrence between chromosomes was assessed by simulating datasets in which the observed substitution counts are defined by different sequence context models. I created three simulated datasets of substitution

counts: one where the true probability of substitution for each k-mer is defined by its 3-mer context, one where it is defined by its 7-mer context, and the other where it is defined by its 9-mer context. I then tested prediction accuracy for the different k-mer models using these simulated datasets to determine whether there is sufficient power to detect an improvement in prediction accuracy given the simulated effect sizes. In the simulated data, the number of substitutions for each 9-mer in each chromosome was determined from a binomial distribution, where the number of trials was equal to the number of each 9-mer, and the probability of success was defined by the observed substitution rates given the 3-mer, 7-mer, or 9-mer context. For example, 3-mer, 7-mer, and 9-mer probabilities of substitution were calculated from the observed data using all chromosomes, and then these probabilities were used to define the probabilities of success for each trial within chromosomes. Given these simulated datasets, prediction accuracy was tested as mean absolute error calculated from train test resampling between chromosomes as described above.

2.5.7 Mutational effect and probability of occurrence

I consider the coding consequences for all possible point mutations across all possible codons. In total, there are 64 codons and 576 possible point mutations. Of these, 23 cause nonsense mutations, 415 cause missense mutations, and 138 cause synonymous mutations. Each codon point mutation has 256 possible 7-mer contexts and corresponding substitution rates (Figure 2.5). I classified each of the 49,152 7-mer substitutions (16,384 7-mers each with 3 possible middle-base substitutions) into one of two categories: potential nonsense substitutions (PNS, n = 5,248) for substitutions with the potential to cause one or more nonsense mutations in

translated sequences; or no-nonsense substitutions (NNSs, n = 19,328) for substitutions that do not have the potential to cause a nonsense mutation (Figure 2.5). Note that each 7-mer substitution has a complementary 7-mer on the reverse strand. For example, the 7-mer CAAAAAA on the forward strand must have the complementary 7-mer TTTTTTG on the reverse strand, and a CAAAAAA > CAAGAAA substitution must result in a complementary TTTTTTG > TTTCTTG substitution. I, therefore, considered the coding consequences of the forward and reverse strands for each 7-mer substitution, and reported the substitution probabilities for adenine and cytosine only, effectively halving the number of substitutions. 7-mer substitutions were further divided by substitution class (A>C, A>G, A>T, C>A, C>G, C>T) and GC content (i.e. the number of G or C bases in the 7-mer). I tested the difference in substitution rates between groups with Wilcoxon signed-rank tests.

Population genetics theory states that selection can overcome genetic drift when the product of the following exceeds one: the effective population size (Ne), the per-nucleotide mutation rate (u), the magnitude of change in the mutation rate (du), the average selection coefficient of a mutation (s), and the number of sites where a mutation would be under selection (Nsites) (Lynch 2010b). I calculated Nsites by summing the number of sites in transcripts where a substitution would cause a nonsense mutation. I considered all canonical transcripts defined by Ensembl that are divisible by three, start with a start codon, and end with a stop codon (n = 16,903). Nsites was calculated by considering all possible substitutions at each site, and for each substitution class independently. I calculated the minimum Nsites required for selection to overcome genetic drift given a conservative effective population size of 30,000 (Phifer-Rixey et al. 2012), a per-nucleotide substitution

rate of 5.4e−9 (Uchimura et al. 2015), an order of magnitude reduction in mutation rate (Hodgkinson and Eyre-Walker 2011), and an average selection coefficient against nonsense substitutions of 0.025 (Gorlov, Kimmel, and Amos 2006). The relationship between du and the product of Ne, u, du, s, and Nsites for each substitution class was tested and is provided in the supplementary material (Supplementary Fig. 2.4).

Chapter 3

Selective constraint against nonsynonymous variation in the mouse lineage

Results from this chapter have been submitted to BioRxiv (Powell et al. 2019). I confirm that I produced all the material with no more input/advice than would be considered appropriate or acceptable for a thesis.

3.1 Abstract

Genic constraint describes the degree to which a protein-coding gene is intolerant of amino-acid sequence altering variation as a consequence of negative selection. Several methods to quantify gene constraint within the human lineage have been applied to exome sequence data. These methods have identified an enrichment of genes that are essential for life and are associated with disease among the most constrained genes. In this chapter, I describe a method to estimate genic constraint in the mouse lineage and name the resulting metric the nonsynonymous observed expected ratio (NOER). I use NOER to assess the relationships between gene constraint and knockout phenotype. NOER outperforms dN/dS when predicting gene knockout lethality and sub-viability, and is positively correlated with pleiotropy

(i.e. a greater degree of constraint is correlated with an increased number of knockout phenotype associations)(Spearman's rho = 0.67, $p < 0.01$). Finally, I show that NOER is positively correlated with human constraint metrics and with the number of pathogenic variants in the human orthologue (rho = 0.27, $p < 0.01$). This correlation suggests that selective constraint can be in part explained by variants in more constrained genes having an increased likelihood of being pathogenic and that this pathogenicity has been maintained between the human and mouse lineages.

3.2 Introduction

Patterns of genetic variation can be used to estimate the selection effects acting across species' genomes and infer the functional importance of regions, specifically with regard to fitness (Guo et al. 2007; Davydov et al. 2010). Regions that are depleted of variation are considered under selective constraint, assuming the depletion is due to variants being deleterious to fitness and therefore purged by negative selection (Edwards et al. 2013; di Iulio et al. 2018). Several methods have been used to quantify selective constraint acting on protein-coding genes. The principle of each method is to quantify the degree to which a gene is depleted of amino-acid sequence altering variation relative to the expected variation given neutral evolution (Bartha et al. 2018). Methods have traditionally compared variation in orthologous genes between species (Subramanian 2011); however, the availability of exome sequences from human populations has enabled several methods for quantifying gene constraint within the human lineage (Samocha et al. 2014; Bartha et al. 2015; Lek et al. 2016; Fadista et al. 2017; Cassa et al. 2017;

Abramovs, Brass, and Tassabehji 2019; Petrovski et al. 2013; Rackham et al. 2015). These metrics have highlighted genes that are essential for life, associated with disease, and enriched for pathogenic variants among the most constrained human genes (Bartha et al. 2018).

Mouse models are critical for genetic, physiological, and biomedical research, in particular for understanding the function of mammalian genes *in-vivo* (Rosenthal and Brown 2007; Justice and Dhillon 2016; Yue et al. 2014; Lloyd et al. 2020). For example, mouse models are being used to assess the function of mammalian genes by characterising the phenotypic consequences of knockouts (i.e. when a gene is made inoperative) (Dickinson et al. 2016; C. L. Smith and Eppig 2012; Karp et al. 2017). These knockout experiments have been used to characterise constrained human orthologues, and have highlighted an enrichment of genes associated with developmental lethality in mice among the most constrained human genes (Bartha et al. 2018). Interspecific constraint has been shown to correlate between orthologues in the mouse-rat and human-chimpanzee lineages (Subramanian 2011); however, gene to phenotype functional pathways can change over evolutionary time, affecting the degree of selective constraint (Liao and Zhang 2006; Pervouchine et al. 2015; Denas et al. 2015). Quantifying gene constraint in the mouse lineage has the potential to help contextualise differences between humans and mice, and inform future clinical research using mouse models. This could be particularly pertinent for mouse humanization using CRISPR/Cas9 (F. Li et al. 2014), and the clinical development of new drugs (Minikel et al. 2019).

In this chapter, I describe and calculate a new metric for quantifying mouse gene constraint named the nonsynonymous observed expected ratio (NOER). I use this

metric to assess the relationships between mouse gene constraint and gene function by considering knockout phenotypes from the International Mouse Phenotyping Consortium (IMPC). NOER outperforms dN/dS when predicting gene knockout lethality, and positively correlates with gene pleiotropy (i.e. more constrained genes have more associated knockout phenotypes). Finally, I highlight the correlation in constraint between mouse and human orthologues, and enrichment of pathogenic variants in the human orthologues of the most constrained mouse genes.

3.3 Results

3.3.1 Quantifying mouse gene constraint against nonsynonymous variation.

I quantified selective constraint across 16,609 mouse genes as the nonsynonymous observed expected ratio (NOER), using single nucleotide variants (SNVs) from wild mice populations (*Mus musculus sp.* and *Mus spretus*) (Harr et al. 2016). NOER quantifies constraint as the ratio of the observed number of nonsynonymous variants to the expected number given an absence of selection. In brief, the expected number of nonsynonymous variants, given no selection in each gene, is predicted using a linear model trained on the number of synonymous variants (presumed not to have been under selection) as a function of the gene's sequence-specific probabilities of incurring a given mutation type (i.e. synonymous or nonsynonymous), controlling for mutability captured by local sequence context and CG methylation status (see methods for more details) (Samocha et al. 2014). Genes with lower NOER have

relatively fewer nonsynonymous variants than expected given no selection and are considered more constrained (Supplementary Figure 3.1). NOER was used to rank genes by their estimated degrees of selective constraint. Genes were binned from 1 to 100 based on their NOER percentile rank, with the most constrained genes in the top percentile and the least constrained in the bottom percentile.

3.3.2 Mouse gene constraint and knockout phenotype

Gene knockout experiments characterise individual genes by their homozygous complete loss-of-function phenotype. I tested three hypotheses. First, genes associated with severe phenotypes such as embryonic lethality are enriched amongst the most constrained mouse genes. Second, genes with no significant knockout phenotype are enriched amongst the least constrained mouse genes. And third, there is a positive correlation between gene constraint and pleiotropy. I consider homozygous knockout phenotypes from the International Mouse Phenotyping Consortium (IMPC), which characterises mouse knockout phenotypes by measuring various parameters across a range of standardised phenotyping procedures (Dickinson et al. 2016).

I classified 5,252 mouse knockouts conducted by the IMPC into four mutually exclusive categories: lethal (L) if they cause preweaning lethality in all knockouts (n = 1,372); sub-viable (SV) if they cause preweaning lethality in one or more, but not all, knockouts (n = 519); viable phenotype (VP) if they are not classified as lethal or sub-viable, but are annotated with one or more other phenotypes (n = 2,993); and viable ‘no phenotype’ (VN) if they are not annotated with any significant phenotypic

deviations from the control group (n = 368). I calculated odds ratios to assess the enrichment of each knockout grouping amongst the 10% most constrained (n = 527) and 10% least constrained (n = 528) IMPC knockouts (Figure 3.1, Supplementary Table 3.1). Knockouts associated with preweaning lethality or sub viability are enriched among the most constrained mouse genes (Figure 3.1, Supplementary Table 3.1). This suggests that nonsynonymous variants in developmental genes are more likely to be purged by negative selection. Knockouts with no annotated phenotype are enriched among the least constrained mouse genes (Figure 3.1). The relative decrease in constraint against nonsynonymous variants in genes with no associated knockout phenotype is likely in part due to these genes having a compensatory network, and/or not playing a critical function in organism fitness. Deviations in phenotype that could result in a fitness cost may, however, have gone undetected as a result of power limitations (Brown and Moore 2012).

I compared NOER with dN/dS to assess its validity. dN/dS is a frequently used measure of the selection effect on a gene, defined as the ratio of nonsynonymous variants per nonsynonymous site (dN) to the number of synonymous substitutions per synonymous site (dS) (Kryazhimskiy and Plotkin 2008). Genes with a lower dN/dS ratio (i.e. fewer nonsynonymous relative to synonymous substitutions) are considered more constrained, and genes that are evolving neutrally are expected to have a dN/dS ratio of 1. To standardise the comparison I consider dN/dS scores between *Mus musculus* and *Mus spretus*, as these are the two most distantly diverged species used to calculate NOER. NOER and dN/dS are positively correlated (rho = 0.54, $p < 2.2e\text{-}16$). I used simple logistic regression to predict knockouts classified as lethal (L) or sub-viable (SV), and viable 'no phenotype' (VN), and assessed model fit and predictive accuracy as the area under the receiver operating

characteristic (ROC). Both NOER (estimate = 0.018±0.001, p = 1.7e-61) and dN/dS (estimate = 0.011±0.001, p = 1.0e-20) are significant predictors of knockout lethality, but predictive accuracy is greater for NOER (ROC = 0.64) compared with dN/dS (ROC = 0.59). Similarly, NOER and dN/dS show the reverse relationship with 'no phenotype' knockouts (NOER estimate = -0.015±0.002, p = 4.9e-14; dN/dS estimate = -0.011±0.002, p = 2.5e-7), but predictive accuracy is modestly greater for NOER (ROC = 0.62) compared with dN/dS (ROC = 0.59). The main difference between NOER and dN/dS is their method for estimating the expected number of nonsynonymous variants in each gene given no selection. The use of a model to predict the average expected number of nonsynonymous single nucleotide variant (SNV) sites per gene for NOER accounts for stochasticity in neutral evolution between genes (Samocha et al. 2014).

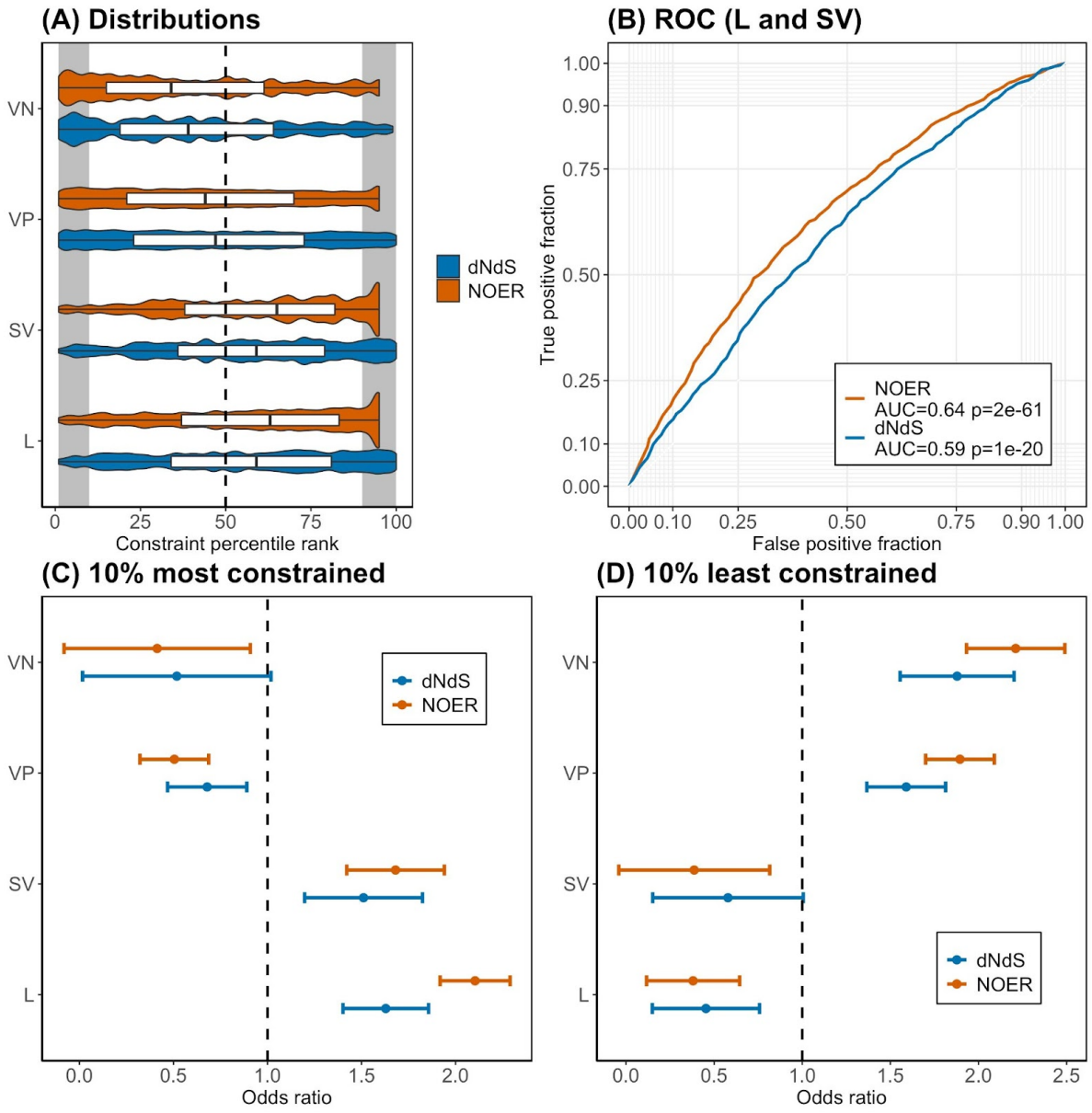


Figure 3.1 – (A) Distributions of gene constraint by lethal (L), sub-viable (SV), viable 'with phenotype' (VP), and viable 'no phenotype' (VN) knockout groupings. Constraint is quantified as the nonsynonymous observed expected ratio (NOER) and dN/dS, with a higher percentile rank indicating a greater degree of constraint. The dashed line indicates the median score for all genes under consideration. (B) ROC curves highlight the accuracy of NOER and dN/dS to predict knockout lethality or sub-viability. I fit two simple logistic regressions to predict knockout lethality or sub-viability, one with NOER and the other with dN/dS as the explanatory variable. NOER has greater predictive accuracy than dN/dS (0.64 compared with 0.59). (C and D) Odds ratios and 95% confidence intervals for the 10% most constrained (n = 527), and 10% least constrained (n = 528) genes across each knockout group. An odds ratio greater than one indicates an enrichment. The most constrained genes are enriched for lethal and sub-viable knockout phenotypes, and the least constrained genes are enriched for genes with no associated phenotype deviation.

Genes can affect multiple, often seemingly unrelated, phenotypes. This phenomenon is known as pleiotropy. Here, I define pleiotropy for each gene as the proportion of phenotypes associated with the IMPC knockout. I calculated pleiotropy for each IMPC knockout as the proportion of phenotyping procedures the knockout was subjected to in which one or more parameters significantly deviated from the control group (i.e. passes a significance threshold of 0.0001). The degree of pleiotropy is positively skewed, with most genes affecting few traits (Figure 3.2). Of the 3,255 knockouts investigated, 368 (11.3%) did not significantly deviate from the control group in any phenotyping procedure, 942 (28.9%) significantly deviated from the control group in one phenotyping procedure, and 1,945 (59.8%) significantly deviated from the control group in more than one phenotyping procedure. Gene constraint is positively correlated with pleiotropy (Spearman's rho = 0.67, p =9.1e-14) (Figure 3.2). To test whether this is because genes are affecting multiple closely related phenotypes, I also calculated the correlation between gene constraint and the number of associated top-level MP terms (i.e. phenotypes that are more distantly related in the ontology), and found that the relationship persists (Spearman's rho = 0.67 p = 4.8e-14) (Figure 3.2). The correlation between gene constraint and pleiotropy may be caused by additive selection effects across multiple phenotypes (Wang, Liao, and Zhang 2010; Fish, Chen, and Capra 2017). For example, the more phenotypes a gene affects, the more likely it is that one or more of the effects caused by loss-of-function variation will be deleterious to fitness (Bauer, Li, and Siegal 2015; Cooper, Ostrowski, and Travisano 2007).

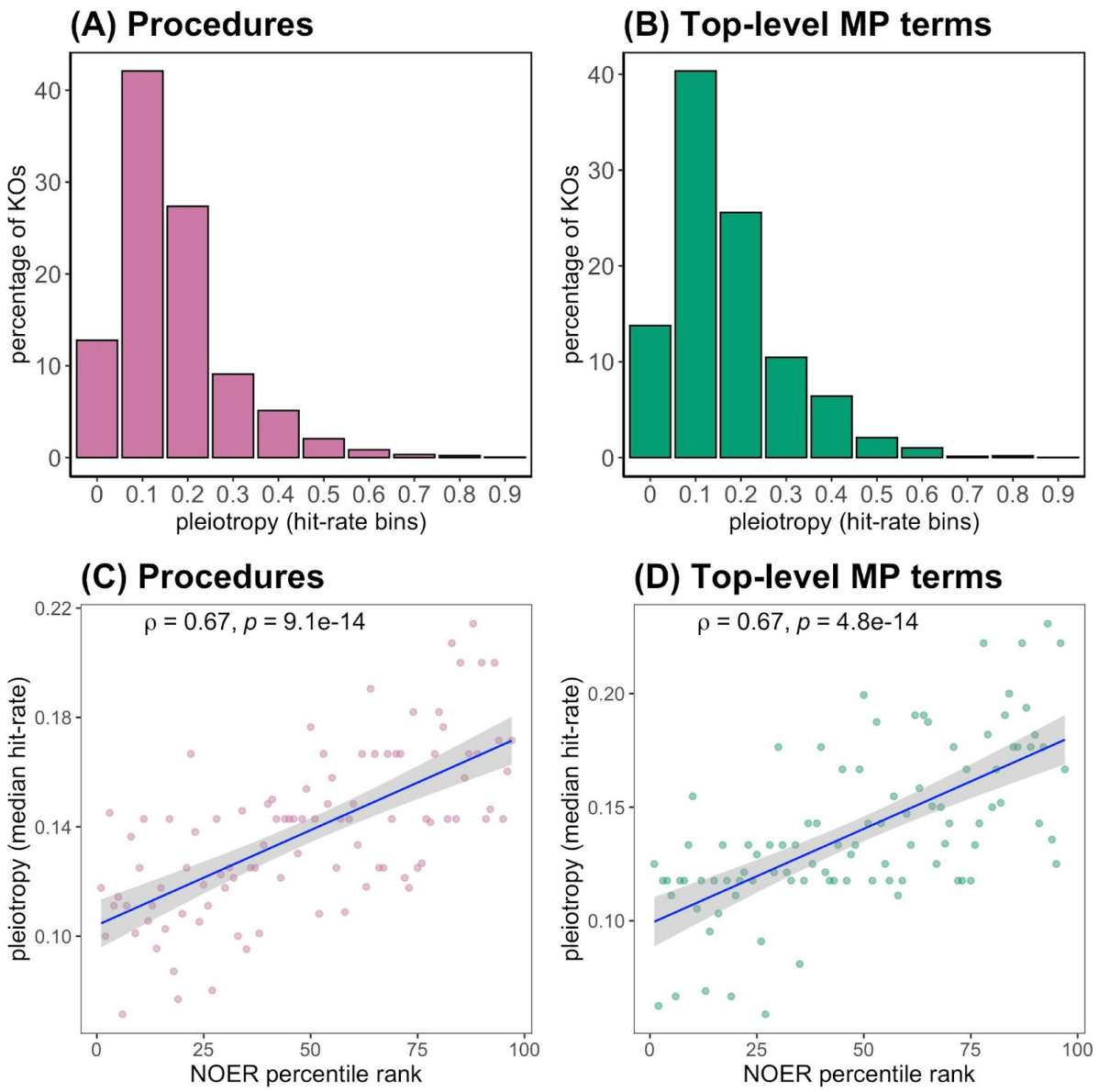


Figure 3.2 – Mouse gene constraint quantified as NOER is positively correlated with an increased number of knockout phenotypes. Two measures of pleiotropy were calculated for each IMPC knockout. First, the proportion of procedures the knockout was subjected to in which one or more parameters significantly deviates from the control group (A and C). Second, the number of top-level mouse phenotype ontology terms associated with the knock-out is divided by the potential number of top-level terms based on the conducted phenotyping tests (B and D). Constraint is quantified as the nonsynonymous observed expected ratio (NOER), with a greater percentile rank indicating a greater degree of constraint. The rank correlation was calculated between (C) the median phenotyping procedures hit-rate per knockout and the NOER percentile bin (Spearman's rho = 0.67, p = 9.1e-14), and (D) the median top-level MP term hit-rate per knockout and the NOER percentile bin (Spearman's rho = 0.67, p = 4.8e-14).

3.3.3 Human pathogenic variant site enrichment

Mouse models can be used to infer the pathobiology of human disease-associated variants, and their potential for this purpose has increased with the development of CRISPR/Cas9 technologies that enable targeted humanisation of the mouse genome. If selective constraint has been maintained between the human and mouse lineages in part because mutations in constrained genes are more likely to cause disease, it is to be expected that mouse gene constraint will positively correlate with the number of pathogenic variants in their human orthologue. To test this, I considered 15,501 human pathogenic variants from the ClinVar database (Landrum et al. 2018) across 7,037 human genes with a one-to-one mouse orthologue. To account for differences in gene length, I averaged pathogenic variants in each gene per kb. There is a significant positive correlation between mouse gene constraint and the mean number of pathogenic variants per kb in the human orthologue (Spearman's rho = 0.27, p = 0.0075) (Figure 3.3). It is possible that this correlation is affected by sampling bias if genes have more recorded pathogenic variants because they have been investigated more thoroughly. To assess this, I also calculated the correlation between mouse gene constraint and the number of benign variants per kb in the human orthologue: if there is sampling bias I would expect to see more benign variants in the genes that have been more thoroughly assessed. However, I observe a negative correlation between mouse gene constraint and the mean number of benign variants per kb in the human orthologue (Spearman's rho = 0.59, p = 2.4e-10) (Figure 3.3). This suggests that the result is not an artefact of sampling bias, but rather that the positive correlation between human gene constraint and the number of known pathogenic variants can be in part explained by variation in more constrained genes having an

increased likelihood of being pathogenic, and therefore purged by negative selection (Bartha et al. 2018). Furthermore, the correlation between mouse gene constraint and the number of known pathogenic variants in the human orthologue suggests that negative selection resulting from pathogenicity in these genes may have been maintained between the lineages.

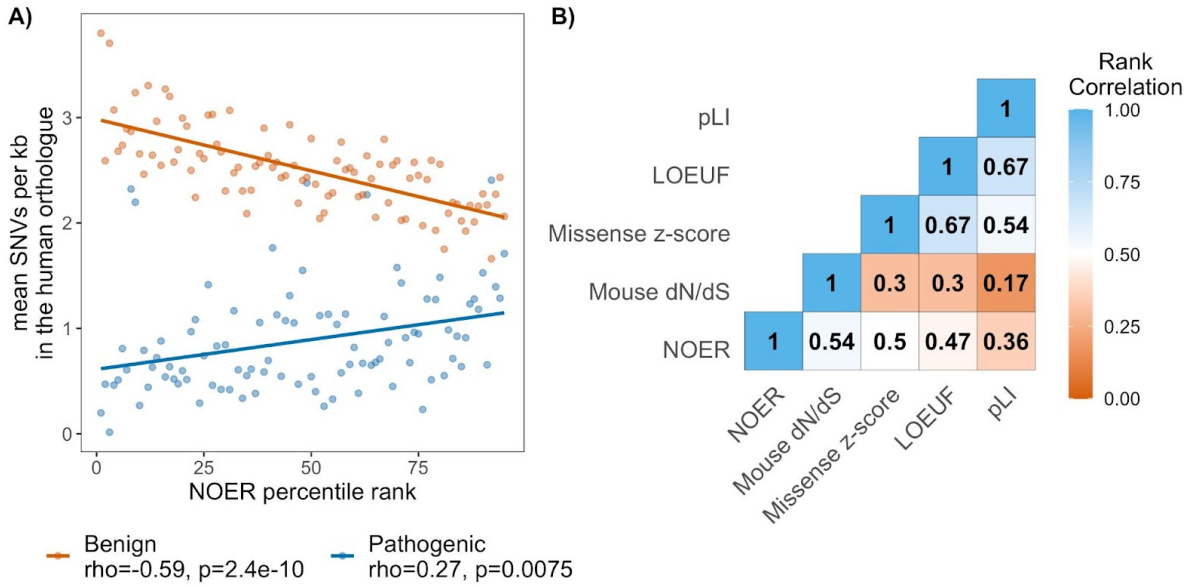


Figure 3.3 – (A) Mouse gene constraint is positively correlated with the number of known pathogenic variants in their human orthologues and negatively correlated with the number of known benign variants. 15,501 pathogenic and 37,032 benign SNV sites were considered from the ClinVar database across 7,037 human genes with a mouse one-to-one orthologue. Constraint is quantified as the NOER, with a greater percentile rank indicating a greater degree of constraint. Rank correlations were calculated between the mean number of pathogenic and benign variants per kb in the human orthologue and the NOER percentile bin. (B) Spearman's Rank correlation matrix of constraint scores for human and mouse one-to-one orthologues. Mouse gene constraint is calculated as the nonsynonymous observed expected ratio (NOER), and dN/dS. Human gene constraint scores include the missense *Z*-score, loss-of-function observed/expected upper bound fraction (LOEUF), and the probability of loss-of-function intolerance (pLI). NOER is significantly (Spearman's rho, $p < 0.0001$) correlated with metrics of human gene constraint and mouse dN/dS.

NOER is positively correlated with metrics of human gene constraint (Figure 3.3); however, there are examples of genes across which the strength of selective

constraint substantially differs between the species. *FRAS1-related* extracellular matrix 3 (*FREM3*) is a gene that is highly intolerant of missense variation ranking among the 7% most constrained human genes (measured as the missense z-score). *FREM3* contains three pathogenic variant sites annotated by ClinVar, in addition to variant sites associated with glycated haemoglobin levels and severe malaria as annotated in the GWAS catalog. By comparison, its murine orthologue *Frem3* is highly tolerant of nonsynonymous variation ranking among the 4% least constrained mouse genes (measured as the NOER). *Frem3* has not yet been knocked out by the IMPC, and currently has no phenotype annotations in the Mouse Genome Informatics database. Calcineurin like EF-hand protein 2 (*CHP2*) and G protein subunit gamma 13 (*GNG13*) are both highly unconstrained in humans, ranking among the 6% and 9% least constrained human genes (measured as the missense *Z*-score), respectively. *CHP2* and *GNG13* have no disease or phenotype associations annotated by Ensembl and do not contain any pathogenic variant sites annotated by ClinVar. By comparison, their murine orthologues *Chp2* and *Gng13* are highly intolerant of nonsynonymous variation ranking amongst the 5% most constrained mouse genes (measured as the NOER). Both *Chp2* and *Gng1* result in sub-viability when knocked out in mice.

3.4 Discussion

In this chapter, I developed a method to rank mouse genes by their intolerance of nonsynonymous mutation using standing variation between populations of wild mice and characterised the most constrained mouse genes by their knockout phenotypes.

The research has two important findings. First, mouse gene constraint is positively correlated with an increased number of knockout phenotypes, suggesting pleiotropic genes are more likely to be under stronger selective constraint. This is possibly due to additive selection effects across the multiple phenotypes of pleiotropic genes (Wang, Liao, and Zhang 2010; Fish, Chen, and Capra 2017). For example, if a loss-of-function or gain-of-function variant affects multiple phenotypes, it is likely there will be a net deleterious effect on fitness resulting in selective constraint (Bauer, Li, and Siegal 2015; Cooper, Ostrowski, and Travisano 2007). This result also has implications for understanding the genetic basis of comorbidity (i.e. the simultaneous presence of two or more diseases in an individual), as genes associated with comorbidities are expected to be pleiotropic (Ellinghaus et al. 2016); (Prieto et al. 2016). Second, mouse gene constraint is positively correlated with an increase in the number of known pathogenic variants in human orthologues. This correlation suggests that selective constraint can be in part explained by variants in more constrained genes having an increased likelihood of being pathogenic (Bartha et al. 2018) and that this pathogenicity has been maintained between the lineages.

Future research could expand on our results by identifying the regions within mouse genes that are most constrained. Gene-wide scores do not capture regional heterogeneity in constraint within protein-coding sequences due to regional differences in the consequences of nonsynonymous variation on protein form and function. Within-gene constrained coding regions have been calculated for human genes, and are enriched for known pathogenic variants (Havrilla et al. 2018). These could be aligned and directly compared with their mouse orthologues to highlight protein domains functionally important to each lineage and used to inform when the

humanisation of mouse genes is appropriate for studying the pathobiology of human variant sites.

In conclusion, quantifying mouse gene constraint has the potential to prioritise future gene knockouts based on their expected phenotypes, with more constrained knockouts being more likely to result in lethality, and less constrained knockouts being less likely to have a detectable phenotype. Furthermore, similarities and differences in constraint between human and mouse orthologues, while hard to disentangle from methodological artefacts, could help contextualise results derived from mouse models of human conditions.

3.5 Methods

3.5.1 Raw data and dependencies

Genetic variation (vcf) and coverage (bam) files were downloaded from http://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/ (Harr et al. 2016). *Mus musculus* to *Mus spretus* dN/dS, human to mouse (*Mus musculus*) orthologues, mouse canonical transcript sequences, and genome-wide methylation of CG dinucleotides in mouse embryonic stem cells were downloaded from Ensembl (v94) (Zerbino et al. 2018; Herrero et al. 2016; Stadler et al. 2011). Measures of intraspecific human constraint, including the missense Z-score (Samocha et al. 2014), loss-of-function observed/expected upper bound fraction (LOEUF) (Karczewski et al., n.d.), and probability of loss-of-function intolerance (pLI) (Lek et al. 2016) was downloaded from the supplementary information of their reference

papers. Human pathogenic and benign variants were downloaded from ClinVar (Landrum et al. 2018) on 02/03/2021. Knockout phenotypes for 5,252 genes were downloaded from the International Mouse Phenotyping Consortium (IMPC)(release 12.0)All data wrangling, quality control, and analyses were performed using custom scripts written in R (v3.6.2) (R Core Team 2017) or bash.

3.5.2 Defining genes, quality variants, and coding consequences

I considered protein-coding genes with a HUGO Gene Nomenclature Committee name and defined the coding sequence for each gene by their Ensembl canonical transcript (v94) (Zerbino et al. 2018). I removed transcripts that did not both begin with a start codon and finish with a stop codon, do not have a length that is divisible by three, or contain any character other than A, T, C, or G. Read-depth at every genomic position was calculated for each individual in the sample from the bam files using SAMtools (H. Li et al. 2009). Genomic positions were classified as having low coverage if they have a read-depth less than 10X in more than 10% of sequenced mice.

I considered genetic variation between 67 mice (*Mus musculus sp*. and *Mus spretus*) from 8 wild populations described in detail by (Harr et al. 2016). I considered all autosomal single nucleotide variants (SNV) with 'PASS' filter status. I classified SNVs as synonymous and nonsynonymous based on their annotated consequences for the amino-acid sequence. The coding consequences of SNVs were determined using the Ensembl Variant Effect Predictor (v94.5) (McLaren et al. 2016). One consequence

was determined per SNV using the "--pick" argument which prioritises annotations by canonical transcript status. Consequences defined as "synonymous_variant", "stop_retained_variant", or "start_retained_variant" were classified as synonymous, and consequences defined as "missense_variant", "stop_gained", "start_lost", "stop_lost", or "protein_altering_variant" were classified as nonsynonymous. After filtering, 501,643 synonymous and 265,343 nonsynonymous SNV sites across 16,609 genes remained for analysis (Supplementary Figure 3.3).

3.5.3 Transcript-specific probabilities of nonsynonymous and synonymous substitution

There is variation between transcripts in the likelihood that a point substitution will cause either a synonymous or nonsynonymous mutation. This variation is a result of differences in the ratio of synonymous to nonsynonymous mutations that can occur from point substitutions in each codon, and variation in the likelihood of different point substitutions (Supplementary Table 3.2). Each codon has nine possible point substitutions as each of the three bases could be substituted with one of the other three bases. For example, there is one point substitution to an AAA codon that causes a synonymous mutation and eight that cause nonsynonymous mutations. In comparison, there are three point substitutions that cause synonymous mutations to CCC codons, and six that cause nonsynonymous mutations. Therefore a random substitution in a CCC codon is three times more likely to cause a synonymous mutation than a substitution in an AAA codon.

There is further variation in the probability of each point substitution captured by

local sequence context, and the methylation state of CG dinucleotides. In the mouse lineage, the 7-mer context (i.e. the three flanking bases on either side of the base of interest) explains more variation in substitution occurrence than either the 3-mer or 5-mer (see chapter 2). Unmethylated CG dinucleotides are approximately eight times less likely to undergo C > T substitutions than methylated CGs (see chapter 2). The majority of CG dinucleotides are methylated across the mouse genome (Stadler et al. 2011) biasing substitution rate estimates that do not consider the methylation status of CG dinucleotides. I, therefore, identified unmethylated CG dinucleotides within each transcript. Methylation status for CG dinucleotides was classified from whole-genome bisulfite sequencing in mouse embryonic stem cells, conducted by Stadler et al (2011). Whole-genome methylation state data is not available for the mouse germline. However, similar patterns of methylation have been observed between mouse embryonic stem cells and sperm cells (Popp et al. 2010). I, therefore, used mouse embryonic stem cells as a proxy for the germline. CG dinucleotides were classified as unmethylated if they have coverage greater than or equal to 5 and a percentage methylated less than or equal to 20. The distribution of CG dinucleotide methylation within the considered transcripts is provided in Supplementary Figure 3.4. In total, 18.6% of the CG dinucleotides within the considered transcripts were classified as unmethylated.

I calculated the probability of synonymous and nonsynonymous mutation for each transcript by: 1) using the 7-mer sequence context to determine the probability of all potential point substitutions as previously calculated in chapter 2. I replaced the 7-mer probability of substitution at unmethylated CG dinucleotide sites with the probability of substitution for unmethylated CG dinucleotides as previously calculated in chapter 2. 2) determining the coding consequences (nonsynonymous

or synonymous) of each potential point substitution; and 3) totalling the probabilities of substitution (defined by the 7-mer context and methylation state) for all possible point substitutions that would cause a synonymous mutation and all possible point substitutions that would cause a nonsynonymous mutation (Supplementary Table 3.2).

3.5.4 Calculating gene-wide intolerance of nonsynonymous variation

I developed a new metric to quantify gene-wide selective constraint against nonsynonymous variation, named the nonsynonymous observed expected ratio (NOER), based on previously published work by (Samocha et al. 2014). Simply put, NOER captures the ratio of the observed number of nonsynonymous SNVs in a transcript relative to the expected number predicted given no selection. NOER is calculated in a two-stage process. First, a 'neutral' model is trained by fitting a linear model to predict the number of synonymous SNVs in each gene as a function of the transcript-specific probability of synonymous mutation and the number of positions classified as 'low coverage' in the transcript (see previous subsections). Synonymous SNVs are assumed to be selectively neutral, and so this model estimates the number of SNVs given no selection. The model closely fits the data with an R^2 of 0.85 and a residual standard error of 11.8. Second, the 'neutral' model is used to predict the 'expected' number of nonsynonymous variants in each gene assuming no selection based on the transcript-specific probabilities of nonsynonymous mutation. NOER is calculated as the ratio of observed to expected nonsynonymous SNVs. Genes with a score of less than one have fewer nonsynonymous variants than expected given no

selection and are considered more constrained. NOER were binned from 1 to 100 based on their percentile rank, with the most constrained genes in the top percentile and the least constrained in the bottom percentile.

The most constrained genes defined by NOER are biased toward either: 1) genes that are short and therefore have a greater probability of containing no nonsynonymous variants by chance and being classified as maximally constrained (i.e. NOER of 0); or 2) longer genes where there is more power to distinguish between the observed and expected number of nonsynonymous variants. I classify the 10% of genes with the lowest NOER scores as highly constrained (n = 1,660). For 305 (18.4%) of these highly constrained genes, the lower bound 95% prediction interval of the expected number of nonsynonymous variants is less than or equal to 0 (i.e. they have a 2.5% chance of containing no nonsynonymous variants based on the neutral model).
A gene must be predicted to contain a minimum of 35 nonsynonymous variants by the neutral model to appear amongst the 10% most constrained genes if they are observed to contain one nonsynonymous variant. Of the 16,609 genes assessed, 4,133 (24.9%) have an expected number of nonsynonymous variants below this threshold. Distributions of gene length by NOER percentile are provided in Supplementary Figure 3.5.

3.5.5 Assessing mouse gene constraint and knockout phenotype

I investigated the relationship between gene constraint measured as the NOER and gene function, by considering knockout phenotypes for 5,252 genes from the

International Mouse Phenotyping Consortium (IMPC)(release 12.0). The IMPC is systematically knocking out all protein-coding genes across the mouse genome. Each knockout in the IMPC is subject to a standardised set of phenotyping procedures at set time points (Dickinson et al. 2016). Each procedure tests multiple parameters, and these parameters are tested for statistical deviation from a control group of non-knockout mice. Parameters that statistically deviate from the control are annotated with Mouse Phenotype Ontology (MP) terms (C. L. Smith and Eppig 2012).

I classified 5,252 knockouts with a NOER into one of four categories. Knockouts were classified as lethal (L) if they are annotated as causing preweaning lethality in all knockouts (n = 1,372); sub-viable (SV) if they are annotated as causing preweaning lethality in one or more, but not all, knockouts (n = 519); viable phenotype (VP) if they are not classified as L or SV, but are annotated with one or more other MP term (n = 2,993); and viable 'no phenotype' (VN) if they were subjected to one or more phenotyping procedure but are not annotated with any MP terms (n = 368) (i.e. knockouts that did not pass a significance threshold of 0.0001 in one of the associated phenotyping parameters). I calculated odds ratios to analyse gene set enrichment for each grouping in the 10% most constrained IMPC knockouts (i.e. nonsynonymous *Z*-score percentile >= 91), and the 10% least constrained genes (i.e. nonsynonymous *Z*-score percentile <= 10) (Supplementary Table 3.1). For example, to assess the enrichment of L knockouts among the most constrained genes I used the formula:

$$(A/B)/(C/D)$$

Where A is the number of knockouts classified as L with a NOER percentile >= 91; B is the number of genes not classified as L with a NOER percentile >= 91; C is the number of genes classified as L with a NOER percentile < 91; B is the number of genes not classified as L with a NOER percentile < 91. 95% confidence intervals were calculated as:

$$1.96*\sqrt{(1/A) + (1/B) + (1/C) + (1/D)}$$

I fit simple logistic regression models to predict knockouts classified as L or SV, and knockouts classified as VN, as a function of gene constraint (measured as NOER and dN/dS). I assessed model fit and predictive accuracy as the area under the receiver operating characteristic (ROC) (Supplementary Table 3.3).

I defined pleiotropy as the proportion of phenotypes associated with the IMPC knockout. Two measures of pleiotropy were calculated for each IMPC knockout. First, the proportion of procedures the knockout was subjected to in which one or more parameters significantly deviates from the control group (i.e. passes a significance threshold of 0.0001). Second, the number of top-level mouse phenotype ontology terms associated with the knock-out is divided by the potential number of top-level terms based on the conducted phenotyping tests. I filtered knockouts that cause lethality, sub-viability, or were subject to less than five phenotyping procedures, as these knockouts are subject to fewer phenotyping procedures which could skew the results. I calculated two Spearman's Rank correlations 1) between the median proportion of phenotyping procedures with one or more significant parameters per mouse knockout and the NOER percentile bin, and 2) between the median proportion of unique top-level mouse phenotype ontology terms per mouse

knockout and the NOER percentile bin. The MP ontology is a directed acyclic graph, and it is possible for one MP term to have multiple top-level terms. I, therefore, counted only one top-level MP term per MP term.

3.5.6 Defining orthologues and correlation with other measures of constraint

Human-mouse orthologues were defined by Ensembl (v94) (Zerbino et al. 2018). I tested whether mouse gene constraint measured as NOER was correlated (Spearman's rank) with *Mus musculus* to *Mus spretus* dN/dS and previously published measures of intraspecific human constraint, including the missense *Z*-score (Samocha et al. 2014), LOEUF (Karczewski et al., n.d.), and pLI (Lek et al. 2016).

3.5.7 Assessing pathogenic variant enrichment in the human orthologues

Human pathogenic and benign variants were downloaded from ClinVar (Landrum et al. 2018) on 02/03/2021. I classified SNV sites with clinical significance labelled as either "Pathogenic" or "Likely pathogenic", as pathogenic, and those with "Benign" or "Likely benign" as benign. All benign and pathogenic SNVs considered have a review status labelled as either "criteria provided, multiple submitters, no conflicts", "criteria provided, single submitter", or "reviewed by expert panel". The Ensembl

canonical transcripts and coding consequences for these SNVs were determined using the Ensembl Variant Effect Predictor (v94.5) (McLaren et al. 2016). I considered SNVs with consequences defined as either “synonymous_variant”, “stop_retained_variant”, “start_retained_variant” “missense_variant”, “stop_gained”, “start_lost”, “stop_lost”, or “protein_altering_variant”. I considered human genes with one or more pathogenic or benign variant sites, and with a one-to-one mouse orthologue for which NOER is calculated. This left 15,501 pathogenic variant sites and 37,032 benign variant sites for analysis across 7,037 human genes. To account for differences in gene length, I averaged pathogenic variants in each human gene per kb. I calculated Spearman’s rank correlations between the mean number of pathogenic variants per kb and the NOER percentile bin and the mean number of benign variants per kb and the NOER percentile bin.

Chapter 4

Human-mouse conservation of non-coding features and disease-associated loci

This chapter is adapted from a manuscript published in Biology Letters, on which I am the first author (Powell et al. 2022). Topologically Associated Domain (TAD) boundary annotations used in the analysis were annotated and provided by Helen Long, and she provided the description of how this was achieved in the methods. I confirm that I produced all other material with no more input/advice than would be considered appropriate or acceptable for a thesis.

4.1 Abstract

Understanding the genetic aetiology of loci associated with disease is crucial for developing preventative measures and effective treatments. Mouse models are used to infer human pathobiology and mechanistic functions of disease-associated loci, but their utility is limited by evolutionary divergence in regulatory mechanisms for pathways of interest. The human and mouse genomes have been extensively annotated for exonic and regulatory elements. However, the extent to which human regulatory annotations and disease-associated loci align to orthologous elements in the mouse genome remains unclear. Addressing this could help prioritise variants for

functional analysis using mouse models. In this chapter, I summarise the alignment of genomic (exonic and multi-cell regulatory) annotations alongside Mendelian and complex disease-associated variant sites between humans and mice. The results highlight the importance of understanding evolutionary divergence in transcription regulation when interpreting functional studies using mice as models for human disease variants.

4.2 Introduction

Understanding the mechanistic function of disease-associated loci is a fundamental challenge for biomedical research and is critical for the development of effective treatments and drug targets (Zuberi and Lutz 2016). Genome-Wide Association Studies (GWAS) have identified a myriad of variant sites associated with the risk of complex diseases (Edwards et al. 2013); however, the causal pathways of these loci remain poorly understood (Gallagher and Chen-Plotkin 2018). This is in part due to the relative difficulty of functional follow-up studies, which is compounded by the small and potentially interactive effects of variants, and the complexity of interpreting the function of non-coding regions, where the majority of GWAS variant sites are found (Cannon and Mohlke 2018).

The mouse is the most commonly used mammalian model for biomedical research (Peters et al. 2007; Eppig et al. 2015; Y. Shen et al. 2012; Vierstra et al. 2014) and has been used to infer the function of human disease variants. Mouse models have been particularly useful for elucidating the function of variants in protein-coding

transcripts, which are highly conserved between the species (Mouse Genome Sequencing Consortium et al. 2002), in addition to loci associated with traits that can only be measured *in vivo* such as body fat distribution or body mass index (Smemo et al. 2014). The mouse is also the only non-human mammal with regulatory feature occupancy data from genomic assays catalogued by the Encyclopedia of DNA Elements (ENCODE) (Yue et al. 2014; ENCODE Project Consortium et al. 2020). It is therefore uniquely suited to serve as a model for understanding regulatory feature function, with further potential for modelling human disease loci through humanization of the mouse genome using CRISPR/Cas9 technologies (Zhu et al. 2019; Devoy et al. 2011; Peters et al. 2007). Studies have mapped human GWAS variants associated with given disease phenotypes to the mouse genome and shown an enrichment in regions linked to transcription regulation (Yue et al. 2014; Cheng et al. 2014; Cusanovich et al. 2018; Hook and McCallion 2020). Studies have also, however, highlighted the substantial divergence in tissue and/or cell-specific transcription regulation between the species (Monaco et al. 2015; Denas et al. 2015; Yue et al. 2014; ENCODE Project Consortium et al. 2020; Lin et al. 2014; Cheng et al. 2014) making it unclear in which instances the mouse can recapitulate mechanisms of human gene expression to sufficiently model the function of human disease-associated genetic variants (Laber and Cox 2017).

The Ensembl Regulatory Build amalgamates datasets from various consortia, including ENCODE, to annotate predicted regulatory sequences across the human and mouse genomes (Zerbino et al. 2015). These annotations are continually updated as more data becomes available, and importantly, have stable identifiers to provide a reference framework for ongoing research. It is, however, currently unreported how human-mouse alignment compares across the spectrum of annotation categories.

Furthermore, it remains unclear how Mendelian and complex disease-associated variant site alignment varies between different regulatory annotations. Addressing these two questions would provide a useful reference point for researchers considering mouse models for human disease-associated loci.

Here, I use genomic annotation from Ensembl to provide a genome-wide overview of sequence alignment for twelve categories of annotation (including exonic and regulatory features) between humans and mice. I assess the alignment of Mendelian and complex disease-associated variant sites between the species across these annotation categories and discuss the implications of our results for the use of mouse models to understand the mechanistic function of human disease loci.

4.3 Results and Discussion

The human and mouse genomes have been annotated genome-wide by the Ensembl Regulatory Build (Zerbino et al. 2015) and GENCODE (Harrow et al. 2012), and I used these two sources to annotate all base-pair positions across the autosomes for both species (see Methods and Supplementary Information). Species genomes that have diverged over evolutionary time can be aligned to identify orthologous loci (R. S. Harris 2007). Throughout this thesis, I define human bases as aligned if they have an orthologous base in the mouse genome (i.e. if they have a corresponding genomic position in the pairwise alignment conducted by Ensembl (Herrero et al. 2016)), independent of whether a point mutation has occurred. I summarise the overall fraction of human bases that align to the mouse genome for each annotation category (Figure 4.1). I describe the fraction of human bases with a given annotation that align

to bases with the same annotation in the mouse genome as having common annotation.

In total, 29.3% of the human autosome aligns with the mouse; however, the proportion of the autosome that can be aligned varies by genomic annotation (Figure 4.1, Supplementary Table 4.3). Human translated exons (95.5%), proximal intronic sequences which include splice sites (77.3%), 3` and 5` untranslated regions (UTRs) (67.6%), and promoters (59.4%) show a relatively higher degree of alignment to the mouse genome than other exonic and regulatory annotations, including proximal and distal enhancers (41.8% and 45.3% respectively), miscellaneous sequences (39.2%), CTCF binding sites (41.4%), and topologically associated domain (TAD) boundaries (32.4%). The fraction of bases that align to the mouse and have common annotation in both species provides a coarse measure of feature conservation. The fraction of common annotation varies by annotation category and is greatest for translated exonic sequences (88.2%) followed by proximal intronic regions (56.0%), UTRs (40.8%), and promoters (38.5%) (Figure 4.1, Supplementary Table 4.3). In contrast, the fraction of common annotation is lower for proximal and distal enhancers (9.2% and 3.1% respectively) (Figure 4.1, Supplementary Table 4.3). This is consistent with previous research highlighting the rapid rate of enhancer turnover relative to promoters across mammalian species (Y. Shen et al. 2012; Villar et al. 2015; Berthelot et al. 2018; Fish, Chen, and Capra 2017; Huh et al. 2018). However, because enhancers can act as tissue-specific cis-regulatory elements (Sethi et al. 2020) and the human and mouse regulatory builds are constructed using different amalgamations of tissue sources (Zerbino et al. 2015), some enhancer annotation and alignment may not have been captured by the multicell annotation model. I assessed the alignment of tissue-specific enhancers by comparing the alignment of

enhancers active in adult heart, liver and spleen samples (Supplementary Table 4). Tissue-specific enhancers have a comparable alignment (ranging from 42.2% to 51.0% for proximal enhancers and 43.5% to 60.2% for distal enhancers) to the multicell model but were less conserved (ranging from 1.6% to 5.1% for proximal enhancers and 0.4% to 2.7% for distal enhancers).

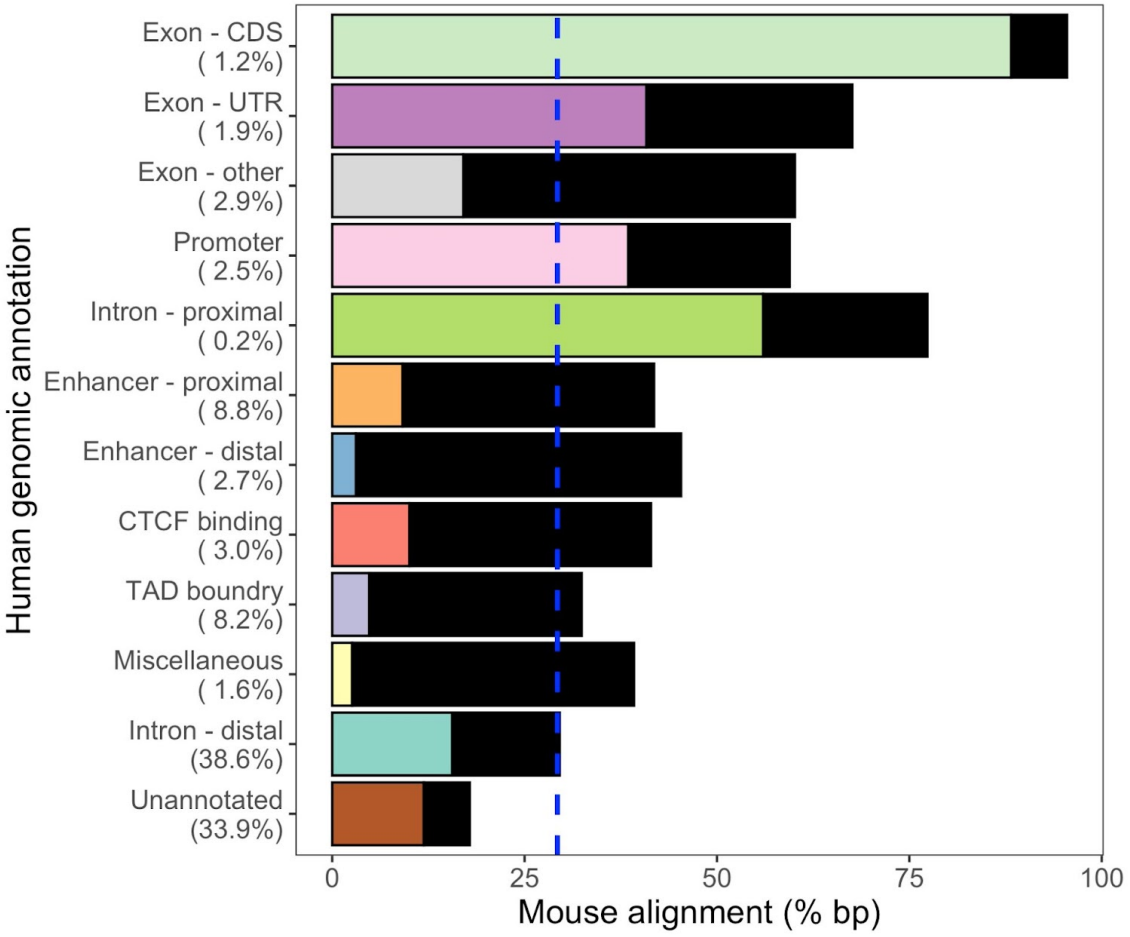


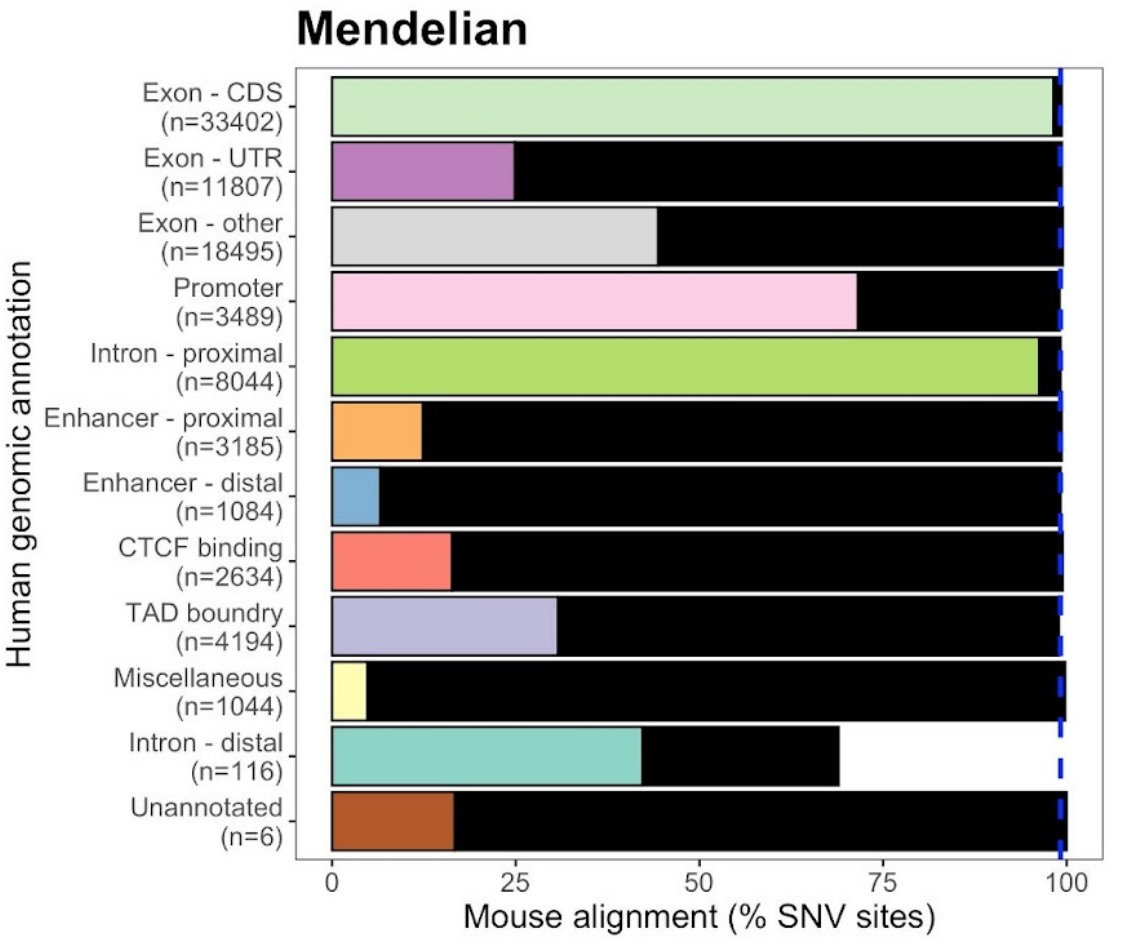
Figure 4.1– Alignment of genomic annotations between humans and mice. Bars represent the percentage of human bases that align with the mouse. Coloured bars represent the percentage of aligned bases with a common annotation (i.e. the same annotation in each species). Black bars represent the percentage of bases that align to a different annotation in the mouse (i.e. do not have a common annotation). The blue line represents the genome-wide percentage of human bases with an orthologous base in the mouse. The genomic coverage for each human annotation is labelled in brackets on the Y-axis. The sum of coverage is greater than 100% due to the overlap

of annotations (Supplementary Table 4.2). Human protein-coding sequences show the greatest alignment to the mouse genome (95.5%). The fraction of human annotation that aligns to the same annotation in mice is highest for protein-coding sequences (88.2%) proximal intronic sequences (56.0%), untranslated regions (UTRs) (40.8%), and promoters (38.5%), and lowest for distal enhancers (3.1%), TAD boundaries (4.8%) and miscellaneous (2.6%).

It is important to determine the similarities and differences in regulatory architecture between humans and mice when considering using a mouse model to infer the mechanistic function of human disease-associated variants (Laber and Cox 2017). I assessed the alignment of human variant sites predicted to cause Mendelian disease and human variant sites associated with complex disease with the mouse genome by considering two datasets: SNV sites predicted to cause Mendelian disease from ClinVar (Landrum et al. 2018) (n=42,039), and SNV sites associated with complex disease from the GWAS Catalog (Buniello et al. 2019) (n=27,794). As expected, both Mendelian and complex disease-associated variant sites in translated human sequences (Exon - CDS) have a high degree of alignment to the mouse genome (99.3% and 95.8% respectively: Figure 4.2, Supplementary Table 4.5). Across non-protein-coding sequences (i.e. loci not classified as Exon - CDS hereafter referred to as non-coding), 98.4% of pathogenic variant sites predicted to follow Mendelian inheritance patterns have an orthologous position in the mouse genome (Figure 4.2, Supplementary Table 4.5). This is significantly more than the genome-wide average of 28.8% for non-coding loci ($z = 139.5$, $p < 1.0e\text{-}300$), and indicates these sites have had a higher probability of being constrained by local purifying selection, potentially as a result of functional importance, since the species divergence. There is, however, variation in the fraction of SNV sites that align to the same annotation in mouse between regulatory elements. For example, 70.8% of Mendelian pathogenic SNV sites in human promoter sequences align to mouse

promoter sequences, yet only12.2% of Mendelian pathogenic SNV sites in human proximal enhancers and 6.5% in human distal enhancers align to loci with the same annotation in mice (Figure 2, Supplementary Table 5). This difference suggests that while these loci may have had a higher probability of preservation due to local purifying selection in both lineages, the active regulatory elements and functional pathways at these variant sites have diverged. It must be noted, however, that some similarities may be missed due to regulatory feature specificity and differences in the tissue amalgamations used to annotate regulatory features.

A significantly smaller fraction of non-coding variant sites associated with complex disease aligns with the mouse genome than non-coding variant sites predicted to follow Mendelian inheritance patterns (36.4% compared with 98.4%, $z = 97.7$, $p < 1.0e\text{-}300$: Figure 4.2, Supplementary Tables 4.5 and 4.6). One explanation for this may be the small effect size of variant sites associated with complex disease having limited fitness effects (Visscher et al. 2017). Distal introns and unannotated regions contain the majority (62.0%) of variant sites associated with complex disease making their effect on transcription regulation difficult to infer. However, a significantly greater fraction of variant sites associated with complex disease in these regions aligns with the mouse genome than the total fraction of bases with these annotations: 34.3% compared with 29.6% ($z = 10.7$, $p = 9.1e\text{-}27$) for distal introns and 25.8% compared with 17.9% ($z = 16.5$, $p < 4.9e\text{-}61$) for unannotated variants (Supplementary Table 4.8). This suggests that the functional role of loci within regions annotated as “Intron - distal” and “Unannotated” has not been captured by the annotation model and may discourage the production of mouse models for these variants.


Mendelian
Human genomic annotation
Exon - CDS (n=33402)
Exon - UTR (n=11807)
Exon - other (n=18495)
Promoter (n=3489)
Intron - proximal (n=8044)
Enhancer - proximal (n=3185)
Enhancer - distal (n=1084)
CTCF binding (n=2634)
TAD boundry (n=4194)
Miscellaneous (n=1044)
Intron - distal (n=116)
Unannotated (n=6)
0
25
50
75
100
Mouse alignment (% SNV sites)

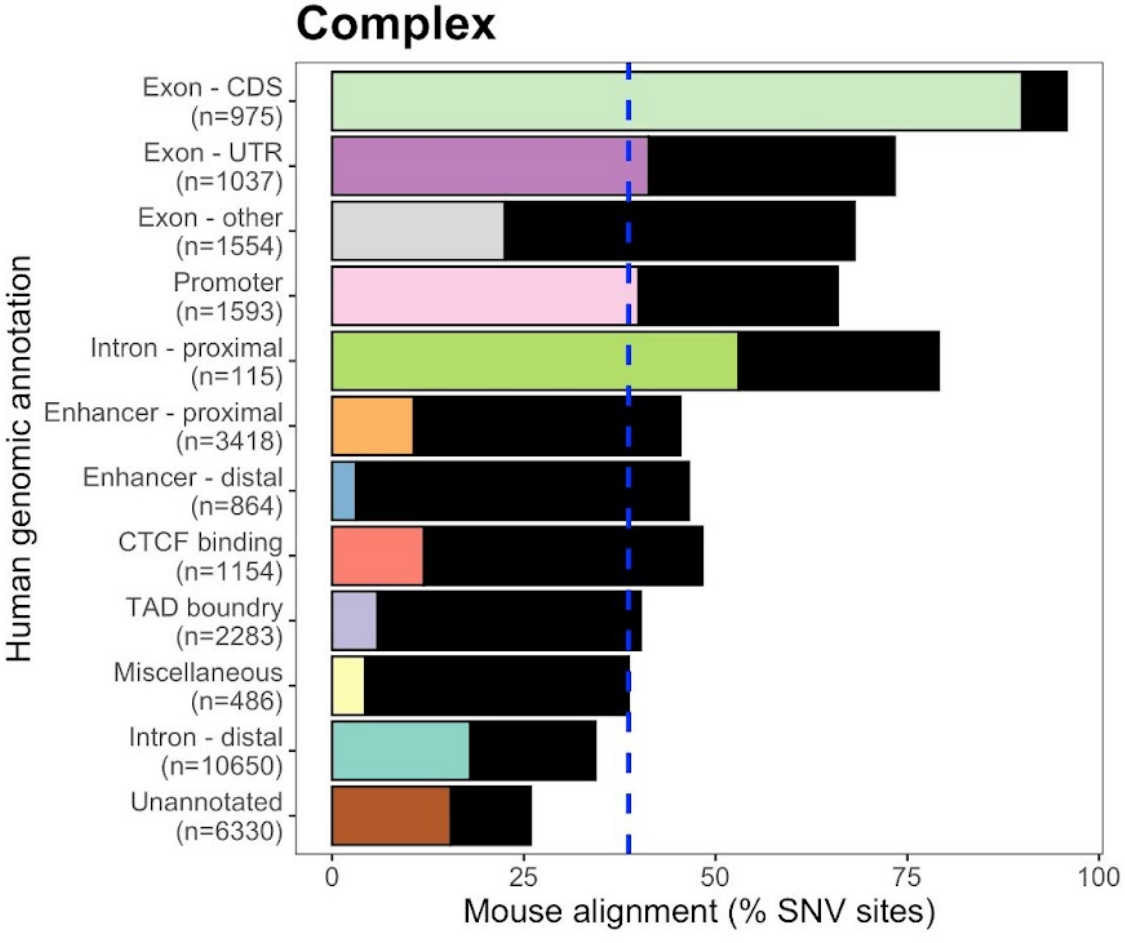

Complex
Human genomic annotation
Exon - CDS (n=975)
Exon - UTR (n=1037)
Exon - other (n=1554)
Promoter (n=1593)
Intron - proximal (n=115)
Enhancer - proximal (n=3418)
Enhancer - distal (n=864)
CTCF binding (n=1154)
TAD boundry (n=2283)
Miscellaneous (n=486)
Intron - distal (n=10650)
Unannotated (n=6330)
0
25
50
75
100
Mouse alignment (% SNV sites)

Figure 4.2 – Alignment of human SNV sites associated with complex disease (GWAS Catalog) and Mendelian disease (ClinVar) between humans and mice. Bars represent the percentage of human variant sites with an orthologous variant site in mice for each genomic annotation. Coloured bars represent the percentage of variant sites for each annotation that aligns to the same annotation in the mouse. Black bars represent the percentage of variant sites for each annotation that aligns to a different annotation in the mouse. Blue lines represent the total percentage of variant sites with a mouse orthologue. Variant sites associated with human Mendelian disease are more conserved between the species than variant sites associated with human complex disease. However, annotation of non-exonic regulatory features (excluding the promoter) is poorly conserved, suggesting functional divergence between the species.

4.4 Conclusions

By comparing the mouse and human genomes, I found that 95.5% of human protein-coding sequence and 28.5% of human non-coding (untranslated) sequence aligns with the mouse genome. Furthermore, 98.4% of human non-coding variant sites associated with Mendelian disease align with the mouse genome, compared with 36.4% of non-coding variant sites associated with complex disease. The degree of overall divergence in the regulatory landscape between humans and mice highlights the importance of understanding the differences between functional pathways of interest when using mouse models to infer human disease mechanisms.

4.5 Methods

4.5.1 Calling topologically associated domains (TADs)

TAD boundaries were annotated by Helen Long and provided in bed file format. In brief, raw Hi-C data was downloaded from Gene Expression Omnibus (GEO) using

the following accessions: GSE86821 (Human ESC (Freire-Pritchett et al. 2017)), GSE58752 (Human liver (Leung et al. 2015)), GSE96107 (Mouse ESC (Bonev et al. 2017)) and GSE104129 (Mouse liver (ZT10) (Kim et al. 2018)). Reads were analysed using the Juicer pipeline aligning to GRCh38 and GRCm38 (Durand et al. 2016). For datasets containing multiple replicates (Mouse ESC and Mouse liver), each replicate was separately analysed using the Juicer pipeline and then merged using the juicer "mega" script. Only reads passing a mapping quality (MAPQ) threshold of 30 were considered and Hi-C maps were normalised using the "Vanilla coverage" (VC) method. TADs were then called in each dataset using the Arrowhead TAD calling algorithm at 50kb bin size with default parameters (Durand et al. 2016). TAD boundaries were defined as the region ±25kb from the start and end of each called TAD. A ±25kb interval was chosen because, as TADs were called on 50kb binned Hi-C data, we reason that the true location of the boundary is likely to fall within this interval[4].

4.5.2 Compiling an annotation model of genomic features across the human and mouse genomes

Human and mouse reference genomes (GRCh38 and GRCm38) were downloaded from Ensembl (v101) (Zerbino et al. 2018). I filtered ambiguous bases referenced as "N", which are primarily located in the telomeres and centromeres, and account for 4.1% and 2.7% of the human and mouse genomes, respectively.

[4] This description for how TAD boundaries were annotated was provided by Helen Long.

I annotated exonic and multi-cell regulatory features autosome-wide for humans and mice by compiling exonic and regulatory annotations from Ensembl (v101) (Zerbino et al. 2015, 2018). I developed the genome-wide multi-cell annotation model defined by (di Iulio et al. 2018), and annotated each nucleotide across the human and mouse autosomes with one or more of the following categories:

1. **"Exon - CDS"** for translated nucleotides in protein-coding exons (including immunoglobulin and T cell receptor genes) as annotated by Ensembl.
2. **"Exon - UTR"** for 5' UTR and 3' UTR nucleotides in protein-coding exons as annotated by Ensembl.
3. **"Exon - other"** for nucleotides in non-protein-coding exons (notably ncRNAs and lncRNAs) as annotated by Ensembl.
4. **"Intron - proximal"** for intronic nucleotides inferred from exon coordinates as annotated in Ensembl, that are located within 10bp of a splice-site position.
5. **"Promoter"** for nucleotides annotated by the Ensembl Regulatory Build.
6. **"Enhancer - proximal"** for nucleotides annotated as promoter flanking by the Ensembl Regulatory Build.
7. **"Enhancer - distal"** for nucleotides annotated as enhancer by the Ensembl Regulatory Build.
8. **"CTCF binding"** for nucleotides annotated by the Ensembl Regulatory Build.
9. **"Miscellaneous"** for nucleotides categorised as unannotated transcription factor binding site or unannotated open chromatin by the Ensembl Regulatory Build.

10. **“TAD boundary”** for nucleotides within 25 kb upstream and downstream of a TAD start or end coordinate.
11. **“Intron - distal”** for intronic nucleotides inferred from exon coordinates as annotated in Ensembl, that are located more than 10bp from a splice-site position, and do not have any other annotation.
12. **“Unannotated”** for all nucleotides without any annotation.

A summary of the genomic coverage for each annotation is provided in Supplementary Table 4.1. Annotation overlap is summarised in Supplementary Table 4.2.

4.5.3 Human and mouse pairwise alignment

Species genomes that have diverged over evolutionary time can be aligned to identify orthologous regions. I used human-mouse pairwise alignment conducted by Ensembl (Herrero et al. 2016) using LastZ (R. S. Harris 2007) to define orthologous bases between humans and mice. I define bases as orthologous if their genomic coordinates align, independent of whether a point mutation has occurred. Orthologous positions are defined as having common annotation if they have the same annotation in both species. The alignment of the human and mouse genomes is summarised in Supplementary Table 4.3.

4.5.4 Tissue-specific enhancers

Enhancers are tissue-specific, and the human and mouse regulatory builds are multicell (ie constructed using different amalgamations of tissue sources) (Zerbino et al. 2015). I, therefore, compared the multicell model to three tissue-specific models. Tissue-specific regulatory builds were downloaded from Ensembl (v101), and include heart, liver, and spleen samples, with the mouse samples taken from adult mice at 8 weeks of age. Enhancers were deemed active if their cell type-specific activity levels are described as either “ACTIVE” or “POISED”, but were otherwise defined as previously described. The alignment of tissue-specific human enhancers with the mouse genome is summarised in Supplementary Table 4.4.

4.5.5 Defining human variant sites associated with Mendelian and complex disease

Human single nucleotide variant (SNV) sites associated with Mendelian disease were downloaded from ClinVar (Landrum et al. 2018) (variant_summary.txt downloaded on 02/03/2021). I considered all SNV sites with clinical significance labelled as either “Pathogenic” or “Likely pathogenic”, and a review status labelled as either "criteria provided, multiple submitters, no conflicts", "criteria provided, single submitter", or "reviewed by expert panel" (n=42,039). Human SNV sites associated with complex disease were obtained from the GWAS Catalog (Buniello et al. 2019), and have a phenotype that is ontologically classified as either disease, disorder, or cancer, and a p-value < e10-8 (n=27,794). The alignment of human SNV sites associated with complex and Mendelian disease is summarised in Supplementary Table 4.5.

4.5.6 Assessing the alignment of human variant sites associated with Mendelian and complex disease to the mouse genome

I assessed the proportions of human Mendelian and complex disease-associated variant sites that align to the mouse genome for each annotation category. For each annotation category I tested three null hypotheses: 1) There is no difference in the proportion of SNV sites associated with Mendelian disease that align to the mouse genome and the proportion of SNVs associated with complex disease that align to the mouse genome (Supplementary Table 4.6); 2) There is no difference in the proportion of SNV sites associated with Mendelian disease that align to the mouse genome and the total proportion of bases that align to the mouse genome (Supplementary Table 4.7); 3) There is no difference in the proportion of SNV sites associated with complex disease that align to the mouse genome and the total proportion of bases that align to the mouse genome (Supplementary Table 4.8). I tested differences in proportions using two-proportion z-tests, using the formula:

$$Z = \frac{P_1 - P_2}{\sqrt{P(1-P)/(\frac{1}{N_1} + \frac{1}{N_2})}}$$

Where P1 and P2 are the proportions for comparison, P is the total proportion (i.e. the sum of the P1 and P2 numerators divided by the sum of the P1 and P2 denominators), and N1 and N2 are the sample size of P1 and P2 respectively (ie their denominators). P-values were calculated from a two-tailed hypothesis. Proportions of conserved and aligned sites can be calculated using Supplementary Tables 4.3 and

4.5, and Z-scores and p-values are provided in Supplementary Tables 4.6, 4.7, and 4.8.

Chapter 5

Characterising the mouse non-coding genome by genetic diversity

5.1 Abstract

Patterns of genetic diversity can be used to infer regions of the genome that are constrained by negative selection and functionally important with regard to organism fitness. Constrained regions are expected to show a relative depletion of variation as new variants are more likely to be deleterious and therefore purged. Genetic variation from whole-genome sequencing in human populations has been used to characterise constrained non-coding regions across the genome. These regions are enriched for functional elements and disease-associated loci and deviate from traditional estimates of between species conservation. Here, I characterise the mouse genome by patterns of genomic variation between populations of wild mice (*Mus musculus sp* and *Mus spretus*). I show that the most constrained regions across the mouse genome are enriched for exonic and regulatory features, as expected. Furthermore, I highlight a positive correlation between regional constraint and the number of conserved human pathogenic variant sites across the mouse genome. My estimates of constraint provide a resource for interpreting the fitness consequences

of genetic variants in mouse non-coding regions and prioritising human disease-associated variants for further analyses using mouse models.

5.2 Introduction

Understanding the functional consequences of genetic variants associated with disease is a priority for biomedical research and facilitates the development of effective drug treatments (Nadeau and Auwerx 2019; Edwards et al. 2013; Crunkhorn 2020). Mouse models are critical to our understanding of mammalian biology and have been used to infer the functional consequences of human variant sites associated with disease *in-vivo* (Nadeau and Auwerx 2019; ENCODE Project Consortium et al. 2020; Dickinson et al. 2016). This is especially the case for variants in protein-coding genes, which are highly conserved between the species (see chapter 4), and where the consequences of mutations on amino-acid sequences are readily understood (Devoy et al. 2011). However, the majority of human variants associated with disease identified via Genome-Wide Association Studies (GWAS) are located in non-protein-coding (hereafter non-coding) regions which constitute around 98% of the genome (Powell et al. 2022; Cannon and Mohlke 2018). The functional consequences of variants in non-coding regions are difficult to interpret, but they are expected to alter organismal phenotype via their effect on gene expression through transcriptional or post-transcriptional regulation (Edwards et al. 2013; Watanabe et al. 2019; Gallagher and Chen-Plotkin 2018). Mouse models contribute to our understanding of regulatory feature function via experimental assays and are the only non-human mammal to have had their regulatory features mapped by The

Encyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium et al. 2020; Yue et al. 2014). These assays have highlighted an enrichment of human non-coding GWAS variants in mouse regulatory features (Yue et al. 2014; Cheng et al. 2014; Cusanovich et al. 2018; Hook and McCallion 2020), but also substantial divergence in regulatory pathways between the species (Monaco et al. 2015; Denas et al. 2015; Yue et al. 2014; ENCODE Project Consortium et al. 2020; Lin et al. 2014; Cheng et al. 2014). Questions remain, therefore, about the effect of variation in non-coding regions across the mouse genome, and mice' suitability for modelling human non-coding variants.

Patterns of genomic variation can be used to infer regions of the genome that are constrained by negative selection and functionally important with regard to fitness (Bartha et al. 2018; Davydov et al. 2010; di Iulio et al. 2018). Constrained regions are expected to show a relative depletion of variation as new variants are more likely to be deleterious and therefore purged. Previous methods have identified constrained loci by considering sequence conservation between distantly related species (Davydov et al. 2010). However, these methods have limited ability to capture the full spectrum of constrained non-coding regulatory sequences, due in part to changes in selection between lineages over evolutionary time and the high turnover rate of regulatory elements (see chapter 4) (Huber, Kim, and Lohmueller 2020; Powell et al. 2022). To minimise the loss of information from considering variation between distantly related species, di lulio *et al.* (2018) used patterns of genomic variation from human populations to identify non-coding regions of the human genome that are depleted of variation and inferred to be constrained by negative selection. The authors found average differences in selective constraint between different regulatory elements, and enrichment of disease-associated variant sites in the most

constrained non-coding regions of the genome (di Iulio et al. 2018). These findings are important for interpreting the human genome, and prioritising variants that are likely to be pathogenic; however, a comparable analysis had not been performed in the mouse. Mapping sequence constraint across the mouse genome could help decipher the relationships between functional elements and fitness, and help determine when mouse models are most likely to be suitable for modelling human pathogenic variant sites.

To provide a resource for interpreting the fitness consequences of genetic variants in mouse non-coding regions, I characterise the mouse genome by patterns of genomic variation between populations of wild mice (*Mus musculus sp* and *Mus spretus*). I show that the most constrained regions across the mouse genome are enriched for exonic and regulatory features, as expected. Furthermore, I highlight a positive correlation between regional constraint and the number of conserved human pathogenic variant sites across the mouse genome.

5.3 Results

5.3.1 Ranking constrained regions across the mouse genome

I built a map of constrained regions covering 52.2% of the mouse autosome by considering patterns of standing variation between populations of wild mice (*Mus musculus sp* and *Mus spretus*). Sequence constraint was characterised in 100 bp regions using a sliding window method. Windows were 1 kb in length sliding every

100 bp with a constraint score, hereafter named the mouse-specific constraint score (MSCS), attributed to the middle 100 bp of each window (Supplementary Figure 5.1). The size of the window and sliding distance was selected as a compromise between resolution, power, and computational complexity. MSCSs were quantified for each 100 bp region by their residual from the regression of the weighted number of single nucleotide variants (SNVs) in each window as a function of the window's weighted probability of substitution. Probabilities of substitution were determined by the heptonucleotide (7-mer) context and methylation state at CG sites calculated in chapter 2. The 16,384 (4^7) distinct 7-mers and the methylation state of CG dinucleotides are not evenly distributed across the mouse genome and show a clear association with different genomic annotations (Figure 5.1). The window's probabilities of substitution were considered to control for differences in mutability between windows which could otherwise bias estimates of constraint. Regions with lower residual values (i.e. residuals < 0) and therefore lower MSCSs (i.e. fewer SNVs than expected given their mutability) are considered more constrained. Regions were ranked by their MSCS with the most constrained regions in the 1st percentile rank and least constrained regions in the 100th percentile rank. Distributions of the weighted number of SNVs and weighted probabilities of substitution across windows are provided in Supplementary Figure 5.2.

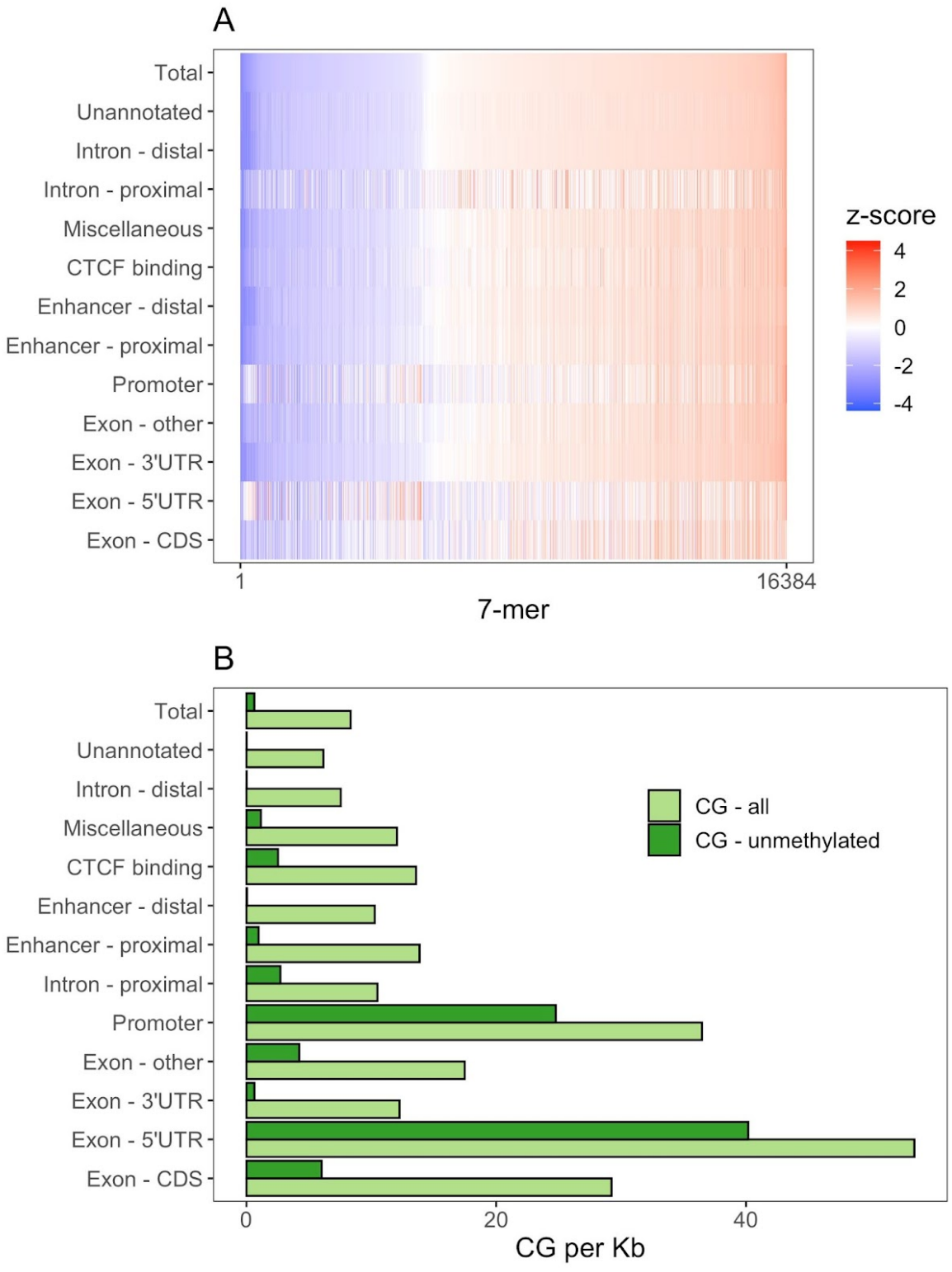


Figure 5.1 – (A) A heat map displaying the relative counts of each heptamer (n = 16,384) across each mouse genomic annotation category and the mouse autosome (Total). Heptamer counts for each category are standardised as Z-scores calculated from log transformed distributions, with a lower Z-score indicating lower occurrence frequency. (B) A bar plot displaying the number of CG and unmethylated CG dinucleotides per kb across each mouse genomic annotation category and the mouse

autosome (Total). Unmethylated CG dinucleotides were defined from bisulphate sequencing (see methods).

5.3.2 Feature composition of constrained regions

I characterised the genomic feature composition of the most and least constrained regions across the mouse autosome defined by MSCS. Genomic features were categorised using annotations from GENCODE and the Ensembl Regulatory Build (see methods). As expected, translated sequences (Exon - CDS) are highly enriched amongst the most constrained regions of the mouse genome (MSCSs in the 1st percentile), with a 133.3-fold increase in genomic territory relative to the least constrained regions (MSCSs in the 100th percentile) and a 10.2-fold increase relative to the autosome wide average (Figure 5.2). Other exonic features including 3' and 5' untranslated regions (UTRs), and intronic regions within 10 bp of an exon splice site (Intron - proximal), are greatly enriched in the most constrained regions relative to the least constrained (35.7, 22.4, and 26.3 fold respectively) (Figure 5.2). This is consistent with previous studies that report highly conserved sequences within 3'-UTRs of vertebrate species, enriched amongst housekeeping genes, and speculate this may reflect their importance in post-transcriptional regulation (Siepel et al. 2005).

After translated sequences, promoters have the greatest enrichment in the most constrained regions, with a 42.0-fold increase relative to the least constrained regions and a 4.6-fold increase relative to the genome-wide average (Figure 5.2). In comparison, distal and proximal enhancers show a 4.4 and 3.9 fold enrichment in

the most constrained regions relative to the least constrained (1.3 and 1.4 fold relative to the genome-wide average) (Figure 5.2). This is consistent with the higher evolutionary turnover rate of enhancers relative to promotors across mammalian species (as described between humans and mice in chapter 4) and indicates stronger purifying selection across promoters compared with enhancers (Villar et al. 2015), potentially as a result of variants having a greater average effect on gene expression.

Finally, unannotated sequences are depleted in the most constrained regions relative to the least constrained (82.0% decrease). It is noteworthy, however, that 10.0% of the most constrained regions across the mouse autosome are unannotated. This has two possible explanations: 1) the regions are functionally important but currently lack functional annotation. For example, Lilue et al. (2018) demonstrated how the GRCm38 reference genome was missing annotation for 62 exonic loci (19 protein-coding genes, 37 lncRNAs, and 7 pseudogenes), including a large 188 exon protein-coding gene (Gm11639) on chromosome 11. Or 2) the regions appear as a consequence of the limited specificity of the MSCS metric. For example, unconstrained regions may appear in the most constrained percentiles if they are in close proximity to highly constrained sequences. Intronic sequences (intron - distal) that are more than 10bp from a splice site and do not have any other annotation have negligible enrichment in the most constrained regions relative to the least (2.5% increase) and are marginally depleted relative to the genome-wide average (24.2% decrease).

I compared the enrichment of genomic features in the most constrained regions of the mouse autosome defined by MSCS with the most conserved regions defined by their average Genomic Evolutionary Rate Profiling (GERP) score. GERP provides an

estimate of selective constraint over a greater evolutionary timescale than MSCS by considering the conservation of sequences between multiple mammalian species. The most constrained regions defined by MSCS are more enriched for exonic and regulatory features than the most constrained regions defined by GERP (Supplementary Figure 5.3). This suggests models that estimate selective constraint over a shorter evolutionary timeframe have a greater capacity to identify regulatory elements, as changes in selection and sequence function can compound over greater evolutionary timeframes.

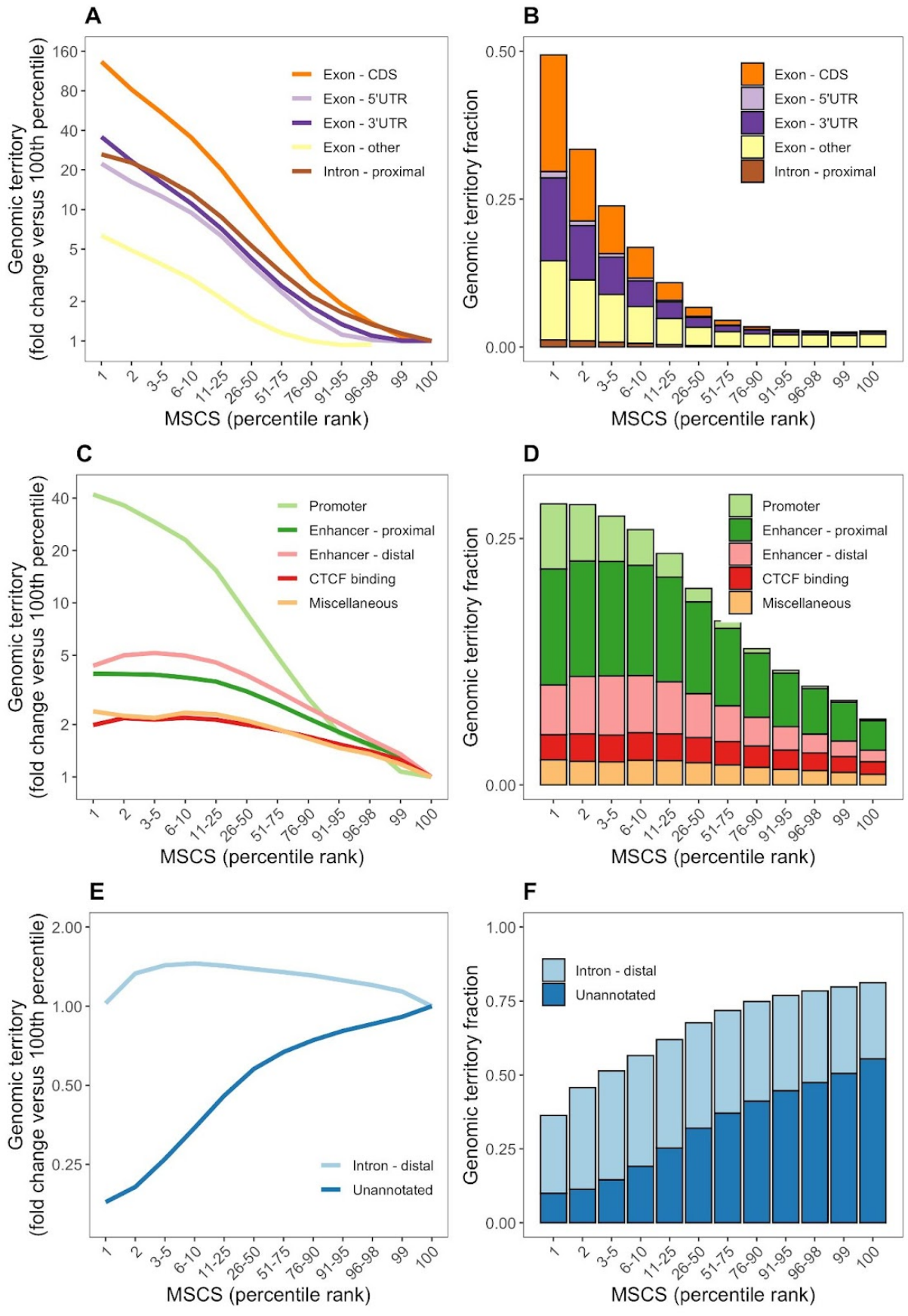

A
Genomic territory
(fold change versus 100th percentile)
160
80
40
20
10
5
2
1
Exon - CDS
Exon - 5'UTR
Exon - 3'UTR
Exon - other
Intron - proximal
1
2
3-5
6-10
11-25
26-50
51-75
76-90
91-95
96-98
99
100
MSCS (percentile rank)
B
Genomic territory fraction
0.50
0.25
0.00
Exon - CDS
Exon - 5'UTR
Exon - 3'UTR
Exon - other
Intron - proximal
MSCS (percentile rank)
C
Genomic territory
(fold change versus 100th percentile)
40
20
10
5
2
1
Promoter
Enhancer - proximal
Enhancer - distal
CTCF binding
Miscellaneous
MSCS (percentile rank)
D
Genomic territory fraction
0.25
0.00
Promoter
Enhancer - proximal
Enhancer - distal
CTCF binding
Miscellaneous
MSCS (percentile rank)
E
Genomic territory
(fold change versus 100th percentile)
2.00
1.00
0.50
0.25
Intron - distal
Unannotated
MSCS (percentile rank)
F
Genomic territory fraction
1.00
0.75
0.50
0.25
0.00
Intron - distal
Unannotated
MSCS (percentile rank)

Figure 5.2 – Genomic feature composition of the most and least constrained regions across the mouse autosome defined by the mouse-specific constraint score (MSCS) percentile rank. Genomic features are defined by their GENCODE and Ensembl Regulatory Build annotations (see methods). Panels B, D, and F show the fraction of genomic territory of each percentile grouping annotated with each genomic feature. Panels A, C, and E show the change in the genomic territory of each percentile grouping relative to the least constrained regions (100th percentile). Note that MSCS percentiles (x-axis) are not uniformly grouped to highlight the changes in the most and least constrained regions.

5.3.4 Human pathogenic variant site enrichment across the most constrained regions of the mouse genome

It is important to understand the evolutionary divergence between humans and mice in order to prioritise when mice may be suitable for modelling the mechanistic function of human variant sites associated with disease, and contextualise inferences about human biology. This is particularly pertinent for non-coding variants as the non-coding genome has diverged at a much greater rate between humans and mice than for protein-coding sequences (see chapter 4) (ENCODE Project Consortium et al. 2020; Cheng et al. 2014). I hypothesised that human variant sites that are associated with disease and align with the mouse genome are more likely to occur in more constrained regions of the mouse genome. To test this hypothesis I compared two sets of human variants that have an orthologous position in the mouse genome: pathogenic single nucleotide variants (SNVs) defined by ClinVar (n = 32,215); and SNVs associated with either disease, disorder, or syndrome, identified via genome wide-association studies (GWAS) (n = 7,776). These variant sets were classified further as coding DNA sequence (CDS) if they occur within translated sequences in the human genome, otherwise as non-CDS. As expected, there is a strong positive

correlation between MSCS and alignment to the human genome, with more constrained mouse regions having greater alignment (Figure 5.3). To eliminate this bias from my results the number of human disease-associated variant sites in each MSCS percentile was adjusted by the proportion of the percentile that aligns with the human genome (see methods). The number of orthologous loci of human variant sites (CDS and nonCDS) associated with disease is highly enriched amongst the most constrained regions of the mouse genome (Figure 5.3). However, this relationship changes when the enrichment is adjusted by the alignment fraction of each MSCS percentile (Figure 5.4). The number of orthologous loci of human CDS variant sites associated with disease is not enriched amongst the most constrained regions of the mouse genome relative to the genome-wide average (3.4% increase and 51.5% decrease for ClinVar and GWAS variants respectively). However, the number of orthologous loci of human CDS variant sites classified as pathogenic are depleted amongst the least constrained regions of the mouse genome relative to the genome-wide average (93.9% decrease). There is a correlation between selective constraint across mouse protein-coding genes and the number of human orthologous pathogenic variant sites they contain (see chapter 3). The lack of correlation between the MSCS percentile and the number of human orthologous pathogenic variant sites is therefore likely a consequence of the limited power of MSCS to distinguish the strength of constraint between protein-coding sequences. There is a clear positive association between mouse regional constraint (MSCS) and the number of aligned non-coding human sites that are associated with disease (Figure 5.4), indicating that human disease-associated variant sites in non-coding regions are more likely to align to regions of the mouse genome that are constrained by negative selection. If this reflects conserved function, this could support the use of mice to model the mechanistic function of given human disease-associated non-coding variants.

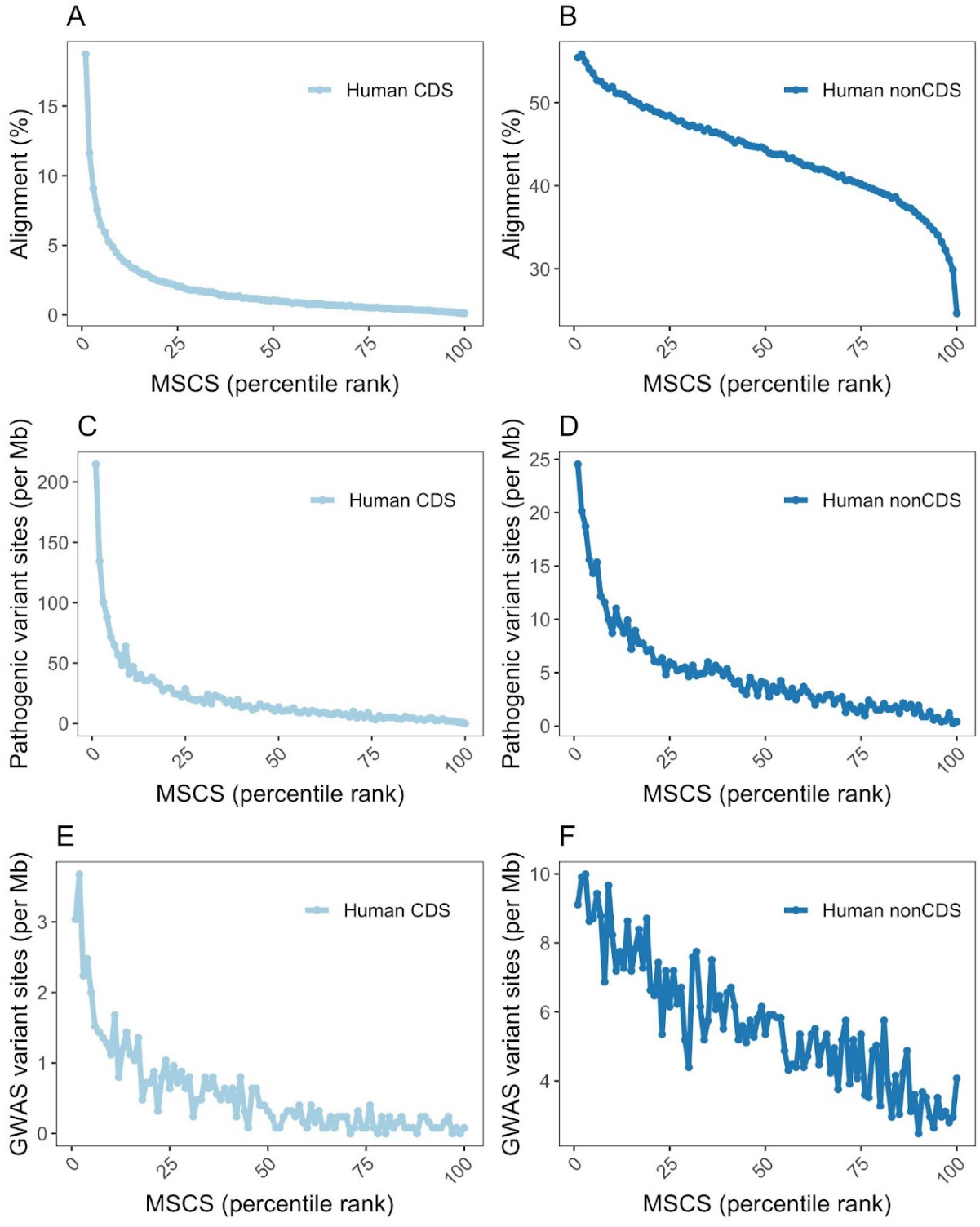


Figure 5.3 – (A and B) Relationship between mouse-specific constraint score (MSCS) percentile rank and alignment to human protein-coding sequences (Human CDS, panel A) and human non-coding sequences (Human non-CDS, panel B). More constrained mouse regions defined by MSCS are therefore more likely to contain

aligned human loci, and this is reflected in the relationship between MSCS and the number of human disease-associated variant sites (C, D, E and F).

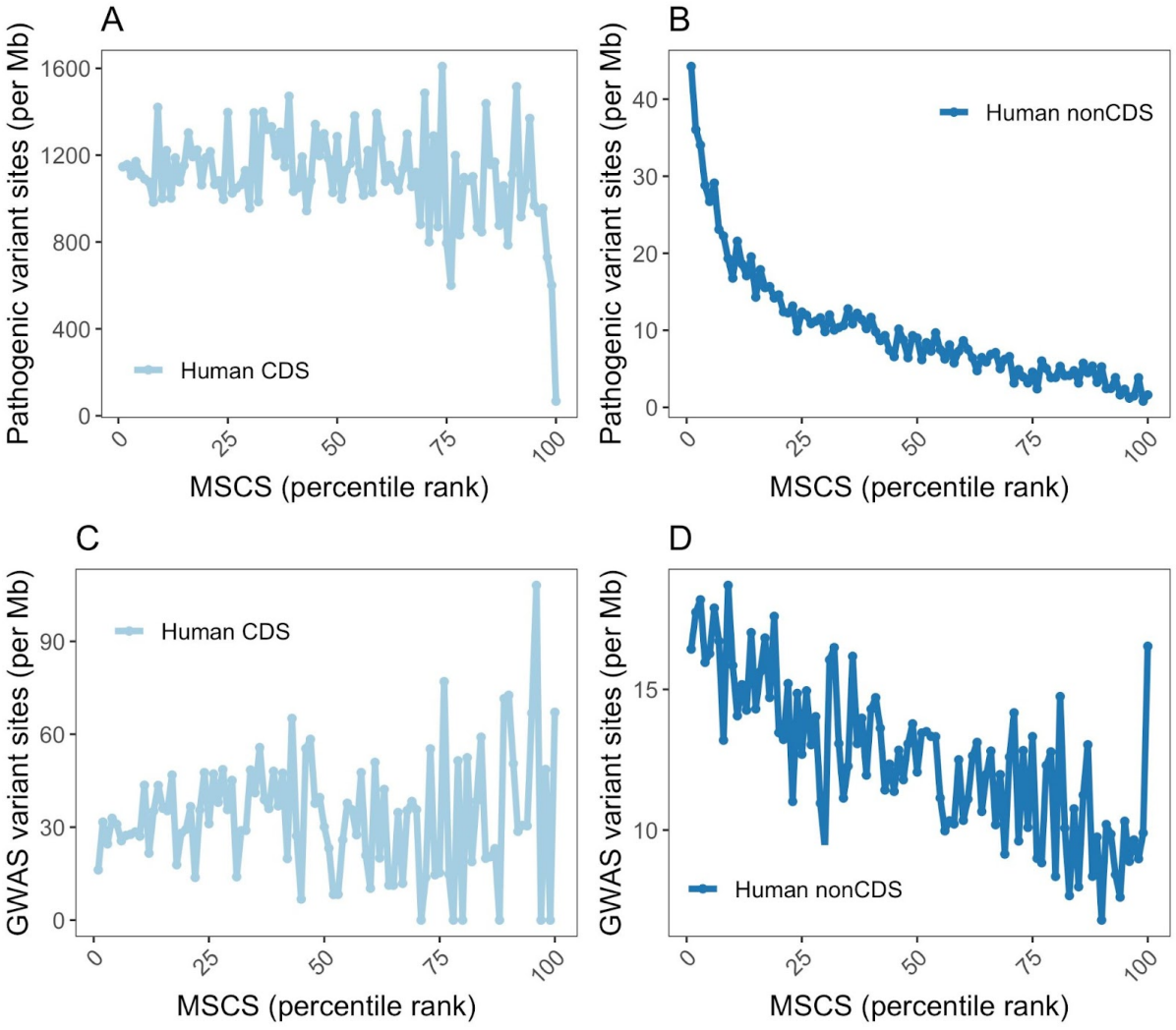


Figure 5.4 – The number of human pathogenic variant sites (A and B) and human GWAS variant sites (C and D) that align to the mouse genome by mouse-specific constraint score (MSCS) percentile rank. Pathogenic single nucleotide variant (SNV) sites are defined by ClinVar (see methods), and defined as human coding-sequence (CDS) (A) (n = 26,181) if they occur within translated sequences and human non-CDS (B) (n = 6,034) if they do not. Human GWAS SNV sites are associated with either disease, disorder, or a syndrome as defined by the GWAS catalog (see methods), and are also split into human CDS (C) (n = 705) and human non-CDS (D) (n = 7,071). A lower MSCS percentile indicates a greater degree of selective constraint (x-axis). The number of variant sites per Mb in each percentile is adjusted by the fraction of bases in the percentile that align between the human and mouse genome (y-axis), as there is a strong positive relationship between constraint and alignment that would otherwise bias the results (see supplementary information).

There is a positive relationship between the number of non-coding pathogenic and GWAS variants and the degree of selective constraint.

5.4 Discussion

I built a map of sequence constraint in 1kb windows across the mouse genome by considering patterns of standing variation between wild mice populations in order to help prioritise and infer functionally important non-coding regions. The most constrained mouse genomic regions defined by my metric are enriched for exonic and regulatory sequences. Furthermore, I demonstrate the number of human non-coding variant sites associated with disease are enriched amongst the most constrained mouse regions.

Previous research has shown promoters are more conserved than enhancers between humans and mice (Villar et al. 2015; Powell et al. 2022), but the drivers of this difference remain difficult to dissect. New enhancers are more likely to form from the exaptation of ancestral DNA and this could contribute to their higher evolutionary turnover rate compared with promoters, which are spatially limited to regions immediately upstream of their target genes (Villar et al. 2015). There are also differences in the sequence specificity of transcription factors (TFs) that bind to enhancers and promoters (Grossman et al. 2018), with binding motifs for transcription initiating TFs that more readily bind to promoters being highly conserved between diverse taxa (Nitta et al. 2015). If these differences result in new variants in promoters having a greater average effect on transcription regulation and

gene expression than those in enhancers, it may contribute toward explaining the differences in sequence constraint between the features.

The degree of sequence divergence in non-coding regions between humans and mice raises questions about when the mouse may be a suitable model to improve our understanding of the functional consequences of non-coding human disease-associated variants (Perlman 2016). I show that human non-coding variant sites associated with disease are enriched amongst the most constrained regions of the mouse genome, whilst controlling for sequence conservation between the species. This suggests they are more likely to occur in functionally active regions, but further research is required to determine on a case-by-case basis which regulatory pathways of interest have been sufficiently conserved. A comparison of sequence composition and constraint across orthologous cis-regulatory elements could contribute to this.

There are several limitations of my constraint metric. First, inferences of functional importance from estimates of sequence constraint rest on the assumption that functionally important regions are less tolerant of variation. Whilst this trend is likely to broadly apply, there are notable exceptions such as major histocompatibility complex (MHC) genes (Liston et al. 2021), and the previously discussed tolerance of enhancer sequences that have been demonstrated to be functionally active through experimental assays. Second, the MSCS is estimated at the kb resolution limiting its ability to detect regional variability at a finer scale. Greater sample sizes of sequenced mice from wild populations will increase the power to distinguish selective constraint at a finer resolution. Third, the MSCS does not account for the effects of linked selection – where selection at a given locus reduces the genetic diversity in linked neutral regions. Background selection and classic sweeps explain approximately 20%

of the variance in genetic diversity at the kb scale in drosophila (Elyashiv et al 2016). In wild mice, genetic diversity is reduced by an average of ~16% in protein-coding sequences, with this reduction incrementing down across the surrounding ~100kb (Halligan et al. 2013), and this pattern can be partially explained by population genetics models combining selective sweeps and background selection (Booker and Keightley 2018). Linked selection has been proposed as a partial explanation for the reduced variation in promoters that flank highly constrained exonic sequences (Hernandez et al. 2011).

In conclusion, my results provide a framework for inferring the fitness consequences of genetic variation across the mouse genome. Mouse models will continue to play a large role in elucidating the function of the non-coding genome. My map of sequence constraint can be used in combination with other data sources such as functional assays and scores of interspecies conservation to help delineate the regulatory landscape across the mouse genome and prioritise variants for functional follow-up studies using mouse models.

5.5 Methods

5.5.1 Raw data and computational dependencies

Genetic variation (vcf) and coverage (bam) files were downloaded from ftp (Harr et al. 2016). Ensembl Regulatory Build and GENCODE annotation were downloaded from Ensembl ftp (v101). The soft-masked mouse reference genome (GRCm38) was downloaded from Ensembl ftp (Zerbino et al. 2018). ENCODE blacklisted regions

were downloaded from ENCODE (Amemiya, Kundaje, and Boyle 2019). Genome-wide methylation of CG dinucleotides in mouse embryonic stem cells was downloaded from Ensembl (v101) (Zerbino et al. 2018; Herrero et al. 2016; Stadler et al. 2011). Genomic Evolutionary Rate Profiling (GERP) scores and GERP conserved regions calculated from 43 eutherian mammals were downloaded from Ensembl (v101) (Zerbino et al. 2018; Davydov et al. 2010). Human-mouse pairwise alignment (GRCh38 and GRCm38) was downloaded from Ensembl (v101) (Herrero et al. 2016). Human variant sites with association from Genome-Wide Association Studies were downloaded from the GWAS catalog (Buniello et al. 2019) (v1.0 downloaded on 30/11/2020). Human pathogenic variant sites predicted to follow Mendelian inheritance patterns were downloaded from ClinVar (Landrum et al. 2018) (variant_summary.txt downloaded on 02/03/2021). All data wrangling, quality control, and analyses were performed using custom scripts written in R (v3.6.2) (R Core Team 2017) or bash.

5.5.2 Defining high-quality variants and genomic loci

I considered genetic variation between 67 mice (*Mus musculus sp*. and *Mus spretus*) from 8 wild populations, described in detail by (Harr et al. 2016). I considered all autosomal single nucleotide variant (SNV) sites with 'PASS' filter status (n = 132,000,731). Read-depth and repetitive elements affect the false positive and false negative rates in genome sequencing (Metzker 2010). Thus, as a filtering step, I evaluated average read depth site-by-site and applied a hard cutoff. Read-depth at every genomic position was calculated for each individual in the sample from the bam files using SAMtools (H. Li et al. 2009). Genomic positions were classified as

having low coverage if they have a read-depth less than 10X in more than 10% of sequenced mice. In total 37.9% of the mouse autosome was classified as “low coverage”. Repetitive elements are soft-masked by Ensembl using the RepeatMasker software (Zerbino et al. 2018); Smit et al 2015). In total 45.9% of the mouse autosome is softmasked. Genomic regions with a high likelihood of erroneous signals in next-generation sequencing experiments (blacklisted regions) are defined by ENCODE (Amemiya, Kundaje, and Boyle 2019). In total 5.5% of the mouse genome is blacklisted.

5.5.3 Determining sequence-specific probabilities of mutation

Local sequence context captures the probabilities of substitution for nucleotides, which can in turn introduce bias into estimates of the expected number of variants in a region given neutrality if not controlled for. The 7-mer context (i.e. the three flanking bases on either side of the base of interest) was shown if chapter 2 to explain the most variance in mutability at the local sequence scale in the mouse lineage. The 16,384 (4^7) distinct 7-mers are not evenly distributed across the mouse genome and show a clear association with different genomic annotations (Figure 5.1). Probabilities of substitution for each base pair across the autosome were determined by their heptamer sequence context (as described in chapter 2). Methylated CG dinucleotides have a high substitution rate, approximately 8-fold greater than unmethylated CG dinucleotides (see chapter 2). As the majority of CG dinucleotides across the genome are methylated, this can bias substitution rate estimates in unmethylated regions. Therefore, the heptamer probabilities of substitution at CG dinucleotides classified as unmethylated were replaced with the probability of

substitution calculated exclusively from unmethylated CG dinucleotides (see chapter 2). Methylation status for CG dinucleotides across the autosome was classified from whole-genome bisulfite sequencing in mouse embryonic stem cells, conducted by Stadler et al (2011). Whole-genome methylation state data is not available for the mouse germline. However, similar patterns of methylation have been observed between mouse embryonic stem cells and sperm cells (Popp et al. 2010), and I, therefore, used mouse embryonic stem cells as a proxy for the germline. CG dinucleotides were classified as unmethylated if they have coverage greater than or equal to 5 and a percentage methylated less than or equal to 20.

5.5.4 Ranking constrained regions across the genome

I quantified constrained regions across the mouse reference autosome using a sliding window method. A window size of 1000 bp, sliding every 100 bp, was selected as a compromise between resolution, power, and computational complexity. For each window, I summed: the weighted count of SNV sites; the weighted probabilities of substitution for each base pair position (see above); the number of blacklisted base pair positions; the number of ambiguous (N) bases; the number of softmasked base pair positions, the number low coverage base pair positions; and the median weighted Genomic Evolutionary Rate Profiling (GERP) score. SNV counts, probabilities of substitution, and GERP scores were weighted based on location in the window, in order to prioritise the middle 100 bp to which the score for the window is attributed. The middle 100 bp for each window was given a weighting of one, and the flanking 450 bp were weighted from one to zero based on their linear distance from the central 100 bp (i.e. base 101 is weighted as 449/450, and base 1000

is weighted as 1/450) (Supplementary Figure 5.1). I filtered all windows with one or more blacklisted or ambiguous (N) base pair positions, more than 90% of base pairs softmasked, or more than 40% of base pairs were categorised as low coverage. This left 52.2% of the mouse autosome for further analysis. I quantified selective constraint for each unfiltered window as the mouse-specific constraint score (MSCS) as the difference between the observed weighted number of SNV sites and the expected number given the window's weighted sequence-specific probability of substitution. To do this, I regressed the weighted number of SNV sites in each window against the window's weighted sequence-specific probabilities of substitution. The residual from this regression was taken as the measure of selective constraint and attributed to the middle 100 bp region of each window. Regions that have a lower residual (i.e. fewer SNV sites than expected) are deemed more constrained. Regions were ranked and binned from 1 (least constrained) to 100 (most constrained) based on their MSCS and average GERP scores.

5.5.5 Genome-wide annotations

I annotated exonic and multi-cell regulatory features autosome-wide for humans and mice by compiling exonic annotations from GENCODE and regulatory annotations from the Ensembl Regulatory Build (v101). I developed the genome-wide multi-cell annotation model defined by di lulio et al (2018), and annotated each nucleotide across the mouse autosome with one or more of the following categories:

1. **“Exon - CDS”** for translated nucleotides in protein-coding exons (including immunoglobulin and T cell receptor genes) as annotated by Ensembl.

2. **“Exon - 5`UTR”** for 5’ UTR nucleotides in protein-coding exons as annotated by Ensembl.
3. **“Exon - 3`UTR”** for 3’ UTR nucleotides in protein-coding exons as annotated by Ensembl.
4. **“Exon - other”** for nucleotides in non-protein-coding exons (notably ncRNAs and lncRNAs) as annotated by Ensembl.
5. **“Intron - proximal”** for intronic nucleotides inferred from exon coordinates as annotated in Ensembl, that are located within 10bp of a splice-site position.
6. **“Promoter”** for nucleotides annotated by the Ensembl Regulatory Build.
7. **“Enhancer - proximal”** for nucleotides annotated as promoter flanking by the Ensembl Regulatory Build.
8. **“Enhancer - distal”** for nucleotides annotated as enhancer by the Ensembl Regulatory Build.
9. **“CTCF binding”** for nucleotides annotated by the Ensembl Regulatory Build.
10. **“Miscellaneous”** for nucleotides categorised as unannotated transcription factor binding site or unannotated open chromatin by the Ensembl Regulatory Build.
11. **“Intron - distal”** for intronic nucleotides inferred from exon coordinates as annotated in Ensembl, that are located more than 10bp from a splice-site position, and do not have any other annotation.
12. **“Unannotated”** for all nucleotides without any other annotation.

A summary of the genomic coverage for each annotation is provided in Supplementary Table 5.1. Annotation overlap is summarized in Supplementary Table 5.2.

5.5.6 Feature composition of constrained regions

I characterised constrained genomic regions by their composition of exonic and regulatory annotations. I calculated the fraction of genomic territory covered by each annotation category for each MSCS and GERP percentile rank across the genome. I calculated the fold-change of genomic territory for each annotation in the most constrained (1st) percentile relative to the least constrained (100th) percentile by dividing the fraction of genomic territory in the 1st percentile by the fraction of genomic territory in the 100th percentile. I calculated the fold-change of genomic territory for each annotation in the most constrained (1st) percentile relative to the genome-wide average by dividing the fraction of genomic territory in the 1st percentile by the fraction of genomic territory across all percentiles.

5.5.7 Human variant enrichment

Human pathogenic and benign variants were downloaded from ClinVar (Landrum et al. 2018). I classified SNV sites with clinical significance labelled as either “Pathogenic” or “Likely pathogenic”, as pathogenic. All pathogenic SNVs considered have a review status labelled as either "criteria provided, multiple submitters, no conflicts", "criteria provided, single submitter", or "reviewed by expert panel".

Human SNV sites associated with complex disease were obtained from the GWAS Catalog (Buniello et al. 2019), and have a phenotype that is ontologically classified as either disease, disorder, or cancer, and a p-value < e10-8. ClinVar and GWAS catalog variants were categorised as human coding-sequence (CDS) if they occur within protein-coding exons annotated by GENCODE and human nonCDS if they do not. Species genomes that have diverged over evolutionary time can be aligned to identify orthologous regions. I used human-mouse pairwise alignment conducted by Ensembl (Herrero et al. 2016) using LastZ (R. S. Harris 2007) to define orthologous bases between humans and mice. I define bases as orthologous if their genomic coordinates align, independent of whether a point mutation has occurred. In total, I considered 26,181 ClinVar and 705 GWAS catalog human CDS SNV sites, and 6,034 ClinVar and 7,071 GWAS catalog human CDS SNV sites that mapped to regions of the mouse autosome with a corresponding MSCS.

Chapter 6

Summary and Conclusions

6.1 Summary of findings

The overarching aim of this thesis was to characterise the mouse genome by patterns of selective constraint in order to 1) identify and rank genes and genomic regions that are functionally important to fitness; 2) characterise relationships between genetic variation in constrained regions and organismal phenotype; and 3) contextualise genomic and functional divergence between mice and humans.

In chapter 2, I quantified the effect of expanding local sequence context (i.e. the composition of a base's flanking nucleotides) and methylation state on the ability to predict substitution occurrence across the mouse genome. I did this to control for differences in sequence mutability when estimating selective constraint in my downstream analyses. I found that the 7-mer context (i.e the composition of the three nucleotides flanking either side of the base of interest) explains 84.6% and 77.8% of the total variance in cytosine and adenine substitution occurrence in the mouse lineage, respectively. This is more than the 3-mer (one nucleotide on either side) and 5-mer (two nucleotides on either side) contexts. Furthermore, 82.5% to 99.7% of the variance in CG>TG substitution rates for CG dinucleotides can be explained by their methylation state, with methylated CG dinucleotides being 8.3-fold more likely to mutate than their unmethylated counterparts. Interestingly, I

highlight a relationship between the probability of negative selection and mutability at the local sequence scale, showing that substitutions in contexts that are more likely to be deleterious to fitness occur at a lower rate across the genome. I postulate that this correlation is in part a consequence of adaptive evolution on mechanisms of DNA repair to minimise the occurrence of mutations in sequence contexts where they are more likely to be deleterious.

In chapter 3, I used the substitution rate estimates determined by the 7-mer context to characterise and rank genes by their relative depletion of nonsynonymous variation in the mouse lineage. I then used these estimates to characterise relationships between genic constraint and gene function. Genes essential to embryonic development (i.e. those that result in preweaning lethality when knocked out) are enriched among the most constrained mouse genes. Furthermore, there is a positive correlation between gene constraint and pleiotropy (Spearman's rho = 0.67, $p < 0.01$), defined as the number of knockout phenotype associations for a gene, which may be due to additive selection effects across the multiple phenotypes of pleiotropic genes (Wang, Liao, and Zhang 2010; Fish, Chen, and Capra 2017). Finally, my mouse gene constraint metric is positively correlated with human constraint metrics and the number of pathogenic variants in the human orthologue. This correlation suggests that selective constraint can be in part explained by variants in more constrained genes having an increased likelihood of being pathogenic and that this pathogenicity has been maintained between the human and mouse lineages.

In chapter 4, I provided an overview of the sequence alignment between humans and mice for twelve categories of genomic annotation, including protein-coding genes

annotated by GENCODE and regulatory features annotated by the Ensembl Regulatory Build. I then assessed the alignment of human Mendelian and complex disease-associated variant sites with the mouse genome across these annotation categories. Substantially more human non-coding variant sites predicted to cause Mendelian disease align with the mouse genome (98.4%) compared to human non-coding variant sites associated with complex disease (36.4%). Interestingly, a greater proportion of human non-coding variant sites associated with complex disease that are located in unannotated regions or intronic regions that are not in proximity to splice sites align with the mouse genome than the neutrally evolving expectation (34.4% compared with 29.6%). This indicates these loci are more likely to have been conserved over evolutionary time. It remains unclear, however, whether this conservation is a result of conserved function.

In my final analysis chapter, I characterised patterns of genetic diversity across the entire mouse genome, including non-coding regions, in order to provide a framework for interpreting the fitness consequences of genetic variants and identifying functionally important regions of the genome. I show that the most constrained regions across the mouse genome are enriched for exonic and regulatory features, as expected, with promoters showing the greatest enrichment relative to other non-coding elements. Furthermore, I highlight a positive correlation between regional constraint and the number of conserved non-coding human pathogenic variant sites across the mouse genome. This suggests that non-coding human pathogenic variant sites are more likely to align with regions of the genome that are functionally important in mice. Further work is required to determine whether genotype-phenotype pathways have been conserved between the species in constrained regions.

6.2 Limitations and future research

6.2.1 Correlation between mutation rate and fitness consequences

The degree to which mutation rates correlate with potential fitness consequences has implications for estimates of selective constraint. In chapter 2, I highlight a relationship between mutability and the probability of negative selection at the local sequence scale – with substitutions in nucleotide contexts that could cause nonsense mutations occurring at a rate that is 31% lower than those in contexts where a nonsense mutation is not possible (regardless of where in the genome it occurs). This result complements several studies, including work in *E. coli* (Martincorena et al. 2012) and *Arabidopsis thaliana* (Monroe et al. 2022), indicating mutations occur at a lower rate in genomic regions subject to negative selection. For example, Monroe et al. (2022) report a roughly two-fold difference in mutation rates in gene bodies relative to their flanking regions, with even lower mutation rates in essential genes. The lower mutation rates correlate with physical features and epigenetic markers at the kb scale and these are hypothesised to provide a basis for adaptive evolution of the mutation rate through their interaction with mechanisms of DNA repair (Monroe et al. 2022). It is currently unclear to what extent these correlations between the probability of negative selection and mutation rate could introduce bias in models

that control for mutability when estimating selection (di Iulio et al. 2018; Samocha et al. 2014; Karczewski et al. 2020; Lek et al. 2016).

My findings provide, to the best of my knowledge, the first evidence that the local mutation rate spectrum has been shaped by adaptive evolution at the local sequence scale. Considered alongside findings in *E. coli* (Martincorena et al. 2012) and *Arabidopsis thaliana* at the kilobase scale (Monroe et al. 2022), this indicates that mutation rate modifiers have evolved to suppress deleterious variation in diverse taxa. How these mechanisms manifest is unclear, but it is likely that they involve DNA repair processes (Gonzalez-Perez, Sabarinathan, and Lopez-Bigas 2019; Supek and Lehner 2015). Similar mechanisms may also target deleterious somatic mutations, such as those associated with cancer genomes (Koh et al. 2021; Alexandrov et al. 2020), and this provides promising directions for future research.

6.2.3 Methodological limitations and comparisons with human constraint scores

Estimates of selective constraint are useful for prioritising regions of the genome that are functional (i.e. they affect phenotype) and may affect disease. This has been demonstrated across the human genome, and in this thesis across the mouse genome, as genes and regions depleted of variation are associated with developmental lethality and disease phenotypes (Bartha et al. 2018; Karczewski et al. 2020; di Iulio et al. 2018). However, not all functional loci are intolerant of variation. Functional loci may be tolerant of variation (unconstrained) if only a small subset of potential mutations affects phenotype, or if new variants affect phenotype but are not

subject to negative selection. For example, certain protein-coding genes, such as those with immunological functions, are highly tolerant of variation (Karczewski et al. 2020; Bartha et al. 2018). Furthermore, loci under positive selection may be variable between populations. For example, house mice on Gough Island (a remote island in the South Atlantic) are substantially larger than mainland mice and display distinct behaviours, and these phenotypic differences have been linked to genetic variation showing evidence of positive selection (Payseur and Jing 2021).

Not all disease phenotypes affect fitness and not all reductions in fitness are caused by disease. Genotypes that cause diseases that typically present after reproduction are less likely to be constrained by negative selection. BRCA1 is an example of a relatively unconstrained gene, as highlighted in the introductory chapter, that contains variant sites that are causally linked to adult-onset hereditary cancer (Hall et al. 1990; Findlay et al. 2018). A recent analysis in a human cohort suggests that 20% of the purifying selection against predicted complete loss-of-function variants in some genes can be explained by reduced fecundity of carriers (Gardner et al. 2022). Gardner et al (2020) show that predicted complete loss-of-function variants present in the UK Biobank (a biomedical database comprising genotypic and phenotypic data for ~500,000 individuals based in the United Kingdom; (Sudlow et al. 2015) are associated with childlessness and provide evidence that this association is a consequence of behavioural and cognitive traits that results in a lower probability of reproductive success. It is possible that some of the purifying selection against nonsynonymous variation across mouse genes may also be a result of sexual selection in wild populations.

The power to estimate selective constraint in the mouse lineage is limited by the availability of whole-genome and whole-exome sequencing data for wild mice. In this thesis, I have used single nucleotide variant sites between 67 wild mice comprising two species (*Mus musculus* and *Mus spretus*) from eight distinct populations (Harr et al. 2016). Wild house mice are approximately 10 times more genetically diverse than humans as a result of their larger effective population size (Halligan et al. 2013). However, whole-exome and whole-genome sequencing of human populations has been far more extensive. For example, The Genome Aggregation Database (gnomAD) contains 14.9 million exonic and 229.9 million genomic variants from sequencing 125,748 exomes and 15,708 genomes. Of the exonic variants, 443,769 are predicted to cause complete loss of function, enabling genetic estimation of selective constraint against null mutations in genes. Larger samples of sequenced mice are required in order to accumulate sufficient power to assess constraint against loss-of-function variants, which tend to be rare.

A notable difference between the samples of genetic variation used to estimate mouse-specific and human-specific constraint is the phylogenetic distance between the sampled individuals. The last common ancestor of all humans (i.e. the most recent ancestor that all individuals in a group share) is predicted to have lived ~200,000 years ago (Zhou and Teo 2016). By comparison, *Mus musculus* and *Mus spretus* are more distantly related, with their last common ancestor predicted to have lived ~2 million years ago. The number of years to the last common ancestor underrepresents the evolutionary distance between the two lineages though, as mice have higher mutation rates and shorter generation times than humans (Cagan et al. 2022; Lindsay et al. 2019). Including *Mus spretus* in the sample of genetic variation increased the number of variants (and therefore the power to identify constrained

regions); however, the increased phylogenetic distance between the lineages will have increased the potential for functional elements and selection coefficients to have changed between the lineages. This is particularly pertinent for noncoding regulatory elements that have a high evolutionary turnover rate. Furthermore, the evolutionary divergence between *Mus spretus* and the mouse reference genome (Mus musculus) may have led to regions of the genome being inaccessible and therefore excluded from the analysis. For example, early sequencing of the *Mus spretus*-derived inbred strain SPRET/EiJ showed that ~23% of its genome could not be mapped to the *Mus musculus* reference genome (Keane et al. 2011).

During quality control filtering I excluded genomic positions that have low coverage (i.e. less than 10X in more than 10% of individual mice in the sample), as low coverage regions are more likely to contain false-negative SNP calls (Metzker 2010). On the other hand, however, excessive coverage may indicate regional mismapping that can result in an enrichment of false-positive SNP calls (Lilue et al. 2018; Nurk et al. 2022). For example, regions with high densities of heterozygous SNPs have been identified across inbred mouse strains when called against the mouse (C57BL/6J) reference assembly (Lilue et al. 2018). Inbred mouse strains are homozygous at almost all loci (Beck et al. 2000), and these heterozygous SNP calls are therefore expected to be an artefact from the mismapping of regions that are not present in the reference assembly such as novel sequences or copy number variations (Lilue et al. 2018). If i was to repeat my analysis, I would filter high-read-depth regions and assess how this affects the results.

Finally, this thesis focuses on single nucleotide variants (SNVs), excluding other variant types such as insertions and deletions (indels), and larger structural variants.

Indels and larger structural variants are less prevalent than SNVs (Keane et al. 2011) as they are more likely to be highly disruptive and purged by negative selection (Karczewski et al. 2020). This is particularly true for indels within protein-coding sequences as these are likely to cause frameshift mutations that result in the complete loss of function of the disrupted gene. Larger structural variants are evolutionarily important. For example, a recent study linked transposable elements, the most common type of structural variant, to gene expression variation between wild populations of *Drosophila melanogaster*, indicating a role in adaptive evolution (Rech et al. 2022). Future research could characterise selective constraint against a broader range of mutation types.

Larger sample sizes of ultra-deep sequencing in wild mice populations would increase the power to calculate constraint metrics and make them more directly comparable to constraint scores calculated from samples of human genetic variation. Future work may include assessing constraint against heterozygous loss-of-function variation for mouse protein-coding genes and identifying constrained coding regions within protein-coding sequences (i.e regions within genes that are intolerant of mutation). It would be interesting to assess whether there are amino-acid sequences that are highly constrained in one species but not the other. Finally, as the annotation of noncoding regulatory elements improves, it would be interesting to characterise the most constrained regulatory elements in the mouse lineage. This would be particularly interesting in mice, as there is the potential to experimentally manipulate these constrained regulatory features in an attempt to infer their function.

6.2.4 Implications of linked selection

One major difficulty in estimating selective constraint from patterns of genetic diversity is teasing apart the contribution of linked selection (i.e. reductions in genetic diversity at neutral sites linked to loci under selection). Reductions in genetic diversity at linked sites can result from positive selection via selective sweeps, and negative selection through background selection (described in detail in the introduction). The degree to which neutral genetic diversity is decreased at linked sites is affected by the selection coefficient, effective population size, recombination rate, and mutation rate (Booker and Keightley 2018). Differences in these factors lead to variation in the magnitude of linked selection across genomes and between species. Previous studies have assessed the contribution of linked selection to the patterns of genetic variation observed across the human and mouse genomes. In humans, there is a marginal difference between the reduction in diversity around synonymous and nonsynonymous mutations, suggesting hard selective sweeps were infrequent in the past 250,000 years of human evolution (Hernandez et al. 2011). Furthermore, an analysis of selective constraint across the human genome found far fewer variants in promoter sequences relative to matched intronic sequences that flank the other side of the exon, suggesting the reduced diversity in promoters could not be solely explained by selection at linked sites (di Iulio et al. 2018). Across the mouse genome, by comparison, genetic diversity is reduced by an average of ~100Kb around protein-coding sequences and ~10Kb around conserved noncoding elements (CNEs) (Halligan et al. 2013). This depletion of variation can be partially explained by population genetics models combining selective sweeps and background selection, with the wider troughs in diversity around protein-coding genes resulting from

stronger selection acting on mutations in these regions compared with mutations in CNEs (Halligan et al. 2013; Booker and Keightley 2018). It is predicted that a greater proportion of nonsynonymous mutations and mutations in CNEs are driven to fixation by positive selection in the mouse lineage compared with humans due to the larger effective population sizes of wild mice, and this could account for a greater effect of linked selection in shaping patterns of genetic diversity (Halligan et al. 2013).

6.3 Concluding remarks

In conclusion, estimates of selective constraint provide insight into the fitness consequences of genetic variation and can be used to prioritise genes or genomic regions that are likely to be functionally important. The research in this thesis provides a further step toward characterising the mouse mutation-rate spectrum at the local-sequence scale and selective constraint across the mouse genome, including protein-coding and non-coding sequences. As more data becomes available from larger sample sizes of ultra-deep sequencing in wild mice populations, estimates of selective constraint across the mouse genome will become more powerful and more equivalent for comparison with human constraint scores. Mouse models provide critical insights into mammalian genotype function including for human disease; however, inferences about human biology from mouse models need to be contextualised with an understanding of the evolutionary divergence between the two species. A deeper understanding of how selective pressures have shaped genotype to phenotype pathways in the mouse lineage will continue to shed light on mammalian evolution and provide context for mouse models of human disease.

Supplementary Information

Chapter 1

Supplementary Table 1.1 – Human constraint scores. Abbreviations: single nucleotide variant (SNV); single nucleotide polymorphism (SNP); allele frequency (AF); Functional Analysis through Hidden Markov Models (FATHMM); protein-truncating variant (PTV); loss-of-function (LoF).

Score	**Genomic Evolutionary Rate Profiling (GERP)**	**dN/dS**	**Residual Variation Intolerance Score (RVIS)**	**EvoTol**
Brief description	Nucleotide position-specific measure of evolutionary constraint between species.	Gene-specific measure of selection on fixed nonsynonymous variation between species.	Gene-specific measure of intolerance to common 'functional' variants.	Gene-specific measure of intolerance to missense variants.
Method summary	Based on multiple sequence alignment. Positions are considered more constrained if they have fewer substitutions than expected given neutrality.	Calculated as the ratio of nonsynonymous variants per nonsynonymous site (dN) to synonymous variants per synonymous site (dS).	Based on the residual from regressing the number of 'functional' SNVs against the total number of SNVs (functional and non-functional) within genes.	Based on the residual from regressing the number of 'functional' SNVs against the total number of SNVs (functional and non-functional) within genes.
Interpretation	Continuous score. A higher value indicates more constraint.	Continuous score >= 0. A lower value indicates more constraint.	Continuous score. A lower value indicates more constraint.	Continuous score. A lower value indicates more constraint.
Type of variation	SNPs	SNPs: synonymous and nonsynonymous.	SNVs (missense, nonsense, splice acceptor, and splice donor are considered 'functional', synonymous are considered non-functional).	SNPs: missense only (classified as functional or non-functional with FATHMM).

Source of variation	Multiple species alignment	Reference genomes (eg human and chimpanzee).	6,503 whole exomes from the ESP.	686,501 SNPs from dbSNP.
AF filter	NA	NA	AF > 0.1	NA
Regional mutability	NA	Controlled for by the rate of substitutions at synonymous sites (dS).	Controlled for by the number of synonymous variants.	NA
Functional correlations	Defines ~1.3 million constrained regions spanning ~7% of the genome. Constrained regions are most strongly enriched for coding exons, followed by UTRs.	NA	The most constrained genes are enriched for genes associated with Mendelian disease (OMIM), particularly haploinsufficenc-y; enriched for knockout lethal mouse orthologues.	The most constrained genes are enriched for genes associated with Mendelian disease (OMIM); enriched for genes associated with epilepsy and congenital heart disease; enriched for nuclear receptors and metabolic processes. The least constrained genes are enriched for olfactory receptors and G protein-coupled receptors.
Other notes	Constrained elements are defined using a permutation test to identify significant stretches of highly constrained (high GERP score) positions.	NA		FATHMM predicts the functional consequences of variants based on the alignment of homologous sequences and conserved protein domains.

Reference	(Davydov et al. 2010)	(Herrero et al. 2016; Kryazhimskiy and Plotkin 2008)	(Petrovski et al. 2013)	(Rackham et al. 2015)

Score	**Missense z-score**	**Probability of haploinsufficiency (Phi)**	**Probability of loss-of-function intolerance (pLI)**	**LoFtool**
Brief description	Gene-specific measure of tolerance of de novo missense mutations.	Gene-specific probability of selection against heterozygous LoF variants.	Gene-specific probability of selection against heterozygous LoF variants.	Gene-specifc rank based on the strength of negative selection acting on LoF variants.
Method summary	Based on the difference between the observed and expected number of missense SNVs given no selection. The expected number is predicted using a mutation model trained using synonymous variants.	Based on a null distribution of expected variants given no selection.	Based on the difference between the observed and expected number of LoF variants given no selection. The expected number is predicted using a mutation model trained using synonymous variants.	Heuristic combining the Samocha et al (2014) mutational model and EvoTol scores.
Interpretation	Continuous Z-score. A higher value indicates more constraint.	Continuous score between 0 and 1. A higher value indicates more constraint.	Continuous score between 0 and 1. A higher value indicates more constraint. Genes are categorised as completely intolerant of heterozygous LoF variants with pLI > 0.9.	Percentile rank. A lower value indicates more constraint.
Type of variation	SNVs: missense and synonymous	PTVs: nonsense and frame-shift (predicted LoF with SnpEff).	PTVs: nonsense, splice, frameshift (predicted LoF with	PTVs: nonsense, splice, frameshift (predicted LoF with

			VEP-LOFTEE).	VEP-LOFTEE).
Source of variation	6,503 whole exomes from the ESP.	10,500 whole exomes from ESP UK10K and CoLaus	60,706 whole exomes from ExAC.	60,706 whole exomes from ExAC.
AF filter	Singletons only.	AF < 0.01	AF < 0.001	NA
Regional mutability	Controlled for by trinucleotide probabilities of mutation and regional divergence between humans and macaques.	Controlled for using transcript-specif-ic de novo mutation rate estimates calculated by Samocha et al (2014)(Samocha et al. 2014).	Controlled for by trinucleotide probabilities of mutation.	
Functional correlations	The most constrained genes are enriched for LoF variants in individuals with autism spectrum disorders or intellectual disability.	The most constrained genes have fewer paralogs; are more central in protein interaction networks; are enriched for knockout lethal mouse orthologues and cellular lethal genes.	The most constrained genes are highly expressed; have more protein-protein interactions; are more likely to be adjacent to GWAS signals; are enriched for knockout lethal mouse orthologues.	The most constrained genes are enriched for genes associated with Mendelian disease (OMIM); enriched for genes expressed in the brain and associated with neurodevelopm-e ntal disorders.

Other notes	The expected number of missense SNVs were predicted using a linear model trained to predict the number of synonymous SNVs as a function of the transcript specific probability of incurring a synonymous substitution (adjusted for read depth), and the regional divergence between humans and macaques.	Generated null distributions of the expected number of PTVs for each gene given no selection.	Probabilities calculated from a three-state Poisson mixture model.	Combines the increased sample size in ExAC with the Samocha et al (2014) mutational model and EvoTol scores.
Reference	(Samocha et al. 2014)	(Bartha et al. 2015)	(Lek et al. 2016)	(Fadista et al. 2017)

Score	**Selective effects of heterozygous protein-truncating variants (Shet)**	**Constrained coding regions (CCRs)**	**Gene variation intolerance ranking (GeVIR)**	**Loss-of-funct-ion observed/exp-ected upper bound fraction (LOEUF)**
Brief description	Gene-specific estimates of the selection and dominance coefficient for heterozygous loss-of-function alleles.	Defines and ranks protein-coding regions within genes that are intolerant of protein-changing variation. Genic constraint is defined by the length of their longest CCR.	Ranks genes based on their intolerance to protein-changing variation.	Gene-specific continuous measure of selection against pLoF variants.

Method summary	Bayesian estimation.	CCRs are defined by the distance between protein-changing variants.	A heuristic that considers the number, length, and conservation of variation intolerant regions within a gene.	Based on the ratio between the observed and expected number of LoF variants given no selection. The expected number is predicted using a mutation model was trained using synonymous variants.
Interpretation	Continuous score between 0 and 1. A higher value indicates more constraint.	Percentile rank. A higher value indicates a longer constrained region.	Percentile rank. A lower value indicates more constraint.	Continuous score >= 0. A lower value indicates more constraint.
Type of variation	PTV SNVs: nonsense and splice (predicted LoF with VEP-LOFTEE).	Protein-changing (missense, nonsense, splice, frameshift).	Protein-changing (missense, nonsense, and frameshift).	PTVs: nonsense, splice, frameshift (predicted LoF with VEP-LOFTEE).
Source of variation	60,706 whole exomes from ExAC.	123,136 whole exomes from gnomAD	123,136 whole exomes from gnomAD	125,748 whole exomes from gnomAD
AF filter	AF < 0.001	NA	NA	AF < 0.001
Regional mutability	Controlled for using transcript-specif-ic de novo mutation rate estimates calculated by Samocha et al (2014)(Samocha et al. 2014).	Controlled for by regressing the CCR length against the number of CG dinucleotides in the CCR.	NA	Controlled for by trinucleotide probabilities of mutation. Distinguishes between methylated and unmethylated CpG sites.

Functional correlations	The most constrained genes are enriched for genes associated with Mendelian disease; enriched for knockout lethal mouse orthologues; enriched for genes known to regulate transcription.	The most constrained coding regions are enriched for pathogenic variants (ClinVar); enriched for mutations associated with developmental disorders.	The most constrained genes are enriched for genes associated with Mendelian disease (ClinVar).	The most constrained genes are enriched for haploinsufficient genes; depleted for olfactory genes; enriched for cellular lethal genes; enriched for knockout lethal mouse orthologues; enriched for genes with a higher number of protein-protein interactions; enriched for genes expressed in a greater number of tissues; enriched for LoF variants in individuals with intellectual disability;
Other notes	NA	CCRs are defined by the bp distance between protein changing variants, excluding introns and correcting for read depth and mutability.	GeVIR scores are calculated for each gene as the sum of the weighted lengths of intragenic coding regions that have 'high-coverage' adjusted by their conservation (mean GERP), and normalized by the total number of regions in the gene including 'low-coverage' regions.	The expected number of pLoF variants is predicted using a mutational model trained on the number of synonymous SNVs. Confidence intervals are calculated for the observed number of variants, with the score based on the upper limit of the 90% interval.

Reference	(Cassa et al. 2017)	(Havrilla et al. 2018)	(Abramovs, Brass, and Tassabehji 2019)	(Karczewski et al. 2020)

Score	**Context-Dependent Tolerance Score (CDTS)**
Brief description	Loci-specific continuous measures of intolerance to variation across the whole genome.
Method summary	Sliding window method where constraint for each window is measured as the difference between the observed number of variants and the expected number given the heptoneucleotide context.
Interpretation	Percentile rank. A lower value indicates more constraint.
Type of variation	SNVs
Source of variation	11,257 whole genomes (Telenti et al. 2016)
AF filter	AF > 0.0001
Regional mutability	Controlled for by heptonucleotide probabilities of mutation.

Functional correlations	The most constrained regions across the genome are enriched for annotated coding and regulatory features (GENCODE and Ensembl Regulatory Build); enriched for pathogenic variants (ClinVar).
Other notes	Windows are 550 bp in length, sliding every 10 bp, with the CDTS score attributed to the middle 10 bp of each window.
Reference	(di Iulio et al. 2018)

Chapter 2

Supplementary Table 2.1 – Mean absolute error (MAE) for each substitution model by substitution class.

Substitution class	Model	MAE	Confidence interval (95%)
A > C	1-mer	280.8	278.5 - 283.1
A > C	A-CG	280.8	278.5 - 283.1
A > C	3-mer	274.1	271.7 - 276.5
A > C	5-mer	259	256.6 - 261.4
A > C	7-mer	241.1	238.6 - 243.6
A > G	1-mer	743.2	739 - 747.4
A > G	A-CG	743.2	739 - 747.4

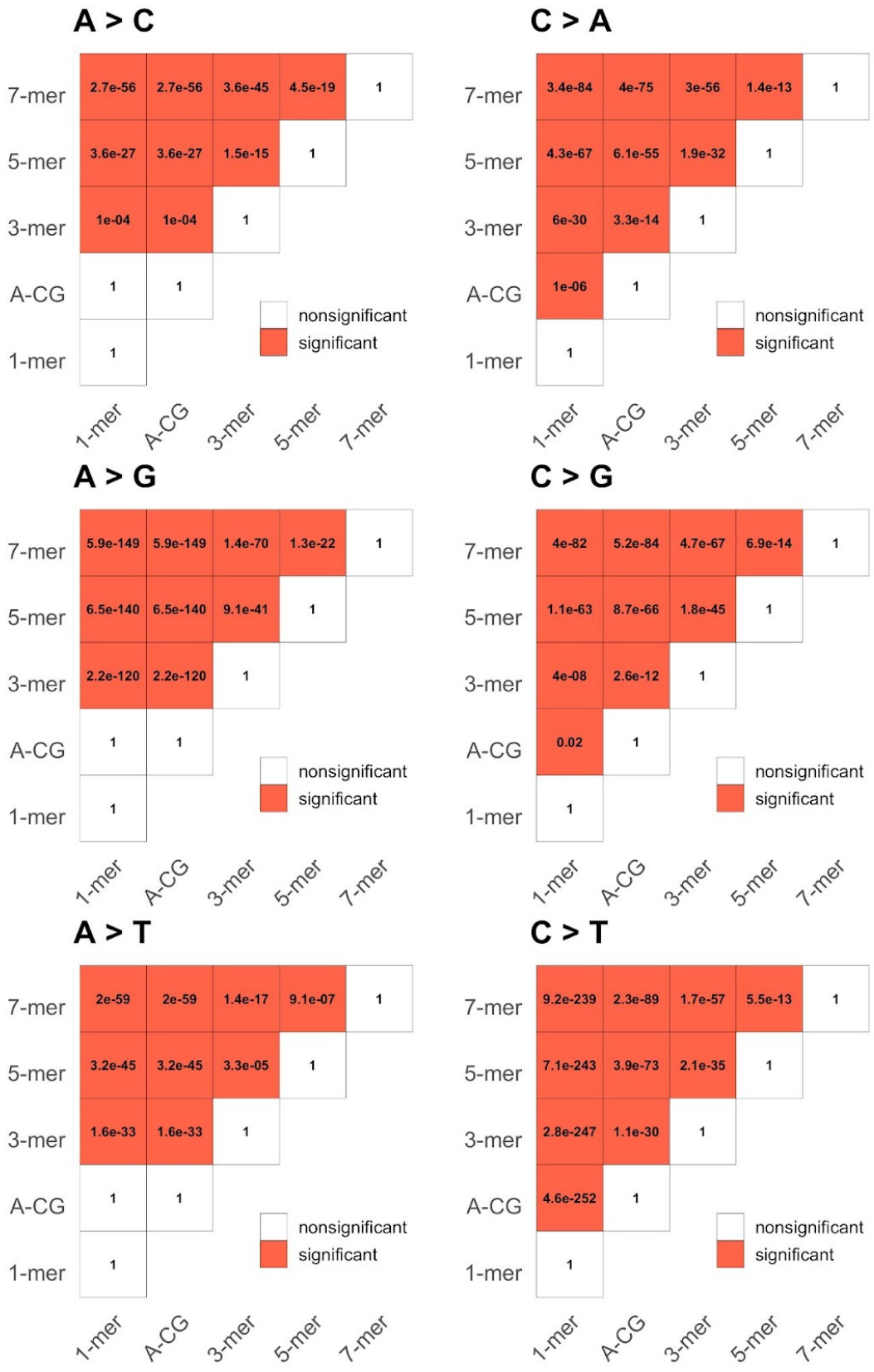
A > G	3-mer	570.3	565.7 - 574.9
A > G	5-mer	511.8	506.9 - 516.7
A > G	7-mer	471.6	466.5 - 476.7
A > T	1-mer	309.6	307.2 - 312
A > T	A-CG	309.6	307.2 - 312
A > T	3-mer	283.2	280.6 - 285.8
A > T	5-mer	275.1	272.4 - 277.8
A > T	7-mer	265	262.2 - 267.8
A > *	1-mer	947.9	942.3 - 953.5
A > *	A-CG	947.9	942.3 - 953.5
A > *	3-mer	736.9	730.9 - 742.9
A > *	5-mer	673	666.7 - 679.3
A > *	7-mer	621.3	614.6 - 628
C > A	1-mer	778.9	774.5 - 783.3
C > A	A-CG	761.8	756.7 - 766.9
C > A	3-mer	731	725.7 - 736.3
C > A	5-mer	671	664.8 - 677.2
C > A	7-mer	635.1	628.8 - 641.4
C > G	1-mer	489.9	486.8 - 493
C > G	A-CG	495.8	492.1 - 499.5
C > G	3-mer	475.6	471.8 - 479.4
C > G	5-mer	423.9	420 - 427.8
C > G	7-mer	400.7	396.7 - 404.7
C > T	1-mer	5560.9	5551.4 - 5570.4
C > T	A-CG	1873	1857.9 - 1888.1

C > T	3-mer	1719.7	1703.9 - 1735.5
C > T	5-mer	1541.7	1525.2 - 1558.2
C > T	7-mer	1448.5	1431.5 - 1465.5
C > *	1-mer	6141.6	6132.1 - 6151.1
C > *	A-CG	2026.5	2010.3 - 2042.7
C > *	3-mer	1842.5	1825.6 - 1859.4
C > *	5-mer	1713.8	1696.5 - 1731.1
C > *	7-mer	1611.6	1593.5 - 1629.7

Supplementary Table 2.2 – Absolute variance explained for each substitution model by substitution class.

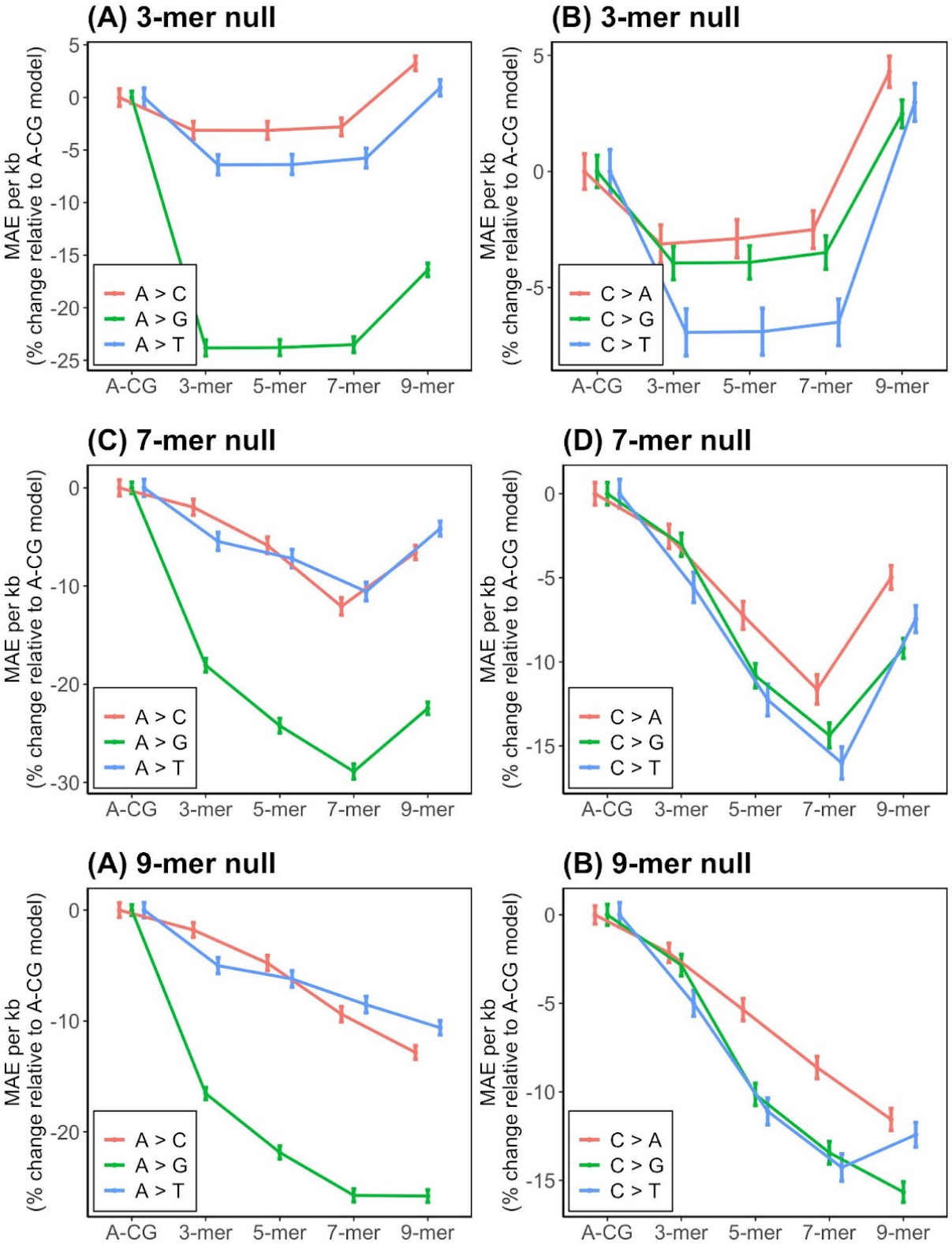
Substitution class	Model	Absolute variance explained (mean)	Absolute variance explained (95% confidence interval)
A > *	1-mer	66	0.1
A > *	3-mer	73.6	0.2
A > *	5-mer	75.9	0.2
A > *	7-mer	77.8	0.2
A > *	9-mer	77.7	0.2
A > *	A-CG	66	0.1
A > C	1-mer	42.9	0.4
A > C	3-mer	44.3	0.4
A > C	5-mer	47.4	0.4
A > C	7-mer	51	0.4
A > C	9-mer	50.3	0.4
A > C	A-CG	42.9	0.4
A > G	1-mer	56.4	0.2
A > G	3-mer	66.6	0.2
A > G	5-mer	70	0.2

A > G	7-mer	72.4	0.2
A > G	9-mer	71.8	0.2
A > G	A-CG	56.4	0.2
A > T	1-mer	47.6	0.4
A > T	3-mer	52.1	0.4
A > T	5-mer	53.5	0.4
A > T	7-mer	55.2	0.4
A > T	9-mer	53.5	0.4
A > T	A-CG	47.6	0.4
C > *	1-mer	41	0.1
C > *	3-mer	82.4	0.1
C > *	5-mer	83.6	0.1
C > *	7-mer	84.6	0.1
C > *	9-mer	83.9	0.1
C > *	A-CG	80.6	0.1
C > A	1-mer	39.1	0.3
C > A	3-mer	43	0.4
C > A	5-mer	47.7	0.4
C > A	7-mer	50.5	0.4
C > A	9-mer	50.1	0.4
C > A	A-CG	40.5	0.4
C > G	1-mer	29.4	0.4
C > G	3-mer	31.5	0.4
C > G	5-mer	39	0.5
C > G	7-mer	42.4	0.5
C > G	9-mer	39.9	0.4
C > G	A-CG	28.6	0.4
C > T	1-mer	34.1	0.1
C > T	3-mer	79.8	0.2
C > T	5-mer	81.9	0.2
C > T	7-mer	83	0.2
C > T	9-mer	82	0.1
C > T	A-CG	77.9	0.1



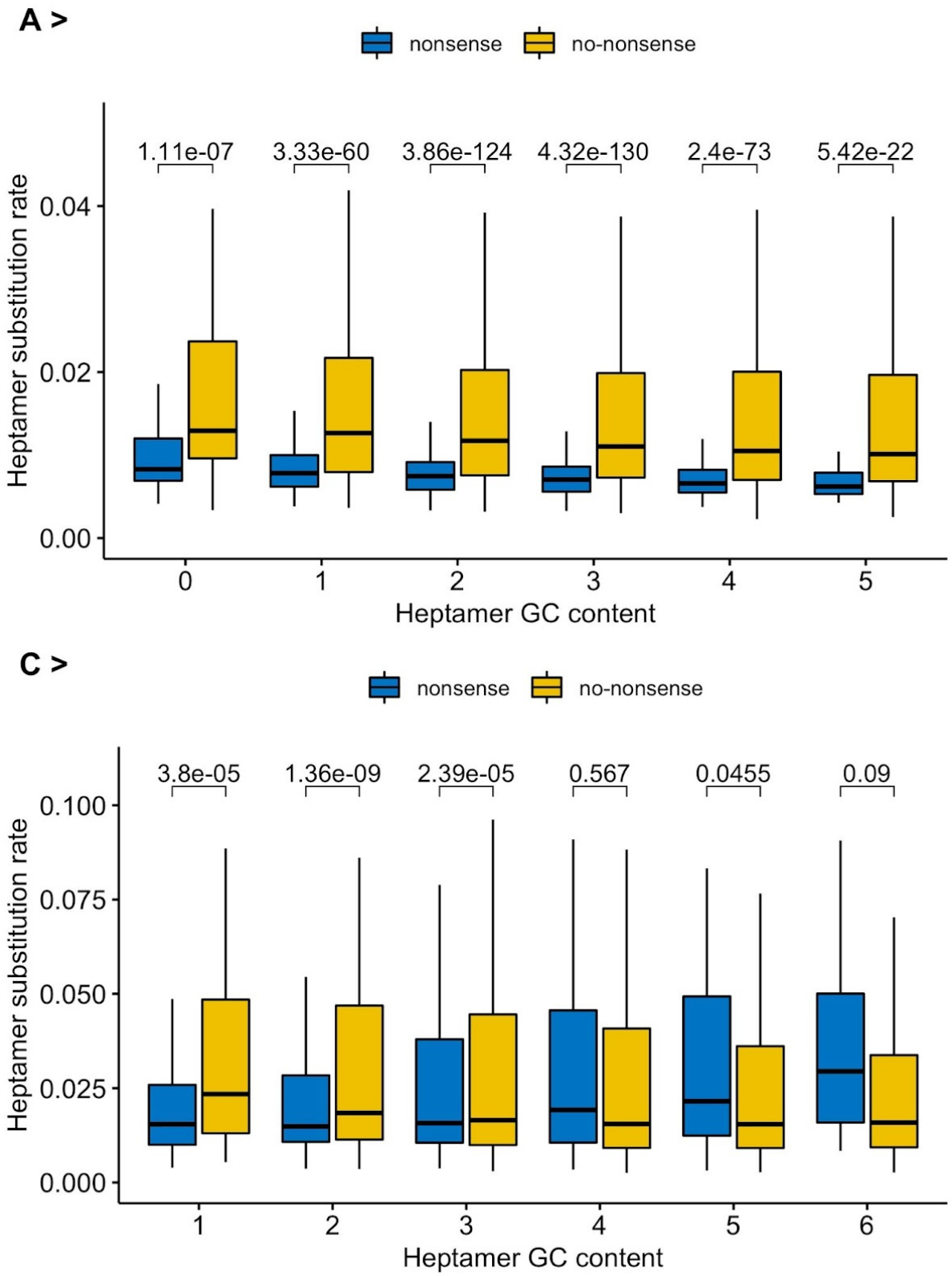
Supplementary Figure 2.1 – P-values for paired t-tests testing for a difference in mean absolute error between models (99 degrees of freedom). P-values are provided

for each substitution class and are in scientific notation rounded to one decimal place. Significance is defined as $p < 0.05$.

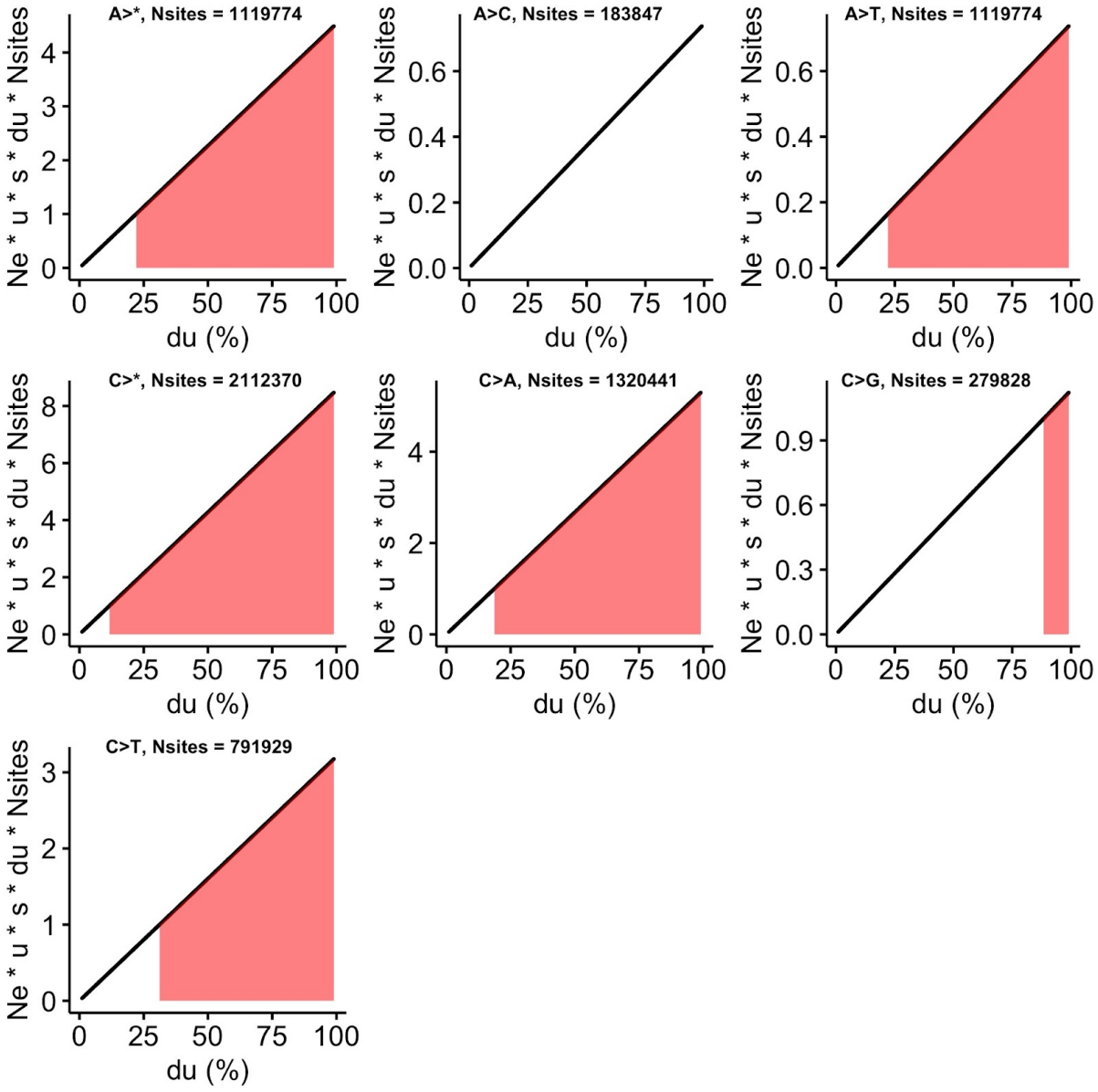


Supplementary Figure 2.2 – Mean absolute error (MAE) per kb relative to the A-CG model given simulated probabilities of substitution determined by the 3-mer

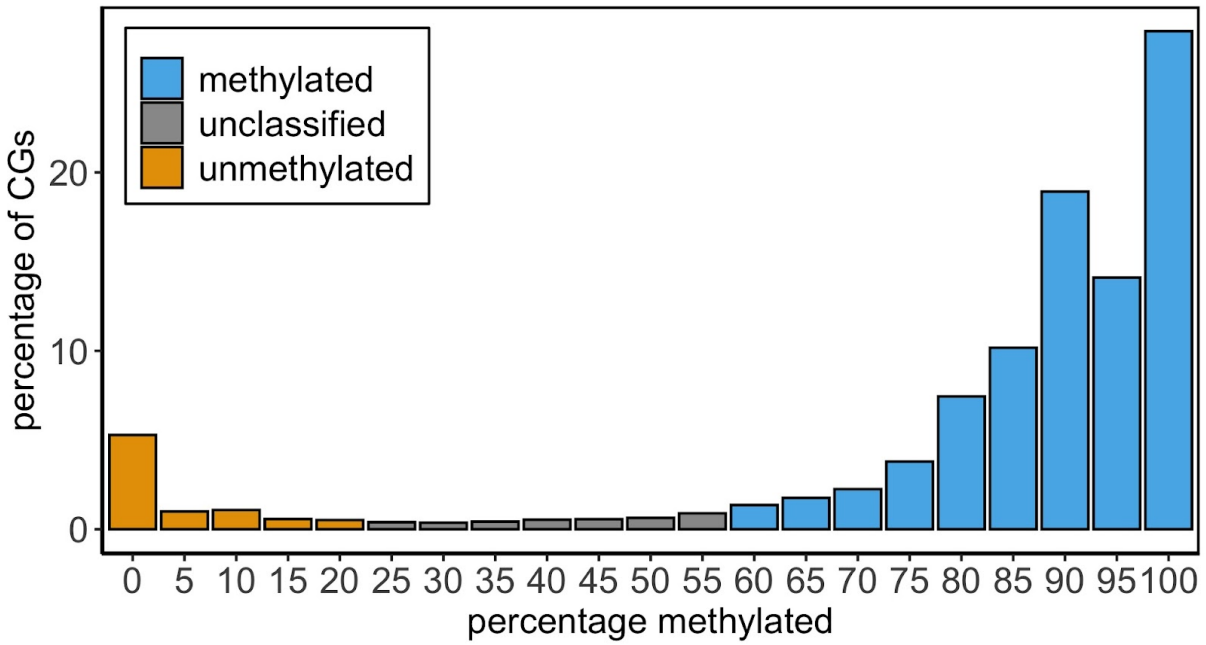
(A and B), 7-mer (B and C), and 9-mer (E and F) contexts for each substitution class. When probabilities of substitution are defined by the 3-mer context, considering additional flanking bases does not improve prediction accuracy. MAE increases, however, for the 9-mer model relative to the 3-mer model for both A and C substitutions ($p < 0.001$, paired t-tests). This is likely caused by increased variance in 9-mer probabilities of substitution between chromosomes, resulting from the greater number of parameters and reduced sample size. When probabilities of substitution are defined by the 9-mer context there is insufficient power to detect significant increases in prediction accuracy for A > G and C > T substitutions relative to the 7-mer model.



Supplementary Figure 2.3 – Distributions of human substitution rates for potential nonsense substitutions and no-nonsense substitutions categorised by the number of C or G nucleotides in the 7-mer (x-axis) for A substitutions (A >) and C substitutions (C >). Differences between groups are tested using Wilcox signed-rank tests.

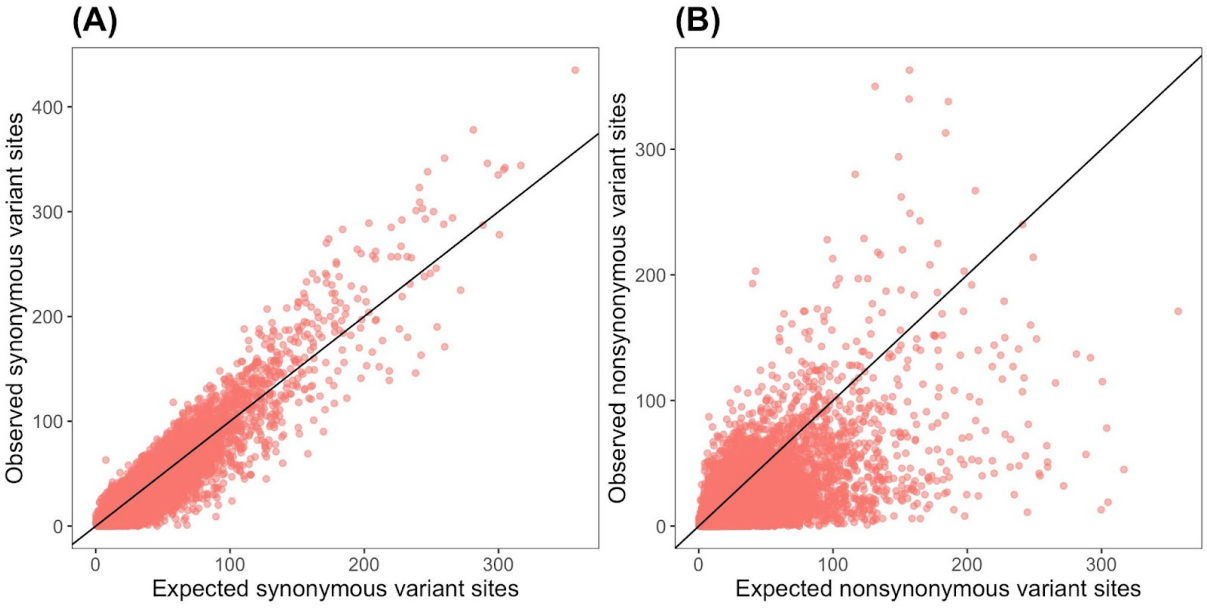


Supplementary Figure 2.4 – Change in the product of the effective population size (Ne), the per-nucleotide mutation rate (u), the magnitude of change in the mutation rate (du), the average selection coefficient of a mutation (s), and the number of sites where a mutation would be under selection (Nsites) (y -axis) across a gradient of du values (x-axis). Population genetics theory states that selection can overcome drift to shape the mutation rate when the product of Ne, u, s, du, and Nsites exceeds one. I consider an effective population size (Ne) of 30,000 (Phifer-Rixey et al. 2012), a per-nucleotide substitution rate (u) of 5.4e−9 (Uchimura et al. 2015), and an average selection coefficient against nonsense substitutions (s) of 0.025 (Gorlov, Kimmel, and Amos 2006). Nsites was calculated for each substitution class by summing the number of sites across transcripts where a substitution would cause a nonsense mutation. The values of du for selection to overcome drift given these parameters are highlighted in red.



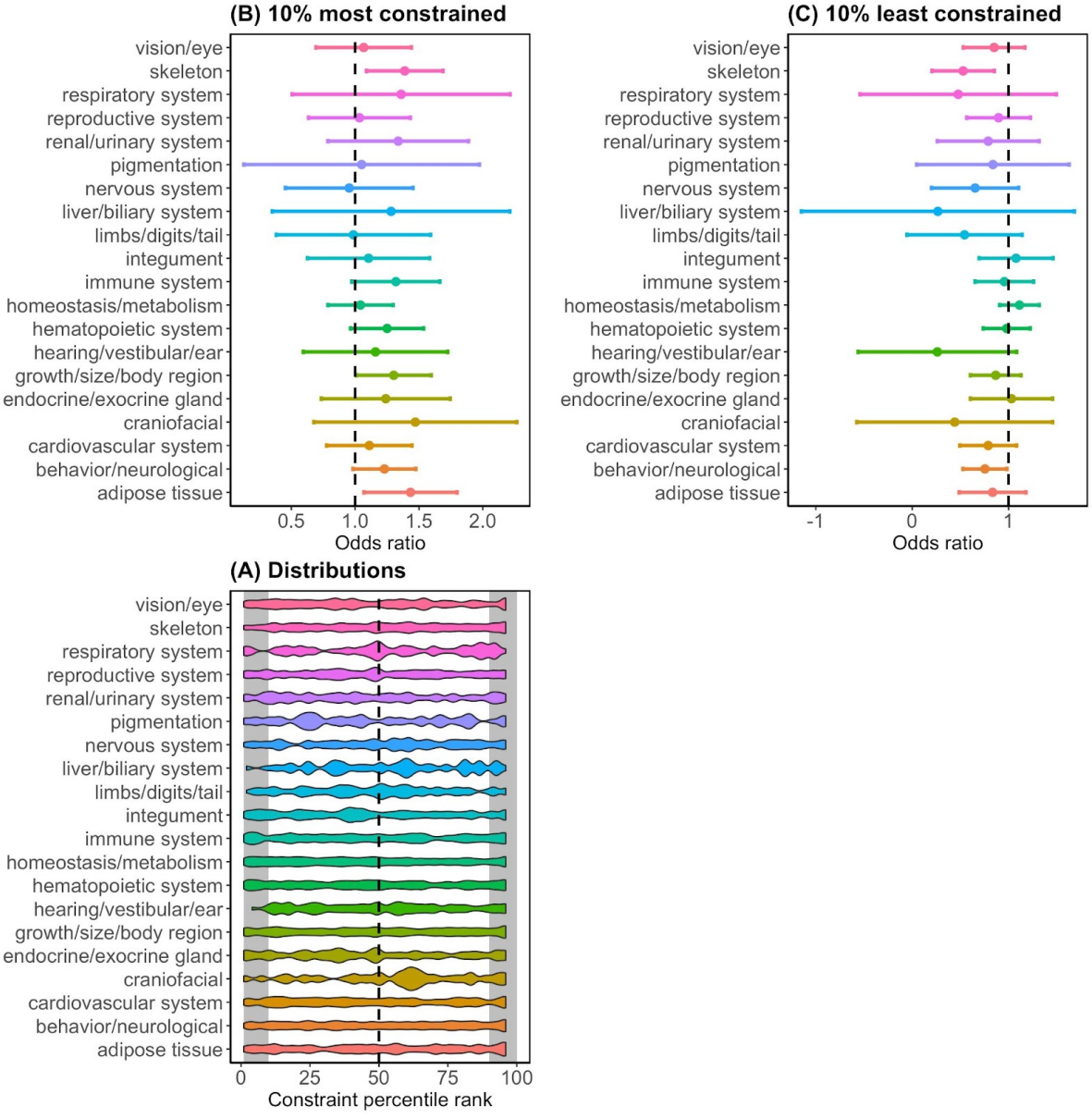
Supplementary Figure 2.5 – Distribution of methylation percentage across CG dinucleotides across the mouse genome. I considered autosomal CG dinucleotides with coverage greater than or equal to five (83.4%). Of these, 8.4% have a methylation percentage less than or equal to 20 and are classified as unmethylated, 87.7% have a methylation percentage more than or equal to 60 and are classified as methylated, and 3.8% are unclassified.

Chapter 3



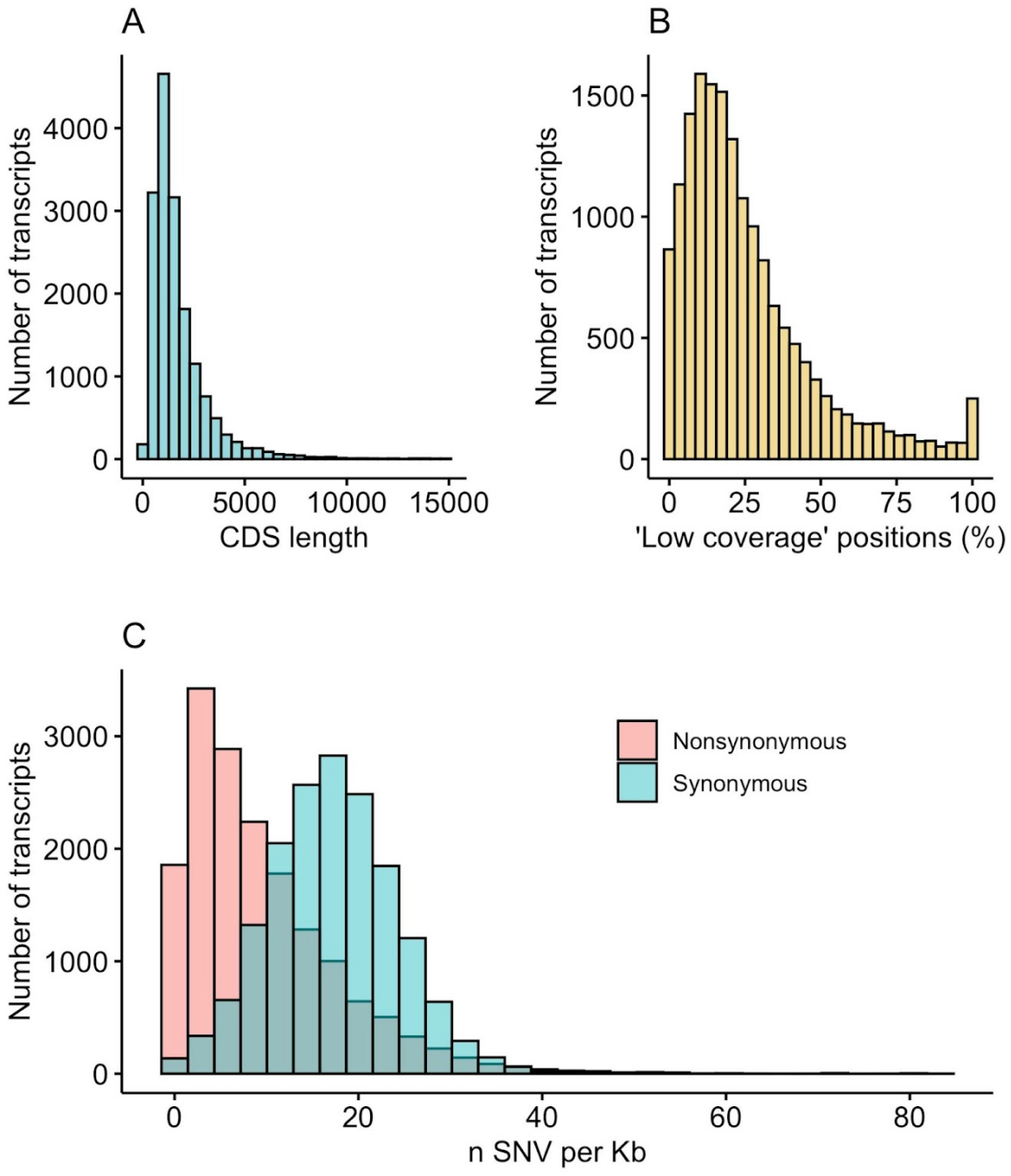
Supplementary Figure 3.1 – Relationship between the observed and expected synonymous (A) and nonsynonymous (B) variant sites for 16,609 mouse genes.

Single nucleotide variant sites are considered between 67 wild mice from 8 populations. A linear model was trained to predict the expected synonymous variant sites in each gene (A). Synonymous variants are assumed not to be under selection, and this model can therefore be used to predict the expected number of nonsynonymous variant sites given neutral evolution (B). Genes that are more constrained have fewer observed nonsynonymous variant sites than expected (B).



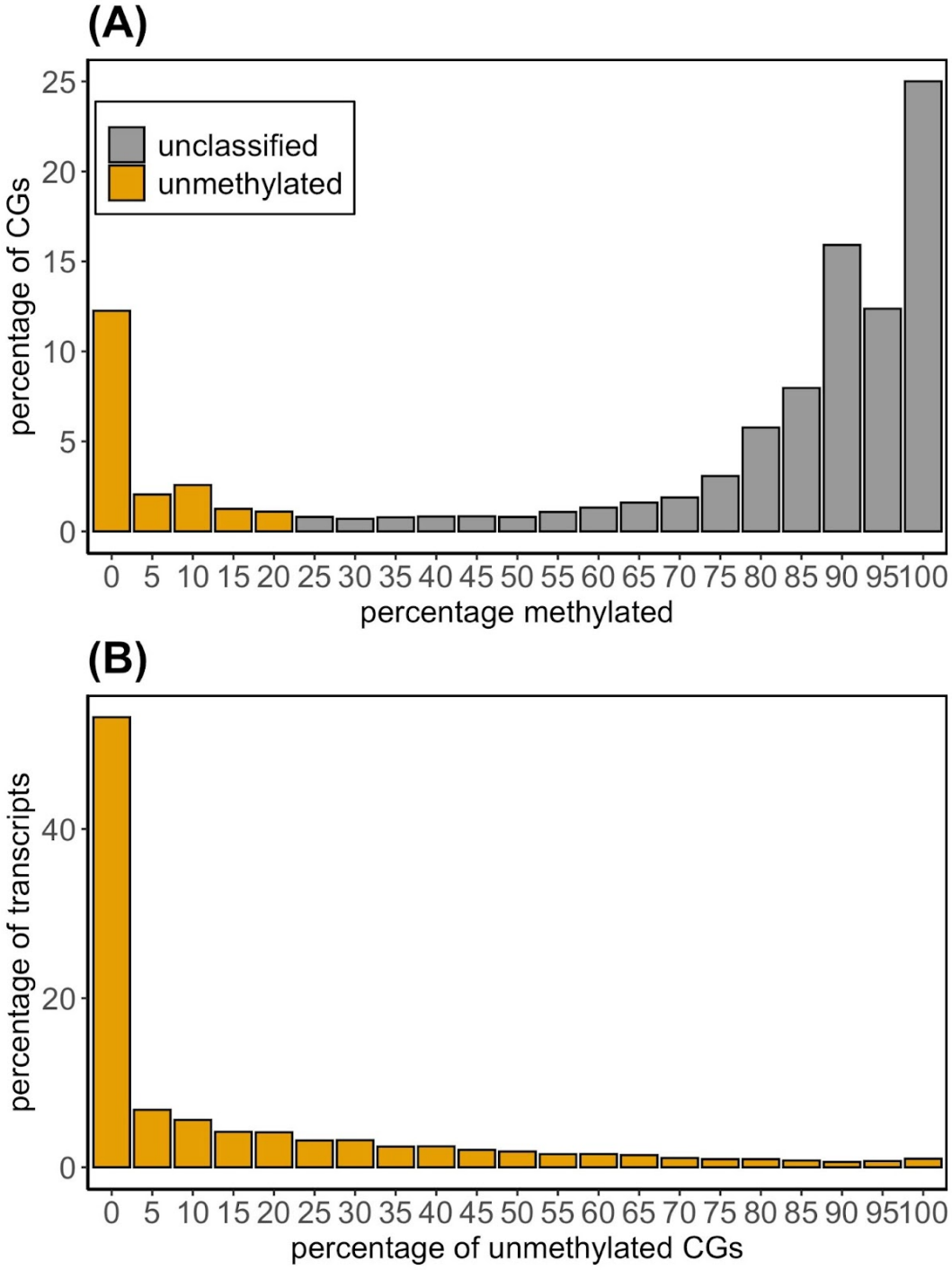
Supplementary Figure 3.2 – (A) Distributions of gene constraint for VP knockouts grouped by their top-level mouse phenotype ontology terms. Constraint is quantified as the nonsynonymous observed expected ratio (NOER) with a higher percentile rank indicating a greater degree of constraint. The dashed line indicates the median score for all genes under consideration. (B and C) Odds ratios and 95% confidence intervals for the 10% most constrained (n = 527), and 10% least

constrained (n = 528) genes across each knockout group. An odds ratio greater than one indicates an enrichment. The most constrained VP genes are enriched for genes associated with adipose tissue and skeleton phenotype terms, and the least constrained genes are not enriched for any top-level phenotype terms.



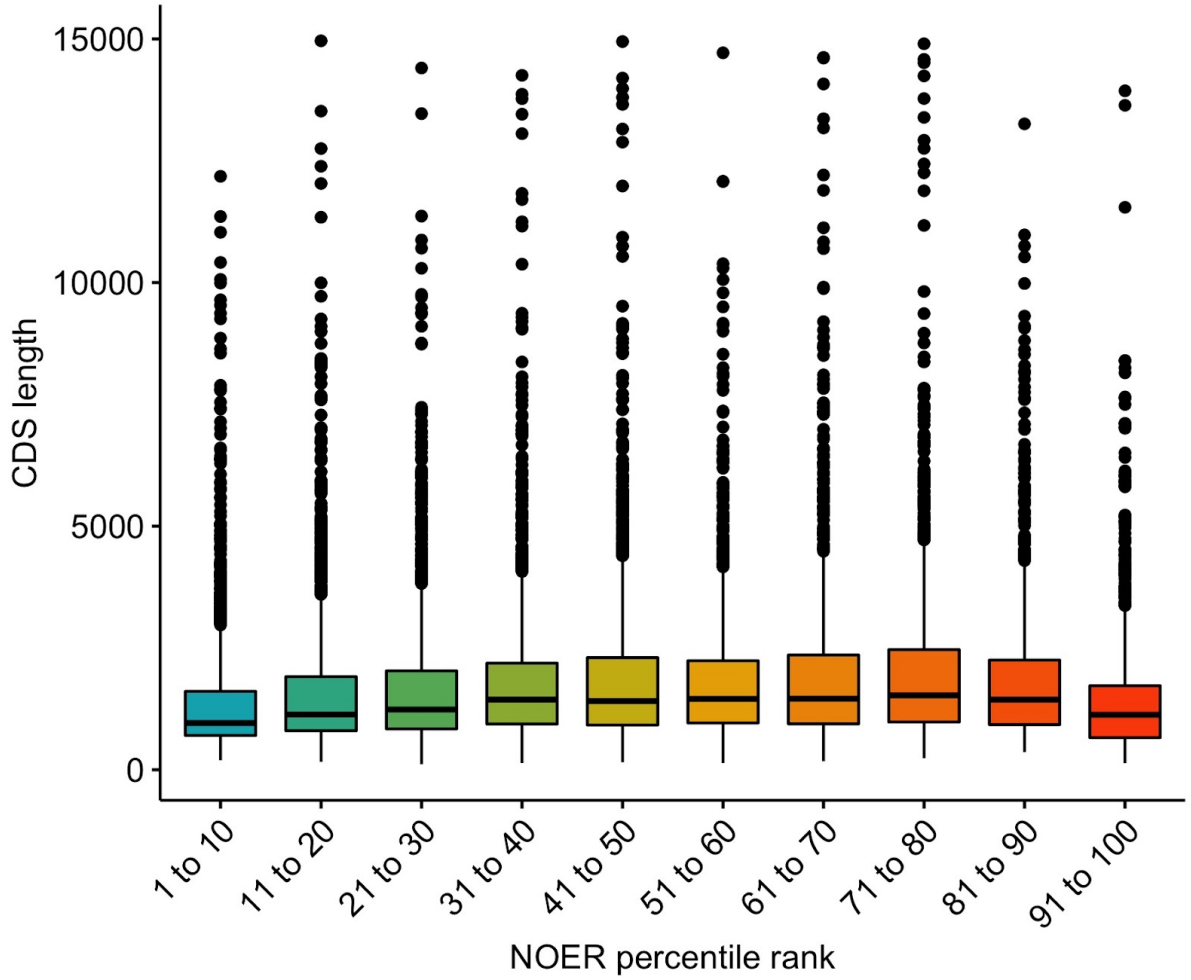
Supplementary Figure 3.3 – A) Distribution of transcripts by their CDS length. B) Distribution of transcripts by their percentage masked as 'low coverage'. Genomic positions were classified as having low coverage if they have a read-depth less than

10X in more than 10% of sequenced mice. C) Distribution of transcritps by their number of nonsynonymous and synonymous variant sites per Kb.



Supplementary Figure 3.4 – (A) The distribution of methylation percentage across CG dinucleotides from 17,526 mouse transcripts. Methylation was classified

from bisulfite sequencing in mouse embryonic stem cells. We considered all CG dinucleotides with coverage greater than or equal to five (91.3%, 825,847/904,384). Of these, 19.2% have a methylation percentage less than or equal to 20 and are classified as unmethylated. (B) The distribution of the percentage of unmethylated CGs across considered transcripts. Most transcripts (50.5%) do not contain any unmethylated CG dinucleotides.



Supplementary Figure 3.5 – Distributions of gene CDS length by NOER percentile rank.

Supplementary Table 3.1 – Odds ratio enrichment of genes by knockout phenotype grouping amongst the most (>=91) and least (<=10) constrained mouse genes. Knockout phenotype groups include lethal (L), sub-viable (SV), viable with no significant phenotype deviation (VN), and viable with a significant phenotype deviation (VP).

Score	Percentile threshold	Phenotype group	Odds ratio	95% confidence Interval

NOER	>=91	L	2.14	1.95 - 2.33
NOER	>=91	SV	1.62	1.36 - 1.88
NOER	>=91	VN	0.47	0 - 0.94
NOER	>=91	VP	0.5	0.32 - 0.68
dN/dS	>=91	L	1.63	1.4 - 1.86
dN/dS	>=91	SV	1.51	1.2 - 1.82
dN/dS	>=91	VN	0.52	0.02 - 1.02
dN/dS	>=91	VP	0.68	0.47 - 0.89
NOER	<=10	L	0.39	0.13 - 0.65
NOER	<=10	SV	0.41	-0.01 - 0.83
NOER	<=10	VN	2.21	1.93 - 2.49
NOER	<=10	VP	1.84	1.65 - 2.03
dN/dS	<=10	L	0.45	0.15 - 0.75
dN/dS	<=10	SV	0.58	0.15 - 1.01
dN/dS	<=10	VN	1.88	1.56 - 2.2
dN/dS	<=10	VP	1.59	1.37 - 1.81

Supplementary Table 3.2 – Transcript-specific probabilities of synonymous and nonsynonymous mutation were calculated by summing the probabilities of all potential substitutions for each mutation type across codons. For example, for codon CCC in the transcript ATG CCC GAA TAG there are three substitutions that cause synonymous substitutions and six that cause nonsynonymous. Therefore the probability of nonsynonymous mutation is P1 + P2 + P3 + P4 + P5 + P6, and the probability of synonymous mutation is P7 + P8 + P9. Note that the final base in the CCC codon is part of a CG dinucleotide (red). If this is classified as unmethylated the 7-mer probabilities of substitution are replaced with the probabilities of substitution for unmethylated CG dinucleotides.

7-mer substitution	Probability	Consequence
ATG CCC G > ATG ACC G	P1	Missense
ATG CCC G > ATG GCC G	P2	Missense
ATG CCC G > ATG TCC G	P3	Missense

TG CCC GA > TG CAC GA	P4	Missense
TG CCC GA > TG CGC GA	P5	Missense
TG CCC GA > TG CTC GA	P6	Missense
G CCC GAA > G CCA GAA	P7	Synonymous
G CCC GAA > G CCG GAA	P8	Synonymous
G CCC GAA > G CCT GAA	P9	Synonymous

Supplementary Table 3.3 – Simple Logistic regression summary statistics and ROC values for predicting mouse knockout outcomes as a function of constraint metrics. Knockout outcomes include lethality (L), lethality or subviability (L or SV), viable with 'no phenotype' deviation (VN), and infertility.

Independe-nt variable	Dependent variable	N	Estimate	Estimate standard error	P value	ROC
L	NOER	1310	0.016252065 39	0.001182655 604	5.69E-43	0.629924186 3
L	Rat-Mus musculus dN/dS	1209	0.009556110 593	0.001179750 072	5.49E-16	0.579551817 4
L	Mus spretus-Mus musculus dN/dS	891	0.009941394 359	0.001357103 87	2.38E-13	0.581738681 5
L	Median GERP	1228	0.007393104 598	0.001183010 035	4.12E-10	0.559821723 9
L	pLI	1268	0.014689417 25	0.001178157 683	1.11E-35	0.619248727 4
L	LOEUF	1268	0.01898048 31	0.001216450 909	6.93E-55	0.650990285 3
L	Missense Z-score	1268	0.013405163 21	0.001175286 125	3.91E-30	0.60907469 4
L or SV	NOER	1805	0.019147414 77	0.00109986 056	7.06E-68	0.651479101 7
L or SV	Rat-Mus musculus dN/dS	1682	0.010754830 46	0.001082674 545	2.97E-23	0.589364068 2

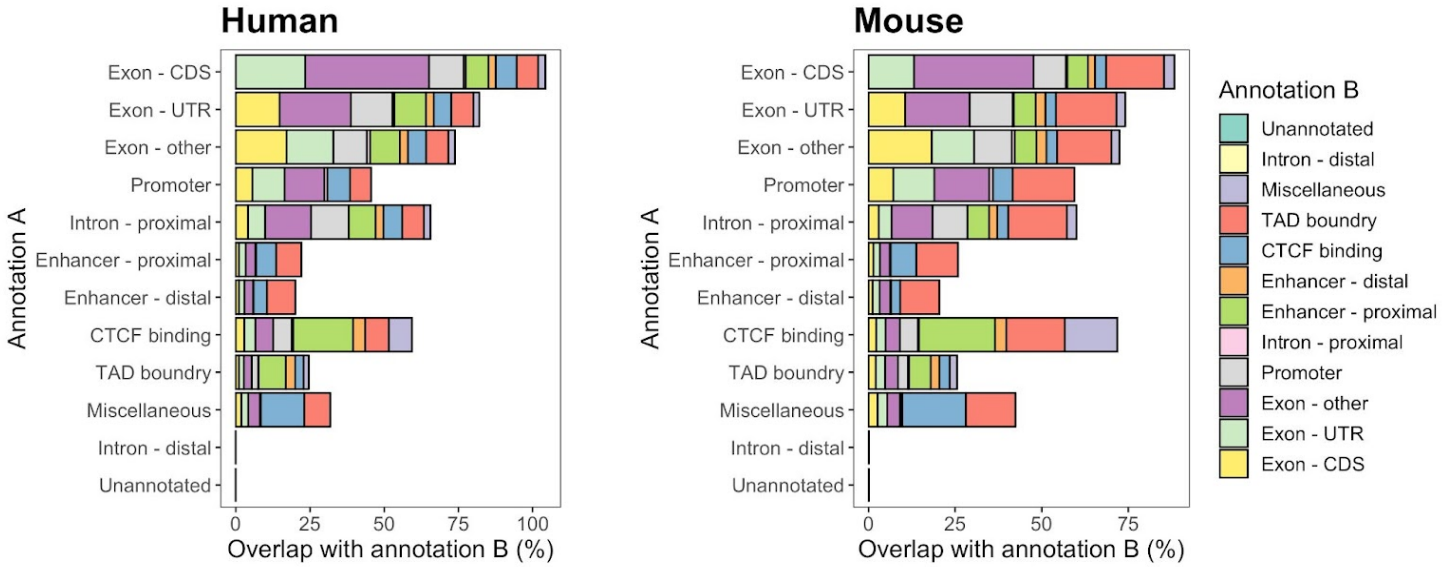
L or SV	Mus spretus-Mus musculus dN/dS	1257	0.012026997 69	0.001233607 784	1.85E-22	0.59839086 32
L or SV	Median GERP	1692	0.00743046 95	0.001083739 875	7.07E-12	0.560057828 8
L or SV	pLI	1753	0.017006479 65	0.001092297 197	1.17E-54	0.636847917 1
L or SV	LOEUF	1753	0.022219936 5	0.001137020 311	4.80E-85	0.674395599 5
L or SV	Missense Z-score	1753	0.015407618 9	0.00108589 6225	1.07E-45	0.624623616 8
VN	NOER	347	-0.01552079 249	0.002041322 94	2.89E-14	0.624156445 8
VN	Rat-Mus musculus dN/dS	313	-0.00937146 3262	0.00206064 4256	5.42E-06	0.578042112
VN	Mus spretus-Mus musculus dN/dS	287	-0.010991171 42	0.002187967 605	5.07E-07	0.589903667 6
VN	Median GERP	324	-0.00952318 6591	0.002066991 626	4.08E-06	0.576177996 6
VN	pLI	346	-0.01274745 27	0.00200240 0509	1.94E-10	0.603569643 9
VN	LOEUF	346	-0.01952757 956	0.002104843 045	1.74E-20	0.653777785 9
VN	Missense Z-score	347	-0.01303128 174	0.002007318 731	8.48E-11	0.60580226 87
Infertility	NOER	246	0.009231396 862	0.00236702 0086	9.62E-05	0.575276792
Infertility	Rat-Mus musculus dN/dS	229	-0.00131904 4656	0.002393118 38	0.581508665 8	0.511072596
Infertility	Mus spretus-Mus musculus dN/dS	190	0.000999174 1461	0.002629316 983	0.703935903 1	0.50834805 42
Infertility	Median GERP	235	0.002293135 543	0.00240645 6472	0.34063586 09	0.518636395 7
Infertility	pLI	238	0.001454688 519	0.002347944 801	0.535548675 6	0.512065376 6
Infertility	LOEUF	238	0.00808067 232	0.002384759 178	0.00070285 07204	0.566667113 6
Infertility	Missense	238	0.00329634	0.001912900	0.08484852	0.539373437

	Z-score		9998	052	43	1

Chapter 4

Supplementary Table 4.1 – Genomic coverage for the human and mouse annotation models. All annotated bases excluding “Intron-distal” or “Unannotated” are annotated as “Functional - all”. All functional bases excluding “Exon - CDS” are annotated as “Functional - nonCDS”. Percentages are rounded to one decimal place.

Annotation	**Mouse genomic coverage (%)**	**Human genomic coverage (%)**
Exon - CDS	1.4	1.2
Exon - UTR	1.8	1.9
Exon - other	2.7	2.9
Promoter	1.8	2.5
Enhancer - proximal	6.0	8.8
Enhancer - distal	2.5	2.7
CTCF binding	2.0	3.0
Miscellaneous	1.7	1.6
TAD boundary	11.4	8.2
Intron - proximal	0.2	0.2
Intron - distal	30.4	38.6
Unannotated	43.6	33.9
Functional - all	26.0	27.5
Functional - nonCDS	24.6	26.3



Supplementary Figure 4.1 – Overlap of genomic annotations for humans and mice. Overlap is calculated as the percentage of bases with annotation A that also have annotation B.

Supplementary Table 4.2 – Overlap of genomic annotations for humans and mice. Overlap is calculated as the percentage of bases with annotation A that also have annotation B. percentages are rounded to one decimal place. Annotation B key: 1. Exon - CDS, 2. Exon - UTR, 3. Exon - other, 4. Promoter, 5. Enhancer - proximal, 6. Enhancer - distal, 7. CTCF binding, 8. Miscellaneous, 9. TAD boundary, 10. Intron - proximal, 11. Intron - distal, 12. Unannotated.

		Annotation B (see key in legend)											
Species	**Annotation A**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Human	Exon - CDS	100	23.4	41.7	11.6	7.7	2.4	7.1	2.5	7.2	0.7	0	0
Human	Exon - UTR	14.7	100.0	24.0	14.1	10.7	2.6	5.9	2.0	7.4	0.6	0	0
Human	Exon - other	17.1	15.7	100.0	11.3	10.0	2.7	6.2	2.2	7.5	1.1	0	0
Human	Promoter	5.6	10.8	13.4	100.0	0	0	7.6	0	7.1	1	0	0
Human	Enhancer - proximal	1.0	2.3	3.3	0	100.0	0	6.8	0	8.5	0.2	0	0
Human	Enhancer - distal	1.0	1.8	2.9	0	0	100.0	4.5	0	9.5	0.2	0	0
Human	CTCF binding	2.8	3.8	6.0	6.3	20.2	4.1	100.0	7.9	7.9	0.4	0	0
Human	Miscellaneous	1.8	2.3	4.0	0	0	0	14.7	100.0	8.7	0.3	0	0
Human	TAD boundary	1.0	1.7	2.6	2.1	9.2	3.1	2.9	1.7	100.0	0.2	0	0

Human	Intron - proximal	4.1	5.8	15.5	12.7	9.0	2.6	6.3	2.2	7.3	100.0	0	0
Human	Intron - distal	0	0	0	0	0	0	0	0	0	0	100.0	0
Human	Unannotated	0	0	0	0	0	0	0	0	0	0	0	100.0
Mouse	Exon - CDS	100.0	13.1	34.6	9.3	6	2.1	3.2	3	16.7	0.4	0	0
Mouse	Exon - UTR	10.6	100.0	18.6	12.3	6.3	2.8	3.0	2.5	17.4	0.4	0	0
Mouse	Exon - other	18.3	12.2	100.0	10.9	6.4	2.8	3.1	2.3	15.7	0.8	0	0
Mouse	Promoter	7.2	11.8	15.9	100.0	0	0	5.7	0	17.9	1.0	0	0
Mouse	Enhancer - proximal	1.4	1.9	2.9	0	100.0	0.1	7.4	0.1	12	0.2	0	0
Mouse	Enhancer - distal	1.2	2.0	3.0	0	0.1	100.0	2.6	0	11.3	0.2	0	0
Mouse	CTCF binding	2.2	2.7	4.1	5.2	22	3.3	100.0	15.1	16.9	0.3	0	0
Mouse	Miscellaneous	2.6	2.7	3.7	0	0.3	0	18.4	100.0	14.3	0.3	0	0
Mouse	TAD boundary	2.1	2.7	3.7	2.9	6.3	2.5	3.0	2.1	100.0	0.3	0	0
Mouse	Intron - proximal	3.0	3.7	11.8	10.1	6.3	2.4	3.2	2.8	16.9	100.0	0	0
Mouse	Intron - distal	0	0	0	0	0	0	0	0	0	0	100.0	0
Mouse	Unannotated	0	0	0	0	0	0	0	0	0	0	0	100.0

Supplementary Table 4.3 – Alignment of human annotation categories to the mouse genome. All bases not annotated as “Intron-distal” or “Unannotated” are annotated as “Functional”. All functional bases excluding “Exon - CDS” are annotated as “Functional - nonCDS”. “Total” refers to all annotated bases.

Annotation	**N human bp**	**N bp aligned to mouse**	**N bp conserved in mouse**	**Bp aligned to mouse (%)**	**Bp conserved in mouse (%)**
Exon - CDS	34252968	32705717	30220215	95.5	88.2
Exon - UTR	54564547	36874207	22275618	67.6	40.8
Exon - other	83458031	50184072	14210208	60.1	17.0
Promoter	70819139	42097585	27232358	59.4	38.5
Enhancer - proximal	254279321	106354426	23270865	41.8	9.2
Enhancer -	78036529	35379976	2399522	45.3	3.1

distal					
CTCF binding	85645600	35476812	8600833	41.4	10.0
Miscellaneous	46045668	18058798	1196714	39.2	2.6
TAD boundary	236803513	76778664	11400803	32.4	4.8
Intron - proximal	5762460	4456928	3226081	77.3	56.0
Intron - distal	1110486349	328329971	173169384	29.6	15.6
Unannotated	974067966	174098559	115871796	17.9	11.9
Functional - all	790747213	338891270	NA	42.9	NA
Functional - nonCDS	756494245	306185553	NA	40.5	NA
Total - nonCDS	2841048560	808614083	NA	28.5	NA
Total	2875301528	841319800	NA	29.3	NA

Supplementary Table 4.4 – Alignment and conservation of tissue-specific human enhancers to the mouse genome.

Annotation	**Tissue**	**Human genomic coverage (%)**	**Bp aligned with mouse (%)**	**Bp conserved in mouse (%)**
Enhancer - proximal	heart	1.7	47.1	5.1
Enhancer - distal	heart	0.3	53.9	2.7
Enhancer - proximal	kidney	0.7	51.0	4.6
Enhancer - distal	kidney	0.2	60.2	1.6
Enhancer - proximal	spleen	1.4	42.2	1.6
Enhancer - distal	spleen	0.1	43.5	0.4

Supplementary Table 4.5 – Summary of SNV sites associated with complex and Mendelian disease that have an orthologous base in mice. All SNV sites not annotated as “Intron-distal” or “Unannotated” are annotated as “Functional”. All SNV sites not annotated as “Exon - CDS” are annotated as “Non - CDS”. “Total” refers to all annotated SNV sites.

Disease class	Annotation	N SNV	N SNV aligned	N SNV conserved	SNV aligned (%)	SNV conserved (%)
Mendelian	Exon - CDS	33935	33709	33326	99.3	98.2
Mendelian	Exon - UTR	12078	11998	3070	99.3	25.4
Mendelian	Exon - other	18877	18772	8423	99.4	44.6
Mendelian	Promoter	3586	3552	2540	99.1	70.8
Mendelian	Enhancer - proximal	3227	3206	394	99.3	12.2
Mendelian	Enhancer - distal	1093	1084	71	99.2	6.5
Mendelian	CTCF binding	2703	2688	437	99.4	16.2
Mendelian	Miscellaneous	1055	1053	52	99.8	4.9
Mendelian	TAD boundary	4201	4157	1291	99.0	30.7
Mendelian	Intron - proximal	8220	8150	7917	99.1	96.3
Mendelian	Intron - distal	117	81	50	69.2	42.7
Mendelian	Unannotated	6	6	1	100.0	16.7
Mendelian	Functional - all	41916	41596	NA	99.2	NA
Mendelian	Functional - nonCDS	7981	7887	NA	98.8	NA
Mendelian	Total - nonCDS	8104	7974	NA	98.4	NA
Mendelian	Total	42039	41683	NA	99.2	NA
Complex	Exon - CDS	1008	966	907	95.8	90.0

Complex	Exon - UTR	1056	778	436	73.7	41.3
Complex	Exon - other	1603	1091	361	68.1	22.5
Complex	Promoter	1632	1078	655	66.1	40.1
Complex	Enhancer - proximal	3467	1566	369	45.2	10.6
Complex	Enhancer - distal	873	404	27	46.3	3.1
Complex	CTCF binding	1182	574	139	48.6	11.8
Complex	Miscellaneous	496	192	21	38.7	4.2
Complex	TAD boundary	2311	928	135	40.2	5.8
Complex	Intron - proximal	116	92	62	79.3	53.4
Complex	Intron - distal	10831	3711	1946	34.3	18.0
Complex	Unannotated	6395	1648	983	25.8	15.4
Complex	Functional - all	10568	5370	NA	50.8	NA
Complex	Functional - nonCDS	9560	4404	NA	46.1	NA
Complex	Total - nonCDS	26786	9763	NA	36.4	NA
Complex	Total	27794	10729	NA	38.6	NA

Supplementary Table 4.6 – Z-scores and corresponding p-values from testing the null hypothesis that there is no difference in the proportion of SNV sites associated with Mendelian disease that align to the mouse genome and the proportion of SNVs associated with complex disease that align to the mouse genome by annotation category. Z-scores are rounded to one decimal place. P-values are given in scientific notation. P-values of 0 are less than 1.00E-300.

Annotation	**Alignment z-score**	**Alignment p-value**	**Conservation z-score**	**Conservation p-value**

Exon - CDS	12.6	3.70E-36	18.2	2.50E-74
Exon - UTR	49.1	0.00E+00	-11.2	5.10E-29
Exon - other	70.6	0.00E+00	17.2	4.90E-66
Promoter	34.9	1.40E-267	21.1	8.10E-99
Enhancer - proximal	49.0	0.00E+00	2.0	4.40E-02
Enhancer - distal	27.2	1.60E-162	3.4	5.70E-04
CTCF binding	39.8	0.00E+00	3.6	3.80E-04
Miscellaneous	28.2	5.60E-175	0.6	5.50E-01
TAD boundary	54.9	0.00E+00	23.2	1.90E-119
Intron - proximal	20.1	8.40E-90	22.6	1.60E-113
Intron - distal	7.9	2.60E-15	6.9	5.10E-12
Unannotated	4.2	3.30E-05	0.1	9.30E-01
Functional - all	145	0.00E+00	NA	NA
Functional - nonCDS	76.0	0.00E+00	NA	NA
Total - nonCDS	97.7	0.00E+00	NA	NA
Total	181.0	0.00E+00	NA	NA

Supplementary Table 4.7 – Z-scores and corresponding p-values from testing the null hypothesis that there is no difference in the proportion of SNV sites associated with Mendelian disease that align to the mouse genome and the total proportion of bases that align to the mouse genome by annotation category. Z-scores are rounded to one decimal place. P-values are given in scientific notation. P-values of 0 are less than 1.00E-300.

Annotation	**Alignment z-score**	**Alignment p-value**	**Conservation z-score**	**Conservation p-value**
Exon - CDS	34.2	1.10E-255	57.0	0.00E+00
Exon - UTR	74.6	0.00E+00	-34.4	5.40E-260
Exon - other	110.3	0.00E+00	100.8	0.00E+00
Promoter	48.3	0.00E+00	39.9	0.00E+00
Enhancer -	66.2	0.00E+00	6.0	1.70E-09

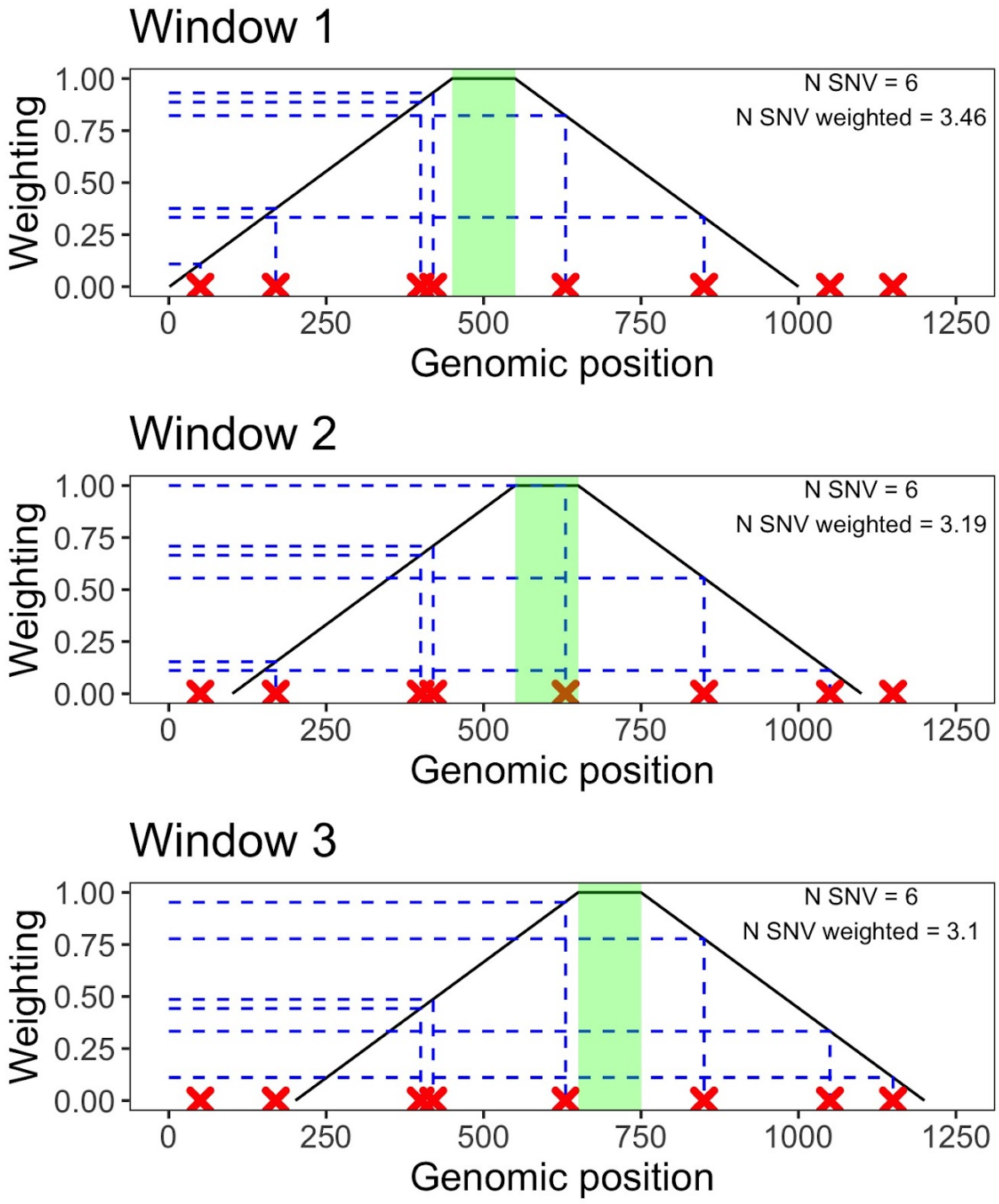
proximal				
Enhancer - distal	35.8	5.70E-280	6.6	5.70E-11
CTCF binding	61.2	0.00E+00	10.6	3.20E-26
Miscellaneous	40.3	0.00E+00	4.8	2.00E-06
TAD boundary	92.1	0.00E+00	78.5	0.00E+00
Intron - proximal	47.2	0.00E+00	73.6	0.00E+00
Intron - distal	9.4	5.40E-21	8.1	5.90E-16
Unannotated	5.3	1.50E-07	0.4	7.20E-01
Functional - all	233.2	0.00E+00	NA	NA
Functional - nonCDS	106.2	0.00E+00	NA	NA
Total - nonCDS	139.5	0.00E+00	NA	NA
Total	315.0	0.00E+00	NA	NA

Supplementary Table 4.8 – Z-scores and corresponding p-values from testing the null hypothesis that there is no difference in the proportion of SNV sites associated with complex disease that align to the mouse genome and the total proportion of bases that align to the mouse genome by annotation category. Z-scores are rounded to one decimal place. P-values are given in scientific notation. P-values of 0 are less than 1.00E-300.

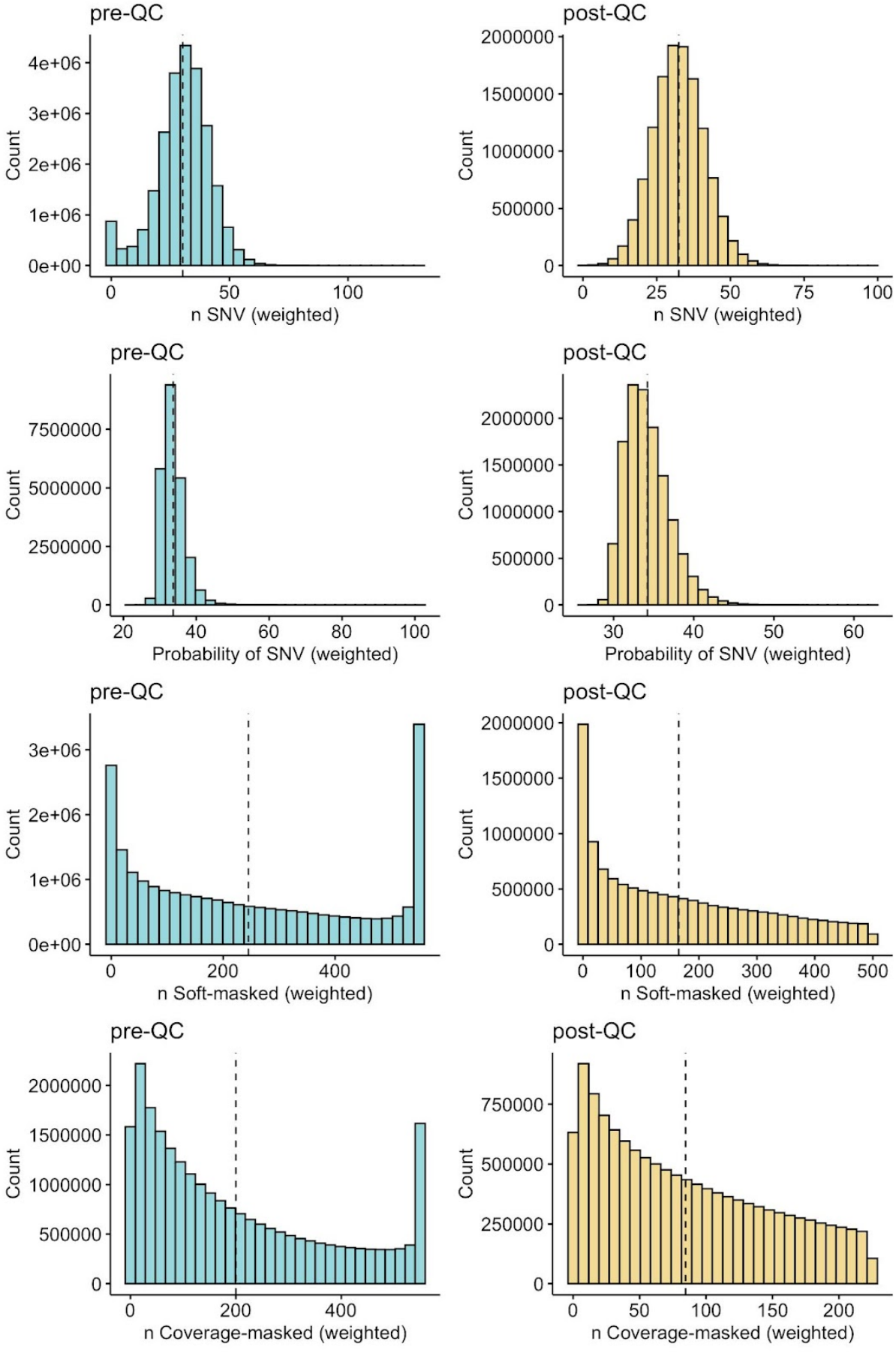
Annotation	**Alignment z-score**	**Alignment p-value**	**Conservation z-score**	**Conservation p-value**
Exon - CDS	0.5	5.90E-01	1.7	8.40E-02
Exon - UTR	4.2	2.30E-05	0.3	7.60E-01
Exon - other	6.5	9.00E-11	5.9	4.90E-09
Promoter	5.4	5.40E-08	1.4	1.60E-01
Enhancer - proximal	4.0	6.60E-05	3.0	2.30E-03
Enhancer - distal	0.6	5.80E-01	0.0	9.80E-01
CTCF binding	5.0	6.30E-07	2.0	4.90E-02
Miscellaneous	-0.2	8.20E-01	2.3	2.20E-02
TAD boundary	7.9	2.00E-15	2.3	2.10E-02

Intron - proximal	0.5	6.10E-01	-0.6	5.80E-01
Intron - distal	10.7	9.10E-27	6.8	1.00E-11
Unannotated	16.5	4.90E-61	8.6	9.00E-18
Functional - all	16.5	2.30E-61	NA	NA
Functional - nonCDS	11.1	8.00E-29	NA	NA
Total - nonCDS	29.0	1.7E-184	NA	NA
Total	34.2	8.20E-257	NA	NA

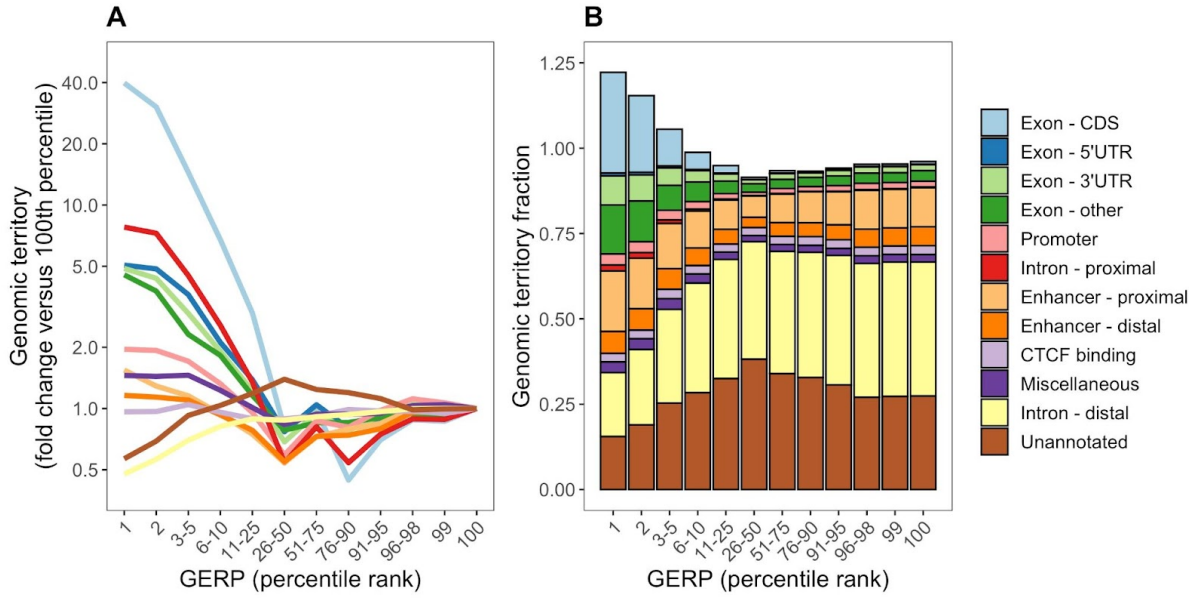
Chapter 5



Supplementary Figure 5.1 – Schematic illustrating how variables were weighted for each window. Windows were 1000 bp in length, sliding every 100 bp. In order to prioritise the middle 100 bp to which the score for the window is attributed, single nucleotide variant counts, probabilities of substitution, and GERP scores were weighted based on location in the 1000 bp window. The middle 100 bp for each window was given a weighting of one, and the flanking 450 bp were weighted from one to zero based on their linear distance from the central 100 bp (i.e. base 101 is weighted as 449/450, and base 1000 is weighted as 1/450).


pre-QC
post-QC
Count
n SNV (weighted)
Probability of SNV (weighted)
n Soft-masked (weighted)
n Coverage-masked (weighted)

Supplementary Figure 5.2 – Distribution of variables across all windows pre and post quality control (QC).



Supplementary Figure 5.3 – Genomic feature composition of the most and least constrained regions across the mouse autosome ranked by their average GERP score. Genomic features are defined by their GENCODE and Ensembl Regulatory Build annotations (see methods). Panel B shows the fraction of genomic territory of each percentile grouping annotated with each genomic feature. Panel A shows the change in the genomic territory of each percentile grouping relative to the least constrained regions (100th percentile). Note that GERP percentiles (x-axis) are not uniformly grouped to highlight the changes in the most and least constrained regions.

Supplementary Table 5.1 – Genomic coverage for the mouse annotation models. Percentages are rounded to one decimal place.

Annotation	Mouse genomic coverage (%)	Human genomic coverage (%)
Exon - CDS	1.4	1.2
Exon - UTR	1.8	1.9
Exon - other	2.7	2.9
Promoter	1.8	2.5
Enhancer - proximal	6.0	8.8

Enhancer - distal	2.5	2.7
CTCF binding	2.0	3.0
Miscellaneous	1.7	1.6
Intron - proximal	0.2	0.2
Intron - distal	30.4	38.6
Unannotated	43.6	33.9

Supplementary Table 5.2 – Overlap of mouse genomic annotations. Overlap is calculated as the percentage of bases with annotation A that also have annotation B. percentages are rounded to one decimal place. Annotation B key: 1. Exon - CDS, 2. Exon - UTR, 3. Exon - other, 4. Promoter, 5. Enhancer - proximal, 6. Enhancer - distal, 7. CTCF binding, 8. Miscellaneous, 9. Intron - proximal, 10. Intron - distal, 11. Unannotated.

	Annotation B (see key in legend)										
Annotation A	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Exon - CDS	100.0	13.1	34.6	9.3	6	2.1	3.2	3	0.4	0	0
Exon - UTR	10.6	100.0	18.6	12.3	6.3	2.8	3.0	2.5	0.4	0	0
Exon - other	18.3	12.2	100.0	10.9	6.4	2.8	3.1	2.3	0.8	0	0
Promoter	7.2	11.8	15.9	100.0	0	0	5.7	0	1.0	0	0
Enhancer - proximal	1.4	1.9	2.9	0	100.0	0.1	7.4	0.1	0.2	0	0
Enhancer - distal	1.2	2.0	3.0	0	0.1	100.0	2.6	0	0.2	0	0
CTCF binding	2.2	2.7	4.1	5.2	22	3.3	100.0	15.1	0.3	0	0
Miscellaneous	2.6	2.7	3.7	0	0.3	0	18.4	100.0	0.3	0	0
TAD boundary	2.1	2.7	3.7	2.9	6.3	2.5	3.0	2.1	0.3	0	0
Intron - proximal	3.0	3.7	11.8	10.1	6.3	2.4	3.2	2.8	100.0	0	0
Intron - distal	0	0	0	0	0	0	0	0	0	100.0	0
Unannotated	0	0	0	0	0	0	0	0	0	0	100.0

References

Abramovs, Nikita, Andrew Brass, and May Tassabehji. 2019. “GeVIR Is a Continuous Gene-Level Metric That Uses Variant Distribution Patterns to Prioritize Disease Candidate Genes.” *Nature Genetics*, December. https://doi.org/10.1038/s41588-019-0560-2.

Adams, David J., Anthony G. Doran, Jingtao Lilue, and Thomas M. Keane. 2015. “The Mouse Genomes Project: A Repository of Inbred Laboratory Mouse Strain Genomes.” *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 26 (9-10): 403–12.

Aggarwala, Varun, and Benjamin F. Voight. 2016. “An Expanded Sequence Context Model Broadly Explains Variability in Polymorphism Levels across the Human Genome.” *Nature Genetics* 48 (4): 349–55.

Aikens, Rachael C., Kelsey E. Johnson, and Benjamin F. Voight. 2019. “Signals of Variation in Human Mutation Rate at Multiple Levels of Sequence Context.” *Molecular Biology and Evolution* 36 (5): 955–65.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. “The Repertoire of Mutational Signatures in Human Cancer.” *Nature* 578 (7793): 94–101.

Amemiya, Haley M., Anshul Kundaje, and Alan P. Boyle. 2019. “The ENCODE Blacklist: Identification of Problematic Regions of the Genome.” *Scientific Reports* 9 (1): 9354.

Barghi, Neda, Joachim Hermisson, and Christian Schlötterer. 2020. “Polygenic Adaptation: A Unifying Framework to Understand Positive Selection.” *Nature Reviews. Genetics* 21 (12): 769–81.

Bartha, István, Julia di Iulio, J. Craig Venter, and Amalio Telenti. 2018. “Human Gene Essentiality.” *Nature Reviews. Genetics* 19 (1): 51–62.

Bartha, István, Antonio Rausell, Paul J. McLaren, Pejman Mohammadi, Manuel Tardaguila, Nimisha Chaturvedi, Jacques Fellay, and Amalio Telenti. 2015. “The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome.” *PLoS Computational Biology* 11 (12): e1004647.

Bauer, Christopher R., Shuang Li, and Mark L. Siegal. 2015. “Essential Gene Disruptions Reveal Complex Relationships between Phenotypic Robustness, Pleiotropy, and Fitness.” *Molecular Systems Biology* 11 (1): 773.

Beck, J. A., S. Lloyd, M. Hafezparast, M. Lennon-Pierce, J. T. Eppig, M. F. Festing, and E. M. Fisher. 2000. “Genealogies of Mouse Inbred Strains.” *Nature Genetics* 24 (1): 23–25.

Begun, D. J., and C. F. Aquadro. 1992. “Levels of Naturally Occurring DNA Polymorphism Correlate with Recombination Rates in D. Melanogaster.” *Nature* 356 (6369): 519–20.

Berthelot, Camille, Diego Villar, Julie E. Horvath, Duncan T. Odom, and Paul Flicek. 2018. “Complexity and Conservation of Regulatory Landscapes Underlie Evolutionary Resilience of Mammalian Gene Expression.” *Nature Ecology & Evolution* 2 (1): 152–63.

Birling, Marie-Christine, Atsushi Yoshiki, David J. Adams, Shinya Ayabe, Arthur L. Beaudet, Joanna Bottomley, Allan Bradley, et al. 2021. “A Resource of Targeted Mutant Mouse Lines for 5,061 Genes.” *Nature Genetics* 53 (4): 416–19.

Bonev, Boyan, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L.

Papadopoulos, Yaniv Lubling, Xiaole Xu, et al. 2017. “Multiscale 3D Genome Rewiring during Mouse Neural Development.” *Cell* 171 (3): 557–72.e24.
Bonhomme, François, Eric Rivals, Annie Orth, Gemma R. Grant, Alec J. Jeffreys, and Philippe R. J. Bois. 2007. “Species-Wide Distribution of Highly Polymorphic Minisatellite Markers Suggests Past and Present Genetic Exchanges among House Mouse Subspecies.” *Genome Biology* 8 (5): R80.
Booker, Tom R., and Peter D. Keightley. 2018. “Understanding the Factors That Shape Patterns of Nucleotide Diversity in the House Mouse Genome.” *Molecular Biology and Evolution* 35 (12): 2971–88.
Bowl, Michael R., Michelle M. Simon, Neil J. Ingham, Simon Greenaway, Luis Santos, Heather Cater, Sarah Taylor, et al. 2017. “A Large Scale Hearing Loss Screen Reveals an Extensive Unexplored Genetic Landscape for Auditory Dysfunction.” *Nature Communications* 8 (1): 886.
Brown, Steve D. M. 2021. “Advances in Mouse Genetics for the Study of Human Disease.” *Human Molecular Genetics* 30 (R2): R274–84.
Brown, Steve D. M., and Mark W. Moore. 2012. “The International Mouse Phenotyping Consortium: Past and Future Perspectives on Mouse Phenotyping.” *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 23 (9-10): 632–40.
Bullaughey, Kevin. 2011. “Changes in Selective Effects over Time Facilitate Turnover of Enhancer Sequences.” *Genetics* 187 (2): 567–82.
Bulmer, M. 1986. “Neighboring Base Effects on Substitution Rates in Pseudogenes.” *Molecular Biology and Evolution* 3 (4): 322–29.
Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019.” *Nucleic Acids Research* 47 (D1): D1005–12.
Cacheiro, Pilar, Violeta Muñoz-Fuentes, Stephen A. Murray, Mary E. Dickinson, Maja Bucan, Lauryl M. J. Nutter, Kevin A. Peterson, et al. 2020. “Human and Mouse Essentiality Screens as a Resource for Disease Gene Discovery.” *Nature Communications* 11 (1): 655.
Cagan, Alex, Adrian Baez-Ortega, Natalia Brzozowska, Federico Abascal, Tim H. H. Coorens, Mathijs A. Sanders, Andrew R. J. Lawson, et al. 2022. “Somatic Mutation Rates Scale with Lifespan across Mammals.” *Nature*, April. https://doi.org/10.1038/s41586-022-04618-z.
Cannon, Maren E., and Karen L. Mohlke. 2018. “Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci.” *American Journal of Human Genetics* 103 (5): 637–53.
Cassa, Christopher A., Donate Weghorn, Daniel J. Balick, Daniel M. Jordan, David Nusinow, Kaitlin E. Samocha, Anne O’Donnell-Luria, et al. 2017. “Estimating the Selective Effects of Heterozygous Protein-Truncating Variants from Human Exome Data.” *Nature Genetics* 49 (5): 806–10.
Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. “The Effect of Deleterious Mutations on Neutral Molecular Variation.” *Genetics* 134 (4): 1289–1303.
Charlesworth, Brian. 2009. “Fundamental Concepts in Genetics: Effective Population Size and Patterns of Molecular Evolution and Variation.” *Nature Reviews. Genetics* 10 (3): 195–205.
Chatterjee, Nimrat, and Graham C. Walker. 2017. “Mechanisms of DNA Damage, Repair, and Mutagenesis.” *Environmental and Molecular Mutagenesis* 58 (5):

235–63.
Chen, Chun-Long, Aurélien Rappailles, Lauranne Duquenne, Maxime Huvet, Guillaume Guilbaud, Laurent Farinelli, Benjamin Audit, et al. 2010. “Impact of Replication Timing on Non-CpG and CpG Substitution Rates in Mammalian Genomes.” *Genome Research* 20 (4): 447–57.
Cheng, Yong, Zhihai Ma, Bong-Hyun Kim, Weisheng Wu, Philip Cayting, Alan P. Boyle, Vasavi Sundaram, et al. 2014. “Principles of Regulatory Information Conservation between Mouse and Human.” *Nature* 515 (7527): 371–75.
Chen, Siwei, Laurent Francioli, Julia Goodrich, Ryan Collins, Qingbo Wang, Jessica Alfoldi, Nicholas Watts, et al. 2022. “A Genome-Wide Mutational Constraint Map Quantified from Variation in 76,156 Human Genomes.” *bioRxiv*. https://doi.org/10.1101/2022.03.20.485034.
Chimpanzee Sequencing and Analysis Consortium. 2005. “Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome.” *Nature* 437 (7055): 69–87.
Churchill, Gary A., David C. Airey, Hooman Allayee, Joe M. Angel, Alan D. Attie, Jackson Beatty, William D. Beavis, et al. 2004. “The Collaborative Cross, a Community Resource for the Genetic Analysis of Complex Traits.” *Nature Genetics* 36 (11): 1133–37.
Comeron, Josep M. 2017. “Background Selection as Null Hypothesis in Population Genomics: Insights and Challenges from Drosophila Studies.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372 (1736). https://doi.org/10.1098/rstb.2016.0471.
Cong, Le, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, et al. 2013. “Multiplex Genome Engineering Using CRISPR/Cas Systems.” *Science* 339 (6121): 819–23.
Cooper, Tim F., Elizabeth A. Ostrowski, and Michael Travisano. 2007. “A Negative Relationship between Mutation Pleiotropy and Fitness Effect in Yeast.” *Evolution; International Journal of Organic Evolution* 61 (6): 1495–99.
Coulondre, C., J. H. Miller, P. J. Farabaugh, and W. Gilbert. 1978. “Molecular Basis of Base Substitution Hotspots in Escherichia Coli.” *Nature* 274 (5673): 775–80.
Crunkhorn, Sarah. 2020. “Using Loss-of-Function Human Mutations to Evaluate Drug Targets.” *Nature Reviews. Drug Discovery* 19 (7): 446.
Cusanovich, Darren A., Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, et al. 2018. “A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.” *Cell* 174 (5): 1309–24.e18.
Cutter, Asher D., and Bret A. Payseur. 2013. “Genomic Signatures of Selection at Linked Sites: Unifying the Disparity among Species.” *Nature Reviews. Genetics* 14 (4): 262–74.
Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. “Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++.” *PLoS Computational Biology* 6 (12): e1001025.
Dejager, Lien, Claude Libert, and Xavier Montagutelli. 2009. “Thirty Years of Mus Spretus: A Promising Future.” *Trends in Genetics: TIG* 25 (5): 234–41.
Denas, Olgert, Richard Sandstrom, Yong Cheng, Kathryn Beal, Javier Herrero, Ross C. Hardison, and James Taylor. 2015. “Genome-Wide Comparative Analysis Reveals Human-Mouse Regulatory Landscape and Evolution.” *BMC Genomics* 16 (February): 87.
Devoy, Anny, Rosie K. A. Bunton-Stasyshyn, Victor L. J. Tybulewicz, Andrew J. H. Smith, and Elizabeth M. C. Fisher. 2011. “Genomically Humanized Mice:

Technologies and Promises." *Nature Reviews. Genetics* 13 (1): 14–20.
Dickinson, Mary E., Ann M. Flenniken, Xiao Ji, Lydia Teboul, Michael D. Wong, Jacqueline K. White, Terrence F. Meehan, et al. 2016. "High-Throughput Discovery of Novel Developmental Phenotypes." *Nature* 537 (7621): 508–14.
Duncan, B. K., and J. H. Miller. 1980. "Mutagenic Deamination of Cytosine Residues in DNA." *Nature* 287 (5782): 560–61.
Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems* 3 (1): 95–98.
Duret, Laurent, and Peter F. Arndt. 2008. "The Impact of Recombination on Nucleotide Substitutions in the Human Genome." *PLoS Genetics* 4 (5): e1000071.
Ebersberger, Ingo, Dirk Metzler, Carsten Schwarz, and Svante Pääbo. 2002. "Genomewide Comparison of DNA Sequences between Humans and Chimpanzees." *American Journal of Human Genetics* 70 (6): 1490–97.
Edwards, Stacey L., Jonathan Beesley, Juliet D. French, and Alison M. Dunning. 2013. "Beyond GWASs: Illuminating the Dark Road from Association to Function." *American Journal of Human Genetics* 93 (5): 779–97.
Ellinghaus, David, Luke Jostins, Sarah L. Spain, Adrian Cortes, Jörn Bethune, Buhm Han, Yu Rang Park, et al. 2016. "Analysis of Five Chronic Inflammatory Diseases Identifies 27 New Associations and Highlights Disease-Specific Patterns at Shared Loci." *Nature Genetics* 48 (5): 510–18.
ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, et al. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." *Nature* 583 (7818): 699–710.
Eppig, Janan T., Judith A. Blake, Carol J. Bult, James A. Kadin, Joel E. Richardson, and Mouse Genome Database Group. 2015. "The Mouse Genome Database (MGD): Facilitating Mouse as a Model for Human Biology and Disease." *Nucleic Acids Research* 43 (Database issue): D726–36.
Esté, José A., and Amalio Telenti. 2007. "HIV Entry Inhibitors." *The Lancet* 370 (9581): 81–88.
Fadista, João, Nikolay Oskolkov, Ola Hansson, and Leif Groop. 2017. "LoFtool: A Gene Intolerance Score Based on Loss-of-Function Variants in 60 706 Individuals." *Bioinformatics* 33 (4): 471–74.
Fairfield, Heather, Anuj Srivastava, Guruprasad Ananda, Rangjiao Liu, Martin Kircher, Anuradha Lakshminarayana, Belinda S. Harris, et al. 2015. "Exome Sequencing Reveals Pathogenic Mutations in 91 Strains of Mice with Mendelian Disorders." *Genome Research* 25 (7): 948–57.
Findlay, Gregory M., Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. 2018. "Accurate Classification of BRCA1 Variants with Saturation Genome Editing." *Nature* 562 (7726): 217–22.
Fish, Alexandra, Ling Chen, and John A. Capra. 2017. "Gene Regulatory Enhancers with Evolutionarily Conserved Activity Are More Pleiotropic than Those with Species-Specific Activity." *Genome Biology and Evolution* 9 (10): 2615–25.
Flint, Jonathan, and Eleazar Eskin. 2012. "Genome-Wide Association Studies in Mice." *Nature Reviews. Genetics* 13 (11): 807–17.
Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. "GENCODE

Reference Annotation for the Human and Mouse Genomes." *Nucleic Acids Research* 47 (D1): D766–73.
Freire-Pritchett, Paula, Stefan Schoenfelder, Csilla Várnai, Steven W. Wingett, Jonathan Cairns, Amanda J. Collier, Raquel García-Vílchez, et al. 2017. "Global Reorganisation of Cis-Regulatory Units upon Lineage Commitment of Human Embryonic Stem Cells." *eLife* 6 (March). https://doi.org/10.7554/eLife.21926.
Friedberg, Errol C. 2003. "DNA Damage and Repair." *Nature* 421 (6921): 436–40.
Fuller, Zachary L., Jeremy J. Berg, Hakhamanesh Mostafavi, Guy Sella, and Molly Przeworski. 2019. "Measuring Intolerance to Mutation in Human Genetics." *Nature Genetics* 51 (5): 772–76.
Gallagher, Michael D., and Alice S. Chen-Plotkin. 2018. "The Post-GWAS Era: From Association to Function." *American Journal of Human Genetics* 102 (5): 717–30.
Galtier, N., F. Bonhomme, C. Moulia, K. Belkhir, P. Caminade, E. Desmarais, J. J. Duquesne, A. Orth, B. Dod, and P. Boursot. 2004. "Mouse Biodiversity in the Genomic Era." *Cytogenetic and Genome Research* 105 (2-4): 385–94.
Gardner, Eugene J., Matthew D. C. Neville, Kaitlin E. Samocha, Kieron Barclay, Martin Kolk, Mari E. K. Niemi, George Kirov, Hilary C. Martin, and Matthew E. Hurles. 2022. "Reduced Reproductive Success Is Associated with Selective Constraint on Human Genes." *Nature*, March, 1–6.
Goios, Ana, Luísa Pereira, Molly Bogue, Vincent Macaulay, and António Amorim. 2007. "mtDNA Phylogeny and Evolution of Laboratory Mouse Strains." *Genome Research* 17 (3): 293–98.
Gojobori, T., W. H. Li, and D. Graur. 1982. "Patterns of Nucleotide Substitution in Pseudogenes and Functional Genes." *Journal of Molecular Evolution* 18 (5): 360–69.
Goldman, N., and Z. Yang. 1994. "A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences." *Molecular Biology and Evolution* 11 (5): 725–36.
Gonzales, Natalia M., Jungkyun Seo, Ana I. Hernandez Cordero, Celine L. St Pierre, Jennifer S. Gregory, Margaret G. Distler, Mark Abney, Stefan Canzar, Arimantas Lionikas, and Abraham A. Palmer. 2018. "Genome Wide Association Analysis in a Mouse Advanced Intercross Line." *Nature Communications* 9 (1): 5162.
Gonzalez-Perez, Abel, Radhakrishnan Sabarinathan, and Nuria Lopez-Bigas. 2019. "Local Determinants of the Mutational Landscape of the Human Genome." *Cell* 177 (1): 101–14.
Gorkin, David U., Iros Barozzi, Yuan Zhao, Yanxiao Zhang, Hui Huang, Ah Young Lee, Bin Li, et al. 2020. "An Atlas of Dynamic Chromatin Landscapes in Mouse Fetal Development." *Nature* 583 (7818): 744–51.
Gorlov, Ivan P., Marek Kimmel, and Christopher I. Amos. 2006. "Strength of the Purifying Selection against Different Categories of the Point Mutations in the Coding Regions of the Human Genome." *Human Molecular Genetics* 15 (7): 1143–50.
Gravel, Simon. 2016. "When Is Selection Effective?" *Genetics* 203 (1): 451–62.
Grossman, Sharon R., Jesse Engreitz, John P. Ray, Tung H. Nguyen, Nir Hacohen, and Eric S. Lander. 2018. "Positional Specificity of Different Transcription Factor Classes within Enhancers." *Proceedings of the National Academy of Sciences of the United States of America* 115 (30): E7222–30.
Guo, Xingyi, Yu Wang, Peter D. Keightley, and Longjiang Fan. 2007. "Patterns of Selective Constraints in Noncoding DNA of Rice." *BMC Evolutionary Biology* 7 (November): 208.

Halligan, Daniel L., Athanasios Kousathanas, Rob W. Ness, Bettina Harr, Lél Eöry, Thomas M. Keane, David J. Adams, and Peter D. Keightley. 2013. "Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents." *PLoS Genetics* 9 (12): e1003995.
Hall, J. M., M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey, and M. C. King. 1990. "Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21." *Science* 250 (4988): 1684–89.
Hanawalt, Philip C., and Graciela Spivak. 2008. "Transcription-Coupled DNA Repair: Two Decades of Progress and Surprises." *Nature Reviews. Molecular Cell Biology* 9 (12): 958–70.
Hardison, Ross C., Krishna M. Roskin, Shan Yang, Mark Diekhans, W. James Kent, Ryan Weber, Laura Elnitski, et al. 2003. "Covariation in Frequencies of Substitution, Deletion, Transposition, and Recombination during Eutherian Evolution." *Genome Research* 13 (1): 13–26.
Harr, Bettina, Emre Karakoc, Rafik Neme, Meike Teschke, Christine Pfeifle, Željka Pezer, Hiba Babiker, et al. 2016. "Genomic Resources for Wild Populations of the House Mouse, Mus Musculus and Its Close Relative Mus Spretus." *Scientific Data*.
Harris, Kelley. 2015. "Evidence for Recent, Population-Specific Evolution of the Human Mutation Rate." *Proceedings of the National Academy of Sciences of the United States of America* 112 (11): 3439–44.
Harris, Kelley, and Jonathan K. Pritchard. 2017. "Rapid Evolution of the Human Mutation Spectrum." *eLife* 6 (April). https://doi.org/10.7554/eLife.24284.
Harris, R. S. 2007. "Improved Pairwise Alignment of Genomic DNA." Ph.D. Thesis, The Pennsylvania State University.
Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project." *Genome Research* 22 (9): 1760–74.
Havrilla, James M., Brent S. Pedersen, Ryan M. Layer, and Aaron R. Quinlan. 2018. "A Map of Constrained Coding Regions in the Human Genome." *Nature Genetics*, December. https://doi.org/10.1038/s41588-018-0294-6.
Helleday, Thomas, Eva Petermann, Cecilia Lundin, Ben Hodgson, and Ricky A. Sharma. 2008. "DNA Repair Pathways as Targets for Cancer Therapy." *Nature Reviews. Cancer* 8 (3): 193–204.
Hellmann, Ines, Kay Prüfer, Hongkai Ji, Michael C. Zody, Svante Pääbo, and Susan E. Ptak. 2005. "Why Do Human Diversity Levels Vary at a Megabase Scale?" *Genome Research* 15 (9): 1222–31.
Henn, Brenna M., Laura R. Botigué, Carlos D. Bustamante, Andrew G. Clark, and Simon Gravel. 2015. "Estimating the Mutation Load in Human Genomes." *Nature Reviews. Genetics* 16 (6): 333–43.
Hernandez, Ryan D., Joanna L. Kelley, Eyal Elyashiv, S. Cord Melton, Adam Auton, Gilean McVean, 1000 Genomes Project, Guy Sella, and Molly Przeworski. 2011. "Classic Selective Sweeps Were Rare in Recent Human Evolution." *Science* 331 (6019): 920–24.
Herrero, Javier, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J. Vilella, et al. 2016. "Ensembl Comparative Genomics Resources." *Database: The Journal of Biological Databases and Curation* 2016 (February). https://doi.org/10.1093/database/bav096.
He, Yupeng, Manoj Hariharan, David U. Gorkin, Diane E. Dickel, Chongyuan Luo, Rosa G. Castanon, Joseph R. Nery, et al. 2020. "Spatiotemporal DNA

Methylome Dynamics of the Developing Mouse Fetus." *Nature* 583 (7818): 752–59.
Hickman, Mark, and John Cairns. 2003. "The Centenary of the One-Gene One-Enzyme Hypothesis." *Genetics* 163 (3): 839–41.
Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, et al. 2006. "The UCSC Genome Browser Database: Update 2006." *Nucleic Acids Research* 34 (Database issue): D590–98.
Hodgkinson, Alan, Ferran Casals, Youssef Idaghdour, Jean-Christophe Grenier, Ryan D. Hernandez, and Philip Awadalla. 2013. "Selective Constraint, Background Selection, and Mutation Accumulation Variability within and between Human Populations." *BMC Genomics* 14 (July): 495.
Hodgkinson, Alan, and Adam Eyre-Walker. 2011. "Variation in the Mutation Rate across Mammalian Genomes." *Nature Reviews. Genetics* 12 (11): 756–66.
Hook, Paul W., and Andrew S. McCallion. 2020. "Leveraging Mouse Chromatin Data for Heritability Enrichment Informs Common Disease Architecture and Reveals Cortical Layer Contributions to Schizophrenia." *Genome Research* 30 (4): 528–39.
Huber, Christian D., Bernard Y. Kim, and Kirk E. Lohmueller. 2020. "Population Genetic Models of GERP Scores Suggest Pervasive Turnover of Constrained Sites across Mammalian Evolution." *PLoS Genetics* 16 (5): e1008827.
Huh, Iksoo, Isabel Mendizabal, Taesung Park, and Soojin V. Yi. 2018. "Functional Conservation of Sequence Determinants at Rapidly Evolving Regulatory Regions across Mammals." *PLoS Computational Biology* 14 (10): e1006451.
Husemann, M., F. E. Zachos, R. J. Paxton, and J. C. Habel. 2016. "Effective Population Size in Ecology and Evolution." *Heredity* 117 (4): 191–92.
Hwang, Dick G., and Phil Green. 2004. "Bayesian Markov Chain Monte Carlo Sequence Analysis Reveals Varying Neutral Substitution Patterns in Mammalian Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 101 (39): 13994–1.
Iulio, Julia di, Istvan Bartha, Emily H. M. Wong, Hung-Chun Yu, Victor Lavrenko, Dongchan Yang, Inkyung Jung, et al. 2018. "The Human Noncoding Genome Defined by Genetic Diversity." *Nature Genetics*, February. https://doi.org/10.1038/s41588-018-0062-7.
Jensen, Jeffrey D. 2014. "On the Unfounded Enthusiasm for Soft Selective Sweeps." *Nature Communications* 5 (October): 5281.
Justice, Monica J., and Paraminder Dhillon. 2016. "Using the Mouse to Model Human Disease: Increasing Validity and Reproducibility." *Disease Models & Mechanisms* 9 (2): 101–3.
Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581 (7809): 434–43.
———. n.d. "Variation across 141, 456 Human Exomes and Genomes Reveals the Spectrum of Loss-of- Function Intolerance across Human Protein-Coding Genes." https://doi.org/10.1101/531210.
Karp, Natasha A., Jeremy Mason, Arthur L. Beaudet, Yoav Benjamini, Lynette Bower, Robert E. Braun, Steve D. M. Brown, et al. 2017. "Prevalence of Sexual Dimorphism in Mammalian Phenotypic Traits." *Nature Communications* 8 (June): 15475.
Keane, Thomas M., Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, et al. 2011. "Mouse Genomic Variation and Its

Effect on Phenotypes and Gene Regulation." *Nature* 477 (7364): 289–94.
Kimura, M. 1977. "Preponderance of Synonymous Changes as Evidence for the Neutral Theory of Molecular Evolution." *Nature* 267 (5608): 275–76.
Kim, Yong Hoon, Sajid A. Marhon, Yuxiang Zhang, David J. Steger, Kyoung-Jae Won, and Mitchell A. Lazar. 2018. "Rev-Erbα Dynamically Modulates Chromatin Looping to Control Circadian Gene Transcription." *Science* 359 (6381): 1274–77.
Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15.
Knight, R. D., S. J. Freeland, and L. F. Landweber. 2001. "Rewiring the Keyboard: Evolvability of the Genetic Code." *Nature Reviews. Genetics* 2 (1): 49–58.
Koh, Gene, Andrea Degasperi, Xueqing Zou, Sophie Momen, and Serena Nik-Zainal. 2021. "Mutational Signatures: Emerging Concepts, Caveats and Clinical Applications." *Nature Reviews. Cancer* 21 (10): 619–37.
Kostka, Dennis, Melissa J. Hubisz, Adam Siepel, and Katherine S. Pollard. 2012. "The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome." *Molecular Biology and Evolution* 29 (3): 1047–57.
Kryazhimskiy, Sergey, and Joshua B. Plotkin. 2008. "The Population Genetics of dN/dS." *PLoS Genetics* 4 (12): e1000304.
Kumar, Sudhir, and Sankar Subramanian. 2002. "Mutation Rates in Mammalian Genomes." *Proceedings of the National Academy of Sciences of the United States of America* 99 (2): 803–8.
Laber, Samantha, and Roger D. Cox. 2017. "Mouse Models of Human GWAS Hits for Obesity and Diabetes in the Post Genomic Era: Time for Reevaluation." *Frontiers in Endocrinology* 8 (February): 11.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.
Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. "ClinVar: Improving Access to Variant Interpretations and Supporting Evidence." *Nucleic Acids Research* 46 (D1): D1062–67.
Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285–91.
Lercher, Martin J., and Laurence D. Hurst. 2002. "Human SNP Variability and Mutation Rate Are Higher in Regions of High Recombination." *Trends in Genetics: TIG* 18 (7): 337–40.
Lercher, M. J., E. J. Williams, and L. D. Hurst. 2001. "Local Similarity in Evolutionary Rates Extends over Whole Chromosomes in Human-Rodent and Mouse-Rat Comparisons: Implications for Understanding the Mechanistic Basis of the Male Mutation Bias." *Molecular Biology and Evolution* 18 (11): 2032–39.
Leung, Danny, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, et al. 2015. "Integrative Analysis of Haplotype-Resolved Epigenomes across Human Tissues." *Nature* 518 (7539): 350–54.
Liao, Ben-Yang, and Jianzhi Zhang. 2006. "Evolutionary Conservation of Expression Profiles between Human and Mouse Orthologous Genes." *Molecular Biology*

and Evolution 23 (3): 530–40.
Li, Feng, Dale O. Cowley, Debra Banner, Eric Holle, Liguo Zhang, and Lishan Su. 2014. "Efficient Genetic Manipulation of the NOD-Rag1-/-IL2RgammaC-Null Mouse by Combining in Vitro Fertilization and CRISPR/Cas9 Technology." *Scientific Reports* 4 (June): 5290.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.
Lilue, Jingtao, Anthony G. Doran, Ian T. Fiddes, Monica Abrudan, Joel Armstrong, Ruth Bennett, William Chow, et al. 2018. "Sixteen Diverse Laboratory Mouse Reference Genomes Define Strain-Specific Haplotypes and Novel Functional Loci." *Nature Genetics* 50 (11): 1574–83.
Li, Mingkun, and Su-Shing Chen. 2011. "The Tendency to Recreate Ancestral CG Dinucleotides in the Human Genome." *BMC Evolutionary Biology* 11 (January): 3.
Lindahl, T., and R. D. Wood. 1999. "Quality Control by DNA Repair." *Science* 286 (5446): 1897–1905.
Lindsay, Sarah J., Raheleh Rahbari, Joanna Kaplanis, Thomas Keane, and Matthew E. Hurles. 2019. "Similarities and Differences in Patterns of Germline Mutation between Mice and Humans." *Nature Communications* 10 (1): 4053.
Lin, Shin, Yiing Lin, Joseph R. Nery, Mark A. Urich, Alessandra Breschi, Carrie A. Davis, Alexander Dobin, et al. 2014. "Comparison of the Transcriptional Landscapes between Human and Mouse Tissues." *Proceedings of the National Academy of Sciences of the United States of America* 111 (48): 17224–29.
Liston, Adrian, Stephanie Humblet-Baron, Darragh Duffy, and An Goris. 2021. "Human Immune Diversity: From Evolution to Modernity." *Nature Immunology* 22 (12): 1479–89.
Lloyd, K. C. Kent, David J. Adams, Gareth Baynam, Arthur L. Beaudet, Fatima Bosch, Kym M. Boycott, Robert E. Braun, et al. 2020. "The Deep Genome Project." *Genome Biology* 21 (1): 18.
Ludwig, Michael Z. 2002. "Functional Evolution of Noncoding DNA." *Current Opinion in Genetics & Development* 12 (6): 634–39.
Lundrigan, Barbara L., Sharon A. Jansa, and Priscilla K. Tucker. 2002. "Phylogenetic Relationships in the Genus Mus, Based on Paternally, Maternally, and Biparentally Inherited Characters." *Systematic Biology* 51 (3): 410–31.
Lynch, Michael. 2010a. "Rate, Molecular Spectrum, and Consequences of Human Mutation." *Proceedings of the National Academy of Sciences of the United States of America* 107 (3): 961–68.
———. 2010b. "Evolution of the Mutation Rate." *Trends in Genetics: TIG* 26 (8): 345–52.
———. 2011. "The Lower Bound to the Evolution of Mutation Rates." *Genome Biology and Evolution* 3 (August): 1107–18.
Makova, Kateryna D., and Ross C. Hardison. 2015. "The Effects of Chromatin Organization on Variation in Mutation Rates in the Genome." *Nature Reviews. Genetics* 16 (4): 213–23.
Makrythanasis, P., and S. E. Antonarakis. 2013. "Pathogenic Variants in Non-Protein-Coding Sequences." *Clinical Genetics* 84 (5): 422–28.
Matassi, G., P. M. Sharp, and C. Gautier. 1999. "Chromosomal Location Effects on Gene Sequence Evolution in Mammals." *Current Biology: CB* 9 (15): 786–91.
McDonald, Michael J., Wei-Chi Wang, Hsien-Da Huang, and Jun-Yi Leu. 2011.

"Clusters of Nucleotide Substitutions and Insertion/deletion Mutations Are Associated with Repeat Sequences." *PLoS Biology* 9 (6): e1000622.
McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1): 122.
Meehan, Terrence F., Nathalie Conte, David B. West, Julius O. Jacobsen, Jeremy Mason, Jonathan Warren, Chao-Kung Chen, et al. 2017. "Disease Model Discovery from 3,328 Gene Knockouts by The International Mouse Phenotyping Consortium." *Nature Genetics* 49 (8): 1231–38.
Metzker, Michael L. 2010. "Sequencing Technologies - the next Generation." *Nature Reviews. Genetics* 11 (1): 31–46.
Minikel, Eric Vallabh, Konrad J. Karczewski, Hilary C. Martin, Beryl B. Cummings, Nicola Whiffin, Jessica Alföldi, Richard C. Trembath, et al. 2019. "Evaluating Potential Drug Targets through Human Loss-of-Function Genetic Variation." *bioRxiv*. https://doi.org/10.1101/530881.
Minikel, Eric Vallabh, Konrad J. Karczewski, Hilary C. Martin, Beryl B. Cummings, Nicola Whiffin, Daniel Rhodes, Jessica Alföldi, et al. 2020. "Evaluating Drug Targets through Human Loss-of-Function Genetic Variation." *Nature* 581 (7809): 459–64.
Moll, C. 2005. San Francisco: Wikimedia Foundation. https://upload.wikimedia.org/wikipedia/commons/thumb/b/b2/Survivorship-bias.svg/1600px-Survivorship-bias.svg.png.
Monaco, Gianni, Sipko van Dam, João Luis Casal Novo Ribeiro, Anis Larbi, and João Pedro de Magalhães. 2015. "A Comparison of Human and Mouse Gene Co-Expression Networks Reveals Conservation and Divergence at the Tissue, Pathway and Disease Levels." *BMC Evolutionary Biology* 15 (November): 259.
Monroe, J. Grey, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, Moises Exposito-Alonso, Marie Klein, et al. 2022. "Mutation Bias Reflects Natural Selection in Arabidopsis Thaliana." *Nature*, January, 1–5.
Mouse ENCODE Consortium, John A. Stamatoyannopoulos, Michael Snyder, Ross Hardison, Bing Ren, Thomas Gingeras, David M. Gilbert, et al. 2012. "An Encyclopedia of Mouse DNA Elements (Mouse ENCODE)." *Genome Biology* 13 (8): 418.
Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420 (6915): 520–62.
Muñoz-Fontela, César, William E. Dowling, Simon G. P. Funnell, Pierre-S Gsell, A. Ximena Riveros-Balta, Randy A. Albrecht, Hanne Andersen, et al. 2020. "Animal Models for COVID-19." *Nature* 586 (7830): 509–15.
Muse, S. V., and B. S. Gaut. 1994. "A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome." *Molecular Biology and Evolution* 11 (5): 715–24.
Nadeau, Joseph H., and Johan Auwerx. 2019. "The Virtuous Cycle of Human Genetics and Mouse Models in Drug Discovery." *Nature Reviews. Drug Discovery* 18 (4): 255–72.
Narasimhan, Vagheesh M., Karen A. Hunt, Dan Mason, Christopher L. Baker, Konrad J. Karczewski, Michael R. Barnes, Anthony H. Barnett, et al. 2016. "Health and Population Effects of Rare Gene Knockouts in Adult Humans with Related Parents." *Science* 352 (6284): 474–77.

Nitta, Kazuhiro R., Arttu Jolma, Yimeng Yin, Ekaterina Morgunova, Teemu Kivioja, Junaid Akhtar, Korneel Hens, et al. 2015. “Conservation of Transcription Factor Binding Specificities across 600 Million Years of Bilateria Evolution.” *eLife* 4 (March). https://doi.org/10.7554/eLife.04837.
Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. “The Complete Sequence of a Human Genome.” *Science* 376 (6588): 44–53.
Orth, Annie, Khalid Belkhir, Janice Britton-Davidian, Pierre Boursot, Touria Benazzou, and François Bonhomme. 2002. “[Natural hybridization between 2 sympatric species of mice, Mus musculus domesticus L. and Mus spretus Lataste].” *Comptes rendus biologies* 325 (2): 89–97.
Payseur, Bret A., and Peicheng Jing. 2021. “Genomic Targets of Positive Selection in Giant Mice from Gough Island.” *Molecular Biology and Evolution* 38 (3): 911–26.
Perlman, Robert L. 2016. “Mouse Models of Human Disease: An Evolutionary Perspective.” *Evolution, Medicine, and Public Health* 2016 (1): 170–76.
Pervouchine, Dmitri D., Sarah Djebali, Alessandra Breschi, Carrie A. Davis, Pablo Prieto Barja, Alex Dobin, Andrea Tanzer, et al. 2015. “Enhanced Transcriptome Maps from Multiple Mouse Tissues Reveal Evolutionary Constraint in Gene Expression.” *Nature Communications* 6 (January): 5903.
Peters, Luanne L., Raymond F. Robledo, Carol J. Bult, Gary A. Churchill, Beverly J. Paigen, and Karen L. Svenson. 2007. “The Mouse as a Model for Human Biology: A Resource Guide for Complex Trait Analysis.” *Nature Reviews. Genetics* 8 (1): 58–69.
Petrovski, Slavé, Quanli Wang, Erin L. Heinzen, Andrew S. Allen, and David B. Goldstein. 2013. “Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes.” *PLoS Genetics* 9 (8): e1003709.
Petruska, J., and M. F. Goodman. 1985. “Influence of Neighboring Bases on DNA Polymerase Insertion and Proofreading Fidelity.” *The Journal of Biological Chemistry* 260 (12): 7533–39.
Phifer-Rixey, Megan, François Bonhomme, Pierre Boursot, Gary A. Churchill, Jaroslav Piálek, Priscilla K. Tucker, and Michael W. Nachman. 2012. “Adaptive Evolution and Effective Population Size in Wild House Mice.” *Molecular Biology and Evolution* 29 (10): 2949–55.
Phifer-Rixey, Megan, and Michael W. Nachman. 2015. “Insights into Mammalian Biology from the Wild House Mouse Mus Musculus.” *eLife* 4 (April). https://doi.org/10.7554/eLife.05959.
Pink, Catherine J., and Laurence D. Hurst. 2010. “Timing of Replication Is a Determinant of Neutral Substitution Rates but Does Not Explain Slow Y Chromosome Evolution in Rodents.” *Molecular Biology and Evolution* 27 (5): 1077–86.
Plenge, Robert M., Edward M. Scolnick, and David Altshuler. 2013. “Validating Therapeutic Targets through Human Genetics.” *Nature Reviews. Drug Discovery* 12 (8): 581–94.
Poetsch, Anna R., Simon J. Boulton, and Nicholas M. Luscombe. 2018. “Genomic Landscape of Oxidative DNA Damage and Repair Reveals Regioselective Protection from Mutagenesis.” *Genome Biology* 19 (1): 215.
Polak, Paz, Rosa Karlić, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael Lawrence, Alex Reynolds, et al. 2015. “Cell-of-Origin Chromatin Organization Shapes the Mutational Landscape of Cancer.” *Nature* 518 (7539): 360–64.

Popp, Christian, Wendy Dean, Suhua Feng, Shawn J. Cokus, Simon Andrews, Matteo Pellegrini, Steven E. Jacobsen, and Wolf Reik. 2010. “Genome-Wide Erasure of DNA Methylation in Mouse Primordial Germ Cells Is Affected by AID Deficiency.” *Nature* 463 (7284): 1101–5.
Pouyet, Fanny, Simon Aeschbacher, Alexandre Thiéry, and Laurent Excoffier. 2018. “Background Selection and Biased Gene Conversion Affect More than 95% of the Human Genome and Bias Demographic Inferences.” *eLife* 7 (August). https://doi.org/10.7554/eLife.36317.
Powell, George, Helen S. Long, Thomas M. M. Versluys, Ann-Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2021. “Mechanisms of DNA Repair Have Evolved to Minimise the Probability of Nonsense Mutations.” *bioRxiv*. https://doi.org/10.1101/2021.06.22.449367.
Powell, George, Helen Long, Louisa Zolkiewski, Rebecca Dumbell, Ann-Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2022. “Modelling the Genetic Aetiology of Complex Disease: Human-Mouse Conservation of Noncoding Features and Disease-Associated Loci.” *Biology Letters* 18 (3): 20210630.
Powell, George, Michelle Simon, Sara Pulit, Ann-Marie Mallon, and Cecilia M. Lindgren. 2019. “Tolerance of Nonsynonymous Variation Is Closely Correlated between Human and Mouse Orthologues.” https://doi.org/10.1101/657981.
Prieto, M. L., E. Ryu, G. D. Jenkins, A. Batzler, M. M. Nassan, A. B. Cuellar-Barboza, J. Pathak, S. L. McElroy, M. A. Frye, and J. M. Biernacka. 2016. “Leveraging Electronic Health Records to Study Pleiotropic Effects on Bipolar Disorder and Medical Comorbidities.” *Translational Psychiatry* 6 (August): e870.
Pritchard, Jonathan K., and Anna Di Rienzo. 2010. “Adaptation - Not by Sweeps Alone.” *Nature Reviews. Genetics* 11 (10): 665–67.
Pritchard, Jonathan K., Joseph K. Pickrell, and Graham Coop. 2010. “The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation.” *Current Biology: CB* 20 (4): R208–15.
Rackham, Owen J. L., Hashem A. Shihab, Michael R. Johnson, and Enrico Petretto. 2015. “EvoTol: A Protein-Sequence Based Evolutionary Intolerance Framework for Disease-Gene Prioritization.” *Nucleic Acids Research* 43 (5): e33.
Ran, F. Ann, Patrick D. Hsu, Jason Wright, Vineeta Agarwala, David A. Scott, and Feng Zhang. 2013. “Genome Engineering Using the CRISPR-Cas9 System.” *Nature Protocols* 8 (11): 2281–2308.
R Core Team. 2017. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Rech, Gabriel E., Santiago Radío, Sara Guirao-Rico, Laura Aguilera, Vivien Horvath, Llewellyn Green, Hannah Lindstadt, Véronique Jamilloux, Hadi Quesneville, and Josefa González. 2022. “Population-Scale Long-Read Sequencing Uncovers Transposable Elements Associated with Gene Expression Variation and Adaptive Signatures in Drosophila.” *Nature Communications* 13 (1): 1–16.
Rettelbach, Agnes, Alexander Nater, and Hans Ellegren. 2019. “How Linked Selection Shapes the Diversity Landscape in Ficedula Flycatchers.” *Genetics* 212 (1): 277–85.
Rosenthal, Nadia, and Steve Brown. 2007. “The Mouse Ascending: Perspectives for Human-Disease Models.” *Nature Cell Biology* 9 (9): 993–99.
Rozman, Jan, Birgit Rathkolb, Manuela A. Oestereicher, Christine Schütt, Aakash Chavan Ravindranath, Stefanie Leuchtenberger, Sapna Sharma, et al. 2018. “Identification of Genetic Elements in Metabolism by High-Throughput Mouse

Phenotyping." *Nature Communications* 9 (1): 288.
Salcedo, Tovah, Armando Geraldes, and Michael W. Nachman. 2007. "Nucleotide Variation in Wild and Inbred Mice." *Genetics* 177 (4): 2277–91.
Samocha, Kaitlin E., Jack A. Kosmicki, Konrad J. Karczewski, Anne H. O'Donnell-Luria, Emma Pierce-Hoffman, Daniel G. MacArthur, Benjamin M. Neale, and Mark J. Daly. 2017. "Regional Missense Constraint Improves Variant Deleteriousness Prediction." *bioRxiv*. https://doi.org/10.1101/148353.
Samocha, Kaitlin E., Elise B. Robinson, Stephan J. Sanders, Christine Stevens, Aniko Sabo, Lauren M. McGrath, Jack A. Kosmicki, et al. 2014. "A Framework for the Interpretation of de Novo Mutation in Human Disease." *Nature Genetics* 46 (9): 944–50.
Schroeder, Jeremy W., William G. Hirst, Gabriella A. Szewczyk, and Lyle A. Simmons. 2016. "The Effect of Local Sequence Context on Mutational Bias of Genes Encoded on the Leading and Lagging Strands." *Current Biology: CB* 26 (5): 692–97.
Schuster-Böckler, Benjamin, and Ben Lehner. 2012. "Chromatin Organization Is a Major Influence on Regional Mutation Rates in Human Cancer Cells." *Nature* 488 (7412): 504–7.
Schwarz, Jana Marie, David N. Cooper, Markus Schuelke, and Dominik Seelow. 2014. "MutationTaster2: Mutation Prediction for the Deep-Sequencing Age." *Nature Methods* 11 (4): 361–62.
Scully, Ralph, Arvind Panday, Rajula Elango, and Nicholas A. Willis. 2019. "DNA Double-Strand Break Repair-Pathway Choice in Somatic Mammalian Cells." *Nature Reviews. Molecular Cell Biology* 20 (11): 698–714.
Ségurel, Laure, Minyoung J. Wyman, and Molly Przeworski. 2014. "Determinants of Mutation Rate Variation in the Human Germline." *Annual Review of Genomics and Human Genetics* 15 (June): 47–70.
Sethi, Siddharth, Ilya E. Vorontsov, Ivan V. Kulakovskiy, Simon Greenaway, John Williams, Vsevolod J. Makeev, Steve D. M. Brown, Michelle M. Simon, and Ann-Marie Mallon. 2020. "A Holistic View of Mouse Enhancer Architectures Reveals Analogous Pleiotropic Effects and Correlation with Human Disease." *BMC Genomics* 21 (1): 754.
Shen, J. C., W. M. Rideout 3rd, and P. A. Jones. 1994. "The Rate of Hydrolytic Deamination of 5-Methylcytosine in Double-Stranded DNA." *Nucleic Acids Research* 22 (6): 972–76.
Shen, Yin, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, et al. 2012. "A Map of the Cis-Regulatory Sequences in the Mouse Genome." *Nature* 488 (7409): 116–20.
Siepel, Adam, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, et al. 2005. "Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes." *Genome Research* 15 (8): 1034–50.
Smemo, Scott, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, et al. 2014. "Obesity-Associated Variants within FTO Form Long-Range Functional Connections with IRX3." *Nature* 507 (7492): 371–75.
Smith, Cynthia L., and Janan T. Eppig. 2009. "The Mammalian Phenotype Ontology: Enabling Robust Annotation and Comparative Analysis." *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 1 (3): 390–99.
———. 2012. "The Mammalian Phenotype Ontology as a Unifying Standard for Experimental and High-Throughput Phenotyping Data." *Mammalian Genome:*

Official Journal of the International Mammalian Genome Society 23 (9-10): 653–68.
Smith, J. M., and J. Haigh. 1974. "The Hitch-Hiking Effect of a Favourable Gene." *Genetical Research* 23 (1): 23–35.
Springer, Mark S., and William J. Murphy. 2007. "Mammalian Evolution and Biomedicine: New Views from Phylogeny." *Biological Reviews of the Cambridge Philosophical Society* 82 (3): 375–92.
Stadler, Michael B., Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, et al. 2011. "DNA-Binding Factors Shape the Mouse Methylome at Distal Regulatory Regions." *Nature* 480 (7378): 490–95.
Stamatoyannopoulos, John A., Ivan Adzhubei, Robert E. Thurman, Gregory V. Kryukov, Sergei M. Mirkin, and Shamil R. Sunyaev. 2009. "Human Mutation Rate Associated with DNA Replication Timing." *Nature Genetics* 41 (4): 393–95.
Stephan, Wolfgang. 2010. "Genetic Hitchhiking versus Background Selection: The Controversy and Its Implications." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365 (1544): 1245–53.
———. 2019. "Selective Sweeps." *Genetics* 211 (1): 5–13.
Subramanian, Sankar. 2011. "High Proportions of Deleterious Polymorphisms in Constrained Human Genes." *Molecular Biology and Evolution* 28 (1): 49–52.
Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *PLoS Medicine* 12 (3): e1001779.
Sung, Way, Matthew S. Ackerman, Samuel F. Miller, Thomas G. Doak, and Michael Lynch. 2012. "Drift-Barrier Hypothesis and Mutation-Rate Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 109 (45): 18488–92.
Supek, Fran, and Ben Lehner. 2015. "Differential DNA Mismatch Repair Underlies Mutation Rate Variation across the Human Genome." *Nature* 521 (7550): 81–84.
Suzuki, Hitoshi, Tomofumi Shimada, Mie Terashima, Kimiyuki Tsuchiya, and Ken Aplin. 2004. "Temporal, Spatial, and Ecological Modes of Evolution of Eurasian Mus Based on Mitochondrial and Nuclear Gene Sequences." *Molecular Phylogenetics and Evolution* 33 (3): 626–46.
Tajima, F. 1989. "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." *Genetics* 123 (3): 585–95.
Tian, Dacheng, Qiang Wang, Pengfei Zhang, Hitoshi Araki, Sihai Yang, Martin Kreitman, Thomas Nagylaki, Richard Hudson, Joy Bergelson, and Jian-Qun Chen. 2008. "Single-Nucleotide Mutation Rate Increases close to Insertions/deletions in Eukaryotes." *Nature* 455 (7209): 105–8.
Uchimura, Arikuni, Mayumi Higuchi, Yohei Minakuchi, Mizuki Ohno, Atsushi Toyoda, Asao Fujiyama, Ikuo Miura, Shigeharu Wakana, Jo Nishino, and Takeshi Yagi. 2015. "Germline Mutation Rates and the Long-Term Phenotypic Effects of Mutation Accumulation in Wild-Type Laboratory Mice and Mutator Mice." *Genome Research* 25 (8): 1125–34.
Vierstra, Jeff, Eric Rynes, Richard Sandstrom, Miaohua Zhang, Theresa Canfield, R. Scott Hansen, Sandra Stehling-Sun, et al. 2014. "Mouse Regulatory DNA Landscapes Reveal Global Principles of Cis-Regulatory Evolution." *Science* 346 (6212): 1007–12.
Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, et al. 2015. "Enhancer Evolution across 20

Mammalian Species." *Cell* 160 (3): 554–66.
Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *American Journal of Human Genetics* 101 (1): 5–22.
Wald, A. 1980. "A Reprint of 'A Method of Estimating Plane Vulnerability Based on Damage of Survivors.'"
Wang, Zhi, Ben-Yang Liao, and Jianzhi Zhang. 2010. "Genomic Patterns of Pleiotropy and the Evolution of Complexity." *Proceedings of the National Academy of Sciences of the United States of America* 107 (42): 18034–39.
Watanabe, Kyoko, Sven Stringer, Oleksandr Frei, Maša Umićević Mirkov, Christiaan de Leeuw, Tinca J. C. Polderman, Sophie van der Sluis, Ole A. Andreassen, Benjamin M. Neale, and Danielle Posthuma. 2019. "A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits." *Nature Genetics* 51 (9): 1339–48.
Whiffin, Nicola, Konrad J. Karczewski, Xiaolei Zhang, Sonia Chothani, Miriam J. Smith, D. Gareth Evans, Angharad M. Roberts, et al. 2020. "Characterising the Loss-of-Function Impact of 5' Untranslated Region Variants in 15,708 Individuals." *Nature Communications* 11 (1): 2523.
Wyrick, John J., and Steven A. Roberts. 2015. "Genomic Approaches to DNA Repair and Mutagenesis." *DNA Repair* 36 (December): 146–55.
Yue, Feng, Yong Cheng, Alessandra Breschi, Jeff Vierstra, Weisheng Wu, Tyrone Ryba, Richard Sandstrom, et al. 2014. "A Comparative Encyclopedia of DNA Elements in the Mouse Genome." *Nature* 515 (7527): 355–64.
Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. "Ensembl 2018." *Nucleic Acids Research* 46 (D1): D754–61.
Zerbino, Daniel R., Steven P. Wilder, Nathan Johnson, Thomas Juettemann, and Paul R. Flicek. 2015. "The Ensembl Regulatory Build." *Genome Biology* 16 (March): 56.
Zhang, Feng, and James R. Lupski. 2015. "Non-Coding Genetic Variants in Human Disease." *Human Molecular Genetics* 24 (R1): R102–10.
Zhou, Jin, and Yik-Ying Teo. 2016. "Estimating Time to the Most Recent Common Ancestor (TMRCA): Comparison and Application of Eight Methods." *European Journal of Human Genetics: EJHG* 24 (8): 1195–1201.
Zhu, Fei, Remya R. Nair, Elizabeth M. C. Fisher, and Thomas J. Cunningham. 2019. "Humanising the Mouse Genome Piece by Piece." *Nature Communications* 10 (1): 1845.
Zitvogel, Laurence, Jonathan M. Pitt, Romain Daillère, Mark J. Smyth, and Guido Kroemer. 2016. "Mouse Models in Oncoimmunology." *Nature Reviews. Cancer* 16 (12): 759–73.
Zuberi, Aamir, and Cathleen Lutz. 2016. "Mouse Models for Drug Discovery. Can New Tools and Technology Improve Translational Power?" *ILAR Journal / National Research Council, Institute of Laboratory Animal Resources* 57 (2): 178–85.

Acknowledgements

I would like to thank my supervisors Professor Cecilia Lindgren, Dr Ann-Marie Mallon, and Dr Michelle Simon for expertly guiding me through each stage of this process. Your feedback and encouragement have been critical to my success, and you have made me a better scientist and independent thinker. I would like to give particular thanks to Michelle for her patience and day-to-day guidance. She has been a calming influence through the highs and lows.

I would also like to extend my appreciation to members of my lab groups at Harwell and the Big Data Institute, particularly Duncan, Saskia, Helen, and Chris, for reading and providing feedback on my thesis. It is by your critical reflections and sober comments that my thesis has been honed into a polished article.

Finally, I would like to thank my family and friends for the contributions you have made to my life throughout a challenging period. My immediate family have been loving and supportive as always, and my friends have kept me sane. A final thanks to Rhuenna – you have been a rock and always make me smile after a long day's work.

I acknowledge that parts of the following thesis chapters have previously been published: chapter 2 on BioRxix (Powell et al. 2019), chapter 3 on BioRxiv (Powell et al. 2021), and chapter 4 in Biology Letters (Powell et al. 2022). In chapter 4, Topologically Associated Domain (TAD) boundary annotations used in the analysis were calculated and provided by Helen Long, and she also provided the description of how this was achieved in the methods (highlighted with footnotes). I confirm that I produced all other published material with no more input/advice than would be considered appropriate or acceptable for a thesis.