Novel techniques for measuring the effect of neighbouring bases on mutation and their applications

Yicheng Zhu

Research School of Biology
The Australian National University

A thesis submitted for the degree of

Doctor of Philosophy

April, 2019

Declaration

The thesis contains no material which has been accepted for the award of any other degree. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due acknowledgement has been made.

Name: Yicheng Zhu

Signature: __________________

Date: 17/02/2020

Acknowledgements

I would like to express my sincere gratitude to Professor Gavin Huttley for his thorough and inspiring supervision. His expertise, advice and help has been extremely valuable. Without him, this project would never have been possible. His enthusiasm for science and broad knowledge has set an example for me and motivated me to pursue an academic career. I am certainly looking forward to opportunities to work with and learn from him in the future.

I am also grateful to the members of my advisory panel: Associate Professor Conrad Burden, Dr Teresa Neeman, Dr Benjamin Kaehler from ANU, and Assistant Professor Von Bing Yap from National University of Singapore. They have provided invaluable support and advice at various stages.

I would particularly like to acknowledge my collaborator, Dr Cheng Soon Ong from Data61, CSIRO for his valuable input to my second project: 'Machine learning techniques for classifying the mutagenic origins of point mutations'. It was a very rewarding and enjoyable collaboration.

I would like to thank all members in Professor Gavin Huttley's laboratory—Helmut Simon, Hua (Emily) Ying, Yurong Tang. Thank you also to all members of the ANU bioinformatic consultancy unit and my friends at the John Curtin School of Medical Research, Aaron Chuah, Bob Buckley, Cameron Jack, Hardip Patel and Stephanie Palmer, for providing a pleasurable environment in which to conduct my research.

I greatly appreciate my husband Yuehua Duan, my parents Guiyang Zhu and Lihua Lyu, and my good friend Åsa pérez-Bercoff for their encouragement, love and unwavering support throughout my research.

Finally, I would like to thank everyone from the John Curtin School of Medical Research and the Research School of Biology for providing such a friendly environment.

Abstract

Understanding factors influencing mutations can improve detection of novel mutations, the diagnostic signatures of disease-causing mutagens, and facilitate the development of more accurate models of genetic divergence. Hypermutability of CpG demonstrates the existence of mutation motifs, sequences of flanking bases that influence point mutation processes. These motifs can also be indicative of specific underlying mutation mechanisms. I developed novel log-linear models for identifying mutation motifs that allow further comparisons of these mutation motifs, and of the complete mutation spectra between samples. Mutation motifs are visualised using a sequence logo type method.

In this thesis, I applied the methods to examine each of the possible 12 point mutations in ≈13.6 million human germline mutations (inferred from single-nucleotide polymorphisms recorded in the Ensembl database) and ≈181,000 melanoma mutations from the COSMIC database.

My method recovered the well-known CpG effect, which a conventional motif detection method failed to do. I established that all point mutations have significant and distinct mutation motifs. While the major effects of flanking bases lie within 2bp of the mutated position, I refute previous reports that the effect magnitude decays monotonically with distance. Comparison between autosomes and X-chromosomes supported a reduced contribution from methylation-induced C→T mutation on the X-chromosome, consistent with a previous prediction.

In addition, analyses of malignant melanoma confirmed reported characteristic features of this cancer, such as strand asymmetry of mutation processes. Further, I found that neighbouring influences in malignant melanoma differ significantly from those affecting germline mutations. Interestingly, for C→T mutation, the CpG effect was no longer evident, and was largely substituted by different neighbouring mechanisms. Moreover, the observed neighbouring influence is able to reflect the chemical influences of mutagenic processes after exposure to ultraviolet light. Based on this observation, I hypothesised that information regarding the mechanistic origin of point mutations is present in surrounding DNA sequences, and sequence neighbourhood can be used to identify the mechanistic origin of particular mutations.

Machine learning classifiers were developed to assess the above hypothesis and discriminate between N-ethyl-N-nitrosourea (ENU)-induced and spontaneous point mutations in the mouse germline. ENU is a synthetic chemical employed in mutagenesis studies, introducing novel point mutations to genomes. My classification results reveal that a combination of *k*-mer size and representation of second-order interactions among nucleotides was able to improve classification performance compared to the naïve classifier approach.

In conclusion, this work demonstrates that neighbouring bases have a profound effect on the occurrence of mutations. The statistical methods reported in this research can be used to examine the role of flanking sequence on mutation processes from polymorphism data, which further enable identification of differences in the operation of mechanisms of mutation between genomic regions, cell types or species. In addition, the machine learning classification results have important implications for

modelling context-dependent effects on sequence evolution.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mutations are changes in DNA sequences. Depending on the consequence of a mutation process, a mutation can be an insertion, a deletion, or a point mutation. Insertions are additions of one or more nucleotide base pairs into a DNA sequence. Deletions refer to mutations where one or more bases are deleted during DNA replication. Point mutations are mutations where a base is replaced with a different base during DNA replication. Mutation is the primary source of genetic variation and a critical contributor to phenotypic diversity (Graur and Li, 2000; Nei, 2013; Peichel and Marques, 2017). Of particular interest to biologists is understanding mutation, and the development of genetic analyses to study factors that influence mutations. Advances in understanding factors that influence mutation tendencies can increase the sensitivity of mutation detection techniques (Gasser and Zhu, 1999; Kidess et al., 2015), identify diagnostic signatures of human disease-causing mutagens (Alexandrov et al., 2013; Nik-Zainal et al., 2012; Peltomaki and Vasen, 1997; Ying and Huttley, 2011) and facilitate development of more accurate models of genetic divergence (Harris, 2015; Huttley, 2004; Kaehler et al., 2017, 2015; Schluter, 2009). Investigations of mutagenesis have revealed that mutation is a complex process and can be induced by multiple factors (Cooke et al., 2003; Helleday et al., 2014). (as a side effect of normal cellular processes, man-made chemicals and/or radiation). There is also evidence for a sequence neighbourhood influence on a given point mutation (Cooper and Youssoufian, 1988; Nevarez et al., 2010; Zhang and Zhao, 2004; Zhao and Boerwinkle, 2002).

Intensive research into mutagenesis has revealed that mutations are affected by sequence neighbourhoods, and these influences are, in some specific instances, well understood in terms of their underlying mechanism. The most well-known examples of the neighbourhood effect is CpG hypermutability (Coulondre et al., 1978). Before this study, it was not clear how general this phenomenon was, including the extent to which neighbourhoods influence mutation and the size of major influential neighbourhoods. To solve these, in Chapter 2, I describe a comprehensive statistical model that facilitated discovery of the neighbouring influences on mutation processes. After the analyses in Chapter 2, the implication is clear, the neighbourhood signals invitation mechanism suggest that this information can be used to infer mutation mechanism from the sequence neighbourhood around point mutations. In Chapter 3, I focus on this property, and use sequence features to build classifiers to discriminate between mutagen-induced and spontaneous point mutations in the mouse germline. Finally in Chapter 4, a hypothesis has been proposed to explain distinctive features of genome organisation—compositional heterogeneity. I apply the methods I have developed to tackle this.

1.1 Overview of Mutations

Mutations are changes within DNA sequences, and DNA is the information system encoding for most living organisms (Graur and Li, 2000). Any variations within that information system, potentially, are of importance in the biology of the organism encoded by that genome. Mutations may interrupt the subsequent replication, transcription and/or translation of a gene, and give rise to a different functioning molecules (Cooke et al., 2003; Helleday et al., 2014). Consequently, an individual may end up with a different phenotype for natural selection to operate upon (Nei, 2013), or the interaction between DNA and its cellular environment may be disrupted and trigger a particular disease. In some cases, it could be cancer (Alexandrov et al., 2016).

1.1.1 Biochemistry of mutations

Mutation is a complex process that is triggered by DNA damage (Hartl et al., 1997; Strauss, 1968). DNA damage can be a point mutation, an insertion or a deletion of base pairs (Hartl et al., 1997). DNA is a biomolecule of two anti-parallel polymers which consisting of repeating monomeric units termed nucleotides. Nucleotides can be further decomposed into a phosphate group, a deoxyribose sugar and a nucleobase. It is the latter that differs between the four so-called bases of DNA and via which hereditary information is encoded (Watson et al., 1953). There are four different bases: adenine (A), guanine (G), cytosine (C) and thymine (T) (Darnell et al., 1990). Depending on numbers of carbon-nitrogen rings, these nucleotides can be classified as purines and pyrimidines. Purines are A and G, comprising two carbon-nitrogen rings and four nitrogen atoms. Pyrimidines are C and T, comprising a single carbon-nitrogen ring and two nitrogen atoms (Darnell et al., 1990). Point mutations are mutations in which, in a sequence, a base is replaced with another base. There are two sub-types of substitution mutations: transitions and transversions. Transition occurs when a purine is substituted with another purine or when a pyrimidine is substituted with another pyrimidine. All other point mutations are referred to as transversions (Darnell et al., 1990). Insertions are mutations in which additional base(s) are inserted into a DNA sequence and may lead to frame shifts (Darnell et al., 1990). Deletions are mutations in which base(s) are removed from a DNA sequence (Darnell et al., 1990). This study focused on point mutations only, because they are the most abundant mutation in human genomes and seem to be the simplest to model (compared to insertions and deletions).

1.1.2 Germline mutations

Depending on the cell type in which a point mutation occurs, a mutation can be classified as germline mutation or somatic mutation. A germline mutation arises from a lesion that occurs in germline cells that is incorrectly repaired. Germline mutations are heritable—that is, they are passed on to offspring (Arnheim and Calabrese, 2009) (see Figure 1.1). When these mutations achieve a population frequency of 1%

they are considered polymorphic within a population. If the variant increases to a frequency of 100%, it becomes a fixed difference between species and is referred to as a substitution (Hartl et al., 1997). This process is illustrated by the schematic in Figure 1.1. Therefore, germline mutations are critical contributors to all heritable phenotypic and functional diversity, including differences in individual athletic ability, susceptibility to diseases, complex phenotypic traits and so on (Bouchard et al., 1997; Ehlert et al., 2013; Gibson, 2009).

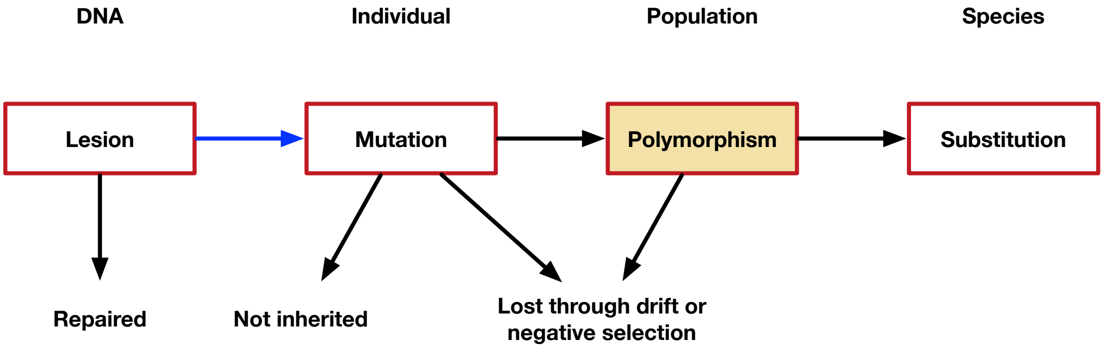


Figure 1.1: Schematic of the possible fates for a DNA lesion. If the lesion is repaired but incorrectly, it becomes a mutation in a sequence. The blue arrow in the figure indicates the erroneous DNA repair, which results in correction of the DNA sequence to a different base. If the mutation is transmitted to the next generation, and achieves a nominal frequency, it becomes a polymorphism within species. Eventually, if the polymorphism gets fixed between species, it becomes a substitution. Any process that influence tendencies indicated by red arrows will manifest genetic variation.

In this thesis, the approach that is employed for understanding and characterising mutagenesis involves the examination of polymorphism data. With reference to Figure 1.1, the fate of a new mutation is governed by processes that range in scale from chemistry of individual nucleotides to the biology of sex and mating system. Any particular biases in the operation of these processes that precede observations made at a particular level should manifest as patterns in the data. Accordingly, in the absence of natural selection, these patterns should reflect only the influence of neutral evolutionary processes and be strongly dominated by any biases originating from mutagenesis itself. I can obtain the polymorphism data directly from the Ensembl database (Flicek et al., 2013).

1.1.3 Somatic mutations

Somatic mutations are not transmittable to the next generation but nonetheless can influence the process of evolution by impacting on disease incidence. The process of somatic mutation is roughly illustrated in Figure 1.2, in a manner analogous to that shown for germline mutations in Figure 1.1. The origins of lesions can be quite cell-type specific, *e.g.*, alcohol for liver cells, or ultraviolet (UV) light for skin cells. If the lesion is subjected to erroneous repair, it becomes a mutation which can be inherited through subsequent mitotic cell divisions and remains in the cell. Some somatic mutations play a substantial role in the aetiology of cancers (Alexandrov et al., 2013). One example of this is hereditary non-polyposis colorectal cancer (HNPCC) (Peltomäki, 2001). From the International Collaborative Group on ICG-HNPCC database, it can be found that most of the mutations associated with HNPCC affect two gene groups associated with mismatch repair (MMR) (Peltomaki and Vasen, 1997). Failure in expression of MMR genes interrupts the MMR process, which allows colorectal cancer to occur (Peltomäki, 2001).

Some cancers are known to exhibit characteristic mutation profiles, which are termed cancer-related diagnostic signatures (Alexandrov et al., 2013). Peltomaki and Vasen (1997) noted that for colorectal cancers, a general pattern of mutation distribution is observed, including a combination of some splice-donor site point mutations, exonic nonsense mutations, and 3-base-pair deletions (Peltomaki and Vasen, 1997). This mutation distribution pattern forms a very distinctive signature between colorectal cancer and healthy tissue. Furthermore, current studies have also discovered that the prevalence of somatic mutations is highly variable between and within cancer classes, but understanding of the origins of this diversity is limited (Alexandrov et al., 2013).

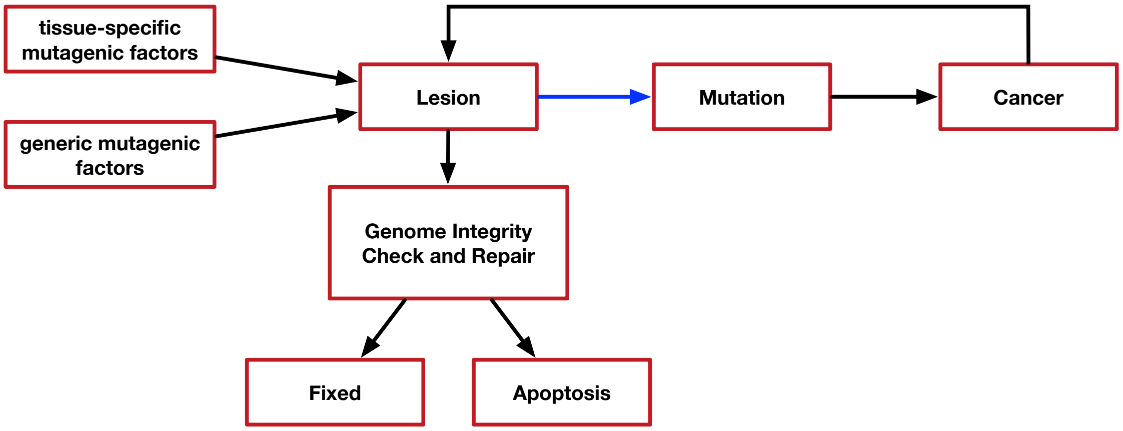

tissue-specific mutagenic factors
generic mutagenic factors
Lesion
Mutation
Cancer
Genome Integrity Check and Repair
Fixed
Apoptosis

Figure 1.2: Schematic of mutagenesis in somatic cells. At the beginning, a DNA lesion is induced by a tissue-specific mutagenic factor. The genomic integrity mechanism then either repairs the lesion by fixing it, or induces the apoptosis. If any of these cases occur, the lesion will be removed. However, if the lesion is repaired incorrectly, it becomes a mutation. The blue arrow in the figure indicates the erroneous DNA repair. Potentially, these mutations can cause disease, including cancer, which contributes to the formation of additional lesions.

1.2 Factors Affecting Mutations

Mutations can be influenced by a multitude of factors, including mutagens, DNA repair mechanisms, the influence of neighbouring DNA sequence on these processes, and attributes of gametogenesis.

1.2.1 Mutagens

Mutagens are agents that influence DNA lesion formation, (*e.g.*, radiation, chemicals, virus and/or bacteria). Different types of mutagens cause lesions at various cell cycle stages via different mutagenesis mechanisms (Helleday et al., 2014; Hoeijmakers, 2001).

Radiation, a physical mutagen that can cause DNA defects, is a form of energy that is given off by radioactive materials, (*e.g.*, X-rays and UV radiation) (Fajardo et al., 2001; Woods and Pikaev, 1994). Radiation was first found to be mutagenic in the 1920s (Duane et al., 1921; von Schwerin, 1920). When a radiation particle or wave penetrates a piece of tissue and hits cellular DNA, the DNA molecule absorbs the energy, and the covalent bonds that exist between the sugar phosphate backbone of the DNA may break (Rak et al., 2015), presenting the cell with the challenge of repairing a single- or double-stranded break in the DNA. Ionising radiation has long been used as a mutagen for genetic analysis and the investigation of gene functions for plants (Hirano et al., 2015; Tanaka et al., 2010) and animals (Bauer et al., 1938). In contrast, exposing cells to UV-radiation results in several types of mutagenic photoproducts of the nucleobases, like pyrimidine dimers and (6-4) photoproducts (Ananthaswamy and Pierceall, 1990; Cadet et al., 1992; Tornaletti and Pfeifer, 1996). In either case, failure to correctly repair this DNA damage results in a mutation (Carty et al., 1995, 1993).

Chemical mutagens are compounds that can increase the frequency of mutations in a genome. For example, 5-bromouracil (5-BU) is a common and widely used base analogue mutagen. During DNA replication, 5-BU mimics the structure of base T and replaces T (Henderson et al., 2003; Terzaghi et al., 1962). Within a DNA sequence, 5-BU exists in three tautomeric forms: keto, enol and ion. Each form has its own base binding preference (Hu et al., 2004). The keto form 5-BU is complementary to base A. Alternatively, the enol and ion forms are complementary to base G, which will cause A:T→G:C transitions (Hu et al., 2004). N-ethyl-N-nitrosourea (ENU), a synthetic alkylating chemical, is another type of commonly used chemical mutagen (Alvarez et al., 2003; Lee et al., 2012; Stottmann and Beier, 2010). ENU is able to transfer its ethyl group to oxygen and nitrogen reactive sites of DNA nucleotide molecules, and produce two forms of alkylation products (Noveroske et al., 2000; Shrivastav et al., 2010). If the DNA integrity system fails to repair these alkylation products, these alkylation products would be incorrectly recognised during the DNA replication process and cause mutations (Justice et al., 1999; Noveroske et al., 2000). A more detailed description of ENU mutagenesis will be presented in Chapter 3.

Viruses and/or bacteria are sometimes considered environmental factors and can also act as mutagens. Many studies have discovered that viral and bacterial infection leads to DNA defects, including chromosomal rearrangements, aneuploidy and polyploidy (Gershenson, 1986; Machida et al., 2010; Nichols, 1970; Storchova and Pellman, 2004). For example, the human papilloma virus (HPV), especially the E6 and E7 types, is able to alter the DNA structure in a cell, and suppress or inhibit the functions of the *p*53 gene (Münger et al., 1989; Narisawa-Saito and Kiyono, 2007; Werness et al., 1990). The *p*53 gene is part of a regulatory pathway that causes damaged cells to undergo apoptosis when damaged DNA cannot be repaired (Riley et al., 2008; Toledo and Wahl, 2006). When the *p*53 gene does not function properly after the HPV infection, the damaged cells will grow uncontrollably instead of dying.

1.2.2 DNA repair

DNA repair processes are crucial to maintaining genomic integrity. Repair of lesions can be critical to ensure the ability of cells to transcribe genes and to replicate. For example, the double strand break repair removes the whole damaged base, including a pentose sugar, nitrogenous base and phosphate group; and fixes a variety of damage, including photoreactivation products from UV damage (Carty et al., 1993; Tornaletti and Pfeifer, 1996) and cytosine deamination which generates a U-G mismatch (Kow, 2002). As alluded to earlier (Figures 1.1, 1.2), in some instances, DNA repair can still result in mutation(s). The variety of DNA repair processes is substantial. Of particular relevance to this work are MMR and transcription coupled repair (TCR).

DNA MMR (Iyer et al., 2006; Li, 2008) corrects mismatched nucleotides arising from replication errors, recombination and several classes of DNA damage. During DNA replication, MMR specifically recognises and repairs errors in the newly synthesised strand that has escaped correction by DNA polymerase proofreading of the DNA lesion. In vertebrates whose genomes are methylated, MMR is responsible for correcting T-G mispairs arising from 5-methylcytosine (5-mC) deamination (Petranović et al., 2000).

The TCR repair pathway operates specifically on the transcribed strand (Feng et al., 2002; Mellon et al., 1987). When RNA polymerase II encounters a lesion in DNA during transcription, it stalls without further elongation. In eukaryotes, a few nucleotides at the 3' end of the nascent RNA are then removed, followed by excision repairs—base excision repair (BER) and nucleotide excision repair (NER)—on the DNA lesion. Once the repair is complete, RNA polymerase II continues transcription. Thus, DNA repair is more frequent and rapid on the transcribed strand of actively transcribed genes. Some cancer types are associated with accumulative mutations interrupting the TCR mechanisms. Hainaut and Pfeifer (2001) discovered that the G→T transversions in smoking-associated lung cancer suggesting a possible contribution of TCR is strand bias in transcribed regions. There is also evidence that this process operates in the germline, with a marked transition from strand parity of the complementary bases between genes to strong disparity in transcribed regions (Touchon et al., 2003).

1.2.3 Sex—male-driven mutation

An excess of mutations originating in males has been known for some time (*e.g.*, Haldane, 1948). It has been argued that this excess originates substantively from differences in gametogenesis between the sexes (Ebersberger et al., 2002; Huttley et al., 2000; Miyata et al., 1987; Vicoso and Charlesworth, 2006). In mammals, the way males and females produce gametes is different (see Huttley et al., 2000). As there are more cell divisions in spermatogenesis compared to oogenesis, a corresponding difference in the number of DNA replications must also occur (Vicoso and Charlesworth, 2006). When coupled with the error-prone nature of DNA replication, female germlines are expected to contribute fewer mutations that transmit to the next generation (Haldane, 1948; Huttley et al., 2000; Vicoso and Charlesworth, 2006).

This hypothesis provides a number of predictions regarding heterogeneity in the genomic distribution of mutation. Y-linked genes spend all their time in males, and genes on autosomes spend half of their time in both males and females. In contrast, X-linked genes spend only one-third of their time in males and two-thirds of their time in females (Ebersberger et al., 2002; Vicoso and Charlesworth, 2006). The evolutionary properties mentioned above suggested that mutation rates occur according to the general inequality Y >A >X (Huttley et al., 2000; Miyata et al., 1987).

It has been argued that the strict DNA-replication hypothesis presented above was simplistic and other differences in cell biology / mutagenesis between males and females could contribute to this male bias. Of particular note, that an elevated level of DNA methylation in the germline of males, compared with that of females, may contribute (Driscoll and Migeon, 1990). The C→T transition occurs more frequently at methylated than at unmethylated cytosine sites (Coulondre et al., 1978; Duncan and Miller, 1980). Therefore, the difference in levels of methylation may account for at least some of the male bias in mutation rates. Analyses of Huttley et al. (2000) showed that elevated DNA methylation in the male germline was partly responsible for the elevated male mutation rate, which support bias in DNA methylation patterns as a factor contributing to a male bias in mutation rates.

1.3 Natural selection

Natural selection operates on target genomic sequences that encode individual phenotypes such as exons, RNAs and regulatory elements. While it does not explicitly influence the process of mutagenesis, it does affect the fate of individual mutations. Depending on the phenotypic effect of the mutation, the effect of natural selection can be suppressive or accelerative on a target genome, and consequently, may alter the genetic variation pattern across that genome (Lohmueller et al., 2011; Siepel et al., 2005). For instance, through *purifying selection*, deleterious mutations that significantly reduce the fitness of an individual are selected against, and prevented from becoming fixed in the population (Hartl et al., 1997). Therefore, natural selection may sculpt genetic variants in a consistent and non-random way. Mutations with positive phenotypic will increase in frequency over multiple generations, whereas mutations with strong negative phenotypic effects are quickly discarded (Hartl et al., 1997).

It is considered extremely unlikely that natural selection influences all mutations that occur in a genome. It has been estimated that only about 3-8 per cent of the DNA in the human genome is subjected to natural selection (Consortium et al., 2002; Graur et al., 2013; Siepel et al., 2005; Woofle et al., 2005). Thus, the vast majority of intronic and intergenic sequences should evolve in a selectively neutral manner.

1.4 Neighbourhood Influence on Mutations

Besides the different factors that may influence the occurrence of a mutation, considerable evidence indicates that the influence of neighbouring bases on point mutations is a general phenomenon (Cooper, 1995; Krawczak et al., 1998; Zhao and Boerwinkle, 2002). Two well-known examples are the CpG hypermutability (Coulondre et al., 1978) and thymine dimers (Boyce and Howard-Flanders, 1964). Beyond this, the assumption of neighbouring influences has been assumed for development of diagnostic signatures of carcinogenic processes (Alexandrov et al., 2013). Therefore, exploring the exact nature and size of neighbouring influence on mutation was one of the primary objectives of this study.

1.4.1 CpG hypermutability and thymine dimers

Early studies on inherited, and thus germline, mutations in humans supported the hypermutability of the CpG dinucleotide as the dominant origin of C→T mutations (Cooper, 1995; Coulondre et al., 1978; Krawczak et al., 1998). Grippo et al. (1968) discovered that the base C in CpG dinucleotide is frequently methylated to 5-mC. Coulondre et al. (1978) found mutation hotspots are associated with 5-mC residues. Cooper and Youssoufian (1988) established that, due to DNA methylase-binding affinity property, 5-mC is hypermutable (Cooper and Youssoufian, 1988) and will spontaneously deaminate to T (*i.e.*, C→T) about six to seven times more rapidly than the average base mutation rate. In addition, the DNA MMR mechanism cannot distinguish between the methylated and non-methylated strands (Drummond and Bellacosa, 2001), which may lead to the repair of T:G mismatches from 5-mC deamination on the opposite strand to produce the G→A mutation (Petranović et al., 2000). As the CpG effect involves different enzymes and metabolic pathways, the CpG effect would be an enzymatic or biochemical influence.

Cyclobutane pyrimidine dimers, most often thymine dimers, are lesions formed by UV radiation (Boyce and Howard-Flanders, 1964; Mouret et al., 2006). After exposure to UV light, energy is absorbed by the hydrogen bonds between the T:A or C:G base pair in DNA. This opens the bond and allows it to react with a neighbouring base. If the neighbouring base is another T or C base, it can form a covalent bond between the two bases. Most commonly, two thymine bases form a tight thymine dimer. If thymine dimers are not repaired before a DNA replication, or RNA transcription process and fail to be repaired, the lesion may be mismatched with the wrong bases and cause mutations (Witkin, 1969). This lesion is likely to induce C→T or CC→TT mutations, which may be associated with skin cancers.

1.4.2 Previous studies regarding neighbourhood influence

Numerous studies found that not only mutations induced by the deamination of methyl-Cytosine (associated with the CpG effect) and UV light (which give rise to thymine dimers), most mutations are affected by neighbouring bases (Alexandrov et al., 2013; Krawczak et al., 1998; Zhao and Boerwinkle, 2002). From analyses of mutations in human disease genes, Krawczak et al. (1998) quantified neighboring influence by contrasting observed base frequencies against an equiprobable frequency distribution via a Euclidean distance. In their results, they inferred the influence of neighbours is confined to the positions immediately flanking the mutated location, and dissipates monotonically while moving away from the mutation point (Krawczak et al., 1998). Zhao and Boerwinkle (2002) adopted an *adhoc* measurement, and demonstrated that these results applied more generally across the genome using human polymorphisms data. In their study, Zhao and Boerwinkle (2002) used just the base frequencies per position except beyond ± 10 bp where averages across position ranges were used.

Additional features of genomic sequence indicate that point mutation processes are likely to be affected by neighbouring bases, and these influences are heterogeneous in their operation across the genome. That mammal genomes are compositionally heterogeneous with respect to G+C bases has been known for some time (for review see Bernardi, 2000). Karlin et al. (1998) assessed genomic compositional contrasts based on dinucleotide and tetranucleotide odd ratios. The results indicate a non-random occurrence of nucleotides exists both within a single species and between species (Karlin et al., 1998). Chor et al. (2009) measured k-mer abundance heterogeneity within and between species, where k-mers are short DNA strings of length k. They discovered that different k-mers distribute unevenly across the genome, and a multimodal shape of k-mer distribution indicates the genetic k-mer distribution heterogeneity exists both within a single species and between species. A potential cause of this compositional heterogeneity is intra-genomic variability in context - dependent mutagenesis. For instance, in the CpG dinucleotide example, the 3' G is important for hypermutability of the 5' C. This relationship leads to the CpG dinucleotide being under-represented in genomic regions that are methylated; while the mutational end-products, TpG and CpA, are correspondingly over-represented. Therefore, neighbouring influence from the 3' G is essential for distributions of these three dimers.

Neighbouring bases were also considered by studies developing methods to identify systematic signatures of human cancer types. Motivated by the distinctive mutagenic biology of cancer, the related methods of Alexandrov et al. (2013) and Shiraishi et al. (2015) approach the problem of resolving the signatures for different mutational processes. Alexandrov et al. (2013) identified 21 signatures. Some signatures were associated with multiple cancer types, and others were highly specific to only one type of cancer. The resultant signatures are a composite of distinct underlying mutational processes operating across multiple types of point mutations;

they can contain instances of different point mutation directions (Alexandrov et al., 2013). These signatures are based on trinucleotides, where the mutated base is central, plus 1 immediate neighbour from each side.

1.5 Information Theory and Sequence Logo

While measuring the neighbouring influence on mutations, Krawczak et al. (1998) and Zhao and Boerwinkle (2002) aligned sequences with the same mutation direction together. These sequences are not evolutionarily related, in the sense of descended from a common ancestor. Instead, the sequences have in common the influence of a comparable mutagenic event and thus, arguably, a functional relationship. Aligning functionally related sequences makes these methods parallel to sequence logo analysis, a method widely employed for examining sequence function. The sequence logo is a visualisation method to understand consensus sequences (Schneider et al., 1986). This technique applies Shannon's information theory to molecular biological problems (Schneider, 2006). However, the method assumes bases in a sequence are equally frequent and occur randomly. As shown by the work of Karlin and Chor (Chor et al., 2009; Karlin et al., 1998), neither of these assumptions are correct.

Inspired by Schneider's sequence logo technique, I develop a novel set of information theory-based techniques to analyse and display the neighbouring association pattern. In this section, I will introduce the information theory and explain in detail how the neighbouring influence is visualised with the new method.

1.5.1 An information theory-based approach to measure neighbouring influence

Information theory, also known as the communication information theory, was developed by Claude E. Shannon to study fundamental limits of signal communication operations (Shannon, 2001). A classic communication system consists of five key components (see Figure 1.3): an information source, transmitter, channel, receiver and destination (Shannon, 2001). At the beginning, the information source produces a message. The newly generated message is then encoded into signals suitable for transmission over a channel by a transmitter. After passing a noisy channel, signals are accepted by a receiver, and a receiver regenerates a message from the received signal and sends the message to its destination (Shannon, 2001). A reliable communication system should receive an equal or similar amount of information as the original information source.

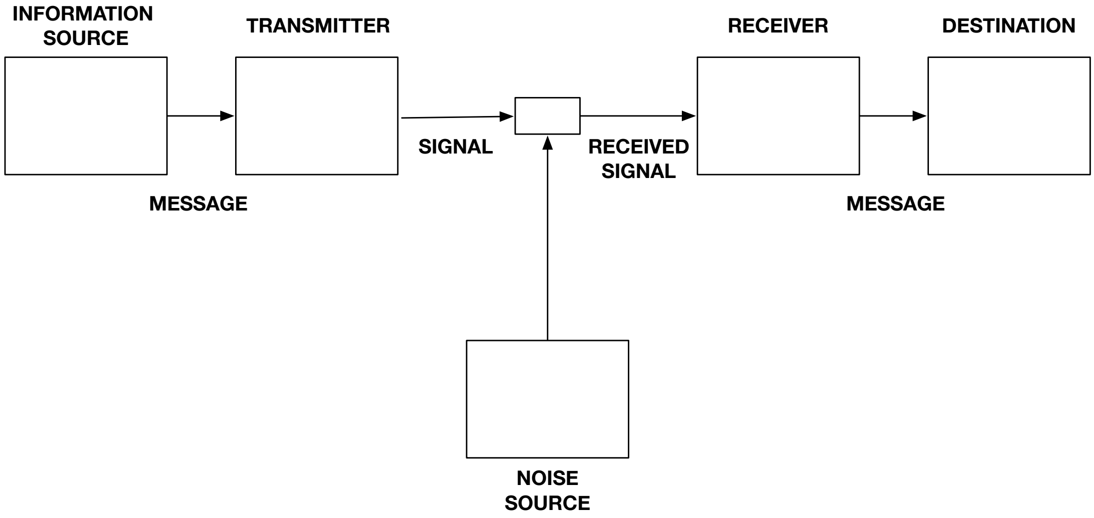


Figure 1.3: The five key components in a communication system according to Shannon. An information source, transmitter, channel, receiver and destination.

The amount of information received at the destination is measured with uncertainty. Shannon's entropy, denoted by H, was created to quantify uncertainty in the following way:

$$H(L) = -\sum_{i=1}^{M} p_i \log p_i \tag{1.1}$$

where p_i is the probability of the i^{th} symbol out of M possible symbols in a message of length L. According to Shannon (2001), information is received when uncertainty is reduced. Before any message is received, the possible outcome is uncertain. Thus, there is a maximum amount of uncertainty. When the message arrives, the amount of uncertainty will decline as the message is received and its content is known. Therefore, the information amount, also known as the mutual information (MI), is the uncertainty difference between an unknown message and the received message. MI is denoted by I and is calculated as a difference:

$$I = H_{before} - H_{after} \tag{1.2}$$

where H is the Shannon's entropy. A reliable communication system is expected to lower uncertainty as much as possible and achieve a higher value for I.

1.5.2 Schneider's sequence logo method

Molecular information theory is an application of information theory in molecular biology (Schneider et al., 1986). The biological genetic information system is an isomorphism of Shannon's information system (see Figure 1.4). DNA is a long molecule that is copied and inherited across generations. It is made of nucleotides that line up in a particular order within this long molecule. The order of these nucleotides carries genetic information and makes the DNA sequence an information source containing essential genetic messages. The sequence message is then transcribed into mRNA via transcription. mRNA plays the channel role, directing genetic messages to ribosomes, which are receivers decoding genetic messages into amino acids via translation (Yockey, 2005).

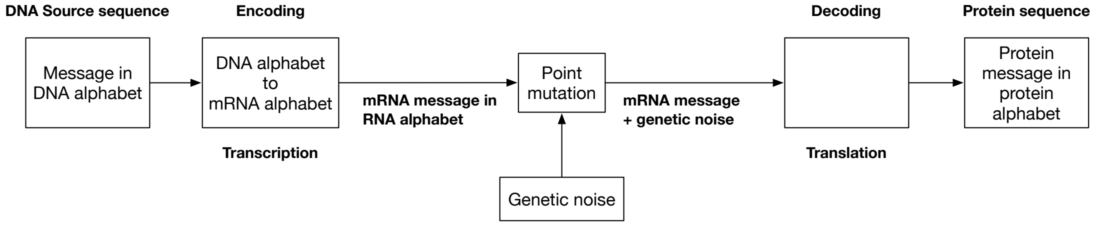


Figure 1.4: The biological genetic information system is an isomorphism of Shannon's information system (Figure 1.3).

Treating genetic information system as analogous to Shannon's information system, Schneider et al. (1986) developed the sequence logos technique to identify consensus sequences by evaluating the information content of corresponding DNA sequences.

Consensus sequences are short sequences found multiple times in the genome; they have the same function regardless of the location difference. By applying the molecular information theory, Schneider discovered *E.coli* ribosomal binding sites (Schneider, 2006), which is one of the many examples of how the approach has been applied. In order to evaluate the information content, hundreds of binding site sequences are aligned together by placing the start codon at the same position (Schneider, 2006), and this method assumes the alignment correctly orients the sequences with respect to the functional encoding. Then, MI is computed per position (i), denoted as $R_{sequence}(i)$, for positions in the sequence alignment using Equation 1.3:

$$R_{sequence}(i) = H_{uniform}(i) - H_{position}(i) \tag{1.3}$$

where $H_{uniform}(i)$ is the per position entropy of DNA sequences with equifrequent and random bases. After application of Equation 1.1, $H_{uniform}(i)$ is computed as $-\sum_{i \in A,C,G,T} p_i \log p_i = -(4 \times \frac{1}{4} \log \frac{1}{4}) = 2$ at each position. $H_{position}(i)$ is the per position entropy of observed DNA sequence alignment.

Coupling this metric with the sequence logo visualisation approach, Schneider's technique has become the most widespread application for discovery of functional motifs (Schneider and Stephens, 1990). The sequence logo approach used the MI statistic to define a stack of colour-coded letters, representing the sequence states, with each letter's height scaled proportionately to its contribution to the total MI (Schneider and Stephens, 1990).

1.6 Thesis Roadmap

The objective of this thesis is to elucidate the information content of neighbouring bases regarding mutation. A broad range of analyses were applied to address questions including: (1) Does neighbourhood affect mutations? (2) How do neighbourhoods affect mutations? (3) What is the size of the neighbourhood influence?

(4) Can we identify characteristic mutation motifs for mutations induced by different mutagenic mechanisms? These questions are evaluated in the following chapters.

Chapter 1 has provided an overview of mutations and mechanisms of mutagenesis. It has introduced the concept of a sequence neighbourhood influence on mutation. The major techniques applied in the thesis are inspired by one of the most widespread information theory-based applications for discovery of functional motifs. Furthermore, methods to measure sequence information for detecting consensus DNA sequence were introduced.

Chapter 2 describes a log-linear modelling approach for examination of mutation processes, and this work has been published. Via application of hierarchical log-linear models, this chapter addresses whether neighbouring bases are associated with mutation direction; in what way neighbouring bases associate between samples; and whether mutation abundance distributes equally between samples. My results replicate the well-known CpG effect, and demonstrate that all point mutations are affected by their neighbouring bases. In addition, the neighbourhood influence does not necessarily decay monotonically with distance.

Chapter 3 evaluates whether the information content of neighbouring bases can be used to distinguish mutagenic origins with machine learning classification algorithms. In this chapter, I draw on my results from Chapter 2 to contrast mutations originating from the multitude of natural mutagenic processes that normally operate in the mouse germline with those induced by the potent chemical mutagen ENU. The results affirm a highly significant influence of neighbouring bases on ENU-induced point mutations and these associations differ from those evident in spontaneous mutations. A logistic regression classifier was proved powerful at discriminating between different mutation classes. In addition, concordance between the feature set of the best classifier and the information content analyses reported in Chapter 2 suggest my results can be generalised to other mutation classification problems.

Chapter 4 describes an application of the log-linear modelling in Chapter 2. In this chapter, I examine and analyse the neighbouring influence on GC biased gene conversion (gBGC)-affected mutations, and the mutation abundance spectra. By contrasting the neighbouring influence between gBGC-affected and non-gBGC-affected mutations, I demonstrate that neighbouring influence exists with gBGC-affected mutations. Furthermore, characteristic mutation motifs are visualised with RE sequence logos. Strikingly, a core prediction made by the biased gene conversion hypothesis is not supported by my analysis. This work has important implications for our understanding of the origins of compositional heterogeneity in mammal genomes.

The conclusion chapter summarises my findings, their implications and discusses potential future work emanating from this thesis.

Chapter 2

Statistical Methods for Identifying Sequence Motifs Affecting Point Mutations

Yicheng Zhu, Teresa Neeman, Von Bing Yap, Gavin Huttley

This chapter has been published in Genetics. The reference for the publication is:

Zhu, Y., Neeman, T., Yap, V. B., and Huttley, G. A. (2017). Statistical methods for identifying sequence motifs affecting point mutations. *Genetics*, 205(2):843–856.

MY CONTRIBUTIONS Professor Gavin Huttley and I conceived and designed the experiments. Dr Teresa Neeman and Dr Von Bing Yap provided statistical advice. In particular, I collected the data, conducted the research, specified the statistical analyses, wrote the software, and drafted the manuscript.

Acknowledgments We thank Jeremy Widman for allowing us to use his Python implementation of logo drawing code for visualisation. We thank Ben Kaehler and Stephen Haslett for their comments on versions of this work.

Supervisor's signature: ____________________

ABSTRACT Mutation processes differ between types of point mutation, genomic locations, cells, and biological species. For some point mutations, specific neighbouring bases are known to be mechanistically influential. Beyond these cases, numerous questions remain unresolved including: What are the sequence motifs that affect point mutations? How large are the motifs? Are they strand-symmetric? Do they vary between samples? In this chapter, we present new log-linear models that allow explicit examination of these questions along with sequence logo style visualisation to enable identification of specific motifs. We demonstrate the performance of these methods by analysing mutation processes in the human germline and in malignant melanoma. Furthermore, we recapitulate the known CpG effect and identify novel motifs, including a highly significant motif associated with A→G mutations. We demonstrate that major effects of neighbours on germline mutation lie within ± 2 bp of the mutating base. Models are also presented to contrast the entire mutation spectra (distribution of the different point mutations). The spectra vary significantly between autosomes and the X-chromosome, with a difference in T→C transition dominating. Analyses of malignant melanoma confirmed reported characteristic features of this cancer including statistically significant strand asymmetry and markedly different neighbouring influences. The methods we present are made freely available as a Python library `https://github.com/HuttleyLab/MutationMotif`.

2.1 Motivation

Understanding the contributions of mutation processes to genetic diversity has broad relevance to topics ranging from estimating genetic divergence (Harris, 2015; Huttley, 2004; Schluter, 2009) to the aetiology of disease (Alexandrov et al., 2013; Nik-Zainal et al., 2012; Peltomaki and Vasen, 1997; Ying and Huttley, 2011). Mutations arise from DNA lesions, and various mechanisms have been characterised that cause DNA lesions (Cooke et al., 2003; Helleday et al., 2014). Similarly, a series of processes repairing DNA lesions have also been described (Helleday et al., 2014). By examining sequence composition, it is obvious that mutagenesis mechanisms differ between genomic locations (Francioli et al., 2015), cell types (Nishino et al., 1996) and species

(Karlin et al., 1998). In some cases, characteristic neighbouring sequence signatures around a mutation might be indicative of the underlying mutagenesis mechanism. For example, as the CpG example (Coulondre et al., 1978) illustrates, the specific CG dinucleotide contributes in predicting the CpG associate C→T mutations. However, besides this well-studied example, in general, by knowing only the starting and ending states of mutation, it is difficult to establish the associated mechanistic origin of that mutation. Motivated by this, we developed a statistical method and associated visualisation approach for revealing signature sequence motifs associated with point mutations. We refer to these signature sequence motifs associated with point mutations as mutation motifs.

The influence of neighbouring bases on mutation can have multiple causes. Due to chemical properties of neighbourhood DNA molecules, when certain bases are located together, they are more vulnerable and may alter the DNA structure much more easily when exposed to mutagen. Adjacent pyrimidines are more susceptible to a dimerisation in the presence of UV light (Brown, 2002, p 426). As the influence of DNA methylase preference for CpG dinucleotide demonstrates, DNA binding properties are likely to be another source of neighbouring influence on mutation. Adjacent bases may have different binding preferences to enzymes involved in different mutagenesis mechanisms. In addition, any sequence motif (*i.e.*, having high binding affinity to enzymes crucial in DNA repair systems) may result in those motifs being under-represented in mutated sequences.

Analysis studies for estimating neighbouring influences on mutation used different approaches and found that neighbourhoods do matter (see Section 1.3.2). However, Krawczak et al. (1998) and Zhao and Boerwinkle (2002) assumed bases are equally frequent and random, which is a common assumption in the extant research. However, it is incorrect because k-mer distribution is non-random across the genome in vertebrates, as well as in many other species. Therefore, these approaches potentially obscure the real signal by confounding it with the non-random occurrence of bases characteristic of DNA sequences. In addition, these results demonstrate the influence of neighbouring bases generalises only to somatic mutations. Early influential work on plant cpDNA demonstrated the generality of neighbouring influences across the tree-of-life (Morton et al., 1997). The limited sample source and ad hoc techniques may also have biased the results, making the conclusion not as generalisable as suggested. Alexandrov et al. (2013) and Shiraishi et al. (2015) considered neighbourhoods while developing mutation signatures. However, they studied a mixture of different processes, rather than focusing on single point mutation directions. Moreover, the exact nature and magnitude of neighbouring influence on mutation remains unsolved.

More recently, the influence of neighbouring bases has been examined using a probability of polymorphism conditioned on the sequence context (Aggarwala and Voight, 2016). A 7-mer context (± 3 bp neighbouring bases on both sides of a mutation) was identified as accounting for a median of 81 per cent of the variability in the probability of polymorphism across point mutations. This result indicates inclusion of larger neighbourhood size (3 bp and greater) interactions, accounting for as much as 50 per cent of the model's predictive power. However, k-mers exhibit a non-random distribution within the human genome (Chor et al., 2009; Karlin, 1998). Moreover, variation in sequence composition is correlated with variation in substitution rate (Hodgkinson and Eyre-Walker, 2011). These findings suggest that by averaging across all occurrences of the sequence context, the results of (Aggarwala and Voight, 2016) could reflect the relationship between genomic location and the probability of polymorphism rather than the mechanistic influence of neighbours on mutation.

Many of the developed techniques are confounded by common properties of genome DNA sequences. The ordering of nucleotides in DNA sequences is not random (Karlin, 1998). For the genomes of many organisms, such as vertebrates, there is also considerable genome variation in k-mer frequencies (Chor et al., 2009; Karlin, 1998). For instance, trinucleotide frequencies within a protein-coding exon are not well explained by the product of their monomer frequencies. Moreover, trinucleotide frequencies can differ between protein coding and non-protein-coding sequences due to the differing influence of natural selection. Thus, neighbour affect analyses on exons may exhibit greater error rates unless such confounding is considered.

Most available methods tackling neighbouring influence on mutation do not distinguish between contributions from independent positions and joint contributions from multiple positions. The independent contribution and joint contribution are different ways in which neighbourhoods can affect mutations. Independent influence is from individual positions and each position will not interfere with the influence level of others. Joint influence is a form of interaction of neighbouring positions compositions (*e.g.*, two-way interaction from any two-position combination, three-way interaction from any three-position combination *etc*). Question like “Are mutations affected by the sequence of bases present at two positions (Zhang and Mathews, 1995)?” have not been addressed in current related studies.

Log-linear models allow flexible parameterisations for hierarchical hypothesis testing of categorical data, and have been previously applied to examination of neighbouring influences (Huttley et al., 2000). Their generality allows for control of potential confounding differences, such as differences in sample size and nucleotide composition. The support for comparing hypotheses in a hierarchical manner enables explicit examination of hypotheses, such as strand symmetry and the absence of higher-order effects, (which have been assumed in other approaches) (Aggarwala and Voight, 2016). Thus, they provide an objective basis for identifying parametrically succinct models.

In this research, we develop log-linear approaches to examine mutation processes. Our work is distinguished from previous methods by conditioning on the mutation event, rather than the sequence context. More importantly, we employ a control distribution matched for genomic location. We present hierarchical hypothesis tests for evaluating whether: (i) neighbouring bases associate with mutation direction; (ii) neighbouring base associations are equal between samples; and (iii) the spectrum of mutations (the relative abundance of the 12 point mutations) are equal between samples. A sequence logo-inspired visualisation approach is also presented. We demonstrate application of the models by applying them to data previously reported to exhibit distinctive mutation processes; namely, germline mutations in different sequence classes (*e.g.*, transcribed, untranscribed) and chromosome classes (*e.g.*, autosome and sex-chromosome), and somatic mutations in cancer. Mutation events in both human germline and somatic tissues are inferred from single-nucleotide genetic variants available in Ensembl. We recapitulate the well-known CpG effect and our results indicate that neighbourhood size can be quite large. In addition, as demonstrated in the A→G transition mutation, the influence of neighbours does not decay monotonically with distance. Furthermore, we show that both independent and dependent position influences contribute to mutational process. Through formal testing of equivalence between samples, we demonstrate significant differences between sequence classes, chromosome classes and between melanoma and germline mutations. Software implementing all these methods, released under an open-source licence, is made available `https://github.com/HuttleyLab/MutationMotif`.

2.2 Materials and Methods

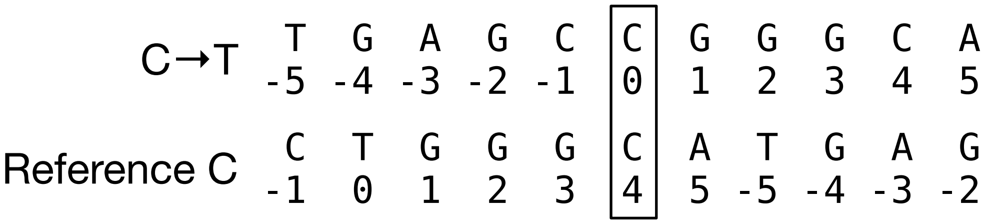
2.2.1 Data sampling

In this research, mutation events in humans were inferred from published genetic variant records. Germline mutations were inferred from single-nucleotide polymorphic (SNP) sites. Somatic mutations were inferred from genetic variants identified in cancers. In both cases, the mutation direction, location and associated flanking sequence were sampled from Ensembl (Flicek et al., 2013) release 79 using PyCogent's Ensembl querying capabilities (Knight et al., 2007). The Ensembl variation database records whether a variant is classified as somatic. We sampled germline SNPs using this flag and required the Ensembl record to indicate that the SNP was validated, had an inferred ancestral allele and its flanking sequence matched the reference genome. For each such filtered SNP, we recorded the alleles, ancestral allele, strand, sequence class (exonic, intronic or intergenic), genomic coordinates and 300 bp of flanking sequence either side of the SNP location.

Sampling somatic genetic variants involved both the COSMIC (Forbes et al., 2014) and Ensembl databases. Complete mutant export data was obtained from COSMIC, which included variant identifiers and the primary pathology from which a variant had been reported. Flanking sequence was derived through the Ensembl records for the variant identifiers, ensuring the record was flagged as somatic and following the same procedure as for germline variants. We restricted our attention to variants identified from malignant melanoma.

2.2.2 Determining base counts

For each mutation direction (*e.g.*, C→T), we obtained base counts from paired mutated and reference base locations. Neighbour positions were indexed relative to the position of the chosen location. For a mutated base, the chosen location was the annotated site of the variant (see Figure 2.1). With knowledge of the mutation direction, a location with the same starting base as that affected by the mutation was randomly sampled within 300 bp of the annotated variant. (*e.g.*, a random choice of a position with a C in the case of a C→T mutation); but excluding the variant location. This is the paired reference base. In each case, a 5 bp long sequence centred on the chosen location was extracted; bases observed per relative position were recorded. These are referred to these as neighbourhoods. As the total number of possible neighbourhoods of a 5 bp-long sequence with a given mutation centred was 256 ($4 \times 4 \times 4 \times 4 = 256$), a single file was written with counts for each of the possible neighbourhood combinations for both the mutated and reference locations. This approach to identifying the reference distribution confers a substantial computational advantage, both in terms of memory required and computing time.



C→T	T	G	A	G	C	C	G	G	G	C	A
	-5	-4	-3	-2	-1	0	1	2	3	4	5
Reference C	C	T	G	G	G	C	A	T	G	A	G
	-1	0	1	2	3	4	5	-5	-4	-3	-2

Figure 2.1: Sampling mutated and reference base neighbourhoods. The neighbourhood of a position at which a C→T mutation occurred is compared with the neighbourhood of a reference occurrence of C randomly selected from within ±300bp of the C→T mutation. (The example sequence is greatly shortened to simplify the figure.) The location of the C→T variant is the central position for the mutated base and is assigned the index 0. The C at position 4 was randomly chosen as the reference location and the sequence is shifted so it is centred on this position (see Section 2.2.2).

2.2.3 Log-linear modelling of neighbour effects

First demonstrated is the general approach of applying log-linear models to understand neighbour influences on mutation, by focusing on the influence of a single neighbouring position. Expected counts are modelled using the Poisson distribution for all log-linear models described in this research. We then considered the extension of comparing neighbour contributions between samples. Both of these analyses are concerned with the independent contribution of bases at a position to mutation status.

For a single position, we evaluated whether *base* and mutation *status* occur independently using a straightforward log-linear model. Under the most saturated log-linear model, the log of the expected frequency f_{is} for *base* i and mutation *status* s can be expressed as

$$\ln f_{is} = \lambda + \lambda_i^{base} + \lambda_s^{status} + \lambda_{is}^{base:status} \tag{2.1}$$

where λ represents the intercept (*i.e.*, common to all counts), λ_i^{base}, the contribution to the frequency of being *base* i, λ_s^{status} the contribution to the frequency of being mutation *status* s, and the interaction between *base* and *status* $\lambda_{is}^{base:status}$. The last

$\lambda_{is}^{base:status}$ term expresses the degree of non-independence between *base* and mutation *status*. The number of levels for each factor are: *base*, four levels (A, C, G, T); and mutation *status*, two levels (mutated, M and reference, R). As the total counts for M and R are identical by design, $\lambda^{status} = 0$ for all s. The fit of a log-linear model is measured as the deviance (D). We specify the null hypothesis that bases occur independently of mutation status by setting $\lambda^{base:status} = 0$ for all i, s. The alternate is the fully saturated model. The difference in D between the null and alternate, nested models, is taken as χ^2 with degrees of freedom equal to the difference in the number of free parameters. In this instance, the degree-of-freedom is 3.

When comparing groups (*e.g.*, autosome versus X-chromosome), we added another factor, λ_g^{group}, to the log-linear model 2.2, to account for the contribution to the frequency of being *group g*. The fully parameterised version of this log-linear model now requires the addition of three interaction parameters: 2 two-way interaction parameters $\lambda^{base:status}$ and $\lambda^{base:group}$, and the three-way interaction parameter $\lambda^{base:status:group}$. The three-way interaction parameter $\lambda^{base:status:group}$ represents the influence of *group* on the *base* : *status* interaction. Therefore, we evaluated the null hypothesis of no difference between samples by setting all $\lambda^{base:status:group} = 0$ and compared this against the fully saturated model. If the *group* factor has only two levels, then the degree-of-freedom for the resulting D is 3.

$$\begin{aligned}\ln f_{isg} = \lambda &+ \lambda_i^{base} + \lambda_s^{status} + \lambda_g^{group} \\ &+ \lambda_{is}^{base:status} + \lambda_{ig}^{base:group} + \lambda_{sg}^{status:group} \\ &+ \lambda_{isg}^{base:status:group}\end{aligned} \tag{2.2}$$

We extended this approach to consider the simultaneous influence on mutation status of bases at multiple positions. To illustrate, consider the two neighbours following the base C in Figure 2.1. There are 16 possible dinucleotides at the +1 and +2 positions. The goal of this model is to establish whether the dinucleotides at these two positions jointly affect the mutation status of C, after considering the independent contributions of these positions. To achieve this, our two-position interaction model extended the independent contribution model 2.1, by adding factors for the additional position and interaction terms between the parameters. The fully saturated two-position interaction model is:

$$\begin{aligned}\ln f_{ijs} = \lambda + \lambda_i^{base_1} + \lambda_j^{base_2} + \lambda_s^{status} + \lambda_{is}^{base_1:status} \\ + \lambda_{js}^{base_2:status} + \lambda_{ij}^{base_1:base_2} + \lambda_{ijs}^{base_1:base_2:status}\end{aligned} \tag{2.3}$$

where λ^{base_1} and λ^{base_2} represent the base contributions at positions 1 and 2. In addition to including factors for the independent contributions of the two positions on mutation status, the $\lambda^{base_1:base_2}$ accounts for non-independent occurrence of bases at the positions—a key property of DNA sequences. The null hypothesis of no interaction between dinucleotides and mutation status is specified by setting all $\lambda^{base_1:base_2:status} = 0$ and comparing this against the fully saturated model. The resulting D has 9 degree-of-freedom. For a given mutation direction, we performed this analysis for all possible combinations of pairs of sites.

These approaches were further extended to consider interactions among three positions, four positions and for comparison of these effects among groups.

2.2.4 Log-linear model of mutation spectra

For analysis of mutation spectra, we evaluated the null hypothesis that the distribution of mutations is the same between groups. The opportunity for a specific mutation direction is affected by the total occurrence of the starting base. This quantity can be difficult to ascertain, such as in cancers in which there may be major genomic rearrangements (*e.g.*, deletions) relative to a reference group. To avoid this uncertainty, we restricted the analysis to point mutations from a specific base, comparing the relative counts of each of the three possible mutations between groups. This is a test of independence between ending base and group.

For a specific base, the log of the expected frequency is defined as:

$$\ln f_{dg} = \lambda + \lambda_d^{direction} + \lambda_g^{group} + \lambda_{dg}^{direction:group} \tag{2.4}$$

where the factor $\lambda^{direction}$ represents the counts of the three different point mutation directions with the same starting base, λ^{group} the counts in the different groups, and $\lambda^{direction:group}$ the interaction between these factors. We specified the null hypothesis of equivalent proportions between groups by setting $\lambda^{direction:group} = 0$. For two groups, comparing against the fully saturated model, the D has 2 degree-of-freedom.

2.2.5 Visualisation

As introduced in Section 1.4, the original sequence logo technique displays motifs using the total MI at a position as the letter stack height. The individual base fraction, contributed to total MI by individual bases, is derived from their individual terms in the MI calculation. We adopted a similar approach in this research. Instead of using MI, we used relative entropy (RE) to measure the relative contribution from each base. The log likelihood ratio D is converted to RE by dividing by twice the sample size. Total RE from a log-linear analysis specifies the letter stack height. We used the terms in the RE equation to determine the proportion of the individual stack height attributable to a specific base. This differed from the conventional sequence logo approach by distinguishing between bases that are under- or over-represented in the mutated class, relative to the unmutated class. Under-represented bases are indicated by a 180° rotation.

Interpretation of the logo is straightforward. A higher RE value indicates that a position has a greater influence level on mutation. The corresponding p-value from the log-linear model, that the data arose under the null hypothesis, indicates whether a stack height reflects a meaningful influence on mutation or not. The magnitudes and orientations of letters further convey meaning in that ordinary letter orientation is indicative of over-representation in the mutated group, while inverted orientation indicates under-representation. Here, we opted to use residuals from the mutated class for display. Using residuals from the unmutated class would generate an image with the opposite letter orientations.

For models measuring dependent neighbourhood effects between multiple positions, we developed multi-position models. For these models (*e.g.*, Equation 2.3), the stack height is identical between the indicated positions. For the two-position model, the characters of the nucleotide pair at the two-position combinations share the same proportion and orientation. For the more complicated analyses involving contrasting neighbour effects between groups, the reference category was the one provided first to the software.

Differences in mutation spectra are visualised using a grid with rows corresponding to the starting base and columns corresponding to the base resulting from the mutation. Each row corresponds to a single log-linear test for equivalent distribution of the possible point mutations from the base indicated by the row label (see Section 2.2.4). The RE for each row is computed from the deviance of the corresponding spectra test. Letter heights for each base are scaled proportionally to the corresponding term in the RE equation. The sum of letter heights in a row is the total RE for that test. Bases over-represented in the reference group are oriented in the conventional manner, while under-represented bases are rotated 180°. In the spectra analysis, the largest base in the grid is the dominant mutation product difference between the groups.

2.2.6 Availability of data and materials

`MutationMotif` is a Python 3.5 compatible library for performing the statistical analyses outlined in this research; it is freely available under an open-source licence. The project homepage is `https://github.com/HuttleyLab/MutationMotif` and the version employed for the reported work is available in Zenodo (DOI 10.5281/ zenodo.166388). It draws on R (Ihaka and Gentleman, 1996) for log-linear modelling, via the glm function, using the rpy2 Python binding to R. Sequence logo's are drawn using custom Python code included in `MutationMotif`. Other dependencies include PyCogent (Knight et al., 2007), `pandas`, `numpy`, `matplotlib` and `scitrack`.

The scripts performing the data sampling and applying the analyses reported in this research are freely available under the GPL at `https://github.com/HuttleyLab/AnalyseMutations`. The version employed for the reported work is available in Zenodo (DOI 10.5281/zenodo.166387). `AnalyseMutations` includes the counts data required by `MutationMotif` and the complete set of results contained in this research. These counts data were produced from data sampled from the Ensembl and COSMIC databases, as described in sec:data. As the data files from which the counts files were produced are so large, they are available separately in Zenodo (DOIs 10.5281/zenodo.53158 `https://zenodo.org/record/53158` and 10.5281/zenodo.53164 `https://zenodo.org/record/53164`) under the Creative Commons Attribution-Share Alike license. Data files are typically in gzip compressed standard formats, including tab delimited text files, FASTA formatted sequence files. Serialised data are stored as JSON or pickle (Python's native serialised format). Supplementary Information 1 contains tables and figures from additional analyses.

2.3 Results

2.3.1 Overview of notation and neighbour effect log-linear models

The notation X→Y refers to a point mutation from starting base X to ending base Y, X→Y* refers to a point mutation and its strand-symmetric counterpart (*e.g.*, C→T* refers to C→T or G→A). The sampled region around a mutated base is called a neighbourhood, with neighbours being the individual positions within the neighbourhood. A mutation motif is a specific neighbourhood that is enriched in mutated sequences compared to the reference distribution.

The log-linear model of neighbour influence evaluates the null hypothesis that a neighbouring base(s) flanking a specific point mutation is the same as that flanking a random occurrence of the starting base. For instance, does the distribution of bases at sites flanking C→T mutations differ from that flanking all C's? As the frequency of bases varies between genomic locations (Bernardi, 2000; Chor et al., 2009; Karlin, 1998), matching of the mutated and reference locations reduces possible confounding. We achieved this matching by deriving a reference location proximal to each mutated location. The sampling process is shown in Figure 2.1. We sampled 300 bp of flanking genomic sequence each side of a variant. Within this segment, we randomly selected another occurrence of the starting base, the same as in the mutation event. Unless stated otherwise, we limited our analysis of neighbouring influence to ±2bp either side of the mutated position, resulting in 256 possible neighbourhoods. For any given mutation direction, counts of these different neighbourhoods are obtained from both the sample centred on the mutated base and the sample centred on a random occurrence of the starting base. These counts are used to construct the contingency tables for the log-linear analysis. This approach achieves the objectives of controlling for compositional variation across the genome and controlling for the non-random occurrence of bases (See Section 2.2.2).

The log-linear models used to examine the effect of neighbours on point mutation include parameters that represent an interaction between neighbouring base(s) and mutation status (see Section 2.2.3). The contribution of this parameter to model fit is measured as a deviance which is used to calculate the corresponding p-value for the null hypothesis. We converted the deviance to RE, as RE measures the information content of the data under the model in a manner that is not sensitive to sample size, thus, allowing comparisons among analyses between different samples with different sample sizes.

As concerned with whether flanking positions individually or jointly affect mutation processes, we described the influence of neighbouring bases as independent or dependent/joint effects respectively. The influence of a base at a single neighbouring position on a point mutation will be referred to as an 'independent' effect. When bases at two or more neighbouring positions influence a point mutation, it will be referred to as a 'dependent' interactive effect or the joint influence of multiple bases. The number of positions involved in a dependent effect is referenced as the 'order' of the interaction. An independent effect, the influence of a single position on mutation, is a first-order effect, and the joint influence of two positions on mutation is a second-order effect. Flanking locations are indexed relative to the mutated position. The immediate flanking 5' base is at position -1, while the immediate flanking 3' base is at position $+1$ (see Figure 2.1). A series of positions are indicated by the relative indices in parentheses; for example, $(-2, -1)$ are two positions 5' to the mutated base. In the case of a dependent effect, the actual positions are not necessarily physically adjacent; for example, $(-2, 2)$.

2.3.2 Log-linear models recapitulate the CpG effect and reveal higher order effects

The analyses below focuses principally on analyses of intergenic autosomal data. We also sampled variants from introns and exons. All results from analysis of other genomic regions have been relegated Supplementary Information 1, as the results are substantively the same as those from the intergenic sequence class.

Given the golden example CpG effect, we benchmarked our newly developed method by examining the influence of neighbouring bases on C→T point mutations in the autosomal intergenic sample. (As none of the strand symmetry tests were significant for the intergenic autosomal mutations, we limited our discussion to the 'plus' strand directions only.) We expected the influence of methylation-induced deamination at CpG to reveal a strong G effect at the $+1$ position (Cooper and Youssoufian, 1988). This prediction was confirmed in the results of the hypothesis test (see Table A.1), and visually in the mutation motif logo (see Figure 2.2B). The analysis established that while all positions made highly significant independent contributions to mutation (all p-values were estimated as $\approx\ 0$; see Table A.1). The magnitude of independent influence from $-2,\ -1$ and $+2$ positions respectively was small compared to that from the $+1$ position. Only one of these was evident in the mutation logo, that of A at the −1 position (see Figure 2.2B). (Results from the equivalent analysis of autosomal exon data are shown in Figure A.1.)

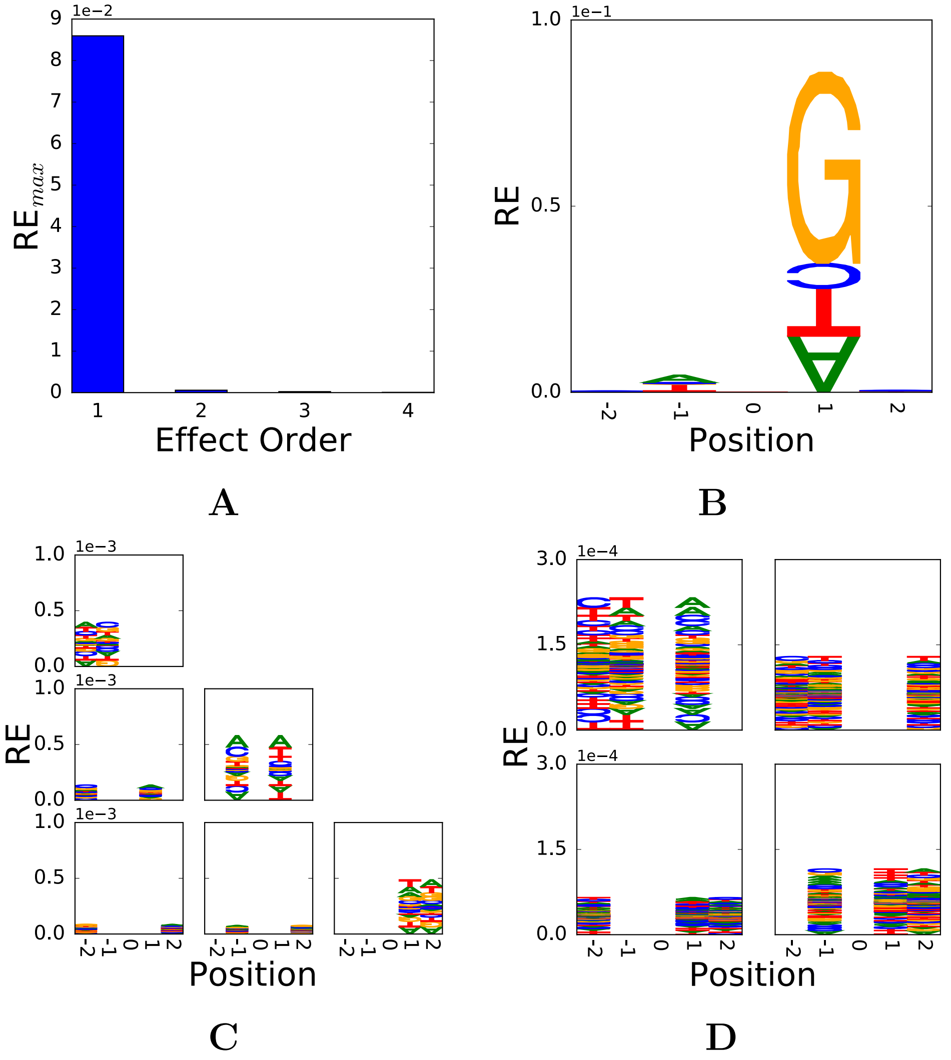


Figure 2.2: Neighbours influence C→T mutations in autosomal intergenic sequences. (**A**) First-order effects are the dominant neighbour influence, RE_{max} (y-axis) is the maximum RE from the possible evaluations for a motif length (x-axis), (**B**) Single-position effects, (**C**) Two-way effects, and (**D**) Three-way effects. For B-D, the y-axis is RE and the x-axis is the position index relative to the mutated base. For details on interpreting the logo see Section 2.2.5.

Specific combinations of bases at multiple positions also significantly affected C→T mutations. All higher-order interactions were statistically significant (all p-values

$< 10^{-22}$, see Table A.1). A feature of the second- and third-order joint effects was that bases physically adjacent to each other or to the mutated position had the strongest association: $(-2, -1)$, $(-1, +1)$, $(+1, +2)$ second-order interactions (see Figure 2.2C and Table A.1), and the $(-2, -1, +1)$ third-order interaction (see Figure 2.2D).

Despite the highly significant associations between combinations of positions and interactions, the independent position contributions dominated. All effect orders were significantly associated with mutation status, even when using the sequential Holm-Šidák correction for 15 tests (Holm, 1979). These results reflect the enormous statistical power resulting from the large sample sizes (*e.g.*, over 1 million C→T intergenic variants). By comparing the maximum RE value (RE$_{max}$) from each effect order (see Figure 2.2A), different RE$_{max}$ magnitudes of these different effects provide a useful indicator of their relative influence. RE$_{max}$(1) is the maximum RE score for independent effects across all positions (*e.g.*, +1 position in this case), RE$_{max}$(2) the maximum RE score from combinations of two positions, and so on for the higher orders. This display established that the 3'-G influence dominates all other neighbouring base effects on C→T mutation. Furthermore, contrasting the RE$_{max}$ values between the point mutations (see Table 2.1) affirms that neighbours have the strongest effect on C→T mutations (see Figure A.2).

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)	**RE**$_{max}$(3)	**Pos.**(3)
A→C	0.0039	-1	0.0016	(+1, +2)	0.0012	(-2, -1, +1)
A→G	0.0188	+1	0.0030	(-2, -1)	0.0007	(-2, -1, +1)
A→T	0.0095	+1	0.0051	(-1, +1)	0.0023	(-1, +1, +2)
C→A	0.0091	+1	0.0044	(-1, +1)	0.0015	(-1, +1, +2)
C→G	0.0054	-2	0.0025	(+1, +2)	0.0010	(-1, +1, +2)
C→T	0.0860	+1	0.0006	(-1, +1)	0.0002	(-2, -1, +1)

Table 2.1: Summary of neighbour associations with plus strand mutations with an autosomal intergenic location. RE$_{max}$(#) is the maximum RE for order # and Pos.(#) the corresponding position(s). All point mutations had at least one significant test after correcting for 15 tests (see Table A.1 in Supplementary Information 1) using the Holm-Šidäk procedure.

2.3.3 A→G mutations are also strongly affected by neighbours

The A→G transition mutation exhibited the next strongest influence of neighbouring bases (see Table 2.1). As for C→T, all effect orders were highly significant after correcting for 15 tests (all p-values $< 10^{-47}$, see Table A.2). All positions showed significant first-order influences, but the $-2, -1, +1$ positions were particularly strong (see Figure 2.3B). Two of these, $(-2, -1)$, also exhibited a prominent second-order interaction (see Figure 2.3C), while all three contributed the strongest third-order interaction (see Figure 2.3D). For A→G mutations, the analysis indicated that while first-order effects dominated, higher-order effects were important factors affecting this mutation direction (see Figure 2.3A). Again, combinations of bases that were physically adjacent were most influential. (Results from the equivalent analysis of autosomal exon data are shown in Figure A.3.)

2.3.4 Transversion mutations are affected by neighbours

All transversion mutations had significant neighbour influences but to a lesser extent than that evident for transition mutations (see Table 2.1 and Figure A.2). The transversion mutations showed $\text{RE}_{max}(1)$ that were 20-fold less than for the C→T mutations. However, higher-order effects were typically more pronounced for transversions than they were for transitions. The A→T and C→A transversion mutations showed the greatest influence of neighbours at all levels. The dominant influences were immediately adjacent to the mutating base, except for C→G, in which position -2 had the strongest effect.

2.3.5 The size of the neighbourhood

The analyses above indicated first-order effects exerted the strongest influence on mutations. Accordingly, we limited our examination of neighbourhood size to first-order effects, and sampled intergenic autosomal variants with a flank size of ± 10bp for an analysis. After correcting for multiple tests, all 20 flanking positions were significant for all point mutations (see Table A.3). This suggests a neighbourhood size $\geq$ 10 bp. In this analysis, the highly significant neighbourhood influence for even

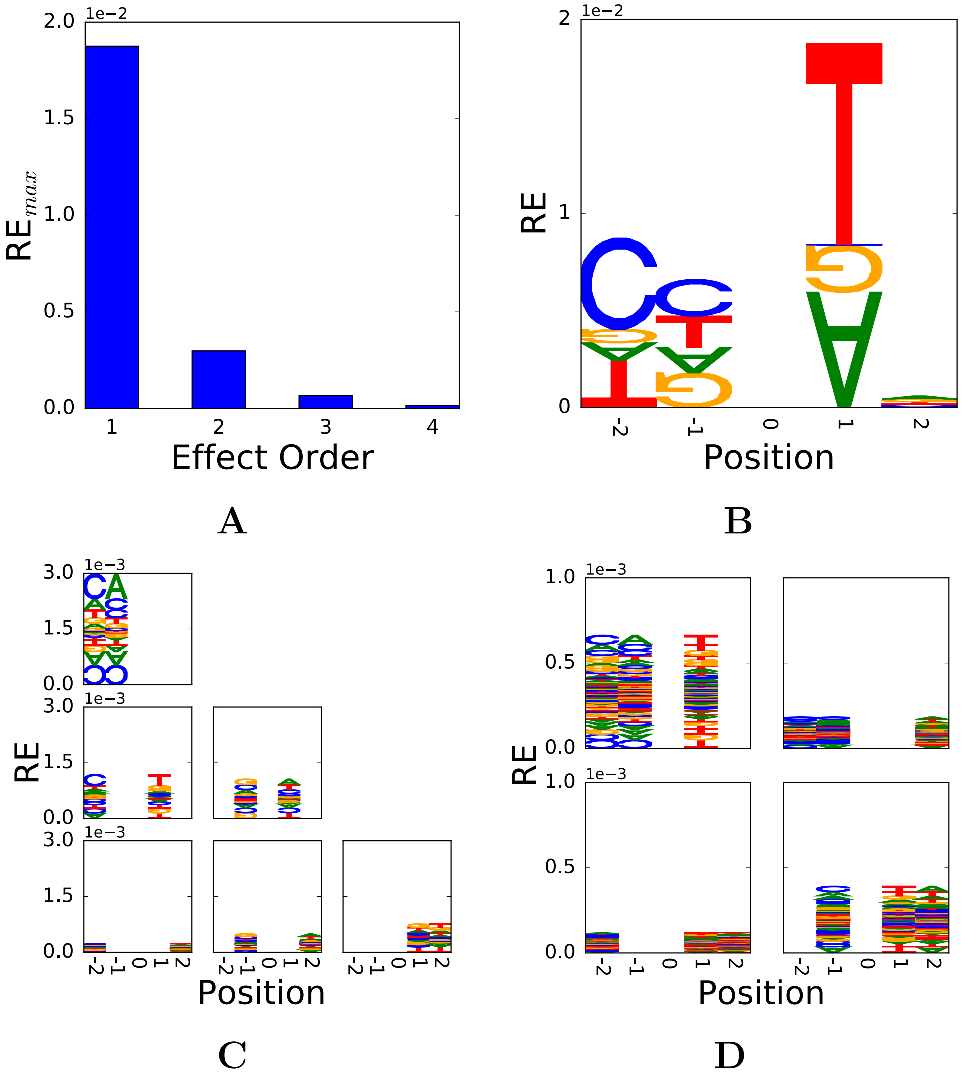


Figure 2.3: Neighbours influence A→G mutation in autosomal intergenic sequences. (**A**) First-order effects are the dominant neighbour influence, (**B**) Single-position effects, (**C**) Two-way effects, and (**D**) Three-way effects. For B-D, the y-axis is RE and the x-axis is the position index relative to the mutated base.

very distant positions reflects the enormous sample sizes employed for this analysis. It does not necessarily reflect the magnitude of a position's influence. Therefore, for each mutation, we estimated the most distant position with an RE that was $\geq 10\%$ of $\text{RE}_{max}(1)$. For the transition mutations, the neighbourhood size was restricted to positions within ± 2 bp (see Figure A.4); while for transversion mutations, the neighbourhood size was within ± 4 bp (see Table A.3).

2.3.6 Some germline point mutations exhibited different neighbouring effects between sequence classes

The operation of transcription-coupled DNA repair processes suggested a possible difference in neighbour effect may exist between transcribed and untranscribed sequences. This predicts a difference in mutation profile between intergenic and intronic sequences. Analysis of neighbour contributions to mutation established that, for first-order effects, every point mutation was significantly different between sequence classes (see Table A.4). For second-order effects, only the transition mutations showed significant differences. In addition, the biggest difference between the regions was for $\text{A}\rightarrow\text{T}^*$. While these effects were highly significant, their $\text{RE}_{max}(1)$ were ≈ 100 fold lower than the overall influence of neighbours on intergenic A→T.

2.3.7 Neighbouring effects differ between chromosome classes

Differences in germline biology between males and females predict distinct mutation profiles between sequences located on the autosomes and X-chromosome (Huttley et al., 2000). To compare the flanking base effect between autosome and X-chromosome mutations, our null hypothesis of no difference in neighbourhood effect between chromosome classes in intergenic sequences was rejected for first-order influences on several of the point mutations, after correcting for 15 tests using the Holm-Šidák procedure (Holm, 1979) (see Table A.5). Interestingly, $\text{A}\rightarrow\text{G}^*$ and $\text{C}\rightarrow\text{T}^*$ showed comparable differences in flanking base effect between the chromosome classes (deviances ≈ 26.0 and ≈ 25.4 respectively). In all cases, the effect exists at the same position as that identified as $\text{RE}_{max}(1)$ in the intergenic analysis (see Table 2.1). While the transition mutations were the most statistically significant, their RE lay within the range of the other point mutations (see Table A.5), indicating that their significance reflects greater abundance and thus, a greater rate.

2.3.8 Analysis of germline mutation spectra

Our log-linear model for analysis of mutation spectra compares counts of point mutations from the same starting base between groups. By considering only mutations from a single base between different locations, differences in the abundance of the starting base between groups are controlled for. This approach can be applied to groups representing different strands, different genomic regions or different biological materials (*e.g.*, germline and somatic).

Our analysis of germline mutation spectra indicated that point mutations were uniformly strand-symmetric, but different between sequence categories. No sequence category exhibited strand asymmetry in mutation spectra for autosomal data. Significant differences in autosomal mutation spectra were evident between intergenic and intronic regions. The major differences were for transversion mutations, specifically C→A and its strand complement (see Table A.6).

Significant differences between chromosome classes were evident (see Figure 2.4 and Table A.7). For the intergenic sequence class, A→G* transition mutations were in strong excess on autosomes compared with X-chromosome (see Figure 2.4). Comparable results were evident for intronic sequences (see Table A.8).

2.3.9 Melanoma mutations exhibit strikingly different neighbour effects and spectra

Mutation processes in malignant melanoma are known to be distinctive and to include strand-asymmetric mutation processes within genes (Pleasance et al., 2010). Our analysis confirmed that the profile of point mutations in the malignant melanoma sample was strikingly different to that of germline mutations (see Tables A.12 and

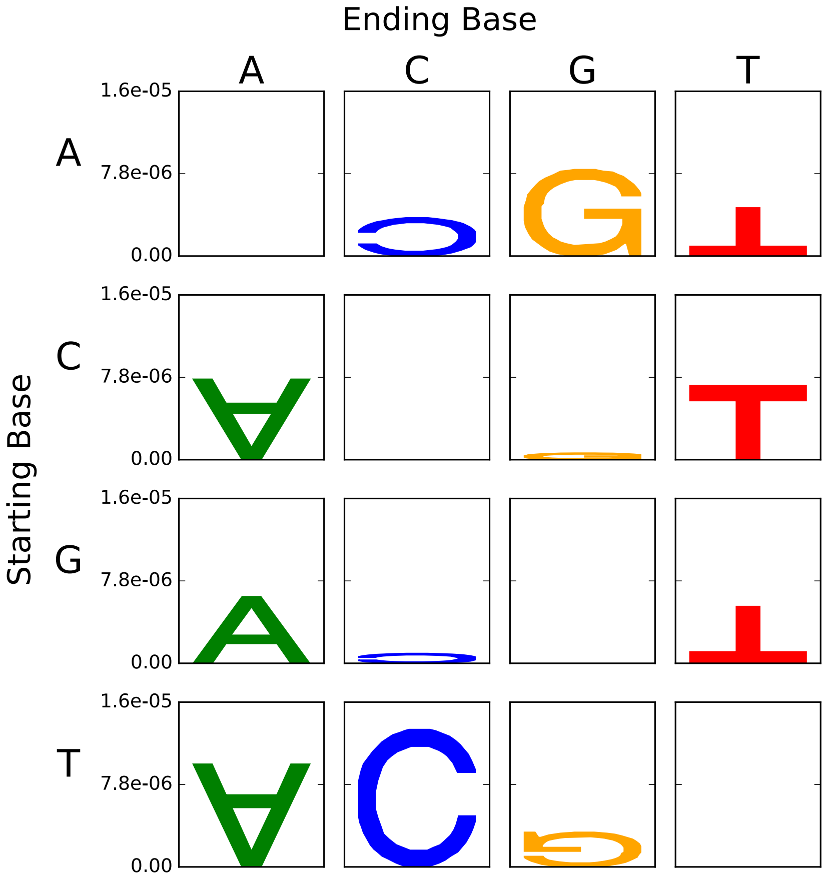


Figure 2.4: Significant differences in mutation spectra between autosomal and X-chromosomal intergenic sequence regions. Starting base, ending base correspond to X, Y respectively in X→Y. The y-axis is RE from the spectra hypothesis test and letter heights are as for the mutation motif logo. Letters in the normal orientation indicate an excess of that mutation direction in autosomal relative to the X-chromosomal mutations. Inverted letters indicate a deficit in autosomal relative to the X-chromosomal mutations.

A.13). The grid of all point mutations (see Figure 2.5) demonstrates that neighbouring influences were most pronounced for C→T point mutations and much stronger influence of neighbouring bases on transversion mutations. The neighbour effects were also significantly strand-asymmetric (see Table A.9), a distinctive characteristic of melanoma. Only substitutions affecting C were significantly different in spectra between strands with the C→T direction being over abundant on the + strand (see Figure 2.6 and Table A.10).

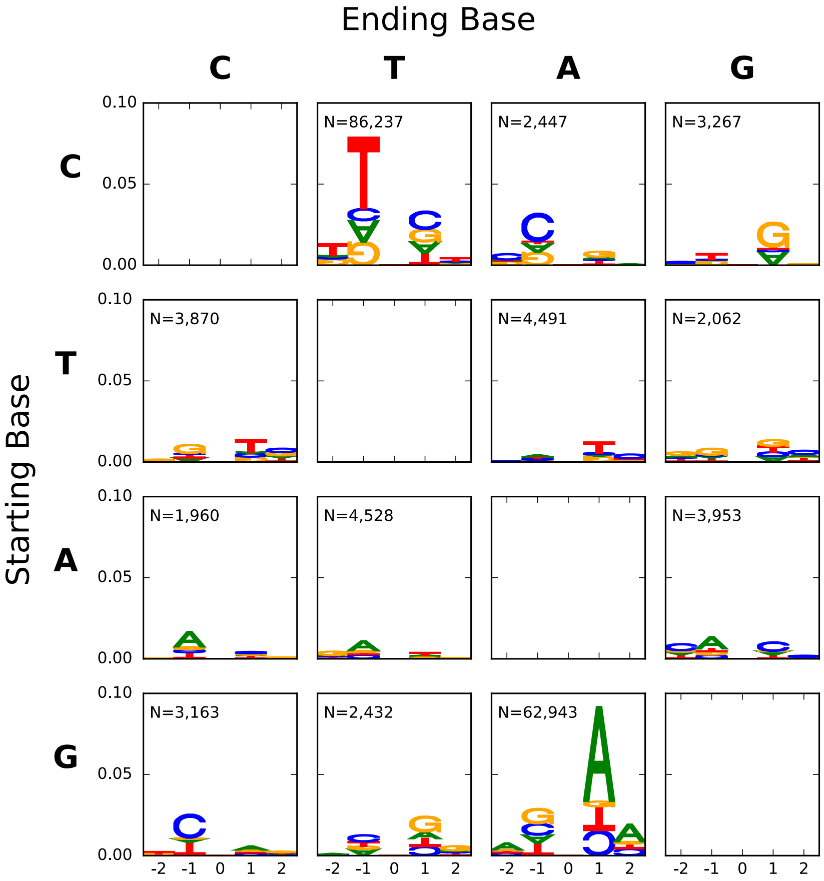


Figure 2.5: Panel of first-order effects from all 12 point mutations from the malignant melanoma sample. Starting base, ending base correspond to X, Y respectively in X→Y. The y-axis is RE and the x-axis is the position index relative to the mutated base. N refers to the number of variants from which the logo was derived.

2.4 Discussion

While it has long been appreciated that sequence neighbourhoods affect point mutations, statistical methods for disentangling how neighbours contribute have been

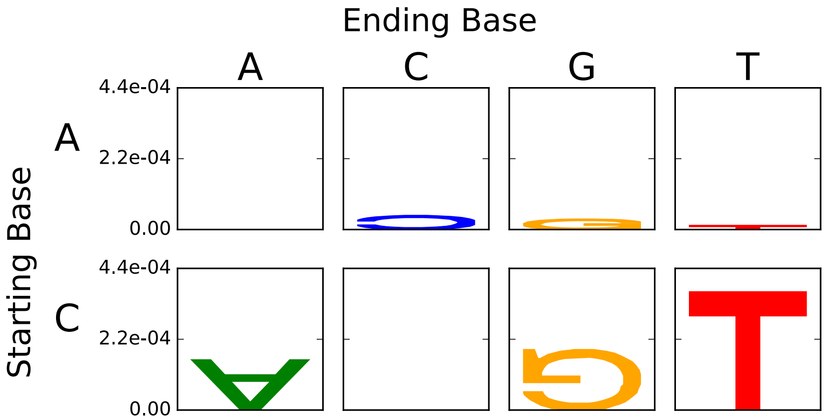


Figure 2.6: Strand asymmetry in malignant melanoma. Only mutations from C were statistically significant. Starting base, ending base correspond to X, Y respectively in X→Y. The y-axis is RE from the spectra hypothesis test and letter heights are as for the mutation motif logo. Letters in the normal orientation indicate an excess of that mutation direction on the + strand. Inverted letters indicate a deficit on the + strand.

limited. We addressed this using a novel determination of the reference distribution and log-linear models. This methodological combination is robust to complexity in the genomic background of nucleotide composition. It further enables hierarchical hypothesis testing for establishing the significance and relative importance of neighbour effects. We illustrated utility of the models by applying them to analyses of mutations from samples reported to exhibit distinctive properties. Our analyses recapitulated well-known effects in terms of neighbour dependence and in terms of differences between genomic regions and somatic and germline, supporting the accuracy of the methods. In addition, our results revealed previously unreported neighbour effects that extends beyond immediate flanking positions. Analyses of mutation spectra complemented the neighbour analyses, confirming known features of point mutations in malignant melanoma and identifying novel differences in germline point mutation abundance between sex-chromosomes and autosomes.

The hypermutability of C→T in CpG dinucleotides is the exemplar of context-dependent mutation and a gold standard that a method of analysis should be able to correctly recover. We established that the conventional sequence logo analysis approach did not recapitulate the dominant influence of a 3'-G (see Figure 2.7). As this method shares the assumption of equifrequent bases with that of (Krawczak et al., 1998), the failure suggests the Euclidean distance approach, which is based on the same assumption, will also be flawed. In contrast, as shown in Figure 2.2 and Table A.1, our analysis successfully recapitulated this known effect. The RE$_{max}$ values (see Figure 2.2B) further affirm C→T as most strongly affected by neighbouring bases.

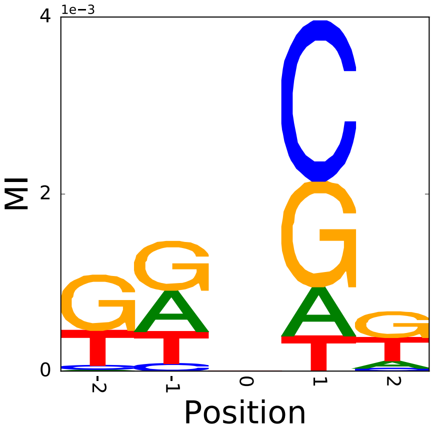


Figure 2.7: The CpG effect on C→T is not revealed by applying the conventional sequence logo method to autosomal intergenic mutations. MI is mutual information.

To sensibly interpret the results of our analyses, we de-emphasised the importance of statistical significance, and instead focused on effect magnitude. Due to the very large number of inferred mutations, our analyses possess considerable power to detect small effects. This is illustrated by the very small p-values associated with, for example, third-order effects for the C→T mutation (see Table A.1). Yet, the magnitude of these effects is relatively small in comparison with the first-order effects (see Figure 2.2A). Consequently, and in addition to considering whether effects are statistically significant according to standard criteria, we contrasted RE statistics to establish relative importance.

Our analysis identified numerous novel properties of neighbouring sequence influence on point mutation in the germline. First, all mutations were significantly affected by neighbouring bases, and transition mutations showed a larger neighbouring influence than those on transversions. Interestingly, as illustrated by the A→G* mutations, these influences did not decay monotonically with distance from the mutation (see Figure 2.3B). This point mutation further illustrated that multiple neighbouring positions can influence mutation outcome. Comparing RE values to those for C→T indicates that the first-order neighbour effects of other point mutations were $\sim 5-20$ fold less, with those values corresponding to A→G and A→C mutations respectively (see Table 2.1). Second, all mutations were significantly affected by higher-order effects (interactions between adjacent bases). These were evident in a manner such that bases contiguous with each other and the mutated location showed the largest RE. This may reflect the importance of interactions among adjacent bases (base-stacking) in affecting DNA stability (Kariin and Burge, 1995; Yakovchuk et al., 2006). For all point mutations, the RE terms from first-order effects were markedly stronger than those for higher order effects. These results were replicated in our analysis of intronic variants (see Table A.11).

The evidence for neighbouring influence on mutation raised the important question of how far these effects of flanking sequence extend? While our results showed strong statistical significance of positions as far as 10 bp from the mutating base (see Table A.3), the relative magnitude of RE values indicated a very rapid decay away from the mutated position. In particular, the magnitude of the effect decayed below an order of magnitude within two bases for transition mutations. This trend is illustrated by the mutation motif logo displays (see Figure A.4). Transversion mutations exhibited a slower decay in effect magnitude (therefore, they have a larger neighbourhood), and these reflect the smaller $\mathrm{RE}_{max}(1)$ of transversions that constitute a less stringent cut-off.

Our results regarding the importance of higher-order interactions indicate that considering 5-mers accounts for the majority of model fit. The deviances from the first-order effects of A→G* and C→T* transition mutations accounted for 81 and 98 per cent of the total deviance respectively in the autosomal intergenic sample.

Inclusion of second-order effects increased both these to $>$ 96% (see Tables A.1 and Table A.2). Across all point mutations in the autosomal intergenic sample, combining first and second-order effects accounted for a median 91 per cent of the total deviance of the 5-mer model. These differences are further illustrated by the motif [C/T]CAAT[C/G/T]N, reported as exhibiting an odds ratio of $\sim$ 6000 for enrichment in mutated sequences (Aggarwala and Voight, 2016). Our results (see Figure 2.3D and Table A.2) identified the CAAT core of this motif as highly significant. However, this is a third-order interaction and the RE for this specific combination of sites is 28-fold less than the strongest first-order effect and accounts for only 1.5 per cent of the total deviance. Our estimated odds ratio for the CAAT mutation motif was $\sim$ 4.0, which was less than the $\sim$ 5.7 odds ratio estimated for the 7-mer of Aggarwala and Voight (2016). (Our odds ratios are closer to what Aggarwala and Voight (2016) termed 'fold change'.)

The profile of somatic mutations is expected to exhibit differences to germline mutations due to requisite defects in DNA repair systems. As reported (Nik-Zainal et al., 2012), such defects are characteristic of cancers. Of the characterised cancers, malignant melanomas exhibit the most distinctive mutation signatures. Included in the distinctiveness of malignant melanoma is a striking strand asymmetry (Pleasance et al., 2010). This putatively derives from UV light-induced formation of pyrimidine dimers. In transcribed regions, NER processes, together with TCR mechanism, is able to repair transcribed strand lesions efficiently. As a consequence of this, mutations are expected to accumulate on the non-transcribed strand. Evidence supporting this, with more C→T mutations on the non-transcribed strand than on the transcribed strand, has been reported (Pleasance et al., 2010).

Our analysis demonstrated that point mutations in melanoma were dependent on neighbours in a manner strikingly different from that of germline processes discussed thus far (see Figure 2.5 and Table 2.2). While C→T mutations were again the point mutation most affected by neighbouring bases, the motif was markedly different to that of the germline process, with a 5'-T showing the greatest influence. This difference indicates that 5mC deamination plays a less prominent role in C→T in melanoma tissue. Since melanoma arises in part due to defect(s) in DNA repair, the distinctive mutation motifs in melanoma indicate either a very effective masking of neighbour effects on lesion formation, or the DNA repair mechanisms inactivated in melanoma are strongly affected by neighbours. Our melanoma analysis also strongly supported strand asymmetry of mutations, with the effect most pronounced for C→T.

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)	**RE**$_{max}$(3)	**Pos.**(3)
A→C	0.0167	-1	0.0101	(-1, +1)	0.0078	(-2, +1, +2)
A→G	0.0135	-1	0.0118	(-1, +1)	0.0051	(-1, +1, +2)
A→T	0.0110	-1	0.0039	(-2, +1)	0.0033	(-2, -1, +1)
C→A	0.0319	-1	0.0102	(-1, +1)	-	-
C→G	0.0264	+1	0.0035	(-1, +1)	0.0041	(-2, -1, +1)
C→T	0.0788	-1	0.0130	(-1, +1)	0.0006	(-2, -1, +1)
G→A	0.0918	+1	0.0090	(-1, +1)	0.0009	(-1, +1, +2)
G→C	0.0254	-1	0.0028	(-2, +1)	0.0043	(-1, +1, +2)
G→T	0.0242	+1	0.0078	(+1, +2)	0.0052	(-1, +1, +2)
T→A	0.0123	+1	0.0042	(+1, +2)	0.0044	(-1, +1, +2)
T→C	0.0135	+1	0.0244	(-1, +1)	0.0057	(-1, +1, +2)
T→G	0.0137	+1	0.0118	(-1, +1)	0.0074	(-2, +1, +2)

Table 2.2: Summary of neighbour associations with mutations in malignant melanoma. RE$_{max}$(#) is the maximum RE for order # and Pos.(#) the corresponding position(s). All point mutations had at least one significant test after correcting for 15 tests (see Table A.1) using the Holm-Šidák procedure. Non-significant results are indicated by '-'.

A major asset to the log-linear modelling framework is the ease of extension to enable comparisons between samples. The utility of this is illustrated above in comparing somatic to germline processes. The appeal of this capability, however, is much broader. It further allows evaluation of the processes that contribute to within-genome heterogeneity in sequence composition. We have illustrated this application by considering genomic regions for which the incidence of mutation processes are known to differ (X-chromosome versus autosomes) or where DNA repair processes are known to differ (transcribed versus untranscribed regions).

Since Haldane (Haldane, 1946, 1935, 1948), studies have identified a systematic tendency for mutations to originate in males. The most popular hypothesis to account for male biased evolution is the mutation-through-DNA-replication hypothesis (Li et al., 2002; Webster et al., 2005). Other, non-replication based, differences in mutation between the sexes have also been proposed (Huttley et al., 2000). Included in these is evidence for elevated methylation of DNA in the male germline. This suggests the relative contribution of 5mC derived lesions will be greater on the autosomes compared to X-chromosome, as the latter spends less time (on average) in males. Our analyses for comparing differences in neighbour influences support this suggestion by showing the existence of distinct 5mC-affecting mutation processes operating between the X-chromosome and autosomes (see Table A.5), including a reduced magnitude of the +1 influence on the X-chromosome. However, this was not the strongest difference in neighbour effect between chromosomal classes; A→G showed the strongest statistical significance while C→G showed the greatest RE. The spectra analyses further emphasised the importance of differences in A→G* point mutations (see Figure 2.4). These results therefore indicate more extensive point mutation differences between these chromosome classes than previously appreciated. Furthermore, they suggest a corresponding diversity in mutational processes between male and female germlines.

Due to localised influence of transcription-coupled DNA repair, differences in operation of DNA repair processes may affect mutation. This process is known to operate in a manner that is strand-asymmetric, and different between transcribed and non-transcribed strand. Differences in base parity – the frequency of A should equal to that of T; the frequency of G should equal to that of C – support an effect of transcription on point mutation (Touchon et al., 2003). Significant differences in neighbour effects for all point mutations were evident between intergenic and intron regions. However, our analysis of strand symmetry for neighbour effects was not significant for intron sequences for any point mutation. This suggests a distinctive mutation profile arising from transcription, rather than the influence of transcription-coupled DNA repair.

We have argued that the matched sampling of the reference distribution in our neighbour analysis is important. Briefly recapitulating that approach, the reference distribution is obtained by randomly selecting a paired reference base within ±300bp of each observed mutation (see Figure 2.1). An alternate to this strategy is to obtain the reference base by randomly selecting from the full genome sequence, instead of the ± 300 bp range. For a given point mutation direction, only the reference counts can differ between the ± 300 bp and genome reference approaches (*i.e.*, the observed counts are identical). Consequently, the statistical inferences will likely differ when the k-mer distribution for a sequence class differs from that of the entire genome. An obvious case in which this condition arises is protein-coding exons. A neighbour analysis of exon sequences where the reference distribution was obtained from the full genome sequence showed significant differences to the ± 300 bp one. The relative importance of each flanking position and/or the identity of bases at those positions differed for all point mutation directions (for a subset, see Figure A.5). These results, and their considerable computational advantages, support using the ± 300 bp reference distribution.

As formulated, the neighbour analyses do not evaluate the relative abundance of mutations between samples. For this purpose, we introduced what we termed the mutation spectrum analysis. As the opportunity for mutation is affected by the frequency of the starting base, and base frequency differs between genomic locations, we performed spectrum analysis for each nucleotide separately. The null hypothesis is a very simple one, that the three possible point mutations from a starting base occur in equal frequency between samples. As such, this spectrum approach does not consider neighbouring base contributions at all; therefore, it is complementary to it.

For each of the above analyses comparing groups, we also undertook mutation spectrum analyses. There were no significant strand differences for autosomal data.

Comparisons between the X-chromosome and autosomes revealed highly significant differences in composition for all bases (see Figure 2.4). The most pronounced difference was an excess of A→G* transition mutations on autosomes. Similarly, all point mutations showed significantly different mutation spectra between intergenic and intronic regions (see Table A.6). In this case, however, the dominant differences were an excess of transversions creating A/T base pairs in intergenic regions, and introns were characterised by an excess of C/G base pair creating mutations.

The methods we present enabled examine mutational processes affecting samples. For the neighbour analyses, the critical properties of our methods derive from the specification of the reference distribution, and utilisation of the well-established log-linear modelling framework. This combination has considerable potential for detailed interrogations of mutation properties, which should further improve understanding of the mechanism of mutations, both germline and somatic. Our application of the method generated mutation motifs consistent with well-known effects. We further revealed a pronounced influence of flanking bases on all point mutation processes. From germline mutations, we have identified a striking dependence of the A→G transition on multiple positions. The mechanistic basis of this mutation motif is unknown.

The neighbour and spectral analyses examined complementary aspects of mutational processes. The former examines the contribution of neighbouring bases to the mutation outcome from a starting base, and the latter considers the breakdown of mutations from a single base. While the p-values from the hypothesis tests are sensitive to sample size, a property that may be proportional to mutation rate, neither approach considers the rate of mutation explicitly.

As with all methods that seek to characterise data arising from unobserved processes, there are interpretation challenges. In both the neighbour and spectral analysis approaches, the data are a composite of mutation events with potentially diverse etiological histories. As a consequence, differences between samples will potentially reflect multiple mechanistic differences. Regardless of these issues, analyses that use measures of genetic distance, such as phylogenetics, cannot rationally rely on models of sequence divergence that assume mutations affect nucleotides independent of their neighbours. Instead, models that accommodate neighbour effects *e.g*, (Hwang and Green, 2004) to at least ± 2 positions will need to be developed in order to reasonably capture the neighbour influences described here.

Chapter 3

Machine Learning Techniques for Classifying the Mutagenic Origins of Point Mutations

Yicheng Zhu, Cheng Soon Ong, Gavin Huttley

For this chapter, the preprint is available in BioRxiv, and the manuscript is under revision for submission. The reference on BioRxiv is:

Zhu, Y., Cheng S. O., and Huttley, G. A. (2018). Machine Learning Techniques for Classifying the Mutagenic Origins of Point Mutations. *BioRxiv*, `https://www.biorxiv.org/content/early/2018/06/08/342618`.

MY CONTRIBUTIONS Professor Gavin Huttley and I conceived and designed the experiments. Dr Cheng Soon Ong provided technical guidance regarding tools for implementing the machine learning algorithms. During my PhD, I conducted the research, specified the statistical analyses, wrote the software, and drafted the manuscript. I note here that as a part of the process of revising this work for publication, the software I originally wrote was revised by Dr Gavin Huttley to extend its functionality. Those revisions were entirely based on my original code.

Acknowledgments We thank Ben Kaehler, Helmut Simon and Hua Ying for comments on versions of the manuscript.

Supervisor's signature: G. Huttley

ABSTRACT There is increasing interest in developing diagnostic signatures that discriminate between individual mutagenic mechanisms. Understanding these signatures can facilitate a range of applications that include identifying population-specific mutagenesis and resolving distinct mutation signatures in cancer samples. Analyses for these applications assume that mutagenic mechanisms have a unique relationship with neighbouring bases that allows them to be distinguished. Direct support for this assumption is limited to a small number of simple cases (*e.g.*, CpG hypermutability). In this study, we have directly evaluated whether the mechanistic origin of a point mutation can be resolved using only sequence context for a more complicated case. We contrasted mutations originating from the multitude of mutagenic processes that normally operate in the mouse germline with those induced by the potent mutagen N-ethyl-N-nitrosourea (ENU). The considerable overlap in the mutation spectra of these two samples make this a challenging problem. Employing the new, robust log-linear modelling method described in the previous chapter, we demonstrated that neighbouring bases contain information regarding point mutation direction that differs between the ENU-induced and spontaneous mutation classes. A logistic regression classifier proved to be substantially more powerful at discriminating between the different mutation classes than were the alternatives. Concordance between the feature set of the best classifier and information content analyses suggests our results can be generalised to other mutation classification problems. We concluded that machine learning can be used to build a practical classification tool to identify the mutation mechanism for individual genetic variants. Software implementing our approach is freely available under the BSD clause-3 licence at `https://github.com/HuttleyLab/mutationorigin`.

3.1 Motivation

In most catalogues of genetic variation, the data consist of variants derived from a mixture of mutagenic processes. Whether analysis of the genetic variants alone allows resolution of the causative mechanism for an individual genetic variant remains an open question. Instances of a singular etiological relationship between point mutation mechanism and flanking sequence are known for only a small number of relatively simple cases. From a biochemical perspective, it seems a reasonable conjecture that the sequence of neighbouring bases should affect mutagenic processes in general. In addition, a related conjecture that knowledge of neighbouring sequence is sufficient to identify the specific mutagenic origin also remains substantively unverified. Methods to discriminate between entire mutation spectra (*e.g.*, those characteristic of cancers) have been developed (Zhu et al., 2017). Methods to estimate the major components of these spectra have also been developed (Alexandrov et al., 2013; Shiraishi et al., 2015). As far as we are aware, there has not been a detailed examination of the relationship between a mutation mechanism and neighbouring bases, with a view to identifying mechanistic origins of individual variants. In this chapter, we employed machine learning methods to address this using a data set of point mutations of known origin. We limited our discussion and analyses to the 12 distinct single-nucleotide point mutations only.

In mammals, mutation processes exhibit considerable heterogeneity that manifests between genomic locations, cell types, disease states and clinical treatments. The within-genome heterogeneity of sequence composition is taken as an indicator of the heterogeneous operation of mutation processes operating in the germline; and multiple factors are implicated in driving this pattern (Hodgkinson and Eyre-Walker, 2011). These include factors that distinguish gametogenesis between the sexes (Huttley et al., 2000), and the localised operation of transcription-coupled DNA repair processes (Svejstrup, 2002). There is also considerable complexity in the origin of mutations affecting somatic tissues. Variation in mutagenesis distinguishes normal cell lineages, as evidenced by the biochemically specified somatic hypermutation that occurs in immune cells (Chahwan et al., 2012). The spectrum of mutations can be a distinctive feature of different cancers (Pleasance et al., 2010), such as the reported excess of $\mathrm{G} \rightarrow \mathrm{T}^*$ transversions (where * indicates a mutation direction and its strand complement) in smoking-associated lung cancer (Hainaut and Pfeifer, 2001). These distinctive mutation spectra may be indicative of tissue-specific exposure to exogenous mutagens; they may also reflect defects in specific DNA repair processes (Viel et al., 2017). In all these cases the catalogue of mutations arises from a mixture of different processes, making assignment of a specific cause to a single mutation challenging.

Germline heterogeneity in mutagenesis has been correlated with a number of genomic features and processes, including the abundance of G and C nucleotides (GC) and sexual dimorphism in gametogenesis. The primary explanation for the positive correlation with GC is that, it reflects a causal relationship with the recombination rate via the process of biased gene conversion (Hellmann et al., 2005; Hodgkinson and Eyre-Walker, 2011; Meunier and Duret, 2004). Differences between the sexes in the spectrum of point mutations leads to differences in GC between chromosomes, based on time spent in the male germline (Huttley et al., 2000).

The process of a mutation can be decomposed into two fundamental steps: lesion formation followed by a failure of DNA repair to reconstitute the original base pair. High exposure of cells to UV light, which elevates formation of dipyrimidine lesions, illustrates the role of lesion creation in mutagenesis (Pfeifer et al., 2005). The accumulation of defects in DNA mismatch repair genes, which contributes to development of colorectal cancer, illustrate the role of defective DNA repair (Viel et al., 2017). In both these cases, the rate at which the different point mutations occur can be affected, highlighting that different types of point mutation can have a common mechanistic origin. As systemic changes to the mutation process are a feature of cancer cells, a primary analysis focus in cancer biology has to been to resolve mutagenic signatures that characterise cancers (Alexandrov et al., 2013; Shiraishi et al., 2015). This work exploits the presumed relationship between point mutation processes and flanking DNA sequence.

The nucleotides flanking a mutated position contain information regarding the mutagenesis process responsible for the change. Hypermutability of the CpG dinucleotide illustrates the relationship between neighbouring bases and point mutation mechanisms. Association of a 3'-G with elevated C→T mutation rates derives from the binding preference of DNA methylases (Krawczak et al., 1998). These enzymes bind to this dinucleotide and modify C to 5mC (5-methyl-cytosine). The resulted 5mC exhibits a 10-fold increase in the spontaneous deamination rate. This effect is so pronounced and is the dominant cause of most C→T mutations in the genome (Zhu et al., 2017). The apparent simplicity of the relationship between C→T point mutations and flanking 3'-G nucleotides reflects the dominance of a single chemical process in creating lesions.

The sequence motifs associated with non-C→T point mutations are more complicated (Zhu et al., 2017), suggesting contributions from multiple mutagenesis mechanisms. In Chapter 2, it was shown from an analysis of millions of human germline mutations that more than one nucleotide at flanking positions was associated with the non-C→T point mutations (Zhu et al., 2017). This observation is consistent with multiple mutation mechanisms contributing to these point mutations. At present, the mechanistic basis underlying these mutation-associated sequence motifs (mutation motifs) remains unknown. Even in the case of cancer, the diversity of defects in DNA repair limit the understanding of possible mechanisms that may be responsible for a specific genetic variant.

The systematic use of mutagens in forward genetic screens provides an opportunity to develop an understanding of the relationship between neighbouring sequence and mutagenesis. ENU is a synthetic alkylating chemical widely employed in mutagenesis studies (Alvarez et al., 2003; Lee et al., 2012; Stottmann and Beier, 2014), causing new germline mutations at $\sim$ 100 times higher rate than the spontaneous mutation rate (Stottmann and Beier, 2014). Exposure to ENU can induce formation of a number of alkylation adducts, including N^1-adenine (e^1A), O^4-thymine (e^4T), O^2-thymine (e^2T), and O^2-cytosine (e^2C) (Noveroske et al., 2000; Shrivastav et al., 2010). If the DNA repair system fails in repairing these adducts, they are mispaired during DNA replication to a non-complementary nucleotide, resulting in a single base change mutation (Justice et al., 1999; Noveroske et al., 2000). The resulting ENU-induced mutations are dominated by $A \rightarrow G^*$ and $A \rightarrow T^*$ mutations, with rare reported occurrences of C → G* mutations (Takahasi et al., 2007).

Whether ENU mutagenesis induces mutations randomly with regard to flanking DNA sequence is debated (Barbaric et al., 2007; Bauer et al., 2015). The unique ENU-induced mutation spectra distribution described above has provided the basis for the ENU-induced variant filtering strategy (Andrews et al., 2012). For example, removing any C → G* transversions, leaves only genetic variants likely to be generated by ENU process; thus, these are candidates for novel phenotypes. This filtering strategy will be referred to as the naïve (classification) method, in which the mutation mechanism is assigned solely on the basis of mutation direction. The approach has high accuracy solely because of the excess of ENU-induced mutations. However, there remains a possibility of misclassification of mutation origin in these studies as some fraction of the point mutations labelled as ENU-induced will instead have originated from non-ENU mutagenesis. If sequence neighbourhood does affect mechanism, then mutation classification techniques that exploit this information should improve over the naïve method.

Machine learning techniques are well suited to the problem of sequence-based classification of samples (Ben-Hur et al., 2008; James et al., 2013). The goal of machine learning classification is to find a rule, based on observed object features, that can assign new objects to one of several classes (James et al., 2013; Sonnenburg, 2008). Machine learning techniques have been applied to a diverse array of sequence-based classification problems ranging from microbial taxon assignment (Bokulich et al., 2018) to the prediction of the position of nucleosomes in eukaryotic cells from ChIP-seq data (Peckham et al., 2007).

In this study, we evaluated whether sequence features can improve the performance of classifiers devised to discriminate between mutagen-induced and spontaneous point mutations in the mouse germline. We affirmed a highly significant influence of neighbouring nucleotides on ENU point mutations, and these associations differ from those evident in spontaneous mutations. Our results reveal that a combination of *k*-mer size and representation of second-order interactions among nucleotides was able to markedly improve classification performance in comparison to the naïve classifier approach. All scripts developed for this work are made available under an open-source licence.

3.2 Materials and Methods

3.2.1 Spontaneous and ENU-induced germline mutation data

We constructed the data set for mutation origin identification from Ensembl release 88 and an ENU variation database from the Australian Phenomics Facility. The number of variants per chromosome are reported in Table B.4 in Supplementary Information 2.

As outlined in the Sections 3.1, we used only the 12 different point mutations for this project and adopted the following notation. The mutation of base X into base Y is indicated by X→Y. Consistent with notations in Chapter 2, we denoted a point mutation and its strand complement using *. For instance, A→G* refers to both A→G and its strand complement T→C.

3.2.1.1 Mouse spontaneous germline variants

The germline spontaneous variant data was obtained from the Ensembl database using EnsemblDb3 (`http://ensembldb3.readthedocs.io`). For each genetic variant, we obtained the SNP name, genomic location, effect and alleles. Only biallelic SNPs were used. As the Ensembl database did not include mutation direction for mouse variants, we computed mutation direction using phylogenetic methods.

Inference of mutation direction was performed using ancestral sequence reconstruction (Yang et al., 1995). The genomic alignments of mouse protein-coding genes and their one-to-one orthologs from the rat and squirrel were sampled from Ensembl using EnsemblDb3. Checks were performed to ensure the obtained syntenic alignments could be used. Specifically, only mouse genetic variants in which the genomic alignment contained unambiguous bases for all species were retained. The genomic alignments were sliced to be centred on a genetic variant. We fitted the HKY85 substitution model (Hasegawa et al., 1985) by maximum likelihood using PyCogent3 (Knight et al., 2007, `http://cogent3.readthedocs.io`), and estimated the most likely base at the mouse variant locus for the common ancestor of mouse and rat. This ancestral base, which matched one of the reported mouse alleles, was taken as the starting base. This allowed inference of the mutation direction that produced the genetic variant.

A total 254,680 validated mouse germline spontaneous variants within protein-coding regions were sampled. These variant records were further separated into sub-categories according to mutation direction and chromosomal location (see Table B.4).

3.2.1.2 ENU variants

ENU induced variant data examined in this study were obtained from the Australian Phenomics Facility website (`https://pb.apf.edu.au/phenbank/download/`). In the database, each genetic variant record includes the variant identifier, genomic location, putative effect, reference base and variant base. Only synonymous and non-synonymous mutations in mouse exonic protein-coding regions were used for this study, because the mutagenised mice were only exon sequenced (Caruana et al., 2013). This resulted in 234,177 ENU-induced mutations. Summary details of ENU variant records regarding mutation direction and the chromosomal location are presented in Table B.4.

3.2.2 Association of neighbouring bases using log-linear modelling

We employed our previously published log-linear methods (Zhu et al., 2017) and corresponding MutationMotif software (`https://github.com/HuttleyLab/MutationMotif`) for evaluating the association of neighbouring nucleotides with spontaneous and

ENU-induced point mutations in the mouse. In summary, these methods allow statistical evaluation of the association between point mutations and bases at individual, or multiple, sequence positions. Furthermore, these methods allow comparisons between samples for these associations. The log-linear models operate by comparing the count of observed bases at a position in sequences for which the point mutation is known against a paired reference distribution of counts from unmutated sequences. The association of bases at a single position with point mutations is referred to as an independent effect, and the influence of bases at two or more positions are referred to as dependent effects. These tests were used to assess the null hypotheses that ENU-induced point mutations occur independently of neighbouring bases. We also tested the null that the neighbouring base effects were the same for ENU-induced and spontaneous point mutations.

Mutation motifs were visualised in a sequence logo style. The stack height in these figures corresponds to RE. Individual letter heights within a stack represent the relative magnitude of the residual from the log-linear model for that letter. Base(s) that are overabundant in mutated sequences are on top, with a normal orientation. Base(s) with letters rotated 180° are underrepresented in mutated sequences.

3.2.3 Prediction of mutation origins

A difference in the association of neighbouring bases with spontaneous and ENU-induced mouse point mutations provides a basis for using machine learning classifiers to predict mutation origin. We considered two scenarios for such analyses. In the first, two mutation classes are known in advance, allowing development of a discriminating function. In the second, we considered the case in which only one mutation class is known in advance, and we sought to identify mutations that were 'outliers' to this known class. Of the numerous alternative machine learning techniques that could be applied to the two-class problem, we employed logistic regression (LR) and Naïve Bayes (NB). We used LR because of its similarity to the log-linear modelling approach described in Chapter 2. NB was chosen, as it is methodologically quite different from LR and has been used extensively for sequence classification. For the one-class problem, we used a support vector machine (SVM). In all cases, we used the open source software library scikit-learn (Pedregosa et al., 2011).

3.2.3.1 Logistic regression

The parametric nature of LR facilitates mechanistic interpretation of the developed classifier (Prosperi et al., 2009; Wålinder, 2014). This is of particular interest in seeking to relate attributes of the biological data to classifier performance. LR is based on the logistic function (James et al., 2013) as shown in Equation 3.1. The response value of LR ranges from 0 to 1. In classification, the probability that an observation belongs to a certain mutation class (*e.g.*, ENU) is expressed in Equation 3.2. We classified mutation with feature sets M as originating by mutation class 1, if $Pr(Y = 1|M)$ is greater or equal to 0.5.

$$F(t) = \frac{1}{1 + e^{-t}}, \tag{3.1}$$

$$Pr(Y = \text{ENU}|M) = \frac{1}{1 + e^{-\beta M}}, \tag{3.2}$$

The approximate probability of a mutation with feature sets M, can be expressed as:

$$\pi_M = P(M) = Pr(\text{Origin} = \text{ENU}|M) \tag{3.3}$$

$P(M)$ ranges between 0 and 1, and the LR expression of $P(M)$ is:

$$logit(P(M)) = \log \frac{P(M)}{1 - P(M)} = (1, M^T)\beta \tag{3.4}$$

or

$$P(M) = \frac{\exp((1, M^T)\beta)}{1 + \exp((1, M^T)\beta)} \tag{3.5}$$

where M is the input feature sets vector of a mutation. β is a parameter weight vector describing how important each feature is. However, a large β may indicate that the associated feature is over-fitted. Also, according to Equation 3.5, we found that different settings of β value will lead to different prediction probability. We wanted our classifier to perform as accurately as possible. Therefore, we needed to find the optimal set of β that generates the maximum prediction probability without over-fitting feature weights. The $\ell 1$ norm ($\ell 1$) regularisation was performed to achieve this.

In this study, we used $\ell 1$ regularisation because it prunes out unneeded features by setting their associated weights to 0. This characteristic allows for better understanding of the contribution of each feature. Mathematically, $\ell 1$ regularised LR classifier by solving the following optimisation problem (Pedregosa et al., 2011):

$$\min_{\beta,C} \sum |\beta| + C \sum \log(exp(-P(M)(M^T\beta + c)) + 1) \tag{3.6}$$

where hyperparameter C is a positive constant that balances how much we should care about when fitting the training data compared to penalising large weights. C was tuned during the cross validation process to maximise the likelihood; the according estimates of β were stored for subsequent use in predicting mutation origin based on the selected feature set.

3.2.3.2 Naïve Bayes

NB classifiers are built upon the assumption of conditional independence of the predictive variables given the class. This assumption is typically violated. However, as our variant data were randomly sampled from different mice, the dependency between mutations is relatively low. Thus the NB classifier was expected to perform reasonably.

To learn information from training samples according defined feature sets, and predict origins of mutation with the NB classifier, similar to the LR classification, mutation data are ultimately represented as a vector of binary feature sets M, including mutation direction and the neighbourhood sequences. In an NB algorithm, the posterior probability a variable was ENU-induced given a feature set is calculated as:

$$Pr(\text{Origin} = 1|M) = \frac{p_{(\text{Origin}=1)} \times p_{(M|\text{Origin}=1)}}{p_{(\text{Origin}=1)} \times p_{(M|\text{Origin}=1)} + p_{(\text{Origin}=0)} \times p_{(M|\text{Origin}=0)}} \tag{3.7}$$

where Origin classes 1, 0 correspond to ENU-induced and spontaneous germline mutations respectively. This product examines all data in the training sample, where m_i represents feature vectors. If the resulting posterior probability is higher than a defined cut-off threshold, a mutation is classified as ENU-induced mutation; otherwise, it is considered a normal mouse germline mutation. To optimise $Pr(\text{Origin} = 1|M)$, key components $p_{(M|\text{Origin})}$ for each origin class (see Equation 3.7) is estimated by a smoothed version of maximum likelihood:

$$p_{(M|\text{Origin})} = \frac{N_{(\text{Origin}\cap m_i)} + \alpha}{N_{(\text{Origin})} + \alpha n} \tag{3.8}$$

where, for each origin class, $N_{(\text{Origin}\cap m_i)}$ is the frequency count of feature $m_i, m_i \in M$ appearing in a sample belonging to that particular origin class. Similarly, $N_{(\text{Origin})}$ is the frequency count of samples belonging to a particular origin class. α is the smoothing factor and the value of α is tuned during the cross validation process to optimise the result, while n is the number of features.

One of the main advantages of NB classifiers is that they are probabilistic models. In addition to predicting the class label of a point mutation, the probability of the class labels is also generated.

3.2.3.3 One-class classification using SVM

The LR classifier and NB classifiers are designed to solve the two-class situation. That is, they are designed to distinguish whether a mutation is a germline spontaneous mutation or an ENU-induced mutation. An interesting possibility that may arise in real studies is that the properties of an alternative mutation mechanism are unknown, but a well-characterised reference data set exists. In that case, we are interested in discovering whether a mutation is likely to be a member of the reference set. In the present case, the reference distribution corresponds to spontaneous germline point mutations, and we wish to determine whether it is possible to successfully identify the ENU-induced mutations.

To address this question, we employed a one-class SVM algorithm to identify whether a mutation is considered a spontaneous mutation given training data and a proposed feature set. The spontaneous mutations are now the target objects and are labelled as group +1. The ENU-induced mutations are outliers and are labelled as group −1. Training of the one-class classifier involves analysis of only spontaneous mutations to learn a classification boundary. To make the one-class SVM classifier results comparable to the LR classifier results, we adopted the linear kernel when constructing the classifier. Thus, there is the following decision function:

$$f(m) = sign(\sum_{i=1}^{n} \alpha_i K_{(m,m_i)} - \rho) \tag{3.9}$$

where $K_{(m,m_i)}$ denotes the linear kernel function which is a linear decision surface separating the class features. α_i are the Lagrange multipliers, and ρ are the parameters of the hyperplane. The classifiers are then applied to the test data to determine the mechanistic origin of a mutation according to the sign of $f(m)$. If $f(m) > 0$, a mutation is classified as a spontaneous mutation, otherwise it is classified as an ENU-induced mutation.

3.2.4 The feature sets employed for classification

The machine learning approaches require numerical representation of the data. The choices of features employed will affect the final performance of a classifier. If the feature is insufficient to describe a data sample, there is not enough information available for a classifier to learn the data structure well. Intuitively, increasing the number of non-correlated features typically increases classification performance. However, if too many features are selected, it is computationally expensive.

We explored four different types of features: mutation direction, independent neighbourhood effects, dependent neighbourhood effects, and GC%. Mutation direction, which is represented by M, is the point mutation direction (*e.g.*, C→T); and there are 12 possible point mutation directions. Independent effects, represented by I, are the influence of bases at flanking positions independent of the bases present at other positions. Dependent effects are indicated by nD, where n is the possible effect order. For example, a second-order dependent effect, represented by 2D, is the influence of the bases at two separate positions. For a 5-mer with the mutation at the central base, there are six possible pairs of second-order position combinations. The fully saturated feature set, represented by fully saturated (FS), contains the mutation direction and all possible independent, dependent features. We further employed a restriction on the dependent effect that the component positions were proximal to each other in the sequence (after excluding the mutated position). We represented this feature set variant using a 'p' suffix (*e.g.*, 2Dp). For a 5-mer, there are three 2Dp features. Each of these features are logical propositions that are represented by a one-hot encoding (see Table 3.1).

We further considered the percentage of G and C nucleotides (GC%) around a point mutation. This property was included, as a significant positive correlation exists between inferred mutation rate and GC% in mammals (Hodgkinson and Eyre-Walker, 2011). The GC% was numerical data obtained from 500 bp flanking sequences around a mutation (500 bp from each side).

(a) Example data

Feature	ENU	Spontaneous
Mutation direction	C→A	G→T
Pos -1	A	G
Pos +1	G	T

(b) One-hot encoded data

Feature	Value	Record 1	Record 2
Variant class	ENU	+1	-1
	Spontaneous	-1	+1
Mutation direction	A→C	-1	-1
	A→G	-1	-1
	A→T	-1	-1
	C→A	+1	-1
	C→G	-1	-1
	C→T	-1	-1
	G→A	-1	-1
	G→C	-1	-1
	G→T	-1	+1
	T→A	-1	-1
	T→C	-1	-1
	T→G	-1	-1
Independent effect, Pos -1	A	+1	-1
	C	-1	-1
	G	-1	+1
	T	-1	-1
Independent effect, Pos +1	A	-1	-1
	C	-1	-1
	G	+1	-1
	T	-1	+1

Table 3.1: One-hot encoding of two mutation records for analysis. (a) An example raw data set containing an ENU and a Spontaneous mutation record. For each record, 1 bp neighbouring bases on both sides are shown (*i.e.*, $k = 3$). Positions -1, +1 are the left and right flanking neighbouring positions respectively. (b) The one-hot encoding of the example data for a M+I classifier. In our notation, the feature 'mutation direction' corresponds to M and the features 'Pos' correspond to I. Within a feature, there are multiple possible values: 12 for the 'mutation direction' feature and four for each 'Pos' feature. For each record (column), only a single row within a feature can equal '+1'.

For feature sets that were strictly categorical, genetic variant data were encoded with the one-hot encoding scheme. One-hot encoding is a process by which categorical variables are converted into binary variables, which is a form that could be provided to machine learning algorithms to do predictions. In this study, we used a $\{+1, -1\}$ encoding for binary features, where $+1$ indicates that the logical proposition is true, and -1 indicates that the logical proposition is false. The application of this process is illustrated, for a small example, in Table 3.1. In this example, the first record was derived from ENU-mutagenised mice. The feature 'variant class' is assigned $+1$ for the ENU value, and -1 for the spontaneous value. This process continues such that for a single record, only one of the possible values of a feature can be assigned $+1$.

As the GC% feature is not categorical, a different numerical representation was employed. The mutation direction features are categorical features, and labelled as $+1$ if true, or -1 if not true. Conversely, the GC% feature is a numerical feature requiring a numerical representation of average GC percentage in neighbouring sequences around a mutation; it ranges from 0 to 100 per cent. As the range of values of raw data varies widely, the proposed classifier may not work properly without normalisation. During a normalisation application, the different numerical scales of GC% and the one-hot encoded categorical feature values were adjusted to a notionally common scale. This leads to these different features having approximately the same effect in the computation of similarity (Aksoy and Haralick, 2001). We used the scikit-learn StandardScaler to obtain a scalar for a normalised transformation of the training data. The scalar derived from the training set was also used to normalise the test data.

3.2.5 Machine learning experimental design

Multiple factors can influence the performance of a classifier. These include the choices regarding the algorithm, the values of associated hyperparameters and the feature set to be used for classifying. In addition, there are design considerations

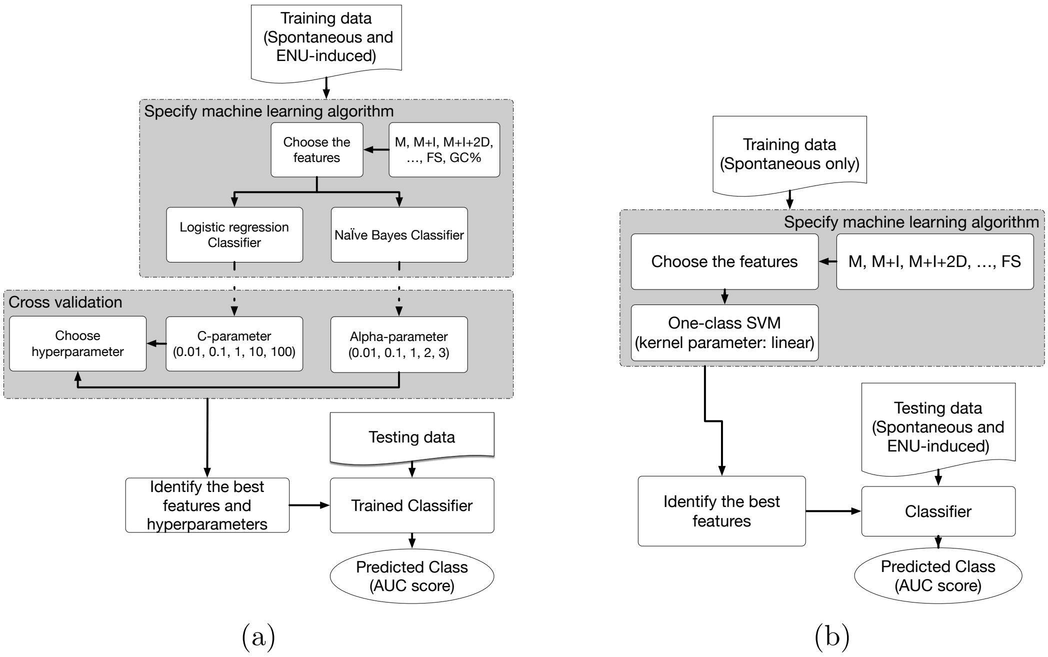

Training data (Spontaneous and ENU-induced)
Specify machine learning algorithm
Choose the features
M, M+I, M+I+2D, ..., FS, GC%
Logistic regression Classifier
Naïve Bayes Classifier
Cross validation
Choose hyperparameter
C-parameter (0.01, 0.1, 1, 10, 100)
Alpha-parameter (0.01, 0.1, 1, 2, 3)
Testing data
Identify the best features and hyperparameters
Trained Classifier
Predicted Class (AUC score)
(a)
Training data (Spontaneous only)
Specify machine learning algorithm
Choose the features
M, M+I, M+I+2D, ..., FS
One-class SVM (kernel parameter: linear)
Testing data (Spontaneous and ENU-induced)
Identify the best features
Classifier
Predicted Class (AUC score)
(b)

Figure 3.1: Overview of classifier algorithm evaluation. (a) Two-class classification includes labelled spontaneous and ENU-induced germline point mutations in the training data. (b) One-class classification includes only spontaneous germline point mutations in the training data. For both approaches, training data were limited to mutations occurring on mouse chromosome 1.

concerning selection of data for training and subsequent testing. The processes employed in this study, for both the one-class and two-class classification problems are illustrated for LR, NB and one-class (OC) in Figure 3.1. Our core algorithm choices are described above. Our experimental design involved training classifiers on data derived from mouse chromosome 1 only. For each algorithm, we used cross validation to tune the hyperparameters and optimise the classifier. For every cross validation iteration, we first performed a random train-test split, and divided datasets into training data and testing data. Then inside the training data, we further split training data to actual training data and validation data (see Figure 3.2). We trained the classifier on training data, set hyperparameters on validation data and finally evaluated classification performance on testing data. Within each validation process, we compared algorithm performance with different hyperparameter values; the hyperparameter generating the best performance for the available data was saved. For each classification experiment, this process was repeated five times.

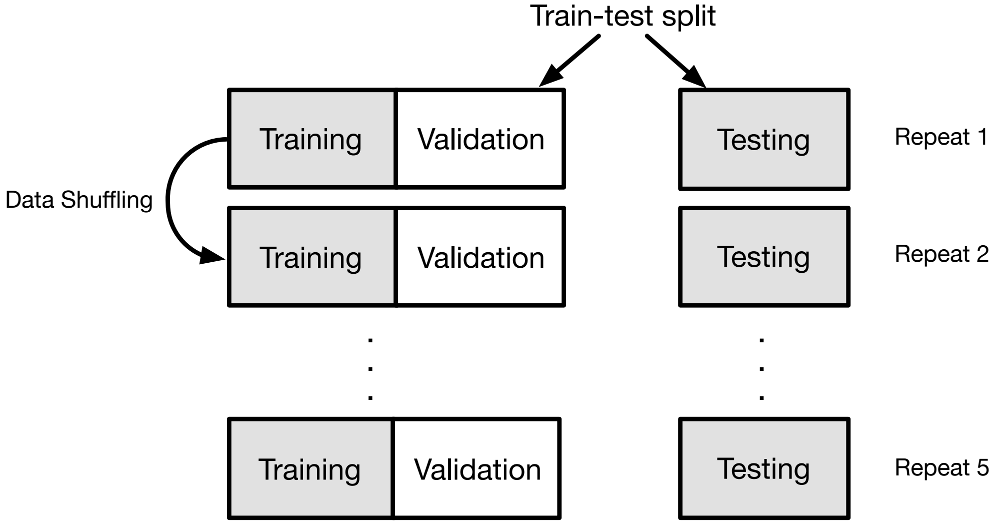


Figure 3.2: Procedure of cross validation. For each cross validation iteration, the data were shuffled and divided into three segments: one for training, one for validation and one for testing. For each experiment, performance of algorithms with different hyperparameters was compared. The best algorithm for the available data was saved. The process was repeated five times.

For the LR classification, the hyperparameter C is the trade-off regularisation parameter that trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims to classify all training examples correctly by giving the model freedom to select more samples as support vectors. We considered candidate C options from the log-scale of: $0.01, 0.1, 1, 10, 100$. The C value that resulted in the best performance was chosen for all subsequent analyses.

For the NB classification, the hyperparameter α is the Laplace parameter used to smooth categorical data. We considered candidate alpha options of: $0.01, 0.1, 1, 2, 3$. The value of α that resulted in the best performance was chosen for all subsequent analyses.

3.2.5.1 Classifier performance evaluation

We evaluated classifier performance using the area under the receiver operating characteristic curve (AUC). The AUC is an estimate of the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. An advantage of using the AUC score as the performance measure is that the score does not require choice of a cutoff threshold. Some other binary classification algorithms compute a series of performance scores (*e.g.*, accuracy, sensitivity, and specificity) and classify based upon whether the score is above a certain threshold. Therefore, as the choice of threshold is of particular importance in these scoring schemes. Shifting of the threshold may dramatically alter the score, and thus, the performance of a classifier. An AUC score has the advantage of illustrating the trade-off between sensitivity and specificity for all possible thresholds rather than just the threshold chosen by the modelling technique. The AUC scores of the different experiments were reported and we interpreted a larger AUC score as indicating better classification performance.

3.2.5.2 The effect of increasing the number of examples during training

The whole classification process is achieved by implementing training and testing phases. In the training phase, a set of data and their respective labels are used to build a classification model. In the test phase, the trained classifier is used to predict new cases. Overlap sampling between training and testing data will make the prediction performance of a classifier overly optimistic because of the overfitting problem. To avoid the overfitting situation, for each experiment, to start with, both ENU-induced and mouse germline mutations should be split into two non-overlapping sets for training and testing from the outset.

The accuracy of a classifier improves with the number of observations used to train the algorithm. This improvement tends to be rapid initially. When the training size is sufficient to a point, the improvement decreases gradually. The ‘learning curve’ is used to describe this phenomenon and estimate the number of samples needed to train a particular classifier to achieve its optimal accuracy (Mukherjee et al., 2003). To plot learning curves and find the desired training size, after selecting a specific classifier and set of features, we used progressively larger samples of observations to train the classifier and plot accuracy performance against the number of training observations.

3.2.5.3 Availability of data and materials

The pre-processed data used in this study are available at Zenodo `https://zenodo.org/record/1204695` under the Creative Commons Attribution-Share Alike licence. Data files are typically gzip compressed standard formats (*e.g.*, tab delimited text files, FASTA formatted sequence files). The source code for a command line application is made available under the BSD clause-3 licence at `https://github.com/HuttleyLab/MutationMotif`. The scripts used to perform the data sampling and analyses reported in this work are freely available at Zenodo `https://zenodo.org/record/3497585`.

3.3 Results

3.3.1 Distinctions between ENU-induced and spontaneous point mutations

A logical requirement for using sequence features to discriminate between samples is that those features differ in abundance between ENU-induced and spontaneous point mutations. We addressed this using two complementary formal hypotheses tests. The ‘spectra’ hypothesis test compares the distribution of point mutation outcomes in the two source materials (see Section 2.2.4). The ‘neighbourhood’ hypothesis test contrasts the association of neighbouring bases with those point mutation outcomes (see Section 2.2.3). In both cases, ENU-induced germline point mutations were obtained from the Australian Phenomics Facility, and spontaneous germline mutations from the Ensembl database.

We employed a log-linear model to test the null of equivalence in mutation spectra between the ENU-induced and spontaneous samples (see Section 2.2.4) (Zhu et al., 2017). This test considers the relative distribution of outcomes from mutations of, for example, the base T. A separate test was employed for each possible starting base. Consistent with Chapter 2, the spectra of ENU-induced and spontaneous point mutations in the mouse were significantly different (see Figure B.1 and Table B.1 in Supplementary information 2). To simplify the following, and as stated in the Section 3.1, we abbreviated the description of a point mutation and its strand complement using the notation X→Y* (*i.e.*, A→G* refers to both A→G and its strand complement T→C). Direct examination of counts for the ENU-induced mutations revealed they were dominated by A→G* and A→T* mutations, with frequencies of 42 and 27 per cent respectively. These contrast with their abundance in mouse spontaneous mutations of 29 and 3.7 per cent respectively. Visualisation of the spectrum analyses (see Figure B.1) reflects these changes in proportion. These differences affirm the basis for the current naïve mutation classification algorithms applied to ENU samples.

The striking difference in mutation spectra was also accompanied by striking differences in the magnitude and identity of neighbouring base influences. The log-linear model is employed here to test the null hypothesis of no relationship between position bases and mutation source groups (see Section 2.2.3). We used position indices relative to the point mutation location, defined as position 0, with negative/positive indices representing 5'-/3'- positions respectively. Consider, for example, the question of whether bases at the 3' position immediately to a point mutation of A→G associate with the mutation. The test assessed the null hypothesis that in sequences in which an A→G mutation occurred, the base counts at the +1 position were equivalent to those at the +1 position for occurrences of A in the reference distribution. This is an example of a single-position (first-order), or independent (denoted I in the modelling notation) effect. We were also able to evaluate whether the joint counts of bases at two positions were equal between the mutated and reference sequence collections (second-order dependence, or 2D). Our previous analyses of spontaneous germline mutations from humans identified neighbour effects as highly influential, and affirmed that independent and second-order effects dominated higher-order effects (Zhu et al., 2017). These analyses are readily extended to comparing equivalence between samples, as is the objective here.

Our analyses established that there were strongly significant differences between the ENU-induced and spontaneous mutations in the identity of the associated mutation motifs and their relative magnitude. To simplify the exposition, we limited this discussion here to a description of the results from the A→G* case, the most abundant ENU-induced point mutation. (All point mutations exhibited strongly significant differences and are summarised in Table B.2 in Supplementary information 2). The maximum RE association of independent positions with A→G was 5-fold larger in the ENU-induced sample. This maximum association was at +1 in the ENU-induced sample, compared with -1 for the spontaneous sample (see Figure 3.3). Using the log-linear model, we rejected the null hypothesis of the equivalence between ENU-induced and spontaneous samples for neighbouring base associations with A→G mutations. While these samples revealed highly significant differences for nearly all effects orders (see Table 3.2), the magnitude of difference was greatest for the I and 2D effects (see Figure B.2). As mentioned previously, these patterns held true for all point mutation directions (see Table B.2).

Position(s)	**Deviance**	**df**	***p*-value**
+2	88.6	3	4.4×10^{-19}
-2	1105.6	3	0.0
+1	1393.7	3	0.0
-1	5693.3	3	0.0
(-2, +2)	12.0	9	0.2145
(-1, +2)	50.3	9	9.4×10^{-8}
(+1, +2)	96.1	9	9.5×10^{-17}
(-2, +1)	123.0	9	3.3×10^{-22}
(-2, -1)	284.1	9	6.2×10^{-56}
(-1, +1)	353.1	9	1.3×10^{-70}
(-2, -1, +2)	41.2	27	0.0396
(-1, +1, +2)	46.9	27	0.0100
(-2, +1, +2)	55.1	27	0.0011
(-2, -1, +1)	62.2	27	0.0001
(-2, -1, +1, +2)	118.6	81	0.0042

Table 3.2: Log-linear analysis comparing neighbour associations between mouse germline and ENU-induced A→G mutations. Deviance is from the log-linear model, with df degree-of-freedom and corresponding *p*-value obtained from the χ^2 distribution. *p*-values below 0.025 were determined to be significant.

Of further relevance to feature selection for classifier design is the physical limit

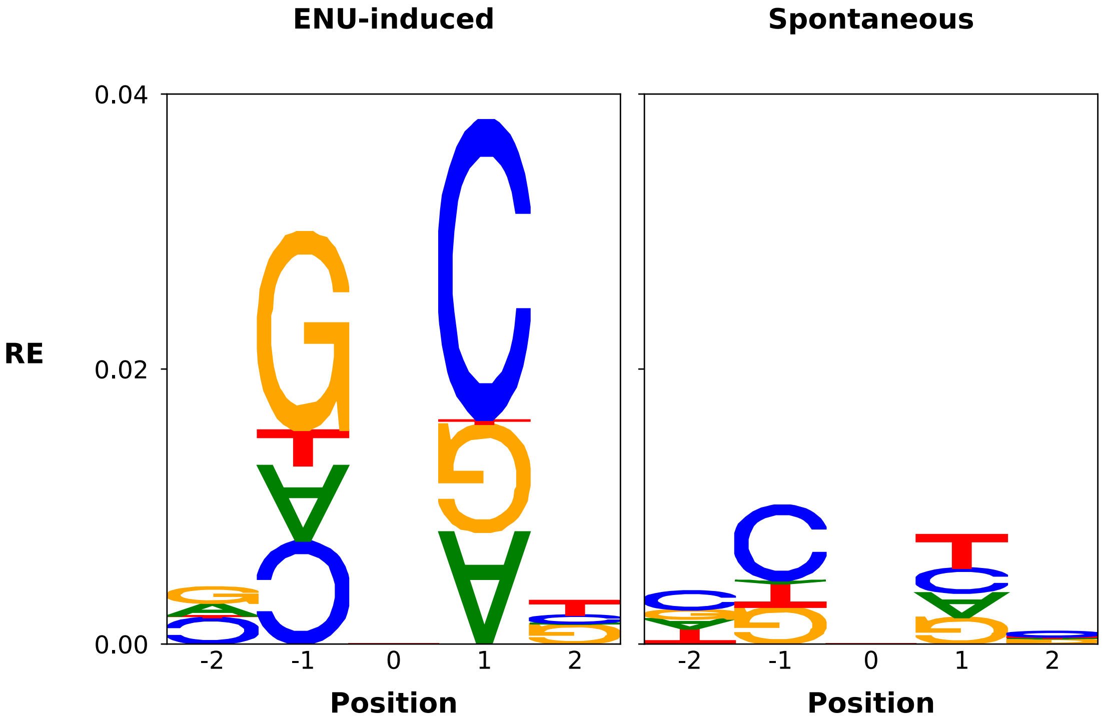


Figure 3.3: Neighbouring base associations significantly differ between ENU-induced and spontaneous germline A→G mutations. Position is relative to the point mutation at position 0. RE is relative entropy, derived from the deviance of the log-linear model (Zhu et al., 2017). Letter height is proportional to the RE term for that base. Normally oriented (180°-rotated) letters represent bases that are positively (negatively) associated with the point mutation. See Section 3.2 for greater detail.

to these associations. Estimation of the physical limit of association from longer flanking contexts was obtained using RE as per Zhu et al. (2017) (see Figure B.3 and Table B.3). The ENU-induced sample showed the physical limit mean, median and standard deviation of 3.2 bp, 2 bp and 1.7 bp respectively. In contrast, the corresponding statistics for spontaneous mutations were 2.9 bp, 2.5 bp and 2 bp. As a consequence of this variability, we considered a range of different neighbourhood sizes in development of the classifiers.

3.3.2 Development of a two-class machine learning classifier

In developing classifiers, we evaluated a collection of algorithms, sample sizes, sequence feature sets, k-mer size and hyperparameter values. Classifier development was strictly limited to data from a single mouse chromosome. We arbitrarily chose chromosome 1, given availability of sufficient data (see Table B.4). We have presented only the LR classifier results in this section. LR was chosen because of its systematically better performance over the NB classifiers, and the interpretability of the resultant classifiers. The NB results are in supplementary information 2.

Classifier performance was measured as the AUC score. For any particular classifier, its performance was measured using the mean AUC ($\overline{\text{AUC}}$) and standard error derived from five replicated AUC measures obtained from the cross validation analysis. A classifier whose $\overline{\text{AUC}}$ score was greater than that of another classifier was taken to be superior, after considering the standard errors.

In the following, we describe the classifier feature sets using a combination of the terms M, I, 2D, FS and GC%. These terms correspond to the mutation direction (M), the set of contributions from independent flanking positions only (I), and the set of contributions arising from all of the possible two-way dependent effects among flanking positions (2D). The FS model contains M and all possible independent and multi-position interactions. (In the regressions, the exact values for the I and D terms depend on the value of k.) In the case of $k = 3$, the FS model is equivalent to model M+I+2D. For a classifier considering a larger k, the FS model would include terms considering higher-order dependent effects contributions. The GC% corresponds to the percentage of G+C nucleotides in flanking DNA sequence.

For LR, we made choices regarding two hyperparameters. ℓ_1 regularisation was chosen as it prunes out unneeded features by setting their associated weights to 0 (Bühlmann and Van De Geer, 2011). This allowed me to establish the features that contribute to the classification. The regularisation parameter C controls overfitting by affecting the trade-off between variance and bias of regression parameter estimates. We selected the value of C that returned the best classifier performance on the validation set.

Comparison of training curves resulting from classifier evaluation indicated that

M+I+2D provided robust performance. The learning curves showed the sensitivity of the classifier performance to training set size, where the training set size is the total of both ENU-induced and spontaneous classes. For the categorical feature sets, we considered four distinct models: M, M+I, M+I+2D and FS. Figure 3.4 indicates that when training size is $> 4,000$ samples, the rate of classifier performance improvement with increasing sample size drops off markedly. For subsequent comparisons, we used classifiers trained on data sets with $\sim$ 16,000 samples as their standard errors allowed greater resolution between feature sets. Of the classifiers that only included categorical features, the naïve classifier employed for classifying ENU-induced mutations, M, was the least accurate. Inclusion of individual position features, represented by I, provided a substantial improvement over M. The best performing classifiers, however, included features representing dependence among positions (see Table B.5). The overlap in standard errors of the $\overline{\text{AUC}}$ for the M+I+2D and FS models (see Figure 3.4) indicate that inclusion of two-way dependence captured most of information contained by the sequence neighbourhood. The value of C that returned maximal performance was consistently 0.1 for all models and all samples that considered higher-order interactions (*i.e.*, 2D and above).

3.3.2.1 Choosing neighbourhood size

As illustrated by the log-linear analyses reported above, the physical limit of neighbouring base influence differs between point mutations and mutation mechanism (see Figure B.3 and Table B.2). Recalling that a symmetric neighbourhood size of 3 equates to $k = 7$, we initially assessed the impact of sequence neighbourhood size by comparing the performance of three different k-mer sizes (3, 5, 7) for the M+I and M+I+2D feature sets. Comparison of learning curves established that for training set sizes $>$ 4k, classifiers based on a 7-mer context performed better than the other two values of k (see Figure 3.5) (For detailed AUC statistics, please see Tables B.6, B.7 and B.5). The impact of choice of k differed between feature sets, with the strongest improvements with increasing k evident for the M+I+2D model.

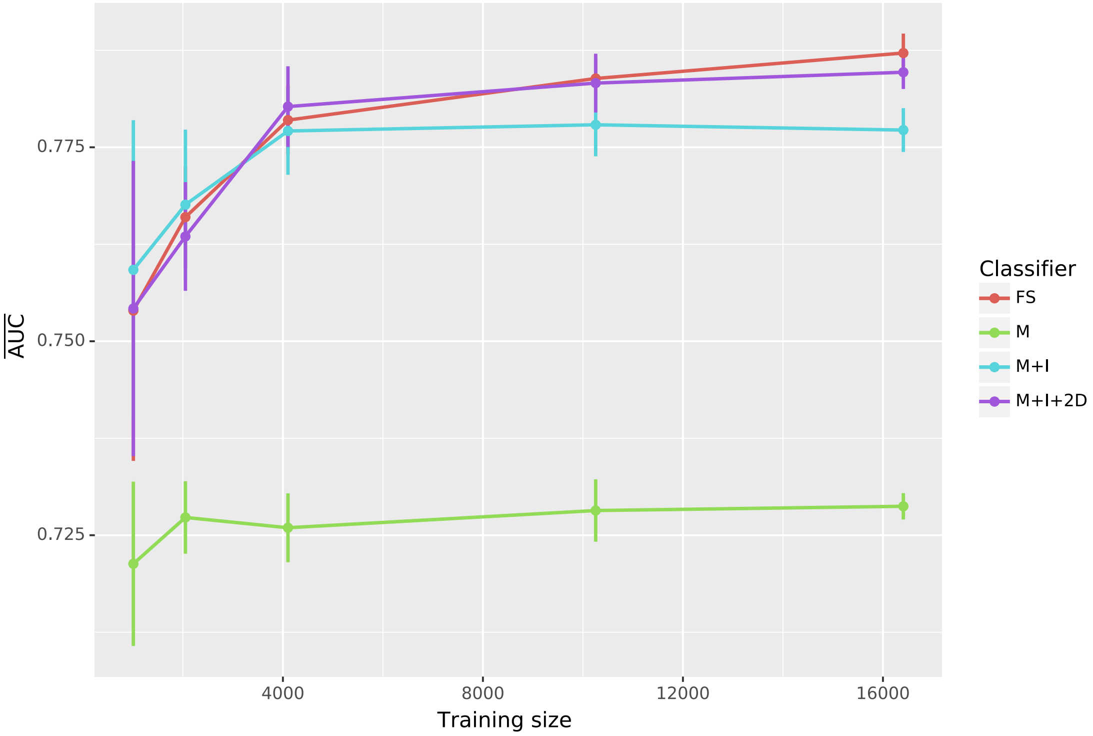


Figure 3.4: Model M+I+2D was sufficient for classifying mutations. Learning curves from training data are shown for four proposed classification models from 7-mers: M, M+I, M+I+2D and FS. The mean ($\overline{\text{AUC}}$) and standard error were calculated from the 5 chromosome 1 training samples. See the text for an explanation of model notation.

These results suggested the need for exploration of larger k. Initial efforts at modest k failed due to excessive memory requirements as the number of 2D parameters increases. Based on the log-linear results presented here and previously (Zhu et al., 2017), which indicate that most information is captured by proximal positions, we considered just the M+I and M+I+2Dp feature sets for $k > 7$. The results reinforced choice of the M+I+2Dp feature set and identified $k = 61$ as an upper limit (see Figure 3.6).

3.3.2.2 Incorporating GC% feature did not improve the classification performance

As described in the Section 3.1, the existence of a correlation between sequence GC% and mutation processes in mammals has been known for some time. Therefore, we considered whether inclusion of GC% as a feature would improve classifier performance. GC% was estimated from ± 500 bp flanking each mutation. Only

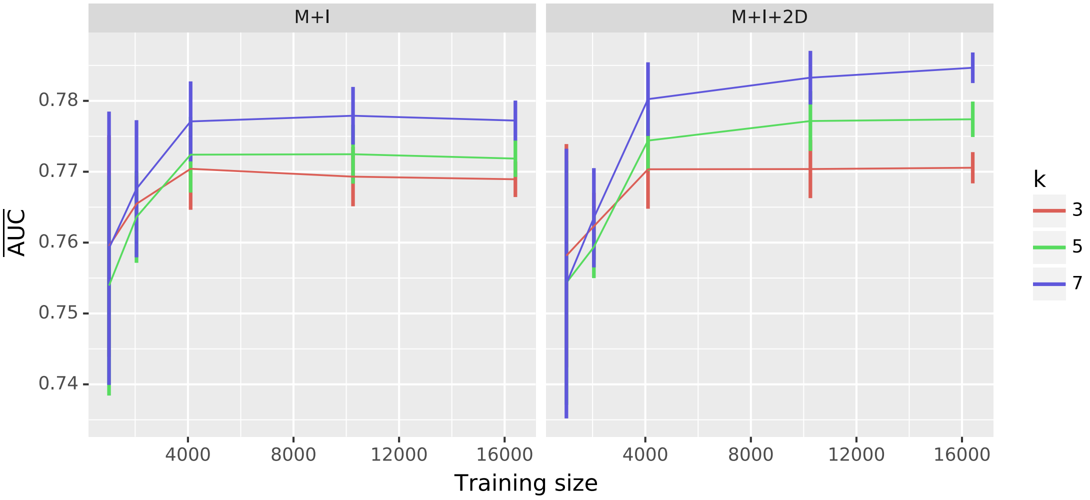


Figure 3.5: Classifier learning curves indicated increasing performance with k. The influence of k-mer choice on learning curves is shown for models M+I and M+I+2D. Plot titles indicate the model evaluated. $\overline{\text{AUC}}$ and the standard error were computed as described in Figure 3.4.

the naïve classifier (M) performance was improved by inclusion of the GC% feature (see Figure B.4). The impact on classifiers containing sequence features ranged from no effect (M+I) to substantially worse (FS). We speculate that the improvement of M+GC% over the M feature set arose because the GC% term indirectly measures the base composition of the immediate neighbourhood captured by the I term.

3.3.2.3 Applying classifier to whole genome

From the classifier development process described above, we selected the LR classifier with $k = 7$, M+I+2D feature set, and hyperparameters ℓ_1, $C = 0.1$ trained on the $\sim 16,000$ data sample from chromosome 1. We then applied this classifier to all mouse point mutations and display the results by chromosome in Figure 3.7. The vertical axis is the AUC score for all chromosomes except chromosome 1 where, because it was used for training, it is the average AUC across the five different cross-validation samples. With a mean and standard deviation of the chromosome AUC scores of 0.8 and 0.01 respectively, the LR M+I+2D classifier has a relatively good performance across the entire genome. Interestingly, the classifier performed worst with chromosome 1 data. The discrepancy of classification accuracy for chromosome 1 was observed when we re-run the analysis with chromosome 2 used for training

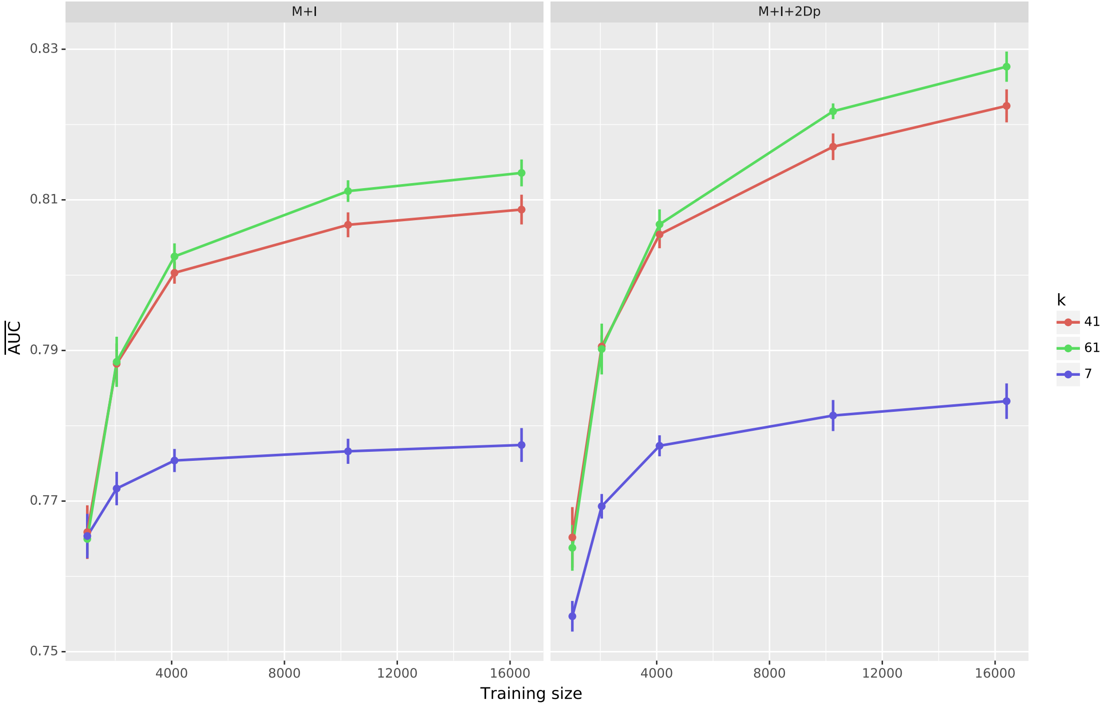


Figure 3.6: Large k and proximal 2D feature sets substantially improved classifier performance. Plot titles indicate the model evaluated. $\overline{\text{AUC}}$ and the standard error were computed as described in Figure 3.4.

(see Figure 3.8).

3.3.2.4 Performance of the OC classifier was substantially worse

We sought to evaluate whether the mutation motifs associated with spontaneous mutations were sufficiently distinctive to allow a machine learning algorithm to effectively identify non-spontaneous mutations. This corresponds to an outlier analysis. We tackled this using an OC Support Vector Machine (SVM). The same feature set choices were considered as for the LR models in a 7-mer context. As shown in Figure 3.9, the M+I+2D feature set demonstrated the best performance. However, all OC classifiers had much lower $\overline{\text{AUC}}$ than even the simplest two-class classifier (M). Furthermore, the OC M+I+2D classifier applied to the entire genome exhibited a systematically lower AUC, compared to the LR classifier (see Figure 3.10).

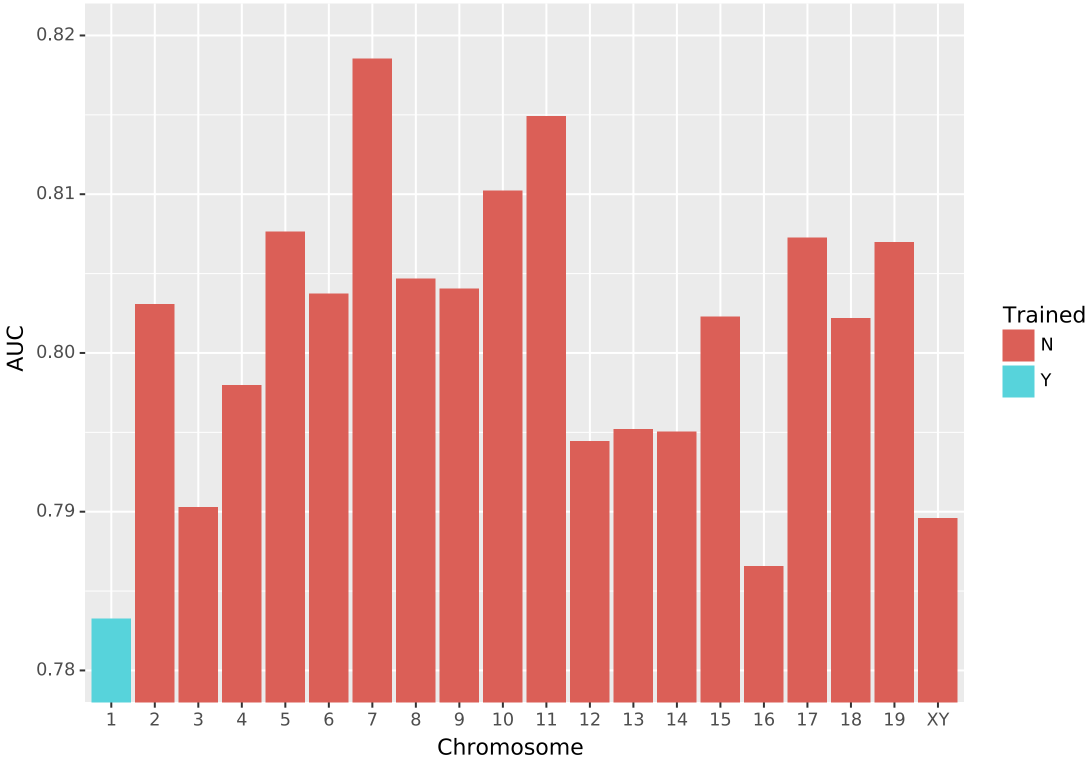


Figure 3.7: Per chromosome classification performance on the mouse genome of the best LR classifier. The classifier was trained on $\sim 16,000$ mutations from chromosome 1 using a 7-mer M+I+2D feature set. The $\overline{AUC}$ score obtained after applying the trained classifier to the remaining mutations (not used for training) from the chromosome 1 is represented by the blue bar.

3.4 Discussion

We have sought to establish the extent to which the etiological relationship between flanking sequence and mutagenesis can be used to identify the mechanism via which individual point mutations originate. Genetic variants in the mouse arising from application of ENU, a potent chemical mutagen, were contrasted with those arising spontaneously. We show that ENU-induced point mutations are strongly associated with neighbouring bases in a manner that differs to their spontaneous counterparts. A two-class classifier performed markedly better to the current standard technique for identifying ENU mutations, and was robust to genomic sequence attributes that have previously been shown to affect mutation processes. Our examination of the potential for machine learning based on the single category of spontaneous germline mutations revealed substantial challenges remain to resolving this more general case.

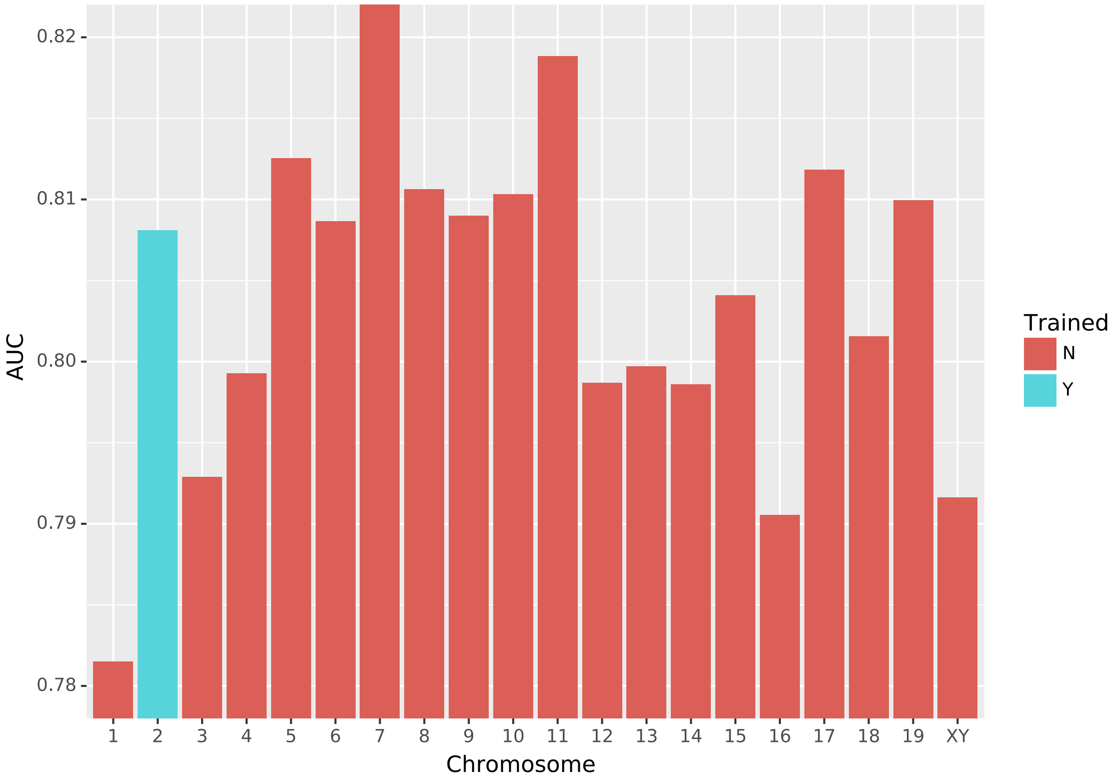


Figure 3.8: Per chromosome classification performance on the mouse genome of the best LR classifier. The classifier was trained on $\sim$ 19,000 mutations from chromosome 2 using a 7-mer M+I+2D feature set. The $\overline{AUC}$ score obtained after applying the trained classifier to the remaining mutations (not used for training) from the chromosome 2 is represented by the blue bar.

Comparison of the mutation spectra between spontaneous and ENU-induced germline mutations supported previous conclusions. The spectral analysis compared the breakdown of single-base mutations from ENU-induced and spontaneous mutations. The proportions of A→G* and A→T* mutations were substantially increased ~1.5 fold and ~7.5 fold respectively, in the ENU-induced and spontaneous sample. These observations are consistent with previous reports (Barbaric et al., 2007; Justice et al., 1999; Noveroske et al., 2000; Takahasi et al., 2007). The abundance of A→G* point mutations in both the ENU-induced and spontaneous samples underscores the challenge of using mutation direction alone for classifying mechanistic origin, as well as the likelihood that such an approach will be error-prone.

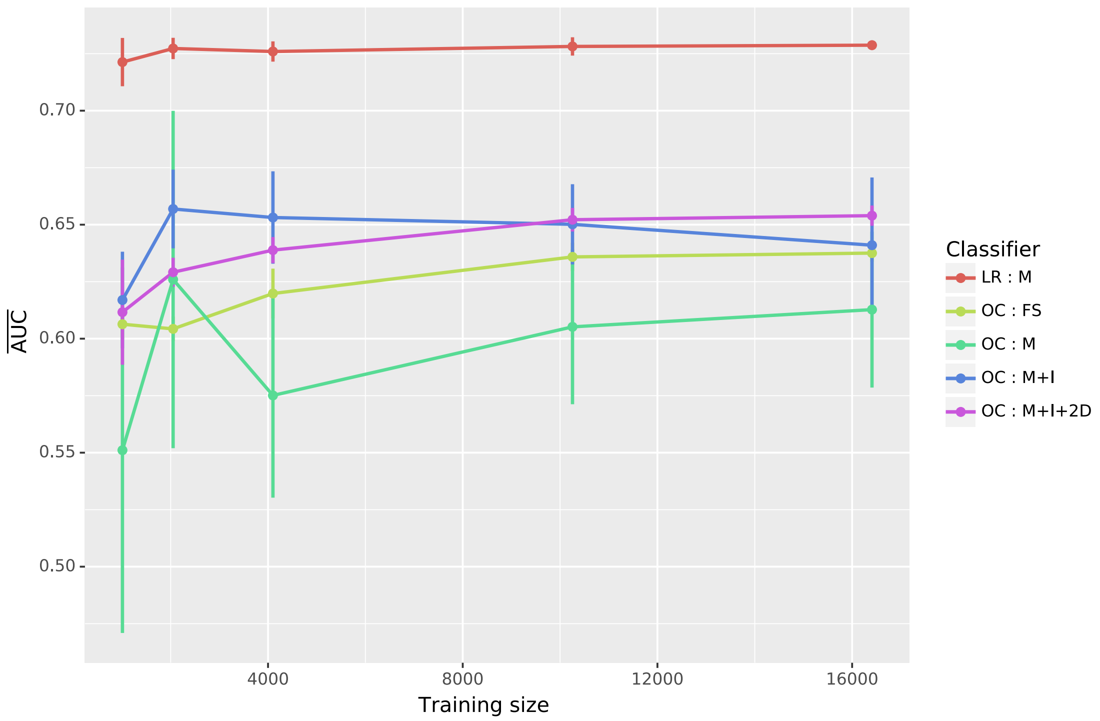


Figure 3.9: The OC SVM classifier performed worse than all LR classifiers. X-axis is the size of the training sample; y-axis is the $\overline{\text{AUC}}$ and standard errors were calculated as per Figure 3.4.

Our analyses established that the DNA sequence flanking ENU-induced mutations contains distinctive information. After correcting for multiple hypothesis tests (Holm, 1979), highly significant associations between neighbouring bases and point mutations were found for the ENU-induced sample, along with highly significant differences in neighbourhood between the ENU-induced and spontaneous mutations. As ENU induces an elevated rate of DNA lesion formation, it seems plausible that these differing neighbouring base associations reflect that mutagenic chemistry. Alternately, they may derive from operation of different DNA repair processes to those typically active in the germline (Noveroske et al., 2000; Shrivastav et al., 2010; Takahasi et al., 2007). In addition to the independent neighbourhood effects, all ENU-induced mutations were found to be significantly associated with higher-order effects. Similar to what was observed from humans (Zhu et al., 2017), the higher-order effects on ENU-induced mutations were evident in a manner such that bases at physically contiguous positions showed the largest RE (see Figure B.2). The latter may reflect the importance of base stacking on helix stability (Yakovchuk et al., 2006)

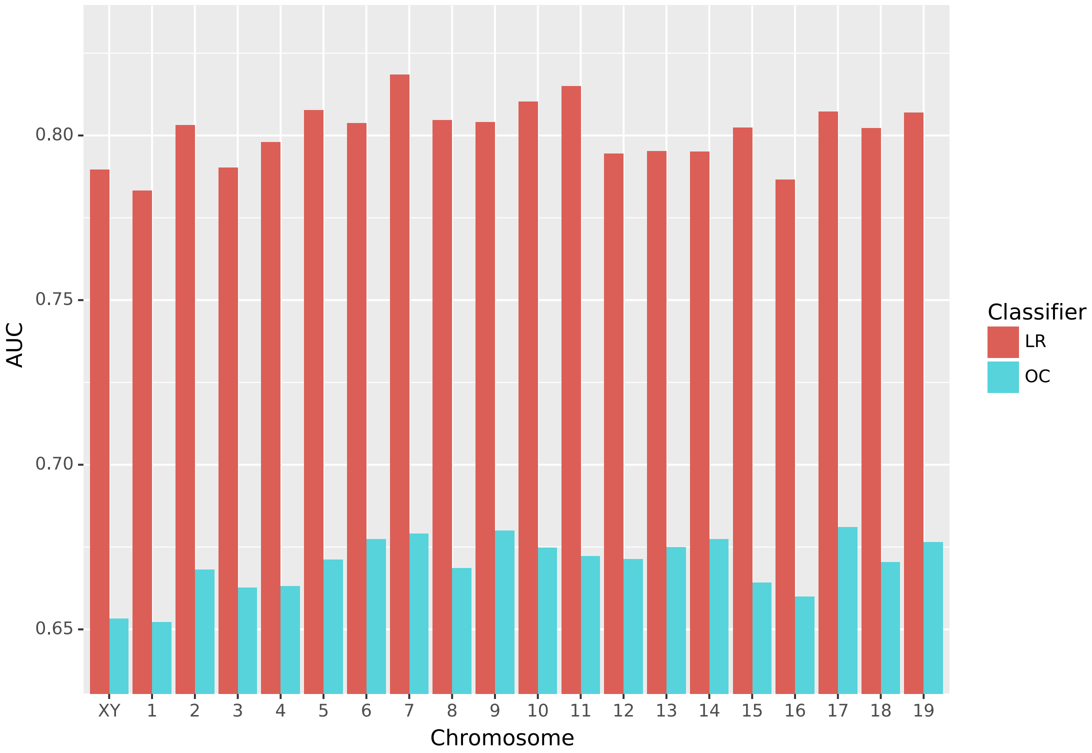


Figure 3.10: The OC SVM classifier performed worse than the LR classifier on the entire genome. LR—logistic regression, and OC—one-class Support Vector Machine classifier. The classifiers were developed on 7-mers with the M+I+2D feature set.

Our analyses of the influence of sequence neighbourhood on ENU-induced point mutations clarified previous reports. Barbaric et al. (2007) found a significant enrichment of base G or base C at one of the two most immediate flanking positions. Their measurement encompassed all 12 mutation types. Therefore, we cannot resolve whether this was a systemic influence of ENU, or one related to a specific point mutation. However, our analyses identified this specific pair of neighbouring bases as significantly associated with ENU-induced A→G*. Our results contradict the claim, by Bauer et al. (2015), that there were no neighbouring base influences. It is also worth noting that those authors did not formally test this hypothesis.

A succinct LR model was capable of strong performance, even when trained on a small fraction of the total data. The current standard naïve classifier, model M, represents the baseline performance. M considers only mutation direction and ignores sequence neighbourhood entirely. The performance ($\overline{\text{AUC}}$) of the M+I+2D feature set on the training data from mouse chromosome 1 was ~ 7 per cent better than that of M. In addition, the FS model exhibited comparable performance as that of M (see Figure 3.4). This observation indicates that including dependent effects with order > 2 confers little benefit to classification performance. This observation is consistent with the results from the log-linear analyses, which showed a small residual deviance after fitting the I+2D model (see Figure B.2).

The GC% statistic, previously correlated with mutation processes in mammals, was determined to be a crude surrogate of more explicit neighbourhood features. GC% is a sequence composition summary statistic. Inclusion of this feature in the classifier only improved the M model. In all other cases, it had no effect or it reduced classifier performance (see Figure B.4). This result emphasises the mechanistic role of individual bases, as reflected by the mutation motifs (see Figure 3.3), rather than a more general property (*e.g.*, the local DNA melting point) of a sequence region.

Application of the developed LR classifier to the whole genome produced a greater performance than was observed on the training chromosome. We evaluated classifier performance on a per-chromosome basis to facilitate evaluation of whether a relationship existed between classifier performance and the distinctive *k*-mer distributions reported for mammal sex-chromosomes (Huttley et al., 2000). The AUC from the combined sex-chromosomes lay within the range of AUC scores from the autosomes, indicating the discriminatory resolution of the classifier was robust to such differences. The observation that both the two-class and one-class classifiers returned their lowest AUC for chromosome 1 data (the part that had not been used for training) is puzzling (see Figures 3.7 and 3.10). The consistency between these very different machine learning algorithms suggests that mouse chromosome 1 presents a particularly challenging case for classification. The basis for this remains unknown.

It is worth noting that the LR classifier is trained using relatively balanced data that is, the number of ENU and germline mutations were comparable in the data set. This design reflects our interest in understanding the sequence factors that affect classifier performance, rather than the specific objective of delivering a classifier for studies employing ENU. In such studies, the mutation classes will be highly imbalanced because they are expected to have many more ENU than spontaneous mutations (up to 100-fold excess). This attribute needs a different trade-off between false positive and false negative predictions from the classifier. There are several extensions to this work that may be useful when a practitioner attempts the class imbalanced task. The first is to consider using a performance metric that is less sensitive to class imbalance (Davis and Goadrich, 2006). The second is to extend the learning method to manage class imbalance during training, either using resampling methods or cost-sensitive methods (Haixiang et al., 2017).

An OC classifier would provide a mean for generic identification of mutations that do not match a designated reference sample. For instance, a forward genetics screen employing ENU where spontaneous mutations are rare. While the outcome of feature selection identified the feature set M+I+2D as the best performing OC classifier, the $\overline{\text{AUC}}$ from the genome was 0.67. This is significantly better than a random guess, but lower than the $\overline{\text{AUC}}$ of ∼84% from the two-class classifier performance. This discrepancy in performance likely reflects the overlap between sequence features of the ENU-induced and spontaneous mouse germline mutations. Since the OC models are trained only on one sample, they are extremely sensitive to irrelevant neighbourhoods, compared to the two-class classifiers. In other words, the presence of ‘noise’ makes it difficult to identify neighbourhoods that are unique to the positive class. Furthermore, the one-hot encoding scheme for OC classification produces a sparse table for the sample size, which can reduce classification performance.

Both the choice of k and the corresponding feature set can improve the results obtained here. For values of k in (3, 5, 7), we considered the full set of alternative feature sets (*i.e.*, M, I, and all possible dependent interaction terms). Classifier performance increased with the value of k. There was a trade-off between classifier performance and memory usage with choice of k. This precluded extending k for the comprehensive feature set comparison. However, we did consider the simpler M+I model for much larger values. The results indicate that additional, potentially quite substantial, gains in performance may be attainable. Learning curve analysis of the M+I model for $k = 61$ returned $\overline{\mathrm{AUC}} = 0.81$ (see Figure B.5). Inclusion of 2D terms was precluded by memory issues. A potential solution to this arose from restricting the D features to those for physically adjacent positions. Our log-linear analyses revealed the strongest information content exists among dependently interacting positions that are physically adjacent with each other and/or the mutating position. Incorporating this into feature selection could significantly improve classifier performance for both the one- and two-class classifier problems.

The performance of the NB classifier was also considered. Generally, it is poorer than LR classifier (see Figure B.6). There have been systematic examinations of differences between LR and NB classifiers (Ng and Jordan, 2002). These differences are due to the different structural assumptions used by the classifiers. LR is a discriminative classifier, and it directly estimates the conditional probability of interest. NB is a generative classifier, estimating both the prior and likelihood before using them to estimate the posterior probability of interest. The design choice of estimating the likelihood makes NB more sensitive to data that violate the Gaussian noise assumption. Therefore, when the underlying data do not exhibit Gaussian noise, LR classifiers have lower asymptotic errors than NB. In addition, if training sizes are relatively large, LR performs better than NB classifiers (Ng and Jordan, 2002).

Our results have established the utility of including a representation of sequence neighbourhoods in classifiers for resolving point mutation origins. Questions remain as to why large k should be so informative, when the analysis of information content of neighbouring bases revealed quite a restrictive limit (see Figure B.3 and Zhu et al., 2017). Perhaps, as speculated previously (Bauer et al., 2015), this reflects broader sequence features correlated with open chromatin status during spermatogenesis. Irrespective of biological mechanism, the marked improvement in classifier performance suggests that further improvements are possible.

We have demonstrated that neighbouring positions can be used to classify the mechanistic origins of mutations using machine learning techniques. The LR classifier can be expressed in relation to the log-linear models; and this relationship allowed dissection of the contribution level between different positions. However, the classifier features used here were mainly designed for two classes. Although we used them for the OC classification as well, and the performance was superior to random guessing, the best customisation of feature selection for the OC classifier remains unresolved. Furthermore, we restricted our consideration to only neighbourhood sizes up to $k = 7$. This reflected a practical barrier to examining larger k due to the feature table becoming too large, and requiring too much memory. Introducing a kernel to the classifier design may be a potential solution to examine a much larger k in a computationally efficient way. Kernel functions can be developed to examine the weighted contributions of different features as a whole, without explicitly computing the feature vectors.

Chapter 4

Is Biased Gene Conversion Responsible for GC Heterogeneity?

Abstract

Concentrations of G+C nucleotides in mammal genomes are referred to as isochores. The currently favoured hypothesis for the origin of isochores is that they derive from the operation of GC-biased gene conversion (gBGC). The proposed mechanism underpinning the gBGC hypothesis is the preferential repair of hetero-duplexes towards the GC base pairs during meiotic recombination. Specific predictions can be made under the gBGC hypothesis. Firstly, given the correspondence between recombination rate and conversion rate, there should be a difference in the mutation spectra between high- and low-recombination regions. Secondly, as the gBGC hypothesis seeks to explain the overabundance of GC nucleotides, it implies that it is operating most frequently in regions with high GC. Thus, the mutation spectra should differ between genomic regions with different GC levels. Thirdly, as common variants are more likely to have experienced gBGC than rare variants, I can expect that the relative abundance of GC-generating mutations will be greatest in common variants compared to the rare variants. These predictions are amenable to interrogation using the statistical methods presented in Chapter 2. For the first two hypotheses, the data were stratified with respect to upper and lower quartiles. Mutations were classified into high-recombination and low-recombination classes according to their

standard recombination rate obtained from the deCODE 2010 recombination map; into high-GC and low-GC classes according to their neighbourhood GC-percentage; and into common and rare classes according to their allele frequencies. Because of the small number of rare genetic variants, a combined analysis, in which data were not subsampled with respect to either recombination rate or GC content, was performed. Based on these data, no significant association was detected between mutation abundance and recombination rate, and between mutation abundance and GC content. However, there was a difference in the mutation spectra of rare and common genetic variants. The A→G* mutations were in strong excess in common variants, and A→T* mutations in strong deficit in common variants. In conclusion, my analyses do not support the relationship between recombination and gBGC events, nor the relationship between sequence GC content and gBGC events. However, there was support for the existence of some mechanism driving increasing GC content, whose influence likely accumulates through time.

4.1 Motivation

Both experimental and evolutionary studies suggest that nucleotide composition heterogeneity is a striking feature of vertebrate genomes. Nucleotide composition in vertebrate genomes is characterised by 'isochores' which are long regions (≥ 300 kb) of relatively homogeneous GC content (Bernardi, 2001; Eyre-Walker and Hurst, 2001; Paces et al., 2004). Although the overall base composition can vary dramatically between genomic locations, the base composition is relatively homogeneous within each isochore component (Bernardi et al., 1985; Eyre-Walker and Hurst, 2001; Paces et al., 2004).

Numerous studies have been done to explore the potential mechanistic basis of isochores (Eyre-Walker and Hurst, 2001; Mugal et al., 2015; Paces et al., 2004; Šmarda et al., 2014). Studies found that the GC content is correlated with many other genomic attributes. These include methylation pattern (Jabbari and Bernardi, 1998), recombination rates (Kong et al., 2002), replication timing (Watanabe et al., 2002), intron length (Duret et al., 1995) and gene density (Consortium et al., 2001). Among these correlations with isochores, the most striking one is that with meiotic recombination (Eyre-Walker and Hurst, 2001; Montoya-Burgos et al., 2003). This has motivated the suggestion that recombination might be responsible for the evolution of base composition.

To address the conjectured relationship between GC content and recombination, I first discuss meiotic recombination. In eukaryotes, most homologous recombinations are initiated by double strand break (DSB), an event initiated by a Zn-finger protein, PRDM9 (Billings et al., 2013; Mugal et al., 2015). PRDM9 recognises specific binding motifs in DNA sequence that are 13 to 17 bp-long (Myers et al., 2010). The zinc fingers of PRDM9 bind to the DNA sequence and trimethylate lysine 4 and lysine 36 of histone H3 (hereafter H3K4me3 and H3K36me3 respectively, see Powers et al., 2016) at nucleosomes next to the binding motif. H3K4me3 and H3K36me3 act like markers for enzymes to recognise and to catalyse meiotic DSB formation. Recombination then takes place to repair the DSB damage. Immediately proceeding recombination, the homologous chromosomes pair up and a heteroduplex complex is formed between them. Then chromosome material is exchanged between the pairs of sister chromatids. This is an opportunity for gene conversion to occur (Szostak et al., 1983). Gene conversion is a process that involves the unidirectional transfer of genetic material between homologous sequences (Chen et al., 2007). It operates on any genetic variant that distinguish the two sister chromatids, a difference that will manifest as a mismatch in the new helices. The DNA MMR system fixes these recombination generated mismatches in a non-random manner which favours certain alleles. This is termed biased gene conversion, or BGC (Friedberg et al., 2005; Spies and Fishel, 2015). In summary, recombination events can be associated with gene conversion events. Changes in the rate of recombination are therefore expected to change the rate of gene conversion.

Several pieces of evidence have been interpreted as supporting a causal relationship between recombination and the evolution of GC content in the mammalian genome. Montoya-Burgos et al. (2003) discovered that the GC content increases after translocation of the mice *Fxy* gene from X chromosome (with low recombination rate) into a pseudoautosomal region (PAR) (with high recombination rate). This was interpreted as resulting from a causal relationship between recombination events and sequence GC content. Kong et al. (2002) in their study discovered a positive correlation between the GC content of genomic DNA and the local rate of crossover. This positive correlation between recombination rate and GC content was revealed by multiple genomic studies in yeast, Drosophila and nematode (Birdsell, 2002; Gerton et al., 2000; Marais et al., 2001; Marais and Piganeau, 2002). Meunier and Duret (2004) found that GC content correlates more strongly with female crossover rates, than with male crossover rates in human. As the average rate of crossover is $\sim$ 1.6 times higher in the autosomes of females than in those of males (Bhérer et al., 2017; Kong et al., 2002), the results of Meunier and Duret (2004) supports the relationship between recombination and GC content. It is also worth noting that these reports were done using estimates of recombination rate predate publication of the high resolution deCode recombination map (Kong et al., 2002).

BGC is considered as an important evolutionary phenomenon that acts like natural selection. BGC is essentially a preferential propagation of one of the genetic variants segregating at a locus into the next generation (Nagylaki, 1983). The theoretical work of Nagylaki (1983) demonstrated the similarity to natural selection. A selectionist interpretation of GC content is that the sequence composition has evolved under direct selective pressure that favours high GC in some genomic regions. However, the fact that there is high GC content in non-coding regions contradicts this model, making it very unlikely.

The GC-biased gene conversion (gBGC) model is a well argued model explaining how, mechanistically, recombination might influence the GC content (Galtier et al., 2001; Hellmann et al., 2005; Hodgkinson and Eyre-Walker, 2011; Meunier and Duret, 2004). This model asserts that GC content is generated via the process of gBGC during recombination (Duret et al., 2002; Webster et al., 2003). gBGC is one type of BGC which favours the accumulation of GC base pairs during meiotic recombination (Galtier et al., 2001). During gBGC, the repair by MMR of a heterozygous mismatch preferentially favours GC gametes over AT gametes. Brown and Jiricny (1989) reported that repair of base-base mismatches, which arise during DNA replication in human cells, favours the retention of G/C bases by a substantial margin. This suggests that the relationship between recombination and gBGC leads to accumulation of GC base pairs within a genomic region if operating consistently throughout evolutionary time.

Another type of BGC mechanism is DSB-related biased gene conversion (dBGC) (Latrille et al., 2017). When the two chromosomes differ due to presence of heterozygous sites, and one of the chromosomes has an active PRDM9 binding site; while the other has fewer PRDM9 binding sites, PRDM9 will preferentially target the chromosome with the active allele and initiate DSB. The sequence with the active allele is then erased and, during recombination, that sequence is replaced by a copy of the less active allele (Boulton et al., 1997; Latrille et al., 2017). This phenomenon of dBGC is also known as the hot-spot conversion paradox (Latrille et al., 2017). dBGC only happens in the PRDM9 binding sites, and it is less common than gBGC. In addition, unlike gBGC, dBGC does not favour the accumulation of particular base pairs (*e.g.*, GC base pairs). Instead, the dBGC process results in a progressive accumulation of inactive PRDM9 target sites. In other words, mechanistically, dBGC is not a process that might influence the GC content, but gBGC is. Therefore, in this work, I am only interested in the gBGC process as it may be responsible for GC content variation within the genome.

The gBGC hypothesis makes a number of explicit predictions that are interrogated in this chapter. Given the relationship between recombination rate and conversion rate, the mutation spectra are expected to differ between high-recombination rate and low-recombination rate regions. In particular, there should be more A/T→G/C mutations in high-recombination regions compared with low-recombination regions.

In addition, as the gBGC hypothesis seeks to explain the overabundance of GC nucleotides, it implies a positive relationship between GC levels and conversion rate. Therefore, I expect there should be more A/T→G/C mutations in high GC areas. Finally, it is expected under this hypothesis that a GC base pair at a polymorphic site is more likely to become fixed. Thus an additional prediction is that the mutation spectrum will differ between SNPs where the minor allele is rare compared with those where the minor allele is relatively common.

There are studies comparing frequencies of substitution between species, and considering GC-content and recombination as factors affecting substitution frequencies (Arnheim and Calabrese, 2009; Hardison et al., 2003; Hellmann et al., 2005). However, these studies do not compare mutation spectra within a single species directly. In addition, there is no evidence that a single factor, such as GC content or recombination rate, can explain the mutation rate variation between species (Arnheim and Calabrese, 2009; Ellegren et al., 2003). Glémin et al. (2015) measures the strength of gBGC in the human genome by analysing the derived allele frequency spectra. However, Glémin et al. (2015) did not explicitly account for the heterogeneity in composition between genomic regions.

In this chapter, I explicitly address the following predictions of the gBGC model using intergenic point mutations in human: (1) there should be a greater relative abundance of GC base pair alleles in common variants relative to rare variants; (2) SNPs in regions of the genome with high recombination rates will exhibit a greater relative abundance of GC base pairs; and (3) a similar relative abundance difference should also distinguish SNPs in genomic regions with high- and low-GC. These hypotheses are examined using the statistical methods presented in Chapter 2. The analyses provide support only for the first hypothesis, challenging the explanatory power of the gBGC mechanism.

4.2 Materials and Methods

Scripts for this project are written in Python, and requires programming using PyCogent (Knight et al., 2007) to query Ensembl (Flicek et al., 2013) and do statistical inference. In addition, formal programming tests were done to guarantee the correctness of the scripts.

In Chapter 2, I described a log-linear modelling method and associated visualisation approach, developed for comparing point mutation spectra between groups. In this study, I used the same statistical method to investigate the relationship between the mutation abundance and different genomic regions. In particular, whether the abundance of G/C-generating mutations is positively associated with recombination rate, with sequence GC content, and with minor allele frequency respectively.

4.2.1 Data sampling

In this work, I am concerned only with germline mutagenesis. As I do not get to directly observe the mutation events within the germline, I rely on an indirect measure. Specifically, that genetic variants that are segregating within human populations originated as mutations within the germline. The direction of the point mutation for individual genetic variant inferred by estimating the ancestral nucleotide. This latter state is already present in the database that I use.

Human germline SNPs were sampled from the international HapMap project phase III, which includes data from 1,397 individuals genotyped at over 1.5 million SNPs (Consortium et al., 2003). The detailed SNP information were obtained from Ensembl (Flicek et al., 2013) release 93, including the mutation direction, location and associated flanking sequence *etc.*, using Ensembldb3 (`http://ensembldb3.readthedocs.io`) querying capabilities. In particular, by ensuring the record was flagged as not somatic and validated, I sampled validated human germline SNPs from the database. For each human SNP record, I obtained the following information: SNP name, chromosomal location, genetic effect, strand, alleles, minor allele frequencies, ancestor base, flanking sequences (250 bp from each side of a SNP), and GC content from neighbourhood sequences. The former eight types of information were obtained directly from querying HapMap GVF files stored in Ensembl's *Homo Sapiens* variation database, and recombination rates for the SNP allocated region.

4.2.2 Choosing gBGC influenced mutations

From the sampled human germline SNPs, I employed a stratified data sampling strategy to minimise the potential confounding influence of other biological factors, and maximise the opportunity to discern the influence of biased gene conversion.

4.2.2.1 Excluding natural selection effects

gBGC acts like positive selection favouring GC alleles, and increases GC abundance in a DNA sequence. To distinguish gBGC effects from the natural selection effects, only SNPs in intergenic regions were sampled. As there is little evidence that natural selection operates to any substantial extent in these regions (Graur et al., 2013), these mutations are more likely to be selectively neutral.

4.2.2.2 Excluding dBGC effects

To avoid confounding with dBGC, variants within the PRDM9 binding motif were excluded. Myers et al. (2010) identified a degenerate 13–bp PRDM9 binding motif, CCNCCNTNNCCNC, that is over-represented in human recombination hotspots. The location of putative PRDM9 binding sites was identified using a position specific scoring matrix (PSSM). The 13-mer PRDM9 binding motif PSSM was obtained from the human PRDM9 binding map published in 2017 (Altemose et al., 2017).

To assess whether a mutation was present within a possible PRDM9 binding motif, for each genetic variant, a 27-bp sequence centred on the variant was extracted and evaluated at every possible position for a match with the PSSM. The sequence was scanned iteratively, and a log-odds score was computed for each possible position. Higher scores indicate a better correspondence to the pattern represented by the

PWM, *i.e.*, more likely to be a PRDM9 binding motif than from the background distribution. The cutoff threshold, arbitrarily defined as 1.78, was used to classify a location as being a PRDM9 binding motif, or not. This cutoff threshold is computed as the log-odds summation of 13-bp motif CCNCCNTNNCCNC, where N is replaced with the least likely base at a designated position. In this study, if a 27 bp sequence contains one or more candidate PRDM9 binding motif, this variant is disregarded.

4.2.3 Variants classification

4.2.3.1 Separating common and rare mutations

In this study, mutations were classified into common and rare groups according to their minor allele frequencies, as well as their existence among different human population groups. I defined common SNPs as those present among all 12 HapMap III human population groups, and with the derived allele frequencies greater than $\geq$50 per cent. Rare SNPs are mutations occurring exclusively in European population with the derived allele frequencies <10 per cent.

4.2.3.2 GC content information

The GC content around a point mutation is calculated from the 500 bp flanking a genetic variant as:

$$GC\% = \frac{G+C}{A+T+G+C} \times 100\% \tag{4.1}$$

According to the neighbourhood GC percentage, I separated the variants into high-GC and low-GC groups. The data were stratified with respect to upper and lower quartiles estimated for the human genome. The lowest quartile, median and highest quartile of neighbourhood GC percentage is 0.358, 0.404 and 0.452 respectively. Specifically, I defined high-GC region as those with GC percentage $\geq$ 0.452; and low-GC regions as those with GC percentage $<$ 0.358.

4.2.3.3 Recombination information

The recombination information was obtained from the deCODE 2010 recombination map (Kong et al., 2010). The recombination map used was an updated female recombination map consisting of 15,257 meiosis events, using genome-wide SNP data.

The recombination rate in deCODE2010 was recorded for each 10 kb bin interval. The recombination rate of a SNP was established by identifying the bin interval within which the SNP was located. The genome coordinates used by deCODE2010 recombination map are for the hg18 assembly (Kong et al., 2010). These coordinates were updated to Genome Reference Consortium GRCh38 using the liftOver tool (`https://genome.ucsc.edu/cgi-bin/hgLiftOver`). The variant location from the GRCh38 maps was used to determine the corresponding recombination rate. The deCODE2010 recombination map does not cover the entire human genome. Only SNPs covered by the deCODE recombination map were included in this study.

According to the located recombination rates, I separated the variants into high-recombination class and low-recombination class in two ways. Firstly, recombination rate were classified in line with Kong et al. (2010). According to those authors, a high-recombination region was defined as those bins with a standard recombination rate $\geq$ 10; and low-recombination region was those with a standard recombination rate < 10. Secondly, similar to the stratification of the GC content, the data were stratified according to upper and lower quartiles. The lowest quartile, median and highest quartile of standard recombination rate is 0.234, 0.911 and 2.568 respectively. Therefore, I defined high-recombination region as those with a standard recombination rate $\geq$ 2.568; and low-recombination regions as those with a standard recombination rate < 0.234.

In this study, SNPs were categorised into common and rare variant groups first, then further categorised according to either recombination rate or GC content. However, due to the limited sample size of rare variants, in the next step, only common variants were subclassified into high- and low-recombination group according to the recombination rate; or into high- and low-GC group according to the neighbourhood GC levels.

4.2.4 Log-linear modelling of neighbour effects

In this work, I used log-linear models comparing mutation spectra between groups described in Chapter 2 to test for the three main hypotheses proposed in the Motivation section respectively. Firstly, I implemented the 'spectra' hypothesis test to compare the distribution of point mutation outcomes between mutations in high- and low-recombination regions. This is to test whether mutations in high recombination show excess of events that create GC base pairs as hypothesised. Secondly, I implemented the analysis to compare point mutation spectra between high-GC and low-GC regions, which is ultimately what the gBGC model is about. In each of these two comparisons, only variants for which the derived allele was common were employed. Thirdly, I implemented the analysis to compare point mutation abundance between common and rare groups. This is to test whether the abundance of mutations producing GC base pairs will be greater with common variation compared to the rare group.

The question concerning the relative abundance of the different mutation types between different specified region groups was examined using the spectral analysis described in chapter 2. This method is appropriate in that it effectively controls for differences in the relative abundance of the bases between groups. It examines the equivalents between groups in the distribution of possible outcomes from mutating a specific base. For this analysis, the null hypothesis is that the abundance of G/C alleles originating from mutation of A/T is the same between the groups. The alternate hypothesis is that the abundance of G/C-generating mutations is different between these regions. For these tests, I specified the null hypothesis by setting the interaction parameter $\lambda^{direction:group}$ (see Equation 2.4) equals to 0.

4.3 Results

4.3.1 Analysis of mutation spectra between common- and rare-mutations

In this study, I sampled 110,903 common variants, and 2,621 rare variants in total (see Table 4.1). Recall that my definition of common and rare variants is different from that conventionally applied. Specifically, this designation is based on the frequency of the derived (or new) allele. Under the biased gene conversion hypothesis, GC base pairs are more likely to become fixed at a polymorphic site and the mutation spectrum is predicted to differ between common and rare variants. gBGC is a process that favours the accumulation of GC base pairs over AT base pairs when fixing heterozygous allele mismatches during recombination. Thus, an excess in A/T→G/C mutation abundance is expected in common variants compared with rare variants. The 'spectra' analysis (see section 2.2.4) was applied to test for this hypothesis.

Mutation	**Common**	**Rare**
A→C	3,496	103
A→G	14,943	401
A→T	1,358	119
C→A	3,393	131
C→G	1,728	105
C→T	14,228	431
G→A	14,924	459
G→C	1,834	125
G→T	35,305	131
T→A	1,349	146
T→C	14,865	361
T→G	3,480	109
Total	110,903	2,621

Table 4.1: By-mutation sample sizes of gBGC-affected common and rare variants.

After correcting for multiple test using the Holm-Šidäk procedure (Abdi, 2010), all mutation directions exhibited significant spectra differences between common and rare variant groups (see Table 4.2 and Figure 4.1). For the common variants, the A→G point mutation and its strand complement T→C were in strong excess, with residual entropy term (RET, see definition in section 2.2.5) value of 0.0033 and 0.005 respectively. Conversely, the A→T point mutation and its strand complement T→A were in strong deficit in common variants compared with rare variants, with RET value of -0.0035 and -0.0048 respectively. This observation is consistent with the prediction under the gBGC hypothesis.

As showed in Table 4.1, the number of rare variants is quite low. As a result, a robust comprehensive analysis is not possible. Therefore, my analyses were limited to those questions which could be addressed using common variants alone.

Direction	**RET**
T→A	-0.0048
A→T	-0.0035
G→C	-0.0027
C→G	-0.0022
C→A	-0.0007
G→T	-0.0003
T→G	-0.0000
A→C	0.0003
C→T	0.0029
G→A	0.0031
A→G	0.0033
T→C	0.0050

Table 4.2: Significant differences in spectra between gBGC common and rare point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. Only RET from the common group are shown. A positive (negative) RET indicates an excess (deficit) of that mutation in the common mutation group. All tests returned p-values that were below the significance threshold (0.05) and thus were statistically significant after correcting for four tests using the Holm-Šidák procedure.

4.3.2 Analysis of mutation spectra between mutations in high- and low-recombination regions

As the mechanism of gBGC is proposed to be explicitly coupled with the rate of recombination, analyses of mutation spectra comparison between mutations in high- and low-recombination region were performed. The gBGC hypothesis predicts that

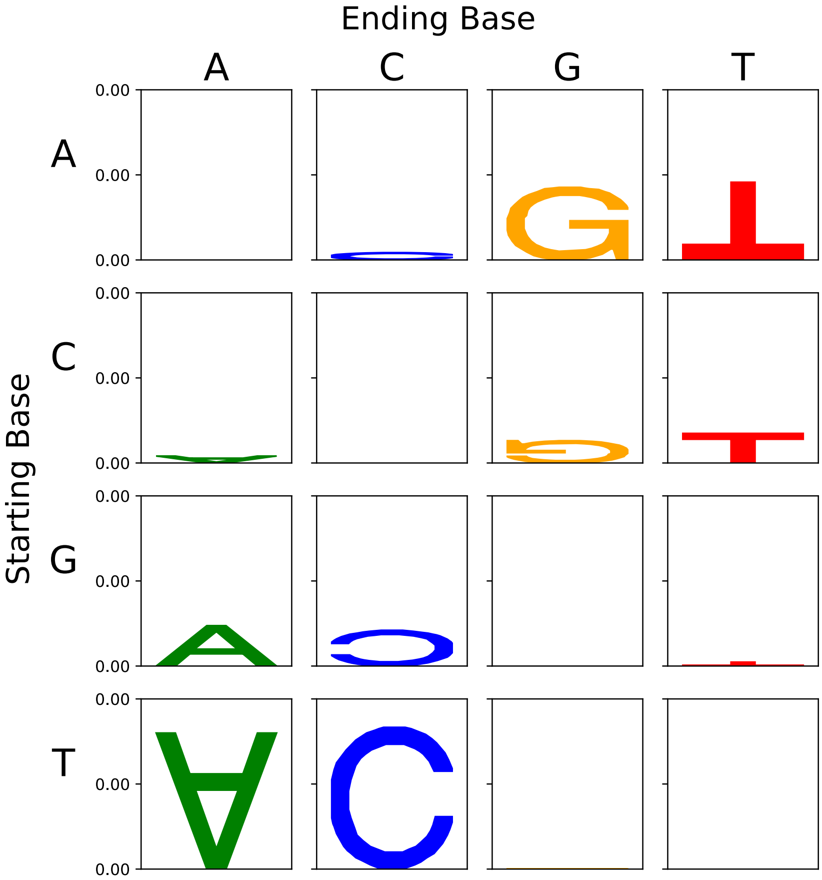


Figure 4.1: Significant differences in mutation spectra between common and rare variant groups. Starting base, Ending base correspond to X, Y respectively in X→Y. The y-axis is RE from the spectra hypothesis test and letter heights are as for the mutation motif logo in section 2.2.5. Letters in the normal orientation indicate an excess of that mutation direction in common relative to rare mutations. Inverted letters indicate a deficit in common relative to rare mutations.

genetic variants in a high-recombination rate region should have greater exposure to biased gene conversion compared with those in low-recombination rate regions. This prediction leads to the expectation that these regions should also differ in the relative abundance of variants fitting the A/T→G/C mutation pattern.

High-recombination regions were defined as those bins with a standard recombination rate ≥10, in line with Kong et al. (2010); and low-recombination rate regions as those with a standard recombination rate <10. With this setting, in total, there were 3,584 common variants located in high-recombination rate regions, and 75,544 common variants locate in low-recombination rate regions. The number of inferred observations per mutation direction are reported in Table 4.3.

Mutation	**High-recombination**	**Low-recombination**
A→C	191	3,305
A→G	783	14,160
A→T	62	1,296
C→A	121	3,272
C→G	72	1,656
C→T	539	13,689
G→A	596	14,328
G→C	67	1,767
G→T	151	3,379
T→A	57	1,292
T→C	762	14,103
T→G	183	3,297
Total	3,584	75,544

Table 4.3: By-mutation sample sizes of common gBGC-affected variants in high- and low-recombination regions respectively. A variant was classified as common if the derived allele had a frequency greater than 50%. Here, the high-recombination region is defined as those bins with a standard recombination rate ≥10; and low-recombination region is defined as those with a standard recombination rate <10.

My log-linear model for analysis of mutation spectra compared counts of point mutations from the same starting base between high- and low-recombination variant groups. Even before correcting for multiple tests using the Holm-Šidăk procedure, no significant difference was detected between these two classes. Thus, I accept the null hypothesis and conclude that the abundance of G/C alleles originating from mutation of A/T do not differ between high- and low-recombination rate regions (see Table 4.4).

Direction	***p*-value**
T→A	0.2933
A→T	0.4398
C→A	0.5698
G→C	0.5278
G→A	0.5278
C→T	0.5698
A→G	0.4398
T→C	0.2933
T→G	0.2933
C→G	0.5698
A→C	0.4398
G→T	0.5278

Table 4.4: No significant differences in spectra between common mutations in high- and low-recombination region. A variant was classified as common if the derived allele had a frequency greater than 50%. Separate log-linear models were used for each starting base (X in X→Y). The p-value was obtained from the χ^2 distribution for a starting base. This table shows the uncorrected results.

As Table 4.3 shows, the high- and low-recombination variants sample size is quite imbalanced, and number of variants sampled from the low-recombination rate regions was approximately 20 times larger than the number sampled from the high-recombination rate regions. To test whether such an imbalanced sample size affect the result, I redefined the definition of high- and low-recombination as upper- and lower-quartile of the standard recombination rate distribution. Specifically, I defined high-recombination regions as those bins with a standard recombination rate $\geq$2.568; and low-recombination regions as those with a standard recombination rate <0.234. With this setting, in total, I sampled 19,773 variants located in high-recombination rate regions, and 19,777 variants located in low-recombination rate regions. Detailed sample sizes per mutation direction are reported in Table 4.5.

Using this different definition for high- and low-recombination rate regions did not

Mutation	**High-recombination**	**Low-recombination**
A→C	960	827
A→G	3,936	3,608
A→T	337	343
C→A	798	909
C→G	413	394
C→T	3,448	3,674
G→A	3,584	3,803
G→C	477	466
G→T	847	916
T→A	307	358
T→C	3,772	3,631
T→G	894	848
Total	19,773	19,777

Table 4.5: By-mutation sample sizes of common gBGC-affected variants in high- and low-recombination regions respectively. A variant was classified as common if the derived allele had a frequency greater than 50%. Here, the high-recombination rate region was defined as those bins with a standard recombination rate ≥2.568; and low-recombination rate region was defined as those with a standard recombination rate <0.234.

change the inference. No single unadjusted *p*-value achieved the nominal significance threshold of 0.05 (see Table 4.6). Together with the previous test, these results do not support a relationship between recombination rate and G/C generating mutations.

4.3.3 Analysis of mutation spectra between mutations in high- and low-GC regions

Given the prediction of the gBGC hypothesis, that high GC content is a product of the operation of this mechanism, I also contrasted the mutation spectra between high- and low-GC regions. G/C generating mutations were predicted to be in excess in high-GC region compared with low-GC region.

High- and low-GC regions were defined using the upper- and low-quartile of GC percentage. High-GC were those with GC percentage ≥ 0.452; and low-GC regions were those with GC percentage < 0.358. With these definitions, a total of 51,459 variants were located high-GC regions and 24,530 variants were located in low-GC regions, see Table 4.7.

Direction	***p*-value**
T→A	0.0516
C→A	0.1138
A→T	0.1704
G→T	0.4232
G→A	0.4232
A→G	0.1704
C→T	0.1138
T→G	0.0516
G→C	0.4232
T→C	0.0516
C→G	0.1138
A→C	0.1704

Table 4.6: No significant differences in spectra between common mutations in high- and low-recombination region. A variant was classified as common if the derived allele had a frequency greater than 50%. Separate log-linear models were used for each starting base (X in X→Y). The p-value was obtained from the χ^2 distribution for a mutation direction.

Mutation spectra analysis was performed to compare counts of point mutations from the same starting base between high- and low-GC regions. A mixed signal was evident based on the nominal p-values alone. For instance, mutations between the C and G nucleotides themselves were nominally different (see Table 4.8). However, after correcting for multiple tests there were no significant differences. This result indicates that the abundance of G/C alleles generating mutation do not differ between these classes of regions (see Table 4.8).

In summary, there was no basis for rejecting the null hypothesis of equivalent abundance in G/C generating mutations between the high-GC and low-GC regions.

Mutation	**High-GC**	**Low-GC**
A→C	2,328	1,026
A→G	9,873	4,488
A→T	883	426
C→A	2,143	1,122
C→G	1,160	493
C→T	9,132	4,548
G→A	9,577	4,727
G→C	1,178	585
G→T	2,269	1,131
T→A	850	444
T→C	9,771	4,482
T→G	2,295	1,058
Total	5,1459	24,530

Table 4.7: By-mutation sample sizes of common gBGC-affected variants in high- and low-GC regions respectively. A variant was classified as common if the derived allele had a frequency greater than 50%. Here, the high-GC region is defined as those with a GC percentage ≥0.452; and low-recombination region is defined as those with a GC percentage <0.358.

Direction	***p*-value**
C→A	0.0050
T→A	0.1072
A→T	0.4299
C→T	0.0050
G→T	0.9679
A→G	0.4299
G→C	0.9679
T→G	0.1072
G→A	0.9679
A→C	0.4299
T→C	0.1072
C→G	0.0050

Table 4.8: No significant differences in spectra between common point mutations in high- and low-GC region. A variant was classified as common if the derived allele had a frequency greater than 50%. Separate log-linear models were used for each starting base (X in X→Y). The p-value was obtained from the χ^2 distribution for a mutation direction. This table shows the results before multiple test correction, only mutations starting from base C and G were nominally different. However, after multiple test correcting using the Holm-Šidäk procedure, no test returned p-values that were below the significance threshold and thus mutation spectra was not significantly different between the two region classes.

4.4 Discussion

A number of reports have presented evidence that has been interpreted as supporting a causative role for GC biased gene conversion and the evolution of isochores. Making claims regarding the mechanistic origins of a genomic feature requires demonstrating the plausibility of the historic operation of the proposed mechanism. Such arguments seem more credible if they are founded on observations from contemporary process operating within a species. In the current work, this perspective has been used to define questions that can be evaluated using genetic variants within humans. The gBGC hypothesis is particularly suited to such examination by virtue of the clear predictions that it makes. The results of my analyses of those predictions, however, do not provide strong support for the hypothesis. I discuss below whether this lack of support reflects limitations of the data or methods.

I have used the gBGC hypothesis to make explicit predictions concerning the enrichment of the G/C base pair at polymorphic sites. My predictions assume that: the age of a variant is positively correlated with its frequency; and, the older a derived variant, the greater the opportunity that it has been subjected to gBGC. With respect to the number of hydrogen bonds, nucleotide base pairs can be categorised as strong (S) and weak (W). The A and T pairing with two hydrogen bonds are W bases; and G and C pairing with three hydrogen bonds are S bases. I first tested the prediction regarding variant age and opportunity for gBGC, and my results do support this relationship. In particular, the abundance of mutations with A/T starting bases differ significantly between common and rare groups. In other words, mutations with W starting bases exhibited the largest spectra differences. Both W→W (A→T and T→A) mutations show a lower abundance and occurs less in the common group (see Table 4.2). For W→S mutations, T→C and its strand complement A→G mutations have a greater abundance and thus greater relative rate in the common group. This result may indicate a gBGC process is preferentially fixing A:C and T:G mismatches by removing W alleles, consistent with the hypothesis that gBGC is the mechanism increasing sequence GC content through time.

While the above result satisfies a fundamental assumption of the hypothesis, it does not account for the heterogeneity in composition between genomic regions, which is the motivation for the hypothesis to start with. Therefore, the most critical test is to establish a relationship exists between the relative abundance of the G/C derived alleles and genomic features most closely associated with the putative mechanism. In other words, a relationship with the rate of recombination. Or, secondarily, with the outcome of the mechanism, the relative abundance of GC base pairs.

Regarding the relationship with the rate of recombination, the prediction of the gBGC model is that the magnitude of the bias in gene conversion should be positively related to rates of recombination. However, comparison of the mutation spectra between high- and low-recombination rate regions did not support this prediction (see Table 4.4 and 4.6). One possible explanation for this result is that the estimation on the recombination rate may be noisy with respect to its historic value. The sequence may have produced G/C alleles via a relationship with recombination rate, and then the recombination rate changed. As a result, the data sampling, which stratified by the estimated contemporary recombination rate, introduces error. Analysing the GC content can reduce the uncertainty about the estimates, as obtaining accurate estimates of GC content from a sequence is straightforward.

Ultimately, the gBGC model is intended to explain the GC percentage variation in the genome. High-GC region are expected to have arisen as a result of more gBGC events. Therefore, if the process is still ongoing, there is expected to be more G/C-enriching mutations in high-GC region than low-GC region. Results of the mutation spectra comparison between high- and low-GC group did not support this prediction (see Table 4.8). One possibility is that processes that gave rise to GC heterogeneity may operate on different timeframes. The process operating now and the ancient process may differ in their relationship to GC content, again resulting in the introduction of noise.

According to how I defined common and rare variants, common genetic variants are more likely to have experienced gBGC compared with rare variants. Therefore, the issue becomes whether or not these classified variants are indicative of the true process that drives the origin of isochore's. There may exist other factors accounting for mutation abundance in different genomic regions. Glémin et al. (2015) measured the strength of gBGC in the human genome by including the derived allele frequency spectra. However, Glémin et al. (2015) did not explicitly evaluate whether the derived allele frequency spectra were different between genomic regions that differ in recombination rate, between regions with different GC content, or between common and rare variant groups. Hardison et al. (2003) studied the covariation in frequencies of substitution between human and mouse, and showed that GC content between human and mouse can only account for part of the variation, but cannot account for all the variation in divergence. However, it is now demonstrated that conventional phylogenetic methods for estimating substitution are systematically biased by changing sequence composition (Kaehler et al., 2017, 2015). So these inferences are likely to be unreliable. In addition, because a mutation spectra difference was observed between the rare and common variants (see Table 4.2), I suggest that the relationships which should exist with recombination rate and/or GC content, are either somehow being obscured by attributes of the data.

The absence of the hypothesised relationship between recombination rate and gBGC was reported by Robinson et al. (2013) for *Drosophila melanogaster*. Kliman and Eyre-Walker (1998) detect significant heterogeneity in GC content among intron segments from genes of *D. melanogaster*. In a population genomic analysis of *D. melanogaster*, a comparison of the site frequency spectrum of GC-enriching versus AT-enriching mutations between different recombination rate categories found no evidence for the relationship between recombination rate and gBGC (Robinson et al., 2013). Robinson et al. (2013) speculated this observation was due to an unclear relationship between gBGC and crossovers or non-crossovers events, *i.e.*, whether gBGC is associated with crossovers only, or non-crossovers only, or both.

However, they assumed recombination is only associated with crossover events. In fact, both crossovers and non-crossovers are potential outcomes of meiotic recombination, and both types of events may involve gene conversion (Duret and Galtier, 2009).

Another possible interpretation for absence of a relationship between recombination rate and mutation spectra is that there may exist additional factors affecting mutation spectra. Ellegren et al. (2003) argued that, it is unlikely that a single factor (*e.g.*, recombination rate) is an adequate predictor that can explain mutation rate variation. The existence of numerous distinct contributors to mutation processes are described and evaluated in Chapter 2. A different type of analysis that treats the recombination rate as a co-factor in the analysis will likely be better powered and thus more informative.

The inconsistencies between the predictions with GC levels and my results may reflect limitations of the data. The data sampled here are polymorphism data. These may result from an allelic- or non-allelic gene conversion process. Allelic gene conversions happen between orthologous during meiotic recombination; whereas non-allelic gene conversions are those between paralogous DNA segments (Chen et al., 2007). Allelic gene conversion is predominantly associate with gBGC (Duret and Galtier, 2009), thus a positive correlation is expected between the degree of GC-enriching mutation and GC content. Studies also found that the non-allelic gene conversion is not GC-biased. Assis and Kondrashov (2011) performed an analysis of Drosophila and primate genomes, and computed frequencies of AT→GC and GC→AT mutations produced by non-allelic gene conversion. Their results revealed the non-allelic gene conversion is AT-biased, rather than GC-biased. If this is also the case in humans, it would be difficult to detect the expected relationship between the mutation abundance and GC content.

In summary, in this work, I have evaluated the relationship of mutation spectra and recombination rate; and the relationship of mutation spectra and sequence GC content, respectively. My analyses do not provide a strong support for the relationships between mutation abundance and genomic regions. There is compelling evidence that mutations abundance differ significantly between rare and common genetic variants. This observation does demonstrate the operation of some molecular process underpinning compositional divergence. Unfortunately, the relationship with recombination rate predicted by the biased gene conversion hypothesis was not supported, even weakly. It is certainly a possibility that this inability to establish a relationship is a consequence of noise in the data regarding the rate of recombination. There are also limitations in the current analysis. My approach involves discarding data in order to achieve a simplified stratification suitable for the analysis method. Development of an alternative methodological approach may be able to make better use of all the information in the data. Finally, it may simply be that gBGC is not the explanation and alternate mechanistic candidates should thus still be considered.

Chapter 5

Conclusion

The main focus of the research reported in this thesis was the influence of neighbourhood sequences on point mutations. Three separate projects were reported in separate thesis chapters. In Chapter 2, I established a robust statistical test for the existence of neighbouring base influence on point mutations, and demonstrated the performance of these statistical methods by analysing mutation processes in human germline and malignant melanoma. In Chapter 3, I asked if there is so much information in neighbouring bases, can this information be used to classify the mechanistic origins of a genetic variant? In particular, I compared neighbouring bases between the ENU-induced and spontaneous mouse mutations, and demonstrated that neighbourhood effects differ between these two classes and have unique relationships with the underlying mutagenic mechanisms that allow them to be distinguished. I further developed a LR classifier to discriminate between ENU-induced and spontaneous mouse mutations. In Chapter 4, I tested three fundamental hypotheses underlying GC biased gene conversion, a mechanism proposed to account for the existence of compositional heterogeneity in GC content, a major feature of the human genome. This chapter is an application of statistical models developed in Chapter 2, demonstrating their value as powerful tools for enhancing our understanding of the origins and implications of genetic variation.

While the nature and size of neighbourhood effect has been extensively studied, my published work has re-defined the scale of these influences. Although many research areas are impacted by this work, it presents a particular challenge for phylogenetic studies. The dominant phylogenetic approaches model sequence evolution on the assumption of strict independence, or at best allows for interactions among non-overlapping trinucleotides. My result shows that the size of neighbourhood effect can reach up to ± 5 bp away from the mutation. In addition, higher-order neighbourhood effects were determined to also be strongly significant. The tendency for physical proximity feature among positions engaged in higher-order effects substantiates the sensible nature of these effects as physical proximity is a primary determinant for chemical interactions. Therefore, a more realistic phylogenetic framework will require consideration of at least the immediate neighbours, which would be a 3-mer; or even better if can go up to 5-mer. The joint effect between neighbours should also be considered.

My results provided insight into how the information present in neighbouring bases can be exploited to further studies of mutagenesis. Mutations may originate from a wide spectra of mutagens and each mutagenic process can operate via distinct mechanisms at the DNA level; neighbourhood sequence associations may reflect the mechanism. My second projects showed that neighbourhood effects between mutations induced by two classes of mutagenic process, ENU-induced and spontaneous, were significantly different. Furthermore, these distinctive neighbourhood effects proved useful as features in machine learning classifiers to successfully discriminate between mutagenic mechanisms. In results that were consistent with the information content analyses presented in Chapter 2, I identified that I+2D effects were a sufficient set of features for discriminating between mutagenic processes. Adding more higher-order effect did not significantly improve classification performance. This feature combination may be a generalisable set of features that can be considered by other studies seeking to capture properties of DNA sequences. Furthermore, I considered a more general case, whether it is possible to train the classifier with one class of data (*e.g.*, the spontaneous germline point mutations), and obtain a classifier to discriminate outliers from that data. The short answer to this question was ‘not very well’. However, this question is the natural perspective from which to maximise the information that can be extracted from a known distribution, and apply it to search for distinctive processes. It remains a viable topic for further investigation.

The analyses of my final project did not concern neighbourhood influences directly, but was concerned with a more fundamental property of the mutation spectra. In particular, I tested the relationship between mutation spectra and different genomic features hypothesised under gBGC model. The method is developed in Chapter 2 and is robust to the variability in nucleotide content between samples, eliminating a potential confounder. My results substantiated the fundamental premise, that GC base pairs increase in the relative abundance through time. However, the conjectured relationship with the rate of recombination was not evident. This suggested that additional work needs to be done in refining the statistical approach such that it takes full use of the available data.

In conclusion, the results presented in this thesis significantly improve the understanding of mutagenesis contributions to the distribution of genetic variation, and how to analyse it. They have substantial implications and direct applications to a number of research problems, including mutation detection technique development, phylogenetic and molecular evolutionary analyses studies and disease aetiology studies. With the continued development of comparative genomic algorithms and the availability of more sequencing data, these methods will continue to be of value due to their statistical robustness, interpretability and their computational tractability.

Appendix A

Supplementary Information 1

Position(s)	Deviance	df	p-value
-2	1574.2	3	0.0
-1	18674.9	3	0.0
+1	346848.0	3	0.0
+2	2174.5	3	0.0
(-2, -1)	1603.1	9	0.0
(-2, +1)	555.5	9	0.0
(-2, +2)	352.7	9	1.7×10^{-70}
(-1, +1)	2341.3	9	0.0
(-1, +2)	315.1	9	1.6×10^{-62}
(+1, +2)	1965.0	9	0.0
(-2, -1, +1)	939.7	27	0.0
(-2, -1, +2)	523.0	27	2.7×10^{-93}
(-2, +1, +2)	264.6	27	7.3×10^{-41}
(-1, +1, +2)	467.8	27	6.5×10^{-82}
(-2, -1, +1, +2)	273.9	81	9.1×10^{-23}

Table A.1: Log-linear analysis of C→T autosomal intergenic mutations. Position(s) are relative to the index position (see Figure 2.1). Deviance is from the log-linear model, with df degrees-of-freedom and corresponding *p*-value obtained from the χ^2 distribution. p-values listed as 0.0 are below the limit of detection.

Position(s)	Deviance	df	p-value
-2	26528.3	3	0.0
-1	20038.7	3	0.0
+1	57037.8	3	0.0
+2	1802.0	3	0.0
(-2, -1)	9058.8	9	0.0
(-2, +1)	3615.8	9	0.0
(-2, +2)	701.1	9	0.0
(-1, +1)	3233.2	9	0.0
(-1, +2)	1516.8	9	0.0
(+1, +2)	2329.1	9	0.0
(-2, -1, +1)	2018.3	27	0.0
(-2, -1, +2)	561.1	27	0.0
(-2, +1, +2)	362.2	27	2.4×10^{-60}
(-1, +1, +2)	1191.2	27	0.0
(-2, -1, +1, +2)	426.5	81	2.3×10^{-48}

Table A.2: Log-linear analysis of A→G autosomal intergenic mutations. Position(s) are relative to the index position (see Figure 2.1). Deviance is from the log-linear model, with df degrees-of-freedom and corresponding p-value obtained from the χ^2 distribution. p-values listed as 0.0 are below the limit of detection.

Direction	RE$_{max}$(1)	RE Dist.	p-val Dist.
A→C	0.0042	4	10
A→G	0.0186	2	10
A→T	0.0093	3	10
C→A	0.0093	4	10
C→G	0.0057	3	10
C→T	0.0861	1	10
G→A	0.0860	1	10
G→C	0.0054	3	10
G→T	0.0091	4	10
T→A	0.0095	3	10
T→C	0.0190	2	10
T→G	0.0039	4	10

Table A.3: The most distant positions from the mutation with RE(1) $\geq$ 10% of RE$_{max}$(1). RE(1) is the first-order RE for the position, and RE$_{max}$(1) the largest RE from a first-order effect for the surveyed positions. RE Dist. is the absolute value of the relative position based on the RE value. p-val Dist. is the corresponding distance based on the p-value. The maximum possible distance is 10. Only point mutations significant after correcting for 20 tests using the Holm-Šidák procedure were considered.

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)
A→C	1.6×10^{-5}	+1	-	-
A→G	4.2×10^{-5}	+1	2.6×10^{-5}	(-2, -1)
A→T	9.3×10^{-5}	-1	1.5×10^{-5}	(-2, +2)
C→A	2.7×10^{-5}	-1	3.4×10^{-5}	(-1, +1)
C→G	3.8×10^{-5}	-1	1.5×10^{-5}	(-2, -1)
C→T	3.2×10^{-5}	+1	1.2×10^{-5}	(-1, +1)

Table A.4: Neighbour associations with point mutations differ between autosomal intronic and intergenic point mutations. As there was no significant strand asymmetry detected for either sequence class, only + strand effects are shown. Only point mutations with at least one significant test after correcting for 15 tests using the Holm-Šidäk procedure are shown. Non-significant results are indicated by '-'.

Direction	**RE**$_{max}$(1)	**Pos.**(1)
A→C	1.7×10^{-5}	+1
A→G	6.3×10^{-6}	+1
C→G	1.4×10^{-5}	+1
C→T	5.0×10^{-6}	+1
G→A	6.2×10^{-6}	+1
T→A	1.6×10^{-5}	+2
T→C	8.3×10^{-6}	-1
T→G	2.1×10^{-5}	-1

Table A.5: Significant differences in neighbour associations between intergenic autosomal and X-chromosomal point mutations. RE$_{max}$(1) the largest RE from a first-order test and Pos.(1) is the corresponding position. Only mutations significant after correcting for the 15 different tests using the Holm-Šidäk procedure are shown.

Direction	**RET**
G→A	-0.0032
A→G	-0.0031
C→T	-0.0031
T→C	-0.0026
C→G	-0.0019
G→C	-0.0017
T→G	-0.0011
A→C	-0.0007
T→A	0.0038
A→T	0.0039
G→T	0.0051
C→A	0.0052

Table A.6: Significant differences in mutation spectra between autosomal intergenic and intronic point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. Only RET from the intergenic group are shown. A positive (negative) RET indicates an excess (deficit) of that mutation in the intergenic group. All tests returned p-values that were below the limit of detection and thus were statistically significant after correcting for four tests using the Holm-Šidäk procedure.

Direction	**RET**
T→A	-0.0004
C→A	-0.0003
G→T	-0.0003
A→T	-0.0002
A→C	-0.0002
T→G	-0.0001
G→C	-0.0000
C→G	0.0000
C→T	0.0003
G→A	0.0003
A→G	0.0004
T→C	0.0005

Table A.7: Significant differences in spectra between autosomal and X-chromosomal intergenic point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. p-value is from the corresponding hypothesis test. Only RET from the autosomal group are shown. A positive (negative) RET indicates a excess (deficit) of that mutation in autosomes. All tests returned p-values that were $\leq 4.7e^{-9}$ and thus were statistically significant after correcting for four tests using the Holm-Šidäk procedure.

Direction	**RET**
T→G	-0.0001
A→C	-0.0001
G→T	-0.0001
C→A	-0.0001
G→C	-0.0001
A→T	-0.0001
T→A	-0.0000
C→G	0.0000
C→T	0.0001
G→A	0.0002
T→C	0.0002
A→G	0.0002

Table A.8: Significant differences in spectra between autosomal and X-chromosomal intronic point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. p-value is from the corresponding hypothesis test. Only RET from the autosomal group are shown. A positive (negative) RET indicates an excess (deficit) of that mutation in autosomes. All tests returned p-values that were $\leq 8.6e^{-5}$ and thus were statistically significant after correcting for four tests using the Holm-Šidák procedure.

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)	**RE**$_{max}$(3)	**Pos.**(3)
A→C	0.0132	-1	0.0093	(-1, +1)	0.0039	(-2, -1, +1)
A→G	0.0134	-1	0.0164	(-1, +1)	0.0032	(-2, -1, +1)
A→T	0.0116	-1	0.0030	(-2, +1)	0.0027	(-2, -1, +1)
C→A	0.0276	-1	0.0076	(-1, +1)	0.0029	(-1, +1, +2)
C→G	0.0259	+1	0.0028	(-1, +1)	0.0025	(-2, -1, +1)
C→T	0.0840	-1	0.0110	(-1, +1)	0.0006	(-2, -1, +1)

Table A.9: Test of strand symmetric neighbourhood associations for malignant melanoma point mutations. RE$_{max}$(#) is the maximum RE for order # and Pos.(#) the corresponding position(s). Only effects significant after correcting for the 15 different tests using the Holm-Šidák procedure are shown.

Direction	**RET**	**p-value**
A→C	-0.0025	0.1650
C→G	-0.0024	8.0×10^{-50}
C→A	-0.0020	8.0×10^{-50}
A→T	0.0007	0.1650
A→G	0.0018	0.1650
C→T	0.0048	8.0×10^{-50}

Table A.10: Differences in spectra between strands for malignant melanoma point mutations. Separate log-linear models were used for the + strand starting bases A and C. RET is the RE term for that row mutation direction. Only RET from the + strand are shown. A positive (negative) RET indicates an excess (deficit) of that mutation on the + strand. *p*-value is from the corresponding hypothesis test. Only mutations from C were significant after correcting for two tests using the Holm-Šidäk procedure.

Direction	$\mathbf{RE}_{max}(1)$	**Pos.**(1)	$\mathbf{RE}_{max}(2)$	**Pos.**(2)	$\mathbf{RE}_{max}(3)$	**Pos.**(3)
A→C	0.0034	-1	0.0016	(+1, +2)	0.0012	(-2, -1, +1)
A→G	0.0205	+1	0.0042	(-2, -1)	0.0007	(-2, -1, +1)
A→T	0.0089	+1	0.0051	(-1, +1)	0.0025	(-1, +1, +2)
C→A	0.0092	+1	0.0035	(-1, +1)	0.0012	(-1, +1, +2)
C→G	0.0049	+1	0.0022	(+1, +2)	0.0008	(-1, +1, +2)
C→T	0.0924	+1	0.0004	(+1, +2)	0.0002	(-2, -1, +1)

Table A.11: Neighbour associations with point mutations within autosomal introns. $\text{RE}_{max}(1)$ is the largest RE from a first-order test and Pos.(1) is the corresponding position. Only mutations significant after correcting for the 15 different tests using the Holm-Šidäk procedure are shown.

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)	**RE**$_{max}$(3)	**Pos.**(3)
A→C	0.0033	-1	0.0009	(-1, +1)	-	-
A→G	0.0023	+1	0.0005	(-1, +1)	0.0004	(-1, +1, +2)
A→T	0.0071	-1	0.0023	(+1, +2)	-	-
C→A	0.0065	-1	0.0013	(-2, -1)	0.0007	(-2, +1, +2)
C→G	0.0007	+1	0.0004	(+1, +2)	0.0006	(-2, -1, +2)
C→T	0.0268	-1	0.0030	(-1, +1)	0.0002	(-2, -1, +1)
G→A	0.0275	+1	0.0017	(-1, +1)	0.0002	(-1, +1, +2)
G→C	0.0008	-1	0.0004	(+1, +2)	0.0006	(-2, -1, +2)
G→T	0.0056	+1	0.0011	(+1, +2)	0.0007	(-1, +1, +2)
T→A	0.0080	+1	0.0018	(-2, -1)	0.0018	(-1, +1, +2)
T→C	0.0023	-1	0.0014	(-1, +1)	0.0005	(-1, +1, +2)
T→G	0.0014	+1	0.0015	(-1, +1)	0.0013	(-2, +1, +2)

Table A.12: Significant differences in the association of neighbours on exonic point mutations between germline and malignant melanoma. RE$_{max}$(1) is the largest RE from a first-order test and Pos.(1) is the corresponding position. Only mutations significant after correcting for the 15 different tests using the Holm-Šidäk procedure are shown. Non-significant results are indicated by '-'.

Direction	**RET**
T→C	-0.0332
A→G	-0.0327
C→G	-0.0109
C→A	-0.0092
G→C	-0.0091
G→T	-0.0080
A→C	0.0045
T→G	0.0061
G→A	0.0263
C→T	0.0346
T→A	0.0624
A→T	0.0624

Table A.13: Significant differences in spectra between germline exon and malignant melanoma point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. Only RET from the melanoma group are shown. A positive (negative) RET indicates an excess (deficit) of that mutation in the melanoma group. All tests returned p-values that were below the limit of detection and thus were statistically significant after correcting for four tests using the Holm-Šidäk procedure.

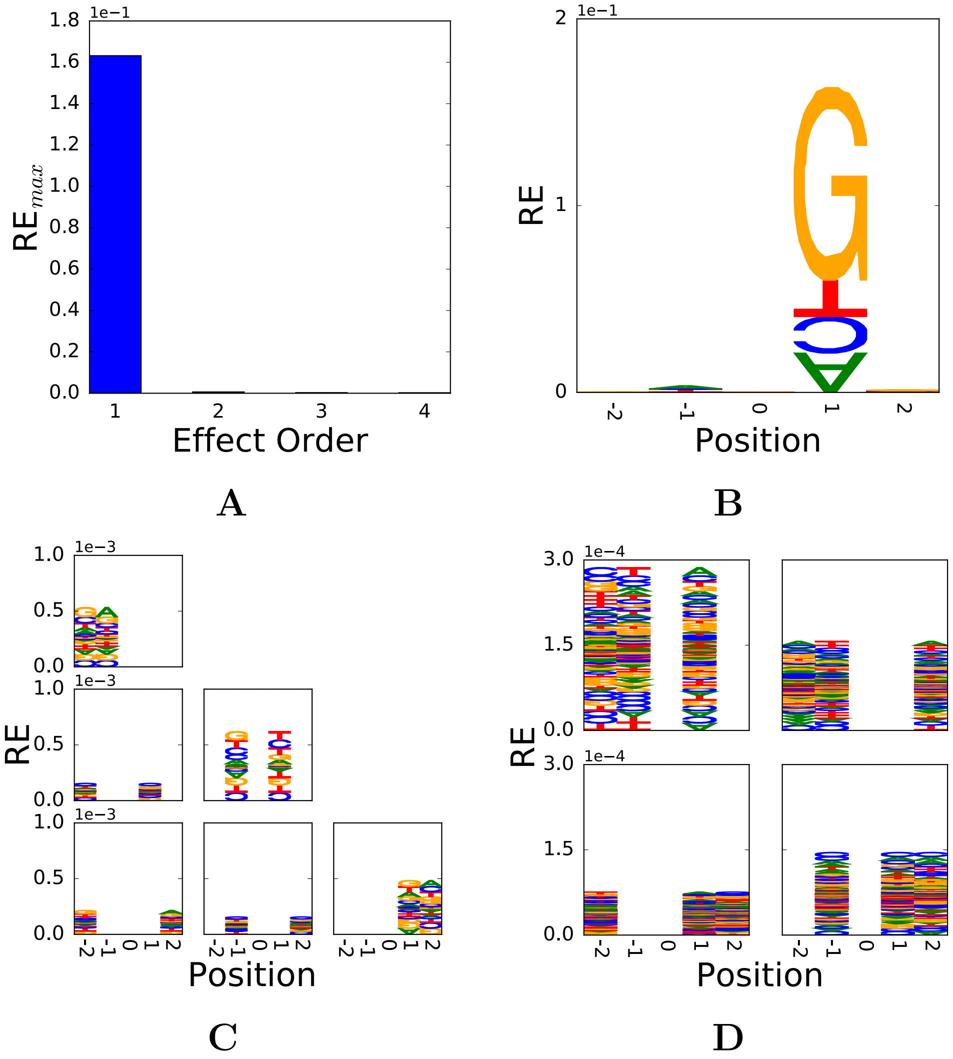


Figure A.1: Flanking influences on C→T mutation in autosomal exon sequences. (**A**) First-order effects are the dominant neighbourhood influence, RE$_{max}$ (y-axis) is the maximum RE from the possible evaluations for a motif length (x-axis), (**B**) Single-position effects, (**C**) Two-way effects, and (**D**) Three-way effects.

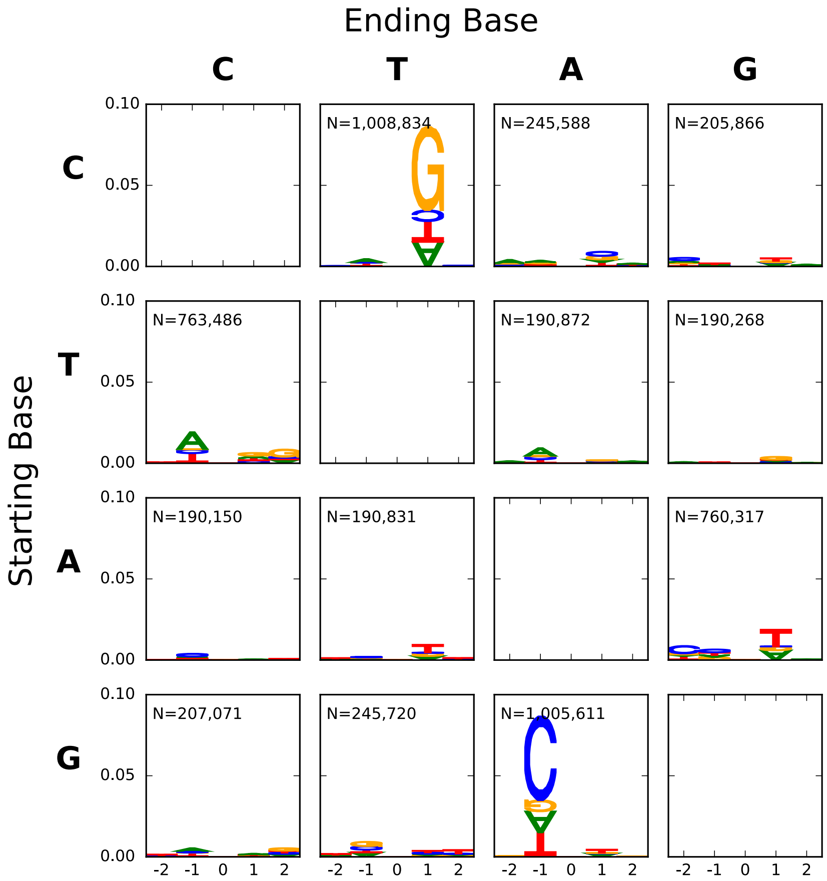


Figure A.2: A panel of all 12 point mutations from autosomal intergenic germline mutations. Text in each panel indicates the number of genetic variants analysed.

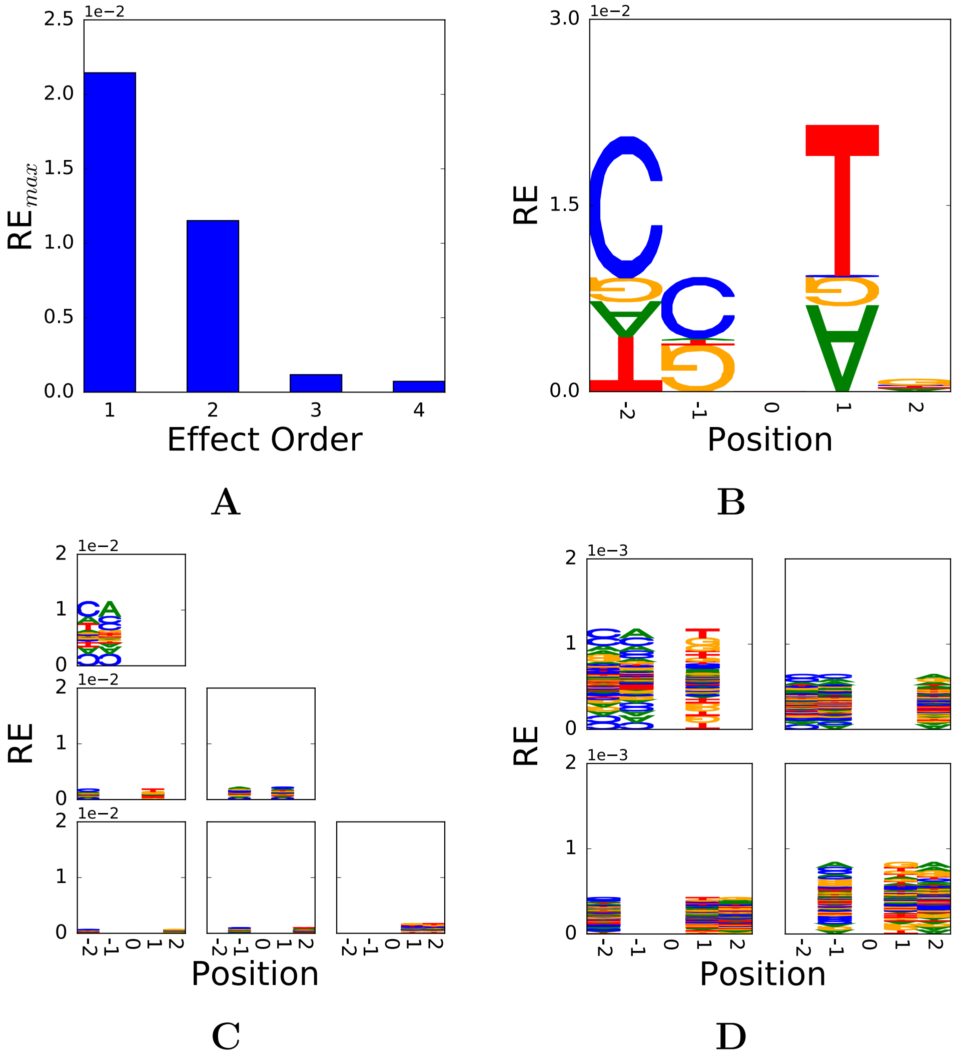
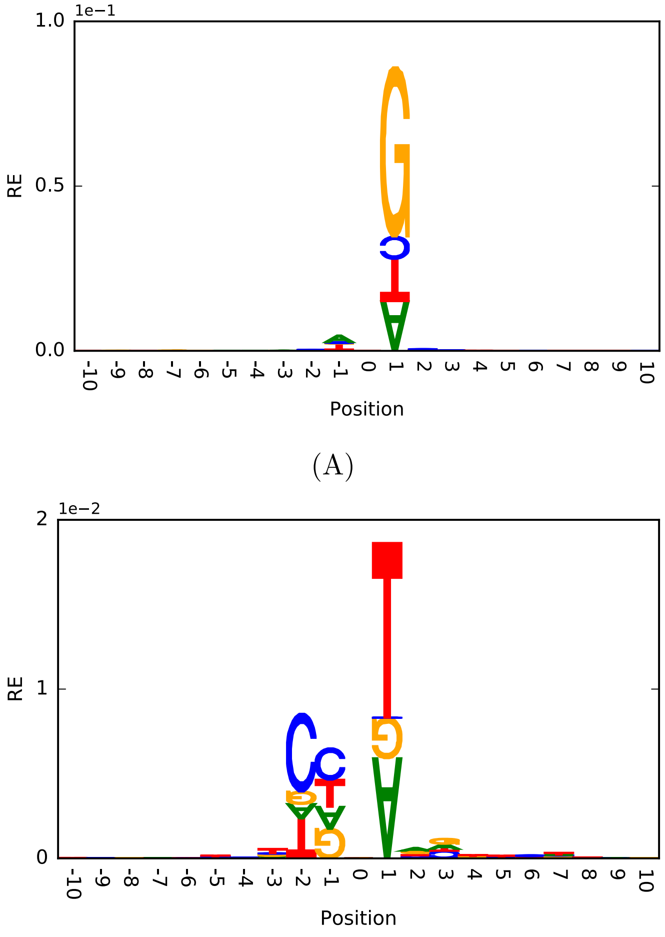


Figure A.3: Flanking influences on A→G mutation in autosomal exon sequences. (**A**) First-order effects are the dominant neighbourhood influence, (**B**) Single-position effects, (**C**) Two-way effects, and (**D**) Three-way effects



(A)

(B)

Figure A.4: The extent of neighbourhood effects on autosomal intergenic mutations. (**A**) C→T, (**B**) A→G.

Mutation motifs estimated using the whole genome as the reference

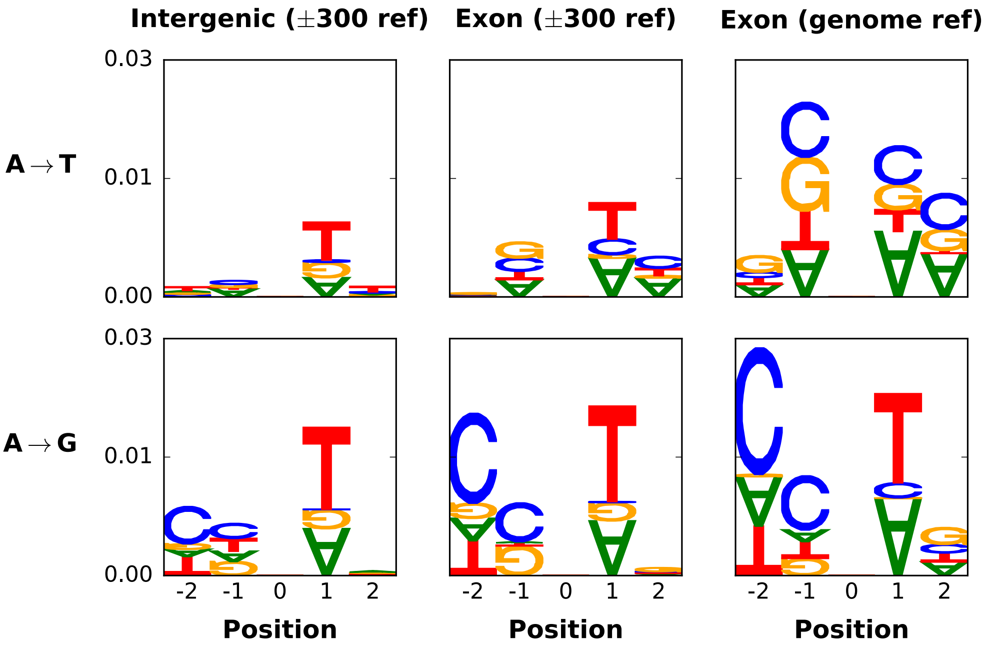


Figure A.5: Using the genome as the reference introduces bias. '± 300bp ref' uses reference bases selected at random within ± 300bp of the mutated base. 'genome ref' uses reference bases selected at random from the entire human genome. Only autosomal mutations were used for the analysis.

Appendix B

Supplementary Information 2

Direction	Class	RET
T→C	ENU	-0.047
A→T	Spontaneous	-0.036
G→T	Spontaneous	-0.036
T→A	Spontaneous	-0.035
A→G	ENU	-0.035
C→A	Spontaneous	-0.034
G→A	ENU	-0.025
C→T	ENU	-0.021
A→C	ENU	-0.018
T→G	ENU	-0.007
G→C	ENU	-0.001
C→G	ENU	-0.001
T→G	Spontaneous	0.009
C→G	Spontaneous	0.022
C→T	Spontaneous	0.022
G→C	Spontaneous	0.023
G→A	Spontaneous	0.027
A→C	Spontaneous	0.027
A→G	Spontaneous	0.039
C→A	ENU	0.052
T→C	Spontaneous	0.055
G→T	ENU	0.063
T→A	ENU	0.066
A→T	ENU	0.067

Table B.1: Comparison of mutation spectra between Spontaneous and ENU-induced germline point mutations. RET values are proportional to deviance generated from the log-linear model (Zhu et al., 2017), and p-value are obtained from the χ^2 distribution. All p-values were below the limit of detection.

Mutation direction	**1st-order**	**2nd-order**	**3rd-order**	**4th-order**
A→C	4	5	3	0
A→G	4	5	4	1
A→T	4	5	2	1
C→A	4	6	4	1
C→T	4	5	4	1
G→A	4	5	4	1
G→T	4	5	2	1
T→A	4	6	2	0
T→C	4	6	4	0
T→G	4	5	3	1

Table B.2: Number of positions showing significant differences between ENU-induced and spontaneous germline point mutations from analysis of 5-mers. A p-value ≤ 0.05 was classified as significant. p-values were from the log-linear analysis.

Direction	**RE**$_{max}$(1)	**RE Dist.**	**p-val Dist.**
A→C	0.0374	6	10
A→G	0.0402	4	10
A→T	0.0638	2	10
C→A	0.0632	2	10
C→T	0.0703	2	10
G→A	0.0710	2	10
G→T	0.0624	2	10
T→A	0.0606	2	10
T→C	0.0395	4	10
T→G	0.0373	6	10

(a) ENU-induced

Direction	**RE**$_{max}$(1)	**RE Dist.**	**p-val Dist.**
A→C	0.0047	8	10
A→G	0.0118	3	10
A→T	0.0194	3	10
C→A	0.0332	4	10
C→T	0.0505	1	10
G→A	0.0508	1	10
G→T	0.0351	3	10
T→A	0.0117	2	10
T→C	0.0152	2	10
T→G	0.0148	2	10

(b) Spontaneous

Table B.3: Longer range neighbourhood effect log-linear analyses results of (a) ENU-induced mutations and (b) germline spontaneous mutations. For both subtables, the most distant positions from the mutation with RE(1) $\geq$ 10% of RE$_{max}$(1). RE(1) is the first-order RE for the position, and RE$_{max}$(1) the largest RE from a first-order effect for the surveyed positions. RE Dist. is the furthest position with an RE value $\geq 0.1 \times$ RE$_{max}$. p-val Dist. is the corresponding distance based on the p-value≤ 0.05. As the analysis was limited to a flank size of 10bp either side of the mutating base, the maximum possible distance is 10.

Chromosome	**ENU-induced**	**Spontaneous**
1	16,977	17,848
2	21,100	20,051
3	11,228	11,713
4	13,973	16,936
5	14,509	16,028
6	13,039	12,097
7	20,864	19,161
8	11,232	13,465
9	14,010	15,662
10	11,315	12,641
11	17,101	19,626
12	8,022	8,817
13	9,085	8,939
14	8,395	8,868
15	9,342	11,079
16	7,266	8,117
17	11,981	12,168
18	6,356	7,732
19	7,529	8,635
XY	853	5,097

Table B.4: By-chromosome sample sizes of genetic variants from the ENU induced and spontaneous germline mutations.

Classifier design	Training size	max AUC	min AUC
M	1,009	0.738	0.711
M	2,050	0.735	0.723
M	4,101	0.730	0.720
M	10,255	0.735	0.725
M	16,408	0.731	0.727
M+I	1,009	0.781	0.737
M+I	2,050	0.778	0.755
M+I	4,101	0.785	0.771
M+I	10,255	0.784	0.774
M+I	16,408	0.780	0.774
M+I+2D	1,009	0.777	0.734
M+I+2D	2,050	0.771	0.755
M+I+2D	4,101	0.788	0.775
M+I+2D	10,255	0.788	0.778
M+I+2D	16,408	0.787	0.783
FS	1,009	0.777	0.733
FS	2,050	0.773	0.757
FS	4,101	0.784	0.773
FS	10,255	0.787	0.780
FS	16,408	0.790	0.784

Table B.5: Summary of AUC scores from LR classifiers using 7-mers.

Classifier design	Training size	max AUC	min AUC
M	1,009	0.738	0.711
M	2,050	0.735	0.723
M	4,101	0.730	0.720
M	10,255	0.735	0.725
M	16,408	0.731	0.727
M+I	1,009	0.777	0.737
M+I	2,050	0.772	0.763
M+I	4,101	0.777	0.762
M+I	10,255	0.775	0.764
M+I	16,408	0.773	0.766
M+I+2D	1,009	0.777	0.736
M+I+2D	2,050	0.767	0.755
M+I+2D	4,101	0.776	0.763
M+I+2D	10,255	0.775	0.766
M+I+2D	16,408	0.774	0.769

Table B.6: Summary of AUC scores from LR classifiers using 3-mers.

Classifier design	**Training size**	**max AUC**	**min AUC**
M	1,009	0.738	0.711
M	2,050	0.735	0.723
M	4,101	0.730	0.720
M	10,255	0.735	0.725
M	16,408	0.731	0.727
M+I	1,009	0.771	0.736
M+I	2,050	0.775	0.758
M+I	4,101	0.779	0.764
M+I	10,255	0.778	0.768
M+I	16,408	0.775	0.769
M+I+2D	1,009	0.774	0.735
M+I+2D	2,050	0.765	0.755
M+I+2D	4,101	0.779	0.769
M+I+2D	10,255	0.781	0.772
M+I+2D	16,408	0.781	0.775
M+I+4D	1,009	0.774	0.734
M+I+4D	2,050	0.771	0.756
M+I+4D	4,101	0.779	0.770
M+I+4D	10,255	0.782	0.773
M+I+4D	16,408	0.782	0.776

Table B.7: Summary of AUC scores from LR classifiers using 5-mers.

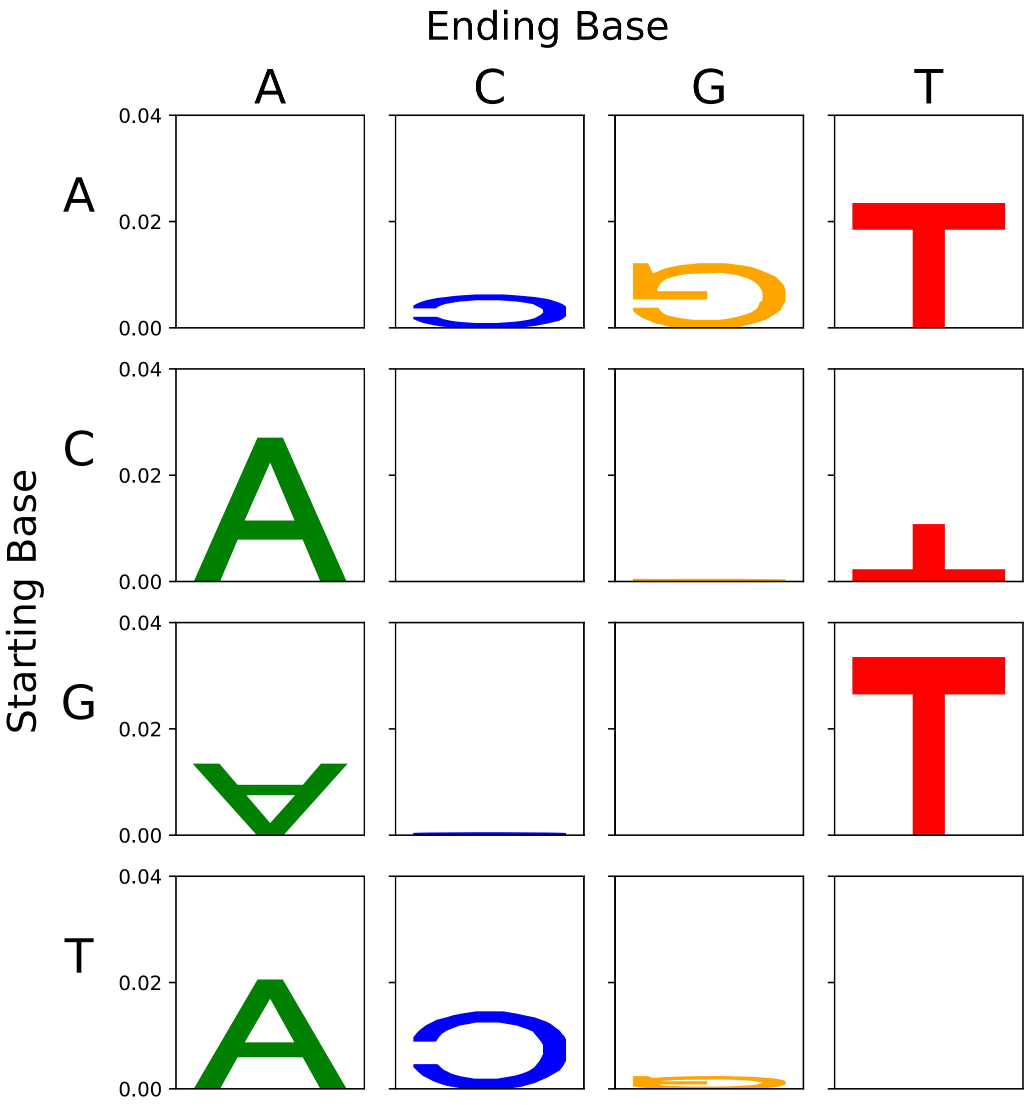


Figure B.1: Confirmation of the mutation spectra difference between the ENU-induced and spontaneous germline mutations. Starting and Ending Base correspond to X, Y respectively in X→Y. The y-axis is RE from the spectra hypothesis test and letter heights are as for the mutation motif logo. Letters in the normal orientation indicate an excess of that mutation direction in ENU-induced mutations relative to the spontaneous mutations. Inverted letters indicate a deficit in ENU-induced mutations relative to the spontaneous mutations. See Zhu et al. (2017) for a more detailed description of the log-linear models.

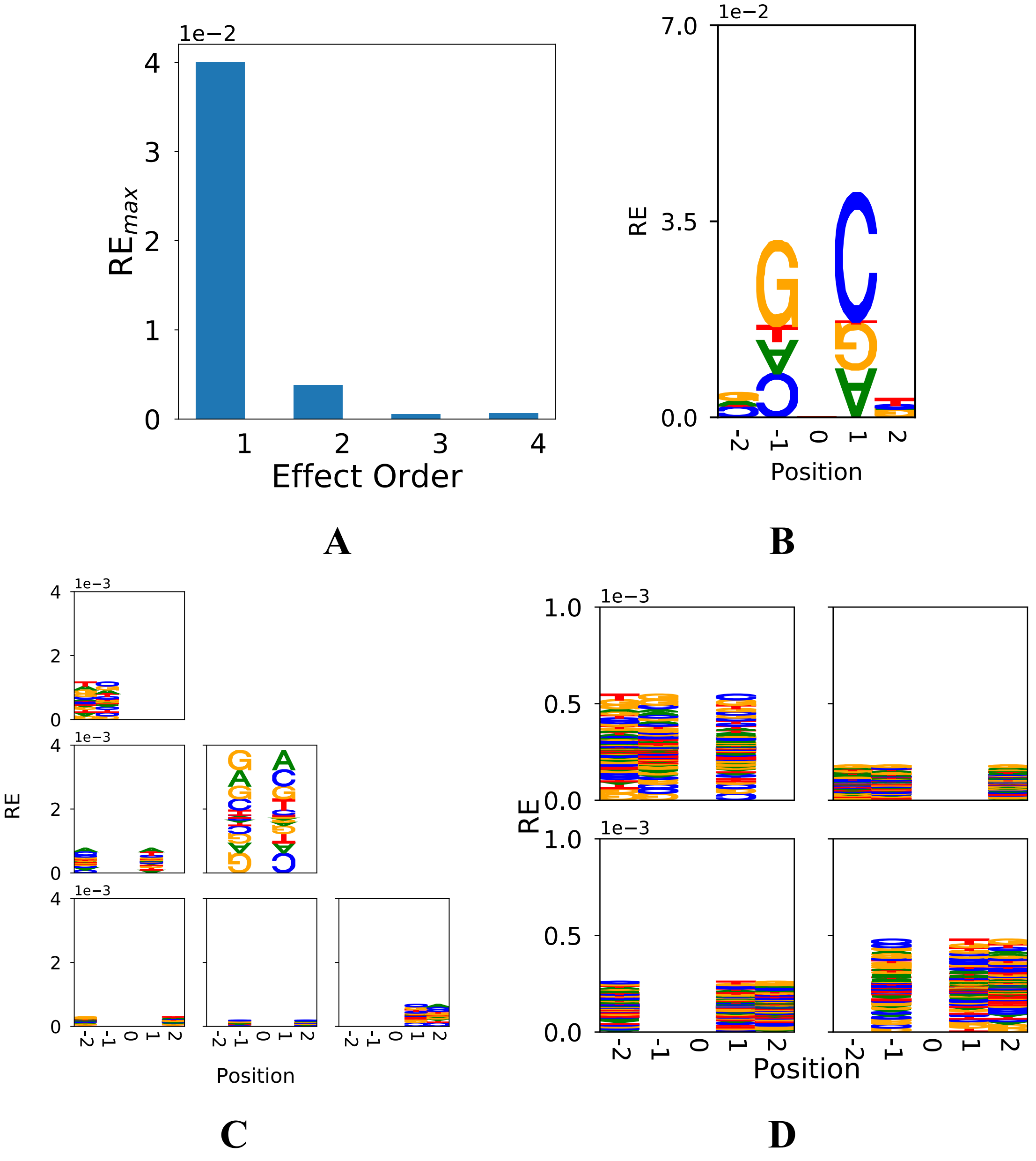
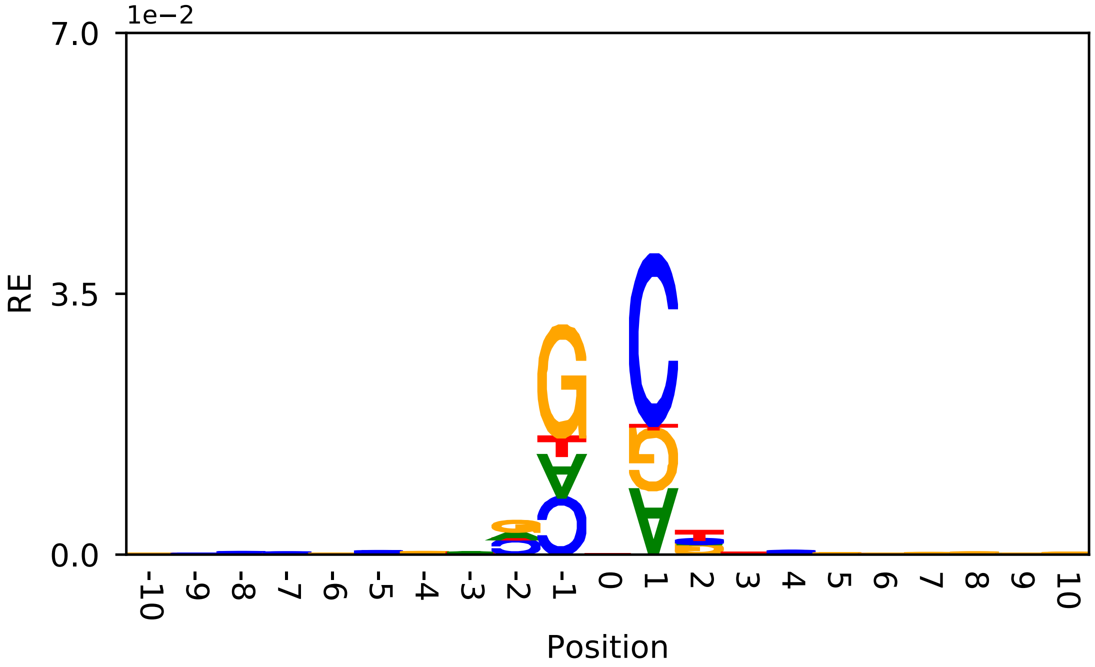
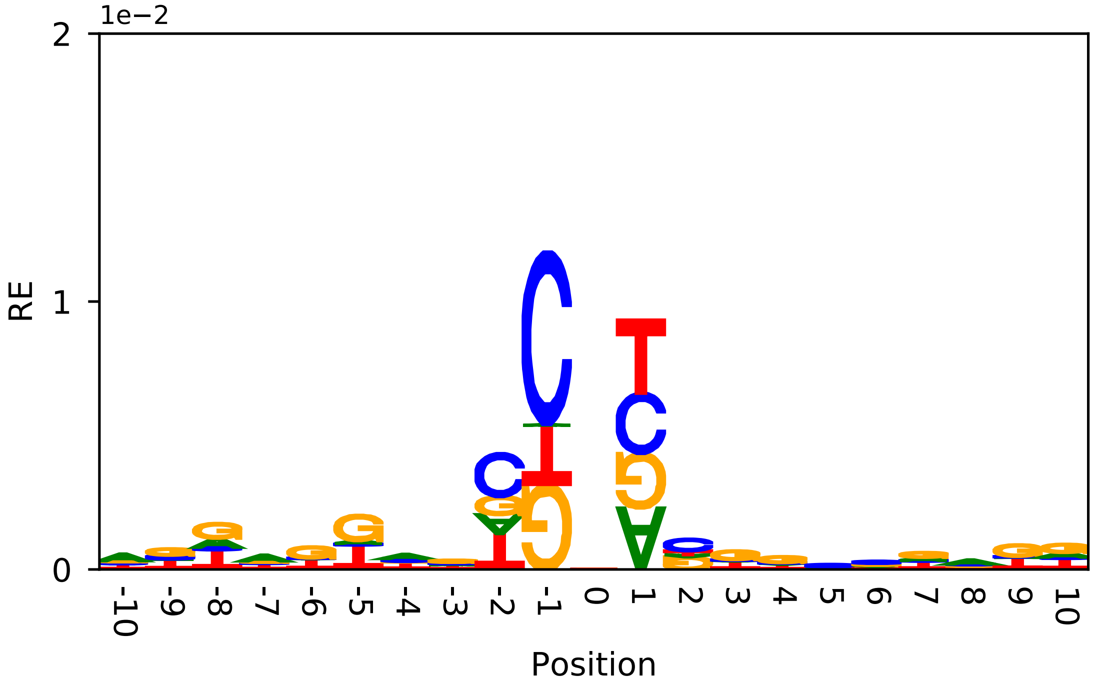


Figure B.2: Independent and second-order position effects dominate ENU-induced A→G point mutations. Note also that RE is largest for dependent effects among positions that are physically contiguous and overlap the mutated position at index 0. (A) Summary of the strength of associations by effect order. RE$_{max}$ is the maximum RE from any analysis for the indicated order; (B) The independent, or first-order, effects; (C) Second-order effects; and (D) Third-order effects.



(A) ENU-induced



(B) Spontaneous

Figure B.3: The physical extent of neighbourhood effects in the mouse. Mutation motifs are drawn from the results of the log-linear analysis of first-order effects (summarised in Table B.3). (A) ENU-induced germline mutations and (B) Spontaneous germline mutations.

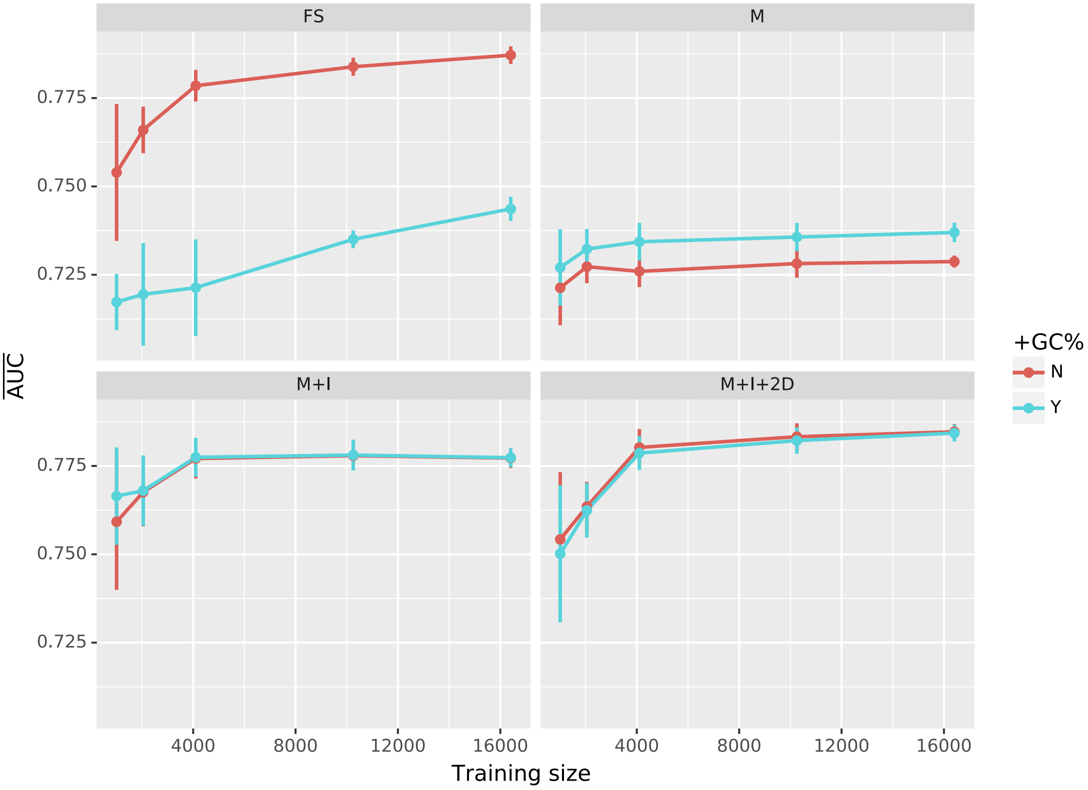


Figure B.4: Inclusion of GC% reduced performance when categorical neighbourhood features were included. The classifier including the GC% feature is indicated by a +GC% value of Y. The value N corresponds to the classifier with the strictly categorical feature set.

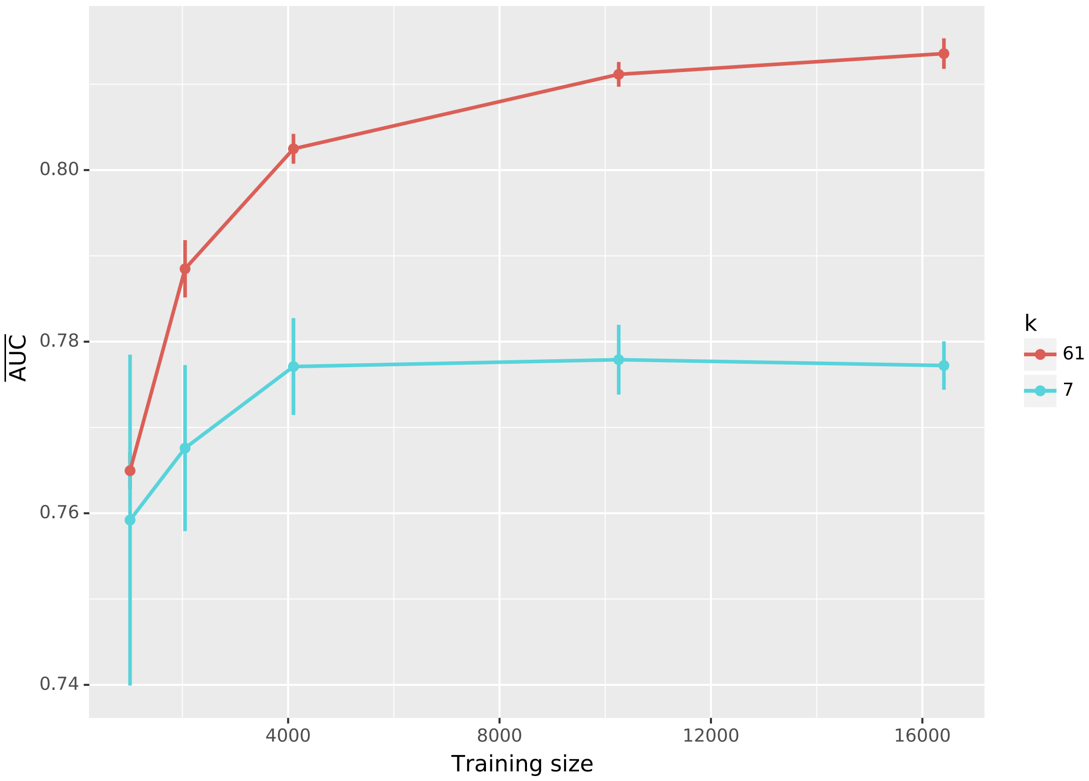


Figure B.5: The LR classifiers for 61-mer performed better than the 7-mer. x-axis is the size of the training sample, y-axis is the mean AUC obtained from implementing the M+I model, and error bars were calculated from the 5 chromosome 1 training samples.

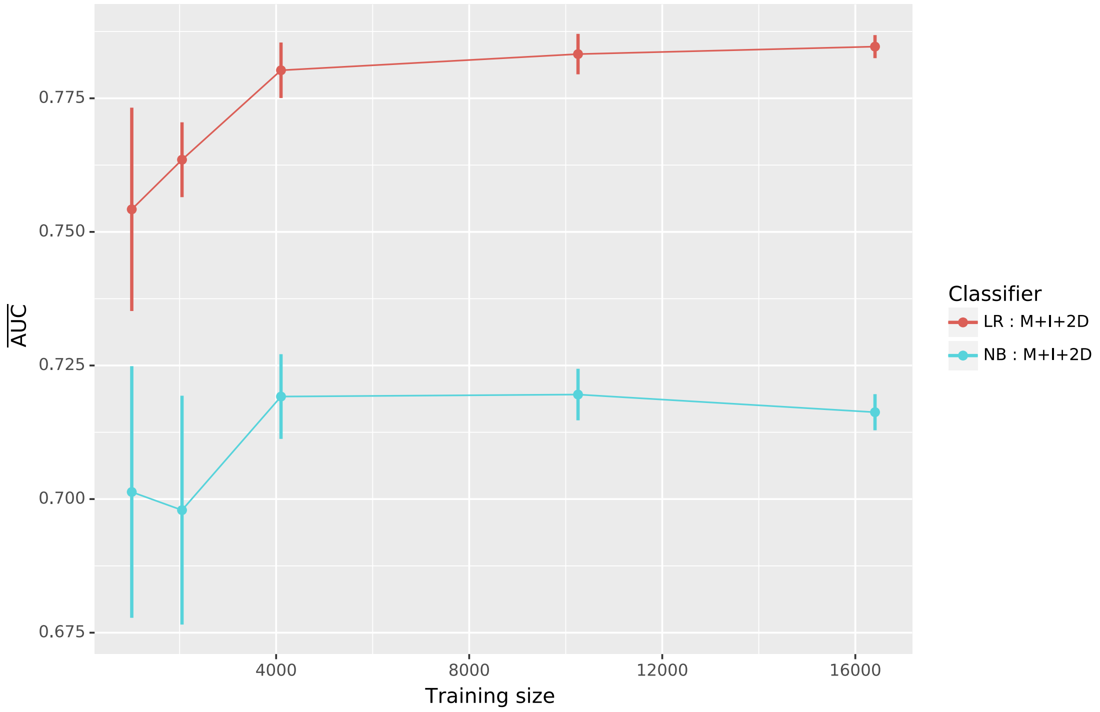


Figure B.6: The LR classifier performed better than the NB classifier. x-axis is the size of the training sample, y-axis is the mean AUC and error bars were calculated from the 5 chromosome 1 training samples.

References

Abdi, H. (2010). Holm's sequential bonferroni procedure. *Encyclopedia of research design*, 1(8):1–8.

Aggarwala, V. and Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature genetics*.

Aksoy, S. and Haralick, R. M. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern recognition letters*, 22(5):563–582.

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. *Nature*.

Altemose, N., Noor, N., Bitoun, E., Tumian, A., Imbeault, M., Chapman, J. R., Aricescu, A. R., and Myers, S. R. (2017). A map of human prdm9 binding provides evidence for novel behaviors of prdm9 and other zinc-finger proteins in meiosis. *Elife*, 6:e28383.

Alvarez, L., Comendador, M., and Sierra, L. (2003). Effect of nucleotide excision repair on enu-induced mutation in female germ cells of drosophila melanogaster. *Environmental and molecular mutagenesis*, 41(4):270–279.

Ananthaswamy, H. N. and Pierceall, W. E. (1990). Molecular mechanisms of ultraviolet radiation carcinogenesis. *Photochemistry and photobiology*, 52(6):1119–1136.

Andrews, T. D., Whittle, B., Field, M., Balakishnan, B., Zhang, Y., Shao, Y., Cho, V., Kirk, M., Singh, M., Xia, Y., et al. (2012). Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. *Open biology*, 2(5):120061.

Arnheim, N. and Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. *Nature Reviews Genetics*, 10(7):478.

Assis, R. and Kondrashov, A. S. (2011). Nonallelic gene conversion is not gc-biased in drosophila or primates. *Molecular biology and evolution*, 29(5):1291–1295.

Barbaric, I., Wells, S., Russ, A., and Dear, T. N. (2007). Spectrum of enu-induced mutations in phenotype-driven and gene-driven screens in the mouse. *Environmental and molecular mutagenesis*, 48(2):124–142.

Bauer, D. C., McMorran, B. J., Foote, S. J., and Burgio, G. (2015). Genome-wide analysis of chemically induced mutations in mouse in phenotype-driven screens. *BMC genomics*, 16(1):1.

Bauer, H., Demerec, M., and Kaufmann, B. P. (1938). X-ray induced chromosomal alterations in drosophila melanogaster. *Genetics*, 23(6):610.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Comput Biol*, 4(10):e1000173.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene*, 241(1):3–17.

Bernardi, G. (2001). Misunderstandings about isochores. part 1. *Gene*, 276(1-2):3–13.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science*, 228(4702):953–958.

Bhérer, C., Campbell, C. L., and Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature communications*, 8:14994.

Billings, T., Parvanov, E. D., Baker, C. L., Walker, M., Paigen, K., and Petkov, P. M. (2013). Dna binding specificities of the long zinc-finger recombination protein prdm9. *Genome biology*, 14(4):R35.

Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular biology and evolution*, 19(7):1181–1197.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., and Caporaso, J. G. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2's q2-feature-classifier plugin. *Microbiome*, 6(1):90.

Bouchard, C., Malina, R. M., and P'Russe, L. (1997). *Genetics of fitness and physical performance*. Human Kinetics.

Boulton, A., Myers, R. S., and Redfield, R. J. (1997). The hotspot conversion paradox and the evolution of meiotic recombination. *Proceedings of the National Academy of Sciences*, 94(15):8058–8063.

Boyce, R. P. and Howard-Flanders, P. (1964). Release of ultraviolet light-induced thymine dimers from dna in e. coli k-12. *Proceedings of the National Academy of Sciences*, 51(2):293–300.

Brown, T. (2002). *Genomes*. Wiley-Liss.

Brown, T. C. and Jiricny, J. (1989). Repair of base–base mismatches in simian and human cells. *Genome*, 31(2):578–583.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cadet, J., Anselmino, C., Douki, T., and Voituriez, L. (1992). New trends in photobiology: Photochemistry of nucleic acids in cells. *Journal of Photochemistry and Photobiology B: Biology*, 15(4):277–298.

Carty, M. P., El-Saleh, S., Zernik-Kobak, M., and Dixon, K. (1995). Analysis of mutations induced by replication of uv-damaged plasmid dna in hela cell extracts. *Environmental and molecular mutagenesis*, 26(2):139–146.

Carty, M. P., Hauser, J., Levine, A., and Dixon, K. (1993). Replication and mutagenesis of uv-damaged dna templates in human and monkey cell extracts. *Molecular and cellular biology*, 13(1):533–542.

Caruana, G., Farlie, P. G., Hart, A. H., Bagheri-Fam, S., Wallace, M. J., Dobbie, M. S., Gordon, C. T., Miller, K. A., Whittle, B., Abud, H. E., et al. (2013). Genome-wide enu mutagenesis in combination with high density snp analysis and exome sequencing provides rapid identification of novel mouse models of developmental disease. *PloS one*, 8(3).

Chahwan, R., Edelmann, W., Scharff, M. D., and Roa, S. (2012). Aiding antibody diversity by error-prone mismatch repair. In *Seminars in immunology*, volume 24, pages 293–300. Elsevier.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10):762.

Chor, B., Horn, D., Goldman, N., Levy, Y., and Massingham, T. (2009). Genomic dna k-mer spectra: models and modalities. *Genome Biol*, 10(10):R108.

Consortium, I. H. et al. (2003). The international hapmap project. *Nature*, 426(6968):789.

Consortium, I. H. G. S. et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860.

Consortium, M. G. S. et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520.

Cooke, M. S., Evans, M. D., Dizdaroglu, M., and Lunec, J. (2003). Oxidative dna damage: mechanisms, mutation, and disease. *The FASEB Journal*, 17(10):1195–1214.

Cooper, D. N. (1995). The nature and mechanisms of human gene mutation. *The metabolic and molecular bases of inherited disease*, pages 259–291.

Cooper, D. N. and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Human genetics*, 78(2):151–155.

Coulondre, C., Miller, J. H., Farabaugh, P. J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli. Nature*, 274(5673):775–780.

Darnell, J. E., Lodish, H. F., Baltimore, D., et al. (1990). *Molecular cell biology*, volume 2. Scientific American Books New York.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

Driscoll, D. J. and Migeon, B. R. (1990). Sex difference in methylation of single-copy genes in human meiotic germ cells: implications for x chromosome inactivation, parental imprinting, and origin of cpg mutations. *Somatic cell and molecular genetics*, 16(3):267–282.

Drummond, J. T. and Bellacosa, A. (2001). Human dna mismatch repair in vitro operates independently of methylation status at cpg sites. *Nucleic acids research*, 29(11):2234–2243.

Duane, W., Palmer, H., and Yeh, C.-S. (1921). A remeasurement of the radiation constant, h, by means of x-rays. *Proceedings of the National Academy of Sciences*, 7(8):237–242.

Duncan, B. K. and Miller, J. H. (1980). Mutagenic deamination of cytosine residues in dna. *Nature*, 287(5782):560.

Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10:285–311.

Duret, L., Mouchiroud, D., and Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in gc-rich isochores. *Journal of Molecular Evolution*, 40(3):308–317.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. (2002). Vanishing gc-rich isochores in mammalian genomes. *Genetics*, 162(4):1837–1847.

Ebersberger, I., Metzler, D., Schwarz, C., and Pääbo, S. (2002). Genomewide comparison of dna sequences between humans and chimpanzees. *The American Journal of Human Genetics*, 70(6):1490–1497.

Ehlert, T., Simon, P., and Moser, D. A. (2013). Epigenetics in sports. *Sports medicine*, 43(2):93–110.

Ellegren, H., Smith, N. G., and Webster, M. T. (2003). Mutation rate variation in the mammalian genome. *Current opinion in genetics & development*, 13(6):562–568.

Eyre-Walker, A. and Hurst, L. D. (2001). The evolution of isochores. *Nature Reviews Genetics*, 2(7):549.

Fajardo, L. F., Berthrong, M., and Anderson, R. E. (2001). *Radiation pathology.* Oxford University Press.

Feng, Z., Hu, W., Komissarova, E., Pao, A., Hung, M.-C., Adair, G. M., and Tang, M.-s. (2002). Transcription-coupled dna repair is genomic context dependent. *Journal of Biological Chemistry*.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2013). Ensembl 2014. *Nucleic acids research*, page gkt1196.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2014). Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43(D1):D805–D811.

Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C. M., Swertz, M., Wijmenga, C., van Ommen, G., et al. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature genetics*.

Friedberg, E. C., Walker, G. C., Siede, W., and Wood, R. D. (2005). *DNA repair and mutagenesis*. American Society for Microbiology Press.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). Gc-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159(2):907–911.

Gasser, R. and Zhu, X. (1999). Sequence-based analysis of enzymatically amplified dna fragments by mutation detection techniques. *Parasitology Today*, 15(11):462–465.

Gershenson, S. (1986). Viruses as environmental mutagenic factors. *Mutation Research/Reviews in Genetic Toxicology*, 167(3):203–213.

Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O., and Petes, T. D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 97(21):11383–11390.

Gibson, W. T. (2009). Key concepts in human genetics: understanding the complex phenotype. In *Genetics and Sports*, volume 54, pages 1–10. Karger Publishers.

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. (2015). Quantification of gc-biased gene conversion in the human genome. *Genome research*, pages gr–185488.

Graur, D. and Li, W.-H. (2000). Fundamentals of molecular evolution.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets:"function" in the human genome according to the evolution-free gospel of encode. *Genome biology and evolution*, 5(3):578–590.

Grippo, P., Iaccarino, M., Parisi, E., and Scarano, E. (1968). Methylation of dna in developing sea urchin embryos. *Journal of molecular biology*, 36(2):195–208.

Hainaut, P. and Pfeifer, G. P. (2001). Patterns of p53 g t transversions in lung cancers reflect the primary mutagenic signature of dna-damage by tobacco smoke. *Carcinogenesis*, 22(3):367–374.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239.

Haldane, J. (1946). The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Annals of eugenics*, 13(1):262–271.

Haldane, J. B. (1935). The rate of spontaneous mutation of a human gene. *Journal of Genetics*, 31(3):317.

Haldane, J. B. S. (1948). Croonian lecture-the formal genetics of man. *Proc. R. Soc. Lond. B*, 135(879):147–170.

Hardison, R. C., Roskin, K. M., Yang, S., Diekhans, M., Kent, W. J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome research*, 13(1):13–26.

Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences*, 112(11):3439–3444.

Hartl, D. L., Clark, A. G., and Clark, A. G. (1997). *Principles of population genetics*, volume 116. Sinauer associates Sunderland, MA.

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, 22(2):160–174.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*, 15(9):585–598.

Hellmann, I., Prüfer, K., Ji, H., Zody, M. C., Pääbo, S., and Ptak, S. E. (2005). Why do human diversity levels vary at a megabase scale? *Genome research*, 15(9):1222–1231.

Henderson, J. P., Byun, J., Takeshita, J., and Heinecke, J. W. (2003). Phagocytes produce 5-chlorouracil and 5-bromouracil, two mutagenic products of myeloperoxidase, in human inflammatory tissue. *Journal of Biological Chemistry*, 278(26):23522–23528.

Hirano, T., Kazama, Y., Ishii, K., Ohbu, S., Shirakawa, Y., and Abe, T. (2015). Comprehensive identification of mutations induced by heavy-ion beam irradiation in a rabidopsis thaliana. *The Plant Journal*, 82(1):93–104.

Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766.

Hoeijmakers, J. H. (2001). Genome maintenance mechanisms for preventing cancer. *nature*, 411(6835):366.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Hu, X., Li, H., Ding, J., and Han, S. (2004). Mutagenic mechanism of the at to gc transition induced by 5-bromouracil: an ab initio study. *Biochemistry*, 43(21):6361–6369.

Huttley, G. A. (2004). Modeling the impact of dna methylation on the evolution of brca1 in mammals. *Molecular biology and evolution*, 21(9):1760–1768.

Huttley, G. A., Jakobsen, I. B., Wilson, S. R., and Easteal, S. (2000). How important is dna replication for mutagenesis? *Molecular biology and evolution*, 17(6):929–937.

Hwang, D. G. and Green, P. (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):13994–14001.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.

Iyer, R. R., Pluciennik, A., Burdett, V., and Modrich, P. L. (2006). Dna mismatch repair: functions and mechanisms. *Chemical reviews*, 106(2):302–323.

Jabbari, K. and Bernardi, G. (1998). Cpg doublets, cpg islands and alu repeats in long human dna sequences from different isochore families. *Gene*, 224(1):123–128.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer.

Justice, M. J., Noveroske, J. K., Weber, J. S., Zheng, B., and Bradley, A. (1999). Mouse enu mutagenesis. *Human molecular genetics*, 8(10):1955–1963.

Kaehler, B. D., Yap, V. B., and Huttley, G. A. (2017). Standard codon substitution models overestimate purifying selection for nonstationary data. *Genome biology and evolution*, 9(1):134–149.

Kaehler, B. D., Yap, V. B., Zhang, R., and Huttley, G. A. (2015). Genetic distance for a general non-stationary markov substitution process. *Systematic biology*, 64(2):281–293.

Kariin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends in genetics*, 11(7):283–290.

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Current opinion in microbiology*, 1(5):598–610.

Karlin, S., Campbell, A. M., and Mrázek, J. (1998). Comparative DNA analysis across diverse genomes. *Annual review of genetics*, 32(1):185–225.

Kidess, E., Heirich, K., Wiggin, M., Vysotskaia, V., Visser, B. C., Marziali, A., Wiedenmann, B., Norton, J. A., Lee, M., Jeffrey, S. S., et al. (2015). Mutation profiling of tumor dna from plasma and tumor tissue of colorectal cancer patients with a novel, high-sensitivity multiplexed mutation detection platform. *Oncotarget*, 6(4):2549.

Kliman, R. M. and Eyre-Walker, A. (1998). Patterns of base composition within the genes of drosophila melanogaster. *Journal of molecular evolution*, 46(5):534–541.

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J. G., Easton, B. C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., et al. (2007). Pycogent: a toolkit for making sense from sequence. *Genome Biol*, 8(8):R171.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. *Nature genetics*, 31(3):241.

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099.

Kow, Y. W. (2002). Repair of deaminated bases in dna12. *Free Radical Biology and Medicine*, 33(7):886–893.

Krawczak, M., Ball, E. V., and Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *The American Journal of Human Genetics*, 63(2):474–488.

Latrille, T., Duret, L., and Lartillot, N. (2017). The red queen model of recombination hot-spot evolution: a theoretical investigation. *Phil. Trans. R. Soc. B*, 372(1736):20160463.

Lee, J., Cox, B. D., Daly, C. M., Lee, C., Nuckels, R. J., Tittle, R. K., Uribe, R. A., and Gross, J. M. (2012). An enu mutagenesis screen in zebrafish for visual system mutants identifies a novel splice-acceptor site mutation in patched2 that results in colobomas. *Investigative ophthalmology & visual science*, 53(13):8214–8221.

Li, G.-M. (2008). Mechanisms and functions of dna mismatch repair. *Cell research*, 18(1):85.

Li, W.-H., Yi, S., and Makova, K. (2002). Male-driven evolution. *Current opinion in genetics & development*, 12(6):650–656.

Lohmueller, K. E., Albrechtsen, A., Li, Y., Kim, S. Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A. F., Grarup, N., et al. (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet*, 7(10):e1002326.

Machida, K., McNamara, G., Cheng, K. T.-H., Huang, J., Wang, C.-H., Comai, L., Ou, J.-H. J., and Lai, M. M. (2010). Hepatitis c virus inhibits dna damage repair through reactive oxygen and nitrogen species and by interfering with the atm-nbs1/mre11/rad50 dna repair pathway in monocytes and hepatocytes. *The Journal of Immunology*, page 1000618.

Marais, G., Mouchiroud, D., and Duret, L. (2001). Does recombination improve selection on codon usage? lessons from nematode and fly complete genomes. *Proceedings of the National Academy of Sciences*, 98(10):5688–5692.

Marais, G. and Piganeau, G. (2002). Hill-robertson interference is a minor determinant of variations in codon bias across drosophila melanogaster and caenorhabditis elegans genomes. *Molecular biology and evolution*, 19(9):1399–1406.

Mellon, I., Spivak, G., and Hanawalt, P. C. (1987). Selective removal of transcription-blocking dna damage from the transcribed strand of the mammalian dhfr gene. *Cell*, 51(2):241–249.

Meunier, J. and Duret, L. (2004). Recombination drives the evolution of gc-content in the human genome. *Molecular biology and evolution*, 21(6):984–990.

Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K., and Yasunaga, T. (1987). Male-driven molecular evolution: a model and nucleotide sequence analysis. In *Cold Spring Harbor symposia on quantitative biology*, volume 52, pages 863–867. Cold Spring Harbor Laboratory Press.

Montoya-Burgos, J. I., Boursot, P., and Galtier, N. (2003). Recombination explains isochores in mammalian genomes. *Trends in genetics*, 19(3):128–130.

Morton, B. R., Oberholzer, V. M., and Clegg, M. T. (1997). The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome. *Journal of molecular evolution*, 45(3):227–231.

Mouret, S., Baudouin, C., Charveron, M., Favier, A., Cadet, J., and Douki, T. (2006). Cyclobutane pyrimidine dimers are predominant dna lesions in whole human skin exposed to uva radiation. *Proceedings of the National Academy of Sciences*, 103(37):13765–13770.

Mugal, C. F., Weber, C. C., and Ellegren, H. (2015). Gc-biased gene conversion links the recombination landscape and demography to genomic base composition: Gc-biased gene conversion drives genomic base composition across a wide range of species. *Bioessays*, 37(12):1317–1326.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying dna microarray data. *Journal of computational biology*, 10(2):119–142.

Münger, K., Phelps, W., Bubb, V., Howley, P., and Schlegel, R. (1989). The e6 and e7 genes of the human papillomavirus type 16 together are necessary and sufficient for transformation of primary human keratinocytes. *Journal of virology*, 63(10):4417–4421.

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, 327(5967):876–879.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences*, 80(20):6278–6281.

Narisawa-Saito, M. and Kiyono, T. (2007). Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: Roles of e6 and e7 proteins. *Cancer science*, 98(10):1505–1511.

Nei, M. (2013). *Mutation-driven evolution*. OUP Oxford.

Nevarez, P. A., DeBoever, C. M., Freeland, B. J., Quitt, M. A., and Bush, E. C. (2010). Context dependent substitution biases vary within the human genome. *BMC bioinformatics*, 11(1):462.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

Nichols, W. W. (1970). Virus-induced chromosome abnormalities. *Annual Reviews in Microbiology*, 24(1):479–500.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993.

Nishino, H., Buettner, V. L., Haavik, J., Schaid, D. J., and Sommer, S. S. (1996). System issues: Spontaneous mutation in big blue® transgenic mice: Analysis of age, gender, and tissue type. *Environmental and molecular mutagenesis*, 28(4):299–312.

Noveroske, J., Weber, J., and Justice, M. (2000). The mutagenic action of n-ethyl-n-nitrosourea in the mouse. *Mammalian genome*, 11(7):478–483.

Paces, J., Zıka, R., Paces, V., Pavlıcek, A., Clay, O., and Bernardi, G. (2004). Representing gc variation along eukaryotic chromosomes. *Gene*, 333:135–141.

Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., Struhl, K., and Weng, Z. (2007). Nucleosome positioning signals in genomic dna. *Genome research*, 17(8):000–000.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Peichel, C. L. and Marques, D. A. (2017). The genetic and molecular architecture of phenotypic diversity in sticklebacks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1713):20150486.

Peltomäki, P. (2001). Deficient dna mismatch repair: a common etiologic factor for colon cancer. *Human Molecular Genetics*, 10(7):735–740.

Peltomaki, P. and Vasen, H. (1997). Mutations predisposing to hereditary nonpolyposis colorectal cancer: database and results of a collaborative study. the international collaborative group on hereditary nonpolyposis colorectal cancer. *Gastroenterology*, 113(4):1146–1158.

Petranović, M., Vlahović, K., Zahradka, D., Dzidić, S., and Radman, M. (2000). Mismatch repair in xenopus egg extracts is not strand-directed by dna methylation. *Neoplasma*, 47(6):375–381.

Pfeifer, G. P., You, Y.-H., and Besaratinia, A. (2005). Mutations induced by ultraviolet light. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 571(1):19–31.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191.

Powers, N. R., Parvanov, E. D., Baker, C. L., Walker, M., Petkov, P. M., and Paigen, K. (2016). The meiotic recombination activator prdm9 trimethylates both h3k36 and h3k4 at recombination hotspots in vivo. *PLoS genetics*, 12(6):e1006146.

Prosperi, M. C., Altmann, A., Rosen-Zvi, M., Aharoni, E., Borgulya, G., Bazso, F., Sönnerborg, A., Schülter, E., Struck, D., Ulivi, G., et al. (2009). Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. *Antivir Ther*, 14(3):433–42.

Rak, J., Chomicz, L., Wiczk, J., Westphal, K., Zdrowowicz, M., Wityk, P., Żyndul, M., Makurat, S., and Golon, Ł. (2015). Mechanisms of damage to dna labeled with electrophilic nucleobases induced by ionizing or uv radiation. *The Journal of Physical Chemistry B*, 119(26):8227–8238.

Riley, T., Sontag, E., Chen, P., and Levine, A. (2008). Transcriptional control of human p53-regulated genes. *Nature reviews Molecular cell biology*, 9(5):402.

Robinson, M. C., Stone, E. A., and Singh, N. D. (2013). Population genomic analysis reveals no evidence for gc-biased gene conversion in drosophila melanogaster. *Molecular biology and evolution*, 31(2):425–433.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science*, 323(5915):737–741.

Schneider, T. D. (2006). Twenty-five years of delila and molecular information theory. *Biological theory*, 1(3):250–260.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *Journal of molecular biology*, 188(3):415–431.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

Shiraishi, Y., Tremmel, G., Miyano, S., and Stephens, M. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet*, 11(12):e1005657.

Shrivastav, N., Li, D., and Essigmann, J. M. (2010). Chemical biology of mutagenesis and dna repair: cellular responses to dna alkylation. *Carcinogenesis*, 31(1):59–70.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050.

Šmarda, P., Bureš, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., Tichỳ, L., Grulich, V., and Rotreklová, O. (2014). Ecological and evolutionary significance of genomic gc content diversity in monocots. *Proceedings of the National Academy of Sciences*, 111(39):E4096–E4102.

Sonnenburg, S. (2008). *Machine Learning for Genomic Sequence Analysis-Dissertation*. PhD thesis, Berlin Institute of Technology.

Spies, M. and Fishel, R. (2015). Mismatch repair during homologous and homeologous recombination. *Cold Spring Harbor perspectives in biology*, 7(3):a022657.

Storchova, Z. and Pellman, D. (2004). From polyploidy to aneuploidy, genome instability and cancer. *Nature reviews Molecular cell biology*, 5(1):45.

Stottmann, R. and Beier, D. (2014). Enu mutagenesis in the mouse. *Current protocols in human genetics*, 82(1):15–4.

Stottmann, R. W. and Beier, D. R. (2010). Using enu mutagenesis for phenotype-driven analysis of the mouse. In *Methods in enzymology*, volume 477, pages 329–348. Elsevier.

Strauss, B. (1968). Dna repair mechanisms and their relation to mutation and recombination. Technical report, Univ. of Chicago.

Svejstrup, J. Q. (2002). Transcription: Mechanisms of transcription-coupled dna repair. *Nature Reviews Molecular Cell Biology*, 3(1):21.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. (1983). The double-strand-break repair model for recombination. *Cell*, 33(1):25–35.

Takahasi, K. R., Sakuraba, Y., and Gondo, Y. (2007). Mutational pattern and frequency of induced nucleotide changes in mouse enu mutagenesis. *BMC molecular biology*, 8(1):1.

Tanaka, A., Shikazono, N., and Hase, Y. (2010). Studies on biological effects of ion beams on lethality, molecular nature of mutation, mutation rate, and spectrum of mutation phenotype for mutation breeding in higher plants. *Journal of radiation research*, 51(3):223–233.

Terzaghi, B. E., Streisinger, G., and Stahl, F. W. (1962). The mechanism of 5-bromouracil mutagenesis in the bacteriophage t4. *Proceedings of the National Academy of Sciences*, 48(9):1519–1524.

Toledo, F. and Wahl, G. M. (2006). Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nature Reviews Cancer*, 6(12):909.

Tornaletti, S. and Pfeifer, G. P. (1996). Uv damage and repair mechanisms in mammalian cells. *Bioessays*, 18(3):221–228.

Touchon, M., Nicolay, S., Arnéodo, A., d'Aubenton Carafa, Y., and Thermes, C. (2003). Transcription-coupled ta and gc strand asymmetries in the human genome. *FEBS letters*, 555(3):579–582.

Vicoso, B. and Charlesworth, B. (2006). Evolution on the x chromosome: unusual patterns and processes. *Nature Reviews Genetics*, 7(8):645.

Viel, A., Bruselles, A., Meccia, E., Fornasarig, M., Quaia, M., Canzonieri, V., Policicchio, E., Urso, E. D., Agostini, M., Genuardi, M., et al. (2017). A specific mutational signature associated with dna 8-oxoguanine persistence in mutyh-defective colorectal cancer. *EBioMedicine*.

von Schwerin, A. (1920). Close to the state and oriented toward the fundamentals-research on radiation and radioactivity in the biosciences 1920-1970.

Wålinder, A. (2014). Evaluation of logistic regression and random forest classification based on prediction accuracy and metadata analysis.

Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., and Ikemura, T. (2002). Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Human Molecular Genetics*, 11(1):13–21.

Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

Webster, M. T., Smith, N. G., and Ellegren, H. (2003). Compositional evolution of noncoding dna in the human and chimpanzee genomes. *Molecular biology and evolution*, 20(2):278–286.

Webster, M. T., Smith, N. G., Hultin-Rosenberg, L., Arndt, P. F., and Ellegren, H. (2005). Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Molecular biology and evolution*, 22(6):1468–1474.

Werness, B. A., Levine, A. J., and Howley, P. M. (1990). Association of human papillomavirus types 16 and 18 e6 proteins with p53. *Science*, 248(4951):76–79.

Witkin, E. M. (1969). Ultraviolet-induced mutation and dna repair. *Annual Review of Genetics*, 3(1):525–552.

Woods, R. J. and Pikaev, A. K. (1994). *Applied radiation chemistry: radiation processing*. John Wiley & Sons.

Woofle, A., Goodson, M., Goode, D., Snell, P., Smith, S., Vavouri, T., McEwen, G., Gilks, W., Walter, K., Abnizova, I., et al. (2005). Highly conserved non coding sequences are associated with developmental control genes in vertebrates. *PloS Biol*, 3:e7.

Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic acids research*, 34(2):564–574.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–1650.

Ying, H. and Huttley, G. (2011). Exploiting cpg hypermutability to identify phenotypically significant variation within human protein-coding genes. *Genome biology and evolution*, 3:938–949.

Yockey, H. P. (2005). *Information theory, evolution, and the origin of life*. Cambridge University Press.

Zhang, F. and Zhao, Z. (2004). The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human snps. *Genomics*, 84(5):785–795.

Zhang, X. and Mathews, C. K. (1995). Natural dna precursor pool asymmetry and base sequence context as determinants of replication fidelity. *Journal of Biological Chemistry*, 270(15):8401–8404.

Zhao, Z. and Boerwinkle, E. (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome research*, 12(11):1679–1686.

Zhu, Y., Neeman, T., Yap, V. B., and Huttley, G. A. (2017). Statistical methods for identifying sequence motifs affecting point mutations. *Genetics*, 205(2):843–856.