

Imperial College London
Institute of Clinical Sciences

Biased generation and retention of oligonucleotide motifs in the human genome

Vassili Feodorovich Kusmartsev

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy from Imperial College London, January 2021

Statement of Originality

I confirm that all work in this thesis is my own, with discussion on inception of study with my supervisor Tobias Warnecke. Any previous literature is appropriately referenced.

Regarding Chapter 1; in the final publication [Kusmartsev et al., 2020] there is additional work looking at somatic mutations in cancer cell lines, done by an MSc student, Magdalena Drozd, that I supervised together with Tobias Warnecke and Benjamin Schuster-Böckler. I have not included this analysis as part of this thesis as it was not my work. A study with such conservative controls to investigate the mutational effect surrounding methylated cytosine at fine-scale resolution, being the couple of bases up and downstream, has not previously been done, and so this work is completely original as of writing.

Regarding Chapter 2; to my current knowledge there have been no other studies utilising derived motif frequencies to investigate the selection driving the avoidance of oligonucleotide motifs. Hence, this work is also entirely original as of writing.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY).

Under this license, you may copy and redistribute the material in any medium or format for both commercial and non-commercial purposes. You may also create and distribute modified versions of the work. This on the condition that you credit the author.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

During evolution, variation is created by mutational processes in an organism's genome, and then these variants remain or are purged as a result of selective pressures and drift. Mutation rates are biased at many scales across the genome, and the selective pressure acting on the generated variation as a result of mutations also varies greatly.

One well studied example of mutational heterogeneity at single-base resolution concerns methylated cytosines, which mutate at an approximately 10-fold higher rate than their unmethylated counterpart. In Chapter 1, I demonstrate that cytosine methylation not only affects mutation risk at the methylated base but also exerts mutational effects on its neighbouring nucleotides. In human, bases surrounding the methylated cytosine accrue fewer SNPs than bases in the vicinity of unmethylated sites. In plants, on the other hand, I demonstrate that the opposite is true. I show that this difference is not driven by differential sampling of sequence or chromatin contexts, and is instead likely due to alternative lesion processing between the species.

Regarding selective heterogeneity, studies in the past have largely focused on differential conservation of genetic elements between species, an approach ill-equipped to capture selection against gaining a novel deleterious function. In Chapter 2, I develop and validate a novel approach to interrogate selection against deleterious gain-of-function motifs in protein coding sequences, which is based on derived allele frequency distributions. I show how derived allele frequencies can be used to detect selection against particular amino acid motifs genome-wide.

This thesis elucidates forces biasing the creation and persistence of genetic variation and highlights the entwined nature of mutational bias, selective pressures, and the technical challenges involved.

Acknowledgements

What an incredible journey these four and a bit years have been! I would like to thank the Medical Research Council (now UKRI) for funding the studentship, and Imperial College London for hosting me, without both of which this adventure would not have been possible.

Firstly, I must express immense gratitude to my supervisor - Toby, you have been an amazing role model throughout the PhD. You always gave me enough agency to develop my own ideas and approaches to answering questions, whilst always correcting course whenever I got side-tracked (as often happened!). I have learned so much under your guidance. You also gave so much support outside of work, from showing me who is also the boss on the minigolf pitch, to offering to “shoot some hoops” in the darkest hours of my thesis write-up. Thank you for everything, and I look forward to working together again in the future.

I have to thank the Molecular Systems research group, who made the office feel so much like home. Especially Sasha, from the moment you befriended me during my interview, you made me feel so welcome at the institute. Val, Maria, Shivani, Antoine, Jake, Katie - thank you all for always keeping an eye out for me, for all the patience and advice you offered for my work that was so different from your own, and all the beers and social times we have shared; I could not have asked for a better group to be a part of.

I would like to thank my other mentors, Matthias and Vahid, and all the other group leaders who have asked such insightful questions during my presentations - you have taught me how to see the pertinent links in my work, as well as how to apply that to research in general.

Thank you to all my other colleagues and friends, both old and new, for making these last four years of my life so much fun. Our computational neighbours, the CRG research group (otherwise known as “Boris’s Babes”): Leonie, James, Nash, Eleni, Radi, Nejc, Nevena, Damir - thank you for creating so many moments that I will cherish. The rest of my PhD cohort: Lisa, Alec, Matt, Helen, Domnhall, Chiara, Irene - thank you for being in this together. And thank you to all my other friends both at the institute, from across ICTEM to the other buildings, and outside - I wish I could list you all, but the list would be too long.

Finally, I’d like to thank my family - the unerring rock that I could always lean upon. My parents - thank you for always checking in, for all the scientific discussions from the viewpoint

of physicists, and for all the advice that I never wanted but that always turned out to be incredibly helpful. My sister Anya and brother-in-law Zoltan for always sharing their experience and knowledge. And special dedication to my late grandma, who was always wise; and my grandpa, who always smiles.

Acronyms

- 5mC: 5-methylCytosine
- APOBEC: apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
- BER: Base Excision Repair
- bp: base pair
- CDF: Cumulative Density Function
- CPD: Cyclobutane Pyrimidine Dimer
- DMF: Derived Motif Frequency spectrum
- ESC: Embryonic Stem Cell
- ExAC: Exome Aggregation Consortium
- k-mer: A motif of length k letters
- gnomAD: genome Aggregation Database
- GoF: Gain of Function
- HMM: Hidden Markov Model
- LoF: Loss of Function
- MMR: Mismatch repair
- NER: Nucleotide excision repair

- nt: nucleotide
- RdDM: RNA-directed DNA methylation
- RR_{met} : Relative mutational risk associated with methylation
- SNP: single nucleotide polymorphism
- WT: Wild-Type

Contents

Statement of Originality	1
Copyright Declaration	2
Abstract	3
Acknowledgements	4
Acronyms	6
Introduction	16
1 Does cytosine methylation affect mutation dynamics beyond the base itself?	21
1.1 Introduction	21
1.1.1 DNA modifications as mutagenic agents	24
1.1.2 Effect of context on CpG mutability	25
1.1.3 Effect of 5mC on context mutability	26
1.2 Results and Discussions	28
1.2.1 Methylation is associated with decreased mutability of neighbouring bases in human	28

1.2.2	Methylation increases incidence of SNPs in neighbourhood of 5mC in plants	31
1.2.3	GC-content, chromatin state, and sequence context do not drive the mutational risk pattern of methylation	34
1.2.4	An interlude: The differences in methylation and demethylation machineries in plants and human	42
1.2.5	The effect of methylation and demethylation machinery in plants	49
1.2.6	Discussion: potential mechanisms and an evolutionary explanation . . .	52
1.2.7	Inconsistency with previous work by Qu et al - Mapping Artifacts	55
1.3	Conclusions	60
1.4	Methods	61
1.4.1	Analysis of relative mutational risk - pipeline	61
1.4.2	Methylation basecalls	62
1.4.3	Population SNP Datasets	63
1.4.4	Chromatin state information	63
1.4.5	Genome versions	63
1.4.6	Genome masking	64
1.4.7	Assignment of plant methylation machinery zones	64
1.4.8	Replicating the Qu et al. protocol	65
1.4.9	Data availability statement	65
1.5	Publications	66

2	Avoidance selection of Gain of Function mutations	67
2.1	Introduction	67
2.1.1	Investigation of selection - conservation of genetic elements	67
2.1.2	Compositional approaches	69
2.1.3	Novel approach - Derived Motif Frequency Spectra	70
2.2	Results	71
2.2.1	How to compare highly skewed distributions?	72
2.2.2	First step: single amino acid motifs	79
2.2.3	Ancestral independence of derived patterns	81
2.2.4	Return to single amino acids	94
2.2.5	Quantifying the effect of the chemical properties of amino acids on their avoidance	101
2.2.6	Extension to longer motif patterns	107
2.2.7	The avoidance of dipeptides	108
2.2.8	Dipeptides as a product of single amino acids	111
2.2.9	Asymmetry of dipeptide avoidance	112
2.2.10	Derived tripeptides	114
2.2.11	Tripeptides as a function of dipeptides	117
2.3	Conclusions	120
2.3.1	Future directions	120
2.3.2	Limits of the approach	121

2.3.3	Conclusion	122
2.4	Methods	123
2.4.1	Derived Motif Frequency Spectra	123
2.4.2	Cumulative Density Functions	124
2.4.3	Ancestral Genome	124
2.4.4	Population Variation Data	125
2.4.5	Coding Sequences	125
2.4.6	Genome versions	126
2.4.7	Data availability statement	126
	Concluding remarks	126
	Bibliography	128

List of Tables

2.1	Top chemical property correlates for each eigenvector	105
-----	---	-----

List of Figures

1.1	A schematic illustrating the CpG pairing method	30
1.2	Relative mutational risk of methylation in human H1 ESCs	32
1.3	Relative mutational risk of methylation in human H1 ESCs. <i>A.thaliana</i> , and rice	35
1.4	Relative mutational risk of methylation at CHG and CHH contexts in <i>A.thaliana</i> and rice	36
1.5	Effect of sequence context, regional GC-content, and chromatin context on RR_{met}	38
1.6	Effect of subsampling sequence contexts between human and <i>A. thaliana</i> on RR_{met}	39
1.7	Relative mutational risk of methylation as a function of motif GC content bins in human, <i>A.thaliana</i> , and rice	40
1.8	Effect of chromatin contexts in <i>A. thaliana</i> on SNP incidence between methy- lated and unmethylated sites	43
1.9	Effect of chromatin contexts in human on SNP incidence between methylated and unmethylated sites	44
1.10	Effect of chromatin contexts in rice on SNP incidence between methylated and unmethylated sites	45
1.11	A schematic illustrating RNA-direct DNA methylation in plants	47
1.12	RR_{met} in <i>A.thaliana</i> as a function of NRPD1, ROS1, or DDM1 activity	51

1.13	Relative mutational risk of methylation in human sperm	58
1.14	Testing confounders in Qu et al. [2012] analysis	59
2.1	Schematic of generating derived motif frequency spectra	72
2.2	Distribution of allele frequencies in gnomAD exome dataset	73
2.3	Correlations of different distribution metrics on derived amino acid codons . . .	75
2.4	Schematic of cumulative density functions	77
2.5	Demonstration of DMF CDF for single amino acids	78
2.6	The DMFs of single amino acids in coding regions	81
2.7	Derived amino acid codons split by the ancestral codon	82
2.8	A PCA of 2 amino acid derived motifs	83
2.9	Ancestral dependence of nonsynonymous derived single amino acid codons with- out CpG deaminations	85
2.10	A PCA of 7-nt ancestral motif SNPs without CpG deaminations, highlighting ApT transitions	86
2.11	Submotifs contributing to ApT transition mutational bias	87
2.12	Loss of ancestral against gain of derived single amino acid codons without CpG deaminations	89
2.13	Effect of ancestral sequence complexity on DMF	91
2.14	Effect of ancestral GC-content on DMF of 3 amino acid long motifs	93
2.15	Effect of ancestral GC-content on DMF of 3 amino acid long motifs	94
2.16	PCA of the effect of ancestral sequence complexity on DMF	95

2.17 Geometric means of derived single amino acids with and without CpG or ApT transition SNPs	97
2.18 Geometric means of derived single amino acid codons with and without CpG or ApT transition SNPs	98
2.19 The proportion of singletons of derived single amino acids with and without CpG or ApT transition SNPs	99
2.20 The proportion of singletons of derived single amino acid codons with and without CpG or ApT transition SNPs	100
2.21 CDF analysis of single amino acids excluding CpG deaminations and some ApT transitions	102
2.22 Comparison of CDFs with the geometric mean of derived single amino acids . .	103
2.23 Significant correlations between derived single amino acid DMFs and their chemical properties	106
2.24 The physico-chemical property eigen values associated with each CDF cluster for single amino acids	108
2.25 Scaling of derived motif occurrences with increasing motif length	109
2.26 Derived dipeptides without CpG deaminations and some ApT transitions	111
2.27 Derived dipeptides as a product of their single amino acid submotifs	113
2.28 Asymmetry within derived dipeptides	115
2.29 Derived tripeptides without CpG deaminations and some ApT transitions . . .	116
2.30 Differences between derived tripeptides and their dipeptide submotifs	119

Introduction

Evolution is a result of spontaneous mutations generating variation across an organism's genome, followed by either selection purging some variants, or the lack thereof allowing other variants to persist and neutrally drift to higher allele frequencies in a population. Understanding evolution has very broad repercussions, from understanding how the diversity of species came about, to understanding how diseases arise, developing novel therapeutic drugs, and predicting the effects of various genetic mutations. A current clear example is the interplay of COVID-19 mutants and predicting the effect of these on our immunity post natural infection or vaccination. Interrogation of evolution at a molecular level has generally been done by looking at genomic sequences. Investigating genomes offers a single snap shot in time of the complicated total summation of mutational biases and selective pressures over evolutionary time. Because of this, it is a great challenge to separate and understand the mutational patterns and selection, each of which also do not remain constant over time. There is a long history of studies attempting to disentangle these evolutionary forces: some have compared the conservation of sequence elements between genomes [Adzhubei et al., 2010, Vaser et al., 2016, Rošić et al., 2018, McDonald et al., 2011, Matassi et al., 1999, Francioli et al., 2015], some have interrogated the composition of sequences within each genome [Rahbari et al., 2016, Brendel et al., 1992, Burge et al., 1992, Fay et al., 2001], some have investigated the occurrence of novel variants using statistical models of variants co-occurring with various genomic features [Chen et al., 2017, Consortium et al., 2010, Aggarwala and Voight, 2016, Cano et al., 2021] or the frequency distributions of variants [Harpak et al., 2016, Drake et al., 2006], and many more approaches. This is further complicated by evidence that selection also acts upon mutation rate itself, changing the rate of evolution of distinct regions of the genome [Lynch, 2010, Lynch et al., 2016]. This question of

how does one separate out the evolutionary forces affecting mutation rate and selection has no definitive single answer, and therefore many researchers utilise novel and innovative approaches to do so. In this thesis I build on this prior knowledge and describe in two Chapters projects where I try to separate both mutational and selective variations to elucidate novel evolutionary patterns.

The generation of variation is biased at many scales across the genome. At centiMorgan scales, mutation rates vary in the order of 10-fold when comparing assumed neutrally evolving synonymous sites [Hodgkinson and Eyre-Walker, 2011, Matassi et al., 1999]. At finer scales, this can be even more exaggerated, with relative mutation rates varying more than 400-fold when comparing mutation rates across different heptamer (7-nt) sequences in human [Carlson et al., 2017]. At individual base pair resolution, DNA modifications are often mutational liabilities [Tomkova and Schuster-Böckler, 2018], whilst being functionally important epigenetic markers [Blow et al., 2016, Akitake et al., 2011, Regev et al., 1998, Bewick et al., 2017, Zhang et al., 2018, Bewick and Schmitz, 2017]. In particular, the most common DNA modification, 5-methylcytosine (5mC), is known to mutate 10-fold faster than its unmethylated counterpart [Duncan and Miller, 1980, Bulmer, 1986, Sved and Bird, 1990]. The context of a 5mC can greatly affect its mutation rate, such as the local GC-content [Fryxell and Moon, 2005], the region's propensity to be single-stranded [Beletskii and Bhagwat, 1996, Ehrlich et al., 1986], and accessibility to DNA repair [Amaral et al., 2017, Supek and Lehner, 2015]. But might not the reverse also be true - that 5mC itself can affect the evolvability of the surrounding DNA? 5mC has been shown to affect the electronic [Sharonov et al., 2003, Banyasz et al., 2016] and mechanical [Murchie and Lilley, 1989, Severin et al., 2011, Ngo et al., 2016] properties of the DNA helix. In addition, the resulting base lesions of methylated and unmethylated cytosine deamination are thought to recruit different DNA repair pathways [Chen et al., 2014], which often use error-prone polymerases to resynthesize the DNA surrounding base pair lesions [Peña-Díaz et al., 2012, McDonald et al., 2011]. The question whether 5mC casts a mutational shadow, impacting the generation of variation in the neighbouring nucleotides surrounding the 5mC itself, will be the focus of Chapter 1.

Similarly to this mutational heterogeneity, there is great selective heterogeneity across the

genome with important genes and regulatory features being conserved, whereas many non-coding sites evolve at faster rates with fewer variants being deleterious and therefore purged from the population. However selection is rarely so clear, with many novel variants falling on the neutral spectrum, forming the basis for the neutral, or nearly neutral, theory of evolution [Kimura, 1983, Ohta, 1992]. An important question in population genomics is how to predict the function of variants: which ones are beneficial, and which ones are deleterious and might lead to disease. Generally this has been addressed using homology-based methods, using comparative genomics and the growing availability of population genomic data [Lindblad-Toh et al., 2011, Ng and Henikoff, 2001, Vaser et al., 2016, Adzhubei et al., 2010]. Unfortunately, these studies often only describe what patterns remain conserved, and are less sensitive to the appearance of new functional patterns [Flanagan et al., 2010]. Other methods comparing the composition within genomes have had success at identifying prevalent avoidance of nucleotide sequences, such as avoidance of transcription factor binding sites [Hahn et al., 2003] and RNA binding sites [Farh et al., 2005, Marco, 2015, Umu et al., 2016]. However, the composition of a genome is the result of the interactions of selection with all the mutational heterogeneity present in genomes. Therefore, only highly penetrant effects that exceed the large combination of mutational bias that exists in the genome can be called with confidence. How does one investigate selection acting on the appearance of new, potentially functional, sequences whilst simultaneously controlling for factors affecting the genomic composition? This question will be the focus of Chapter 2.

In this thesis I scrutinize the generation and retention of sequence patterns in human genomes at the finest scale - at a resolution of 1-10 base pairs. Specifically, I investigate the mutational pattern associated with cytosine methylation not at the base itself, but on the surrounding bases. I compare this pattern in human to that in plants to gain insight into the mechanistic origins of the effect and the differences in the evolutionary regime across kingdoms. Next, I interrogate the retention of amino acid and nucleotide sequences after they arose by mutation, and ask which sequence motifs are prevented from rising to high frequency in the human population. I approach the problem in a novel way, attempting to account for varying mutation rate in the ancestral genome. I focus on single amino acids, dipeptides, and tripeptides in coding

contexts of the human exome, and elucidate patterns relating to the amino acids' chemical properties and asymmetry in their appearance in short motifs.

Chapter 1

Does cytosine methylation affect mutation dynamics beyond the base itself?

1.1 Introduction

For evolution to occur there must be heritable variation between the individuals in a population. One fundamental type of variation occurs at the nucleotide level where bases change as a result of mutations, and these mutations must occur in the germline of the organism in order to be heritable. The null hypothesis might be that mutation rates are consistent across the genome, occurring uniformly and randomly. However, many studies have demonstrated that there is much heterogeneity in mutation rate in both germline and somatic tissues [Chen et al., 2017, Hodgkinson and Eyre-Walker, 2011, Carlson et al., 2017, Tomkova and Schuster-Böckler, 2018, Chen et al., 2014, Vogel and Kopun, 1977, Ségurel et al., 2014, Cooper and Krawczak, 1989]. For example, relative mutation rates can vary more than 400-fold in human germline, when looking at rare SNP incidence in 7-mer sequence motifs [Carlson et al., 2017].

What is the origin of such large heterogeneity? Mutation rates in both germline and soma

correlate with a variety of genomic factors, leading to variability at a wide range of scales. The strongest correlations at a larger genomic scale appear to be DNA replication timing [Stamatoyannopoulos et al., 2009] and error-prone repair processes [Supek and Lehner, 2015, Poulos et al., 2017, Feng and Michaels, 2015, Kramer et al., 1984, Allshire and Madhani, 2018]. These processes can be intertwined, with the example that mismatch repair (MMR) can be coupled to DNA replication timing [Supek and Lehner, 2015]. This can lead to certain patterns in gene localisation, such as active genes that perform essential functions are euchromatic and located in early replicating regions where MMR activity is highest. Recombination rates also vary at the megabase (Mb) and kilobase (kb) scale [Myers et al., 2005], and the associated GC-biased gene conversion can influence the fixation of deleterious SNPs [Duret and Galtier, 2009, Galtier et al., 2009]. Chromatin structure also has a strong effect with nucleosomes shown to affect the occurrence of mutations positively and negatively [Chen et al., 2012], as well as exhibiting a 10 bp periodicity of mutations due to DNA minor groove orientation around the nucleosome [Pich et al., 2018]. The orientation of the modified base with respect to the nucleosome can affect mutation rates, with bases facing away from the surface exhibiting a 42-fold higher deamination rate than bases positioned on the histone core [Song et al., 2011, Cannistraro et al., 2015]. Larger scale chromatin structure is also important, where condensed heterochromatin regions seem to accumulate more mutations, with the reason thought to be that such regions are less accessible to DNA repair machinery [Makova and Hardison, 2015, Schuster-Böckler and Lehner, 2012, Gonzalez-Perez et al., 2019]. This creates a dynamic balance between local nucleosomes and potentially other proteins protecting the local sequence from mutations, but also preventing the long term efficacy of repair machinery from correcting any mutations that still occur. Such a balance is also seen in transcribed regions where the single-strandedness of the transcription process may amplify mutational effects (e.g. deamination risk), whilst also being subject to improved repair due to transcription-coupled repair processes [Kim et al., 2007, Aguilera, 2002, Mellon et al., 1987, 1986, Balajee and Bohr, 2000]. While most of these correlations are generally consistent between germline and soma, germline mutation rates correlate strongly with expression levels, whereas the inverse is true for somatic mutation rates, which decrease with expression levels [Chen et al., 2017]. However, in this study, I

will focus on another mutational bias that exist in the genome at a much more fine-grained, base-pair resolution: DNA modifications.

DNA modifications are crucial for genomic regulation, creating variation and control beyond the genomic sequence itself. The most common DNA modification in many eukaryotes is cytosine methylation (5mC) and this is used for genomic regulation in a large variety of organisms, ranging from single-celled bacteria and archaea, to diverse invertebrate, vertebrates, plants, and fungi [Blow et al., 2016, Akitake et al., 2011, Regev et al., 1998, Bewick et al., 2017, Zhang et al., 2018, Bewick and Schmitz, 2017]. In general, it is firstly deposited by *de novo* methyltransferases [Hata et al., 2002, Okano et al., 1999, Bouyer et al., 2017], and then perpetuated by maintenance methyltransferases that target hemi-methylated DNA strands following DNA replication [Goll and Bestor, 2005]. In mammals, these methyltransferases usually target the palindromic sequence CpG [Bird, 2002]. In plants, the CpHpG and asymmetric CpHpH contexts are also methylated [Zhang et al., 2018]. Similarly, 5mC can be removed by targeted active processes, where the methylated base is enzymatically excised by glycosylases and replaced by an unmethylated cytosine [Zhang et al., 2018, Wu and Zhang, 2017, Zhu, 2009], or via non-specific dilution post replication when maintenance methylation is disabled [Piccolo and Fisher, 2014].

In bacteria, methylation is used for protecting their DNA from their own set of restriction enzymes whilst allowing them to digest foreign unmethylated DNA [Meselson et al., 1972]. In eukaryotes that use cytosine methylation, 5mC is often associated with repressive chromatin states, particularly around repetitive and transposable elements. Due to this general pattern, it is thought that DNA methylation acts to protect the organism's genome from retroviral elements and contributes to genomic stability [Piccolo and Fisher, 2014]. DNA methylation, however, has evolved other regulatory functions in a species-specific manner, particularly in regards to gene expression. In the phylum of arthropods, honey bees differentially methylate genes in their brain, which leads both to irreversible specialisation to the queen and worker castes, and reversible behavioural switching between subcastes [Lyko et al., 2010]. In plants, DNA methylation at CpGs is most prominently used to suppress transposon expression [Ji et al., 2019, Zakrzewski et al., 2017], but is also found in the bodies of broadly expressed genes

[Zemach et al., 2010, Takuno and Gaut, 2011, 2013] and represses expression of other genes. One such example is the establishment of flower symmetry in *Linaria vulgaris*, where methylation of the single gene *Lcyc* reversibly silences it and leads to a change of flower symmetry from bilateral to radial [Cubas et al., 1999]. In mammals, DNA methylation has a similar function to methylation in plants in regards to silencing transposable elements [Morris et al., 2004, Bestor and Bourc'his, 2004, Zamudio et al., 2015a] and repressing gene expression [Weber et al., 2005, Zamudio et al., 2015a], but is more ubiquitous [Suzuki and Bird, 2008, Feng et al., 2010]. Specifically, in human at least 80% of CpGs are methylated [Lister et al., 2009], in contrast to only 22% of CpGs in *A. thaliana* [Feng et al., 2010]. One important usage of such ubiquitous CpG methylation is for X-chromosome inactivation in human females to control the dosage of X-localised genes [Riggs and Pfeifer, 1992].

1.1.1 DNA modifications as mutagenic agents

In spite of its widespread role in genome regulation, DNA modifications can also be mutagenic liabilities. 5mC is the earliest discovered and best studied example of this. It has been known since the 1980s that 5mCs at CpG dinucleotides tend to mutate more frequently than unmethylated cytosines. This is explained two-fold: Firstly, 5mC is liable to spontaneously deaminate around 2 to 3 times faster than unmethylated cytosine [Wang et al., 1982, Ehrlich et al., 1986]. Secondly, deamination of 5mC leads to a thymine rather than a uracil [Barker et al., 1984, Coulondre et al., 1978, Cooper and Krawczak, 1989], resulting in less efficient repair. Uracil gets recognised by high-fidelity uracil-deglycosylases, leading to efficient base-excision repair (BER) [Savva et al., 1995]. On the other hand, the T:G mismatch will only be recognised by the mismatch repair (MMR) pathway which, despite in mammals preferentially repairing the thymine to a cytosine, still has only 90% efficiency at hemimethylated sites [Kramer et al., 1984, Jones et al., 1987]. The ensuing error in mismatch repair would cause a guanine to adenine substitution on the opposite strand. This has been demonstrated extensively in experimental setups. In vitro studies showed that 5mC deaminates at significantly higher rates than regular cytosines using single-stranded DNA and bacterial systems [Wang et al., 1982, Ehrlich

et al., 1986]. Many studies also showed the same effect in vivo for *E.coli* [Coulondre et al., 1978, Kramer et al., 1984], humans [Barker et al., 1984, Cooper and Krawczak, 1989], and *Neurospora crassa* [Selker and Stevens, 1985]. Studies also saw this effect manifest as genomic depletions of CpGs as well as corresponding enrichments in CpA and TpG, which are the result of CpG transition mutations [Selker and Stevens, 1985, Mazin and Vanyushin, 1987, Cooper and Krawczak, 1989]. This pattern was again seen when looking at human clinical variant data, showing most disease causing variants were CpG transitions, occurring at a much higher rate than expected [Vogel and Kopun, 1977, Cooper and Youssoufian, 1988]. Such CpG depletions were also seen genome-wide in *A. thaliana* in the 1001 genomes project [Weigel and Mott, 2009]. More recently, studies have utilised the ever growing sample sizes of parent-child trios [Jónsson et al., 2017] and population SNP datasets to see an increased frequency of 5mC deamination polymorphisms [Ségurel et al., 2014, Francioli et al., 2015, Siva, 2008].

1.1.2 Effect of context on CpG mutability

The extent to which CpG mutation rate increases upon methylation is, to some degree, context-dependent. Fryxell and Moon in 2005 showed a strong inverse correlation between cytosine mutability and GC-content [Fryxell and Moon, 2005]. Cytosine deamination has been shown to occur 140 times more frequently in single-stranded DNA as opposed to double-stranded DNA [Beletskii and Bhagwat, 1996, Ehrlich et al., 1986], and so Fryxell and Moon connected their results to the observed reduced rate of DNA melting at G-C base pairs [Leroy et al., 1988]. There has also been much research into other context-related effects on the mutability at the cytosine itself. A strand-specific asymmetry was seen regarding mutations at 5mC sites during replication with increased misincorporation of A opposite 5mC on the leading strand [Tomkova et al., 2018]. UV radiation is also thought to increase mutation of 5mC more than unmethylated cytosine. 5mC has a *circa* 5-fold higher molar absorption coefficient at 290nm wavelength radiation [Sharonov et al., 2003, Rochette et al., 2009], leading to more frequent generation of cyclobutane pyrimidine dimers (CPDs). These CPDs have a deamination half-life of hours to days, compared to thousands of years for unmethylated cytosine and 5mC.

Sequence context seems to have a large effect on this rate, with the TCG sequence exhibiting an approximately 50 times faster rate than CCG [Cannistraro and Taylor, 2009]. Smoking increases mutation rates of 5mC more than unmethylated cytosine. The large carcinogenic DNA adduct BPDE (Benzo(a)pyrene diolepoxide) from smoking preferentially binds guanine in a CpG context, and even more frequently if the cytosine is methylated [Denissenko et al., 1997]. In contrast, APOBEC enzymes seem to exhibit much greater efficiency with deamination of unmethylated cytosine compared to methylated cytosines, ranging in the orders of 10-100 fold depending on the precise APOBEC enzyme involved [Nabel et al., 2012, Fu et al., 2015, Kamba et al., 2015]. Chromatin state also affects the mutability of 5mC: There is differential preference for methylated and unmethylated cytosines to occur in the different chromatin contexts. These different contexts greatly affects the accessibility of methylation-induced lesions to BER and MMR, which are the main pathways by which these lesions are repaired [Supek and Lehner, 2015].

1.1.3 Effect of 5mC on context mutability

Context affects mutability at methylated sites; however, how methylated sites affect the mutability of their context is much less understood. There is a known 'next-neighbour' effect of methylation whereby 5mC has electronic and conformational effects on its neighbours. One outcome of this is that 5mC correlates to greater susceptibility to UV-induced formation of dipyrimidines [Banyasz et al., 2016]. As mentioned above, this is due to 5mC having a greater absorption coefficient to 290nm wavelength radiation [Sharonov et al., 2003]. It is not unreasonable to consider that addition of a chemical moiety to DNA could affect the mutability of the bases further in proximity. 5mC is known to alter the mechanical properties of the DNA helix itself. It affects the dynamics of strand-separation of the DNA helix, making it more rigid and less likely to melt. This was shown in experiments looking at cruciform extrusion of bacterial plasmids [Murchie and Lilley, 1989], PAGE electrophoresis [Shimooka et al., 2013], single molecule force experiments [Severin et al., 2011], and single-molecule cyclization assays [Ngo et al., 2016]. There could also be differential recruitment of trans factors which could affect

mutability, or the T:G mismatches that arise at mCpGs could be repaired in a manner that affects neighbouring bases. Qu et al. investigated this by overlapping SNP calls in human and medaka genomes with methylated and unmethylated regions derived from bisulphite sequencing experiments [Qu et al., 2012]. They found a 1.5 fold increase in SNP incidence 10bp around methylated sites compared to unmethylated sites. They hypothesized that after spontaneous deamination of 5mC, an alternative mismatch repair pathway removes more than just the mismatch leading to error prone resynthesis of the erased gaps, as opposed to base excision repair which is usually error free [Robertson et al., 2009, Hegde et al., 2008]. A similar mechanism was hypothesized by Anthony Furano’s group, who found that the increasing density of CpGs in younger L1 transposons correlated with increased non-CpG mutation rates in the same young L1 transposons [Walser et al., 2008, Walser and Furano, 2010]. They investigated this further by deliberately introducing mismatches into a SV40-episome in multiple human cell lines [Chen et al., 2014]. They saw a statistically significant increase in mutations at neighbouring bases surrounding these deliberate mismatches and presented evidence that both mismatch repair and APOBEC were required.

These observations point to the phenomenon that methylation affects the mutability of neighbouring bases, but, as highlighted above, many factors can affect local and regional variation in mutation rate. To implicate methylation as the causative force of mutation requires careful control of context. Methylation has specific functionality and is distributed non-randomly across the genome, whether it is mainly at condensed heterochromatin in mammals [Rabinowicz et al., 2003], or near transposable elements in plants [Kim and Zilberman, 2014, Ikeda and Nishimura, 2015]. Therefore, depending on how these sites are sampled, the effect can be significantly confounded by regional effects and/or selection [Zhao and Jiang, 2007, Schuster-Böckler and Lehner, 2012, Sung et al., 2015].

Here, I use data on rare polymorphisms in human, *A. thaliana*, and rice to quantify the mutational effect methylation has on adjacent bases. Controlling for sequence, chromatin state, and well-mapped genome regions, I find, in contrast to Qu et al [Qu et al., 2012], that methylation has a protective mutational effect in human, reducing the incidence of SNPs by roughly 15% in the adjacent 1 - 2 bases surrounding 5mC sites. I track this discrepancy to their study to

the fact that the authors included higher frequency SNPs from the HapMap dataset, where common mCpG mutations were being mischaracterised as unmethylated CpGs, confounding the trend. However, I also find that methylation increases mutability locally to the 5mC in both *A. thaliana* and rice. Integrating data on RNA-dependent DNA methylation (RdDM) and active demethylation mutants in *A. thaliana* [Tang et al., 2016], I show that demethylation and RdDM activity dose-dependently reduce the mutational risk of methylation on neighbouring nucleotides. In addition, sites requiring the chromatin remodeler DDM1 for methylation show a higher mutational risk, implying condensed chromatin blocks access to some form of DNA repair. My work highlights that methylation has a broader mutational effect than at the base itself, but it is very dependent on context and species-specific modes of lesion formation and further repair.

1.2 Results and Discussions

1.2.1 Methylation is associated with decreased mutability of neighbouring bases in human

To explore how methylation affects mutation rates proximal to CpGs, I first considered the human system. To approximate mutation rate, I used rare SNP data, which I define as singletons (SNPs that occur at allele count 1), from the gnomAD dataset [Karczewski and Francioli, 2017] and overlapped this with methylation sites. Singleton SNPs constitute a better proxy for germline mutational processes than common SNPs or substitutions, which more strongly reflect longer-term selective processes and gene conversion events [Rahbari et al., 2016]. In addition to rare SNPs, to infer mutation rates from population genetics data at methylated and unmethylated CpGs ideally germline methylation data is required. Unfortunately, no high-quality germline methylation data from human was available at the time the study commenced. However, human H1 embryonic stem cells (ESCs) have previously been shown to be a robust proxy for germline methylation and hydroxymethylation state when inferring mutational effects, elucidated by comparing patterns of SNP incidence in H1 ESCs with differentiated neuronal

cells and cells in early spermatogenesis [Supek et al., 2014]. CpG basecalls for H1 ESCs were obtained from bisulphite sequencing experiments by Lister et al., [Lister et al., 2009]. Only CpG sites with more than 10 reads in the sequencing experiments were used. I defined methylated and unmethylated sites as those with $> 70\%$ and $< 20\%$ methylated reads respectively (see Methods). To control for local sequence context and GC content, methylated and unmethylated sites were paired to create two sets of sites where the distribution of sequence contexts were identical, with the only difference being the methylation status. This is because both sequence context and GC content have been shown to correlate with either CpG mutability directly [Zhao and Jiang, 2007, Fryxell and Moon, 2005] or with mutation rates and SNP calling accuracy generally [Francioli et al., 2015, Carlson et al., 2017]. Each methylated and unmethylated site was paired by matching two main criteria: Firstly, the adjacent 4 bases to the focal CpG had to match. I did not extend the sequence context to longer motifs because a heptanucleotide context was previously shown to account for more than 80% of variability in mutation rate variation across the human genome [Aggarwala and Voight, 2016], and extending the constraint to longer sequences drastically reduced sample size. Furthermore, motifs that contained extra CpGs other than the focal CpG were excluded to prevent other nearby CpGs with known or unknown methylation status confounding the results. Secondly, I required the pairs to have the same chromatin state (Figure 1.1), as defined by a widely used hidden Markov model (HMM) that combines information from ChIP-seq experiments containing histone marks, methylation, and DNA accessibility [Ho et al., 2014]. Mutation rates are known to vary greatly with chromatin context [Schuster-Böckler and Lehner, 2012]. Specifically, heterochromatic regions seem to accumulate more mutations than more open DNA regions that are more accessible to DNA repair machinery [Supek and Lehner, 2015, Fortuny and Polo, 2018]. Chromatin context can also capture other effectors of DNA mutability, including replication timing and transcriptional activity [Stamatoyannopoulos et al., 2009, Poulos et al., 2017, Beletskii and Bhagwat, 1996, Aguilera, 2002]. If multiple such pairs exist I paired the ones closest in genomic distance to each other.

Doing so, I obtained 60,589 pairs of matched sites. I then calculated the relative mutational risk associated with methylation, RR_{met} as the ratio of SNP incidence near methylated CpGs

compared to unmethylated CpGs:

$$RR_{met} = \frac{\frac{SNP_{met}}{N_{met}}}{\frac{SNP_{unmet}}{N_{unmet}}}$$

where SNP_{met} and SNP_{unmet} are the number of SNPs observed near, defined as 1-3bp from, methylated and unmethylated pairs respectively, and N_{met} and N_{unmet} are the number of sites, which conveniently are equal due to being paired by context, but are still important for assessing the error estimate. Therefore $RR_{met} > 1$ implies higher relative mutability of the neighbourhood near a methylated CpG and $RR_{met} < 1$ being a reduction in neighbourhood mutability. I calculated this using derived singleton SNPs across 15,708 whole genomes from the gnomAD database [Karczewski and Francioli, 2017], where a derived SNP is one where the allele is different compared to the ancestral sequence [Siva, 2008]. Any SNP that was a transition at a TpG or CpA was excluded to avoid confounding for potentially mischaracterised CpG deaminations.

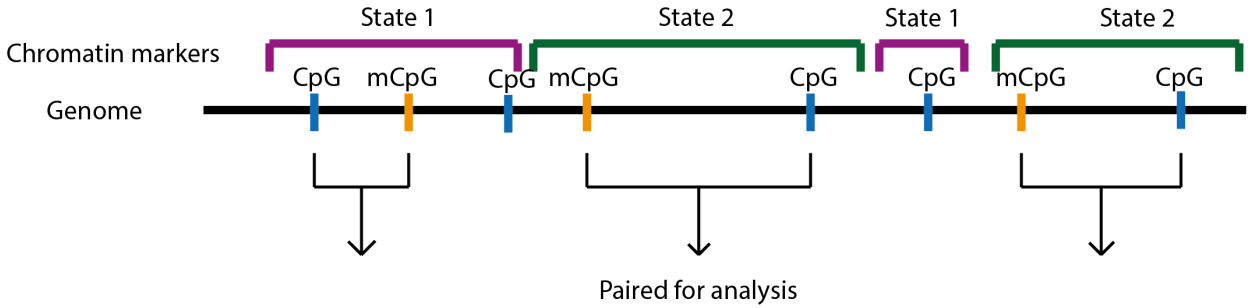


Figure 1.1: A schematic illustrating the CpG pairing method described. Unmethylated and methylated CpG sites in the genome were paired by identical up and downstream 4bp, and by identical chromatin state. If there are multiple partners, the nearest CpGs were selected.

Following this approach, I found a reduction in the incidence of SNPs adjacent to methylated CpGs compared to unmethylated CpGs (Figure 1.2). The average RR_{met} was 0.886 across all 3 up and downstream bases from the focal CpG and mutation types (transitions and transversions) ($p = 2.73 \times 10^{-21}$ using the Z-test for proportions). This means there were 886 SNPs in the ± 1 -3bp neighbourhood of methylated CpGs for every 1000 SNPs in the neighbourhood of unmethylated CpGs. One important consideration is whether common SNPs at the CpG itself

could be a source of apparent reduced mutation frequency, due to scenarios where two or more SNPs in a particular sequenced read prevents it from being aligned or passing quality control. This would lead to reads with more mutations being discarded, and leading to a reduction in mutational incidence adjacent to highly mutable loci, such as methylated CpGs. To test this explicitly, I investigated whether limiting my analysis to methylated and unmethylated sites with a mutation at the focal CpG affects this pattern. Upon investigation, most of the paired sites used in the analysis contained a SNP of any allele frequency at either the C or the G in the focal CpG (N = 62,312 of methylated sites, N = 45,612 of unmethylated sites). Limiting the analysis to these loci, I again see the same reduction of mutational incidence adjacently to the focal CpG when looking at both SNPs of higher allele frequencies, or singleton SNPs. When I select for only common SNPs occurring at the focal CpG (i.e. SNPs with allele frequency greater than 5%) to more likely select for SNPs coincident in a specific read since the phase of the SNPs cannot be extracted from gnomAD, the sample drops dramatically and instead enriches for unmethylated sites (N = 1441 for methylated sites, N = 16464 for unmethylated sites). This effect is discussed further in section 1.2.7, but is likely due to methylated sites that have already mutated to a thymine in the sample, but is still described as a cytosine in the reference, causing bisulphite sequencing to incorrectly designate the base as an unmethylated cytosine where in fact it is a methylated cytosine that has already mutated. Despite this, even when selecting for sites with common SNPs at the CpG, the pattern remains qualitatively the same.

1.2.2 Methylation increases incidence of SNPs in neighbourhood of 5mC in plants

To test the intrinsic or species-specific components of this apparent protective effect in the neighbourhood of CpG methylation, I applied this same method to two plants: *A. thaliana* and rice (*Oryza sativa*). I had also considered investigating the effect in mouse and *Apis mellifera* (honey bee) since both have CpG methylation and SNP data available. However, these latter two were lacking in sequencing depth and volume, and - for mouse - appropriate polymorphism

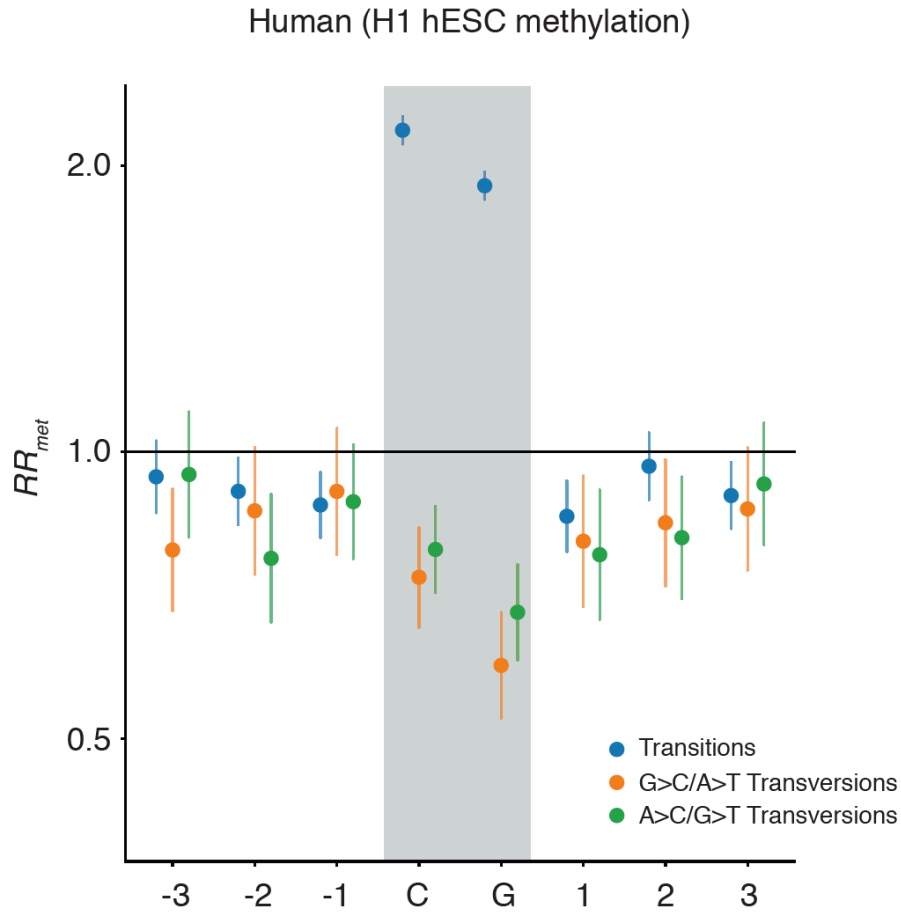


Figure 1.2: The relative mutational risk of methylation (RR_{met}) at CpGs and neighbouring nucleotides in human H1 ESCs. Vertical bars are confidence intervals on RR_{met} computed using the delta method.

data, and so I kept further analysis to *A. thaliana* and rice. The methylomes used were from *A. thaliana* Ws0 global seed stage [Lin et al., 2017], and *O. sativa* 3 week old leaf tissue [Stroud et al., 2013]. As discussed previously, ideally I would use the germline methylation to infer the mutational effect of methylation via population SNPs. However, plants do not have a segregated germline, and based on past literature the CpG methylome does not change much during embryo to adult tissue development in plants [Zhang et al., 2018, Bouyer et al., 2017]. Chromatin state information was again obtained from HMMs of ChIP-seq data in each species [Liu et al., 2017], and the singleton SNPs were obtained from the 1001 Arabidopsis genomes and 3000 Rice genomes projects respectively [Weigel and Mott, 2009, Li et al., 2014]. Further to this, to avoid biases arising from mischaracterised bases that occur in repetitive regions, I also applied RepeatMasker [Smit et al., 2015] (see Methods) to the methylation and SNP calls in the plants, as well as to human for completeness.

In contrast to what we saw in human, there was an increase in mutability near methylated CpGs in both *A. thaliana* and rice. I found an overall $RR_{met} \pm 1\text{-}3\text{bp}$ of 1.28 in *A. thaliana* ($p = 2.06 \times 10^{-89}$) and 1.31 in *O. sativa* ($p = 5.43 \times 10^{-15}$) (Figure 1.3A). The estimates in rice were noisier due to relatively smaller number of matched pairs, with 121,774 pairs in *A. thaliana* and 42,779 pairs in *O. sativa*. The highest RR_{met} was associated with C to T transitions ($RR_{met} = 1.38$ in *A. thaliana*; $RR_{met} = 1.71$ in *O. sativa*). RR_{met} decreases as a function of distance from the focal CpG. I repeated the human analysis applying RepeatMasker, despite the human reference genome and gnomAD already being repeat-masked, and found the same trend as before for human H1 ESCs (Figure 1.3A).

As 5mC is not limited to CpGs in plants, I repeated the same analyses for the CpHpG and CpHpH contexts (where H = A, C, or T) independently. In both cases, I found the same pattern where SNP incidence was significantly increased directly adjacent to the 5mC ($RR_{met} > 1$ Figure 1.4). Methylation is much rarer in these contexts however, occurring at CpHpG and CpHpH contexts 7% and 1.6% of the time compared to 24% at CpGs in *A. thaliana* [Lin et al., 2017, Cokus et al., 2008, Feng et al., 2010]; and similarly 21% and 2.2% compared to 59% in *O. sativa* [Feng et al., 2010]. In addition, methylation in these contexts was greatly skewed to certain functional contexts, leading to much smaller sample sizes of paired sites for controlled

analysis. I therefore focus on CpG sites for greater statistical power and sample size for the remainder of this chapter. Nonetheless, the fact that the behaviour of all three contexts is so similar implies that the effect is likely not specific to molecular processes unique to the CpG context, and more an intrinsic result of cytosine methylation itself.

Since previously I had only looked at SNPs, it raises the question whether insertions and deletions (indels) are similarly affected. Because certain DNA repair mechanisms are known to affect indel rates, investigating such patterns could offer insight into what mechanism is behind the respective increase and decrease in mutability surrounding methylated sites in plants and human respectively. Similar to SNPs, I focused on singleton indels and further confined the analysis to single-base indels so that the sequence context between the pairs can be compared. A larger insertion or deletion could drastically change the surrounding sequence, and therefore not be representative of the paired site. Indels are rarer than SNPs, and so I calculated only two RR_{met} estimates: for the CpG itself, and for all neighbouring bases up and downstream ($\pm 1-3\text{bp}$) combined. In humans, I find a reduced incidence of indels for the methylated sites both at the CpG itself and surrounding it (Figure 1.3B). In *A. thaliana*, RR_{met} is higher than for human, with $RR_{met} > 1$ for deletions affecting CpG-proximal sites. There was not enough indel information to carry out informative analysis in rice.

1.2.3 GC-content, chromatin state, and sequence context do not drive the mutational risk pattern of methylation

As mentioned in the introduction, sequence, GC-content, and chromatin state can drastically affect the mutation rate at individual bases [Fryxell and Moon, 2005, Makova and Hardison, 2015]. Therefore, to test whether the change in pattern seen between plants and humans is due to differential sampling of these regions, I stratified the matched pairs and their RR_{met} by sequence context, local/regional GC content, and chromatin state. Firstly, the diversity of sequences captured is similar between the 3 species, with a large number of unique sequence contexts contributing to the global RR_{met} ($>15,000$ for each species, Figure 1.5A,E,I). More explicitly, when dividing the sequence context of the pairs into sets grouped by the central 4

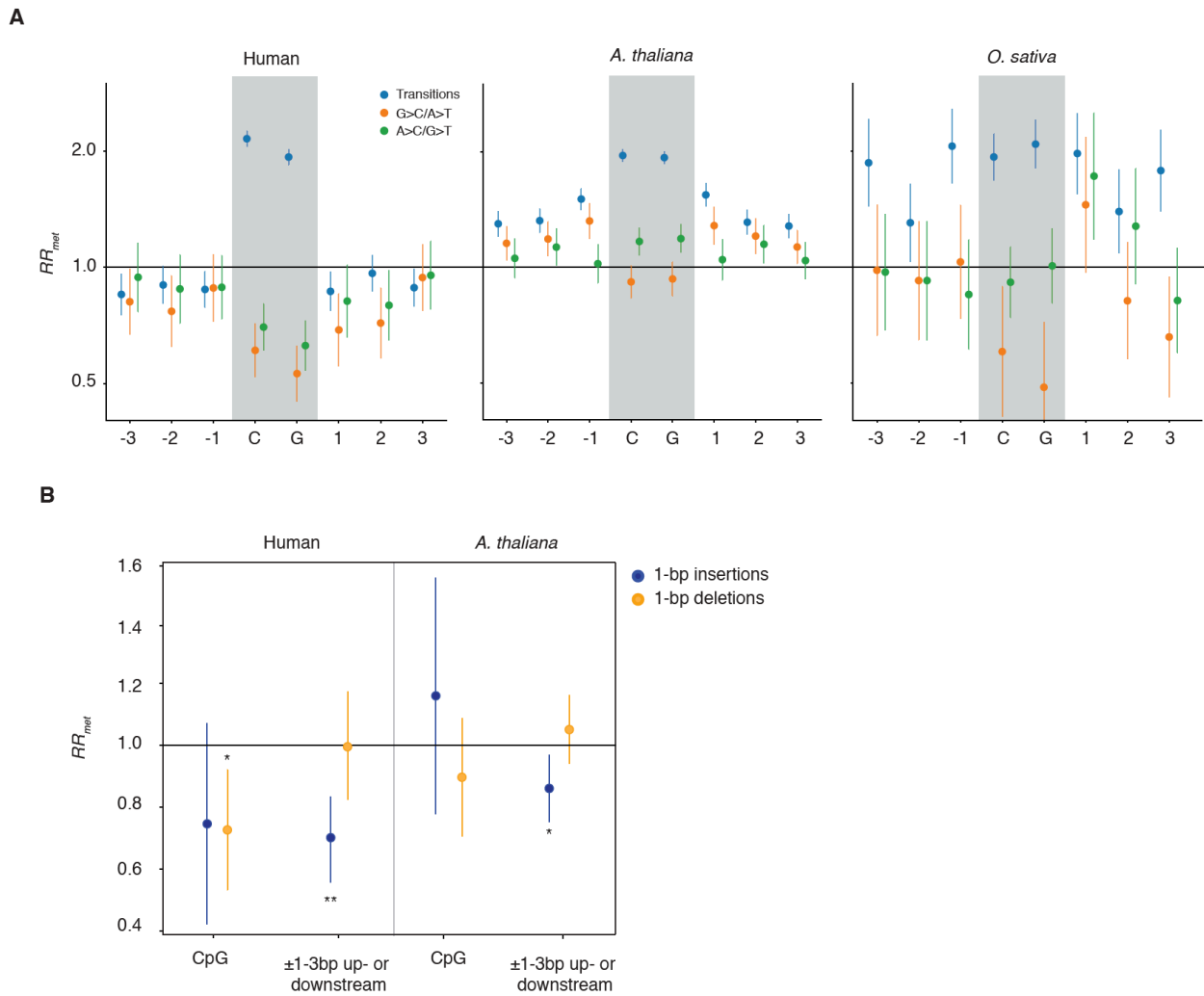


Figure 1.3: **(A)** The relative mutational risk of methylation (RR_{met}) at CpGs and neighbouring nucleotides in human, *A.thaliana*, and rice. Only SNPs and CpG sites in repeat-masked sequences were included for all species. Vertical bars are confidence intervals on RR_{met} computed using the delta method. **(B)** RR_{met} as applied to indels in human and *A.thaliana*. * $p < 0.05$, ** $p < 0.005$.

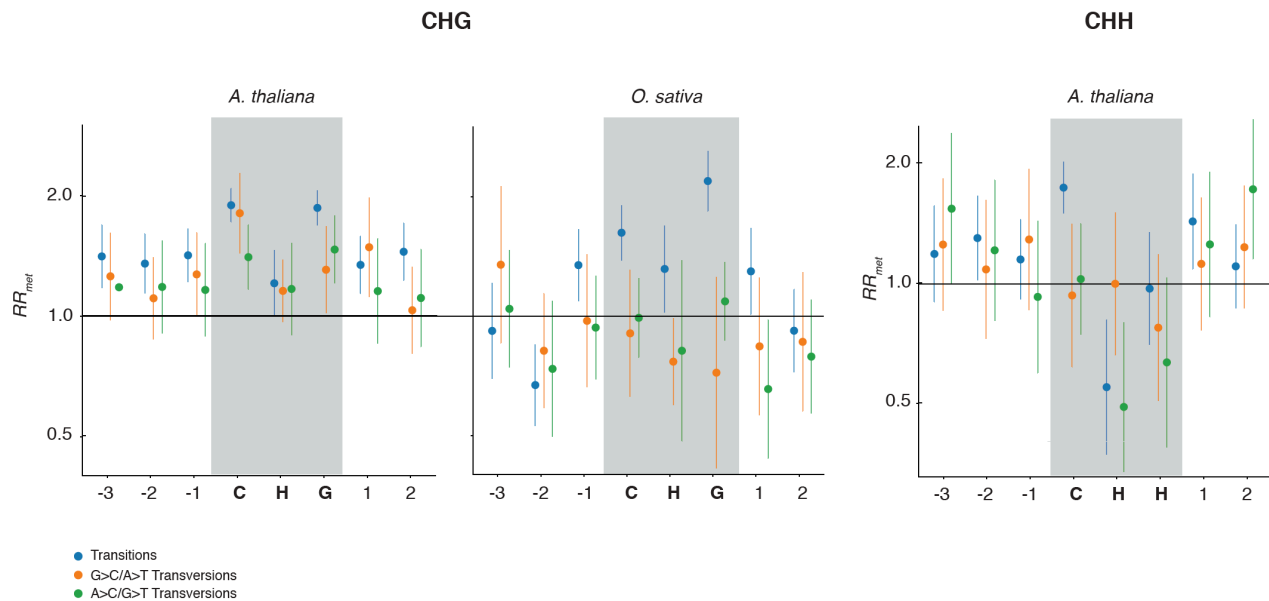


Figure 1.4: The relative mutational risk of methylation (RR_{met}) at CpHpG and CpHpH sites and neighbouring nucleotides in *A. thaliana* and rice. The significance of the overall RR_{met} estimates are: CpHpG *A. thaliana* $p = 3.8 \times 10^{-19}$; CpHpH *A. thaliana* $p = 1.8 \times 10^{-8}$; CpHpG *O. sativa* $p = 0.043$. Vertical bars are confidence intervals on RR_{met} computed using the delta method.

nucleotides of the motif, the pattern of lower mutability around methylated CpGs in human and higher mutability in *A. thaliana* remains (Figure 1.5B,F). For rice, the sample size was too small to stratify into sequence contexts, leading to a noisy pattern (Figure 1.5J). This suggests that global RR_{met} estimates are not driven by a narrow subset of sequence contexts, despite there being variability in the RR_{met} for individual sequence contexts. To test this explicitly, I randomly subsampled the matched pairs from both human and *A. thaliana* to create two sets with identical sets of sequence compositions. More clearly as an example: if the sequence ‘TATACGTATA’ occurred 100 times in human but 400 times in *A. thaliana*, I would subsample the sites in *A. thaliana* to select only 100 of them. Doing so, RR_{met} remains > 1 in *A. thaliana* and < 1 in human (Figure 1.6).

I next considered whether stratifying by GC-content, both locally and regionally, would affect RR_{met} . GC content was previously implicated in modulating CpG deamination rates since DNA duplexes with lower GC content were more likely to spontaneously melt, and single-stranded DNA has a much higher deamination rate than double-stranded DNA [Leroy et al., 1988, Fryxell and Moon, 2005]. When looking at regional GC content (± 100 bp), I see in *A. thaliana* and rice

that RR_{met} is highest at regions with lower GC content, and this risk drops off with increasing GC content (Figure 1.5C,K). This is in line with what is expected with deamination rates at CpGs themselves, and so implies a collateral model in plants, where the deamination of the CpG correlates with greater mutagenesis in the proximity. This does not seem to be the case for human, which consistently has an $RR_{met} < 1$ for the range of GC contents sampled (Figure 1.5G), implying there is some uncoupling between focal CpG deamination lesions and further collateral damage. This same pattern is seen when looking at local, motif-level, GC content (Figure 1.7).

Finally, I looked at whether differential sampling of chromatin states was driving the difference in effect between plants and humans. I stratified the paired sites by the chromatin state they were found in, as described by HMM calculations (see Methods). Then, for each chromatin state, I counted the number of singleton SNPs that occurred in the ± 1 -3bp surrounding either methylated or unmethylated CpG sites, normalised by the number of paired sites. In *A. thaliana*, although the mutation rate generally seems heterogeneous, the deviation from the diagonal where mutation rate $k_{methylated} = k_{unmethylated}$ seems consistent for all chromatin states except genic transcribed regions (Figure 1.8, blue), which lie on the diagonal. This might suggest that transcription-coupled repair negates the intrinsic mutational liability of methylation on its neighbourhood in *A. thaliana*. The similarity in mutational bias towards methylated sites between accessible sites and repressed H3K9me2 marked sites implies that the mutational liability is not restricted to closed heterochromatic regions, which might have been expected if we base the model on the fact heterochromatic regions are less accessible to DNA repair, causing CpG deamination lesions to be repaired less frequently [Makova and Hardison, 2015, Schuster-Böckler and Lehner, 2012]. In human (Figure 1.9) and rice (Figure 1.10), mutation rates associated with different chromatin states are more clustered together, with the outliers being of very low sample size. From this, I conclude that chromatin state is not the main differentiator of the RR_{met} effects seen between plants and human. This is further summarised in Figures 1.5 D, H, L, where I grouped the chromatin states by the coloured labels from Figures 1.8, 1.9, 1.10. We see that in human and rice, the RR_{met} is consistent between all the chromatin functional groups, and in *A. thaliana* we again see the higher RR_{met} in all groups apart from

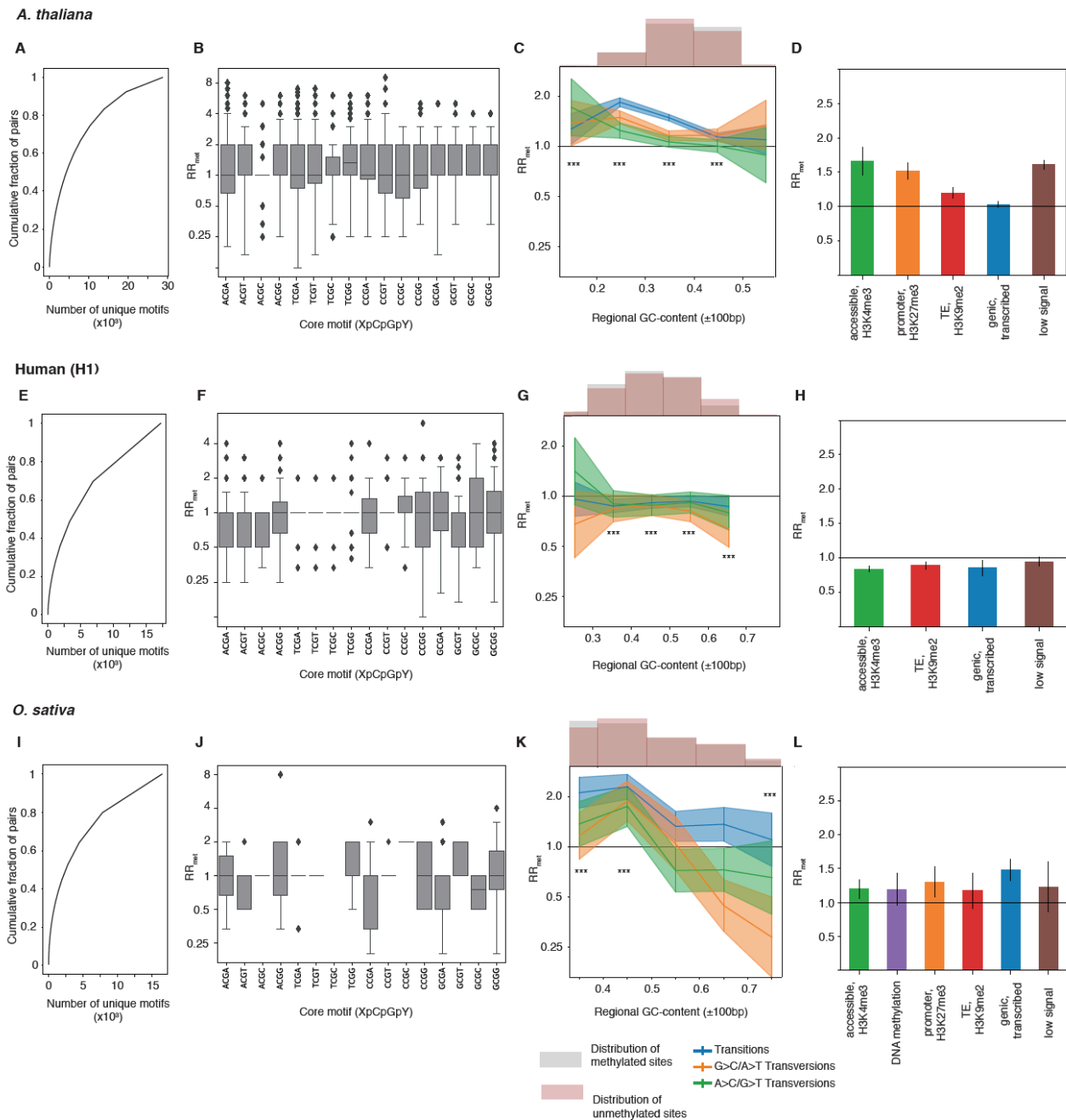


Figure 1.5: The relative mutational risk of methylation (RR_{met}) at CpG-neighbouring nucleotides in human, *A. thaliana*, and rice as a function of sequence context (B,F,J), regional (± 100 bp) GC content (C,G,K) and chromatin state classes (D,H,L). (A,E,I) The number of unique sequence contexts that contribute to genome-wide RR_{met} estimates in each species. (B,F,J) The RR_{met} for each sequence context grouped by core sequence context (XpCpGpY), displayed as weighted boxplots, with weights corresponding to the number of pairs associated with a given context. (C,G,K) Methylated and unmethylated paired sites were split and independently binned based on their regional GC content and RR_{met} calculated based on these bins. The distribution of regional GC content for methylated and unmethylated sites is shown by the histograms above each panel. (D,H,L) The RR_{met} for each broad chromatin state class, as defined based on the underlying original chromatin states (see Figure 1.8 for further breakdown). $***p < 0.001$. Vertical bars are confidence intervals on RR_{met} calculated using the delta method.

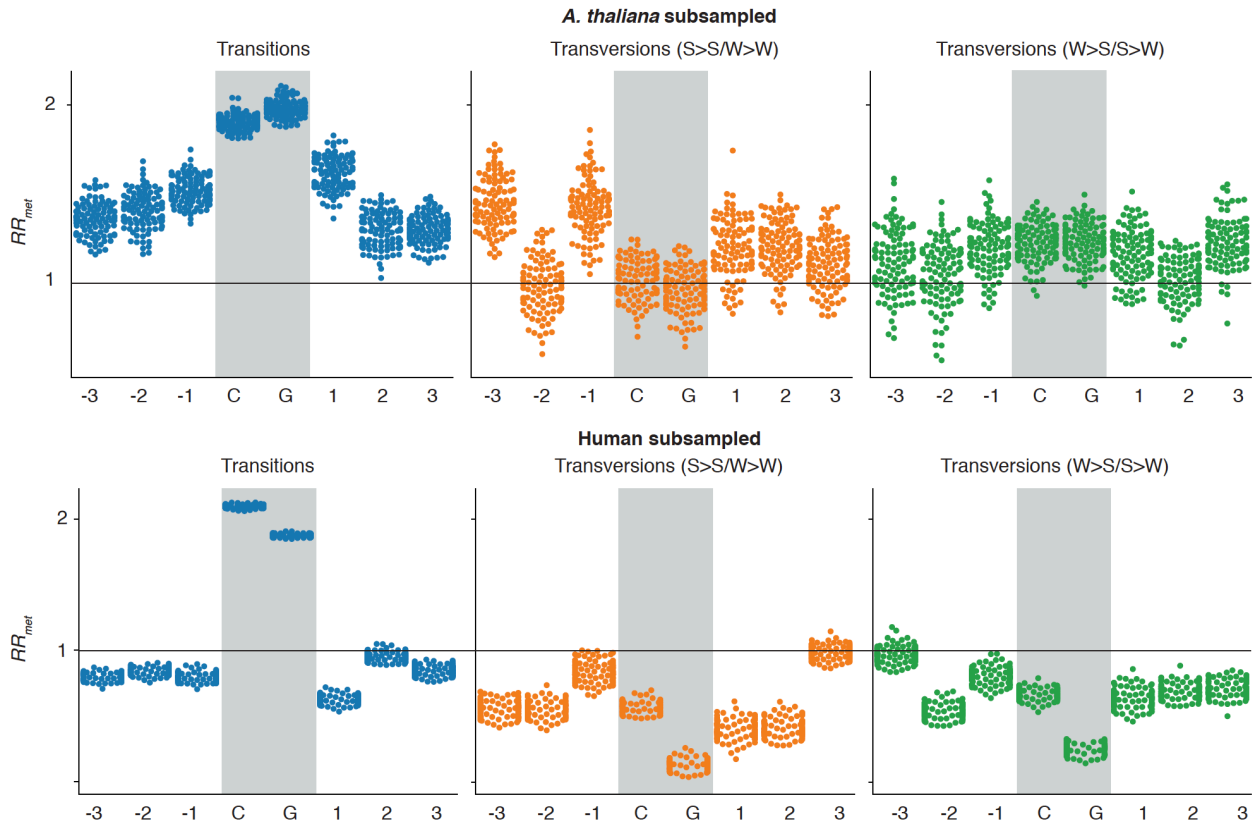


Figure 1.6: RR_{met} of subsampled motifs between human and *A. thaliana*. The number of occurrences for each observed nucleotide motif was counted in both human and *A. thaliana*. For each motif, the species' set with the larger number was randomly subsampled to the set size of the other species. After this, the RR_{met} was calculated for the reduced set of human and *A. thaliana* methylated and unmethylated sites. This process was repeated 100 times and each bootstrap was plotting as a point on the Figure.

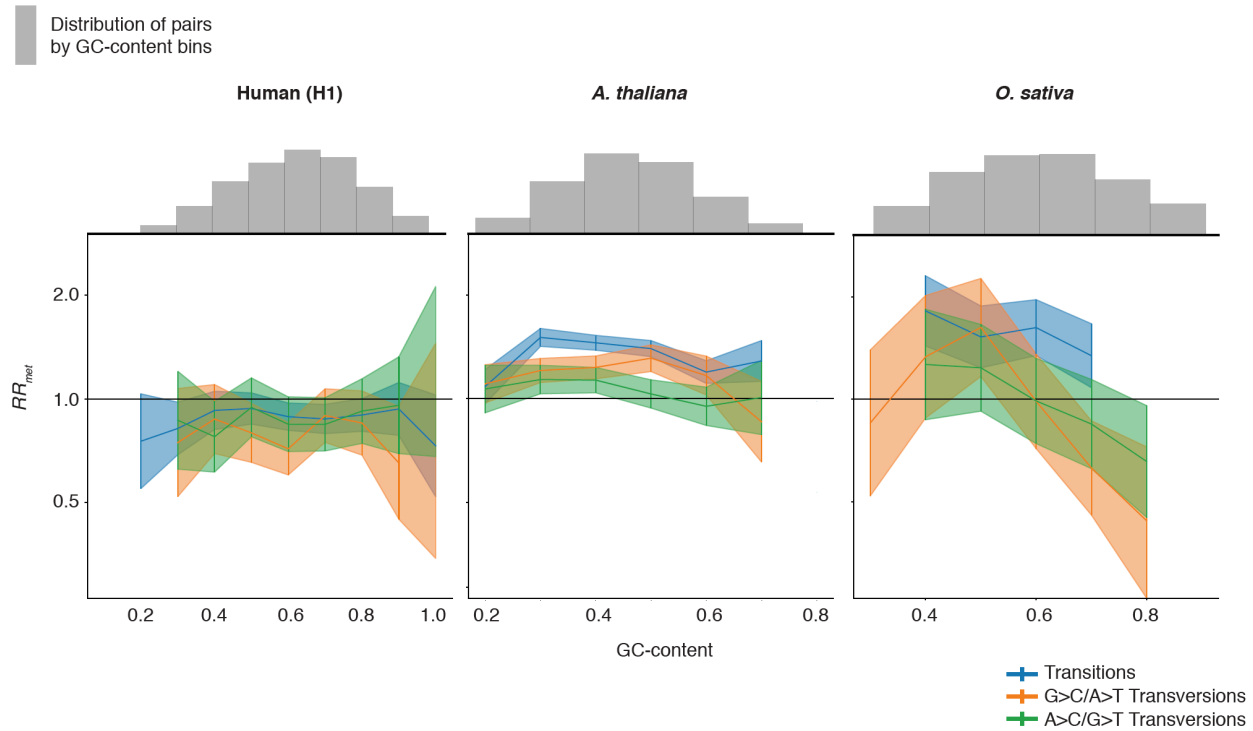


Figure 1.7: The relative mutational risk of methylation (RR_{met}) at CpG-neighbouring nucleotides in human, *A. thaliana*, and rice as a function of motif GC content. The paired site motifs were binned by GC content, and for each individual bin the RR_{met} was recalculated. RR_{met} estimates are omitted where error bars were larger than the plotting range. The histograms show the relative sample size of each bin. Vertical bars are confidence intervals on RR_{met} computed using the delta method.

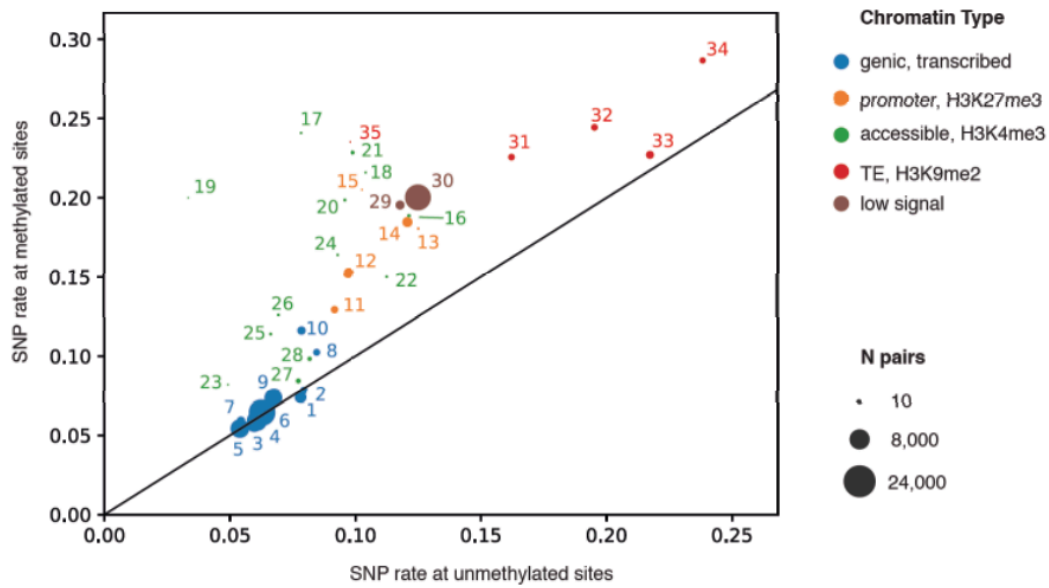
genic transcribed regions. Interestingly, here we see a reduction in RR_{met} in TE H3K9me2 repressed regions. This leads to the idea that heterochromatic regions, which generally are less accessible to other enzymatic processes, are more protected from the mutational liability of methylation on the neighbourhood in *A. thaliana*.

An important mechanism to consider is whether recombination rates and patterns are affected by the distribution of methylation. For example, if recombination is inhibited by methylation, there might be less incident damage due to double-stranded breaks around methylated CpG sites, leading to the reduction in neighborhood mutability that we see in the human data. However, in humans recombination hotspots are positively correlated with methylation-associated SNPs [Sigurdsson et al., 2009]. Later studies have shown that recombination correlates more with active chromatin states H3K4me3 and H3K27me3 [Zeng and Yi, 2014, Zamudio et al., 2015b], and after recombination methylation marks are then applied to silence the downstream product of double-stranded breaks in case the repaired product is deleterious to the cell, for example with a mutant protein or a transposable element [Cuozzo et al., 2007, Zamudio et al., 2015b]. In either case, the methylation marks that we see in both sperm germline and H1 ESCs are established post meiotic recombination, and would include sites that might have undergone recombination. Since I still see a reduction in neighborhood mutability in human this would either imply counterintuitively that at sites of recombination there is a reduced incidence of mutations, or more likely that recombination is not driving the main effect that I capture. Even if we look at the chromatin marks that correlate best with recombination in humans, we do not see a separate pattern at H3K4me3 and H3K27me3 chromatin marks, implying enrichment of recombination hotspots is not driving the reduction in neighborhood mutability around methylated sites 1.9. In Arabidopsis recombination is much more enriched in the euchromatic arms, and suppressed in heavily methylated, heterochromatic regions [Mirouze et al., 2012, Choi et al., 2013]. Again where recombination is most likely there seems to be less incident mutations, further implying that recombination is not a driving factor of the enrichment of mutations surrounding methylated bases in plants. Another potential risk factor is the accidental alkylation product of 3-methylcytosine (3mC) that sometimes gets incorporated by de novo methyltransferases [Rošić et al., 2018]. 3mC often results in double-stranded breaks

(DSBs) in the DNA, and is a high mutation risk that species either co-evolve alkylation damage repair machinery for, or lose methylation as an epigenetic mechanism entirely [Rošić et al., 2018]. The proportion of 3mC incorporated during methylation in human is not known. It is generally thought that such lesions are repaired by DNA direct reversal repair pathways by alpha-ketoglutarate-dependent dioxygenase (AlkB) proteins and as such do not require DNA synthesis and therefore error-free [Gutierrez and O'Connor, 2021], although in some cases when the load of alkylation is too large polymerases are involved in its repair [Furrer and van Loon, 2014]. If DNA methylation leads to the incorporation of 3mC, then one would assume this might happen similarly in both humans and plants, leading to DSBs occurring at methylated sites in both kingdoms. Therefore, if 3mC is driving the difference in mutational effect on the neighborhood surrounding the 5mC, it must involve downstream processing and repair. However, both arabidopsis and humans have several AlkB homologues [Meza et al., 2012, Marcinkowski et al., 2020], and therefore it is unlikely that a difference in alkylation processing is driving the large difference between humans and plants that I see. In addition, if DSBs downstream of a 3mC lesion were driving the difference, then one would also see a pattern emerge as a result of recombination DSBs, which we do not.

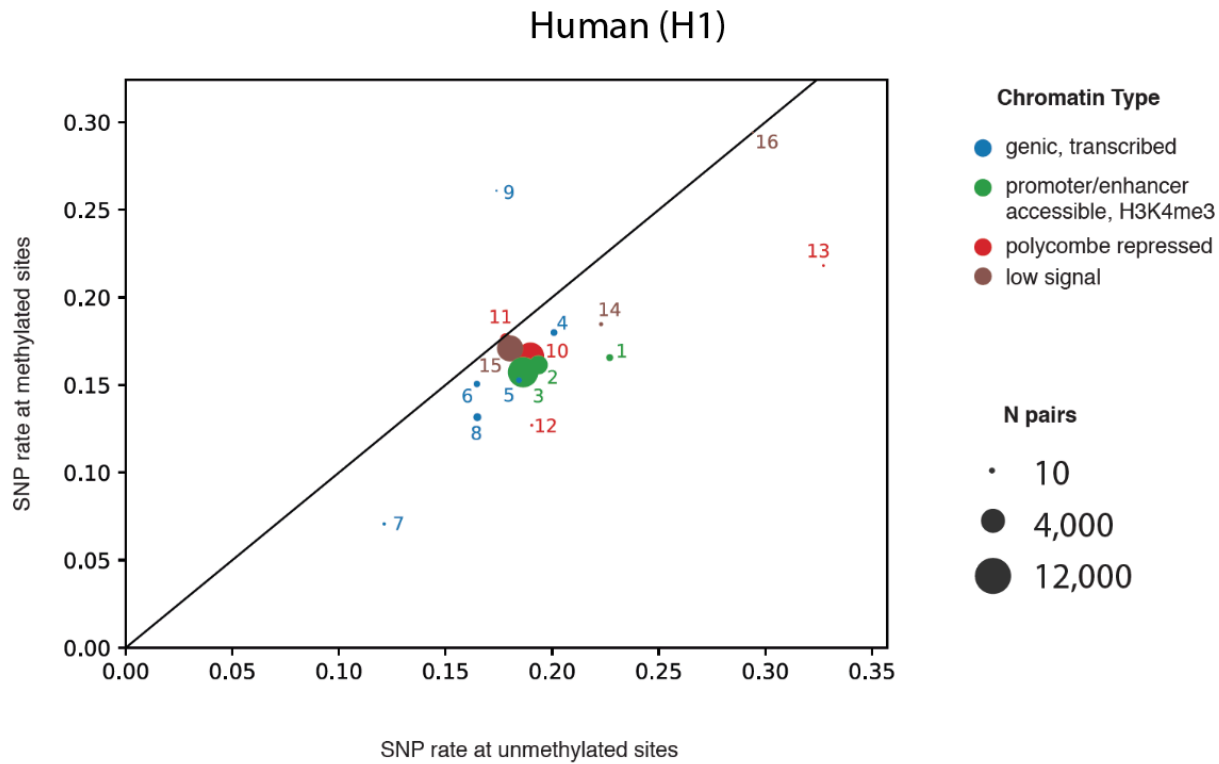
1.2.4 An interlude: The differences in methylation and demethylation machineries in plants and human

So why might we see differences between humans and plants? One possibility is that there are differences in regards to how cytosine methylation is deployed. Despite methylation generally being used in both kingdoms for repression of genes and transposable elements, the exact nuances in deposition, ubiquity, and regulation vary. In plants, a key role of methylation is in silencing transposable elements [Kim and Zilberman, 2014, Ikeda and Nishimura, 2015]. It also can appear repressing the promoters and transcription of individual genes [Zemach et al., 2010], and is interestingly also found in constitutively expressed, highly conserved genes [Takuno and Gaut, 2013, 2011]. In mammals, it is much more common to be found at promoters for silencing genes [Morselli et al., 2015, Schübeler, 2015], although it is also used for protection of the

A. thaliana

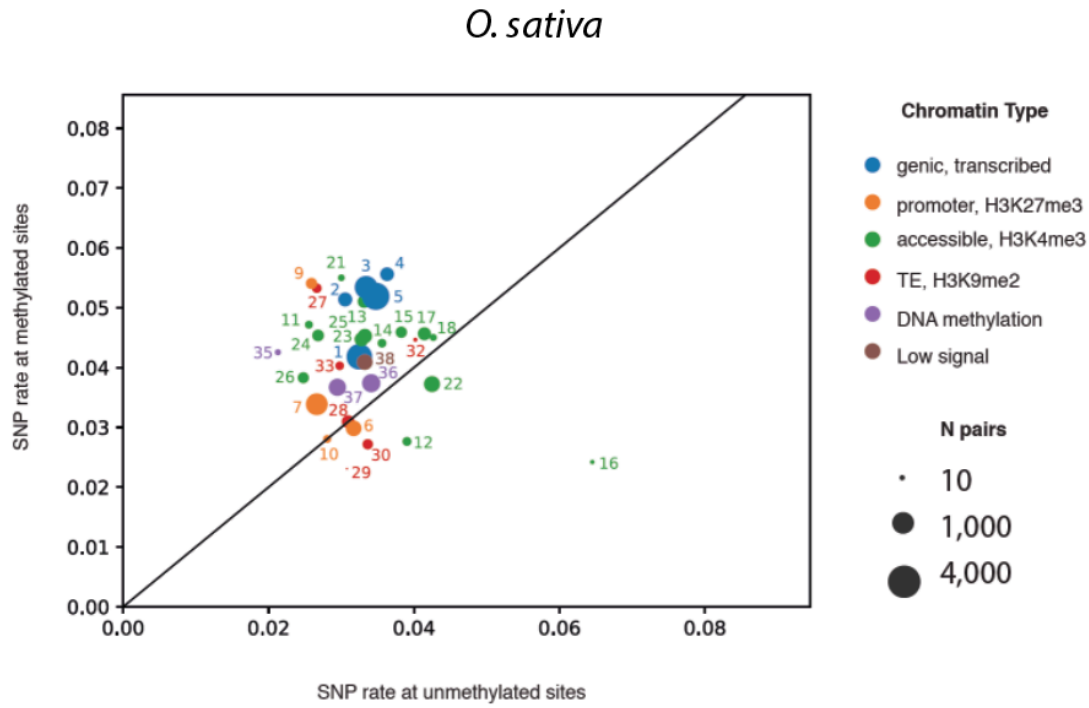
State	Preferential epigenetics marks	Preferential location
State 1	H3.3	3'UTR
State 2	H3.3, histone acetylation,H3K4me2,H2A.Z	CDS,3'UTR
State 3	H3K4me1,H3.3,H3.1	CDS
State 4	H3K4me1,H3.3	CDS,intron
State 5	H3K4me1,H3K36me3,H3.3,H3.1	CDS
State 6	H3K4me1,H3K36me3	Intron
State 7	H3K4me1,H3K36me3,H3K4me2	CDS,intron
State 8	H3K4me1,H3K4me2,H2A.Z	CDS
State 9	H3K4me1	Intron
State 10	H2A.Z	CDS,intron
State 11	H3K27me3,H2A.Z,H3K4me2	CDS
State 12	H3K27me3,H2A.Z	Promoter,CDS,intron,intergenic
State 13	H3K27me3,H2A.Z	Promoter,intergenic
State 14	H3K27me3	Promoter,intergenic
State 15	H3K27me3,accessible DNA	Promoter,intergenic
State 16	accessible DNA	Promoter,intergenic
State 17	accessible DNA	Promoter
State 18	accessible DNA	Promoter
State 19	accessible DNA	Promoter,snRNA
State 20	accessible DNA	Promoter
State 21	accessible DNA	Promoter
State 22	histone acetylation,H3K4me2	coding gene,miRNA,snoRNA
State 23	accessible DNA,H3K36ac,H3K56ac,H4K16ac,H3K4me3	Promoter,5'UTR
State 24	accessible DNA,histone acetylation,H3K4me3	5'UTR,snoRNA
State 25	histone acetylation,H3K4me3,H3K4me2,H2A.Z	Intron
State 26	histone acetylation,H3K4me3,H3K4me2,H2A.Z	CDS
State 27	H3K4me2,histone acetylation,H3K4me3,H2A.Z	CDS
State 28	H3K4me3,H3K4me2,H2A.Z	Intron
State 29	weak signal	Intergenic
State 30	rare signal	Intergenic
State 31	DNA methylation,H3K9me2,H3K27me3	Intergenic,miRNA
State 32	DNA methylation,H3K9me2	Intergenic,TE
State 33	H3K9me2,DNA methylation	TE
State 34	H3K9me2,DNA methylation,H3K27me1	TE,miRNA
State 35	H3K9me2,DNA methylation,H2A.X	Intergenic,pericentromere
State 36	CENH3,H3K9me2,DNA methylation,accessible DNA	rRNA,tRNA,centromere

Figure 1.8: The SNP incidence in methylated and unmethylated matched pairs in the different chromatin states of *A. thaliana* as defined by HMM calculation [Liu et al., 2017] (see Methods). Chromatin state labels are defined in the table underneath. The chromatin states are coloured by the major functional relevance of the chromatin marks, as I defined from reviewing of literature. The size of the point shows the number of matched pairs in each chromatin state.



State	Preferential location
State 1_Pro	Promoter
State 2_Enh1	Enhancer
State 3_Enh1	Enhancer
State 4_Egn1	Expressed gene
State 5_Egn2	Expressed gene
State 6_Egn3	Expressed gene
State 7_Egn4	Expressed gene
State 8_Egn5	Expressed gene
State 9_Egn6	Expressed gene
State 10_Rep1	Polycomb repressed
State 11_Rep2	Polycomb repressed
State 12_Het1	Heterochromatin
State 13_Het2	Heterochromatin
State 14_Low1	Low signal
State 15_Low2	Low signal
State 16_Low3	Low signal

Figure 1.9: The SNP incidence in methylated and unmethylated matched pairs in the different chromatin states of human as defined by HMM calculation [Ho et al., 2014] (see Methods). Chromatin state labels are defined in the table underneath. The chromatin states are coloured by the major functional relevance of the chromatin marks, as I defined from reviewing of literature. The size of the point shows the number of matched pairs in each chromatin state.



State	Preferential epigenetics marks	Preferential location
State 1	weak histone acetylation	intron,promoter,3'UTR,intergenic
State 2	H3K36me1,histone acetylation	promoter,3'UTR,intergenic
State 3	H3K36me2,histone acetylation	3'UTR
State 4	H3K36me2,H3K36me3,H3K4me2	exon
State 5	H3K36me3,H3K4me2	intron
State 6	H3K27me3	intergenic,intron,promoter
State 7	H3K27me3	intergenic,intron,promoter
State 8	H3K27me3	intergenic,intron,promoter
State 9	H3K27me3,H2A.Z	intergenic,intron,promoter
State 10	H3K27me3,H2A.Z,H3K9ac,H3K23ac,H3K4me3,H3K4me2	exon
State 11	H3K9ac,H3K23ac,H3K4me3,H3K4me2,H3K27me3,H2A.Z	exon,promoter
State 12	H2A.Z,histone acetylation,H3K4me2,H3K4me3	exon,promoter
State 13	histone acetylation,H3K4me3,H3K4me2,H3K36me3,H3K36me2,H2A.Z	5'UTR
State 14	H3K4me3,H3K4me3,H3K36me3,H2A.Z,histone acetylation	intron
State 15	H3K4me3,H3K4me3,H3K36me3	intron
State 16	H3K4me3,histone acetylation,H2A.Z	exon,promoter,5'UTR
State 17	histone acetylation,accessible DNA,H3K4me2,H3K4me3	promoter,exon,3'UTR
State 18	histone acetylation,accessible DNA,H2A.Z,H3K4me3	5'UTR,promoter,intergenic
State 19	DNA methylation,H3K4me3,H4K12ac,H4K16ac,OsNF-YB1,OsZIP	centromere,intergenic
State 20	accessible DNA,H2A.Z,histone acetylation,H3K27me3,DNA methylation	intergenic,promoter,intron
State 21	accessible DNA	intergenic,promoter
State 22	accessible DNA	intergenic,promoter
State 23	accessible DNA	intergenic,promoter
State 24	accessible DNA	intergenic,promoter
State 25	H2A.Z,H3K4me2,H3K4me3,histone acetylation	promoter,intergenic,intron,3'UTR
State 26	DNA methylation,OsNF-YB1	intergenic,promoter,intron
State 27	H3K9me2,DNA methylation,OsNF-YB1	intergenic,TE,promoter,exon
State 28	H3K9me2,DNA methylation,Mnase	TE,intergenic,intron,promoter,exon
State 29	DNA methylation,H3K9me2,CENH3,Mnase	TE,exon,promoter,intron,intergenic
State 30	H3K9me2,DNA methylation	TE,exon,intron,promoter,intergenic
State 31	H3K9me2,CENH3,DNA methylation	TE,intron,intergenic,exon,promoter
State 32	H3K9me2,CENH3,DNA methylation,Mnase	TE,intergenic,intron,exon,promoter
State 33	DNA methylation,H3K9me2	TE,exon,intergenic,promoter,intron
State 34	DNA methylation,H3K36,H3K4me2	intron
State 35	DNA methylation,H3K27me3,H3K36me1,H3K36me2,H3K36me3	intergenic,intron,promoter
State 36	DNA methylation,H3K36me1	intergenic,promoter,intron
State 37	DNA methylation	intergenic,intron,promoter
State 38	rare signal	intergenic,promoter,intron

Figure 1.10: The SNP incidence in methylated and unmethylated matched pairs in the different chromatin states of rice as defined by HMM calculation [Liu et al., 2017] (see Methods). Chromatin state labels are defined in the table underneath. The chromatin states are coloured by the major functional relevance of the chromatin marks, as I defined from reviewing of literature. The size of the point shows the number of matched pairs in each chromatin state.

genome from transposons [Zamudio et al., 2015a]. DNA methylation is more likely to colocalise with repressive histone modifications around TEs in Arabidopsis and rice [Liu et al., 2017], as opposed to around exons and H3K36me3 dense regions in human [Collings and Anderson, 2017]. This is further demonstrated in the stoichiometry of DNA methylation: humans methylate the majority of their genome (80% of their CpGs) apart from CpG islands, with very little CpHpG or CpHpH methylation [Lister et al., 2009]. On the other hand, plants methylate many fewer of their CpGs (24% and 63% in rice), as well as utilising CpHpG and CpHpH contexts for methylation much more and more specifically towards TE regions [Vanyushin, 2006, Feng et al., 2010]. Another potential skewing factor could be a technical bias when pairing sites. Since in the human genome, there are many more methylated sites compared to unmethylated sites, I am limited by where the unmethylated sites occur. To find the nearest methylated pair, I have to start the search from each unmethylated base. This may bias which methylated CpGs are selected during the pairing protocol, and so the majority of our sampled set will be dictated by where unmethylated CpGs occur. In plants the percentage of methylated CpGs is lower (24% of CpGs in Arabidopsis and 63% in rice), therefore unmethylated sites would be a smaller limiting factor. This could potentially lead to any effect that we see being an artifact of sampling from regions that are nearest to sites stringently regulated to be unmethylated in human, compared to the more common plant unmethylated sites. However, the fact that we previously saw no connection between different chromatin regions in both plants and human implies that this effect is not merely due to differential sampling of the genome, and is a real intrinsic biological difference between lesion formation and processing in plants and human.

Another difference occurs in the deposition and removal of CpG methylation. In both plants and mammals, DNA methylation is catalysed by DNA methyltransferases using S-adenosyl-L-methionine as the methyl donor, and active demethylation involves base excision repair [Zhang et al., 2018]. However, there are some clear differences that have been very well studied in Arabidopsis. Despite there being some de novo methylation by the methyltransferase DNMT3 in humans, the majority of methylation is perpetuated via maintenance methylation by methyltransferase DNMT1 of hemi-methylated strands accrued post DNA replication [Goll and Bestor, 2005]. The DNMT3 family of methyltransferases are mainly expressed in early development

to establish genomic imprinting and other germline methylation patterns in mammals [Hata et al., 2002, Okano et al., 1999]. On the other hand, de novo methylation is much more prevalent in plants, being applied via RNA-directed DNA methylation (RdDM, Figure 1.11). This mechanism can target all contexts of cytosine methylation, not simply symmetric patterns like CpG. De novo DNA methylation in plants is triggered by short non-coding 24-nucleotide RNAs [Bouyer et al., 2017]. These RNAs are generated by the transcriptional activity of PolV and PolIV, before being cleaved into smaller fragments by Dicer-like proteins. The resulting short, 24-nt RNAs are then loaded back onto target sites by Argonaute proteins, which leads to the recruitment of the methyltransferase DRM2. This is an important mechanism to accurately target silencing to active TEs: whether it is targeting TE RNA fragments transcribed during global demethylation in the endosperm developmental stage [Ji et al., 2019, Hsieh et al., 2009], or during regular plant life. Expressed TEs will be targeted for degradation, and the short RNA products of such degradation are used to localise de novo methylation to suppress aberrant TEs.

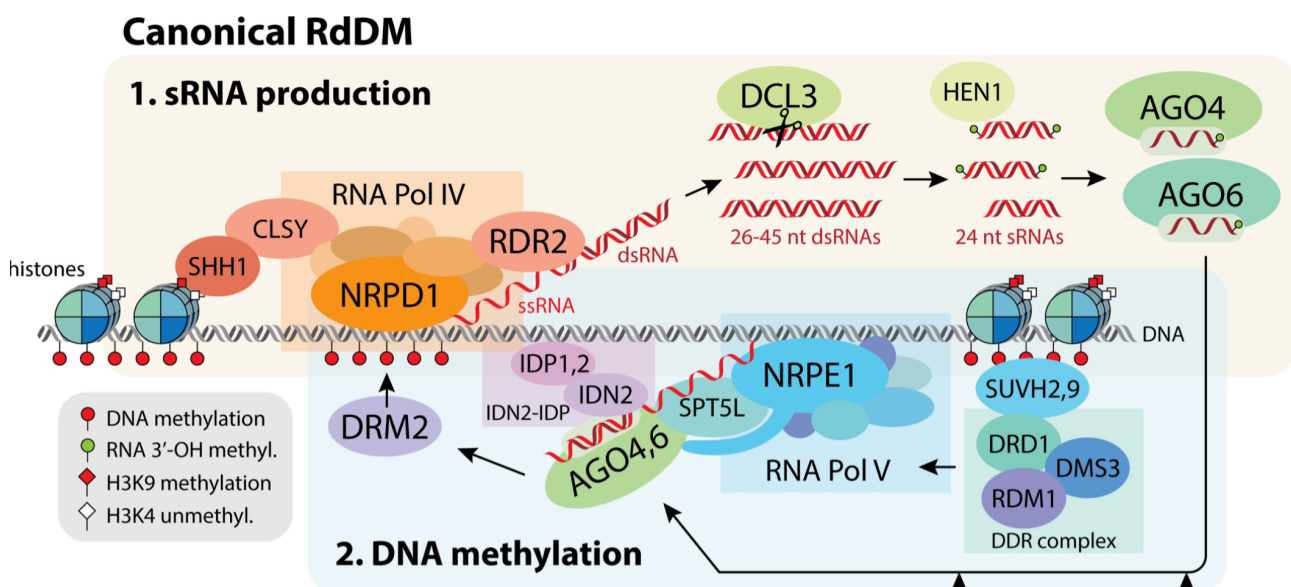


Figure 1.11: An overview of canonical RdDM, taken from Erdmann and Picard [2020]. The canonical RdDM pathway can be broken into 2 parts: 1) short 24-nt RNA production; 2) targeting DNA methylation to complementary sites to the short RNAs.

In addition to differences in methylation of target sequences, there are also differences in the demethylation of 5mC in both plants and human. Both groups have active DNA demethylation. In human, demethylation occurs via ten-eleven translocation (TET) enzymes which oxidize 5mC bases to yield 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine. These

oxidised derivatives can then be excised by DNA glycosylases [Zhang et al., 2018, Wu and Zhang, 2017]. Previous work had shown that sites that spend a larger proportion of their time in a hydroxymethylated state are subject to different mutation dynamics [Supek et al., 2014], so this might be a contributing factor. In plants active DNA demethylation occurs via direct excision of the 5mC base using 5mC DNA glycosylases such as ROS1 (Repressor of Silencing 1) or DME (Demeter) [Zhu, 2009]. This lack of oxidised derivatives of 5mC might contribute to a difference in mutational dynamics, however the fraction of sites that persist to be measured as 5hmC sites in human is only 0.1% of cytosines, which is unlikely to explain the large difference between plants and humans. The resulting abasic sites, however, might be the driving factor. The abasic sites produced as a product of DNA glycosylase activity are repaired by base excision repair (BER). There are two main flavours of this pathway: Short-Patch (SP-BER) where only a single nucleotide is excised and replaced, and Long-Patch (LP-BER) where a longer tract of nucleotide resynthesis occurs [Dianov et al., 2003]. This resynthesis of a longer section of DNA following LP-BER could be error-prone, which could explain the risk to the neighbourhood surrounding 5mC-related lesions. These flavours exist in both human and plants, as seen by experiments in cell extracts [Córdoba-Cañero et al., 2009, Martínez-Macías et al., 2013], and the decay pattern of 5mC neighbourhood risk is consistent with a localised repair process. However, if these repair processes are involved in altered neighbourhood mutability around CpGs, why might one see opposite mutational pattern of the same lesion in human and plants? The mutation rate at a site reflects a combination of two distinct factors: the relative risk of forming a lesion and what pathway is chosen downstream. Firstly, in our example this is the relative rates of unmethylated cytosine deamination versus methylated cytosine deamination (if the neighbourhood liability of mutation is truly linked to focal CpG deamination), and determines the amplitude of the effect. One might reasonably assume this to be approximately constant between the differing species, based *a priori* on deamination being a chemical event, on past experimental results seeing the consistently higher deamination rates of 5mC, as well as in my protocol seeing the same RR_{met} of around a 2-fold increase at the focal CpG between the species in Figure 1.3. Which leads onto the second factor: what pathway is chosen downstream of the lesion? In our specific case, one would ask: do U:G mismatches or T:G mismatches get handled more efficiently? If T:G

mismatches (the result of 5mC deamination) are handled more efficiently with less mutational impact than U:G (the result of unmethylated C deamination) in the proximity, it would follow that $RR_{met} < 1$. If on the other hand, U:G is handled more efficiently, then T:G mismatches might cause a broader mutational shadow, leading to $RR_{met} > 1$. Finally, if U:G mismatches are repaired less efficiently in mammals but more efficiently in plants, this might account for the opposing trends in neighbourhood mutability we observe.

1.2.5 The effect of methylation and demethylation machinery in plants

To investigate any links between methylation/demethylation and RR_{met} , I studied the methylomes of several *Arabidopsis* methylation and demethylation machinery mutants [Wierzbicki et al., 2012]. I utilised the changes in methylation in both a *nrpd1* mutant (NRPD1 is a subunit of PolIV that confers its specificity) and a *ros1* mutant to identify sites that are targeted by RdDM methylation and active demethylation. In the *nrpd1* mutant, the regions where the level of methylation decreases can be assumed to be the regions where RdDM acts. In the *ros1* mutant sites where the level of methylation increases can be labelled as the targets of active demethylation. At some sites both active demethylation and RdDM can act. This is especially common in plants and is a key role of active demethylation, since RdDM will spread from transposons and ROS1 activity is required to keep nearby genes active, leading to much higher turnover of methylation and subsequent demethylation at such sites. If in the WT the active demethylation outweighs the RdDM, then no methylation is seen and no decrease in methylation is possible with *nrpd1* knockout. Therefore, I also compared the *ros1/nrpd1* double mutant to the *ros1* single mutant to highlight these hidden RdDM sites. In addition to *ros1* and *nrpd1* mutants, I also had access to bisulphite-sequencing data in *A. thaliana* of *ddm1* and *drd* mutants from a study by Zemach et al [Zemach et al., 2013]. DDM1 is a chromatin remodeler and is necessary for methylation of heterochromatic regions in *A. thaliana*, and *drd* mutants disrupt the initiation of RdDM process at both euchromatic and heterochromatic TEs. Therefore, using the requirement of DDM1 for methylation, I could define RdDM targets that

are accessible to the methylation and/or repair machinery, versus targets that require heterochromatic remodelling to do so, adding an extra step into the lesion processing that could end up being rate limiting for the repair pathway. From the comparisons of these mutants, I defined sites as follows: a ROS1 site was one where methylation was greater in the *ros1* mutant than the wild-type (WT); an NRPD1 site was one where the methylation was lower in the *nrpd1/ros1* double mutant compared to the *ros1* knockout in order to capture both regular and cryptic RdDM sites; a DDM1 site was one where the *ddm* mutant had lower methylation and the *drd* mutant had no effect. Since the sample size for pairing by both chromatin and methylation machinery zones became too small, I paired sites based solely on their methylation machinery zone instead, and the SNP incidence was recalculated as done previously.

After doing so, I found that regions that were targeted by both ROS1 and NRPD1 had a lower mutational effect of methylation than regions that do not (Figure 1.12A). This result does not support a model where high turnover is at fault for a high RR_{met} . Interestingly sites requiring DDM1 for methylation have a higher RR_{met} (Figure 1.12A). This creates a picture where closed heterochromatin that needs DDM1 to unravel it seems to sequester and collect more mutational damage at sites proximal to 5mC compared to sites that are naturally more accessible, such as the sites that can be accessed by enzymatic machinery such as ROS1 and NRPD1. When pairing by both chromatin and (de)methylation type site, even though the sample sizes were small and the errorbars large, the pattern was qualitatively the same (not shown). To see if this effect is proportional to the rate or amount of activity of each methylation or demethylation process, the SNPs and sites were binned by the size of methylation changes caused by knocking out *ros1* and *nrpd1*. The *ros1* mutant was compared to the *nrpd1/ros1* double knockout mutant to capture both cryptic and regular RdDM sites. The higher the decrease in methylation in the double knockout, the greater the RdDM activity. For ROS1 activity I compared the *ros1* mutant to the WT, and where the methylation increased the most were the sites with greatest ROS1 activity. As shown in Figure 1.12B, in the quadrant with greatest ROS1 activity and greatest NRPD1-mediated RdDM (top-right), the mutational effect of methylation is lowest.

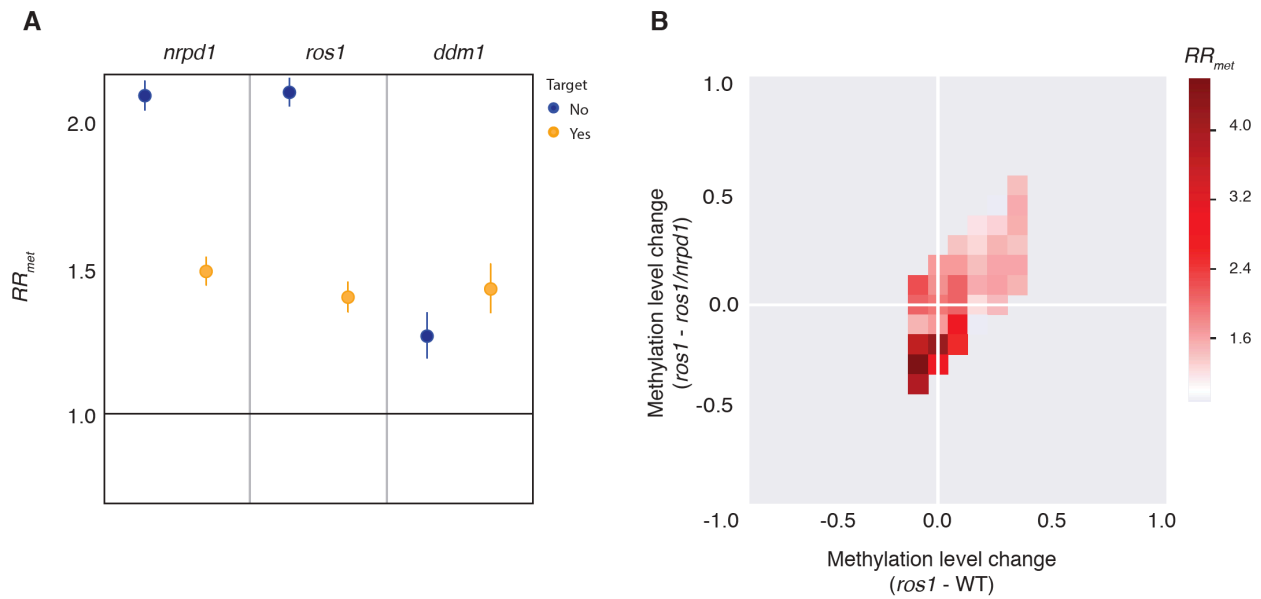


Figure 1.12: **(A)** The relative mutational risk of methylation (RR_{met}) at CpG-neighbouring ($\pm 1-3$ bp) nucleotides in *A.thaliana* as a function of whether a given site is targeted by RNA-directed DNA methylation (involving *nrpd1*), the DNA glycosylase ROS1, or the chromatin remodeler DDM1 (see 2.4 for how targets and non-targets were defined). $p < 10^{-60}$ for all sets. Vertical bars are confidence intervals on RR_{met} computed using the delta method. **(B)** RR_{met} as a function of relative methylation change in *ros1* deletion mutants compared to the corresponding wildtype strain, and compared to *nrpd1/ros1* double mutants. Sites with more methylation in *ros1* compared to the WT are sites of high ROS1 activity, and sites with more methylation in *ros1* compared to *nrpd1/ros1* are sites where the RNA-directed DNA methylation machinery is active.

1.2.6 Discussion: potential mechanisms and an evolutionary explanation

Seeing a higher RR_{met} at heterochromatic methylated sites requiring DDM1 and a lower RR_{met} at sites accessible to both RdDM and demethylation machinery is inconsistent with a model where the increased mutability is tied to more numerous methylation and demethylation cycles. On the contrary, it implies that regions more accessible to enzymatic action have a lower mutational risk proximal to cytosine methylation, and regions locked away in closed chromatin have a propensity to accumulate more mutations. Therefore, these results are more consistent with a model where the increased RR_{met} is tied to downstream repair, where regions more accessible would be easier to repair and therefore have a lower RR_{met} . Following from that observation, the difference in pattern seen between humans and plants might be explained by the preference of downstream repair pathways for lesions at methylated and unmethylated sites. This returns us to the earlier discussion in 1.2.4, where I mentioned that the rate of lesion formation asserts the amplitude of mutational risk, and bias in repair pathway choice might denote the relative direction of mutational enrichment comparing two lesions. In particular, I will focus on the SP and LP-BER pathways which are known to be differentially recruited to different lesions. In human, 8-oxoguanine lesions are thought to mostly trigger SP-BER [Fortini et al., 1999], whereas U:G mismatches, the product of unmethylated cytosine deamination, are mostly associated to LP-BER. Bennett et al. studied the preference of repair of U:G mismatches in mouse embryonic fibroblasts, where they show that in 80% of cases resynthesis post uracil removal by uracil DNA glycosylase exceeded a single nucleotide [Bennett et al., 2001]. This phenomenon might be tied to a biased interaction of the glycosylase with PCNA, which is the sliding clamp for the endonuclease and polymerase involved in initiating LP-BER [Fortini and Dogliotti, 2007]. Furthermore, such uracil-initiated LP-BER has been associated with a higher mutational risk at neighbouring sites in other studies. Chen et al. used a SV40 episome that could replicate in human cells as a vector to test the repair upon artificially introduced mismatches [Chen et al., 2014]. They found both T:G and U:G mismatches were associated with collateral damage to the neighbourhood surrounding them, however, the risk of mutagenesis

surrounding U:G mismatches was approximately 7-fold higher than for T:G mismatches. This is consistent with my findings in Figure 1.2, that even though T:G mismatches would occur more frequently, the greater liability of U:G mismatches on the sequences in which they occur is greater, leading to $RR_{met} < 1$.

Not only might LP-BER fit the pattern regarding mutational properties, but the distance of the effect from the focal CpG is similar to the estimated repair tract lengths of LP-BER. In both human and *A. thaliana*, the RR_{met} around the focal CpG quickly returns towards the baseline after roughly 2 to 3 bases, suggesting the respective mutational protection and liability is a relatively local phenomenon. In SP-BER, only a single nucleotide is added in place of the excised abasic site [Dianov et al., 1992, Singhal et al., 1995, Kubota et al., 1996]. By tracking repair at U:G mismatches, LP-BER has been observed to produce repair tracts of up to 3bp in *A. thaliana* [Córdoba-Cañero et al., 2009, Martínez-Macías et al., 2013], and similar tracts of 2-13bp in mammals [Klungland and Lindahl, 1997, Fortini and Dogliotti, 2007, Krokan and Bjørås, 2013]. These repair tracts appear consistent in length to the effect I see looking at the RR_{met} trend in both human and plants proximally to focal CpGs in this study. Other potential repair mechanisms include nucleotide excision repair (NER) and mismatch repair (MMR), but for these mechanisms repair tracts are typically much longer. In NER, repair tracts are generally longer than 20bp [Brookman et al., 1996], and even longer in MMR, where they extend to several hundred bases or more [Fang and Modrich, 1993].

Seeing as both human and plants have a different repertoire of DNA glycosylases, it is conceivable that these glycosylases might be differentially recruited to lesions formed at methylated or unmethylated sites as a result of cytosine deamination. A possible hypothesis would be that the RR_{met} effect is based on differential recruitment of the LP-BER pathway to lesions at 5mC or unmethylated cytosine sites in plants and human. This is further reinforced by the pattern I see when looking at single-nucleotide deletion rates which are similar to those seen in other experiments investigating LP-BER [Bennett et al., 2001, Lyons and O'Brien, 2010]. Despite this, the precise mechanism is still unclear. In the study mentioned previously [Chen et al., 2014], the authors suggest that the excess mutability requires both BER and MMR and propose a model similar to how somatic hypermutation occurs at immunoglobulin genes. In these events,

base-excised lesions are hijacked by the MMR machinery which had also been previously shown to occur at U:G mismatches [Schanz et al., 2009, Peña-Díaz et al., 2012]. Such MMR-hijacking has also been seen in the context of active demethylation [Grin and Ishchenko, 2016], and the mutational load is attributed to the interaction of the APOBEC family of cytidine deaminases which couple with MMR machinery and attack single-stranded cytosines. However, I see no enrichment of the classical APOBEC mutational signature [Alexandrov et al., 2013] which is an enrichment of C to T changes in a TpCpN context. One explanation might be that APOBEC-related excess mutagenesis is simply a specific manifestation of a more general liability that I observe in my analysis, a product of being more fragile in a single-stranded state. Or perhaps there are other mechanisms and pathways at play.

From an evolutionary standpoint, an argument can be made that the observable pattern might be evolved rather than random. Selecting for differing mutational burdens via the choice of repair pathways might be adaptive. Due to the much lower stoichiometry of CpG methylation in plants, CpG methylation is relatively more specifically entwined with the functional context of the genome, in contrast to the ubiquity of methylation in the human genome. In plants, methylation is more targeted towards TEs and gene bodies [Zhang et al., 2018]. Plant genomes may have had no selective pressure against the higher mutational burden of the neighbourhood in these regions and might even benefit from higher mutation rates where those lead to the suppressed TEs becoming disabled faster. A supporting pattern between mutation rates across the genome and the fitness consequences of mutations in sub-regions of the genome were seen in *Arabidopsis* more generally [Monroe et al., 2020]. Mutations occurring in suppressed TEs have little if any fitness consequence for the plant, and so an increased mutation rate is likewise not avoided at such locations. Methylation in plants is also found in the gene bodies of highly conserved and constitutively expressed genes. However, these might be protected by transcription-coupled repair, as seen by the great reduction in RR_{met} in transcribed regions shown in Figure 1.5D. Therefore, there might be limited selective pressure to reduce the mutational liability proximal to methylated cytosines more genome-wide, since gene bodies were already protected by an established mechanism. In contrast, in mammals methylation is ubiquitous and the underlying sequence often remains important. This is explicitly highlighted by

the strategies of X-chromosome inactivation and genomic imprinting, where important development genes need to be silenced for an extended period of time before being passed onto the person's progeny. This might explain why elevated mutational loads in the neighbourhood of methylated cytosines might be much less tolerated in humans, influencing the evolution of more efficient repair of 5mC lesions. An interesting point to note is that, during post-replicative repair in mammals, G:T mismatches at hemi-methylated sites are converted back to G:C with 90% efficiency, showing some adaptation of the system to the predictable incidence of several-fold greater deamination rates of 5mC compared to regular cytosine [Brown and Jiricny, 1987]. No such bias has been observed in plants. However, this has only been investigated in tobacco protoplasts [Inamdar et al., 1992].

1.2.7 Inconsistency with previous work by Qu et al - Mapping Artifacts

My results in human ($RR_{met} < 1$ at sites flanking the CpG) contradict a prior analysis of genome-wide SNP incidence near human methylated and unmethylated sites [Qu et al., 2012]. As highlighted in the introduction, Qu et al. found more mutations around methylated sites in the human genome. However, their analytical pipeline had four important distinctions from my approach: the bisulphite-sequencing dataset, the human polymorphism dataset, defining regions proximal to focal CpGs, and read coverage of reliable methylation calls. To elaborate, the differences were as follows:

1. Qu et al. used methylation calls from sperm cells [Molaro et al., 2011], as opposed to H1 ESCs that I used.
2. Instead of pairing sites by sequence and chromatin context, they grouped the bases ± 10 bps from focal CpGs into what they term *CpG blocks*. In these blocks, they average the methylation levels and read counts, and define methylated blocks as those with overall CpG methylation levels $\geq 80\%$ and unmethylated blocks as ones with $\leq 20\%$

methylation calls. They then computed an overall SNP incidence in each block, as opposed to by residue. They do discount transitions at CpG/TpG/CpA sites, like in my pipeline.

3. They analysed sites with ≥ 5 reads, as opposed to my 10 read cut-off.
4. They used SNPs from the HapMap project (CEU population [Consortium et al., 2003]) which mostly samples common SNPs ($> 5\%$ allele frequency) as opposed to the SNPs from the much larger gnomAD database which are predominantly rare.

Firstly, I re-ran my analysis whilst using the methylation calls from the sperm bisulphite-sequencing experiments. I obtain a very similar trend ($RR_{met} = 0.877, p = 1.68 \times 10^{-160}$ Figure 1.13), implying that the differences between sperm cell and H1 ESC methylation is not the driver of the discrepancy. Also, similar results are obtained when filtering sites with ≥ 5 reads instead of ≥ 10 reads (overall $RR_{met} = 0.889, p = 3.19 \times 10^{-81}$) which also increases the number of paired sites to 162,156 pairs. Unsurprisingly, there is also no difference when requiring the different cut-off of 80% for methylated sites instead of my cut-off of 70%, due to the significant bimodal pattern of CpG methylation calls.

Seeing as my protocol seems robust to differences in dataset usage from Qu et al's work, the next step was to reproduce the original finding in full using the same conditions. Doing so, I replicate their results qualitatively, showing an RR_{met} of 1.4 (Figure 1.14A). This did not change when using H1 instead of sperm data, re-confirming that methylation differences between the two sets are not confounders of the pattern. Even applying various genome-masks, including repeat-masking [Smit et al., 2015] or the DNase accessibility mask from ENCODE [Consortium et al., 2012] (defined by the 1000 Genomes project [Siva, 2008], see section 2.4), has little effect. However, when considering SNPs from the 1000 Genomes project or from gnomAD, RR_{met} reverses back to being < 1 as seen in my original analytical pipeline (overall $RR_{met} = 0.971$ and $RR_{met} = 0.87$ for 1000 Genomes and gnomAD respectively, Figure 1.14).

This drastic change in behaviour raises the question: how does the pattern observed change as a function of allele frequency? In my pipeline I used singleton SNPs from gnomAD as a proxy

of mutation rate, which would have a frequency of 1 in approximately 31,000 alleles. However, in the CEU population of HapMap the total allele number is only 440, and therefore SNPs from this dataset would generally capture more common SNPs. To test this, I investigated the SNP rate of the focal CpG itself as a function of increasing allele frequency. In doing so, an unintuitive but striking pattern is seen (Figure 1.14B). For SNPs that are relatively rare with allele frequency less than 10%, the expected pattern is seen with more SNPs occurring at 5mC sites compared to unmethylated sites. However, as the SNPs rise to higher allele frequencies, the pattern reverses with more SNPs occurring at unmethylated sites compared to methylated ones. Based on decades of research into methylcytosine deamination, this result is unlikely to be correct at a biological level and instead highlights a potential technical issue. I suggest that the following might be at fault. First, we must recall that bisulphite treatment of DNA converts unmethylated cytosines to uracils that will be sequenced as thymines later, whereas methylated cytosines remain as cytosines. These sequencing reads must then be mapped against the human reference genome. But what happens if the reference genome is not identical to the sequenced genome? Both these methylation call datasets were sequenced from a specific sample, either from H1 hESC [Lister et al., 2009] or from pooled DNA from two anonymous sperm donors [Molaro et al., 2011], and neither will be completely identical to proposed human reference genome. This can lead to mapping errors of methylated or unmethylated sites whenever there is a CpG in the reference but a TpG SNP in the H1 cell assays or sperm donor genomes. The bisulphite-sequencing data is blind to these SNPs and so the site will be called as an unmethylated CpG. This problem is exacerbated by the fact that these TpG SNPs likely resulted from deamination of what used to be a methylated CpG. *Ad continuum*, this leads to an enrichment of the most mutable methylated CpGs in the unmethylated CpG set, thereby skewing downstream analysis. Importantly, this issue is only noticeable when analysis depends on variants that are relatively common on average, as is the case for HapMap. When the vast majority of SNPs are singletons or rare (as in the case in 1000 Genomes and especially gnomAD), it becomes very unlikely for a singleton SNP to have polarised the methylation status of a base.

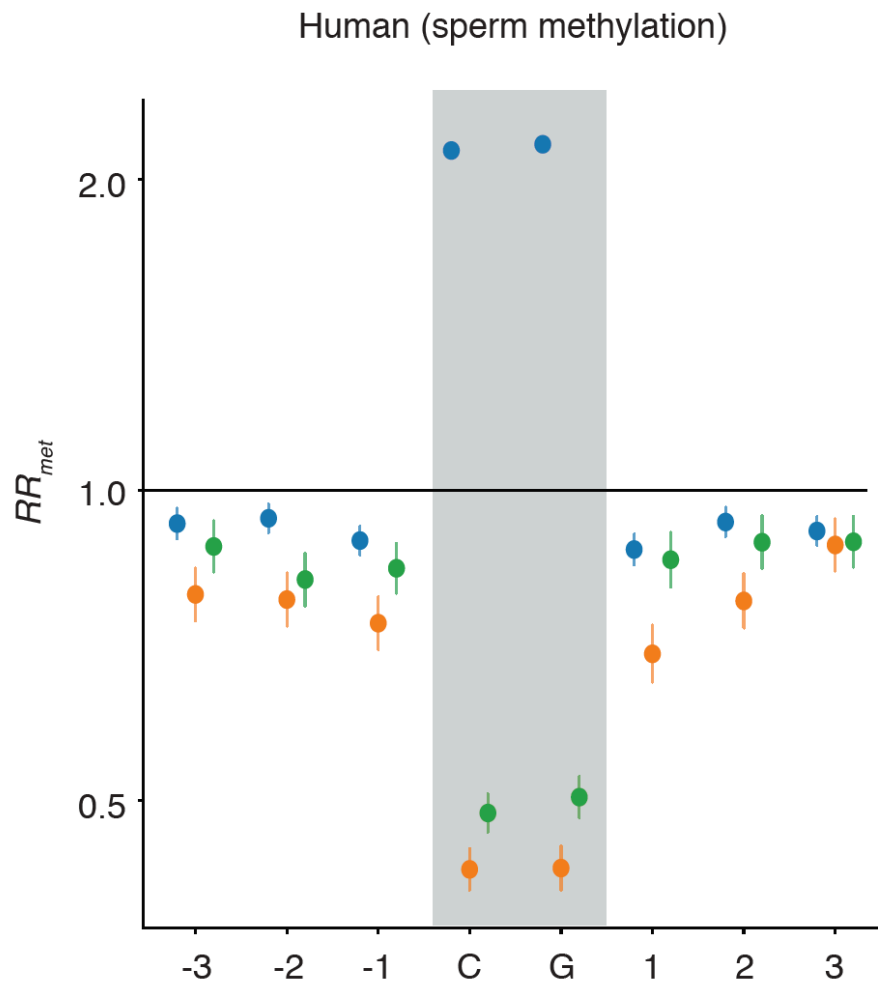


Figure 1.13: The relative mutational risk of methylation (RR_{met}) at CpGs and neighbouring nucleotides in human sperm cells. Vertical bars are confidence intervals on RR_{met} computed using the delta method.

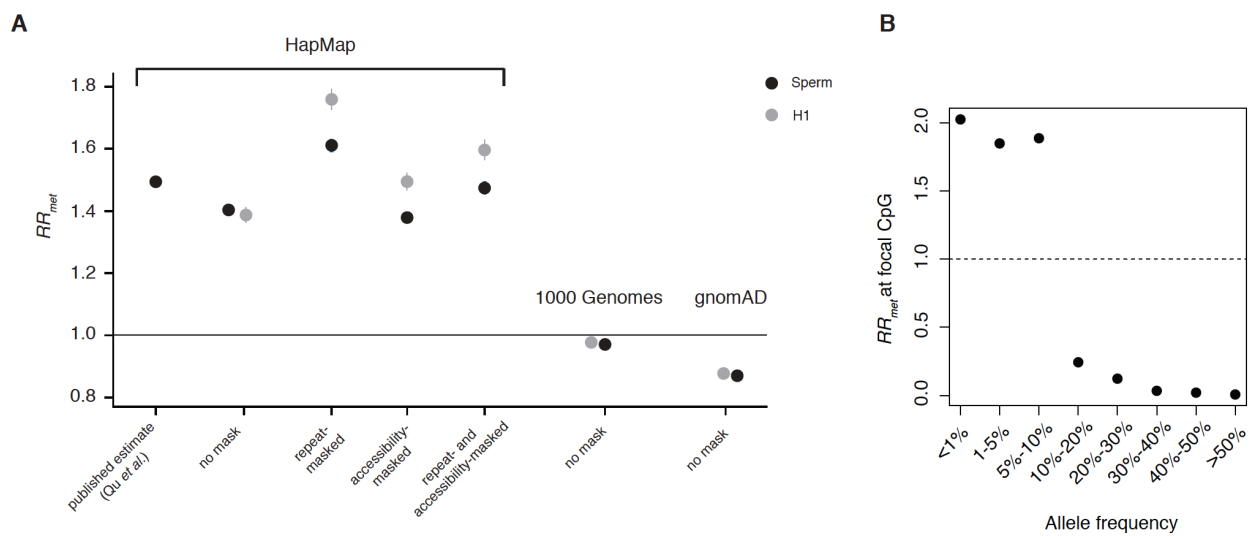


Figure 1.14: **(A)** Effects of dataset choice and filtering on the relative mutational risk of methylation (RR_{met}) at CpG-neighbouring nucleotides in human. RR_{met} estimates are for all SNPs, i.e. transitions and transversions combined. Vertical bars are confidence intervals on RR_{met} computed using the delta method. **(B)** at focal CpGs as a function of allele frequency.

1.3 Conclusions

In conclusion, during this study I demonstrated that cytosine methylation is associated with an effect on the mutability of bases adjacent to the cytosine itself. In humans the bases surrounding a methylated cytosine accumulate fewer mutations than the bases surrounding unmethylated cytosines. This is in contrast to previous work by Qu et al [Qu et al., 2012], but I narrow down the discrepancy to the use of common SNPs in their analysis which led to mischaracterisation of methylated sites as unmethylated ones. Extending the analysis to the plants *A. thaliana* and rice, I see the opposite effect with bases surrounding methylated cytosines accruing more mutations than those adjacent to unmethylated cytosines. This striking difference is not attributable to sequence context, chromatin state, or GC contents, and implies a species-specific difference which I suggest might be tied to disparate recruitment of downstream repair pathways. Interestingly, accessibility of site to enzymatic methylation or demethylation machinery in *A. thaliana* reduces the mutational risk of methylation.

From these observations, the patterns are broadly consistent with the action of LP-BER. LP-BER had been shown to have a 7-fold greater mutagenic effect in the neighbourhood surrounding U:G mismatches compared to the mutagenic effect surrounding T:G mismatches [Chen et al., 2014], in line with the pattern I see in human. LP-BER also has a similar repair tract length to the effect length I see in both plants and humans, limited to the neighbouring 3-10 bases at most. From this I hypothesise that the change in direction of relative mutational risk between methylated and unmethylated bases in humans and plants is due to biases in which downstream factors each glycosylase interacts with. This would lead to a preference for different lesion repair pathways with different mutational risks. From an evolutionary perspective, I propose the idea that plants have had less selective pressure to evolve more fidelity in the repair of methylated cytosines, with methylated cytosines being more specifically located to transposable elements where mutations in the neighbourhood of 5mC might even be beneficial, and gene bodies of constitutively expressed genes already being sufficiently protected by transcription-coupled repair. In contrast, in humans where methylation is more ubiquitous and used for repression of large segments of the genome where the underlying sequence is still

important (such as in X-chromosome inactivation), there would be greater selective pressure to develop more safeguarding surrounding 5mC epigenetic marks.

Further investigation is warranted to understand the complete mechanistic underpinnings of methylation-associated mutability and the downstream involvement of DNA repair pathways. Mutants affecting different sets of repair pathways, in particular LP-BER, in human, *A. thaliana*, and rice genomes would be incredibly useful to tie specific mutational signatures to specific differences in these pathways between plants and mammals. In addition, with increasing volumes of bisulphite-sequencing and population genome sequencing, hopefully other species and organisms can be analysed to see when these differing epigenetic strategies evolved, and would offer great insight into the evolution of both different repair mechanisms as well as the appearance of epigenetic strategies in genome regulation. Despite these unknowns, my work provides strong evidence that methylation casts a longer mutational shadow than initially thought, disparately shaping the emergence of variation across the genomes of different eukaryotes.

1.4 Methods

1.4.1 Analysis of relative mutational risk - pipeline

I pursued a matched pairs approach to generate two sets of methylated and unmethylated cytosines which have the same composition of sequences and chromatin context. To do so, I paired each unmethylated cytosine with the nearest methylated cytosine that a) had an identical sequence ± 4 bp from the focal CpG (or, in addition, CpHpG/CpHpH in plants), and b) had an identical chromatin state as defined by HMM calculation (see below). I defined methylated sites as those cytosines with a ≥ 0.7 ratio of methylated reads to unmethylated reads in the bisulphite-sequencing experiments (see below). Likewise, unmethylated sites were defined as those with a ratio of ≤ 0.2 . Sites with methylation stoichiometry between 0.2 and 0.7 were removed, but as methylation is strongly bimodal, this excludes relatively few sites.

Any sequences that included nearby CpGs were removed to avoid confounding the effect of the focal CpG with proximal CpGs, irrespective of whether their methylation state was known or not. Subsequently singleton SNPs from the respective population genetic databases were overlapped with the paired sites, ignoring the first and final bases of the paired sequence since the sequence context on the fringe cannot discount the presence of CpGs at the edges of our pairs, which would affect the incidence of SNPs at the edges of our pairs. Also, any transition that occurred at the T in a TpG or at the A in a CpA were also ignored. This is because of the possibility that the reference allele was misinferred and that the TpG>CpG mutation documented was actually the reverse. This would lead to mischaracterisation of a methylated site with a SNP to an unmethylated CpG due to the bisulphite sequencing ignoring the TpG site in the reference. Therefore, to avoid such bias these sites were simply removed.

1.4.2 Methylation basecalls

The methylation basecalls were obtained from already published bisulphite sequencing experiments: [Lister et al., 2009] for Human H1 embryonic stemcells, [Molaro et al., 2011] for human sperm cells, [Kawakatsu et al., 2017] for *A. thaliana* Ws0 accession wild type global seed stage methylation state, and [Stroud et al., 2013] for methylation calls of 3 week old leaf tissue in rice. Data accessions in the Data Availability section below. Only autosomal chromosomes were used to avoid the alternative selective regime that affects allosomes from skewing any effect seen. Ideally germline methylation state would have been used for rice as well, but the data was not available. Basecalls were only used if they were covered by at least 10 reads in each set. Methylated sites were defined as positions where the methylation stoichiometry was greater than 70%, and unmethylated sites as positions with a stoichiometry of less than 20%. Due to the bimodal distribution of methylation this excluded very few positions, but also prevented the accidental use of sites of unclear methylation status.

1.4.3 Population SNP Datasets

The species for analysis were chosen based on the quality and availability of population genetics SNP databases. The respective datasets used were gnomAD for human [Karczewski et al., 2016], 1001 genomes project for *A. thaliana* [Weigel and Mott, 2009], and the 3000 rice genomes project for *O. sativa* [Li et al., 2014]. From these sets, only singleton SNPs were used to better approximate mutation rates. This was to avoid selective pressures or neutral drift skewing the association of allele frequency to certain sequence, methylation, or chromatin features. For higher frequency variants, it is more likely that selective pressures or biased segregation (biased gene conversion) have spread them through the population and so no longer cleanly imply mutation rate. Singleton SNPs are those with allele counts of 1 in organisms that do not self-fertilize (human and rice), and 2 for those that do (*A. thaliana*). In addition, ideally the SNPs would be compared to the ancestral state to only include the derived singletons. This would improve the mutation rate approximation since the assumption that since the last common ancestor it is most likely that only a single mutation could have occurred in the space of that evolutionary time to form the singleton at said position. This was done for human, where the ancestral genome is well inferred [Paten et al., 2008]. This was not available for *A. thaliana* and rice, therefore singletons were selected as singletons from any sampled accession.

1.4.4 Chromatin state information

The chromatin states were obtained from published datasets for human [Ho et al., 2014] and plants [Liu et al., 2017] that utilised Hidden Markov Models to calculate the association of histone marks from ChIP-seq experiments to different regions in the genome.

1.4.5 Genome versions

The human genome version hg19 [Consortium et al., 2001], the arabidopsis TAIR10 [Lamesch et al., 2011], and rice IRGSP version 1.0 [Kawahara et al., 2013] were used. For any dataset not

in this version, the UCSC LiftOver utility was used to convert the genomic coordinates [Kuhn et al., 2012].

1.4.6 Genome masking

The human genome masks were obtained from the UCSC genome browser page for the RepeatMasker mask of the human hg19 genome version [Consortium et al., 2001], and the accessibility mask was obtained from the ftp page of the 1000 genomes project [Siva, 2008]. The rice IRGSP version 1 [Kawahara et al., 2013] was already soft-masked by RepeatMasker. For *A. thaliana*, and for completeness for the other species, I used RepeatMasker manually to generate another masked version.

1.4.7 Assignment of plant methylation machinery zones

The methylomes of several *A. thaliana* methylation and demethylation machinery mutants were available from studies by Wierzbicki et al., [Wierzbicki et al., 2012] and Zemach et al., [Zemach et al., 2013]. I used knock-out mutants of *ros1*, *nrpd1*, *nrpd1/ros1* double knockout, *ddm*, and *drd* mutants to define regions in the *A. thaliana* genome that were targeted by the respective enzymatic machinery. Sites were defined as follows:

- ROS1 target: methylation in the *ros1* mutant was greater than in the WT
- NRPD1 target: methylation in the *nrpd1/ros1* double knockout was lower than in the *ros1* deletion strain, to get both regular and cryptic RdDM targets
- DDM target: methylation in the *ddm* mutant was lower than in the WT, and the *drd* mutant had no effect

1.4.8 Replicating the Qu et al. protocol

The Qu et al. protocol was copied exactly from their published article [Qu et al., 2012] and details clarified via email correspondence. For each basecall in the respective methylation dataset, the bases up and downstream of the CpG were all collected into what the authors termed ‘CpG blocks’. If an unmethylated CpG was within 20bp of a methylated CpG, the bases inbetween were split equally between the corresponding unmethylated and methylated region. CpG blocks were considered with ≥ 5 read coverage. Methylated CpG blocks had methylation ≥ 0.8 and unmethylated CpG blocks ≤ 0.2 . SNPs of any allele frequency were counted in each region. Transitions occurring at CpG, TpG, and CpA dinucleotides were removed to avoid the confounding of methylated CpGs themselves skewing the perceived effect on the proximal variation surrounding the CpGs.

1.4.9 Data availability statement

I confirm that all data analysed in this study is publicly available and links to the original sources are provided throughout. The code used to carry out all analysis is available on my Github page: https://github.com/vfkusmartsev/Proximal_Mutations_5mC. The datasets used are as follows:

- Human polymorphism data:
 - HapMap <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>
 - 1000 Genomes project <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>
 - gnomAD <https://gnomad.broadinstitute.org/>
- *A. thaliana* polymorphism data <https://1001genomes.org.data/GMI-MPI/releases/v3.1>
- Rice polymorphism data <https://snp-seek.irri.org/>

- Human methylation data https://neomorph.salk.edu/human_methylome/data.html
 - H1 hESC https://neomorph.salk.edu/human_methylome/data.html
 - sperm cell (GEO GSE30340)
- *A. thaliana* methylation data:
 - wildtype Ws0 global seed stage (GEO GSM1664380)
 - *ros1* (GEO GSM1859475)
 - *nrpd1* (GEO GSM1859476)
 - *nrpd1/ros1* (GEO GSM1859478)
 - *ddm* (GEO GSM1014117)
 - *drd* (GEO GSM1014120)
- Rice methylation data:
 - 3 week old leaf tissue (GEO GSM1039487)

1.5 Publications

Kusmartsev V, et al. Cytosine Methylation Affect the Mutability of Neighboring Nucleotides in Germline and Soma. Genetics 2020; 214:809-823

Chapter 2

Avoidance selection of Gain of Function mutations

2.1 Introduction

Evolution of any organism is driven by the generation of variation via mutational processes, followed by either genetic drift due to random sampling or directed sampling due to selective pressures on this variation. This permits or prevents variants from accumulating in a population. In Chapter 1 I discussed biases on the generation of variation as a result of DNA methylation. Now I will move on to the next layer of evolution: analysing the selection upon stochastically generated variation. Specifically, I will focus on what sequence patterns are avoided and remain at low population frequency once they appear by chance.

2.1.1 Investigation of selection - conservation of genetic elements

There have been many studies investigating selective pressures acting on different genomes. Typically these studies have focused on conservation of sequence elements between species [Lindblad-Toh et al., 2011]. More conserved regions are described to be functionally important, since evolutionary pressures are purging any variation that happens to disrupt these functional

sites. The lack of variation in conserved sites is often not simply a result of reduced mutation rate, as seen by comparing the incidence of new derived alleles between conserved and non-conserved noncoding regions [Drake et al., 2006]. There are many methods that have used such conservation to predict the functional outcome of mutations. One such method is *Sorting Intolerant From Tolerant* (SIFT Ng and Henikoff [2001], Vaser et al. [2016]) which does so by searching and comparing a given protein sequence to closely related sequences that might share similar functions, and calculating the probability for a substitution. Another such method is PolyPhen [Adzhubei et al., 2010], which predicts the functional consequences of mutations by comparing sequence homology and structural features between the wild type and the mutant protein.

Conservation, however, only shows what regions of the genome have well established functions that have survived many generations of mutational bias and genetic drift. Conservation-based approaches fail to capture a potentially large segment of selection: the *avoidance* of functional motifs that create new or strengthen previous interactions. For example, an extreme scenario would be a previously non-functional sequence gaining a transcription factor binding site following a mutation. Mutations that cause such a functional motif to appear will be defined as *Gain of Function* (GoF) mutations. A GoF mutation differs from a *Loss of Function* (LoF) mutation in that a LoF mutation disrupts an existing conserved area of the genome that, in the context of our understanding, has a selected function. In a LoF paradigm, the conserved sequence is the most important and deviations are generally undesired. On the other hand, in a GoF paradigm, the ancestral sequence might be non-functional and selection principally acts on the new derived motif. Selection acting on GoF mutations will act not only in regions that are already functional and have been for a while (and therefore likely conserved across lineages), but in addition in other regions that have no currently apparent functions. Detecting selection against GoF mutations using conservation-based approaches might be difficult. Starting from a conserved functional site any or most mutations are detrimental, whereas starting from a random non-functional sequence there might only be a single mutation that yields a particular GoF sequence. Consequently, the power to differentiate selection against this mutation from variation in the mutation rate is likely low. How then can one study what GoF motifs are

avoided?

2.1.2 Compositional approaches

There have been previous methods pioneered by Samuel Karlin [Brendel et al., 1992, Burge et al., 1992] that interrogate the enrichment, and more importantly the depletion, of sequence motifs beyond conserved regions of the genome. These methods investigate the composition of the genome by comparing the actual number of occurrences of a motif to the expected number of said motif. These expectations are often calculated either from the product of submotifs [Burge et al., 1992] or from randomised shuffled sequences whilst keeping (di-)nucleotide frequencies constant [Kandel et al., 1996]. The current knowledge of the selection against the appearance of sequence patterns is largely limited to use of these global compositional methods. Avoidance is seen, for example, of Shine-Dalgarno sequences in bacterial coding sequences [Itzkovitz et al., 2010], transcription factor binding sites in 52 different Eubacteria and Archaea [Hahn et al., 2003], and of microRNA binding sites in mRNAs that are expressed at high levels in tissues at the same time as the corresponding microRNA [Farh et al., 2005, Marco, 2015]. It has also been shown that ncRNA sequences avoid having complementary sequences to common, highly-expressed mRNAs in bacterial and archaeal cells [Umu et al., 2016]. These RNA-level selective pressures have more recently been seen as both enrichments and depletions of sequences targeted by RNA-binding proteins, where sequences recognized by non-coding RNA-binding proteins are depleted in coding regions [Savisaar and Hurst, 2017].

Such compositional approaches have demonstrated that avoidance selection is prevalent. However, there are some major issues that limit their sensitivity and the inferences that can be drawn from them. Firstly, the overall depletion of particular patterns in a genome might be reflecting mutational biases as opposed to selection. For example, methylated cytosines are hypermutable in the context of CpGs by being liable to deaminate to a thymine [Cooper and Youssoufian, 1988]. This leads to a depletion of CpG dinucleotides in genomes where methylation in this context is common. Not only are methylated cytosines hypermutable, but they also lead to a broader, species-specific, mutational shadow beyond the base itself, as discussed

in Chapter 1. These mutational biases can occur at all scales of the genome, from individual, adjacent nucleotides, up to spanning millions of nucleotides and between whole chromosomes [Hodgkinson and Eyre-Walker, 2011]. Thus, the depletion or enrichment of motifs could in many instances be due to such mutational heterogeneity as opposed to any functional avoidance. Another inherent problem with looking at global genomic composition is that any pattern seen is due to a cumulative effect of mutation, random drift, gene conversion, and selection over a long period of evolutionary time. These patterns might include archaic remnants of selective pressures and evolutionary phenomena from ages past, in addition to ongoing avoidance selection. The selection acting on GoF mutations will often be weak, with the exception of obviously strong interactions such as nonsense mutations and the gain of RNA binding sites, as are captured by compositional approaches. Therefore, the set of compositional methods which only compare the composition within genomes are inadequate for progressing our understanding of avoidance selection beyond highly penetrant deleterious effects.

2.1.3 Novel approach - Derived Motif Frequency Spectra

I set out to develop new measures of ongoing avoidance selection. In particular, I have developed a new *Derived Motif Frequency Spectrum* (DMF) method. This method, the logic of which I will describe in detail below, will utilise the growing amount of variation data in populations to better investigate how derived mutations propagate or are prevented from propagating through the population. *A priori* the propagation of newly gained motifs is an excellent proxy of selective pressures and should allow statistically strong analysis of avoidance selection. I will focus only on sites that contain a derived mutation. Given that a site already has a mutation, we should - under an infinite sites model - be able to effectively account for differences in mutation rate. Here, I develop and test the DMF method, focusing on coding regions and validating the effectiveness and power of the method for single amino acids.

2.2 Results

Firstly, I must explain what I mean by and how to generate a derived motif frequency spectrum for a particular sequence motif. For a given motif, for example ‘ATA’, one searches in the population dataset, in this case gnomAD [Karczewski and Francioli, 2017], for every SNP that results in the sequence ‘ATA’. These alleles are then compared to the ancestral allele. The ancestral sequence in this case is the one used in the 1000 Genomes Project (1kG, [Siva, 2008]), extracted from the database ensembl compara 59. It was generated using the EPO (Enredo-Pecan-Ortheus) pipeline comparing 6 other primates to the human genomic sequence. Derived ‘ATA’ motifs can include, for example, ancestral ‘CTA’ sequences becoming ‘ATA’ or ancestral ‘ATG’ sequences becoming ‘ATA’. In this way, there are nine different single point mutations possible to create the 3-letter derived motif ‘ATA’. Each SNP that generates this particular derived motif has a specific allele frequency with which it occurs in the population of human genomes sampled. Here, in gnomAD the sample contains 125,748 human exomes, therefore the allele frequency varies from $\frac{1}{125748*2}$ since humans are diploid, to 1 where the derived allele is fixed compared to the ancestral sequence. These allele frequencies for a particular derived motif can be combined into a frequency distribution which I will term the *derived motif frequency spectrum* (DMF, Figure 2.1).

The goal of this work is to utilise these DMFs to infer the avoidance of newly occurring patterns and understand the selective pressures against the gain of these motifs. In order to infer selection on a given motif, I will compare these DMF distributions between one another. A schematic example is given in Figure 2.1 where I have depicted mock distributions for derived TCG and GTG motifs. We can see that the TCG DMF is shifted towards the left, towards lower allele frequencies. This shows that if a SNP results in a derived TCG motif, it remains more often at lower allele frequencies in the human population. The implication of this left shift is that there is greater selective pressure against the TCG motif, preventing it spreading to higher allele frequencies. However, these mock distributions are very far from the truth and the comparison is less obvious, as I will discuss in the next section.

One question that might arise is what is the expectation or null hypothesis to which I will

compare these DMF distributions. Alternative studies have generated this null hypothesis by firstly characterising the genome-wide mutation spectra, for example using non-negative matrix factorization on trinucleotide SNP context [Mathieson and Reich, 2017] or using the occurrence of rare SNPs in different length sequence motifs [Carlson et al., 2017]. I purposefully do not attempt to characterise the broad derived mutational spectrum of humans diverging from primates since this would also capture large scale selection and gene conversion events. Mutational spectra are also incredibly heterogeneous across the genome, as discussed in Chapter 1. In this study, I forgo comparing the DMF distributions to any previous expectation based on mutation rates or genome-wide composition to investigate selection in a novel way, focusing only on sites where a mutation has already occurred.

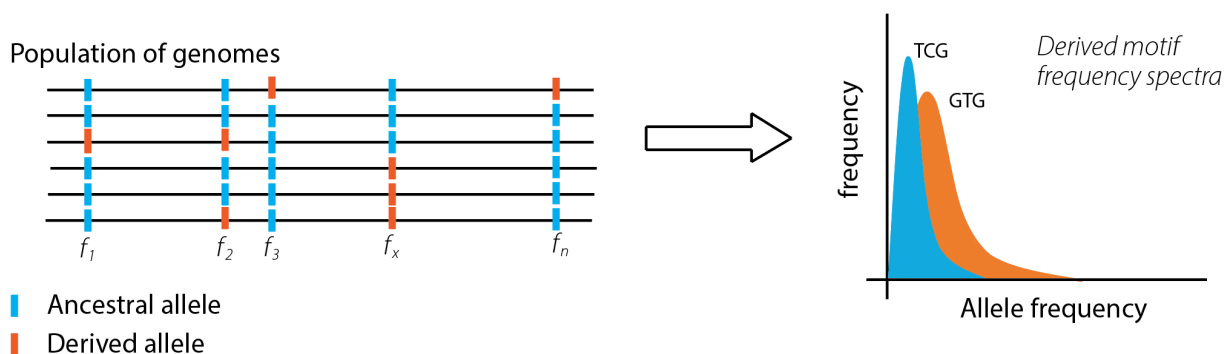


Figure 2.1: A schematic showing the general approach for generating a derived motif frequency spectrum for a particular nucleotide sequence pattern. For every position a SNP causes the ancestral allele to become the queried derived motif, the frequency of that derived SNP is collated into a derived motif frequency spectrum.

2.2.1 How to compare highly skewed distributions?

Before I can fully delve into investigating the avoidance of derived motifs, it must be considered how best to compare these frequency spectra. A first look into the distributions of allele frequencies in the gnomAD dataset shows an incredibly skewed distribution of allele frequencies, with around 65% of the SNPs occurring as singletons in the dataset and almost all of the alleles occur at less than 1% allele frequency (Figure 2.2A). In addition, there is also a second, much smaller peak seen at allele frequencies equal to 1, where the variant has progressed to fixation in the human genome (Figure 2.2B). Due to this bimodal and highly skewed distribution, the

commonly used distribution metrics of mode and median cannot be used since they would be equal to $\frac{1}{2N}$ - the equivalent to a singleton in the dataset. The arithmetic mean is also likely to be of limited use for our purposes because it is affected by high value outliers such as the fixed alleles in the population. It is therefore more likely to capture forces that drive new variants to fixation rather than those that prevent their spread through the population.

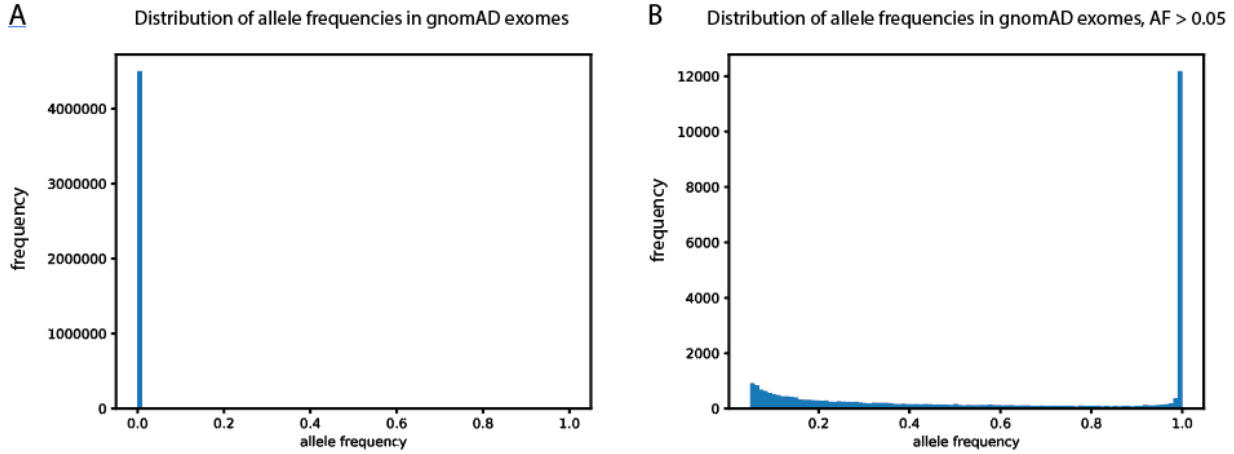


Figure 2.2: The distribution of allele frequencies in the exome dataset of gnomAD. A) Showing all allele frequencies, with 100 bins. B) The distribution of common alleles, with allele frequency greater than 5%, 100 bins.

A natural point of comparison is to extend from the arithmetic mean to investigate the use of the other Pythagorean means: the geometric and harmonic means. These means are defined as follows:

$$\mu_a = \frac{\sum x_i}{n}, \quad \mu_g = \sqrt[n]{\prod x_i}, \quad \mu_h = \frac{n}{\sum \frac{1}{x_i}}$$

where μ_a is the standard, arithmetic mean; μ_g is the geometric mean defined as the n th root of the product of a given set of observed values; and μ_h is the harmonic mean defined as the reciprocal of the arithmetic mean of the reciprocals of the values. Between these three, for a set of positive observed values, the harmonic mean is always the smallest, the arithmetic mean is always the largest, and the geometric mean is in the middle. In other words, the arithmetic mean is most greatly skewed by high value outliers and the harmonic mean is the least skewed by such values. Another way to protect the DMF metric from the skew of high

frequency outliers is to use the regular arithmetic mean but on alleles with some frequency threshold, such as 10%. In addition to these Pythagorean means, I will also compare them to a simple heuristic: the proportion of singletons (or doubletons) in a distribution. A singleton is defined as a SNP with allele count equal to 1 in our sample size of 251,596 exomic alleles and a doubleton is a SNP with allele count equal to 2. Proportion of singletons is a common metric used in evolutionary studies to infer information from rare variants [Tennessen et al., 2012, Nurminsky et al., 2001, Whittam et al., 1986, Wakeley et al., 2001, Verrelli et al., 2008]. As seen in Figure 2.3, the metrics that co-vary the most are the harmonic mean, geometric mean, low frequency arithmetic mean, and proportion of singletons. The proportion of doubletons struggles to distinguish synonymous from nonsynonymous variants at all. The arithmetic mean is greatly affected by outliers and therefore I do not consider it a useful metric for further analysis. Interestingly, there is a forked pattern when comparing the arithmetic mean to the harmonic and geometric means. Investigating the branch with higher geometric mean (and higher harmonic mean) more closely, the points with highest values are motifs containing CpAs, followed by four motifs containing ApT, and the rest in that branch ending on a T nucleotide. This will be investigated and a reason attributed thoroughly in a later section. This forked pattern is not seen when investigating the arithmetic mean on low frequency alleles, implying there is some distinction between the information captured between the geometric and harmonic means and the limited set for the arithmetic mean. Regardless, using a cut-off frequency might prove useful in later analysis. Specifically, I am interested in lower frequencies to observe the avoidance of derived motifs and the arithmetic mean greatly confounds our desired analysis unless a frequency cut-off is used, which artificially restricts the data we use. Between the harmonic and geometric means, the geometric mean has slightly greater variance and therefore might offer more resolution to visualise the avoidance of derived motifs which could be a weak signal in the dataset. Subsequently, I will utilise the proportion of singletons and the geometric means for further analysis and in cases where the error for the geometric point estimate is difficult to calculate the low frequency arithmetic mean is also a useful substitute.

Another method to capture more information than through a single point estimate (average or ratio) is to create a vector containing the entire distribution of frequencies for the motif.

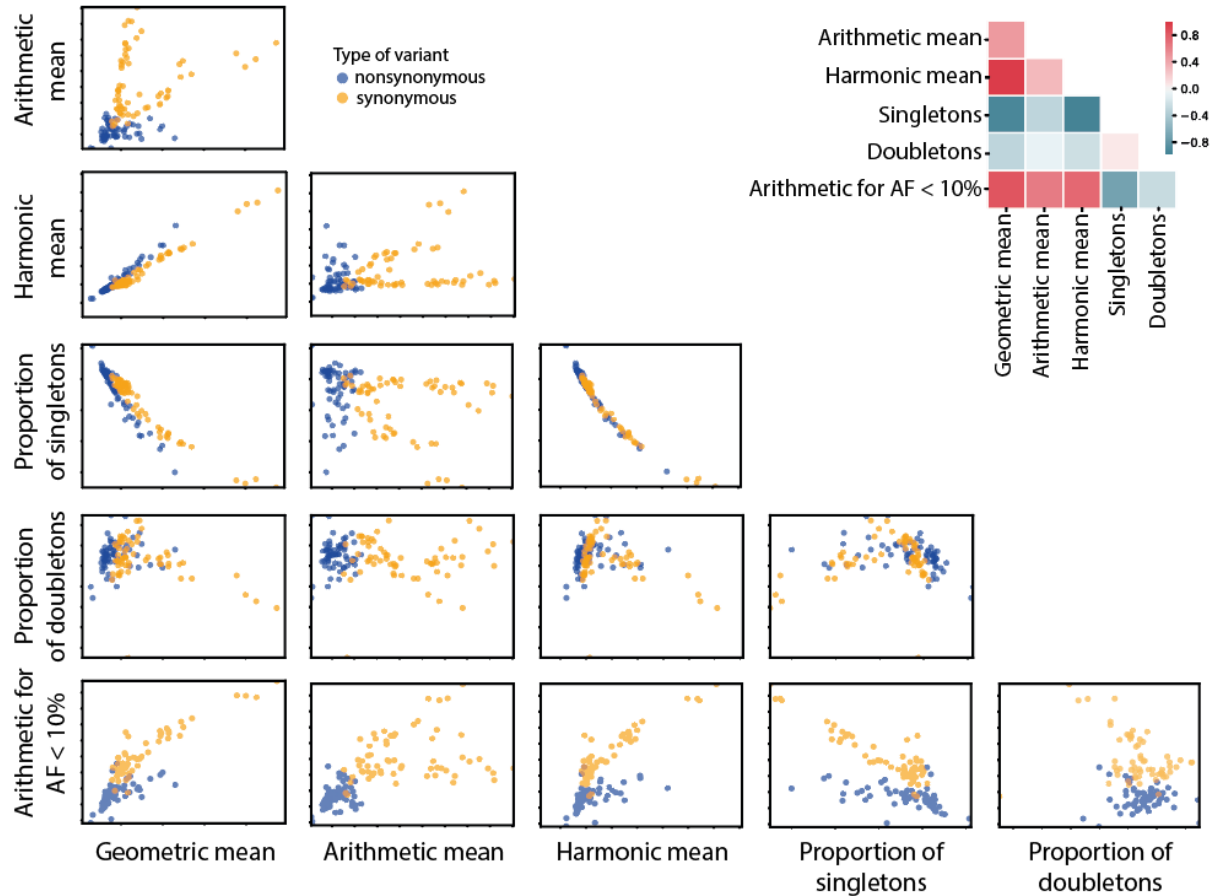


Figure 2.3: A diagram illustrating the covariance of the arithmetic, arithmetic only with SNPs with allele frequency less than 10%, geometric, and harmonic means with the proportion of singletons and doubletons when comparing the frequency spectra of derived single amino acid codons from the exome dataset of gnomAD.

One such technique is to make a cumulative density function (CDF) for each motif, where the vector would be of length M , where M is the maximum allele count in the dataset (251,596 in this case). Then for every allele count i possible ($1, 2, 3, \dots M$), the i -th vector element is the cumulative number of occurrences of that motif at that allele count in the dataset. To avoid skews due to various sample sizes of each derived motif, I normalise the vector so that the vector elements range from 0 to 1. One can then compare the distance between these vectors to show the distance between the avoidance selection acting on each motif. A schematic of what these look like is given in Figure 2.4A. However, this is not how the real DMFs look like as actual CDFs in our dataset are heavily skewed towards the lower allele counts, shown in Figure 2.4B. Due to the vast majority of the information being in the lower allele counts and with higher allele counts being incredibly sparse, by plotting CDFs normally as allele count versus cumulative frequency it becomes impossible to discern the distance between the lines by eye, as shown in Figure 2.4B.

Due to this heavy skew in allele frequencies, I decided to plot the CDF vectors as $\log_{10}(1 - f)$, where f is the cumulative density. This transformation spreads the vectors apart more at the low allele counts, aiding visualisation (Figure 2.5A). Here, we can zoom further into the grey dashed box in Figure 2.5A, obtaining Figure 2.5B. Each line represents the distribution of allele frequencies for a derived single amino acid codon, and the more negative the line, the rarer the derived allele is. These CDFs can then be used for further analysis such as applying clustering methods to their similarity matrices to get dendrograms like Figure 2.5C, illustrating which derived amino acids behave in similar ways. Not only are CDFs good for visualisation of similarities and comparisons in the data, but there are proper statistical tools for comparing cumulative distributions. Figure 2.5A shows that the method captures biologically meaningful information: derived stop codons (*) are greatly avoided and occur mostly as rare variants in the database. The next most avoided cluster of amino acids contain cysteine (C), phenylalanine (F), and tryptophan (W), all of which are used for very specific functions in proteins. For example cysteine being the only amino acid with the capacity for disulphide bonds, and tryptophan and phenylalanine containing aromatic rings that have a delocalised π -electron system, granting the residues important chemical and electronic properties. On the

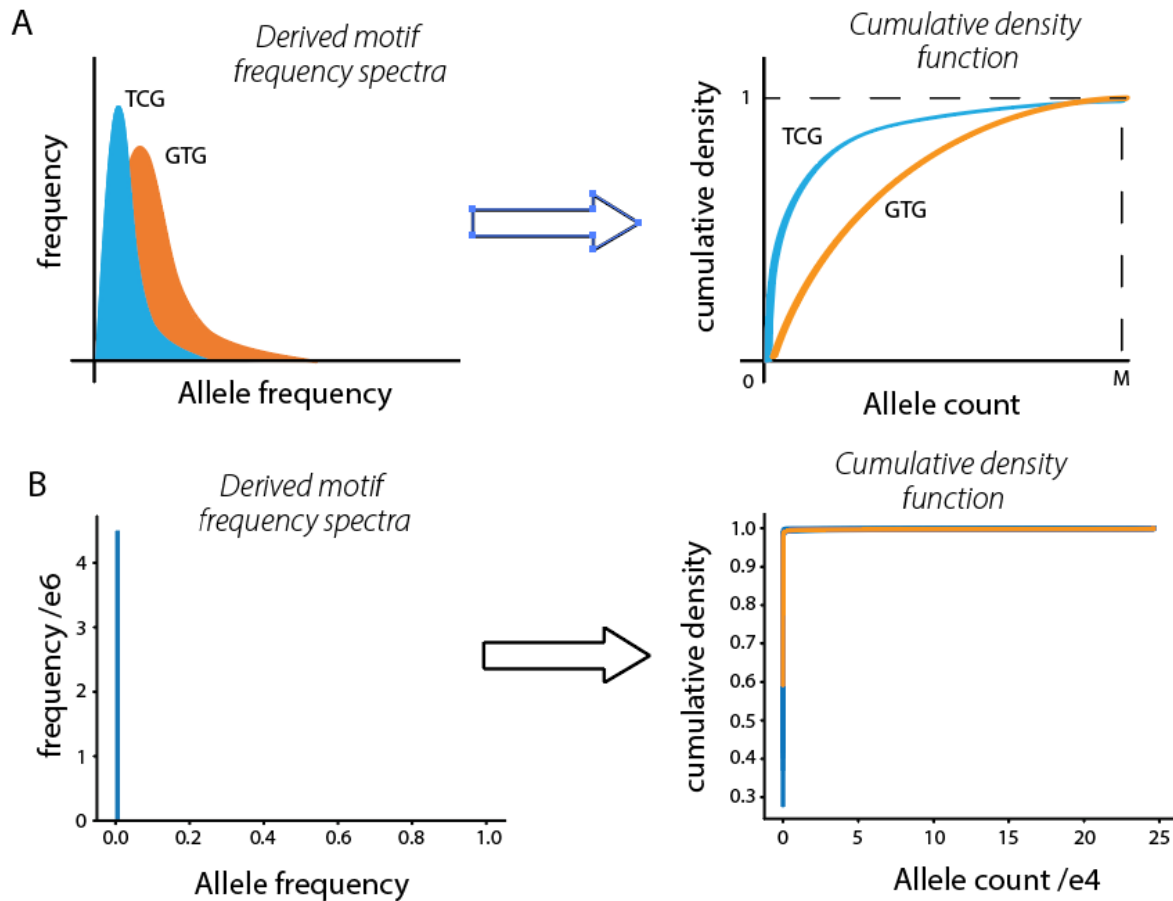


Figure 2.4: A schematic showing how a DMF spectrum can be converted into a cumulative density function (CDF), to allow easier comparison between different derived motifs. The CDFs are vectors of length M , where M is the maximum allele count in the dataset, in our case 251,596. The values of the vector vary from 0, up to 1, where one means all derived alleles occur below or equal to that allele count. The real DMFs and CDFs for derived single amino acids are shown in B, illustrating the severe skew of the dataset.

other hand, the hydrophobic residues alanine (A) and methionine (M) are the least avoided, suggesting that mutations towards these amino acids exert the least negative effect on average on the folding or functionality of proteins. Derived single amino acids and their properties will be discussed in greater detail in the next section.

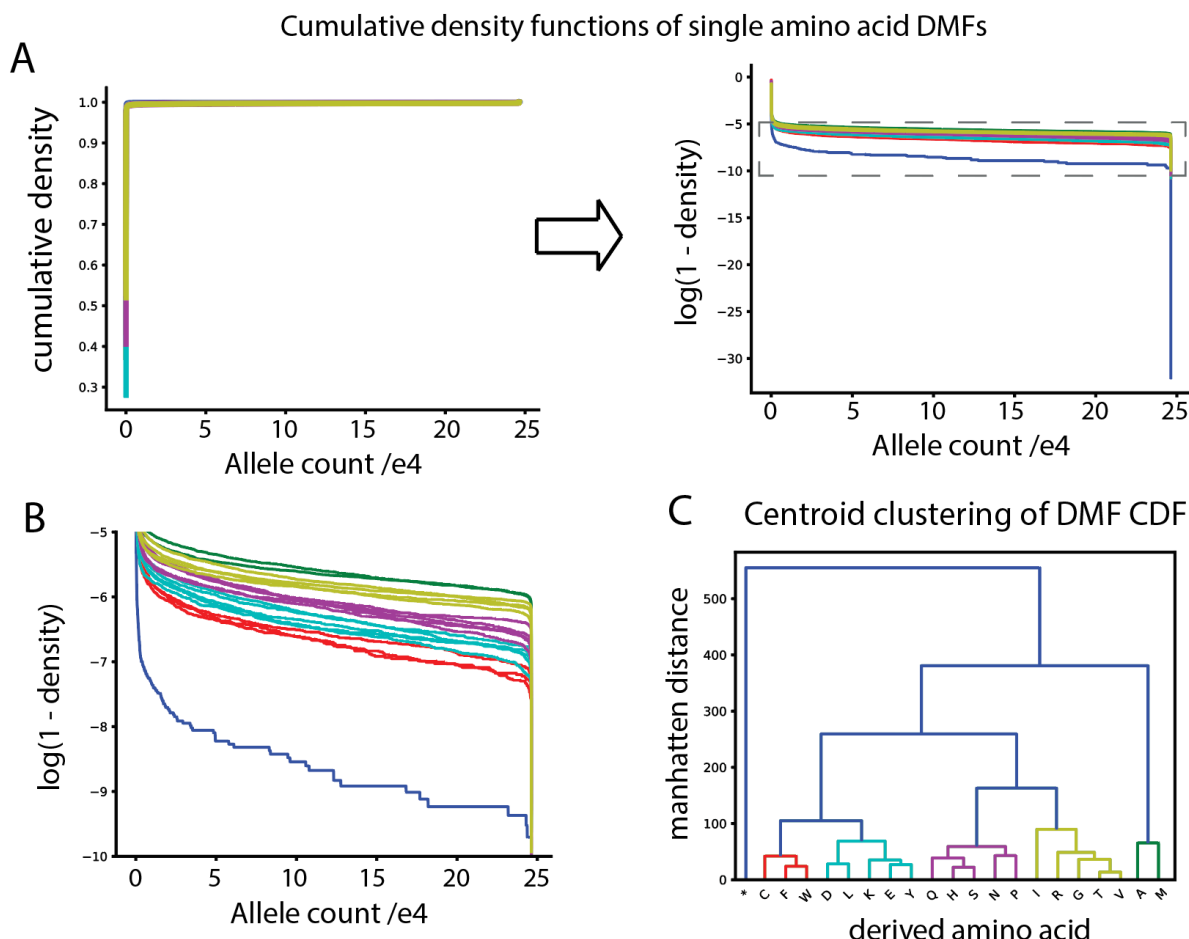


Figure 2.5: [A demonstration of the use of CDF for single amino acids. A) CDFs of single amino acid and stop codon (*) DMFs. By plotting the cumulative density as $\log_{10}(1 - f)$, we can better visualise the low frequency variants, where the majority of the data lies. B) Same as right-hand side of A, but zoomed in on the grey dashed box, to better visualise the separation in CDF between the derived amino acids and stop codons. The colours correspond to the clusters shown in C. C) A dendrogram showing the clusters of amino acids and stop codons based on their DMF CDFs. A centroid clustering method was used on the manhattan distance between the CDF vectors.

CDF vectors have length $2N$ (in our case 251,596). This is manageable for single amino acids or derived codons where there are 21 and 64 vectors respectively in total. However, as I extend the analysis to longer derived motifs, the number of CDFs increases exponentially and analysis quickly becomes computationally onerous. One way to circumvent this is to exploit the sparsity

of the vector at higher allele frequencies and shorten the vector by binning. Alternatively, I may select components where the most information is contained, for example the first 20 allele frequencies, or the first 10 and last 10 due to the bimodality of avoided and fixed variants. This will be explored later once we get to extending the motif length for analysis in section 2.2.6.

2.2.2 First step: single amino acid motifs

In order to both validate the method and identify confounding effects, the first point of analysis will focus on derived single amino acids in the exome portion of the human gnomAD dataset. I will use the exome-only dataset due to the much higher sample depth with 125,748 human samples being incorporated versus the 15,708 whole genomes gnomAD v2 contains. Since conception of the project, gnomAD v3 has been released, including 71,702 whole genomes but no exomes. Therefore, I still have greater sample depth in the exome set of v2 compared to the whole genomes of v3 and will leave analysis of the newer v3 version for future study where the focus might extend to whole genomes and non-coding regions.

For all analyses below, I used the longest available transcript from BioMart [Durinck et al., 2005] for each protein-coding region (see Methods). I ensured correct phase of the SNP in relation to the coding phase to extract the correct derived amino acid as it would occur in that coding context. I separated SNPs based on whether they cause nonsynonymous or synonymous changes to the amino acid sequence. This is due to both sets being under different selective regimes. Both synonymous and nonsynonymous sites otherwise capture effects occurring at the RNA or DNA level, relating to pre-transcriptional events such as transcription factor binding, regulatory events, or epigenetic marks; or post-transcriptional events such as RNA-processing, RNA-folding, and the effect of miRNA or ncRNA target sites. In addition to these effects, nonsynonymous changes capture protein-level selective pressures occurring on the new changed, derived phenotype of the protein due to the altered amino acid residue. Therefore, changes at synonymous sites can be used to focus in on protein-level effects. Initially, I will ignore the exact codon that leads to the derived amino acid in cases where there are multiple codons contributing to a derived amino acid.

Plotting the nonsynonymous geometric means of DMFs versus the synonymous means, most nonsynonymous changes occur with lower frequency than synonymous changes (Figure 2.6A). This is expected as nonsynonymous changes are selected against on all three levels: protein-level, RNA-level, and DNA-level, whereas synonymous changes are largely not selected at the protein-level (synonymous changes only affect the phenotype of the derived protein when considering its folding kinetics, which is influenced by codon usage being linked to elongation rate [Walsh et al., 2020, Sauna et al., 2007, Buhr et al., 2016]). Derived stop codons (*) are clearly avoided, which is a useful sanity check. Likewise, codons leading to large aromatics such as phenylalanine (F) and tyrosine (Y) or charged residues such as aspartic acid (D) and asparagine (N) seem to be most avoided. Oddly, glutamine (Q) has a greater DMF in the nonsynonymous context compared to the synonymous. To investigate whether a specific derived codon was driving this pattern, I plotted the individual 3-nt derived codons in both contexts against each other (Figure 2.6B). It appears that both derived CAG and CAA codons occur above the diagonal in the scatterplot. In addition, when considering whether synonymous directionality might skew the synonymous values to the lower end, I see no particular bias with CAG>CAA nonsynonymous to synonymous ratio being 1.420 and CAA>CAG having a ratio of 1.170, both greater than 1.0 indicating a higher DMF in nonsynonymous contexts. Beyond these derived Q codons, the outliers for the larger nonsynonymous geometric means either contain derived TpGs or CpAs. Likewise for the greater synonymous geometric means the greatest outliers contain a derived CpA as the 2nd and 3rd nucleotides. This pattern of CpAs and TpGs being enriched at higher DMFs might imply some mutational bias, as it is reminiscent of methylated CpG deamination to TpG (or CpA on the opposite strand) as discussed in Chapter 1. The assumption that mutational biases are removed using this DMF approach would only hold true if a specific site is only mutated once in our sample and allele counts greater than 1 are identical by descent, in accordance with the *infinite sites model*. As we will see in the next section, the infinite sites assumption does not seem to hold for mutations at CpG sites, in line with prior work [Harpak et al., 2016]. Interestingly, the CDF clusters (Figure 2.5C) do not seem to match the pattern seen when investigating the geometric means (Figure 2.6). This implies that maybe some information about the derived CDF is being lost when only using a single point estimate.

Further, there is a larger spread in synonymous geometric means across amino acids. However, there are some other non-obvious trends in the data, that require extra validation before further assertion of hypotheses or conclusions.

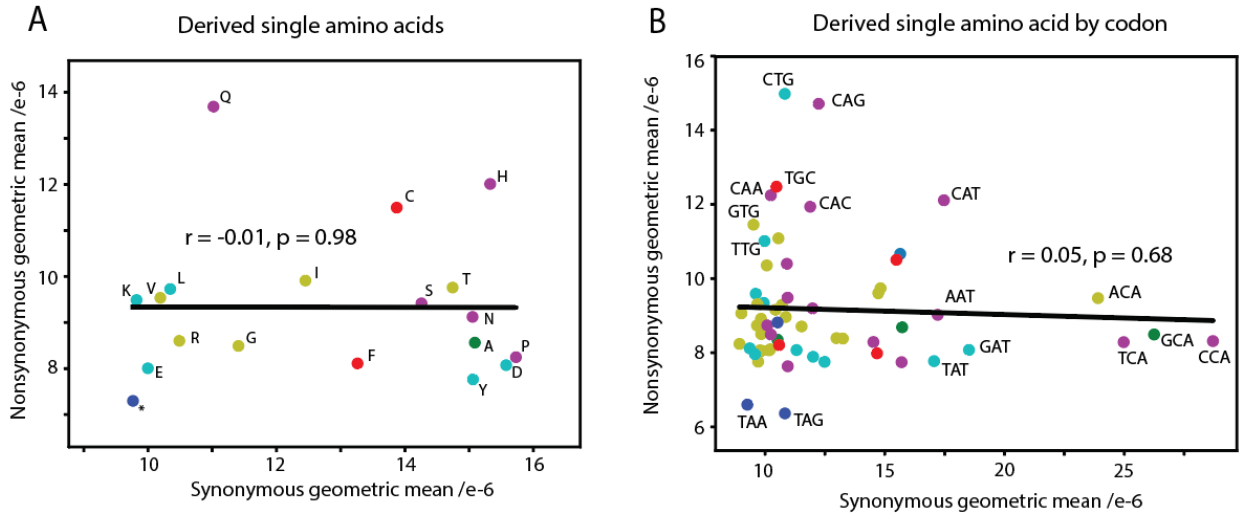


Figure 2.6: Scatterplots comparing the DMF geometric mean for derived single amino acids (A) and their triplet codons (B) in nonsynonymous (y axis) and synonymous (x axis) contexts. The derived amino acids and codons are labelled respectively. The colours correspond to the colours of the clusters in Figure 2.5C.

2.2.3 Ancestral independence of derived patterns

In using DMFs to measure avoidance selection, it is assumed that the DMF is only dependent on the new derived motif in a GoF paradigm, as opposed to an ancestral motif having greater mutational rate or selection acting on the ancestral pattern in a LoF paradigm. In order to investigate if particular ancestral-derived pairs of codons are driving any patterns seen in the DMFs, I separated each derived codon by the 9 ancestral codons from which it might have arisen. I restricted the analysis to only nonsynonymous SNPs to avoid different selective regimes affecting the trends. In addition, synonymous changes have limited ancestral alleles and so would limit the analysis of general ancestrally-induced biases. As illustrated in Figure 2.7, many derived amino acids have a single ancestral codon from which the derived SNP has much higher average frequency than the other ancestral-derived pairs. When looking at the greatest outliers with geometric mean greater than 0.000015 (labelled in Figure 2.7), the vast majority

are CpG>TpG (or the reverse strand equivalent CpG>CpA) transition mutations, with the exception of ACC>CCC and GTC>ATC. The latter is still a transition and might be due to a confounding CpG extending prior to the GTC>ATC mutation in sufficient genomic contexts to bias the DMF. For stop codons, ancestral independence generally holds true, demonstrating that it does not matter what ancestral sequence the derived stop comes from, the gain of a stop codon is avoided. There is also variation between amino acids. For example, the derived alanine codons are consistently different to derived cysteine codons irrespective of the ancestral codon they came from. However, these between-amino-acid differences are much more subtle for amino acids compared to the stop codon, and therefore it is important to get rid of ancestrally-dependent outliers.

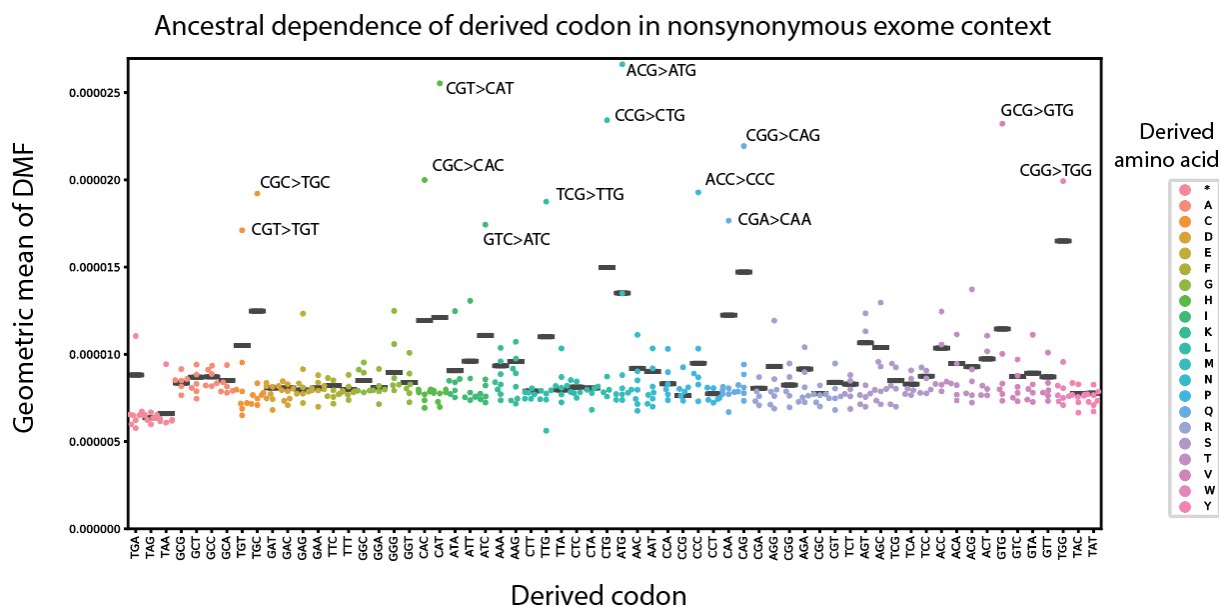


Figure 2.7: A swarmplot showing each derived amino acid codon split by its ancestral codon. All points are from the nonsynonymous derived context. The points are coloured by derived amino acid. The horizontal black lines indicate the overall geometric mean of the derived amino acid, ignoring the ancestral codon. The highest outliers are labelled.

Further evidence highlighting the skew that CpG deaminations have on the DMFs can be seen from principal component analysis. To better illustrate this, I extended the derived motif length to 2 amino acids to increase sample size. The coding phase was established in the same way as for single amino acids, using the longest available transcript from BioMart. Any derived SNP therefore is part of 2 derived motifs: the pattern upstream and the pattern downstream.

Using only nonsynonymous derived SNPs, for each ancestral-derived pair I generated an allele count vector similar to the CDFs described in section 2.1. To allow faster computational processing, I binned the allele counts into 1000 equally sized bins, which due to the large skew of the distributions resulted in only 77 unique, non-overlapping bins. This kept all the information of the most common allele counts (those relating to lower frequencies such as singletons, doubletons, etc.) whilst combining the sparser, higher frequencies together. I then performed singular value decomposition on these allele count vectors and plotted the top two principal components, which accounted for 95.5% and 3.9% of the total variance respectively. Colouring the data points by the presence of a CpG deamination in the derived motif versus not having one, there is a clear distinction in the PCA (Figure 2.8A). CpG deaminations were defined by looking at the genomic context of the ancestral motif with an additional nucleotide either side to account for edge-CpG cases. If in the ancestral genomic sequence there was a CpG with a SNP that resulted in a TpG or a CpA, it was labelled as a CpG deamination. Looking at the loadings of each allele count bin in the PCA (Figure 2.8B), we see that singletons have the greatest weight for the 1st principal component, whereas all the higher allele counts compose the 2nd principal component and are greatly enriched for CpG deaminations.

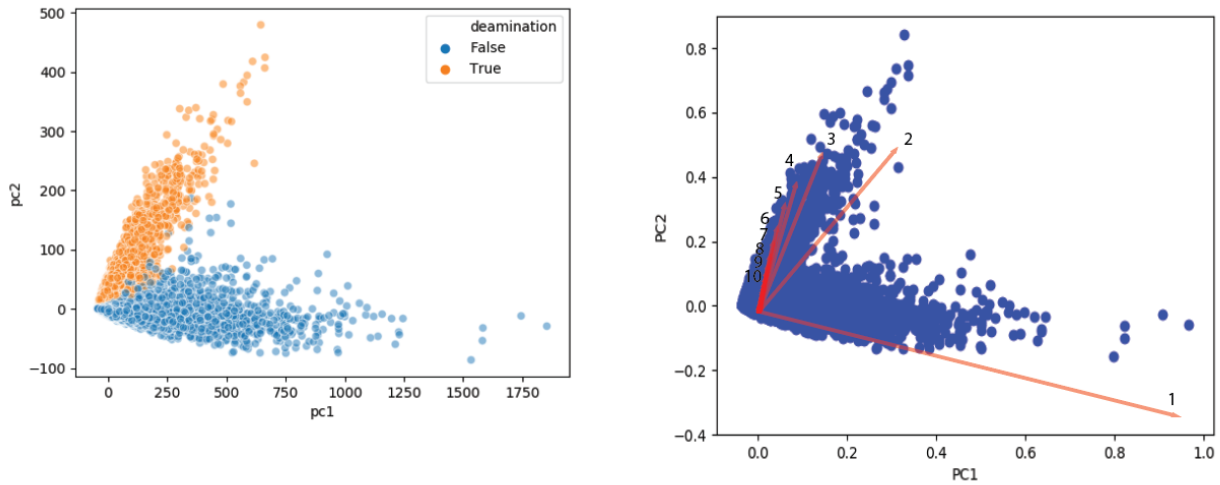


Figure 2.8: PCA of dipeptide ancestral-derived pairs. A) A scatterplot showing the top 2 principal components of eigenvector analysis on the derived 2 amino acid motif set. B) Same as A, except with the loadings of the top 10 allele counts added.

From these results, it becomes apparent that it is imperative to account for these CpG deaminations. In an ideal world, derived allele frequencies should not be affected by mutation rate.

However, in a sample of sufficient size this does not hold true for sites of extremely high mutation rate, such as methylated CpG deaminations. At such sites, mutation rate is high enough to generate independent (as in, not identical by descent) mutations at the same site across a population of individuals. This mutational recurrence was also noted by Lek et al. [2016], with them independently observing 43% of validated *de novo* events from other studies of parent-offspring trios. In addition, they found that 87% of previously reported *de novo* CpG transitions at synonymous sites were also observed in the exome aggregation consortium (ExAC) dataset, implying CpG mutational saturation is occurring in datasets of such size. The most conservative method would be to simply exclude them from further analysis. Removing CpG deamination SNPs and repeating the analysis from Figure 2.7, the ancestral-derived pairs become much closer in DMF values (Figure 2.9). In addition to hypermutable CpG sites, ApT sites have also been implicated to be hypermutable, being enriched for A>G and A>T substitutions [Aggarwala and Voight, 2016]. Looking at the outliers with highest geometric mean in Figure 2.9, there are a couple of outlier ApT ancestral mutations, with AAT>AGT, ATA>GTA, ATT>GTT, and ATT>ACT all showing the enrichment for A>G substitutions that Aggarwala and Voight noticed. However, these are not the most extreme outliers, with the other outliers seemingly returning ancestral sequence to a more repetitive sequence in the case of ACC>CCC, GTG>GGG, and GTG>GGG. However, as was the case for CpG deamination SNPs, one would also expect ApTs in other sequence contexts to be visibly higher, and that does not seem to be the case.

To explicitly investigate the effect of ApT-containing motifs I grouped SNPs in the derived single amino acid dataset by the 7nt ancestral genomic context surrounding each SNP and performed a PCA, colouring the derived SNPs by the presence of an ApT to GpT transition mutation (Figure 2.10A). ApT transition mutations seem to be shifted towards higher values in the 2nd principal component, which corresponds to only 0.4% of the variance compared to the 1st principal component explaining 99.4%. Investigating the loadings contributing to these 2 principal components, it is evident that all allele counts contribute towards the 1st principal component, but especially the lowest 10 allele count bins (Figure 2.10B). In contrast, the greatest contribution to the 2nd principal component were the highest allele count bins,

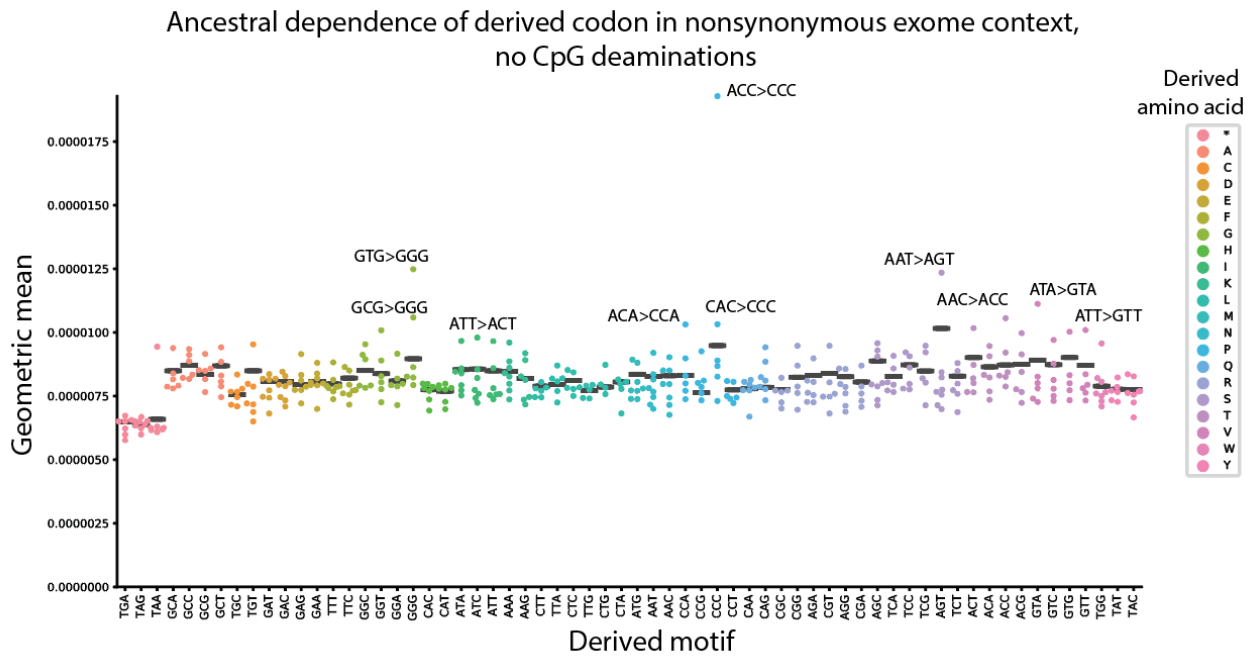


Figure 2.9: A swarmplot showing each derived amino acid codon split by its ancestral codon. All points are from the nonsynonymous derived context and all SNPs that might be explained by CpG deaminations were removed. The points are coloured by derived amino acid. The horizontal black lines indicate the overall geometric mean of the derived amino acid, ignoring the ancestral codon.

being bins 74 to 77 out of a total of 77 allele count bins. SNPs reaching these high allele counts are generally fixed in our human population compared to the ancestral genome. This suggests that mutational bias does not play a role here. In principle, selective sweeps or other segregating bias could have pushed these SNPs to fixation. However, we consider it more likely that these sites reflect errors of unknown origin in the generation of the ancestral genome.

One question is whether specific sequence contexts are contributing to this effect. To investigate this possibility, I selected all ApT transition SNPs which were two standard deviations above the mean value of the 2nd principal component. I then counted the occurrence of the 4 and 3 nucleotide submotifs, including the ApT. When looking at a single base up and downstream of the ApT (Figure 2.11A,B), it appears that the letter upstream of the ApT is most important with CAT motifs being greatly enriched (Figure 2.11A). In contrast, looking at the base downstream, the counts are evenly distributed (Figure 2.11B). Extending this to 4-nt submotifs, the greatest enrichment seen is for the previously reported mutation-promoting CAAT motif [Aggarwala and Voight, 2016, Medvedeva et al., 2013] and the CTAT motif. This is par-

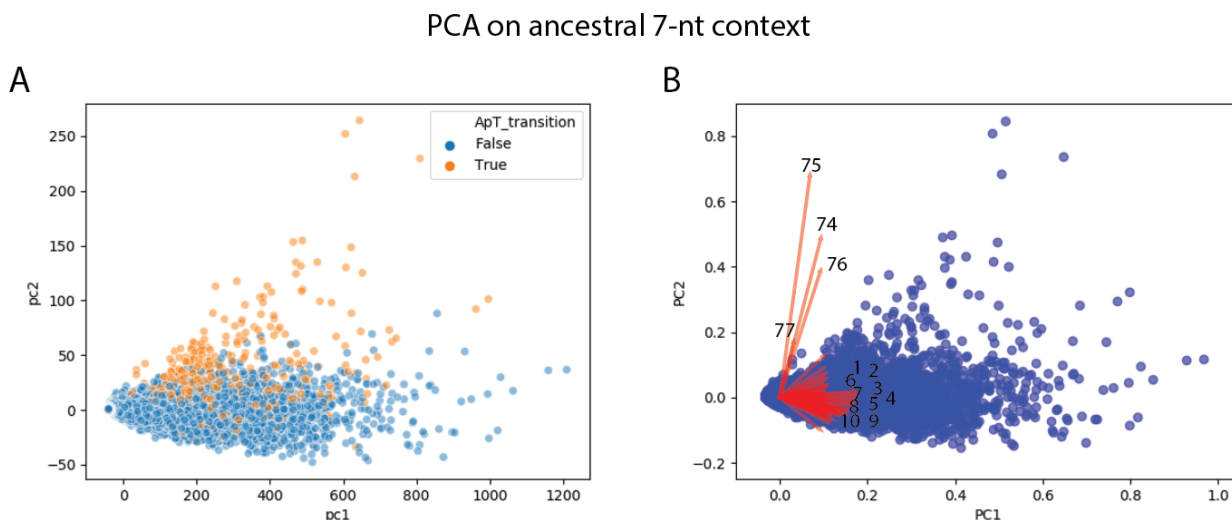


Figure 2.10: Scatterplots of the first 2 principal components of a PCA on SNPs mutating away from a 7-nt ancestral sequence. A) The points are stratified by the presence or lack of an ApT transition mutation. B) The similar to A, but with the loadings of each allele count bin drawn, with the highest and lowest allele count bins labelled.

ticularly interesting given there is no currently known mutagenic process that targets CAAT sequences and our earlier results imply that it is not a mutational bias. The next most enriched motifs are the four CAT containing motifs, matching what is seen in Figure 2.11A, and further illustrating that the base before the CAT is irrelevant. For the 2 bases downstream, again the counts are evenly distributed, solidifying that the downstream bases are less important for ApT transitions (Figure 2.11D). For completeness, looking at a single base both up and downstream of the central ApT it again is seen that CAT containing motifs are the most prevalent (Figure 2.11E), followed by TAT and AAT containing motifs that are likely enriched for the CAAT and CTAT motifs. Therefore, for further analysis it seems pertinent to mainly consider CAAT, CTAT, and CAT containing ancestral sequences for exclusion.

Another way to visualise the ancestral independence of derived motifs is to plot the average frequency for the loss of the ancestral motif versus the average frequency for the gain of the derived motif. This directly compares the GoF and LoF paradigms, and allows comparison of what selective or mutational pressure is driving the loss of the ancestral motif independent of the outcome, versus what is avoiding the new derived motif independent of the source. As shown in Figure 2.12, one can correctly capture that derived stop codons are avoided, but interestingly

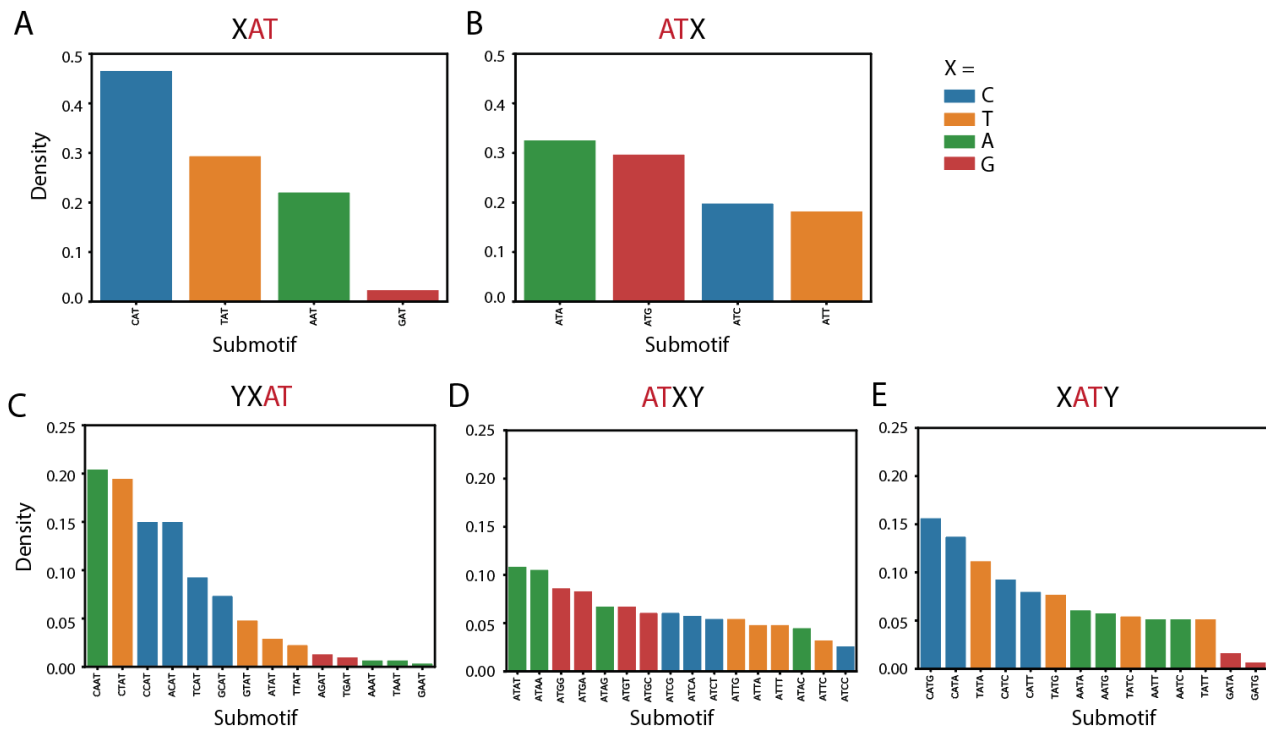


Figure 2.11: Barplots showing the distribution of ApT-transition containing motifs 2 standard deviations above the mean in the 2nd principal component from Figure 2.10B. A,B) These show the distribution of 3-nt submotifs, containing the base upstream (A) and downstream (B) of the mutating ApT. C,D,E) These show the distribution of 4-nt submotifs, split by the 2 bases upstream (C), 2 bases downstream (D), and bases either side (E) of the mutating ApT. The bars are coloured by the nearest base to the ApT as per subplots A and B, in the exception of E where the base upstream was prioritised for the bar colour.

loss of the stop codon has DMF values within the bulk cloud of frequencies for loss of motifs, implying that there is no selective or mutational pressure avoiding stop codons that already exist in the genome. Motifs shifted further to the right in Figure 2.12A (left in B) are more frequently lost in the ancestral compared to their gain in the derived context. Motifs shifted far to the right are therefore implicated in ancestral source effects, such as mutational biases, rather than selective pressures. When colouring the points by their content of hypermutable CpG and ApT dinucleotides (Figure 2.12) it is seen that CpG-containing motifs are generally in the same cloud as all the other points, which means that after my explicit removal of CpG deaminations, no more mutational biases are obvious in that context. Looking at the motifs containing ApT dinucleotides (green), a trend is seen of the loss of these motifs rising to higher allele frequency compared to their gain. This is even more clearly seen when looking at the proportion of singletons (Figure 2.12B). In addition to this shift in ApT-containing dinucleotides, within the set of ApT and CpG-containing motifs one may notice that the most right-shifted motifs in Figure 2.12A are either exclusively AT-rich or GC-rich. Combined with the couple of repetitive sequence outliers in Figure 2.9, this raises the possibility that there is a skew due to repetitive regions or regions of low sequence complexity. Repetitive regions could be a hotspot for errors in SNP calling, artificially increasing the frequency of SNPs and therefore the frequency of particular derived motifs.

To investigate this possibility, I calculated the linguistic sequence complexity of the motifs. This is defined as the product of the vocabulary usage of all potential words:

$$LSC = \prod_{i=1}^W U_i$$

where i is the length of the word, W is the maximum length of word, and U is the number of words of length i that could possibly occur in the sequence. For example, if all 4 nucleotides are used (in motifs at least 4 letters in length), $U_1 = 1.0$. Trifonov initially described the approach for use in DNA sequences [Trifonov, 1990] and the algorithm was optimised so that words longer than 3 need not be used for sequences shorter than 18 bases [Gabrielian and Bolshoy, 1999]. As the length of the word increases, the number of possible words becomes exponentially larger

Loss and Gain of single amino acid codons

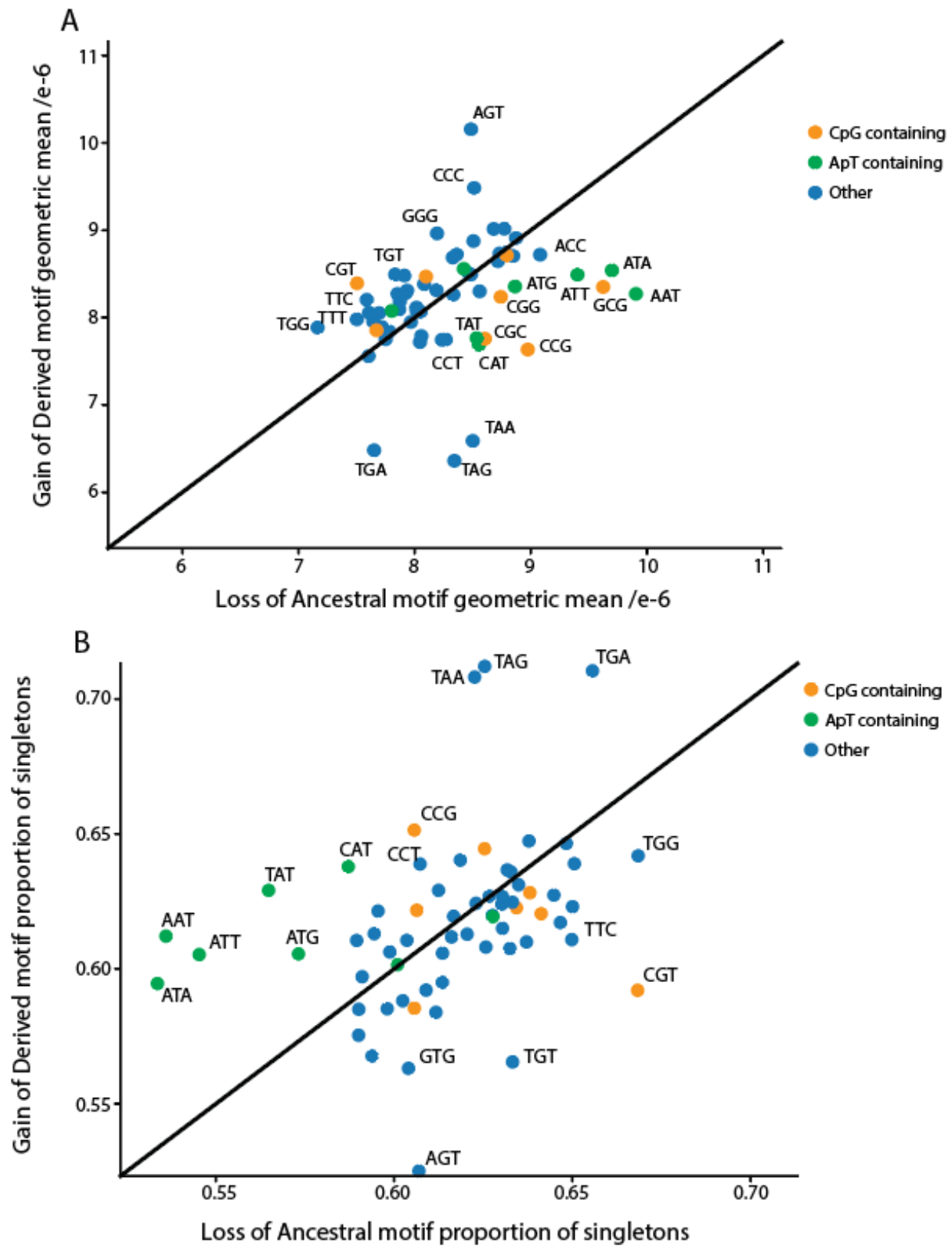


Figure 2.12: Scatterplots of the geometric means (A) and proportion of singletons (B) of the gain of the derived motif versus the loss of the ancestral. ApT motifs are shown in green and CpG containing motifs in orange. Motifs furthest from the diagonal are labelled. All SNPs that might be from CpG deaminations were removed from the analysis for this plot.

and therefore the usage becomes limited by the length of the sequence. For example, for $i = 2$, there are 16 dinucleotide words, but if the motif sequence is shorter than 16, it is only possible to have $[length(sequence) - (i - 1)]$ words in it. To demonstrate, for the motif *ATACGA*: $U_1 = 1.0$, $U_2 = 1.0$ since all 5 dinucleotides in the sequence are different, and $U_3 = 1.0$ since all 4 trinucleotides in the sequence are different; leading to $LSC_{ATACGA} = 1.0$. In contrast for motif *CGCGCG*: $U_1 = 0.5$, $U_2 = \frac{2}{5}$, $U_3 = \frac{2}{4}$; leading to $LSC_{CGCGCG} = 0.1$.

Working with derived nucleotide sequences for single or double amino acid motifs does not allow enough range of sequence complexity values, so I extended this analysis to motifs of 3 amino acids, or 9 nucleotides. There are significant correlations between both geometric means and proportion of singletons with the sequence complexity of the ancestral motif (Figure 2.13). This illustrates that the more simple and repetitive the sequence, the greater the chance of an apparent SNP occurring there at higher frequency. This could be due to some biological reason: Maybe repetitive sequences incur higher mutation rates - Aggarwala and Voight saw outliers in substitution rates in polyA and polyT sequences, which they ascribe to ‘slippage’ of protein machinery during DNA replication [Aggarwala and Voight, 2016]. Another biological reason might be that the collection of mutations in repetitive regions might be beneficial in disabling transposable elements, which are often repetitive in nature and this therefore invokes selection for increased mutation rates (or less selection for high-fidelity repair, as discussed in Chapter 1). However, the observation might perhaps more likely be due to technical difficulties in both sequencing and mapping repetitive regions in the human genome.

Another aspect of the sequence complexity is whether it is AT-rich sequences, GC-rich sequences, or simply the complexity itself driving the skew. If the skew is due to a sequence being entirely AT (or GC for that matter), one would expect a plot of DMF versus GC-content to be a progressively increasing trend. On the other hand, if the GC-content does not matter and it is simply a question of sequence complexity, there would be a U-shaped curve with peaks at the extreme values of GC-content. Plotting derived DMF against ancestral GC-content of the motifs in the same way I plotted the DMFs versus the ancestral sequence complexity, there appears to be a curve, implying sequence complexity itself is the driver, not simply GC-content (Figure 2.14). The patterns hold true both for nonsynonymous and synonymous contexts, as

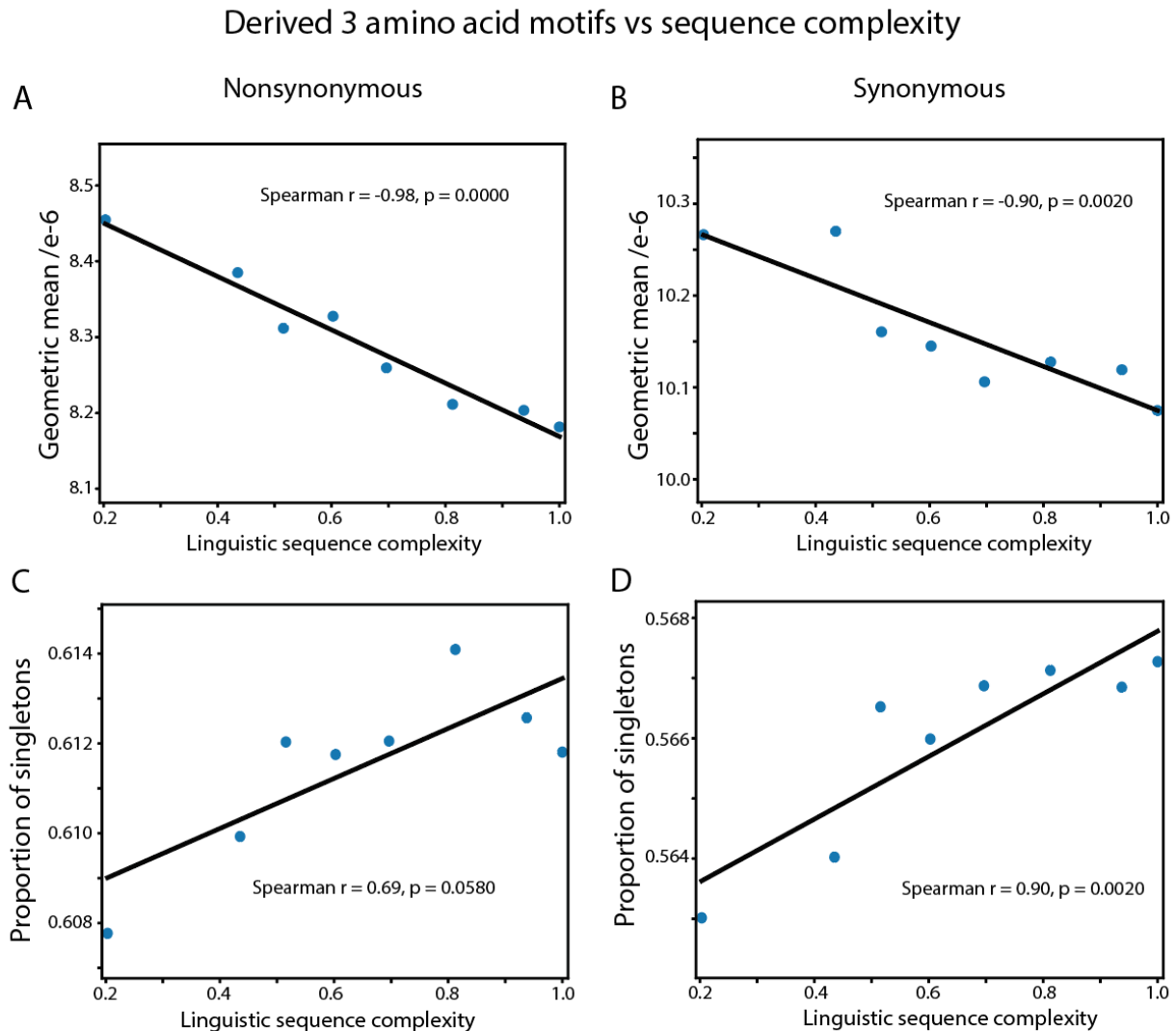


Figure 2.13: Scatterplots comparing the effect of ancestral sequence complexity on DMF. These compare the nonsynonymous (A,C) and synonymous (B,D) DMFs of 3 codon motifs by virtue of their geometric means (A,B) and proportion of singletons (C,D) against their ancestral sequence complexity. Motifs were binned by ancestral sequence complexity. Each subplot contains the spearman rank correlation and associated probability, as well as a line of best fit calculated using an ordinary least squares regression. No SNPs that were ancestral CpG>TpG were included.

well as using both DMF geometric mean and the proportion of singletons as the metric. However at the extreme GC-content values there is still a bias for higher DMF (and lower proportion of singletons) at high GC-content versus low GC-content. One likely candidate explanation is GC-biased gene conversion [Galtier et al., 2009]. There is much evidence suggesting that at recombination sites there is a weak bias towards fixing GC-rich instead of AT-rich alleles [Duret and Galtier, 2009]. This would lead to SNPs converting AT-containing sequences to GC-containing sequences more frequently than the reverse, increasing allele frequency of GC-containing derived motifs. This is especially true for repetitive regions where recombination is more likely to occur. The apparent curve might simply be due to larger error due to lower sample size at the extreme GC-content values. Due to the large skew, it was not possible to visualise the geometric error of the geometric point estimate. To address this more simply, I plotted boxplots for each GC-content bin for derived tripeptides in nonsynonymous and synonymous contexts (Figure 2.15). Due to the very heavy skew, I only selected derived motifs with frequency less than 1% and did not visualise the outliers beyond the tails of the boxplots, since if I did the boxes would not be visible. Doing so, there seems to be a slight curve in the inter-quartile ranges for nonsynonymous derived motifs, but no such trend in the synonymous case. This implies that this skew is only a strong factor for higher frequency outliers, which is the set of outliers I am less interested in since I want to investigate the avoidance of sequence patterns as the initial biological question.

In an attempt to elegantly control for this skew, I performed a PCA decomposition of the allele count vectors grouped by ancestral-derived pairs. For each ancestral sequence allele count vector, I computed the sequence complexity and compared the correlations between the pairs principal components versus its sequence complexity. As shown in Figure 2.16A, there is no obvious trend between the top two principal components and the ancestral sequence complexity. These two components explain 82.5% and 7.2% of the variance respectively. Principal component 1 is mostly made up of the lower allele frequency counts, with the greatest loadings being from the lowest 10 allele count bins; whereas principal component 2 is made up of the highest allele count loadings (Figure 2.16B). There are significant ($p < 0.001$) correlations between the ancestral sequence complexity and 25 of the 76 principal components, including the top 13

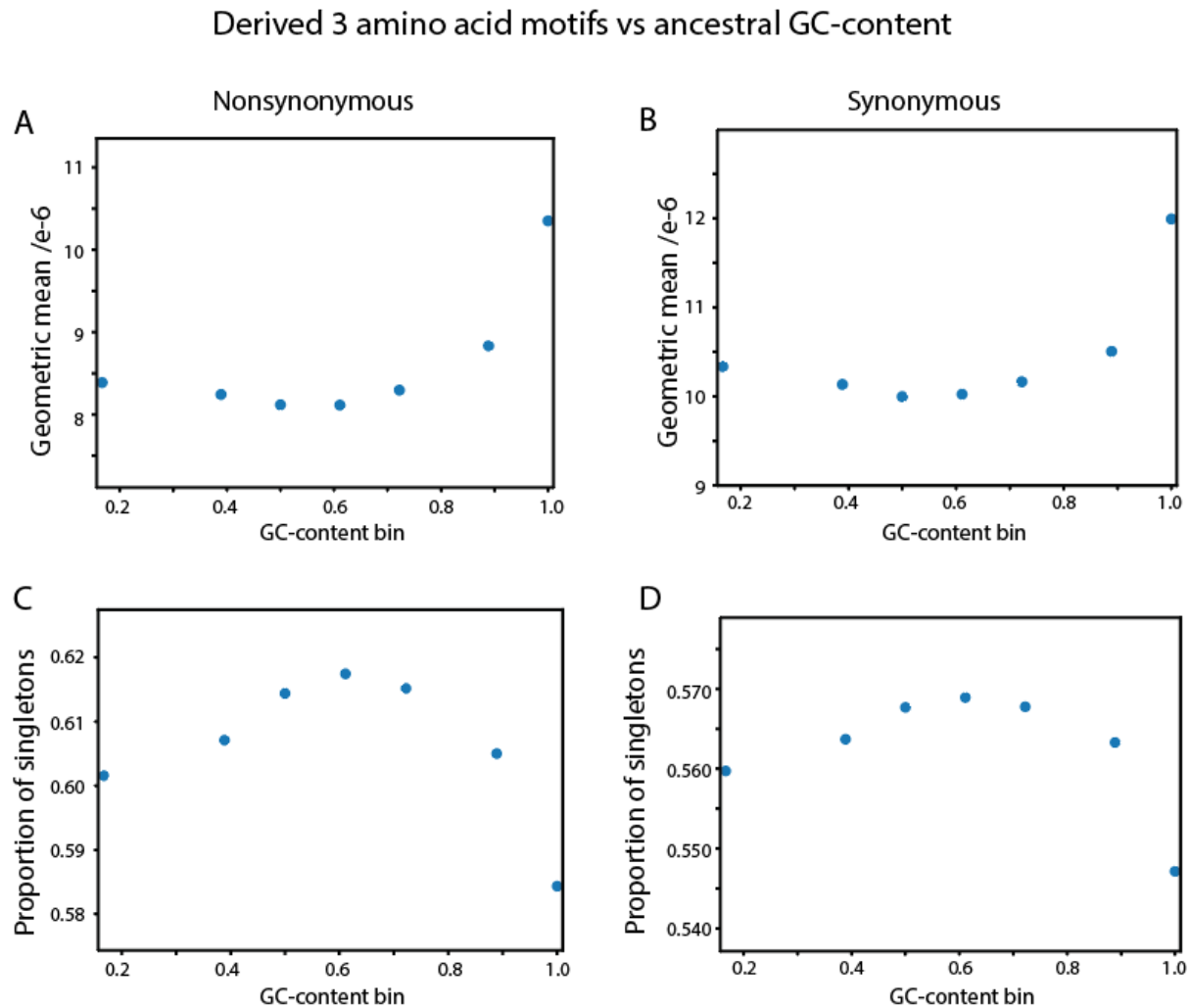


Figure 2.14: Scatterplots showing how DMF varies with ancestral motif GC-content. Left column (A,C) are nonsynonymous changes, and right column (B,D) are synonymous changes. First row (A,B) shows how the DMF geometric mean varies as a function of ancestral GC-content. Second row (C,D) shows how the proportion of singletons varies as a function of ancestral GC-content.

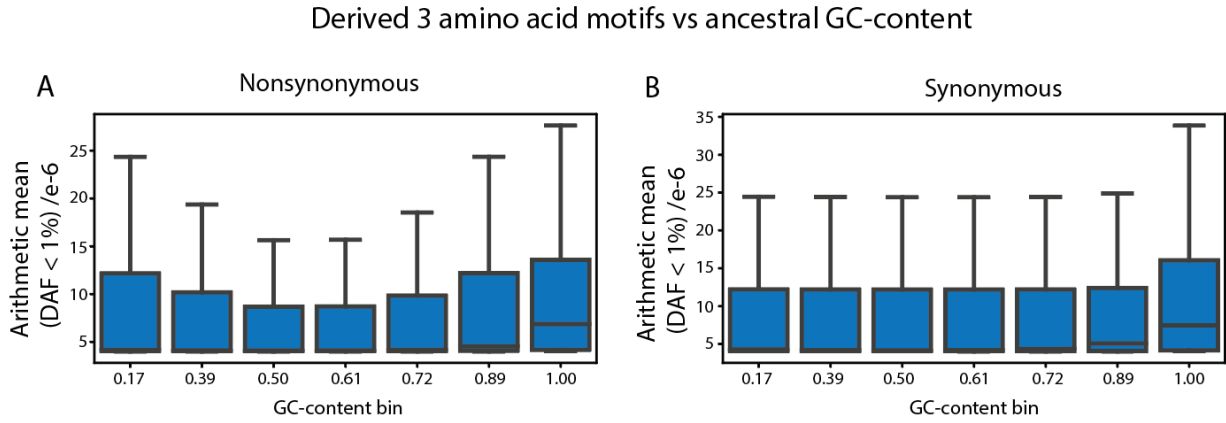


Figure 2.15: Boxplots showing how low frequency (less than 1% derived frequency) arithmetic means varies with ancestral motif GC-content. Left subplot (A) is for nonsynonymous changes, and right subplot (B) are synonymous changes. Outlier motifs are not shown.

components. However, the pearson r for all these correlations never exceeds $r = -0.072$ for the strongest correlation with principal component 1 ($p = 0.0$) which as mentioned before accounts for 7.2% of the entire variance in the CDFs of ancestral-derived pairs. In simpler terms, this effect is important to account for and a significant skewing factor, but likely more important at higher motif lengths and in non-coding sequence and the correlations are not strong enough to be elegantly removable via a PCA. Given that the skewing effects of repetitive sequence is relatively weak, I decided not to remove these contexts but will have to bear them in mind when interpreting altered DMFs.

2.2.4 Return to single amino acids

Following the above systematic search for potential mutational or technical confounders, it is pertinent to reanalyse the derived single amino acids with and without the major skewing factor of CpG deaminations, as well as verifying whether ApT transitions in the ancestral contexts of CAAT, CTAT, and CAT are a significant confounder for inferring selection from DMFs. Therefore, for all SNPs in the dataset, the ancestral genomic context was used to firstly remove all CpG to TpG SNPs on both the Watson and Crick strands as well as categorise SNPs into either being ApT to GpT transitions in the CAAT, CTAT, and CAT contexts, or not. Since there was no elegant way to control for sequence complexity according to the previous section,

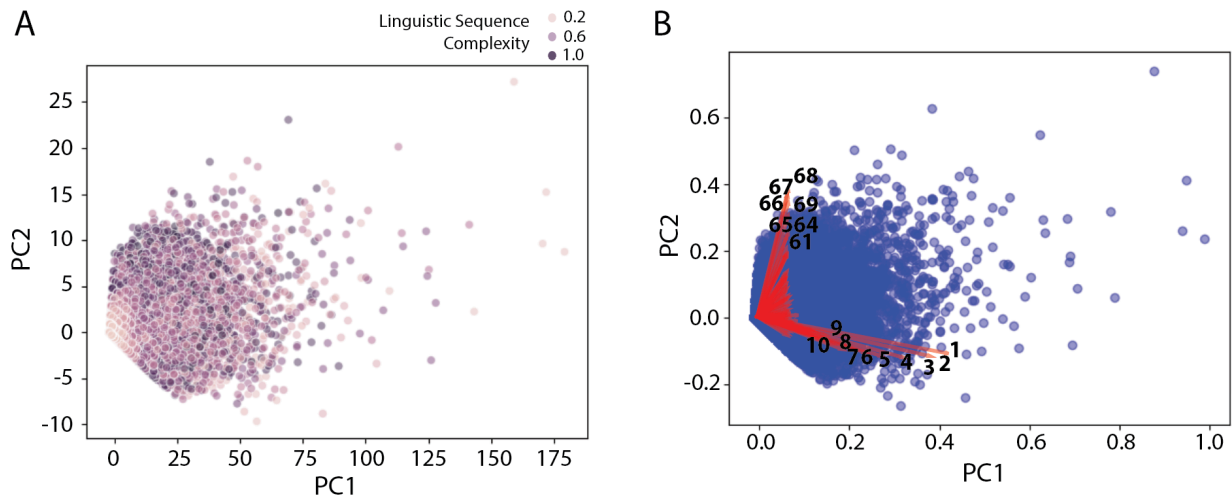


Figure 2.16: The top 2 components of PCA decomposition on the allele count vectors grouped by ancestral-derived pairs. A) The points are coloured by the sequence complexity of the ancestral motif. B) Similar to A but with principal components normalised, and the loading of each allele count bin shown, with the largest contributors labelled. No SNPs that were ancestral CpG>TpG were included.

nothing beyond the already applied RepeatMasker was done for this (see Methods). Firstly, comparing the DMFs for all SNPs with and excluding CpG to TpG transitions, there are no longer any motifs with higher nonsynonymous DMF than synonymous DMF and the DMF point estimates become clustered much more closely (Figures 2.17A,B 2.19A,B). The derived stop is now much more clearly separated from the general cloud of derived amino acids as expected from the nonsense outcome of a derived premature stop codon. Removing ApT transitions, the scatterplot of nonsynonymous DMF compared to synonymous DMF remains largely very similar (Figures 2.17B, 2.19B). Looking instead at the underlying derived codons for each amino acid, we see a larger amount of variance in both synonymous and nonsynonymous geometric means and proportions of singletons when including all SNPs (Figures 2.18A, 2.20A) with a clear outlier cluster of CpA motifs. These return back to the general cloud of points once CpG transitions are removed (Figures 2.18B, 2.20B), instead leaving higher frequency GpT and ApC motif outliers in the nonsynonymous and synonymous directions. These also return to the main cloud of derived codon point estimates upon removal of ApT to GpT transitions in the CAAT, CTAT and CAT context (Figures 2.18C, 2.20C). This implies that the reason for these points being outliers was not some selective pressure affecting the occurrence of the codon itself, but

them being the result of some SNP-promoting effects, with ancestral ApTs being characterised as ApC SNPs - the result on the reverse strand of an ApT to GpT transition mutation. The derived CAG outlier is likely composed of CAAT to CAGT mutations and this is also seen looking at the nonsynonymous outlier of derived AGT also no longer being an outlier once CAAT to CAGT mutations were removed. Interestingly the repetitive derived codon of CCC remains an outlier for higher synonymous and nonsynonymous DMF and is likely a product of the low sequence complexity bias we saw in the previous section. Because of these results, I will also exclude ApT to GpT SNPs for all further downstream analysis, specifically in the CAAT, CTAT, and CAT ancestral contexts.

Even though the correlation is nominally not significant, the negative trend between nonsynonymous and synonymous DMFs is intriguing. One would expect codon usage bias to be visible in Figures 2.18C and 2.20C, but there appears to be no obvious trend. Indeed, the negative correlation is no longer present when looking at the underlying derived codons. More detailed analysis into human codon usage biases, and the propensity of nonsynonymous SNPs to move from an optimised codon for one amino acid towards a not-optimal codon for a different amino acid, needs to be considered and tested.

As introduced in section 2.2.1, another way to investigate the avoidance of derived motifs is to cluster on the CDFs built from their DMFs. Comparing the CDF clusters we obtained pre- and post-removal of CpG and ApT transitions (Figure 2.5), we now see methionine being more avoided (Figure 2.21A,C). This is likely due to the only codon for methionine being ATG and so would be greatly affected by high mutability of ACG towards ATG. Alanine is now visibly different to other amino acids, being the smallest chiral amino acid and least avoided post-removal of CpG and ApT transition biases. Otherwise, there are not many differences in the nonsynonymous clusters. Stop codons still have by far the most avoided CDF and greatest Manhattan distance from the other clusters. Other amino acids still cluster similarly, such as cysteine clustering with phenylalanine and tryptophan, and isoleucine being close in distance to glycine and arginine. Focusing on the transition-bias controlled CDF clusters (Figure 2.21), the most avoided cluster (bar the stop codons) seems to contain amino acids that are often involved in catalytic activities such as histidine, tyrosine, tryptophan, cysteine, and phenylalanine. For

Figure 2.17: Scatterplots comparing the geometric mean of derived single amino acids in the non-synonymous and synonymous contexts. A) All SNPs in the exome set of the gnomAD database, as in Figure 2.6. B) All CpG to TpG (and CpG to CpA) mutations based on the ancestral genome were removed. C) All ApT to GpT (and ApT to ApC) mutations in the CAAT, CTAT, and CAT ancestral genomic contexts were removed, in addition to CpG to TpG mutations. All amino acids and stop codons are labelled.

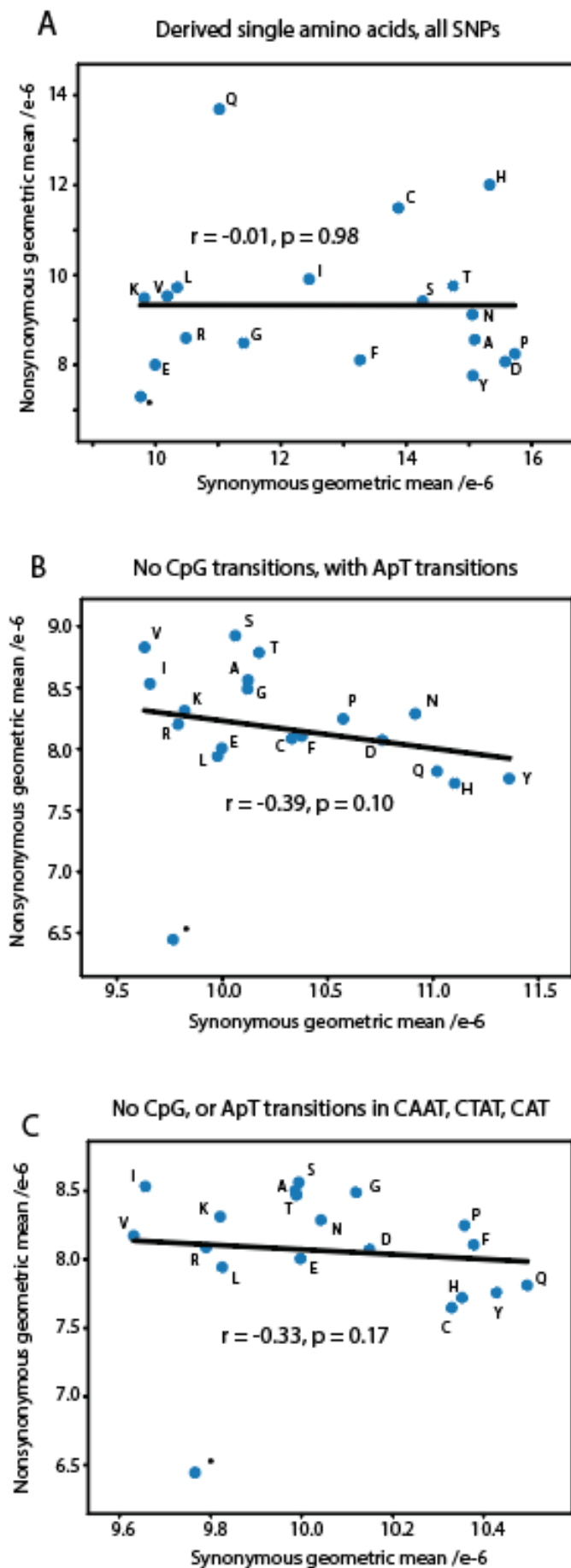


Figure 2.19: Scatterplots comparing the geometric mean of derived single amino acids in the non-synonymous and synonymous contexts. A) All SNPs in the exome set of the gnomAD database, as in Figure 2.6. B) All CpG to TpG (and CpG to CpA) mutations based on the ancestral genome were removed. C) All ApT to GpT (and ApT to ApC) mutations in the CAAT, CTAT, and CAT ancestral genomic contexts were removed, in addition to CpG to TpG mutations. All amino acids and stop codons are labelled.

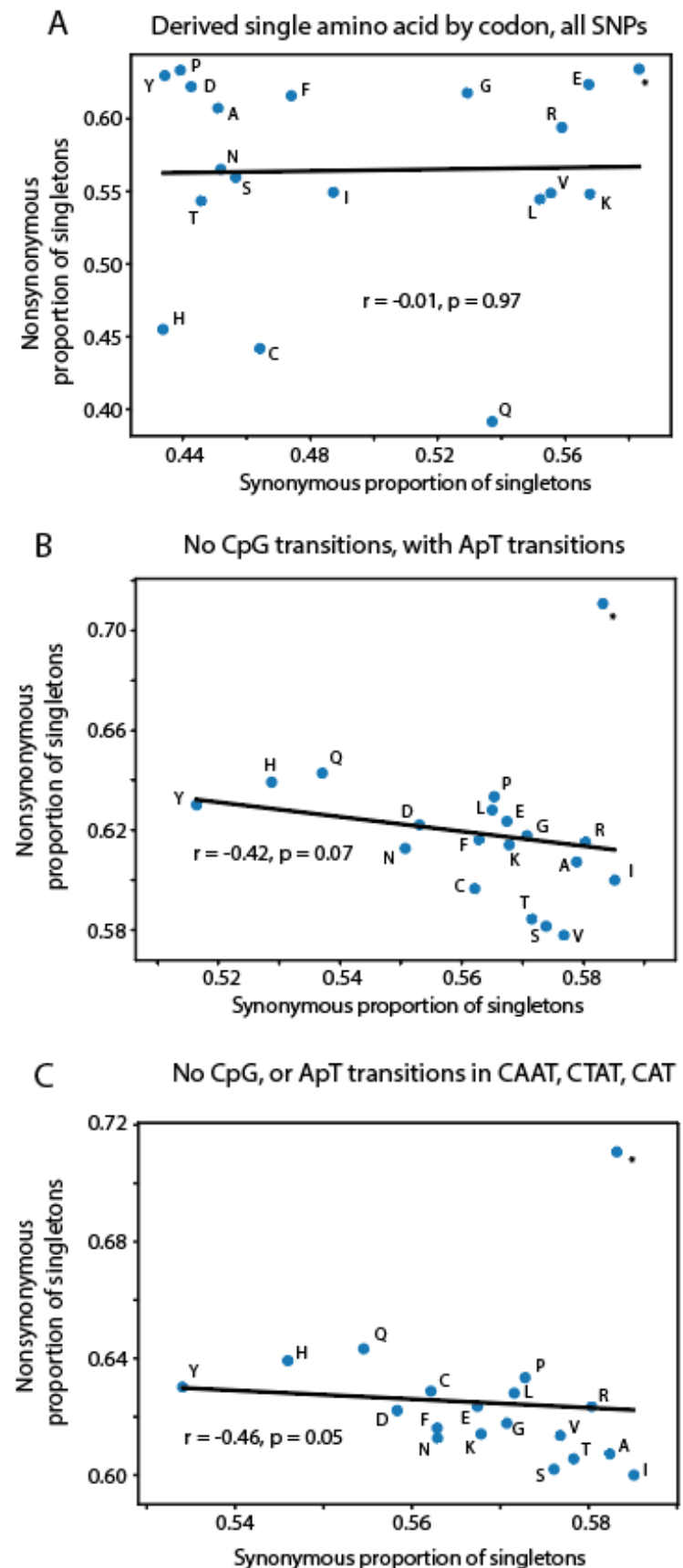
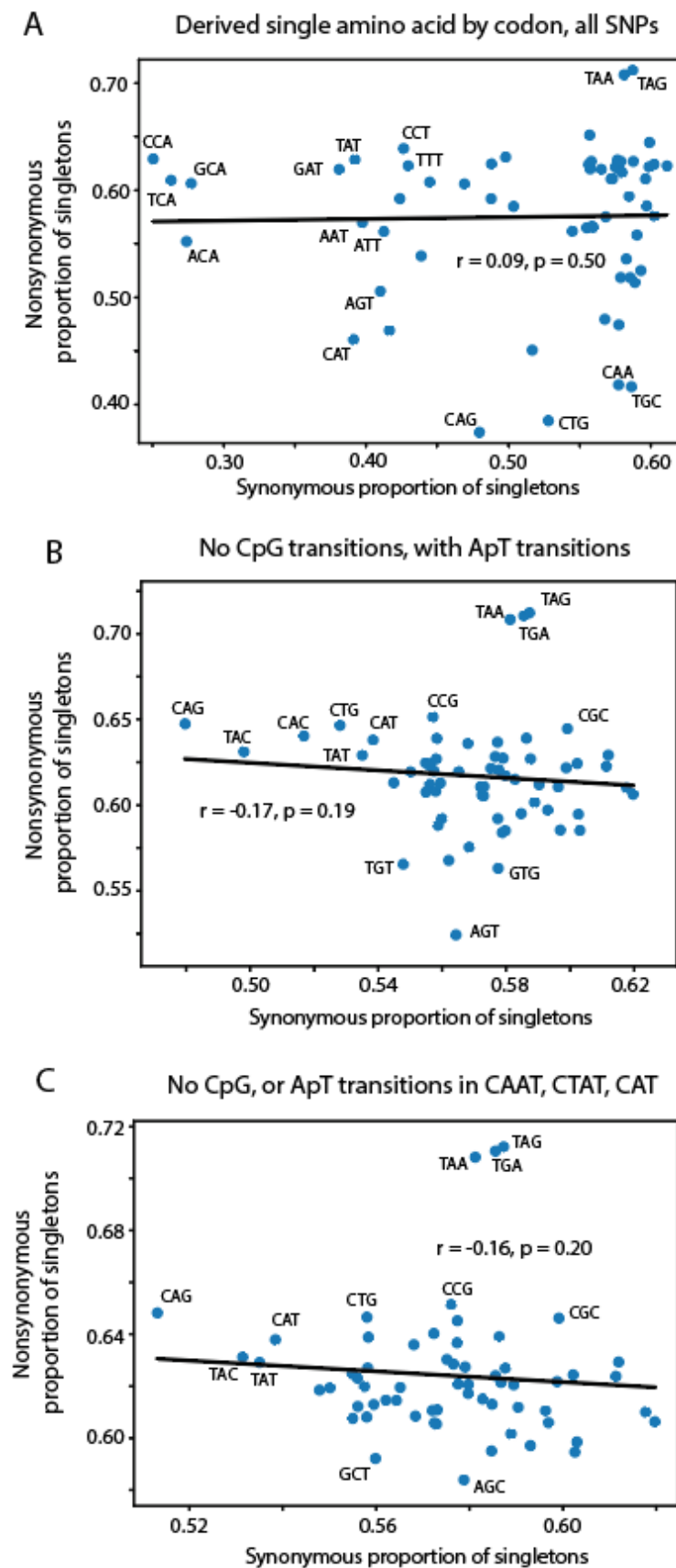


Figure 2.20: Scatterplots comparing the geometric mean of derived single amino acid codons in the nonsynonymous and synonymous contexts. A) All SNPs in the exome set of the gnomAD database, as in Figure 2.6. B) All CpG to TpG (and CpG to CpA) mutations based on the ancestral genome were removed. C) All ApT to GpT (and ApT to ApC) mutations in the CAAT, CTAT, and CAT ancestral genomic contexts were removed, in addition to CpG to TpG mutations. Some outlier codons are labelled.



the other nonsynonymous clusters the trend is less discernible, with quite disparate amino acids appearing close in CDF. For example, glycine is clustering with arginine as one of the highest frequency CDFs and lysine clustering with leucine. Some of this structure is lost in the clustering based on CDFs generated on the derived synonymous context (Figures 2.21B, D). However, there are a couple of amino acids that keep similar clusters in the synonymous context, for example serine and threonine from the green cluster; and aspartate, leucine, and valine from the red cluster. This might be an extension of the negative trend between the nonsynonymous and synonymous DMFs observed in Figures 2.17C, 2.19C where amino acids with mid-range DMFs cluster in both nonsynonymous and synonymous contexts due to remaining in the middle of the DMFs.

It is important to compare the point estimate of the geometric mean with the combined information of the CDFs. To do so, I plotted the geometric mean of derived nonsynonymous amino acids against the cumulative values taken as a slice in the CDF vector. Here I used the cumulative proportion of derived alleles with counts less than 100,000, since in Figures 2.21 and 2.5 that is a point where the cumulative lines seem to be stable in their divergence. Doing so, I see that before removal of the factors skewing DMFs, there was no correlation between the geometric mean and the corresponding CDF slice (Figure 2.22A), whereas after controlling for these a strong correlation is seen (spearman rank of -0.80, $p = 1.1e - 5$ Figure 2.22B). This implies that after removal of the skewing biases, the point estimate of the geometric mean now generally returns consistent information compared to use of the entire CDF distribution.

2.2.5 Quantifying the effect of the chemical properties of amino acids on their avoidance

To see whether the hypotheses regarding the functional aspects of the individual amino acids hold true, I compared the nonsynonymous DMFs to the physico-chemical properties of the amino acids. These chemical properties were described by calculating the top 5 eigenvectors from 553 datasets of experimentally derived chemical properties of amino acids from the Amino Acid Index [Kawashima and Kanehisa, 2000] as done previously by Venkatarajan and Braun

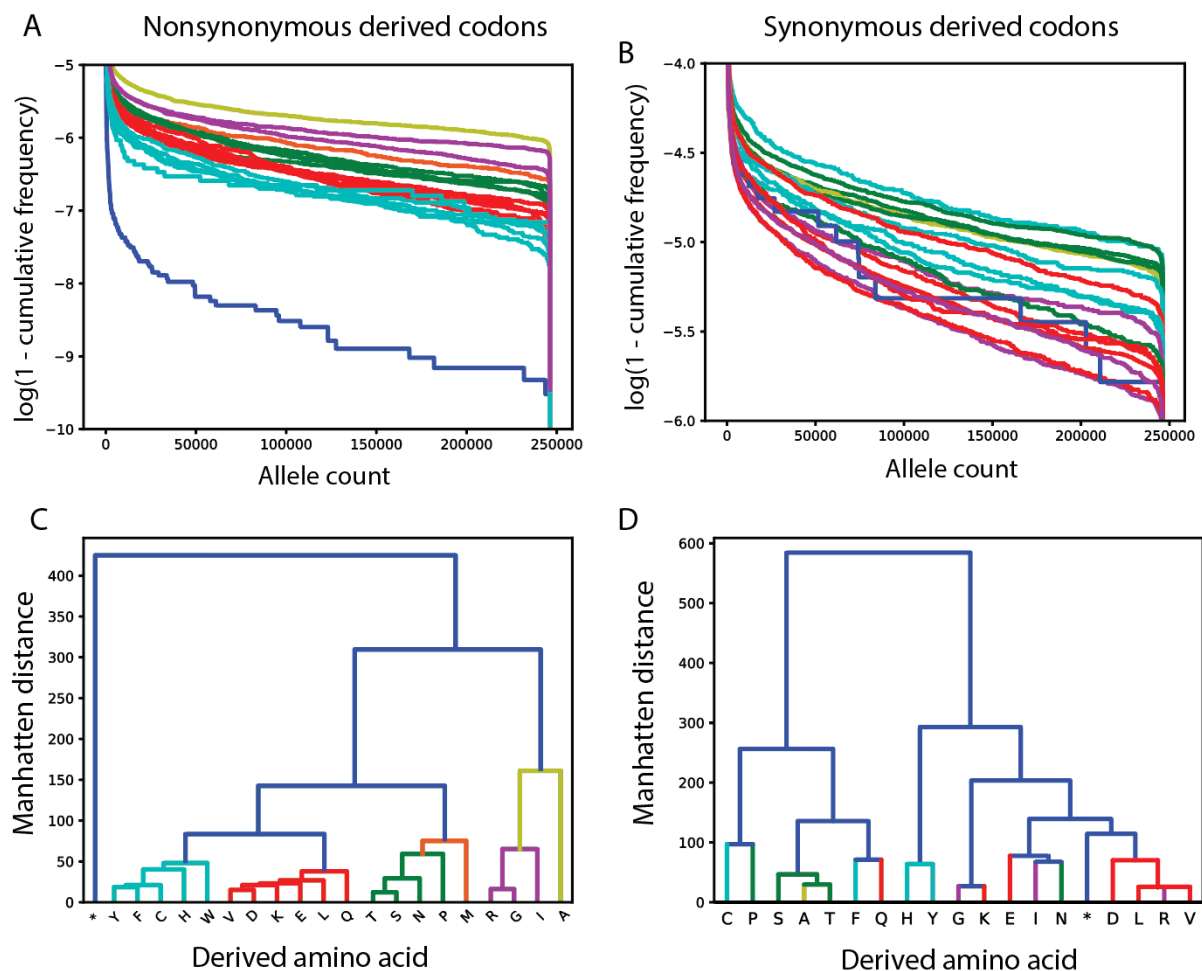


Figure 2.21: The CDF analysis of derived single amino acids after controlling for SNP biases. The CDFs (A,B) and hierarchical dendrograms (C,D) for nonsynonymous (A,C) and synonymous (B,D) derived single amino acids and stop codons. The colours for each amino acid and stop codon correspond to the clusters for nonsynonymous derived codons, depicted in subplot C. In C,D the y axis is plotted as $\log(1-y)$ to better visualise the separation between the CDFs at rare allele frequencies. The dendrograms were generated by using centroid clustering on the Manhattan distance between each CDF vector.

Comparing the geometric mean and CDFs of derived single amino acids

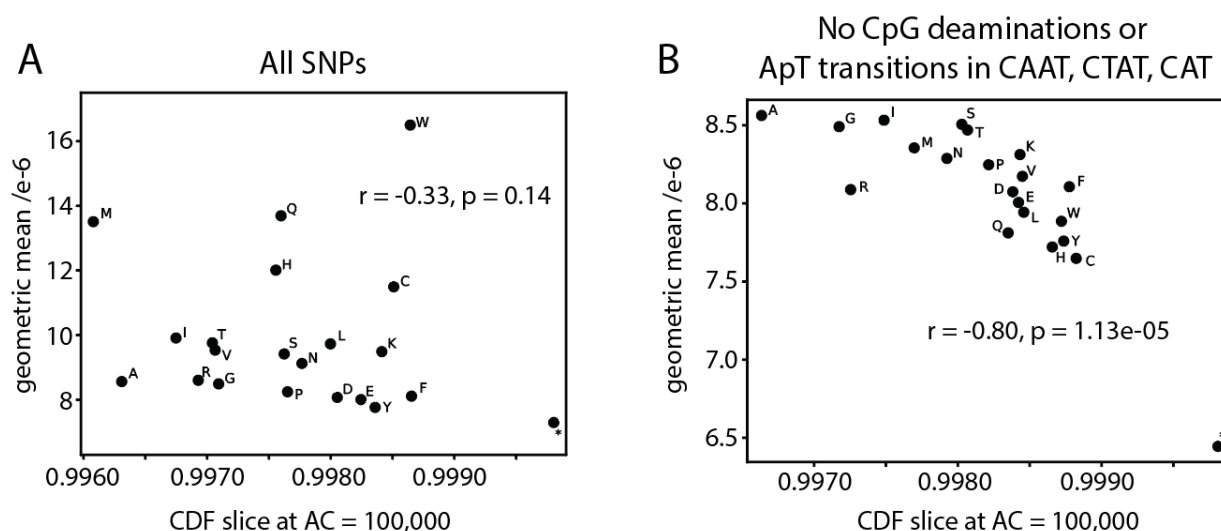


Figure 2.22: Scatterplots comparing the point estimate geometric mean of derived single amino acids in the nonsynonymous context with their corresponding CDF value at allele counts less than 100,000. This slice point was chosen based on Figures 2.5 and 2.21, since at allele counts up to 100,000 the CDF lines seem to be stably divergent. A) A scatterplot comparing the geometric mean and the CDFs before removal of CpG deaminations and ApT transitions in the CAAT, CTAT, and CAT contexts. B) A scatterplot comparing the geometric mean and the CDFs after removal of CPG deaminations and ApT transitions. Spearman ranks and their p-values are denoted on the plots.

[2001] but with an updated Amino Acid Index database. The top 5 eigen vectors capture 98.75% of the information contained in all 553 datasets and therefore we can limit the analysis to these 5. These top 5 vectors correlate the most with hydrophobicity, steric size, α -helical propensity, linker propensity, and a collection of other structural properties as shown in Table 2.1. Calculating the spearman rank correlations for these five eigenvectors with the DMF geometric mean and proportion of singletons for nonsynonymous derived motifs resulted in 3 comparisons being significant or close to significant, and these are shown in Figure 2.23. The third eigenvector, which represents the α -helical propensity of an amino acid, correlated with both DMF metrics (Figure 2.23A,C). A study by Abrusan and Marsh saw that α -helices can tolerate more mutations than β -sheets due to the greater number of inter-residue contacts stabilising missense mutants in the helical structure [Abrusán and Marsh, 2016]. My result is consistent with that observation, with amino acids with greater propensity to form an α -helix being less avoided than those with less such propensity. In addition, the geometric mean correlates with the steric size of amino acids arising in the nonsynonymous context where larger amino acids are associated with rarer DMFs (Figure 2.23E). This also makes sense, since larger amino acids can not fit in the same protein pockets and folds as smaller amino acids and therefore carry greater disturbance to the protein fold when arising by chance [Huang and Nau, 2003]. On the other hand, smaller amino acids offer greater structural flexibility and would be less detrimental to a protein's fold. This is also consistent with previous literature, where the steric size of an amino acid and the hydrophobicity described the most variance of the differences between amino acids [Venkatarajan and Braun, 2001, Micheletti et al., 1998].

This quantification of amino acid physico-chemical properties can also be done on the CDFs clusters seen in Figure 2.21. I compared the physico-chemical properties of each amino acid in each cluster. This was done using the same eigenvector decomposition from earlier. In Figure 2.24 there does not seem to be a single property that coherently explains any of the 6 clusters (when ignoring derived stop codons as the 7th cluster). However, for some individual clusters, the amino acids in that cluster do tend to have similar values for particular eigen vectors. α -helical propensity seems to be generally congruent within each cluster (Figure 2.24A) and this is coherent with the strong correlations seen between α -helical propensity and nonsynonymous

Eigenvector	Rank	Chemical property and reference	Correlation
Hydrophobicity-related Eigenvector	1	14Å contact number (Nishikawa-Ooi, 1986)	r = 0.96
	2	Effective partition energy (Miyazawa-Jernigan, 1985)	r = 0.95
	3	Interactivity scale obtained by maximizing the mean of correlation	r = 0.94
	4	Interactivity scale obtained from the contact matrix (Bastolla et al., 2005)	r = 0.94
	5	Hydrophobicity scales (Ponnuswamy, 1993)	r = 0.93
Size-related Eigenvector	1	Distance between C α and centroid of the side chain (Levitt, 1976)	r = 0.88
	2	STERIMOL length of the side chain (Fauchere et al., 1998)	r = 0.88
	3	Radius of gyration of the side chain (Levitt, 1976)	r = 0.87
	4	Average weighted eccentricity based on the atomic number	r = 0.85
	5	The number of bonds in the longest chain (Charton-Charton, 1983)	r = 0.83
α -helical propensity Eigenvector	1	Normalized frequency of turn (Crawford et al., 1973)	r = 0.73
	2	Conformational parameter of β -turn (Beghin-Dirkz, 1975)	r = 0.69
	3	Information measure for turn (Robson-Suzuki, 1976)	r = 0.68
	4	Information measure for loop (Robson-Suzuki, 1976)	r = 0.67
	5	Information measure for N-terminal turn (Robson-Suzuki, 1976)	r = 0.67
Linker propensity Eigenvector	1	Linker propensity from medium dataset	r = 0.83
	2	Linker propensity from 3-linker dataset (George-Heringa, 2003)	r = 0.82
	3	pK-N (Fasman, 1976)	r = 0.81
	4	Linker propensity from all dataset (George-Heringa, 2003)	r = 0.79
	5	Normalized positional residue frequency at helix termini C	r = 0.76
Structural-related Eigenvector	1	Information measure for N-terminal helix (Robson-Suzuki, 1976)	r = 0.73
	2	Principal property value z3 (Wold et al., 1987)	r = 0.66
	3	Average weighted atomic number or degree based on atomic number in the graph	r = 0.64
	4	Normalized frequency of N-terminal helix (Chou-Fasman, 1978b)	r = 0.61
	5	Hydrophobicity coefficient in RP-HPLC, C4 with 0.1% TFA/MeCN/H2O	r = 0.57

Table 2.1: Top chemical property correlates for each eigenvector

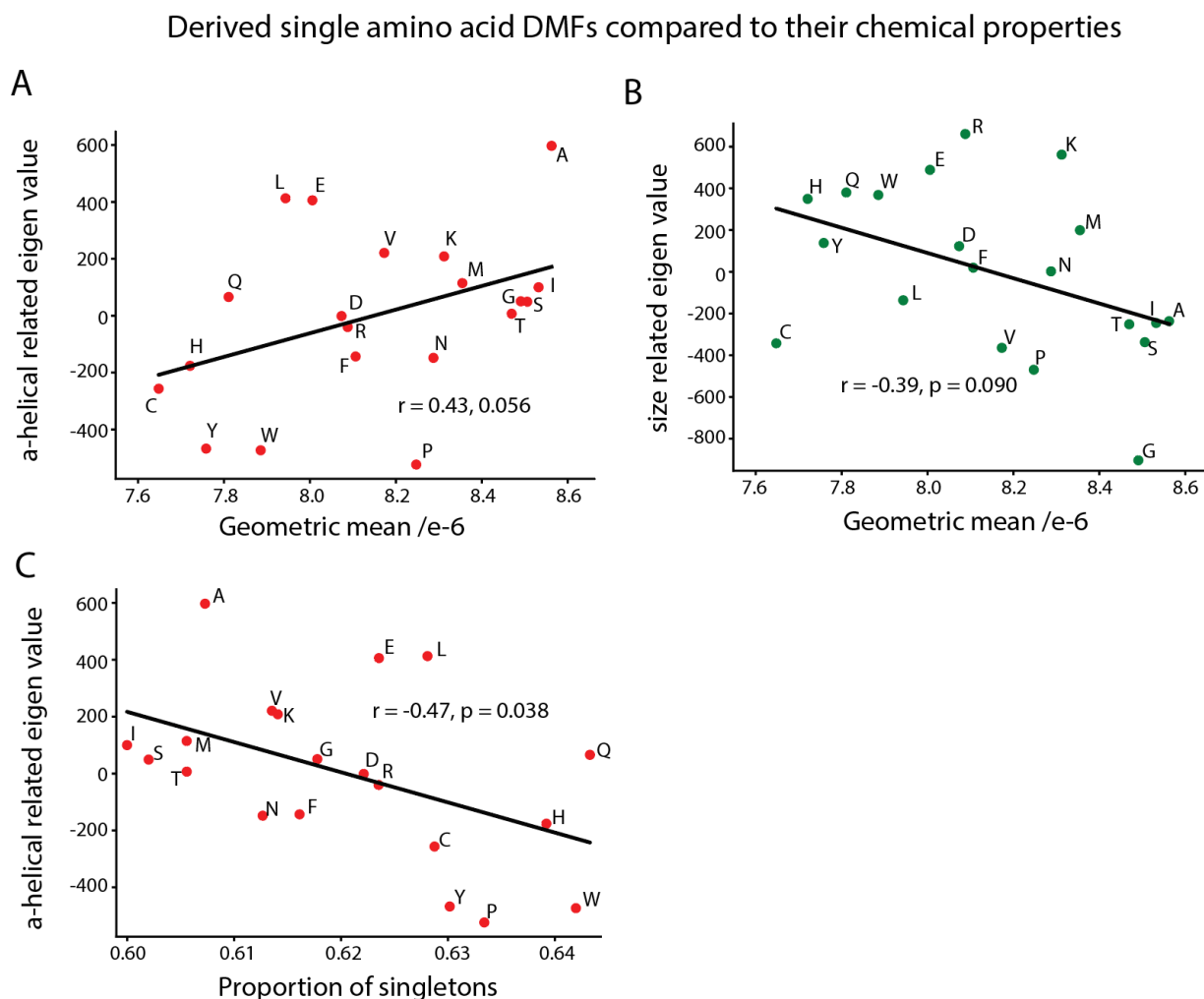


Figure 2.23: Scatterplots showing the significant correlations of single amino acid DMFs with the chemical property eigenvectors, calculated as described in the text and methods. The spearman rank correlations and p-values are written on the figure. Only nonsynonymous amino acid DMFs are shown. The colour of the points depict the eigen component being compared, and the black line is a line of best fit calculated as a 1st order polynomial using the ordinary least squares regression.

DMFs in Figures 2.17C, 2.19C. The red cluster, containing amino acids glutamine (Q), leucine (L), glutamate (E), lysine (K), aspartase (D), and valine (V), seems to be best described by having average values for α -helical propensity and linker-propensity, and extreme values in both directions regarding hydrophobicity. Clustering by the physico-chemical eigenvectors themselves does not succeed in creating these clusters, implying that there is more involved in driving particular amino acids to particular DMFs than their intrinsic physico-chemical properties. Other contributors might include the effect of post-translational modifications, where in certain proteins a derived amino acid might be liable to be targeted for a modification such as phosphorylation, making the derived amino acid behave quite differently to if it arose in an unmodified way. Another might be the context the amino acid arises in, which both includes the amino acid sequence surrounding it as well as the functional context it appears in. For example an amino acid occurring at a catalytic site will likely behave very differently to one arising in a hydrophobic stretch of a transmembrane protein. The observed DMF trends are not strong enough to make conclusions, especially for single amino acids where there are a large variety of such possible contexts for each amino acid. One way to get more insight into the idiosyncrasies of derived amino acid functionality is to look beyond single amino acids towards combinations of them.

2.2.6 Extension to longer motif patterns

As the derived motif is extended, the total number of motifs present at each length becomes exponentially larger. For amino acid motifs of length k the total number of motifs equates to 20^k and for nucleotide codon motifs of length $3k$ the total number of motifs varies as 4^{3k} . Therefore, despite the large size of the human genome there quite rapidly comes a motif length where certain motifs no longer appear at sufficient frequency for statistical analysis. This is exacerbated by the requirement that we need a SNP to have generated the derived motif from the ancestral and not simply for the motif to be present in the genome at a conserved site. To determine at what length we may test derived motifs, I plotted the sample size of each derived amino acid motif of length k amino acids (Figure 2.25). Here we see that the sample size drops

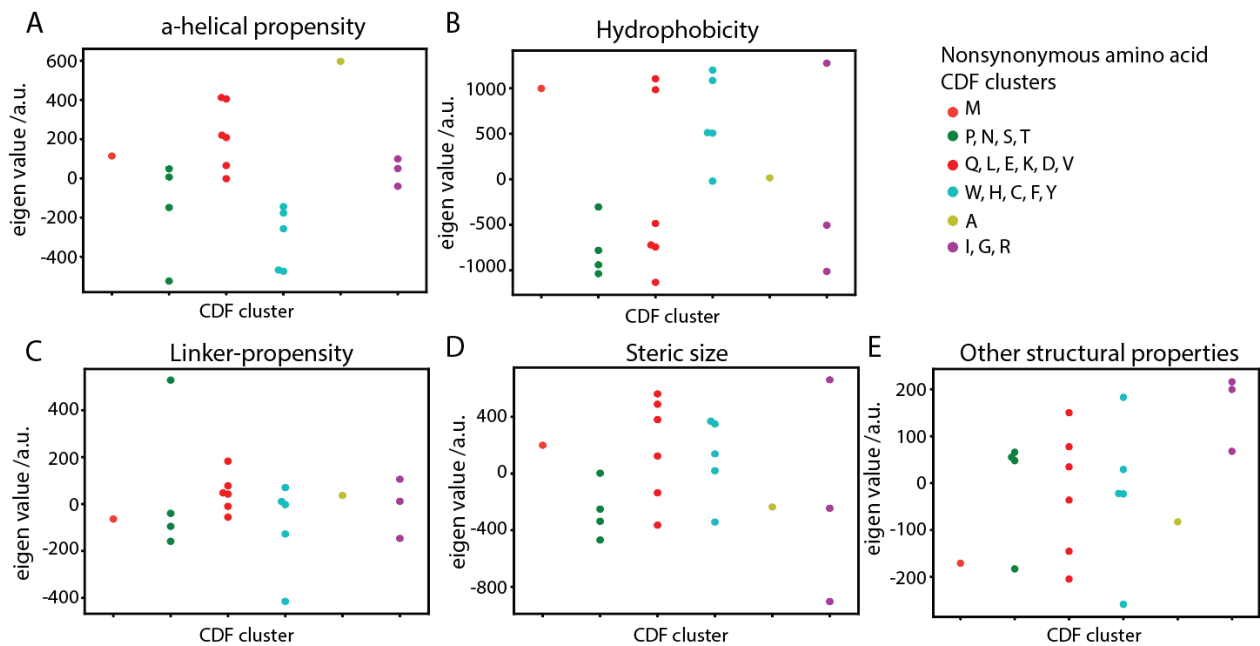


Figure 2.24: Scatterplots showing the eigen values of each amino acid in each CDF cluster for nonsynonymous derived single amino acids. The eigen values relate to the physico-chemical properties of α -helical propensity (A), hydrophobicity (B), linker-propensity (C), steric size (D), and other structural properties (E). The CDFs here are ones generated excluding all CpG to TpG SNPs, and ApT to GpT SNPs in the CAAT, CTAT, and CAT ancestral genomic contexts.

rapidly as the motif length is increased, from an average sample size of 10,000 occurrences when the motif is 2 amino acids long to only 10 occurrences of each motif when the motif is 5 amino acids long. Without having to group motifs by *a priori* ideas such as chemical properties to strengthen sample power, it seems the maximum motif length I may analyse at first are motifs of length 3 or 4, giving sample sizes in the 1000s and 100s per motif respectively. The broader goal of this study is to extend the analysis of DMFs to longer motifs to further elucidate the driving forces of avoidance. *Nota bene* the DMF approach and all associated analyses are still under testing and development, but in the next sections I will present some preliminary findings towards this goal.

2.2.7 The avoidance of dipeptides

The first natural step after investigating derived single amino acids is to look at derived dipeptides. Initially, I visualised the derived dipeptides in the same way I did for derived single amino acids - plotting the DMF geometric means and proportion of singletons between their

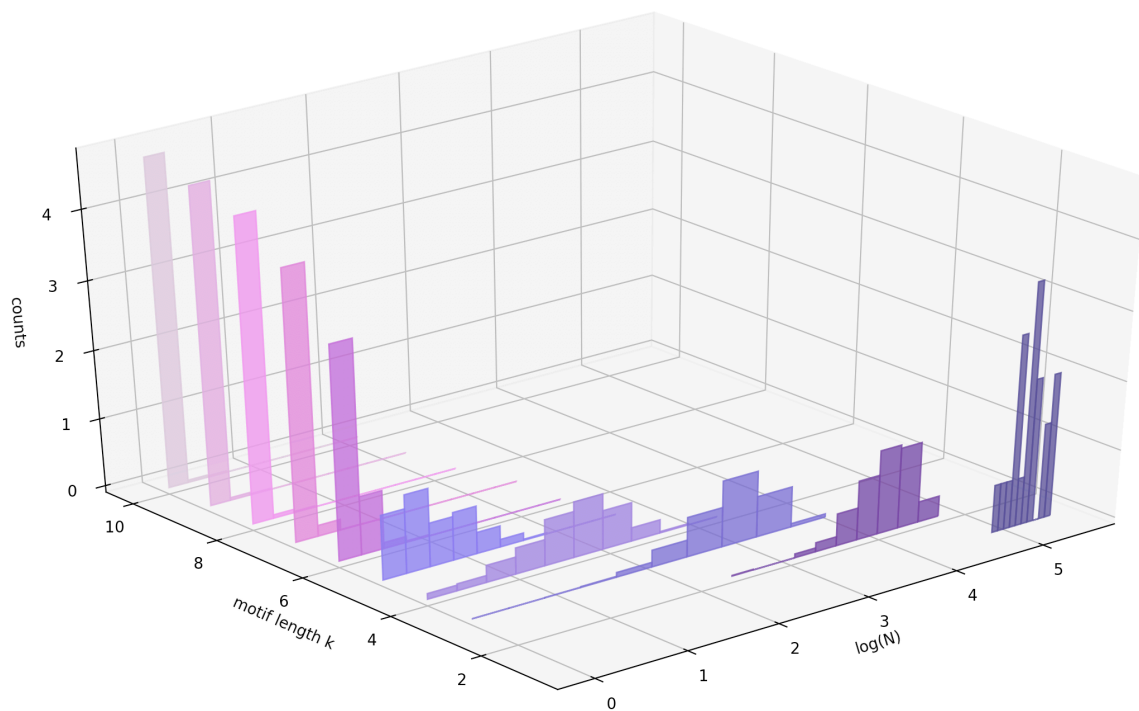


Figure 2.25: A 3D histogram showing the log number of occurrences (N) of each derived amino acid motif versus the length of the motif. For each sub-histogram, the sample number was split into 10 bins.

occurrence in a nonsynonymous or synonymous context (Figure 2.26). Like for single derived amino acids, all CpG to TpG SNPs as well as ApT to GpT SNPs in the CAAT, CTAT, and CAT contexts were removed, and only derived dipeptides with more than 100 occurrences in gnomAD were included to avoid low sample size outliers from skewing the analysis. When looking at both the DMF geometric means (Figure 2.26A) and proportion of singletons (Figure 2.26B), it is similarly seen that all derived motifs containing a stop codon (except G*) occur at lower DMFs and higher proportion of singletons respectively. This suggest, unsurprisingly, that dipeptides will - to some extent - inherit the avoidance profile of their component amino acids. When looking at the derived geometric means, the next most avoided group when compared to their synonymous context are tryptophan-containing (W) motifs (Figure 2.26A). Tryptophan is the largest hydrophobic residue and therefore its occurrence in a nonsynonymous context is likely a large hindrance to the appropriate folding and structure of the final protein. What is perhaps more interesting is that W-containing motifs also appear at high DMF values and with a low proportion of singletons in some contexts such as WG/GW or TW/WT. It is worth clarifying that this includes mutations at either the former or latter amino acid. Looking into the WG and GW outliers, 69.4% of the mutations are for the derived glycine, and it is therefore less surprising that those are more tolerated. Regardless, this highlights that one can capture trends beyond simple addition of the patterns for single amino acids and this will be addressed in further detail below. In Figure 2.26B, it is in addition seen that some tyrosine (Y)-containing derived dipeptides are more avoided, having a higher proportions of singletons. Tyrosine is the second largest hydrophobic amino acid residue and so this is consistent with an idea that larger hydrophobic residues are avoided when occurring out of context. Another feature both W and Y residues have in common is that they are aromatic, containing a complete π -electron orbital system. This idea would extend to the other residues seen clustering with both W and Y from Figures 2.21, 2.24 when looking at derived single amino acids, since histidine (H) and phenylalanine (F) also are aromatic. The negative correlation between nonsynonymous and synonymous DMFs is significant for dipeptides, showing a spearman rank of -0.11 ($p = 0.029$) and -0.33 ($p = 1.2e - 11$) for the dipeptide geometric means and proportion of singletons respectively when ignoring stop codon containing motifs. When looking at the derived codons underlying

these motifs, no obvious outlier codons were driving these trends.

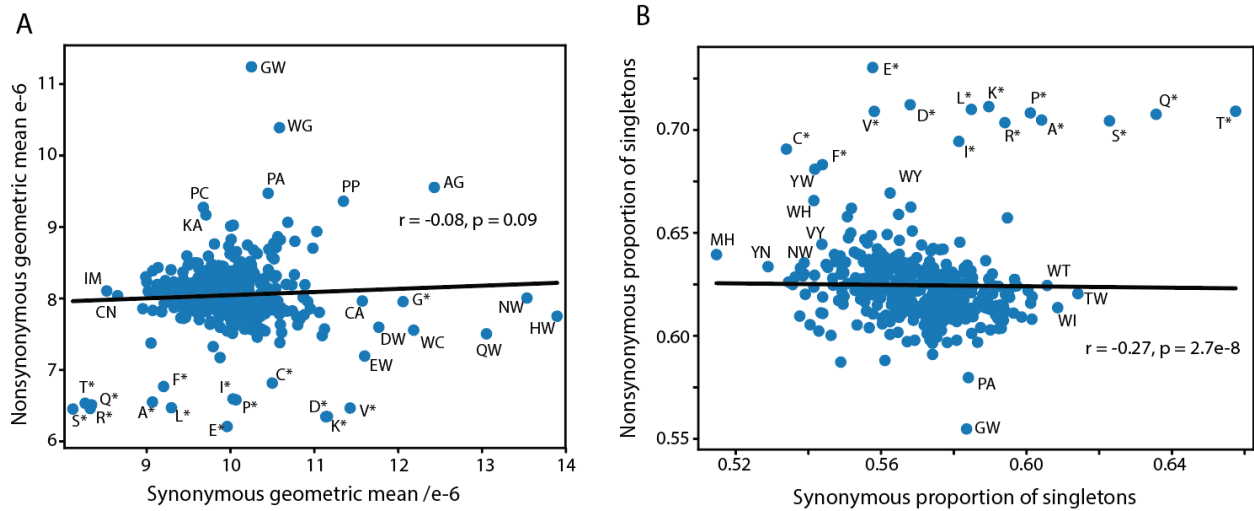


Figure 2.26: Scatterplots comparing the DMF geometric mean (A) and proportion of singletons (B) of derived dipeptide motifs between their nonsynonymous and synonymous contexts. All SNPs equivalent to CpG deaminations, and ApT transitions in CAAT, CTAT, and CAT ancestral genomic contexts, were excluded. Motifs more than 2 standard deviations from the mean value of the points on each axis are labelled. Only derived motifs with more than 100 occurrences were included for analysis, to avoid outliers due to small sample size.

2.2.8 Dipeptides as a product of single amino acids

Derived dipeptides also can be decomposed into their two comprising amino acids. This offers the first chance to compare whether avoidance of a complete motif is different to the smaller parts composing it. As a first pass, I plotted the DMF metrics for each dipeptide (black) and their constituent single amino acids (blue, magenta), visualised in Figure 2.27A and B. The points were sorted by their respective DMF metric. A small subsection of derived dipeptides are more avoided than their comprising single amino acids, labelled with a grey box in Figure 2.27A and B. I zoomed in on these to investigate how they behave as a function of their submotifs. The simplest way to check the difference in avoidance between the single amino acid submotifs and the complete dipeptide would be to sum the differences between the DMF geometric mean (or proportion of singletons) of the dipeptide XY, and its submotifs X and Y. A way to picture this is as the perimeter of a triangle, where the vertices are the derived motifs XY, X, and Y respectively, and so for brevity I will term this metric the *Triangle score*. If XY behaves similarly to X and Y, the Triangle score will be small. If XY is avoided much more than both

X and Y, the value will be larger. Likewise, if only one of the submotifs is much more avoided than the other and the dipeptide rescues that avoidance, it will similarly have a larger sum of differences. In Figure 2.27C and D, I sorted the most avoided dipeptides by their Triangle score and two modes of epistatic interactions are seen. Firstly, some dipeptides mainly behave like one of the constituent amino acids. For example for the motifs PY and CT, the avoidance of the dipeptide is much more similar to the avoidance of the P (proline) and C (cysteine) than to Y (tyrosine) and T (threonine) respectively. This shows that for these avoided motifs the actual avoidance is due to the containment of the single heavily avoided amino acid rather than the entire dipeptide itself. In contrast, some dipeptides have much greater avoidance than either constituent single amino acid as in the examples EW, YW, and FD. For these, this is *prima facie* evidence that the dipeptide itself is avoided, maybe by creating some new interaction that the individual amino acids could not achieve. However, a question that is important to answer before making conclusions is whether this pattern accurately reflects the avoidance of the dipeptide, or merely the reflection of what constituent amino acid substitution is most common. A consistent way to control for this will be a necessary continuation of this analysis.

2.2.9 Asymmetry of dipeptide avoidance

Another useful aspect of investigating longer derived motifs is the opportunity to investigate symmetry and order of amino acids in derived motifs. Derived dipeptides are the earliest example of this, allowing the exploration of whether a derived motif AB would behave differently to the derived motif BA. I compared the arithmetic difference between a dipeptide arising in its forward form (AB) versus the dipeptide arising in its reverse form (BA). I call the resulting difference the motif's *symmetry excess*. Plotting and sorting them by this DMF excess, a curve is seen with two parts: a part generally flat until a much more rapid increase (Figure 2.28A). To investigate the most asymmetric dipeptides, I selected the outliers with DMF excess greater than $5e-7$ (grey line in Figure 2.28A) and separated these 17 dipeptides by their derived codons (Figure 2.28B) to see if particular codons or nucleotide effects might be driving the discrepancy. Generally, there do not seem to be many outlier codons except 'GGCTAA' for

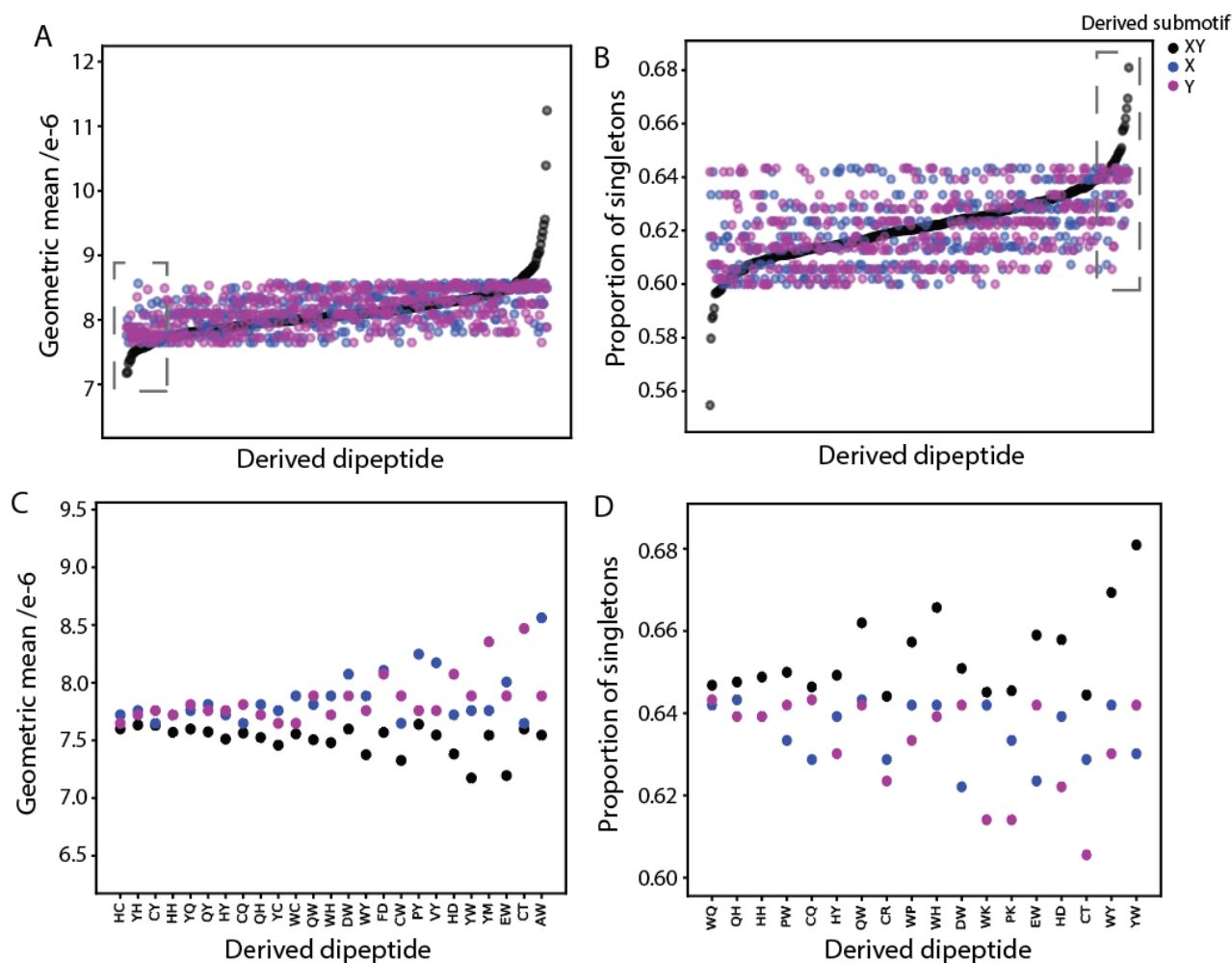


Figure 2.27: Scatterplots investigating dipeptides as a product of their submotifs. These are showing the geometric mean (A,C) and proportion of singletons (B,D) for derived dipeptides XY and their comprising single amino acids X and Y. A,B) Scatterplots showing all derived dipeptides, sorted by the DMF metric. Blue points are the first amino acid in the dipeptide, and magenta points are the second. The grey box depicts motifs where the dipeptide is more avoided than either submotif, and are the motifs illustrated in C and D. C,D) Scatterplots zooming into the grey boxes from A and B, with points sorted their Triangle score (see text).

derived G*'dipeptides. Why this might be such a high DMF outlier is not clear. In the codons comprising derived G* or *G motifs, there are other more repetitive nucleotide motifs, so this outlier is unlikely to be due to low complexity regions. The other higher DMF outliers include some repetitive nucleotide motifs such as 'TGGGGG' for WG dipeptides and 'GCCCCC' for PA dipeptides. Out of the 17 dipeptides with the greatest DMF excess between the two directions of the motif, W-containing and G-containing motifs seem to be the most numerous. In the W-containing motifs, it consistently appears to be the case that, among the most asymmetric dipeptides, the W occurring downstream is more avoided than an N-terminal W (Figure 2.28C). I speculate here that there is some greater constraint on the residues appearing at the N-terminus of such a large hydrophobic residue as tryptophan, maybe due to the biophysics of folding a tryptophan containing peptide, or some other steric interactions being greater in that direction along a peptide. Oddly, similarly for G-containing motifs there is greater avoidance of the G occurring at the C-terminal end of the dipeptide albeit less clearly than for the W-containing motifs (Figure 2.28). The constraint in this case might be due to some constraint on maintaining chirality of the peptide chain. This is supported by the GA motif behaving oppositely to this trend, since alanine is the smallest chiral amino acid. However, before being able to infer anything, I would need to ensure that this asymmetry cannot simply be explained by the tryptophan alone, and so would need integration with the prior analysis of investigating submotifs.

2.2.10 Derived tripeptides

The next natural step is to extend the derived motif one amino acid further - towards derived tripeptides. Collating the derived tripeptides in the same way as for single amino acids and dipeptides, that is excluding all CpG to TpG transitions and ApT to GpT transitions in the CAAT, CTAT, and CAT ancestral genomic contexts, and using the exome portion of the gnomAD dataset, Figure 2.29 is obtained. The variance of tripeptide DMFs is visibly increased in comparison to the dipeptides and single amino acids comprising it, as one would expect since a smaller motif would be sampling more functional contexts than a larger motif. This seems

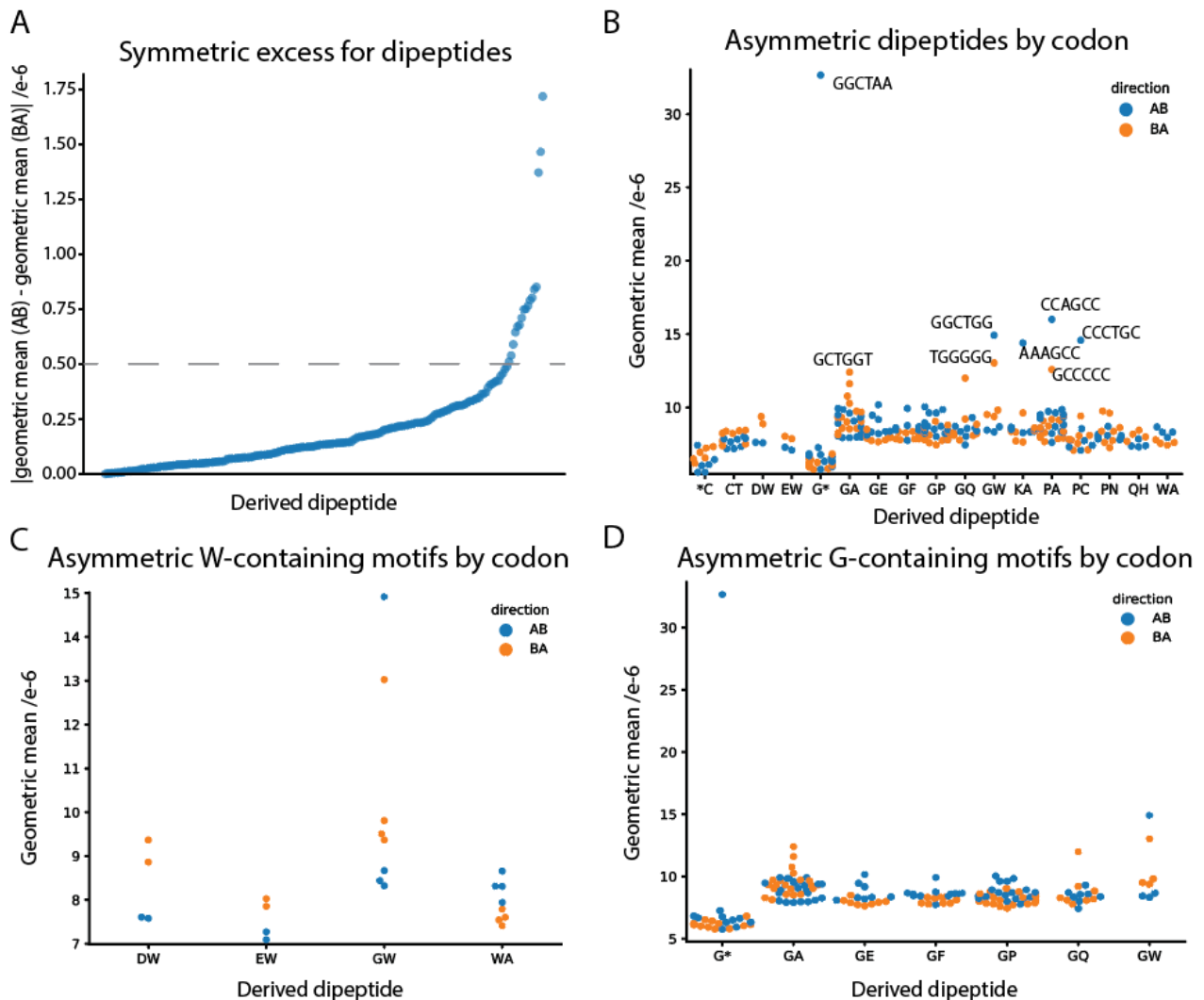


Figure 2.28: Scatterplots investigating the difference between the DMFs for hypothetical motif AB compared to its reverse BA. A) A scatterplot showing the DMF excess between the DMF geometric mean for the derived dipeptide and the DMF of its reverse, for derived dipeptides in the nonsynonymous context. The dashed grey line is a threshold of DMF excess, which was used to select dipeptides for subplot B. B) The top outliers from A (with excess >0.0000005) separated by the derived nucleotide codons composing them. Only nucleotide codons with sample size greater than 100 were included. The top outlier codons are labelled. C,D) Zoomed in version of B, focusing on W-containing and G-containing dipeptides respectively. All SNPs equivalent to CpG deaminations, and ApT transitions in CAAT, CTAT, and CAT ancestral genomic contexts. Only derived motifs with more than 100 occurrences were included for analysis, to avoid outliers due to small sample size.

exacerbated in some cases for high DMF motifs, which is not ideal due to the avoided motifs being the motifs of interest in this study, not the motifs rising to higher allele frequencies or becoming fixed in the population. The highest outliers of WLГ, GWG, and PCH for the nonsynonymous DMF geometric mean, VAG and MAG for synonymous DMF geometric mean (Figure 2.29A), appear to be from very repetitive nucleotide sequence contexts with all codons being particularly GC-rich. Explicitly for the tripeptide MAG the derived codons look repetitive, with repeating ‘TGG’ units separated only by individual nucleotides as in the example ‘ATGGCTGGG’. However, again I mainly care about the other end of the spectrum: what motifs are avoided.

When looking at the proportion of singletons (Figure 2.29B), there are a couple of derived tripeptides being avoided beyond 4 standard deviations of the mean in both the nonsynonymous and synonymous direction. In the nonsynonymous direction, we see avoidance of YCT, YTM, CQI, and QFD, which all are mostly composed of quite reactive amino acid residues. These include the thiol cysteine (C); the residues tyrosine (Y) and threonine (T) that can be phosphorylated or can act as nucleophiles; the basic and acidic residues glutamine (Q) and aspartate (D); and the aromatics tyrosine and phenylalanine (F).

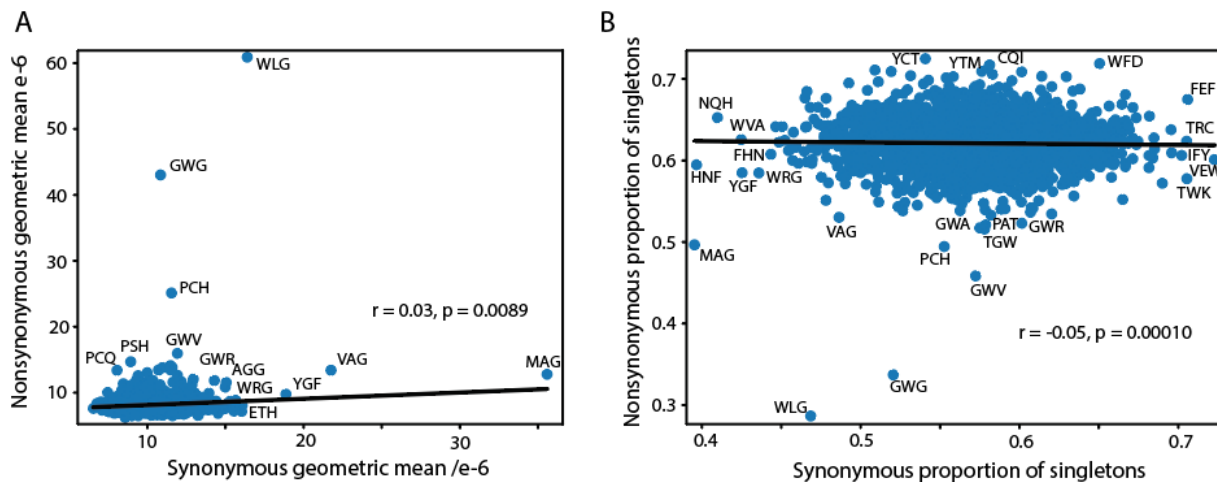


Figure 2.29: Scatterplots investigating derived tripeptides. These compare the DMF geometric mean (A) and proportion of singletons (B) of derived tripeptide motifs between their nonsynonymous and synonymous contexts. All SNPs equivalent to CpG deaminations, and ApT transitions in CAAT, CTAT, and CAT ancestral genomic contexts, were excluded. Motifs more than 4 standard deviations from the mean value of the points on each axis are labelled. Only derived motifs with more than 100 occurrences were included for analysis, to avoid outliers due to small sample size.

2.2.11 Tripeptides as a function of dipeptides

Similar to dipeptides, I interrogated tripeptides as the combination of their constituent dipeptides. As a first pass, I plotted the DMF metrics for each dipeptide (blue, magenta) composing the most avoided derived tripeptides (black), visualised in Figure 2.30A and B. I fixed the SNP to have to occur in the central amino acid of the tripeptide, so only tripeptide-dipeptide triplets that have the SNP at the same position are considered. This is to control for avoidance of the fringe amino acid skewing analysis. The most avoided tripeptides were defined as ones with nonsynonymous geometric mean less than $7e-6$ or having a proportion of singletons greater than 0.7. As shown in Figure 2.30A and B, it appears that derived tripeptides are generally more avoided than their submotifs. This might be due to there being less occurrences for each tripeptide than there are for the composing dipeptides. This leads to the DMF metrics having more extreme values due to individual outliers within the DMF distribution having greater weight. Alternatively, a more biological reason might be due to increasingly longer motifs having much more specific roles in the functionality of a protein than single amino acids or dipeptides, leading to selection being greater. For example, having 3 large tryptophan residues in a row would be much more detrimental to a region of a protein that needs to be hydrophilic than a single tryptophan would be.

For each of the most avoided tripeptides in Figures 2.30A and B, I calculated the Triangle score between the tripeptide and both dipeptides and sorted the motifs by that value (Figures 2.30C,D). From this, I selected the tripeptides that seemed to have the greatest differences with their dipeptide submotifs and visualised these top outliers in the same way as subplots A and B, sorted by their Triangle score (Figures 2.30E,F). Even though there is not much overlap between the top outliers in E and F, there is a global spearman correlation of $\rho = 0.45, p = 1.6e^{-304}$ between the Triangle scores calculated from the geometric mean or the proportion of singletons, showing there is generally good agreement between the two metrics. Again two modes of epistasis are seen as for the dipeptides previously. For some tripeptides, such as PPF and GPY, one of the underlying dipeptides (PF and PY respectively) is the biggest contributor to the avoidance of the total tripeptide. For other tripeptides, neither dipeptide is particularly

avoided, yet the total tripeptide is greatly avoided. For example, TKH is much more avoided than either TK or KH, and likewise HHG is much more avoided than HH and HG. There are likely many more examples of epistasis that can be elucidated by such analysis. Maybe, when looking at tripeptides that are not avoided but whose dipeptide submotif are, we could elucidate what amino acids can rescue detrimental sequence patterns, or even permit peptide sequences to reach certain evolutionary paths. Further analysis in this direction is a future goal of this study.

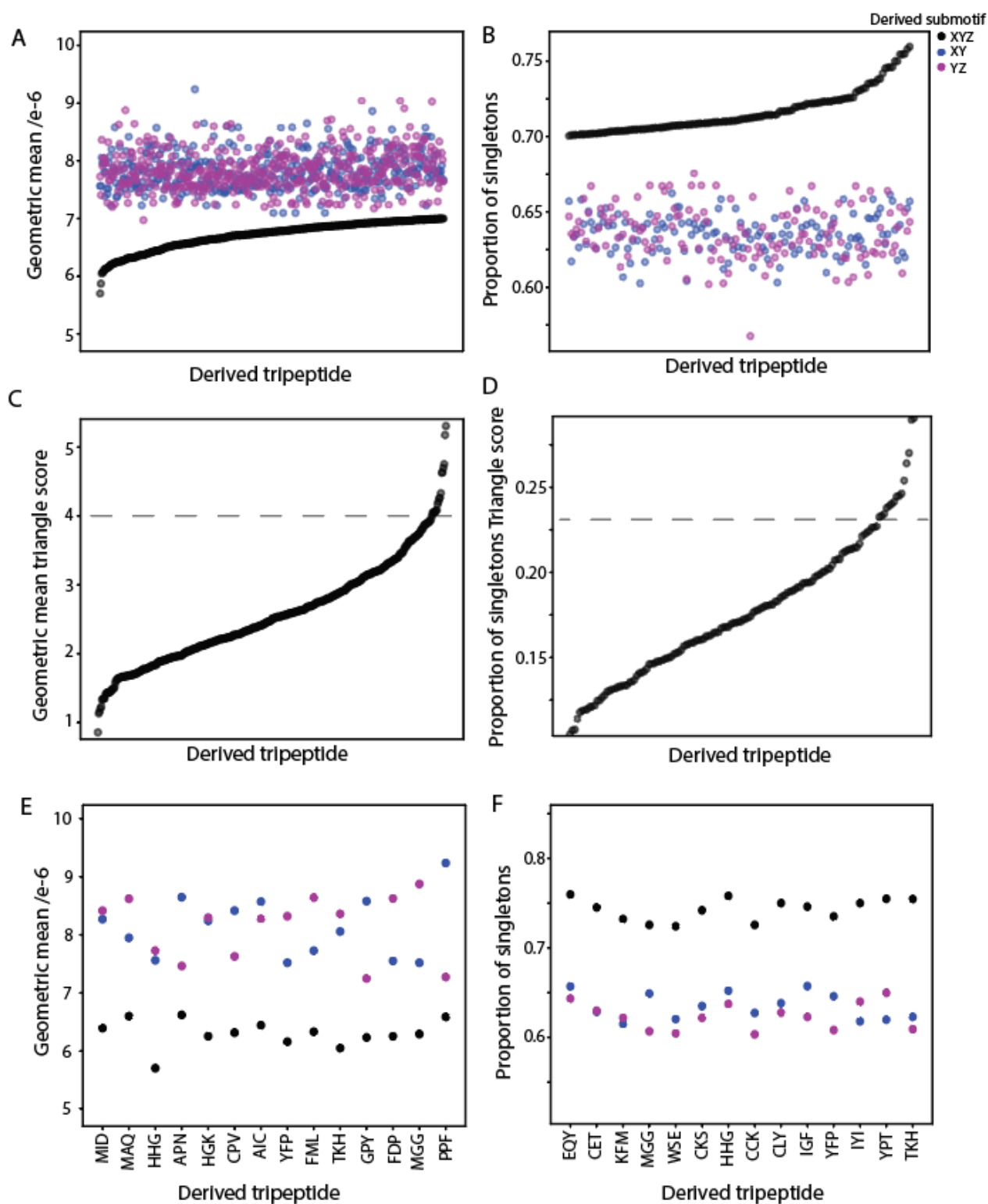


Figure 2.30: Scatterplots examining the difference in avoidance of tripeptides and their dipeptide submotifs. A,B) Scatterplots showing for each tripeptide (XYZ, black) the DMF of its first (XY, blue) and second (YZ, magenta) dipeptide submotifs. The most avoided tripeptides were chosen, with geometric mean $<7e-6$ and proportion of singletons >0.7 . C,D) For the same tripeptides in A,B the total difference between the DMFs of the tripeptide and its submotifs, sorted by this sum. The larger the sum, the more different the 3 DMF values are. E,F) Plots illustrating the DMF values for the tripeptide and dipeptide trios from C,D with the largest sum of differences ($>4e-6$ in C; >0.2 in D), in the same way as subplots A and B. All derived tripeptides and dipeptides datasets were processed in the same way, with all CpG to TpG, and ApT to GpT transitions in CAAT, CTAT, and CAT contexts excluded, and only motifs with greater than 100 sample size were used, and only SNPs in a nonsynonymous context were considered.

2.3 Conclusions

2.3.1 Future directions

The objective of this study was to identify specific motifs that are avoided in human and to unravel the biological basis for such avoidance. The first step of identifying the reason for avoidance of a motif is to discern what is the main unit or subunit of the motif that is driving the avoidance. Continuing on from the end of the results section, investigation of the avoidance of a motif compared to the avoidance of its submotif offers the clearest view on this question. If I see no avoidance for either submotif XY or YZ, but there is avoidance of XYZ, that is *prima facie* evidence for specific selection against XYZ. Likewise, if I see avoidance of one of the submotifs matching the avoidance of XYZ, then the submotif is the driver and there is less need to consider the longer motif. In doing so, it will be important to control for the prevalence of a particular amino acid substitution driving the overall avoidance of the total motif. In addition, there might be other scenarios where addition of an amino acid to an avoided motif might rescue its avoidance. Similarly, combining this with looking at the symmetry of the pattern's avoidance will offer insights whether the avoidance is due to the proximity of the amino acids to one another in sequence, or whether there is either a translational or fold-related reason for their avoidance. Therefore, future work will begin with further comparison of avoided motifs to both short comprising submotifs, and to longer motifs of which it is a part of.

Looking at the avoidance of a pattern's submotifs helps us identify which part of each avoided motif is truly driving the avoidance and allows us to more precisely interrogate the functionality of the motif. This clearly leads me to the next step of future work: asking what function specific avoided motifs contribute towards. To investigate whether we do capture such functionalities for the amino acid motifs, the first point of call would be to overlap the most avoided derived motifs with databases that are associated with protein functionality. These databases would include functional annotations, like those from UniProt [Consortium, 2019]. The UniProt knowledgebase contains protein functional information in the form of both syntax summaries,

or structured vocabularies such as the Gene Ontology (GO) system, providing many terms with which to try to find patterns in the functionalities of avoided motifs. Another method would be to probe where the avoided motifs commonly occur in protein structures. There is a plethora of protein structures available on the Protein DataBase (PDB, Berman et al. [2002]). Do the motifs occur in α -helices or β -sheets? Previous literature found that α -helices are more robust to mutations [Abrusán and Marsh, 2016], and I found a correlation between single amino acid avoidance and their α -helical propensity. Do they occur in amphipathic regions or parts that are deeply buried in the protein? I can further explore the cost of the derived motifs on structure by investigating protein fold prediction databases, such as PFD 2.0 [Fulton et al., 2007], SCOP [Murzin et al., 1995], or PFDB [Manavalan et al., 2019], to see whether certain avoided motifs are crucial for establishing particular folds. Finally, it might be interesting to determine whether avoided motifs often occur in the catalytic sites of proteins. UniProt has vocabulary for annotating catalytic residues, denoting whether a residue or motif is part of a charge relay system, an electrophile, nucleophile, proton donor, or proton acceptor. There is also the Genetic Active Site Search (GASS-WEB) web-server [Moraes et al., 2017] as well as the Catalytic Site Atlas [Furnham et al., 2014], which offer additional avenues for assigning active site residues and motifs according to evolutionary algorithms or from experimentally determined atomic structures respectively. It is worth noting that short 3-4 amino acid long motifs might not sufficiently recur in catalytic portions, so this latter idea is not a main line, but a complementary one, for future exploration.

2.3.2 Limits of the approach

Currently, the approach is limited to short, 3 to 4 amino acid long motifs due to sample size limitations and the weak strength of selection. However, this limit is a dynamic one, and as more genomes are sequenced and more mutations are characterised this approach will only become more powerful at discerning slight selective differences in longer length motifs.

Finally, it is important to note that amino acid usage and function is incredibly idiosyncratic, and this approach of creating frequency distributions will inevitably conflate amino acid motifs

occurring in different regions and functional categories of proteins. One possibility to separate these different contexts out would be to attempt to use mathematics to visualise whether a DMF of a protein is in fact a sum of different distributions, where each distribution is the motif occurring in a different context with a different evolutionary avoidance. However such an approach is currently outside my expertise. This might be achievable with various machine learning approaches and other decomposition methods. In addition, due to the heavy skew of our dataset towards singletons, it is already hard to describe these distributions using statistical models between the various amino acid motifs, let alone different evolutionary contexts within the same amino acid motif.

2.3.3 Conclusion

Here, I have demonstrated that by analyzing population genomic data one can make inferences about the evolutionary selective pressures avoiding particular motifs. I demonstrate that sites with sufficiently high mutation rate, specifically CpG to TpG transitions, break the infinite sites assumption and skew DMFs. In addition, sites that might have an incorrectly inferred ancestral state, due to either low sequence complexity or ApT transitions at CAAT, CTAT, or CAT sites, also have an effect on DMFs. Controlling for these confounders, I find that using DMFs can correctly return trends consistent with past literature. For example, the derived frequencies of nonsynonymous single amino acids negatively correlate with the steric size of the amino acid as well as correlate with the α -helical propensity of the amino acid. However, simply looking at the chemical properties does not fully account for the trends, implying the DMF approach is capturing more information. I then show preliminary evidence of extending the DMF approach to dipeptides and tripeptides. Comparison of both dipeptides and tripeptides to their comprising submotifs gleans more evidence for what part of the derived motif is avoided. Likewise, comparing the forward and reverse version of a derived motif showed asymmetric avoidance events, implying directionality is important in the appearance of certain motifs. Further interrogation of longer avoided derived motifs using these approaches will prove a powerful approach for elucidation of what mutations lead to a *gain of function*. The power of

this method will only be getting stronger as more sequencing and population variation data becomes available. In addition, the method is not limited to human, and can be used in any organism that has the necessary data available. The DMF method pipeline offers a simple interface to statistically interrogate which motifs are avoided in which contexts, and will provide great advancement to the screening of potentially functional motifs, and in understanding what their roles appear to be in biological systems.

2.4 Methods

2.4.1 Derived Motif Frequency Spectra

A *derived* allele is defined as a variant in the present population that is different to the ancestral allele. When extended to the neighbouring nucleotides of the variant, derived *motifs* can be analysed. For a given motif, for example ‘TCG’, I collect any SNP that leads to this motif’s appearance in the genome relative to the ancestral. The exact ancestral allele is not important, whether in this example it’s ‘ACG’ or ‘TCC’, if the SNP results in the derived motif ‘TCG’ existing in the population, those SNPs are used. The frequencies of each of these derived alleles can be pooled into a distribution called the *derived motif frequency specturm* (DMFs). By comparing these spectra between various derived motifs, the relative selective pressures can be inferred. For example, in schematic 2.1, the DMFs of the derived motif ‘TCG’ are shifted left towards the lower allele frequencies compared to the DMFs of ‘GTG’, implying that TCG is more avoided if it was to appear by chance in the genome compared to GTG. The most important advantage of this method compared to the previously used Karlin methods is that it only considers loci where a mutation has already occurred. This rules out differential mutational biases as the cause of motif depletion if the mutations do not break the infinite sites assumption. Only variants from the autosomes were used initially for method development and analysis, since sex chromosomes are under different selective regimes. We also applied a mappability genome mask [Consortium et al., 2010] using BedTools from the UCSC Toolkit to ensure the SNPs analysed are from confidently mapped regions of the human genome, in order

to reduce the chance of technical mapping errors skewing the analysis.

2.4.2 Cumulative Density Functions

One way to capture more information about a DMFs distribution rather than use a single point estimate such as the average would be to create a cumulative density function. This is essentially a vector of length M , where M is the maximum allele count in the database (see schematic in Figure 2.4). Since the maximum allele count in the gnomAD database is very large ($M = 251,596$) and is very sparse at the higher counts, I can shorten the vector by binning. Attempting to bin the cumulative density functions into 1000 equally sized bins led to only approximately 76 quantitatively different bins being created, due to the strong skew in the distributions towards singleton alleles and the very long and sparse tails. These vectors range from 0 to 1, where the value 1 entails that all derived SNPs for the particular derived motif exist at and below that allele count. These CDFs are not only good for visualisation, but there also exist proper statistical tools for comparing cumulative distributions.

2.4.3 Ancestral Genome

The ancestral human genome used was the one generated during the Thousand Genomes Project (1kG, Consortium et al. [2010]). It was generated using the Enredo-Pecan-Ortheus pipeline: Enredo is the program that is used to detect collinear segments; Ortheus is the an alignment modeler that infers the ancestral states; and the Pecan program generates the alignments. These ancestral calls were performed comparing six species of primates. For an example for how such an inference is made, let us take a random site X in the human genome. If this base is 'C' in human, and 'T' in chimp, one could go one step further and look at site X in macaque. If in macaque, site X is a 'C', then the human-chimp ancestral base was a 'C'. Next, the confidence in the ancestral call is given by how well the calls agree at different ancestral levels. For example, in using a human-chimp-macaque alignment to get the ancestral state of human, one would compare the ancestral base of the human-chimp ancestor, to the base in

chimp and to the base in the human-chimp-macaque ancestor. If all three sequences agree, the ancestral site can be called with high confidence. If only two calls agree, the site is a low-confidence call, and I have removed derived motifs in such locations for further analysis. The ancestral alignment is mapped to the human hg19 genome version.

2.4.4 Population Variation Data

The dataset used was the second version from the Genome Aggregation Database (gnomAD, Karczewski and Francioli [2017]). The v2 dataset contained SNPs from both 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals. For greater statistical power, all analysis was focused on the exome dataset. Again, only SNPs on the 22 autosomes were used, due to the sex chromosomes known to be under very different selective pressures and therefore should be processed separately. The *variant call format* (.vcf) files were used from <https://gnomad.broadinstitute.org/downloads/> and were already mapped to the hg19 human genome version. These vcf files are very large in memory due to containing large amount of metadata. I processed these files to only use SNPs, ignoring indels and multiple variants, as well as to reduce them to a more usable size. See file *processing_gnomad.vcf.py* on my Github linked below for further details.

2.4.5 Coding Sequences

To look at motifs at the amino acid level, the coding start and end sites for all human genes were downloaded from BioMart (Durinck et al. [2005], <https://m.ensembl.org/biomart/martview/>). For genes that coded for multiple transcripts, the transcript of longest length was used for consistency. These were converted from human genome version GRCh38 to hg19 using the UCSC LiftOver utility.

2.4.6 Genome versions

The version of the human genome used was the hg19 version [Consortium et al., 2001]. For any dataset not in this version, the UCSC LiftOver utility was used to convert the genomic coordinates onto this genome version [Kuhn et al., 2012].

2.4.7 Data availability statement

I confirm that all data analysed in this study is publicly available and links to the original sources are provided throughout. The code used to carry out all analysis is available on my Github page: https://github.com/vfkusmartsev/Derived_Motif_Frequencies. The datasets used are as follows:

- gnomAD `gs://gnomad-public/release/2.0.2/vcf/genomes/gnomad.genomes.r2.0.2.sites.chr*.vcf.bgz`
- human ancestral sequence `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/`
- human GRCh37 (hg19) genome version `ftp://ftp.ensembl.org/pub/grch37/current/fasta/homo_sapiens/dna/`
- BioMart coding sequences `https://m.ensembl.org/biomart/martview/`
- Amino Acid Index `ftp://ftp.genome.jp/pub/db/community/aaindex/`

Concluding remarks

Summary of Thesis Achievements

In the course of this thesis, I have conducted two separate but coherent studies where I investigated the biased generation of oligonucleotide variation around methylated cytosines and the biased avoidance of sequence motifs if these patterns were to arise by chance.

Specifically, in Chapter 1, I have demonstrated that cytosine methylation indeed has a broader mutational shadow cast beyond the base itself. In human, bases surrounding the 5mC exhibit reduced incidence of SNPs, in contrast to prior studies where the use of common SNPs lead to mischaracterisation of the most mutable methylated sites as unmethylated ones. In plants, I illustrate the opposite effect with an increase in mutations adjacently. I narrow down this effect to be most pronounced in regions least accessible to enzymatic methylation machinery and ones requiring the chromatin remodeler DDM1.

In Chapter 2, I develop and validate a novel approach to investigate the selection acting on the gain of mutations, by considering *derived motif frequencies* (DMFs) at sites where a mutation has already occurred since our primate ancestral genome. I discover, despite the method *a priori* accounting for variable mutation rate by focusing on sites where one has already occurred, very high mutation rates and incorrect inference of ancestral state can skew the results by breaking the identity-by-descent assumption. These biases include potential methylated CpG deaminations, ApT transitions in the sequence contexts CAAT, CTAT, and CAT, and ancestral sequence complexity. After noting and controlling for these biases, I validate the DMF approach in single amino acid, dipeptide, and tripeptide derived motifs, and illustrate the power of the

method for investigating fine-scale avoidance of patterns based on their chemical properties, asymmetrical appearance in protein sequence, and the epistatic effect of their submotifs.

Together, these Chapters elucidate interesting patterns biasing the creation and persistence of various oligonucleotide motifs, as well as highlighting the difficulty with investigating the entwined relationships of mutational bias, selective pressures, and technical challenges of sequence mapping and inference of ancestral genome states.

Future Work

The work has but scratched the surface of the species-specific importance of epigenetic markers on local mutation rate and the importance of controlling for varying mutation rate in understanding evolutionary trends.

From Chapter 1, there remains more work to understand the molecular mechanism driving the reduction in mutation rate adjacently to methylated cytosines. Is it due to more repair at 5mC lesions? Or is it due to greater error-prone repair at unmethylated lesions? And for plants, had they evolved an elevated mutation rate surrounding methylated CpG as an adaptive answer to the methylation of transposable elements? At what stage in the divergence of plants and animals did the epigenetics and evolvability of the epigenetics change?

In Chapter 2, I have demonstrated the effectiveness of using DMFs for understanding small scale avoidance of patterns. Now comes the extension of the method to separate out avoided patterns by their functional consequence in humans: Are certain families of genes under more selective pressures? Can we characterise the idiosyncrasy of pattern appearance in different functional contexts? Is it possible to predict the effects of longer, or unknown, sequence motifs based on products of their submotifs?

Bibliography

- G. Abrusán and J. A. Marsh. Alpha helices are more robust to mutations than beta strands. *PLoS computational biology*, 12(12):e1005242, 2016.
- I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.
- V. Aggarwala and B. F. Voight. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature genetics*, 48(4):349, 2016.
- A. Aguilera. The connection between transcription and genomic instability. *The EMBO journal*, 21(3):195–201, 2002.
- C. M. Akitake, M. Macurak, M. E. Halpern, and M. G. Goll. Transgenerational analysis of transcriptional silencing in zebrafish. *Developmental biology*, 352(2):191–201, 2011.
- L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.
- R. C. Allshire and H. D. Madhani. Ten principles of heterochromatin formation and function. *Nat. Rev. Mol. Cell Biol.*, 19(4):229–244, 2018. ISSN 1471-0072. doi: 10.1038/nrm.2017.119. URL <http://dx.doi.org/10.1038/nrm.2017.119><https://www.ncbi.nlm.nih.gov/pubmed/29235574><https://www.nature.com/articles/nrm.2017.119AllPapers/A/AllshireandMadhani2018-Tenprinciplesofheterochromatinformationandfunction.pdf>.

- N. Amaral, T. Ryu, X. Li, and I. Chiolo. Nuclear Dynamics of Heterochromatin Repair. *Trends Genet.*, 33(2):86–100, 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2016.12.004. URL <http://dx.doi.org/10.1016/j.tig.2016.12.004><https://www.ncbi.nlm.nih.gov/pubmed/28104289><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5285325>[https://linkinghub.elsevier.com/retrieve/pii/S0168-9525\(16\)30170-6](https://linkinghub.elsevier.com/retrieve/pii/S0168-9525(16)30170-6)<http://www.sciencedirect.com/science/article/p>.
- A. S. Balajee and V. A. Bohr. Genomic heterogeneity of nucleotide excision repair. *Gene*, 250(1-2):15–30, 2000.
- A. Banyasz, L. Esposito, T. Douki, M. Perron, C. Lepori, R. Improta, and D. Markovitsi. Effect of C5-Methylation of Cytosine on the UV-Induced Reactivity of Duplex DNA: Conformational and Electronic Factors. *J. Phys. Chem. B*, 120(18):4232–4242, 2016. ISSN 1520-6106. doi: 10.1021/acs.jpcc.6b03340. URL <http://dx.doi.org/10.1021/acs.jpcc.6b03340><https://www.ncbi.nlm.nih.gov/pubmed/27075054><https://dx.doi.org/10.1021/acs.jpcc.6b03340>AllPapers/B/Banyaszetal. 2016-EffectofC5-MethylationofCytosineontheUV-InducedReactivityofDuplexDNA-Confor.
- D. Barker, M. Schafer, and R. White. Restriction sites containing cpg show a higher frequency of polymorphism in human dna. *Cell*, 36(1):131–138, 1984.
- A. Beletskii and A. S. Bhagwat. Transcription-induced mutations: increase in c to t mutations in the nontranscribed strand during transcription in escherichia coli. *Proceedings of the National Academy of Sciences*, 93(24):13919–13924, 1996.
- S. E. Bennett, J.-S. Sung, and D. W. Mosbaugh. Fidelity of uracil-initiated base excision dna repair in dna polymerase β -proficient and-deficient mouse embryonic fibroblast cell extracts. *Journal of Biological Chemistry*, 276(45):42588–42600, 2001.
- H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

- T. Bestor and D. Bourc'his. Transposon silencing and imprint establishment in mammalian germ cells. In *Cold Spring Harbor symposia on quantitative biology*, volume 69, pages 381–388. Cold Spring Harbor Laboratory Press, 2004.
- A. J. Bewick and R. J. Schmitz. Gene body dna methylation in plants. *Current opinion in plant biology*, 36:103–110, 2017.
- A. J. Bewick, K. J. Vogel, A. J. Moore, and R. J. Schmitz. Evolution of dna methylation across insects. *Molecular biology and evolution*, 34(3):654–665, 2017.
- A. Bird. Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21, 2002.
- M. J. Blow, T. A. Clark, C. G. Daum, A. M. Deutschbauer, A. Fomenkov, R. Fries, J. Froula, D. D. Kang, R. R. Malmstrom, R. D. Morgan, et al. The epigenomic landscape of prokaryotes. *PLoS genetics*, 12(2), 2016.
- D. Bouyer, A. Kramdi, M. Kassam, M. Heese, A. Schnittger, F. Roudier, and V. Colot. DNA methylation dynamics during early plant life. *Genome Biol.*, 18(1):179, 2017. ISSN 1465-6906. doi: 10.1186/s13059-017-1313-0. URL <http://dx.doi.org/10.1186/s13059-017-1313-0><https://www.ncbi.nlm.nih.gov/pubmed/28942733><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5611644><https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1313-0><https://doi.org/10.1186/s13059-017-1313-0>
- V. Brendel, P. Bucher, I. R. Nourbakhsh, B. E. Blaisdell, and S. Karlin. Methods and algorithms for statistical analysis of protein sequences. *Proceedings of the National Academy of Sciences*, 89(6):2002–2006, 1992.
- K. W. Brookman, J. E. Lamerdin, M. P. Thelen, M. Hwang, J. T. Reardon, A. Sancar, Z.-Q. Zhou, C. A. Walter, C. N. Parris, and L. H. Thompson. Ercc4 (xpf) encodes a human nucleotide excision repair protein with eukaryotic recombination homologs. *Molecular and cellular biology*, 16(11):6553–6562, 1996.

- T. C. Brown and J. Jiricny. A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell*, 50(6):945–950, 1987.
- F. Buhr, S. Jha, M. Thommen, J. Mittelstaet, F. Kutz, H. Schwalbe, M. V. Rodnina, and A. A. Komar. Synonymous codons direct cotranslational folding toward different protein conformations. *Molecular cell*, 61(3):341–351, 2016.
- M. Bulmer. Neighboring base effects on substitution rates in pseudogenes. *Molecular biology and evolution*, 3(4):322–329, 1986.
- C. Burge, A. M. Campbell, and S. Karlin. Over-and under-representation of short oligonucleotides in dna sequences. *Proceedings of the National Academy of Sciences*, 89(4):1358–1362, 1992.
- V. J. Cannistraro and J.-S. Taylor. Acceleration of 5-methylcytosine deamination in cyclobutane dimers by g and its implications for uv-induced c-to-t mutation hotspots. *Journal of molecular biology*, 392(5):1145–1157, 2009.
- V. J. Cannistraro, S. Pondugula, Q. Song, and J.-S. Taylor. Rapid deamination of cyclobutane pyrimidine dimer photoproducts at tcg sites in a translationally and rotationally positioned nucleosome in vivo. *Journal of Biological Chemistry*, 290(44):26597–26609, 2015.
- A. V. Cano, H. Rozhoňová, A. Stoltzfus, D. M. McCandlish, and J. L. Payne. Mutation bias shapes the spectrum of adaptive substitutions. *bioRxiv*, 2021. doi: 10.1101/2021.04.14.438663.
- J. Carlson, A. E. Locke, M. Flickinger, M. Zawistowski, S. Levy, R. M. Myers, M. Boehnke, H. M. Kang, L. J. Scott, J. Z. Li, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *bioRxiv*, page 108290, 2017.
- C. Chen, H. Qi, Y. Shen, J. Pickrell, and M. Przeworski. Contrasting determinants of mutation rates in germline and soma. *Genetics*, 207(1):255–267, 2017.
- J. Chen, B. F. Miller, and A. V. Furano. Repair of naturally occurring mismatches can induce mutations in flanking dna. *Elife*, 3:e02001, 2014.

- X. Chen, Z. Chen, H. Chen, Z. Su, J. Yang, F. Lin, S. Shi, and X. He. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science*, 335(6073):1235–1238, 2012.
- K. Choi, X. Zhao, K. A. Kelly, O. Venn, J. D. Higgins, N. E. Yelina, T. J. Hardcastle, P. A. Ziolkowski, G. P. Copenhaver, F. C. H. Franklin, et al. Arabidopsis meiotic crossover hot spots overlap with h2a. z nucleosomes at gene promoters. *Nature genetics*, 45(11):1327–1336, 2013.
- S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–219, 2008. ISSN 0028-0836. doi: 10.1038/nature06745. URL <http://dx.doi.org/10.1038/nature06745><https://www.ncbi.nlm.nih.gov/pubmed/18278030><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2377394><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2377394/AllPapers/C/Cokusetal.2008-Shotgunbisulphitesequencingof>.
- C. K. Collings and J. N. Anderson. Links between dna methylation and nucleosome occupancy in the human genome. *Epigenetics & chromatin*, 10(1):18, 2017.
- . G. P. Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061, 2010.
- E. P. Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.
- I. H. Consortium et al. The international hapmap project. *Nature*, 426(6968):789, 2003.
- I. H. G. S. Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.
- U. Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.
- D. N. Cooper and M. Krawczak. Cytosine methylation and the fate of cpg dinucleotides in vertebrate genomes. *Human genetics*, 83(2):181–188, 1989.

- D. N. Cooper and H. Youssoufian. The cpg dinucleotide and human genetic disease. *Human genetics*, 78(2):151–155, 1988.
- D. Córdoba-Cañero, T. Morales-Ruiz, T. Roldán-Arjona, and R. R. Ariza. Single-nucleotide and long-patch base excision repair of dna damage in plants. *The Plant Journal*, 60(4):716–728, 2009.
- C. Coulondre, J. H. Miller, P. J. Farabaugh, and W. Gilbert. Molecular basis of base substitution hotspots in escherichia coli. *Nature*, 274(5673):775, 1978.
- P. Cubas, C. Vincent, and E. Coen. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, 401(6749):157, 1999.
- C. Cuzzo, A. Porcellini, T. Angrisano, A. Morano, B. Lee, A. Di Pardo, S. Messina, R. Iuliano, A. Fusco, M. R. Santillo, et al. Dna damage, homology-directed repair, and dna methylation. *PLoS Genet*, 3(7):e110, 2007.
- M. F. Denissenko, J. X. Chen, M.-s. Tang, and G. P. Pfeifer. Cytosine methylation determines hot spots of dna damage in the human p53 gene. *Proceedings of the National Academy of Sciences*, 94(8):3893–3898, 1997.
- G. Dianov, A. Price, and T. Lindahl. Generation of single-nucleotide repair patches following excision of uracil residues from dna. *Molecular and cellular biology*, 12(4):1605–1612, 1992.
- G. L. Dianov, K. M. Sleeth, I. I. Dianova, and S. L. Allinson. Repair of abasic sites in dna. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 531(1-2):157–163, 2003.
- J. A. Drake, C. Bird, J. Nemesh, D. J. Thomas, C. Newton-Cheh, A. Reymond, L. Excoffier, H. Attar, S. E. Antonarakis, E. T. Dermitzakis, et al. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature genetics*, 38(2):223, 2006.
- B. K. Duncan and J. H. Miller. Mutagenic deamination of cytosine residues in dna. *Nature*, 287(5782):560–561, 1980.

- L. Duret and N. Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10:285–311, 2009.
- S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. Biomat and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.
- M. Ehrlich, K. F. Norris, R. Y. Wang, K. C. Kuo, and C. W. Gehrke. Dna cytosine methylation and heat-induced deamination. *Bioscience reports*, 6(4):387–393, 1986.
- R. M. Erdmann and C. L. Picard. Rna-directed dna methylation. *PLoS genetics*, 16(10): e1009034, 2020.
- W.-H. Fang and P. Modrich. Human strand-specific mismatch repair occurs by a bidirectional mechanism similar to that of the bacterial reaction. *Journal of Biological Chemistry*, 268(16):11838–11844, 1993.
- K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel. The widespread impact of mammalian micrnas on mrna repression and evolution. *Science*, 310(5755):1817–1821, 2005.
- J. C. Fay, G. J. Wyckoff, and C.-I. Wu. Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234, 2001.
- S. Feng, S. J. Cokus, X. Zhang, P.-Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, et al. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, 107(19):8689–8694, 2010.
- W. Feng and S. D. Michaels. Accessing the Inaccessible: The Organization, Transcription, Replication, and Repair of Heterochromatin in Plants. *Annu. Rev. Genet.*, 49: 439–459, 2015. ISSN 0066-4197. doi: 10.1146/annurev-genet-112414-055048. URL <http://dx.doi.org/10.1146/annurev-genet-112414-055048><https://www.ncbi.nlm.nih.gov/pubmed/26631514><http://arjournals.annualreviews.org/doi/full/10.1146/>

annurev-genet-112414-055048?url{_}ver=Z39.88-2003{\&}rfr{_}id=ori:rid:
crossref.org{\&}rfr{_}dat=cr{_}pub{_%}3Dpubmedhtt.

- S. E. Flanagan, A.-M. Patch, and S. Ellard. Using sift and polyphen to predict loss-of-function and gain-of-function mutations. *Genetic testing and molecular biomarkers*, 14(4):533–537, 2010.
- P. Fortini and E. Dogliotti. Base damage and single-strand break repair: mechanisms and functional significance of short-and long-patch repair subpathways. *DNA repair*, 6(4):398–409, 2007.
- P. Fortini, E. Parlanti, O. M. Sidorkina, J. Laval, and E. Dogliotti. The type of dna glycosylase determines the base excision repair pathway in mammalian cells. *Journal of Biological Chemistry*, 274(21):15230–15236, 1999.
- A. Fortuny and S. E. Polo. The response to DNA damage in heterochromatin domains. *Chromosoma*, 2018. ISSN 0009-5915. doi: 10.1007/s00412-018-0669-6. URL <http://dx.doi.org/10.1007/s00412-018-0669-6><https://www.ncbi.nlm.nih.gov/pubmed/29594515><https://dx.doi.org/10.1007/s00412-018-0669-6>AllPapers/F/FortunyandPolo2018-TheresponsetoDNAdamageinheterochromatindomains.pdf.
- L. C. Francioli, P. P. Polak, A. Koren, A. Menelaou, S. Chun, I. Renkens, C. M. Van Duijn, M. Swertz, C. Wijmenga, G. Van Ommen, et al. Genome-wide patterns and properties of de novo mutations in humans. *Nature genetics*, 47(7):822, 2015.
- K. J. Fryxell and W.-J. Moon. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.*, 22(3):650–658, 2005. ISSN 0737-4038. doi: 10.1093/molbev/msi043. URL <http://dx.doi.org/10.1093/molbev/msi043><https://www.ncbi.nlm.nih.gov/pubmed/15537806><https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msi043>AllPapers/F/FryxellandMoon2005-CpGmutationratesinthehumangenomearehighlydependenton.
- Y. Fu, F. Ito, G. Zhang, B. Fernandez, H. Yang, and X. S. Chen. Dna cytosine and methyl-

- cytosine deamination by apobec3b: enhancing methylcytosine deamination by engineering apobec3b. *Biochemical Journal*, 471(1):25–35, 2015.
- K. F. Fulton, M. A. Bate, N. G. Faux, K. Mahmood, C. Betts, and A. M. Buckle. Protein folding database (pfd 2.0): an online environment for the international foldomics consortium. *Nucleic acids research*, 35(suppl_1):D304–D307, 2007.
- N. Furnham, G. L. Holliday, T. A. de Beer, J. O. Jacobsen, W. R. Pearson, and J. M. Thornton. The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic acids research*, 42(D1):D485–D489, 2014.
- A. Furrer and B. van Loon. Handling the 3-methylcytosine lesion by six human dna polymerases members of the b-, x-and y-families. *Nucleic acids research*, 42(1):553–566, 2014.
- A. Gabrielian and A. Bolshoy. Sequence complexity and dna curvature. *Computers & chemistry*, 23(3-4):263–274, 1999.
- N. Galtier, L. Duret, S. Glémin, and V. Ranwez. Gc-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics*, 25(1):1–5, 2009.
- M. G. Goll and T. H. Bestor. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.*, 74:481–514, 2005. ISSN 0066-4154. doi: 10.1146/annurev.biochem.74.010904.153721. URL <http://dx.doi.org/10.1146/annurev.biochem.74.010904.153721><https://www.ncbi.nlm.nih.gov/pubmed/15952895>http://arjournals.annualreviews.org/doi/full/10.1146/annurev.biochem.74.010904.153721?url={_}ver=Z39.88-2003{\\&rfr{_}id=ori:rid:crossref.org{\\&rfr{_}dat=cr{_}pub{\\}%}3D.
- A. Gonzalez-Perez, R. Sabarinathan, and N. Lopez-Bigas. Local determinants of the mutational landscape of the human genome. *Cell*, 177(1):101–114, 2019.
- I. Grin and A. A. Ishchenko. An interplay of the base excision repair and mismatch repair pathways in active dna demethylation. *Nucleic acids research*, 44(8):3713–3727, 2016.
- R. Gutierrez and T. R. O'Connor. Dna direct reversal repair and alkylating agent drug resistance. *Cancer Drug Resistance*, 4, 2021.

- M. W. Hahn, J. E. Stajich, and G. A. Wray. The effects of selection against spurious transcription factor binding sites. *Molecular biology and evolution*, 20(6):901–906, 2003.
- A. Harpak, A. Bhaskar, and J. K. Pritchard. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS genetics*, 12(12):e1006489, 2016.
- K. Hata, M. Okano, H. Lei, and E. Li. Dnmt3l cooperates with the dnmt3 family of de novo dna methyltransferases to establish maternal imprints in mice. *Development*, 129(8):1983–1993, 2002.
- M. L. Hegde, T. K. Hazra, and S. Mitra. Early steps in the dna base excision/single-strand interruption repair pathway in mammalian cells. *Cell research*, 18(1):27, 2008.
- J. W. K. Ho, Y. L. Jung, T. Liu, B. H. Alver, S. Lee, K. Ikegami, K.-A. Sohn, A. Minoda, M. Y. Tolstorukov, A. Appert, S. C. J. Parker, T. Gu, A. Kundaje, N. C. Riddle, E. Bishop, T. A. Egelhofer, S. S. Hu, A. A. Alekseyenko, A. Rechtsteiner, D. Asker, J. A. Belsky, S. K. Bowman, Q. B. Chen, R. A.-J. Chen, D. S. Day, Y. Dong, A. C. Dose, X. Duan, C. B. Epstein, S. Ercan, E. A. Feingold, F. Ferrari, J. M. Garrigues, N. Gehlenborg, P. J. Good, P. Haseley, D. He, M. Herrmann, M. M. Hoffman, T. E. Jeffers, P. V. Kharchenko, P. Kolasinska-Zwierz, C. V. Kotwaliwale, N. Kumar, S. A. Langley, E. N. Larschan, I. Latorre, M. W. Libbrecht, X. Lin, R. Park, M. J. Pazin, H. N. Pham, A. Plachetka, B. Qin, Y. B. Schwartz, N. Shores, P. Stempor, A. Vielle, C. Wang, C. M. Whittle, H. Xue, R. E. Kingston, J. H. Kim, B. E. Bernstein, A. F. Dernburg, V. Pirrotta, M. I. Kuroda, W. S. Noble, T. D. Tullius, M. Kellis, D. M. MacAlpine, S. Strome, S. C. R. Elgin, X. S. Liu, J. D. Lieb, J. Ahringer, G. H. Karpen, and P. J. Park. Comparative analysis of metazoan chromatin organization. *Nature*, 512(7515):449–452, 2014. ISSN 0028-0836. doi: 10.1038/nature13415. URL <http://dx.doi.org/10.1038/nature13415><https://www.ncbi.nlm.nih.gov/pubmed/25164756><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4227084><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4227084/>AllPapers/H/Hoetal.2014-Comparativeanalysisofmetazoanch.
- A. Hodgkinson and A. Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature reviews genetics*, 12(11):756, 2011.

- T.-F. Hsieh, C. A. Ibarra, P. Silva, A. Zemach, L. Eshed-Williams, R. L. Fischer, and D. Zilberman. Genome-wide demethylation of arabidopsis endosperm. *Science*, 324(5933):1451–1454, 2009.
- F. Huang and W. M. Nau. A conformational flexibility scale for amino acids in peptides. *Angewandte Chemie International Edition*, 42(20):2269–2272, 2003.
- Y. Ikeda and T. Nishimura. The Role of DNA Methylation in Transposable Element Silencing and Genomic Imprinting. In O. Pontes and H. Jin, editors, *Nuclear Functions in Plant Transcription, Signaling and Development*, pages 13–29. Springer New York, New York, NY, 2015. ISBN 9781493923861. doi: 10.1007/978-1-4939-2386-1_2. URL https://doi.org/10.1007/978-1-4939-2386-1_{_}2http://dx.doi.org/10.1007/978-1-4939-2386-1_{_}2AllPapers/I/IkedaandNishimura2015-TheRoleofDNAMethylationinTransposableElementSilencingandGenomic.pdf.
- N. M. Inamdar, X.-Y. Zhang, C. L. Brough, W. E. Gardiner, D. M. Bisaro, and M. Ehrlich. Transfection of heteroduplexes containing uracil · guanine or thymine · guanine mispairs into plant cells. *Plant molecular biology*, 20(1):123–131, 1992.
- S. Itzkovitz, E. Hodis, and E. Segal. Overlapping codes within protein-coding sequences. *Genome research*, 2010.
- L. Ji, S. M. Mathioni, S. Johnson, D. Tucker, A. J. Bewick, K. Do Kim, J. Daron, R. K. Slotkin, S. A. Jackson, W. A. Parrott, et al. Genome-wide reinforcement of dna methylation occurs during somatic embryogenesis in soybean. *The Plant Cell*, 31(10):2315–2331, 2019.
- M. Jones, R. Wagner, and M. Radman. Mismatch repair of deaminated 5-methyl-cytosine. *Journal of molecular biology*, 194(1):155–159, 1987.
- H. Jónsson, P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, E. Hjartarson, M. T. Hardarson, K. E. Hjorleifsson, H. P. Eggertsson, S. A. Gudjonsson, et al. Parental influence on human germline de novo mutations in 1,548 trios from iceland. *Nature*, 549(7673):519, 2017.

- K. Kamba, T. Nagata, and M. Katahira. Catalytic analysis of apobec3g involving real-time nmr spectroscopy reveals nucleic acid determinants for deamination. *PLoS One*, 10(4):e0124142, 2015.
- D. Kandel, Y. Matias, R. Unger, and P. Winkler. Shuffling biological sequences. *Discrete Applied Mathematics*, 71(1-3):171–185, 1996.
- K. Karczewski and L. Francioli. The genome aggregation database (gnomad). *MacArthur Lab*, 2017.
- K. J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D. M. Ruderfer, D. Kavanagh, T. Hamamsy, M. Lek, K. E. Samocha, B. B. Cummings, et al. The exac browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research*, 45(D1):D840–D845, 2016.
- Y. Kawahara, M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu, S. Zhou, et al. Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data. *Rice*, 6(1):4, 2013.
- T. Kawakatsu, J. R. Nery, R. Castanon, and J. R. Ecker. Dynamic dna methylation reconfiguration during seed development and germination. *Genome biology*, 18(1):171, 2017.
- S. Kawashima and M. Kanehisa. Aaindex: amino acid index database. *Nucleic acids research*, 28(1):374–374, 2000.
- M. Y. Kim and D. Zilberman. DNA methylation as a system of plant genomic immunity. *Trends Plant Sci.*, 19(5):320–326, 2014. ISSN 1360-1385. doi: 10.1016/j.tplants.2014.01.014. URL <http://dx.doi.org/10.1016/j.tplants.2014.01.014><https://www.ncbi.nlm.nih.gov/pubmed/24618094>[https://linkinghub.elsevier.com/retrieve/pii/S1360-1385\(14\)00029-6](https://linkinghub.elsevier.com/retrieve/pii/S1360-1385(14)00029-6)AllPapers/K/KimandZilberman2014-DNAmethylationasasystemofplantgenomicimmunity.pdf.
- N. Kim, A. L. Abdulovic, R. Gealy, M. J. Lippert, and S. Jinks-Robertson. Transcription-associated mutagenesis in yeast is directly proportional to the level of gene expression and influenced by the direction of dna replication. *DNA repair*, 6(9):1285–1296, 2007.

- M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.
- A. Klungland and T. Lindahl. Second pathway for completion of human dna base excision-repair: reconstitution with purified proteins and requirement for dnase iv (fen1). *The EMBO journal*, 16(11):3341–3348, 1997.
- B. Kramer, W. Kramer, and H.-J. Fritz. Different base/base mismatches are corrected with different efficiencies by the methyl-directed dna mismatch-repair system of e. coli. *Cell*, 38(3):879–887, 1984.
- H. E. Krokan and M. Bjørås. Base excision repair. *Cold Spring Harbor perspectives in biology*, 5(4):a012583, 2013.
- Y. Kubota, R. A. Nash, A. Klungland, P. Schär, D. Barnes, and T. Lindahl. Reconstitution of dna base excision-repair with purified human proteins: interaction between dna polymerase beta and the xrcc1 protein. *The EMBO journal*, 15(23):6662–6670, 1996.
- R. M. Kuhn, D. Haussler, and W. J. Kent. The ucsc genome browser and associated tools. *Briefings in bioinformatics*, 14(2):144–161, 2012.
- V. Kusmartsev, M. Drożdż, B. Schuster-Böckler, and T. Warnecke. Cytosine methylation affects the mutability of neighboring nucleotides in germline and soma. *Genetics*, 214(4):809–823, 2020.
- P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, et al. The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic acids research*, 40(D1):D1202–D1210, 2011.
- M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O’Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.
- J. L. Leroy, M. Kochoyan, T. Huynh-Dinh, and M. Guéron. Characterization of base-pair

- opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. *Journal of molecular biology*, 200(2):223–238, 1988.
- J.-Y. Li, J. Wang, and R. S. Zeigler. The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience*, 3(1):8, 2014.
- J.-Y. Lin, B. H. Le, M. Chen, K. F. Henry, J. Hur, T.-F. Hsieh, P.-Y. Chen, J. M. Pelletier, M. Pellegrini, R. L. Fischer, et al. Similarity between soybean and arabidopsis seed methylomes and loss of non-cg methylation does not affect seed development. *Proceedings of the National Academy of Sciences*, 114(45):E9730–E9739, 2017.
- K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476, 2011.
- R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315, 2009.
- Y. Liu, T. Tian, K. Zhang, Q. You, H. Yan, N. Zhao, X. Yi, W. Xu, and Z. Su. PCSD: a plant chromatin state database. *Nucleic Acids Research*, 46(D1):D1157–D1167, 10 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx919. URL <https://dx.doi.org/10.1093/nar/gkx919>.
- F. Lyko, S. Foret, R. Kucharski, S. Wolf, C. Falckenhayn, and R. Maleszka. The honey bee epigenomes: differential methylation of brain dna in queens and workers. *PLoS biology*, 8(11):e1000506, 2010.
- M. Lynch. Evolution of the mutation rate. *TRENDS in Genetics*, 26(8):345–352, 2010.
- M. Lynch, M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704, 2016.
- D. M. Lyons and P. J. O’Brien. Human base excision repair creates a bias toward- 1 frameshift mutations. *Journal of Biological Chemistry*, 285(33):25203–25212, 2010.

- K. D. Makova and R. C. Hardison. The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics*, 16(4):213, 2015.
- B. Manavalan, K. Kuwajima, and J. Lee. Pfdb: A standardized protein folding database with temperature correction. *Scientific reports*, 9(1):1–9, 2019.
- M. Marcinkowski, T. Pilżys, D. Garbicz, J. Steciuk, D. Zugaj, D. Mielecki, T. J. Sarnowski, and E. Grzesiuk. Human and arabidopsis alpha-ketoglutarate-dependent dioxygenase homolog proteins—new players in important regulatory processes. *IUBMB life*, 72(6):1126–1144, 2020.
- A. Marco. Selection against maternal microrna target sites in maternal transcripts. *G3: Genes, Genomes, Genetics*, 5(10):2199–2207, 2015.
- M. I. Martínez-Macías, D. Córdoba-Cañero, R. R. Ariza, and T. Roldán-Arjona. The dna repair protein xrcc1 functions in the plant dna demethylation pathway by stimulating cytosine methylation (5-mec) excision, gap tailoring, and dna ligation. *Journal of Biological Chemistry*, 288(8):5496–5505, 2013.
- G. Matassi, P. M. Sharp, and C. Gautier. Chromosomal location effects on gene sequence evolution in mammals. *Current Biology*, 9(15):786–791, 1999.
- I. Mathieson and D. Reich. Differences in the rare variant spectrum among human populations. *PLoS genetics*, 13(2):e1006581, 2017.
- A. Mazin and B. Vanyushin. Loss of cpg dinucleotides from dna. 3. methylation and evolution of histone genes. *MOLECULAR BIOLOGY*, 21(3):566–574, 1987.
- M. J. McDonald, W.-C. Wang, H.-D. Huang, and J.-Y. Leu. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol*, 9(6):e1000622, 2011.
- S. A. Medvedeva, A. Y. Panchin, A. V. Alexeevski, S. A. Spirin, and Y. V. Panchin. Comparative analysis of context-dependent mutagenesis using human and mouse models. *BioMed research international*, 2013, 2013.

- I. Mellon, V. A. Bohr, C. A. Smith, and P. C. Hanawalt. Preferential dna repair of an active gene in human cells. *Proceedings of the National Academy of Sciences*, 83(23):8878–8882, 1986.
- I. Mellon, G. Spivak, and P. C. Hanawalt. Selective removal of transcription-blocking dna damage from the transcribed strand of the mammalian dhfr gene. *Cell*, 51(2):241–249, 1987.
- M. Meselson, R. YuAN, and J. Heywood. Restriction and modification of dna. *Annual review of biochemistry*, 41(1):447–466, 1972.
- T. J. Meza, M. N. Moen, C. B. Vågbø, H. E. Krokan, A. Klungland, P. E. Grini, and P. Ø. Falnes. The dna dioxygenase alkbh2 protects arabidopsis thaliana against methylation damage. *Nucleic acids research*, 40(14):6620–6631, 2012.
- C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno. Steric constraints in model proteins. *Physical review letters*, 80(25):5683, 1998.
- M. Mirouze, M. Lieberman-Lazarovich, R. Aversano, E. Bucher, J. Nicolet, J. Reinders, and J. Paszkowski. Loss of dna methylation affects the recombination landscape in arabidopsis. *Proceedings of the National Academy of Sciences*, 109(15):5880–5885, 2012.
- A. Molaro, E. Hodges, F. Fang, Q. Song, W. R. McCombie, G. J. Hannon, and A. D. Smith. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, 146(6):1029–1041, 2011.
- J. G. Monroe, T. Srikant, P. Carbonell-Bejerano, M. Exposito-Alonso, M.-L. Weng, M. T. Rutter, C. B. Fenster, and D. Weigel. Mutation bias shapes gene evolution in arabidopsis thaliana. *BioRxiv*, 2020.
- J. P. Moraes, G. L. Pappa, D. E. Pires, and S. C. Izidoro. Gass-web: a web server for identifying enzyme active sites based on genetic algorithms. *Nucleic acids research*, 45(W1):W315–W319, 2017.
- K. V. Morris, S. W.-L. Chan, S. E. Jacobsen, and D. J. Looney. Small interfering rna-induced transcriptional gene silencing in human cells. *Science*, 305(5688):1289–1292, 2004.

- M. Morselli, W. A. Pastor, B. Montanini, K. Nee, R. Ferrari, K. Fu, G. Bonora, L. Rubbi, A. T. Clark, S. Ottonello, S. E. Jacobsen, and M. Pellegrini. In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. *Elife*, 4:e06205, 2015. ISSN 2050-084X. doi: 10.7554/eLife.06205. URL <http://dx.doi.org/10.7554/eLife.06205><https://www.ncbi.nlm.nih.gov/pubmed/25848745><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4412109><https://doi.org/10.7554/eLife.06205><http://dx.doi.org/10.7554/eLife.06205.001AllPapers/M/Morsellietal.2015-Invi>.
- A. I. Murchie and D. M. Lilley. Base methylation and local dna helix stability: effect on the kinetics of cruciform extrusion. *Journal of molecular biology*, 205(3):593–602, 1989.
- A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, et al. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995.
- S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, 2005.
- C. S. Nabel, H. Jia, Y. Ye, L. Shen, H. L. Goldschmidt, J. T. Stivers, Y. Zhang, and R. M. Kohli. Aid/apobec deaminases disfavor modified cytosines implicated in dna demethylation. *Nature chemical biology*, 8(9):751–758, 2012.
- P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001.
- T. T. Ngo, J. Yoo, Q. Dai, Q. Zhang, C. He, A. Aksimentiev, and T. Ha. Effects of cytosine modifications on dna flexibility and nucleosome mechanical stability. *Nature communications*, 7:10813, 2016.
- D. Nurminsky, D. De Aguiar, C. D. Bustamante, and D. L. Hartl. Chromosomal effects of rapid gene evolution in drosophila melanogaster. *Science*, 291(5501):128–130, 2001.
- T. Ohta. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*, 23(1):263–286, 1992.

- M. Okano, D. W. Bell, D. A. Haber, and E. Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.
- B. Paten, J. Herrero, S. Fitzgerald, K. Beal, P. Flicek, I. Holmes, and E. Birney. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research*, 18(11):1829–1843, 2008.
- J. Peña-Díaz, S. Bregenhorn, M. Ghodgaonkar, C. Follonier, M. Artola-Borán, D. Castor, M. Lopes, A. A. Sartori, and J. Jiricny. Noncanonical mismatch repair as a source of genomic instability in human cells. *Molecular cell*, 47(5):669–680, 2012.
- F. M. Piccolo and A. G. Fisher. Getting rid of dna methylation. *Trends in cell biology*, 24(2):136–143, 2014.
- O. Pich, F. Muiños, R. Sabarinathan, I. Reyes-Salazar, A. Gonzalez-Perez, and N. Lopez-Bigas. Somatic and germline mutation periodicity follow the orientation of the dna minor groove around nucleosomes. *Cell*, 175(4):1074–1087, 2018.
- R. C. Poulos, J. Olivier, and J. W. H. Wong. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic Acids Res.*, 45(13):7786–7795, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx463. URL <http://dx.doi.org/10.1093/nar/gkx463><https://www.ncbi.nlm.nih.gov/pubmed/28531315><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5737810><https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx463>AllPapers/P/Poulosetal.2017-Theinteractionbet.
- W. Qu, S.-I. Hashimoto, A. Shimada, Y. Nakatani, K. Ichikawa, T. L. Saito, K. Ogoshi, K. Matsushima, Y. Suzuki, S. Sugano, H. Takeda, and S. Morishita. Genome-wide genetic variations are highly correlated with proximal DNA methylation patterns. *Genome Res.*, 22(8):1419–1425, 2012. ISSN 1088-9051. doi: 10.1101/gr.140236.112. URL <http://dx.doi.org/10.1101/gr.140236.112><https://www.ncbi.nlm.nih.gov/pubmed/22689467><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409255><http://>

[//genome.cshlp.org/cgi/pmidlookup?view=long{\&}pmid=22689467](http://genome.cshlp.org/cgi/pmidlookup?view=long{\&}pmid=22689467)<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409255/>.

- P. D. Rabinowicz, L. E. Palmer, B. P. May, M. T. Hemann, S. W. Lowe, W. R. McCombie, and R. A. Martienssen. Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res.*, 13(12):2658–2664, 2003. ISSN 1088-9051. doi: 10.1101/gr.1784803. URL <http://dx.doi.org/10.1101/gr.1784803><https://www.ncbi.nlm.nih.gov/pubmed/14656970><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403807><http://genome.cshlp.org/cgi/pmidlookup?view=long{\&}pmid=14656970><http://genome.cshlp.org/content/13/12/2658.abstract><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409255/>.
- R. Rahbari, A. Wuster, S. J. Lindsay, R. J. Hardwick, L. B. Alexandrov, S. Al Turki, A. Dominiczak, A. Morris, D. Porteous, B. Smith, et al. Timing, rates and spectra of human germline mutation. *Nature genetics*, 48(2):126, 2016.
- A. Regev, M. J. Lamb, and E. Jablonka. The role of dna methylation in invertebrates: developmental regulation or genome defense? *Molecular Biology and Evolution*, 15(7):880–891, 1998.
- A. D. Riggs and G. P. Pfeifer. X-chromosome inactivation and cell memory. *Trends in Genetics*, 8(5):169–174, 1992.
- A. Robertson, A. Klungland, T. Rognes, and I. Leiros. Dna repair in mammalian cells. *Cellular and molecular life sciences*, 66(6):981–993, 2009.
- P. J. Rochette, S. Lacoste, J.-P. Therrien, N. Bastien, D. E. Brash, and R. Drouin. Influence of cytosine methylation on ultraviolet-induced cyclobutane pyrimidine dimer formation in genomic dna. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 665(1-2):7–13, 2009.
- S. Rošić, R. Amouroux, C. E. Requena, A. Gomes, M. Emperle, T. Beltran, J. K. Rane, S. Linnett, M. E. Selkirk, P. H. Schiffer, et al. Evolutionary analysis indicates that dna alkylation damage is a byproduct of cytosine dna methyltransferase activity. *Nature genetics*, 50(3):452–459, 2018.

- Z. E. Sauna, C. Kimchi-Sarfaty, S. V. Ambudkar, and M. M. Gottesman. The sounds of silence: synonymous mutations affect function. 2007.
- R. Savisaar and L. D. Hurst. Both maintenance and avoidance of rna-binding protein interactions constrain coding sequence evolution. *Molecular biology and evolution*, 34(5):1110–1126, 2017.
- R. Savva, K. McAuley-Hecht, T. Brown, and L. Pearl. The structural basis of specific base-excision repair by uracil–dna glycosylase. *Nature*, 373(6514):487, 1995.
- S. Schanz, D. Castor, F. Fischer, and J. Jiricny. Interference of mismatch and base excision repair during the processing of adjacent u/g mispairs may play a key role in somatic hypermutation. *Proceedings of the National Academy of Sciences*, 106(14):5593–5598, 2009.
- D. Schübeler. Function and information content of dna methylation. *Nature*, 517(7534):321–326, 2015.
- B. Schuster-Böckler and B. Lehner. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412):504–507, 2012. ISSN 0028-0836. doi: 10.1038/nature11273. URL <http://dx.doi.org/10.1038/nature11273><https://www.ncbi.nlm.nih.gov/pubmed/22820252><https://www.nature.com/articles/nature11273>AllPapers/S/Schuster-BöcklerandLehner2012-Chromatinorganizationisamajorinfluenceonregionalmuta
- L. Ségurel, M. J. Wyman, and M. Przeworski. Determinants of mutation rate variation in the human germline. *Annual review of genomics and human genetics*, 15:47–70, 2014.
- E. U. Selker and J. N. Stevens. Dna methylation at asymmetric sites is associated with numerous transition mutations. *Proceedings of the National Academy of Sciences*, 82(23):8114–8118, 1985.
- P. M. Severin, X. Zou, H. E. Gaub, and K. Schulten. Cytosine methylation alters dna mechanical properties. *Nucleic acids research*, 39(20):8740–8751, 2011.

- A. Sharonov, T. Gustavsson, S. Marguet, and D. Markovitsi. Photophysical properties of 5-methylcytidine. *Photochemical & Photobiological Sciences*, 2(4):362–364, 2003.
- Y. Shimooka, J.-i. Nishikawa, and T. Ohyama. Most methylation-susceptible dna sequences in human embryonic stem cells undergo a change in conformation or flexibility upon methylation. *Biochemistry*, 52(8):1344–1353, 2013.
- M. I. Sigurdsson, A. V. Smith, H. T. Bjornsson, and J. J. Jonsson. Hapmap methylation-associated snps, markers of germline dna methylation, positively correlate with regional levels of human meiotic recombination. *Genome research*, 19(4):581–589, 2009.
- R. K. Singhal, R. Prasad, and S. H. Wilson. Dna polymerase conducts the gap-filling step in uracil-initiated base excision repair in a bovine testis nuclear extract. *Journal of Biological Chemistry*, 270(2):949–957, 1995.
- N. Siva. 1000 genomes project, 2008.
- A. Smit, R. Hubley, and P. Green. Repeatmasker open-4.0. 2013–2015, 2015.
- Q. Song, V. J. Cannistraro, and J.-S. Taylor. Rotational position of a 5-methylcytosine-containing cyclobutane pyrimidine dimer in a nucleosome greatly affects its deamination rate. *Journal of Biological Chemistry*, 286(8):6329–6335, 2011.
- J. A. Stamatoyannopoulos, I. Adzhubei, R. E. Thurman, G. V. Kryukov, S. M. Mirkin, and S. R. Sunyaev. Human mutation rate associated with dna replication timing. *Nature genetics*, 41(4):393, 2009.
- H. Stroud, B. Ding, S. A. Simon, S. Feng, M. Bellizzi, M. Pellegrini, G.-L. Wang, B. C. Meyers, and S. E. Jacobsen. Plants regenerated from tissue culture contain stable epigenome changes in rice. *elife*, 2:e00354, 2013.
- W. Sung, M. S. Ackerman, J.-F. Gout, S. F. Miller, E. Williams, P. L. Foster, and M. Lynch. Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation-Accumulation Experiments. *Mol. Biol. Evol.*, 32(7):1672–1683, 2015. ISSN 0737-4038. doi: 10.1093/molbev/msv055. URL <http://www.pubmedcentral.nih>.

gov/articlerender.fcgi?artid=PMC4476155<http://dx.doi.org/10.1093/molbev/msv055><https://www.ncbi.nlm.nih.gov/pubmed/25750180>AllPapers/S/Sungetal.
2015-AsymmetricContext-DependentMutationPatternsRevealedthroughMutat.

F. Supek and B. Lehner. Differential dna mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521(7550):81–84, 2015.

F. Supek, B. Lehner, P. Hajkova, and T. Warnecke. Hydroxymethylated Cytosines Are Associated with Elevated C to G Transversion Rates. *PLoS Genet.*, 2014. doi: 10.1371/journal.pgen.1004585. URL <http://dx.doi.org/10.1371/journal.pgen.1004585AllPapers/S/Supeketal>.
2014-HydroxymethylatedCytosinesAreAssociatedwithElevatedCtoGTransversionRates.pdf.

M. M. Suzuki and A. Bird. Dna methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476, 2008.

J. Sved and A. Bird. The expected equilibrium of the cpg dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences*, 87(12):4692–4696, 1990.

S. Takuno and B. S. Gaut. Body-methylated genes in arabidopsis thaliana are functionally important and evolve slowly. *Molecular Biology and Evolution*, 29(1):219–227, 2011.

S. Takuno and B. S. Gaut. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences*, 110(5):1797–1802, 2013.

K. Tang, Z. Lang, H. Zhang, and J.-K. Zhu. The dna demethylase ros1 targets genomic regions with distinct chromatin modifications. *Nature plants*, 2(11):16169, 2016.

J. A. Tennessen, A. W. Bigham, T. D. O’Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *science*, 337(6090):64–69, 2012.

- M. Tomkova and B. Schuster-Böckler. DNA Modifications: Naturally More Error Prone? *Trends Genet.*, 2018. ISSN 0168-9525. doi: 10.1016/j.tig.2018.04.005. URL <http://dx.doi.org/10.1016/j.tig.2018.04.005><https://www.ncbi.nlm.nih.gov/pubmed/29853204>[https://linkinghub.elsevier.com/retrieve/pii/S0168-9525\(18\)30081-7](https://linkinghub.elsevier.com/retrieve/pii/S0168-9525(18)30081-7)AllPapers/T/TomkovaandSchuster-Böckler2018-DNAModifications-NaturallyMoreErrorProne.pdf.
- M. Tomkova, M. McClellan, S. Kriaucionis, and B. Schuster-Böckler. Dna replication and associated repair pathways are involved in the mutagenesis of methylated cytosine. *DNA repair*, 62:1–7, 2018.
- E. N. Trifonov. Making sense of the human genome. *Structure and methods: proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics held at the State University of New York at Albany, June 6-10, 1989*/edited by RH Sarma & MH Sarma, 1990.
- S. U. Umu, A. M. Poole, R. C. Dobson, and P. P. Gardner. Avoidance of stochastic rna interactions can be harnessed to control protein expression levels in bacteria and archaea. *Elife*, 5, 2016.
- B. F. Vanyushin. DNA methylation in plants. *Curr. Top. Microbiol. Immunol.*, 301:67–122, 2006. ISSN 0070-217X. URL <https://www.ncbi.nlm.nih.gov/pubmed/16570846>.
- R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, and P. C. Ng. Sift missense predictions for genomes. *Nature protocols*, 11(1):1, 2016.
- M. S. Venkatarajan and W. Braun. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. *Molecular modeling annual*, 7(12):445–453, 2001.
- B. C. Verrelli, C. M. Lewis Jr, A. C. Stone, and G. H. Perry. Different selective pressures shape the molecular evolution of color vision in chimpanzee and human populations. *Molecular biology and evolution*, 25(12):2735–2743, 2008.

- F. Vogel and M. Kopun. Higher frequencies of transitions among point mutations. *Journal of molecular evolution*, 9(2):159–180, 1977.
- J. Wakeley, R. Nielsen, S. N. Liu-Cordero, and K. Ardlie. The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *The American Journal of Human Genetics*, 69(6):1332–1347, 2001.
- J.-C. Walser and A. V. Furano. The mutational spectrum of non-cpg dna varies with cpg content. *Genome research*, 20(7):875–882, 2010.
- J.-C. Walser, L. Ponger, and A. V. Furano. Cpg dinucleotides and the mutation rate of non-cpg dna. *Genome research*, 18(9):1403–1414, 2008.
- I. M. Walsh, M. A. Bowman, I. F. S. Santarriaga, A. Rodriguez, and P. L. Clark. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proceedings of the National Academy of Sciences*, 117(7):3528–3534, 2020.
- R. Y.-H. Wang, K. C. Kuo, C. W. Gehrke, L.-H. Huang, and M. Ehrlich. Heat-and alkali-induced deamination of 5-methylcytosine and cytosine residues in dna. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 697(3):371–377, 1982.
- M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schuebeler. Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. *Nature genetics*, 37(8):853–862, 2005.
- D. Weigel and R. Mott. The 1001 genomes project for arabidopsis thaliana. *Genome biology*, 10(5):107, 2009.
- T. S. Whittam, A. G. Clark, M. Stoneking, R. L. Cann, and A. C. Wilson. Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphism. *Proceedings of the National Academy of Sciences*, 83(24):9611–9615, 1986.
- A. T. Wierzbicki, R. Cocklin, A. Mayampurath, R. Lister, M. J. Rowley, B. D. Gregory, J. R. Ecker, H. Tang, and C. S. Pikaard. Spatial and functional relationships among

pol v-associated loci, pol iv-dependent sirnas, and cytosine methylation in the arabidopsis epigenome. *Genes & development*, 2012.

X. Wu and Y. Zhang. Tet-mediated active dna demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 18(9):517, 2017.

F. Zakrzewski, M. Schmidt, M. Van Lijsebettens, and T. Schmidt. DNA methylation of retrotransposons, DNA transposons and genes in sugar beet (*Beta vulgaris* L.). *Plant J.*, 90(6):1156–1175, 2017. ISSN 0960-7412. doi: 10.1111/tpj.13526. URL <http://dx.doi.org/10.1111/tpj.13526><https://www.ncbi.nlm.nih.gov/pubmed/28257158><http://doi.wiley.com/10.1111/tpj.13526><http://onlinelibrary.wiley.com/wol1/doi/10.1111/tpj.13526/fullpdf><https://api.wiley.com/onlinelibrary/tdm/v1/articles/10.1111{\%}2Ftpj.135>.

N. Zamudio, J. Barau, A. Teissandier, M. Walter, M. Borsos, N. Servant, and D. Bourc’his. DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes Dev.*, 29(12):1256–1270, 2015a. ISSN 0890-9369. doi: 10.1101/gad.257840.114. URL <http://dx.doi.org/10.1101/gad.257840.114><https://www.ncbi.nlm.nih.gov/pubmed/26109049><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4495397><http://www.genesdev.org/cgi/pmidlookup?view=long{\&}pmid=26109049><http://genesdev.cshlp.org/content/29/12/1256.abstract>.

N. Zamudio, J. Barau, A. Teissandier, M. Walter, M. Borsos, N. Servant, and D. Bourc’his. Dna methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes & development*, 29(12):1256–1270, 2015b.

A. Zemach, M. Y. Kim, P. Silva, J. A. Rodrigues, B. Dotson, M. D. Brooks, and D. Zilberman. Local DNA hypomethylation activates genes in rice endosperm. *Proc. Natl. Acad. Sci. U. S. A.*, 107(43):18729–18734, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1009695107. URL <http://dx.doi.org/10.1073/pnas.1009695107><https://www.ncbi.nlm.nih.gov/pubmed/20937895><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2972920><http://www.pnas.org/cgi/pmidlookup?view=long{\&}pmid=20937895><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2972920/>.

- A. Zemach, M. Y. Kim, P.-H. Hsieh, D. Coleman-Derr, L. Eshed-Williams, K. Thao, S. L. Harmer, and D. Zilberman. The arabidopsis nucleosome remodeler ddm1 allows dna methyltransferases to access h1-containing heterochromatin. *Cell*, 153(1):193–205, 2013.
- J. Zeng and S. V. Yi. Specific modifications of histone tails, but not dna methylation, mirror the temporal variation of mammalian recombination hotspots. *Genome biology and evolution*, 6(10):2918–2929, 2014.
- H. Zhang, Z. Lang, and J.-K. Zhu. Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.*, 19(8):489–506, 2018. ISSN 1471-0072. doi: 10.1038/s41580-018-0016-z. URL <http://dx.doi.org/10.1038/s41580-018-0016-z><https://www.ncbi.nlm.nih.gov/pubmed/29784956>AllPapers/Z/Zhangetal.2018-DynamicsandfunctionofDNAmethylationinplants.pdf.
- Z. Zhao and C. Jiang. Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions. *Mol. Biol. Evol.*, 24(1):23–25, 2007. ISSN 0737-4038. doi: 10.1093/molbev/msl156. URL <http://dx.doi.org/10.1093/molbev/msl156><https://www.ncbi.nlm.nih.gov/pubmed/17056644><https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msl156><https://academic.oup.com/mbe/article/24/1/23/1074623>AllPapers/Z/ZhaoandJiang2007-Methylation-.
- J.-K. Zhu. Active dna demethylation mediated by dna glycosylases. *Annual review of genetics*, 43:143–166, 2009.