Mapping the Landscape of Mutation Rate Heterogeneity in the Human Genome: Approaches and Applications

By

Jedidiah E. Carlson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2018

Doctoral Committee:

Professor Jun Z. Li, Co-Chair
Professor Sebastian K. Zöllner, Co-Chair
Professor Gonçalo Abecasis
Assistant Professor Saher Sue Hammoud
Associate Professor Hyun Min Kang

Jedidiah E. Carlson

jedidiah@umich.edu

ORCID iD: 0000-0002-1363-872X

This dissertation is dedicated in loving memory to Dr. Adam Johnson, who taught me how to swear like a scientist.

ACKNOWLEDGMENTS

Human population genetics is the study of where we came from and how we got here, so it is only fitting for me to acknowledge the numerous friends, family members, collaborators, and mentors whose support has made this journey possible. I am honored to have spent the last five years under the mentorship of my co-advisers, Dr. Sebastian Zöllner and Dr. Jun Li. Your enthusiastic collaboration was the spark that ignited my deep interest in mutation rate research, and you have consistently offered a wealth of wisdom, patience, and motivation throughout my graduate studies. I also wish to thank my committee members, Dr. Gonçalo Abecasis, Dr. Sue Hammoud, and Dr. Hyun Min Kang, for your detailed and perceptive feedback on my research, and the unique perspectives they bring to the field of human genetics. In addition, the members of the BRIDGES and TOPMed consortia have played an integral role in the work presented here, both through their generous provision of data and in their conceptual contributions to Chapters 2, 3, and 5. I must also express my deep appreciation to Dr. Michael Boehnke and the Genome Science Training Program for generously supporting me through the first three years of graduate school. This funding opportunity played a major role in my decision to attend the University of Michigan and allowed me the intellectual freedom to pursue many of the research questions that eventually evolved into this dissertation.

I will not venture to thank by name the many friends who have, each in their own unique way, made the last five years not only bearable, but enjoyable. It is not for lack of appreciation, but rather the fact that population geneticists are notorious for miscalculating important things by

factors of two, so I would undoubtedly manage to overlook at least half of the people whose friendship has motivated and inspired me to make it to this point. To the past and present members of the Zöllner and Li labs: thank you for the many engaging conversations about science, sharing numerous of time-saving tips and tricks, never hesitating to criticize my ideas, and supplying all sorts of globally-sourced snacks in lab meetings. It has been so much fun to be immersed in an environment full of deeply intelligent peers who support and strengthen one another, and I look forward to our paths crossing again. To my friends in the 2013 Biostatistics cohort (and affiliates thereof): I never could have imagined that I would so quickly find a close-knit community of friends in graduate school, but here we are, five years later, and you feel like family. I will cherish all the memories of horrible stats pun parties (the puns were horrible, not the parties), nights crammed around a table at Bill's, attending various weddings (four and counting!), and basking in the reflected glory of your academic and personal success. I tell myself that I will try to forget all the memories involving our collective stress and anxiety, but in all honesty, I will probably cherish those, too, simply because you were a part of them. To my life-long friends who have been around since before graduate school: I am so grateful that we have managed to remain close, even several states away. Thank you for always making the effort to see me whenever and wherever we get the chance.

Finally, I wish to thank my family for the integral role they have played in both my formal and informal education. To my parents, Ken and Carolyn Carlson, and my sister, Dr. Ingrid Wurpts: you are truly the first scientists to ever collaborate with me. Whether we are cooking, exploring nature, or discussing books together, you have always inspired my curiosity and excitement for discovery. Your resilience, patience, and kindness are qualities that I hope to continually emulate.

Most of all, thanks to my beloved wife, Dr. Emma Beyers-Carlson, for journeying through graduate school and beyond with me. The last five years have had many moments of doubt and uncertainty, but you have been (and continue to be) unwaveringly constant and supportive, even while enduring the ups and downs of your own doctoral studies. I am in awe of your poise and determination in the face of challenges, your generous and selfless spirit, and your commitment to leaving the world better than you found it. Thank you for making sure that every single day is punctuated with laughter, good food, and the motivation to achieve what we set out to accomplish—I am so indescribably grateful for all that you are.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

All heritable genetic variation is ultimately the result of mutations that have occurred in the past. Understanding the processes which determine the rate and spectra of new mutations is therefore fundamentally important in efforts to characterize the genetic basis of heritable disease, infer the timing and extent of past demographic events (e.g., population expansion, migration), or identify signals of natural selection. This dissertation aims to describe patterns of mutation rate heterogeneity in detail, identify factors contributing to this heterogeneity, and develop methods and tools to harness such knowledge for more effective and efficient analysis of whole-genome sequencing data.

In Chapters 2 and 3, we catalog granular patterns of germline mutation rate heterogeneity throughout the human genome by analyzing extremely rare variants ascertained from large-scale whole-genome sequencing datasets. In Chapter 2, we describe how mutation rates are influenced by local sequence context and various features of the genomic landscape (e.g., histone marks, recombination rate, replication timing), providing detailed insight into the determinants of single-nucleotide mutation rate variation. We show that these estimates reflect genuine patterns of variation among de novo mutations, with broad potential for improving our understanding of the biology of underlying mutation processes and the consequences for human health and evolution. These estimated rates are publicly available at http://mutation.sph.umich.edu/.

In Chapter 3, we introduce a novel statistical model to elucidate the variation in rate and spectra of multinucleotide mutations throughout the genome. We catalog two major classes of multinucleotide mutations: those resulting from error-prone translesion synthesis, and those resulting from repair of double-strand breaks. In addition, we identify specific hotspots for these unique mutation classes and describe the genomic features associated with their spatial variation. We show how these multinucleotide mutation processes, along with sample demography and mutation rate heterogeneity, contribute to the overall patterns of clustered variation throughout the genome, promoting a more holistic approach to interpreting the source of these patterns.

In chapter 4, we develop Helmsman, a computationally efficient software tool to infer mutational signatures in large samples of cancer genomes. By incorporating parallelization routines and efficient programming techniques, Helmsman performs this task up to 300 times faster and with a memory footprint 100 times smaller than existing mutation signature analysis software. Moreover, Helmsman is the only such program capable of directly analyzing arbitrarily large datasets. The Helmsman software can be accessed at https://github.com/carjed/helmsman.

Finally, in Chapter 5, we present a new method for quality control in large-scale whole-genome sequencing datasets, using a combination of dimensionality reduction algorithms and unsupervised anomaly detection techniques. Just as the mutation spectrum can be used to infer the presence of underlying mechanisms, we show that the spectrum of rare variation is a powerful and informative indicator of sample sequencing quality. Analyzing three large-scale datasets, we demonstrate that our method is capable of identifying samples affected by a variety of technical artifacts that would otherwise go undetected by standard ad hoc filtering criteria. We have implemented this method in a software package, Doomsayer, available at https://github.com/carjed/doomsayer.

Chapter I.

Introduction

In the early 20th century, pioneers in the field of genetics synthesized Gregor Mendel's laws of genetic inheritance and Charles Darwin's theory of evolution into a unified theoretical framework, commonly known as the modern evolutionary synthesis [1]. The modern evolutionary synthesis postulates that genetic variation in the population—and, over a longer timespan, gradual evolutionary change—fundamentally originates with the acquisition of new heritable mutations. Hence, a deep understanding of the processes which generate these mutations (and the patterns of variation in rate and spectra thereof) is indispensably important in efforts to characterize the genetic basis of heritable disease and phenotypic variation [2], infer the timing and extent of past demographic events such as population expansion and migration [3], identify signals of natural selection [4], and numerous other active areas of research in the field of genomics.

Many distinct endogenous and environmental sources of mutation have been documented (as reviewed in [5]). Notable examples include a propensity for C>T mutations at CpG dinucleotides due to spontaneous deamination at methylated cytosines [6, 7], G>T mutations resulting from oxidative damage of guanine [8], C>T mutations resulting from exposure to UV radiation [9], and DNA mispairing during replication [10]. Many of the context-dependent

effects that have been observed, however, have yet to be linked to specific mechanisms [5, 11–13]. More fundamentally, the full extent of variation in mutation rates has not been exhaustively characterized, so our knowledge of the factors that contribute to the mutational landscape of the human genome remains rudimentary.

Despite the field's long-standing appreciation for the diversity and regional variation of mutation processes, it is only within the last decade, with the widespread availability of whole-genome sequencing data and increasingly powerful computational resources, that cataloging the fine-scale variation in mutation patterns (and understanding the implications for human health and evolution) has become possible [12–14]. This dissertation aims to describe patterns of mutation rate heterogeneity in detail, identify factors contributing to this heterogeneity, and develop methods and tools to harness this knowledge for more effective and efficient analysis of whole-genome sequencing data.

In Chapter 2, we present a detailed account of factors contributing to fine-scale patterns of variation in single-nucleotide mutation rates throughout the genome. This study leverages a powerful new approach to studying mutation patterns: rather than relying on *de novo* mutations ascertained from trio sequencing data (which are too sparse to precisely quantify granular mutation patterns), or common variants from unrelated individuals (which are affected by natural selection and biased gene conversion, processes that are difficult to disentangle from the underlying mutation signal), we analyze a large collection of extremely rare variants (ERVs) ascertained from a large-scale whole-genome sequencing dataset. These ERVs have the benefit of being abundant throughout the genome—in our dataset, we observe nearly 36 million ERVs in 3,560 unrelated individuals. Further, theoretical models suggest that nearly all ERVs have arisen very recently in human history, thus their distribution is unlikely to have been strongly affected

by natural selection or biased gene conversion, so, in a sufficiently large sample, they accurately represent mutation events that have occurred within the last several dozen generations [15].

These data enable us to estimate how the mutation rate varies with respect to the local sequence context, considering up to 3 bases upstream and 3 bases downstream from the mutation site. We observe a remarkable heterogeneity in the mutation rates of these 7-mer sequence motifs, demonstrating that the broader sequence context at any given site is a key determinant of that site's mutability. This has important implications for how we understand the specificity of damage and repair mechanisms in the genome and how these processes may have evolved over time.

We then implement a series of regression models to show that the mutation rate at any given sequence motif is further influenced by various features of the genomic landscape (e.g., histone marks, recombination, replication timing), providing further insight into the determinants of mutation rate variation. Though previous studies have implied that the presence of a given feature will tend to result in a general increase or decrease in mutation rates (e.g., [13, 14]), we show that the mutagenic effect of a feature can in fact vary according to the sequence motif, suggesting that the mutational efficiencies of damage and repair mechanisms are influenced by nuanced characteristics of the genomic landscape.

Finally, we present conclusive evidence that our ERV-derived mutation rate estimates are consistently more accurate at describing bona fide *de novo* mutation patterns than estimates derived from more common variants. This result underscores the benefits of using ERVs to understand ongoing mutation patterns in the human population and lays the groundwork for future tangential research questions concerning how the subsequent processes of natural

selection, biased gene conversion, and mutation rate evolution are also impacted by variation in local sequence context and genomic features.

In Chapter 3, we elucidate the variation in rate and spectra of multinucleotide mutations (MNMs) throughout the genome, again leveraging rare singleton SNVs from whole-genome sequencing data as a proxy for recent mutation events. MNMs are essentially defined as two or more closely-spaced point mutations that occur simultaneously as part of a single mutation event [16], and are important to consider as a distinct class of mutations for several reasons. First, MNMs likely arise via a handful of specific mutational pathways and thus represent a unique outcome of genome instability [16]; second, although MNMs might superficially be considered indistinguishable from multiple closely-spaced single-nucleotide mutations arising through independent mutation events, MNMs have a distinct transversion-rich mutation signature [17], which increases their likelihood of pathogenic outcomes, making them uniquely implicated in the genetic architecture of human disease [18]; third, because MNMs generate multiple simultaneous changes to the genome, they may lead to more rapid evolutionary changes [19]; finally, failure to account for the spatially non-independent nature of MNMs can lead to false inference of signals of selection or past demographic events [20, 21].

We propose a novel statistical approach to infer the patterns and properties of MNMs in the human germline. Specifically, we model the spatial distribution of ERVs throughout the genome as a mixture of exponential processes, where each process is assumed to generate a subset of mutations occurring at a unique range of spatial proximities. We show that the properties of mutations attributable to two particular processes inferred through this statistical model are consistent with two well-known multinucleotide mutation processes: error-prone translesion synthesis (TLS) and repair of double-strand breaks (DSB) [21–23]. We provide a

detailed account of how MNMs resulting from these processes vary in their intrinsic properties and throughout the genome. This analysis shows that the spectrum of TLS-associated multinucleotide mutations varies dramatically as the inter-mutation distance increases, potentially indicating distinctive signatures of specific translesion polymerases. In addition, we describe how these TLS- and DSB-associated multinucleotide mutations vary throughout the genome and provide evidence that particular genomic features are associated with these patterns of regional variation. We conclude this chapter by exploring how the spatial clustering of rare variants is jointly affected by multinucleotide mutations, regional mutation rate heterogeneity, sample demography, and other factors, thereby demonstrating that a holistic understanding of these factors is necessary in efforts to interpret clustering patterns of rare variants throughout the genome.

The field of human genomics is currently faced with several challenges in overcoming the substantial—and often unforeseen—computational bottlenecks that can occur when analyzing increasingly massive high-throughput sequencing datasets [24]. Modern catalogs of human genetic variation can easily exceed tens or hundreds of millions of variants, ascertained in the genomes of thousands or even hundreds of thousands of individuals. Most non-trivial computational tasks applied to such data must therefore be carefully optimized to minimize processing time, memory usage, and disk input/output bottlenecks. In Chapter 4, we focus on a particular computational method known as mutation signature analysis, first proposed by Alexandrov et al. [25], which has become a fixture in cancer genomics pipelines [26]. The purpose of mutation signature analysis is to jointly evaluate multiple cancer genomes and, based on the observed patterns of variation, make inference about the causal mutation mechanisms [25]. Most published programs for applying these methods, however, were developed and

optimized for relatively small datasets [27–31]. These programs are typically implemented as web servers, where users must upload their data over the internet [30, 31], or as packages in the R programming language, where the entire dataset must be loaded into memory to be analyzed [27–29]. Given the massive size of modern sequencing datasets, it is effectively impossible to perform mutation signature analysis using such programs—web servers typically only accept data uploads on the scale of megabytes (and even if larger datasets were accepted, they would be subject to severe network bandwidth bottlenecks), and all but the most powerful servers lack the physical memory necessary to analyze large datasets directly in memory, as required by existing implementations in R.

To overcome these computational challenges, we develop Helmsman, a highly efficient software tool to infer mutational signatures in arbitrarily large datasets of genetic variation. For datasets small enough to be tractably analyzed with existing mutation signature analysis programs, Helmsman performs this task up to 300 times faster and with a 100-fold reduction in memory usage. Moreover, because the runtime and memory footprint of Helmsman scale linearly and independently with the number of variants and sample size, respectively, it is the only program currently available that is capable of directly performing mutation signature analysis on datasets where the size of the input data exceeds the physical memory capacity of the computer. The scalable nature of Helmsman ensures that mutation signature analysis will be tractable even for datasets larger than those presently available.

A deep understanding of mutation rate heterogeneity also has important implications concerning the technical tasks of variant detection and quality control in high-throughput sequencing [32, 33]. In Chapter 5, we present a novel method to leverage detailed knowledge of mutation patterns as an effective and interpretable indicator of sample quality control in large-

scale whole-genome sequencing datasets. Just as the concept of mutational signatures can be used to describe the mutational mechanisms operative in cancer genomes (as in Chapter 4), we show that signatures of various technical artifacts can manifest among the rare SNV spectra of individual samples or sequencing batches. Our method implements a combination of mutation signature analysis algorithms and unsupervised anomaly detection techniques to infer the presence of such error signatures and flag samples that are enriched for these signatures. Applying this method to multiple large-scale sequencing datasets, we demonstrate that these inferred error signatures often correspond to known error biases induced by various technical issues. Furthermore, we show that samples affected by these artifacts often went undetected by standard sample-level quality control measures, underscoring the efficacy of our method.

Modern sequencing technology has enabled researchers and clinicians to catalog and study variation throughout the genome at an unprecedented scale. Because all genetic variation fundamentally originates with the acquisition of new mutations, an intimate knowledge of the processes which generate these mutations is essential to ensuring the observed patterns of genetic variation are interpreted accurately. This dissertation presents a detailed account of the patterns and properties of mutation processes operating throughout the human genome, many of which have heretofore not been characterized. We introduce powerful new computational methods that exploit this knowledge of mutation patterns to analyze large-scale genomic data more effectively and efficiently. Collectively, these studies have widespread implications both for understanding the biology of mutation processes and reshaping the ways in which we identify and interpret variation in the human genome.

Chapter II.

Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans

Introduction[1]

Germline mutagenesis is a fundamental biological process, and a major source of all heritable genetic variation (see [5] for a review). Mutation rate estimates are widely used in genomics research to calibrate variant calling algorithms [34], infer demographic history [35], identify recent patterns of genome evolution [36], and interpret clinical sequencing data to prioritize likely pathogenic mutations [37]. Although mutation is an inherently stochastic process, the distribution of mutations in the human genome is not uniform and is correlated with genomic and epigenomic features including local sequence context [12, 38], recombination rate [39], and replication timing [14]. Hence, there is considerable interest in studying the regional variation and context dependency of mutation rates to understand the basic biology of mutational processes and to build accurate predictive models of this variability.

---

[1]This chapter is published as Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nat Commun. 2018;9:3753. A full list of co-authors is provided in the published manuscript.

The gold standard for studying the germline mutation rate in humans is direct observation of *de novo* mutations from family-based whole-genome sequencing (WGS) data [14, 40–42]. These studies have produced accurate estimates of the genome-wide average mutation rate ($\sim 1 - 1.5 \times 10^{-8}$ mutations per base pair per generation) and uncovered some of the mutagenic effects of genomic features. However, the inherently low germline mutation rate means family-based WGS studies detect only 40-80 *de novo* mutations per trio sequenced [14, 40, 42], making it difficult to accumulate a dataset large enough to precisely estimate mutation rates and spectrum at a fine scale and identify factors that explain genome-wide variability in mutation rates.

Other data sources for studying mutation patterns include between-species substitutions or within-species polymorphisms [11, 12, 39, 43–45]. However, because these variants arose hundreds or thousands of generations ago, their distribution patterns along the genome have been influenced by the subsequent long-term actions of many evolutionary forces, such as natural selection and GC-biased gene conversion (gBGC), a process in which recombination-induced mismatches are preferentially repaired to G/C base pairs, resulting in an overabundance of common A/T-to-G/C variants [41, 46, 47]. A further complication of estimating mutation rates with ancestrally older variants is that the endogenous mutation mechanisms themselves have likely evolved over time [48], so patterns of variation observed among these data may not necessarily reflect ongoing mutation processes in the present-day population. To minimize the confounding effects of selection, studies that estimated mutation rates from these data tended to focus on intergenic non-coding regions of the genome, which are less often the target of selective pressure. Nevertheless, even putatively neutral loci may be under some degree of selection [49–51], and are susceptible to the confounding effects of gBGC and evolving mutation processes.

Consequently, these processes bias the resulting distribution of variation, making it difficult to determine which trends are attributable to the initial mutation processes, and which to subsequent evolutionary factors.

We therefore adopt an approach that relies exclusively on extremely rare variants (ERVs) to study innate mutation patterns across the genome. Here we exploit a collection of ~35.6 million singleton variants discovered in 3,560 sequenced individuals from the BRIDGES study of bipolar disorder (corresponding to a minor allele frequency of 1/7120=0.0001404 in our sample). Compared to between-species substitutions or common SNVs, these ERVs are extremely young on the evolutionary timescale (in a comparably-sized European sample, one study estimated the expected age of a singleton to be 1,244 years [52]), making them much less likely to be affected by evolutionary processes other than random genetic drift [5, 15, 41, 46]. ERVs thus represent a relatively unbiased sample of recent mutations and are far more numerous than *de novo* mutations collected in family-based WGS studies.

Our results show that mutation rate heterogeneity is primarily dependent on the sequence context of adjacent nucleotides, confirming the findings of previous studies [12–14]. However, we demonstrate that our ERV-derived mutation rate estimates can differ substantially from estimates based on ancestrally older variants. Evaluating these differences in an independent dataset of ~46,000 *de novo* mutations, collected from two published family-based WGS studies [14, 42], we find that ERV-derived estimates yield a significantly more accurate portrait of present-day germline mutation rate heterogeneity. We further refine these estimates of context-dependent mutability by systematically estimating how mutation rates of different sequence motifs are influenced by genomic features in wider surrounding regions, including replication timing, recombination rate, and histone modifications. Remarkably, we find that the direction of

effect for some genomic features depends on the actual sequence motif surrounding the mutated site, underscoring the importance of jointly analyzing sequence context and genomic features. Accounting for these granular effects of the genomic landscape provides even greater accuracy in describing patterns of variation among true *de novo* mutations. Our results suggest that trends of variation throughout the genome are shaped by a diverse array of context-dependent mutation pathways. This high-resolution map of mutation rate estimates, along with estimates of the mutagenic effects of genomic features, is available to the community as a resource to facilitate further study of germline mutation rate heterogeneity and its implications for genetic evolution and disease.

Results

ERV data source and quality control

In the *Bipolar Research in Deep Genome and Epigenome Sequencing (BRIDGES)* study, we sequenced the genomes of 3,716 unrelated individuals of European ancestry to an average diploid-genome coverage of 9.6x. We identified and removed 156 samples which appeared to be technical outliers, resulting in a final call set of 35,574,417 autosomal ERVs from 3560 individuals (**Methods**). Due to the relatively low coverage of our sample, we likely failed to detect millions more ERVs—a recent study [53] estimated the discovery rate for singletons in a sample of 4,000 whole genomes at 10x coverage to be ~65-85%. Quality control measures indicate that the ERVs we detected are high quality, with a Transition/Transversion (Ts/Tv) ratio of 2.00, within the commonly observed range for single nucleotide variants (SNVs) from WGS data [54] (**Table A.1**). Application of the 1000G strict accessibility mask [55] (which delineates

the most uniquely mappable genomic regions) or a more stringent mapping quality score filter (MQ>56) did not appreciably change the Ts/Tv ratio (1.97-2.01) (**Table A.1**). We estimate fewer than 3% of the 35,574,417 ERVs are false positives (**Appendix A**), similar to the validated singleton error rates of other sequencing studies using a similar technology [55–57]. In addition, we present evidence that erroneous calls among the ERVs are unlikely to be biased by motif-specific genotyping error, mapping error, or mispolarization (**Appendix A**).

Context-dependent variability in mutation rates

The nucleotides surrounding a mutated site are a well-known predictor of variability in mutation rates across the genome [12, 13, 41]. The most detailed such analysis to date [12] considered the nucleotides up to 3 positions upstream and downstream from a variant site (i.e., a 7-mer sequence context), and estimated substitution probabilities per heptameric motif using 7,051,667 intergenic SNVs observed in 379 Europeans from phase 1 of the 1000 Genomes Project (hereafter referred to as the "1000G mutation rate estimates"). These estimates have the potential problem of being derived from variants across the entire frequency spectrum: among the intergenic SNVs used to estimate these rates, singletons and doubletons account for only ~25% [12], so most variants occur at a higher frequency and thus likely arose hundreds or thousands of generations in the past. Over such a long time span, variants affected by cryptic selection, gBGC, or other evolutionary processes are more likely to have been fixed or disappeared, altering the distribution of observable variation.

Because ERVs are assumed to have occurred very recently in human history, we asked if ERV-based mutation rate estimates differed from the 1000G estimates, and if so, whether our revised estimation strategy more accurately represents basal mutation processes. To answer these

questions, we first used the BRIDGES ERVs to estimate mutation rates according to mutation type (e.g., A>C, A>G, and so on) and local sequence context, considering the bases up to 3 positions upstream and downstream from each variant site (**Methods**). We refer to a mutation of a given type centered at a given sequence motif as a "mutation subtype" (e.g., C[A>C]G is a 3-mer subtype). Note that we are not estimating an absolute per-site, per generation mutation rate, but rather the relative fraction of each subtype containing an ERV within the BRIDGES data. We refer to rates calculated in this manner as "relative mutation rates," and estimated these rates for all possible 1-, 3-, 5-, or 7-mer subtypes.

ERV-derived relative mutation rate estimates for the six basic 1-mer mutation types reflect the expected higher mutability for transitions relative to transversions [5]. Splitting each mutation type into more granular subtypes reveals how additional patterns of mutation rate heterogeneity emerge as broader sequence contexts are incorporated (**Fig. 2.1; Fig. A.1**). Our ERV-based estimates confirm nearly all the hypo- or hypermutable motifs reported in previous studies [11, 12]. A subset of these are highlighted in **Fig. 2.1a**, including lower relative mutation rates for NNN[C>T]GCG subtypes and A>G subtypes in motifs containing runs of 4 or more A bases (shown in green boxes), and higher relative mutation rates for N[A>G]T, N[C>T]G, and CA[A>G]TN subtypes (pink boxes). Another notable example of context-dependent hypermutability is the set of NTT[A>T]AAA subtypes (**Fig. 2.1b**), also described previously [12]. Despite A>T mutations having the lowest relative mutation rate among 1-mer types, its NTT[A>T]AAA subtypes have a >6-fold higher rate than the 1-mer A>T relative mutation rate.

Overall, the ERV-derived 7-mer relative mutation rates span a >400-fold range from 0.0003 (CGT[A>T]CCG) to 0.1416 (ATA[C>T]GCA). For every 3-mer subtype, we found overwhelming evidence for heterogeneity in the relative mutation rates among their 16 respective

5-mer constituents (chi-squared tests; all $P < 10^{-231}$). Further, 1522 (99%) of the 1536 5-mer subtypes had significantly heterogeneous rates among their respective 7-mer constituents (chi-squared tests; $P < 0.05$) (**Methods**).

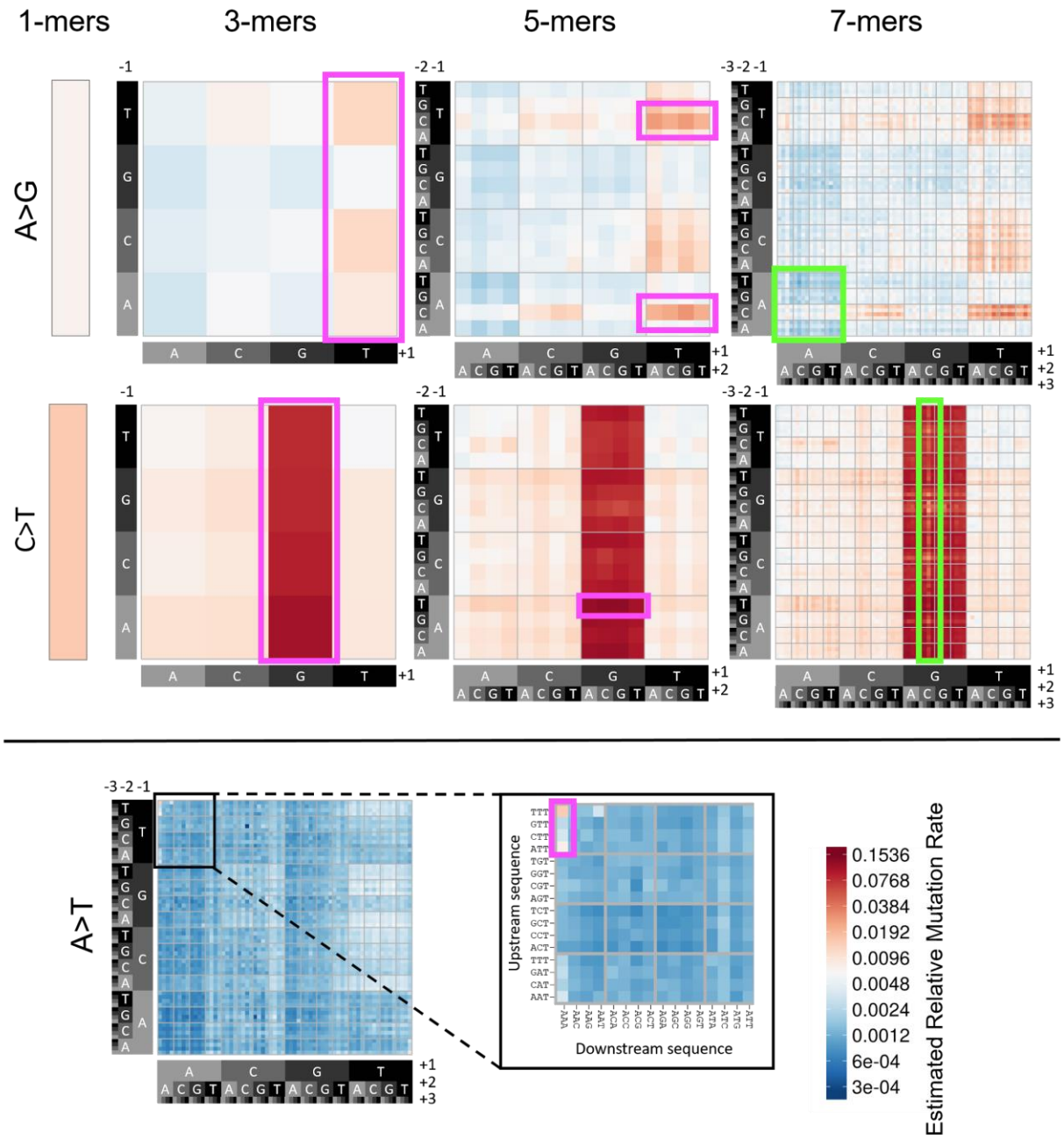


Figure 2.1 Mutation rates vary according to sequence context (a) Heatmap of estimated relative mutation rates for all possible for A>G and C>T transition subtypes, up to a 7-mer resolution (High-resolution heatmaps for all possible subtypes are included in **Fig. A.1**). The leftmost panels show the relative mutation rates for the 1-mer types, and the subsequent panels to the right show these rates stratified by increasingly broader sequence context. Each 4x4 grid delineates a set of 16 subtypes, defined by the

upstream sequence (y-axis) and downstream sequence (x-axis) from the central (mutated) nucleotide. Boxed regions indicate motifs previously identified by Aggarwala and Voight as hypermutable (pink) or hypomutable (green), relative to their similar subtypes. **(b)** Zoomed-in view showing hypermutable NTT[A>T]AAA subtypes relative to other 7-mer A>T subtypes.

Mutation rate estimates differ between ERVs and common SNVs

We next compared the 7-mer relative mutation rates, estimated either from the BRIDGES ERVs or 1000G intergenic SNVs, to determine if our ERV-based estimates differ from previously reported patterns of mutation rate heterogeneity. Across all 24,576 7-mer mutation types, relative mutation rates were highly correlated between the two sets of estimates (Spearman's r=0.95; **Fig. 2.2a**). However, when stratified by mutation type, these correlations were often much weaker (r=0.42 to 0.92; **Fig. 2.2b**). Considering differences in the estimated rates for each individual 7-mer subtype, we found 13% of 7-mer subtypes had differences of 50% or more between the two estimates after normalization. These discrepancies did not occur randomly across subtypes (**Fig. 2.2c**). For example, relative mutation rates for CpG>ApG and CpG>GpG transversions were respectively 26% and 39% higher in the 1000G estimates compared to the ERV-derived estimates. Sequence context also affects relative mutation rate estimates for A>C and A>G subtypes: 1000G-derived estimates were significantly higher than ERV-derived estimates among GC-rich motifs (4-6 G/C bases in the +/-3bp flanking sequence) compared to low-GC motifs (3 or fewer flanking G/C bases) (t-tests; $P < 8.0 \times 10^{-30}$) (**Fig. A.2**; **Table A.2**). This observation is consistent with the known correlation between GC content and biased gene conversion [47, 58], though other evolutionary processes may also have contributed.

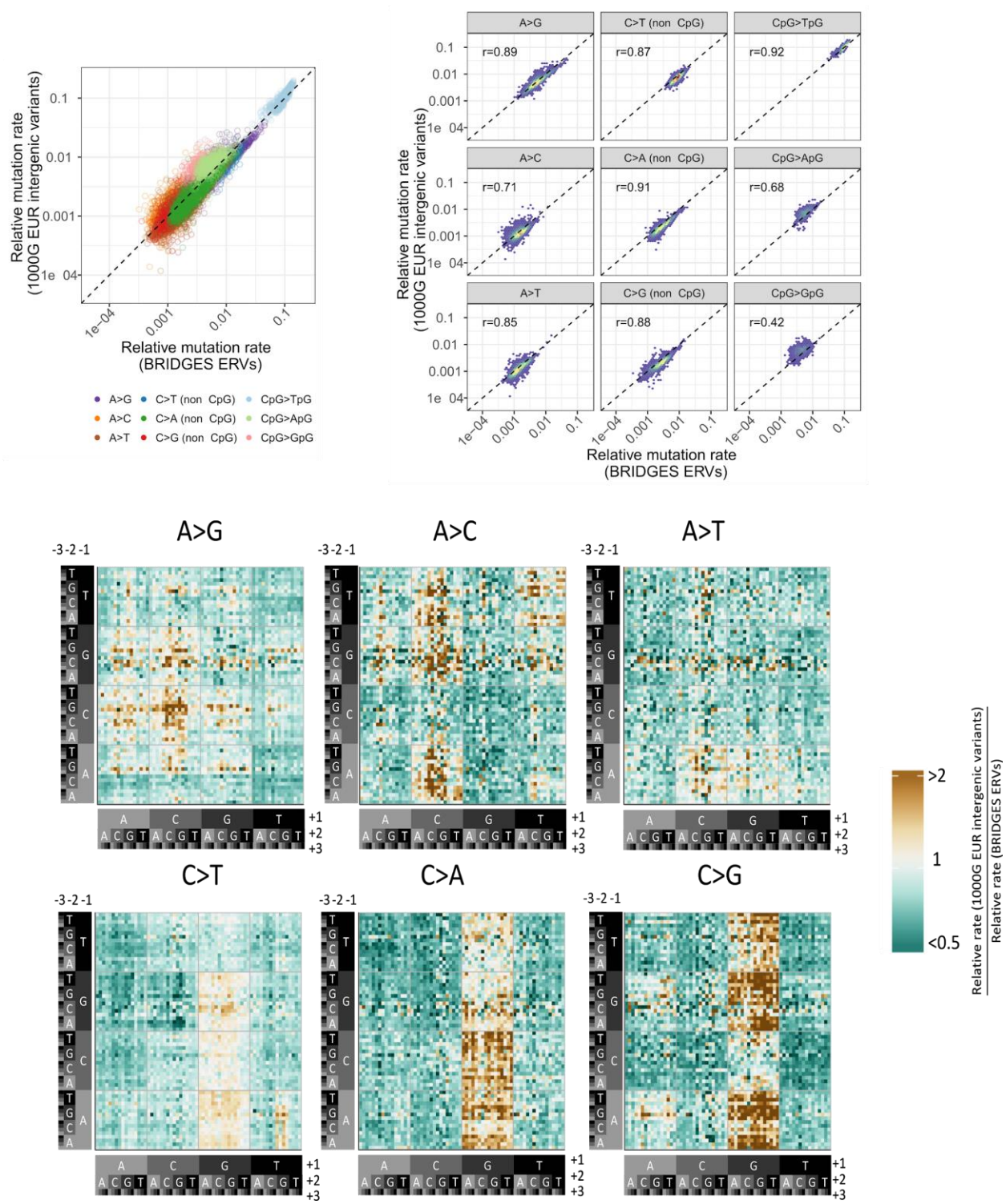


Figure 2.2 Discordance between ERV-estimated and common SNV-estimated mutation rates (a) Relationship between 7-mer relative mutation rates estimated among BRIDGES ERVs (x-axis) and the 1000G intergenic SNVs (y-axis) on a log-log scale. We note that the strength of this correlation is driven by hypermutable CpG>TpG transitions. **(b)** Type-specific 2D-density plots, as situated in the scatterplot of **(a)**. The dashed line indicates the expected relationship if no bias is present. **(c)** Heatmap showing ratio between the relative mutation rates for each 7-mer mutation subtype. Subtypes with higher rates among the 1000G SNVs (relative to ERV-derived rates) are shaded gold, and subtypes with lower rates in the 1000G SNVs are shaded green. Relative differences are truncated at 2 and 0.5, as only 2.5% of subtypes showed differences beyond this range.

We considered the possibility that these patterns of dissimilarity were simply due to technical differences between the BRIDGES and 1000G samples. To address this concern, we estimated 7-mer relative mutation rates using 12,088,037 variants with a minor allele count ≥10 (MAC10+) in the BRIDGES sample and compared these estimates to the ERV-derived and 1000G-derived estimates (**Appendix A**). Importantly, the MAC10+ 7-mer relative mutation rates were more closely correlated with the 1000G-derived estimates (overall: r=0.98; **Fig. A.3a**; type-specific: r=0.87-0.98; **Fig. A.3b**), than with the ERV-derived estimates (overall: r=0.95; **Fig. A.4a**; type-specific: r=0.45-0.95; **Fig. A.4b**). Like the 1000G estimates, the MAC10+ estimates also showed higher rates of CpG transversions and A>G/A>C mutations in GC-rich motifs (**Fig. A.4c**), but between the MAC10+ and 1000G estimates, these differences were absent or much weaker (**Fig. A.3c**).

Collectively, these results suggest that the dissimilarities between ERV-based and common SNV-based estimates are driven not by differences in the data source or analysis pipeline, but by differences in the allele frequencies of the variants used to estimate the rates. There are two plausible explanations for these differences: either 1) the ancestrally older variants included in the 1000G data are under the influence of evolutionary processes that have altered the relative frequencies among subtypes, or 2) even after our careful data cleaning and filtering, certain sequence motifs are enriched for false positive or false negative sequencing errors in the BRIDGES ERVs.

These scenarios can be tested by comparing which set of estimates better describes the observed distribution of true *de novo* mutations. We reasoned that if biased sequencing errors have occurred, such spurious effects would occur more frequently among BRIDGES ERVs, as errors must be present in multiple individuals to manifest among the common variants included

in the 1000G data. In such a scenario, we would expect the 1000G estimates to explain the distribution of true *de novo* mutations more accurately. In contrast, if the relative mutation rate estimates have been influenced by evolutionary processes, such biases should have a stronger effect on the 1000G estimates and the ERV-derived estimates would provide a better fit.

ERVs accurately predict *de novo* mutations

We implemented this validation strategy by comparing how accurately different sets of relative mutation rate estimates predicted the incidence of 46,813 bona fide *de novo* mutations collected from two family-based WGS datasets: The Genomes of the Netherlands (GoNL) project [14] and the Inova Translational Medicine Institute Preterm Birth Study (ITMI) [42] (**Methods; Fig. A.5**). We set these *de novo* mutations against a randomly-selected background of 1 million non-mutated sites, then applied logistic regression models using each set of relative mutation rate estimates (either ERV-based estimates at varying K-mer lengths, or 1000G-based 7-mer estimates) to predict the log-odds of observing a *de novo* mutation at each of the 1,046,813 sites. We evaluated model performance by two likelihood-based goodness-of-fit statistics: the Akaike information criterion (AIC), and Nagelkerke's pseudo- R^2 (**Methods**). Each model has one parameter, so the AIC of each model is $-2 \cdot log - likelihood + 2$.

Among ERV-based K-mer models, goodness-of-fit improved consistently with consideration for longer motifs, with the 7-mer model producing the best fit overall (**Table 2.1**). These trends did not change when varying the number of non-mutated sites (**Table A.3**) nor when applied exclusively to either the GoNL or ITMI mutations (**Table A.4**), indicating the regression was not merely fitting to cryptic errors in the validation data. To assess if our results are affected by mapping artifacts, we also re-estimated the ERV-based 7-mer relative mutation

rates after applying the 1000 Genomes strict accessibility mask (**Appendix A**). The masked and unmasked 7-mer rates are highly concordant, and most discrepancies appear to be an artifact of sampling variation due to fewer ERVs in the masked data (**Fig. A.6**). When applied to predict the *de novo* mutations, the masked rates produced a worse fit than the unmasked rates (**Table 2.1**), suggesting that the reduction in ERVs caused by applying the mask has a larger effect on the precision of our estimates than any mapping artifacts present in the unmasked data. We next analyzed each mutation type separately to determine if the same trend of improved goodness-of-fit using longer K-mers held for different mutation types. In each of these type-specific validation models, the ERV-based 7-mer relative mutation rate estimates provided a significantly better fit than estimates in smaller K-mers (**Table A.5**).

We then compared the goodness-of-fit of the BRIDGES ERV-based K-mer models with the 7-mer model based on 1000G intergenic SNVs. Although Aggarwala and Voight demonstrate that the 1000G 7-mer model significantly improves on 5-mer or 3-mer models [12], our results show that all ERV-based models (except the 1-mer model) predict *de novo* mutations more accurately than 1000G 7-mer model (**Table 2.1**). Considering each mutation type separately (**Table A.5**), we find that the performance of the 1000G 7-mer model is particularly weak among certain mutation classes: for A>C and A>G types, the 1000G 7-mer models provide a worse fit than ERV-derived 5-mer models, and for A>T and CpG>GpG types the fit is worse than ERV-derived 3-mer models. In each of the other C>N types, the 1000G 7-mer model performs comparably to the ERV-derived 7-mer model, indicating the inferred mutation patterns of these types are mostly consistent between the two datasets. These results thus support a scenario where, due to the influence of GC-biased gene conversion [46] or changing mutation processes [48], type- and subtype-specific patterns of variation among the 1000G-derived

estimates are less accurate than ERV-derived estimates in capturing ongoing patterns of germline mutability.

Table 2.1 Goodness-of-fit statistics for mutation rate estimates applied to *de novo* testing data

Mutation rate estimation strategy			AIC	ΔAIC†	AIC rank*	Nagelkerke's R^2
Subtype length	Study	Variant type				
1-mers	BRIDGES	ERVs	353,896	21,575	7	0.088
3-mers	BRIDGES	ERVs	335,319	2998	4	0.118
5-mers	BRIDGES	ERVs	332,861	540	3	0.124
7-mers	BRIDGES	ERVs	332,321	0	1	0.126
7-mers	BRIDGES	ERVs (passing 1000G strict mask)	332,582	261	2	0.125
7-mers	BRIDGES	MAC10+	342,886	10,565	5	0.103
7-mers	1000G	Intergenic SNVs [12]	344,003	11,682	6	0.100

†difference in AIC from the baseline BRIDGES 7-mer model
*lower AIC rank indicates better model performance

Subtype-specific mutagenic effects of genomic features

Family-based sequencing studies have been instrumental in identifying genomic features that are associated with variation in the germline mutation rate [13, 14, 41]. However, these studies have only described the marginal effects of features on the entire spectrum of mutation and have not assessed if the effect of a genomic feature might vary according to the local sequence context. To determine how the mutation distribution varies across the genomic landscape, we selected 14 genomic features (**Table A.6**) and estimated the joint effects of these features on the mutation rate of each 7-mer subtype using multiple logistic regression (**Methods**). Subtypes with few observed ERVs have little power to detect significant

associations, so we estimated the effects of features only for the 24,396 of 24,576 (99.3%) 7-mer subtypes with at least 20 observed ERVs, resulting in 392,128 parameter estimates (**Fig. A.7**). We note that >84% of the 7-mer subtypes we evaluated contained >10 times as many ERVs as parameters estimated, so these estimates are unlikely to be an artifact of overfitting. To identify significant effects among the many associations tested, we applied a false discovery rate (FDR) cutoff of 0.05 to the p-values for each feature across all subtype-specific estimates. Of the 24,396 7-mer subtypes analyzed, 3,481 had at least one genomic feature significantly associated with mutability, with 6,152 significant associations among 392,128 tests.

Three features (H3K9me3 peaks, recombination rate, later replication timing) were associated with higher relative mutation rates across nearly all significantly associated 7-mer subtypes (**Fig. 2.3a**), consistent with previously reported mutagenic effects of these features: H3K9me3 marks are one of the strongest predictors of somatic SNV density [59, 60], and recombination and late replication timing are both correlated with higher germline mutation rates [14, 39]. In addition, four features (H3K36me3 peaks, DNase hypersensitive sites [DHS], GC content, CpG islands) were each associated with both higher and lower relative mutation rates, depending on the mutation type and, in some cases, the sequence motif. These features have been previously implicated in variation in germline or somatic mutation rates, but only as marginal effects, not type- or subtype-specific. H3K36me3 has been shown to regulate DNA repair machinery *in vivo* [61, 62]. DNase hypersensitivity was previously reported to be associated with increased germline mutation rates [13], though cancer genome studies have claimed DHS are susceptible to both increased and decreased somatic mutation rates [63, 64]. CpG islands were associated with ~3-fold lower mutation rates in 99% (1015/1024) of CpG>TpG 7-mer subtypes,

consistent with known patterns of DNA hypomethylation in CpG islands [65], but are associated with higher relative mutation rates in subtypes of other types.

Finally, for CpG>TpG transition subtypes, lamin-associated domains were associated with higher relative mutation rates and three histone marks (H3K4me1, H3K4me3, and H3K27ac) were associated with lower relative mutation rates (**Fig. 2.3b**). These results are consistent with published findings of correlations between these features and DNA methylation: lamin-associated domains were previously found to associate with focal DNA hypermethylation in colorectal cancer [66], and H3K4me1, H3K4me3, and H3K27ac are known markers of DNA hypomethylation [67, 68]. Exonic regions were associated with lower relative mutation rates for ~26% of CpG>TpG subtypes (**Fig. 2.3b**), consistent with findings of lower somatic SNV density in gene-rich regions [59], though it is unclear if this is also due to DNA hypomethylation.

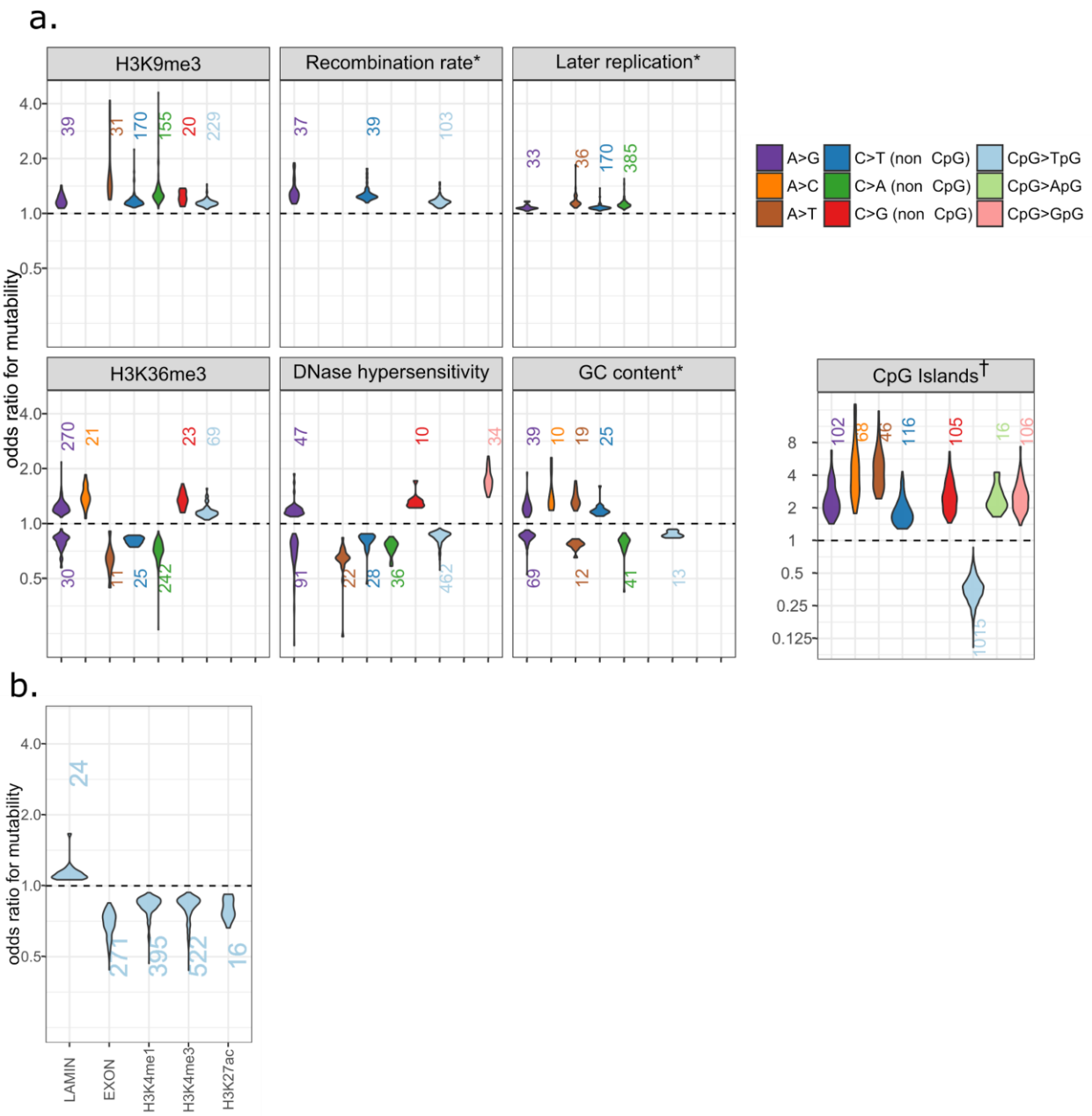


Figure 2.3 Distributions of statistically significant mutagenic effects of genomic features. (a) Effects of 7 genomic features where associations with multiple mutation types were detected. For features with bidirectional effects, we separately plotted distributions of positive associations (OR > 1; above dashed line) and negative associations (OR < 1; below dashed line). The number of 7-mer subtypes within each type for which that feature is statistically significant in a positive or negative direction is shown above or below each distribution. Distributions are only shown for types with 10 or more 7-mer subtypes associated in the same direction. *Odds ratios for the 3 continuously-valued features (recombination rate, replication timing, and GC content) indicate the change in odds of mutability per 10% increase in the value of that feature. †Effects in CpG islands tend to be stronger than other features, so are shown on a wider scale. **(b)** Distributions of significant mutagenic effects for the 5 features only associated with CpG>TpG transitions.

Estimated effects of features predict *de novo* mutations

We applied these 7-mer+features mutation rate estimates to predict the GoNL/ITMI *de novo* mutations, using the same evaluation framework described earlier. Model fit statistics indicate that the rates estimated from 7-mer sequence context and genomic features describe the distribution of *de novo* mutations significantly better than the 7-mer-only estimates (**Fig. 2.4**). When partitioned by mutation type, inclusion of genomic features improves model fit for 8 of the 9 basic mutation types. These differences tend to be weaker among transversion types, likely because there were fewer *de novo* mutations of these types available (**Fig. 2.4**). Including genomic features had the largest effect on the prediction of CpG>TpG transitions, consistent with the expected associations between certain features and DNA methylation. Comparing the distribution of predicted mutations across basic types under different models, we find that all models generally recapitulate the observed distribution of *de novo* mutations, but the 1000G 7-mer model predicts a notably higher proportion of CpG>NpG mutations (**Fig. A.8a**). Stratifying by 3-mer subtype, the 1000G 7-mer predictions also tend to be more dissimilar from the *de novo* distribution than ERV-based 7-mer+features predictions (**Fig. A.8b**).

To further demonstrate that effects of genomic features described in **Fig. 2.3** are supported by bona fide *de novo* mutation data, we pooled all subtypes found to be associated with each feature in a positive or negative direction and respectively tested for an enrichment or depletion of GoNL/ITMI *de novo* mutations in regions covered by that feature (**Methods**). We found 10 of the 20 tests were statistically significant in the expected direction (chi-squared tests; $P < 0.05$), confirming that, at a coarse level, many of the subtype-specific effects of genomic features inferred using ERVs are recapitulated among true *de novo* mutations (**Table A.7**).

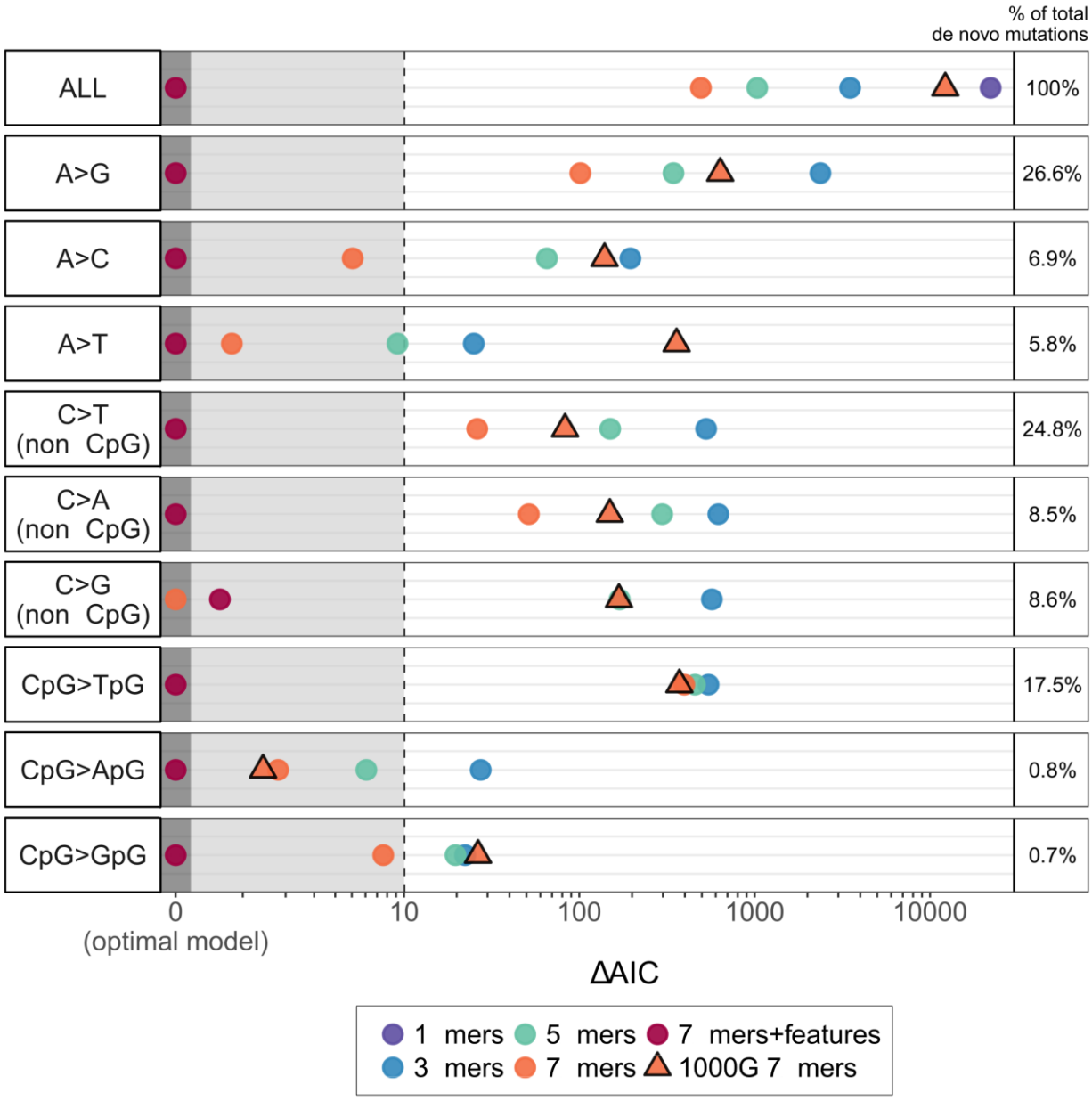


Figure 2.4 Comparison of goodness-of-fit for different mutation rate estimation strategies. For each mutation type and each model i, we calculated $\Delta AIC_i = AIC_i - AIC_{min}$ as a measure of relative model performance, with lower values of ΔAIC indicating better fit to the GoNL/ITMI *de novo* mutation data. ΔAIC is shown on the horizontal axis on an arcsinh scale. For each mutation type, the best-fitting model thus has a $\Delta AIC = 0$. Models with $\Delta AIC < 10$ (grey-shaded area) are considered comparable to the optimal model, whereas models with $\Delta AIC > 10$ are considered to explain substantially less variation than the optimal model [69].

Germline mutation rates mirror somatic mutation processes

The rate heterogeneity between mutations of the same type suggests that distinct mutation mechanisms underlie some of the feature-subtype associations detected by our model. However, mechanisms for specific mutation signatures have mostly been studied for somatic mutations in cancer, and the degree to which these mechanisms affect germline mutations is generally unknown. In the following, we show two examples where the germline mutation rates from our data are consistent with mutation mechanisms observed in cancer. Moreover, we hypothesize a previously undescribed mechanism for germline point mutations.

In cancer genomes, H3K36me3-marked regions are targeted by the error-prone DNA polymerase eta (POLH, also known as pol η) [62]. Human POLH is particularly biased towards generating A>G mutations at sites flanked by weak (A or T, denoted as W) bases [70]; consequently, H3K36me3-marked regions are enriched for W[A>G]W mutations in various cancers [62]. In our data, among the 403 7-mer subtypes showing significant positive associations with H3K36me3-marked regions, a significant majority (270, or 67%) are A>G subtypes (exact binomial test; $P < 1.09 \times 10^{-111}$). Within the 270 positively-associated A>G subtypes, 175 (65%) are W[A>G]W 3-mer subtypes, significantly more than expected by chance (exact binomial test; $P < 4.12 \times 10^{-43}$).Thus, our results suggest the H3K36me3-mediated POLH mutation signature also appears in the germline.

Active transcription factor binding sites (i.e., occurring in DHS) are also prone to elevated somatic mutation rates in various cancers, likely because bound transcription factors make DNA inaccessible to nucleotide excision repair (NER) machinery [64, 71]. For example, the CCAAT motif is a highly specific binding target for the trimeric nuclear factor Y (NF-Y) complex [72], and active NF-Y binding sites show a >3.2-fold enrichment for somatic mutations

in melanomas [64]. Our results indicate that transcription factor binding may also explain motif-specific hypermutability in the germline. Among the 7-mer subtypes positively associated with DHS, CCA[A>G]TNN subtypes show a 1.1 to 1.3-fold enrichment (Wald test; $P < 2 \times 10^{-4}$), and the CCA[A>G]TNN *de novo* mutation rate in the GoNL/ITMI dataset is 1.7-fold higher when occurring within DHS versus non-DHS regions (1-df chi-squared test; $P < 0.0055$).

Finally, we and others [12] observed that NTT[A>T]AAA subtypes have >6-fold higher mutation rates than other A>T subtypes (**Fig. 2.1b**). We note that the TTAAAA hexamer is the canonical insertion target for Long Interspersed Element 1 (LINE-1, or L1) retrotransposons, and is nicked by the L1-encoded ORF2p endonuclease at the antisense 3'-ApT-5' dinucleotide [73]. These nicks produce T-rich 3' flap structures, which can be recognized and removed by NER machinery, inhibiting L1 insertional mutagenesis, but leaving an A-rich single-strand break [74]. In transcriptionally active regions of the genome, such lesions are usually repaired by high-fidelity NER pathways [75], but in nucleosomal DNA, where NER activity is impaired, the lesions are likely bypassed by error-prone translesion synthesis (TLS) polymerases [64]. Our results show NTT[A>T]AAA mutations are reduced >3-fold when occurring in DHS (Wald test; $P < 2.0 \times 10^{-26}$). We hypothesize that the context-dependent mutation signature in our data is the result of damage induced by L1 retrotransposons and subsequent errors of the TLS polymerase. This model is consistent with observing higher NTT[A>T]AAA mutation rate outside of DHS, where NER activity may be impaired and lesions must be bypassed by error-prone TLS during replication. Additionally, according to the "A-rule" [76], TLS polymerases preferentially pair abasic sites with adenine. Hence, mutations generated by errors of the TLS polymerase explain the preponderance of A>T (but not A>G or A>C) mutations at the NTTAAAA motif.

Discussion

The main motivation of our study is to understand the genome-wide variation of germline mutation rates in humans. We bring to this task two innovations: first, we take advantage of large-scale WGS data, focusing on extremely rare variants as a potentially more powerful data source than currently available collections of *de novo* mutations [13, 14, 40, 42] or common variants [11, 12]. Second, building upon previous attempts to holistically model the relationship between sequence context, genomic features, and mutation rate, we estimate fine-scale mutagenic effects of multiple genomic features. Unlike previous studies, which estimated the impact of genomic features by treating all single-nucleotide mutation subtypes in aggregate [13], we allow for the possibility that mutation rates of sequence motifs are differentially affected by these features.

Our results not only confirm the previously reported hypermutable effects of specific sequence contexts and genomic features, but also demonstrate that many feature-associated effects previously only described in somatic cells are present in the germline. Moreover, our approach identifies certain genomic features, including H3K36me3 peaks, DNase hypersensitive sites, and CpG islands, that may act to both suppress and promote mutability depending on the mutation type and sequence context, providing insight into the causal mechanisms of germline mutation rate heterogeneity across the genomic landscape.

The subtype-specific effects of genomic features we report likely represent only a fraction of the effects across the genome, due to the limited power of detecting associations among rarer subtypes. A larger dataset of ERVs will likely reveal additional cases of association and will enable further study of mutation patterns among longer sequence motifs, additional genomic features, and interactions or nonlinear effects thereof. We also note several of the

genomic features used in our study were assayed in somatic cell lines or aggregated over multiple cell types. The currently available data for these features only crudely approximates the true genomic variation in germ cells, so the effects we estimated have likely regressed towards the mean. Generating precise maps of genomic features within male and female germ cell lineages may further uncover mutagenic mechanisms unique to the germline. Despite these limitations, the fine-scale effects of sequence context and genomic features reported here provide the most accurate map to date of germline mutation variation, as demonstrated by their improved ability to predict genuine *de novo* mutation patterns.

Even without accounting for the effects of genomic features, our ERV-derived mutation rate estimates for 7-mer subtypes are consistently more accurate than those based on mostly common SNVs from 1000 Genomes Project data [12]. Remarkably, even coarser estimates—the ERV-derived 5-mer and 3-mer rates—predict the spectrum of *de novo* mutations more accurately than the 1000G 7-mer estimates, demonstrating the merit of ERVs as a refined data resource for studying innate mutation patterns. Some of the improvement is likely the result of reduced sampling error, as our ERV dataset is larger than the 1000G dataset. Nevertheless, this result has two important implications. First, it suggests that high-frequency variants in presumably neutral genomic regions are influenced by biased evolutionary processes, such as selection and gBGC, or these variants arose via past mutational processes that are now inactive [48]. Second, this reaffirms the high quality of ERVs in our data: the potential errors due to calling or mapping biases among these ERVs are likely weaker than the evolution-driven biases affecting the older variants. The larger sample, young allelic age, and high quality of ERVs together result in a demonstrably more accurate appraisal of recent or ongoing patterns of mutability than common SNVs.

Because the germline mutation rate is a critical parameter in the study of genetic variation, we envision a wide range of applications that stand to benefit from incorporating our genome-wide map of mutation rate estimates. Currently, many methods that rely on simulating "baseline" mutations, such as the pathogenicity scoring algorithm *CADD* [77] and coalescent simulator *ms* [78], do not account for context-dependent mutation rate differences. Likewise, clinical applications for differentiating disease-causing mutations from background variation require a precise estimate of the expected *de novo* mutation rate, but even the most advanced of these only consider differences in 3-mer or 7-mer sequence contexts, and are based on intergenic SNVs from 1000 Genomes data [12, 79]. Incorporating more accurate sequence- and feature-dependent estimates of mutation rates may lead to more realistic simulations and greater confidence in the inferences made by these methods. Another relevant area of research where our results might be applicable is the study of how germline mutation mechanisms have evolved over time [48, 80, 81]. If mutator phenotypes have frequently arisen throughout the evolutionary history of humans (as hypothesized by [48]), the effects of mutational modifiers have likely been extremely subtle, manifesting as granular context-specific mutation signatures. Our results, which describe the present-day pattern of mutation rate heterogeneity in Europeans, provide a wealth of potential hypotheses for investigating how these mutation processes have been shaped by past evolution.

To facilitate the use of our genome-wide mutation rate estimates in other analysis and simulation pipelines, we have created a genome browser track to visualize these estimates at a single-base resolution alongside other genomic data. Ultimately, the refined mutation patterns from ERVs and the detailed dissection of context-feature effects serves as a quantitative

foundation for better understanding the molecular origins of mutation rate heterogeneity and its consequences in heritable diseases and human evolution.

Methods

Sample description

The BRIDGES sample contains 3,927 unrelated European American bipolar disorder cases and controls. The cases and controls from the Centre for Addiction and Mental Health (CAMH) in Toronto (n=830), the Institute of Psychiatry, Psychology and Neuroscience (IoPPN) and King's College London in London, U.K. (n=845) [82], the Genomic Psychiatry Cohort (GPC) (n=1,151) [83], and the Prechter Repository (n=363) [84] were collected as previously described, as were the STEP-BD cases (n=304), obtained from the NIMH repository [85], and the Minnesota Center for Twin and Family Research (MCTFR) study controls (n=434) [86]. In all studies, DNA was extracted from blood-based samples. All human research was approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki. All participants provided written informed consent.

Sample library preparation

The concentration of each DNA sample was measured by fluorometric means (PicoGreen, Thermo Fisher, Woburn, MA, USA) followed by agarose gel electrophoresis to verify the integrity of DNA. Six-hundred nanograms of DNA was sheared with acoustic shearing (Covaris, Woburn, MA, USA) to an average size of 400nt. Following shearing, the samples are

transformed to a sequencing library using standard protocols to create a paired-end library. Briefly, sheared DNA was end-repaired, A-tailed and ligated with Illumina adaptors (New England Biolabs, Ipswitch, MA, USA). Following ligation, indexed primers were used to amplify the final libraries for each sample. Each sample received two indexes: 96 i7 indexes were used to identify each sample in each 96-well reaction plate while a single i5 index was used for each plate. This combination of indexes uniquely coded all samples in the project when both the i7 and i5 indexes were read during sequencing. Following six cycles of PCR (Kapa Biosystems, Wilmington, MA, USA), libraries were purified and quality controlled by assaying the final library size using the Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and quantitating the final library via real-time PCR (Kappa Biosciences). A single peak between 300-400bp indicates a properly constructed and amplified library ready for sequencing. PCR cycles for amplification are kept to a minimum to minimize PCR duplication rate and maximize library complexity.

Sequencing

Sequencing was performed per Illumina protocol, essentially as described by [87]. Libraries were pooled in sets of 12 samples and each pool sequenced on a single lane of a HiSeq 2500 flowcell using version 3 Illumina chemistry at paired-end 100nt read lengths. Each library pool was loaded at 13pM to generate 160-180M paired reads per lane. Multiple flowcells of the library pools were performed to generate a final data set with an average coverage of 9.6x per sample.

Sample filtering and data quality control

Among the 3,927 samples attempted, three failed library preparation and were not sequenced. We removed an additional 162 samples due to quality issues: five with imbalanced read counts between read 1 and read 2, four with improperly generated BAM files, 16 that had an average coverage <3x, and 137 due to high contamination (FREEMIX or CHIPMIX score >3% using VerifyBAMID [88]). For samples that failed for multiple reasons, we report a single category for simplicity.

Among these 3,762 samples, reads were mapped to Build 37 of the human reference genome (including decoy sequence [55]), with alignment and variant calling performed using the GotCloud pipeline [33]. After variant calling, we applied additional sample-level filtering as described below to obtain the 3,716 included in our analysis. We first excluded 10 case samples that were not phenotyped as type 1 bipolar disorder (removed solely for consistency with ongoing analyses of the BRIDGES data that do require phenotypes). We identified and removed an additional 23 samples that showed evidence of sample swaps in VerifyBAMID [88], but had not been excluded from variant calling. We next computed continental-ancestry PCA coordinates by projecting BRIDGES samples in the coordinate space of the 1000 Genomes phase 1 samples [89]. We dropped 11 samples identified as PC ancestry outliers, defined by PC1<0.01 or PC2<0.025. We then checked for relatedness using the $\hat{\pi}$ statistic (i.e., estimation of pairwise identity-by-descent based on LD-pruned SNPs), computed in plink [90]. Nearly all pairwise sample comparisons were consistent with being unrelated, with $\hat{\pi} < 0.05$ for 99.9% of sample pairs. Two samples were dropped due to relatedness, as the $\hat{\pi}$ between these was 0.5, indicating the two were full siblings.

These filters reduced the sample to 3,716 individuals, in which we called 37,470,516 autosomal singleton SNVs in the mappable genome (i.e., non-N reference bases in the GRCh37 reference genome) that passed the variant-level filtering criteria implemented in the GotCloud pipeline [33]. Prior to performing our analyses, we examined how these 37.5 million ERVs were distributed across individual samples to identify and remove individuals that showed abnormal patterns of variation due to systematic sequencing errors or batch effects. In brief, we adapted the non-negative matrix factorization (NMF) technique described by [91] to summarize the distribution of ERVs unique to each individual as a composite of 3 distinct "signatures." For each of the 3,716 individuals in our sample, we calculated a vector of 96 3-mer relative mutation rates (described below) using only the ERVs observed in that individual, generating a 3,716 x 96 rate matrix. Decomposition of this matrix via NMF produces a 3,716 x 3 matrix describing the relative contribution of each signature to the observed mutation spectrum per individual. Because we assume the relative mutation rate of any given subtype should be similar across individuals, it follows that the contribution of a given NMF signature should also be similar. We removed 156 individuals where one or more signatures had a contribution >2 standard deviations away from the mean contribution of that signature calculated across all individuals, reasoning that ERVs observed in these individuals are more likely to be errors. The final sample used in our analyses thus consists of 3,560 individuals, in which we identified 35,574,417 singletons. Additional details of this filtering strategy are described in **Appendix A**.

Mutation subtypes and calculation of relative mutation rates

Each of the 35,574,417 singletons can be classified into one of 6 basic mutation types, defined by the reference and alternative allele: A>C, A>G, A>T, C>T, C>G, and C>A. The

notation of A>C includes both A-to-C mutations and complementary T-to-G mutations. For each mutation type, we further define a set of mutation subtypes by the bases flanking the variant site. Since there are 4 possible bases at both the +1 position and the -1 position, there are 4x4=16 possible 3-mers containing each basic mutation type at the central position, producing 6x16=96 3-mer subtypes. Likewise, there are $6x4^4$=1,536 5-mer subtypes, and $6x4^6$=24,576 7-mer subtypes. To simplify notation, we denote a subtype by the sequence motif containing either an A or a C as the reference base at the central position (e.g., either CGT[A>X]TCG or CGT[C>X]TCG).

For each K-mer subtype, we divided the number of ERVs observed at the central position of the K-mer by the number of times the K-mer is seen in the mappable autosomal regions of the reference genome; we term this proportion the *estimated relative mutation rate*. K-mers in the reference genome were counted by a 1-bp sliding window, so that every possible occurrence of that K-mer was accounted for (e.g., a run of 4 As is counted as two AAA 3-mers shifted by one base). For example, we observed 7,548 C>T or G>A autosomal singletons occurring in an ATACGCA or TGCGTAT 7-mer motif (the underlined base indicates the variant site) and there are 53,314 such motifs in the autosomal reference genome where this subtype of mutation could be observed, yielding a relative mutation rate estimate of 7,548/53,314=0.1416 for the ATA[C>T]GCA subtype.

Testing for heterogeneity of relative rates

As each K-mer can be split into 16 possible (K+2)-mers that share the same internal motif but differ in their terminal bases, the relative mutation rate for each K-mer subtype is the weighted mean of the rates found among its 16 possible (K+2)-mer constituent subtypes. To

assess the heterogeneity of relative mutation rates among each set of 16 (K+2)-bp constituent subtypes that share the same K-bp motif, we performed a chi-squared test for uniformity of these rates, with each test having 15 degrees of freedom.

Mutation prediction model and validation

To evaluate the accuracy of different mutation rate estimation strategies, we applied the estimated rates to predict the incidence of 46,813 *de novo* mutations using logistic regression. These *de novo* mutations were published by two independent studies: 11,020 *de novo* mutations detected in 258 Dutch families by the Genomes of the Netherlands (GoNL) project [14], and 35,793 *de novo* mutations from 816 families sequenced by the Inova Translational Medicine Institute (ITMI) Premature Birth Study [42]. We combined the observed mutations with 1 million randomly selected sites from the mappable autosomal regions of the reference genome to serve as a non-mutated background, reasoning that ~20 non-mutated sites for each actual de novo mutation would be sufficient to minimize sampling noise in the set of non-mutated sites; we also repeated this procedure with 500,000, 2 million, and 3 million randomly selected sites to tell if the trends we observed were affected by the size of the non-mutated background. Because each non-mutated site can be ambiguously considered as the background for 3 different mutation types, we divided the 1 million non-mutated sites into 3 non-overlapping sets. We designated A/T and C/G reference bases in the first set (consisting of 333,334 unique sites) as non-mutated A>G and C>T types, respectively, and so on for the second set (A>C or C>G types), and the third set (A>T or C>A types), each of which contained 333,333 unique sites. Hence, we considered a total of 1,046,813 testing sites (1,000,000 unmutated sites and 46,813 de novo mutations), each with one possible mutation event, in our prediction models.

Now let $i = \{1, \ldots, 1046813\}$ be an index for the 1,046,813 testing sites. We coded $d_i = 1$ if site i is a de novo mutation and $d_i = 0$ otherwise. If a set of estimated relative mutation rates reflects the underlying mutation process, we expect that the odds of a given site for carrying a *de novo* mutation increases with the estimated relative mutation rate of that site. To assess this expectation for all sets of mutation rate estimation strategies (e.g., ERV-based or 1000G-based 7-mer estimates), we annotated each testing site i with the relative mutation rate estimated under strategy M ($r_{i,M}$), and used logistic regression to model the probability of a *de novo* mutation at each site as a function of these rate estimates, where α_0 is the intercept term and α_1 is the regression coefficient:

$$ln(\frac{Pr(d_i = 1)}{Pr(d_i = 0)}) = \alpha_0 + \alpha_1 r_{i,M} \qquad (1)$$

The probability of a mutation at each testing site can then be calculated as:

$$Pr(d_i = 1) = \frac{1}{1 + e^{\alpha_0 + \alpha_1 r_{i,M}}} \qquad (2)$$

The overall likelihood of model M, given the observed data, is the product of the probability values over all 1,046,813 sites:

$$L_M = \prod_{d_i=1} \frac{1}{1 + e^{\alpha_0 + \alpha_1 r_{i,M}}} \prod_{d_i=0} \frac{e^{\alpha_0 + \alpha_1 r_{i,M}}}{1 + e^{\alpha_0 + \alpha_1 r_{i,M}}} \qquad (3)$$

Using this likelihood, we evaluated model fit by the Akaike Information Content (AIC), where p is the number of parameters in equation (1) (because all models are based on a single covariate of mutation rates, $p = 1$ in all cases):

$$AIC_M = 2p - 2ln\,(L_M) \qquad (4)$$

For each model, we also calculate Nagelkerke's R^2:

$$R_M^2 = \frac{1 - \{\frac{L_0}{L_M}\}^{2/N}}{1 - \{L_0\}^{2/N}} \qquad (5)$$

Here, L_0 is the likelihood of a null intercept-only model with no covariates.

Because these likelihood-based goodness-of-fit statistics are calculated across all the basic mutation types combined, they do not provide information about which types benefit most strongly from using expanded sequence motifs. For example, it is possible that any improvement to the overall goodness-of-fit is elicited by context-dependent heterogeneity of a single mutation type, whereas other types might not be significantly affected by using longer sequence motifs, and do not contribute to the improved model fit. To identify these type-specific trends, we stratified our testing data by each of the basic mutation types. To account for the known hypermutability of cytosine at CpG dinculeotides, we separated C>T, C>G, and C>A mutations into CpG and non-CpG types, for a total of 9 basic mutation types. For each type, we repeated the 3-mer, 5-mer, and 7-mer models on only the sites of that type. Within each set of type-specific models, we again compared the goodness-of-fit using AIC and Nagelkerke's R^2. Note that because the absolute values of AIC and Nagelkerke's R^2 are a function of the number of data points included in the model, these statistics cannot be directly compared between type-specific models, where the number of data points vary.

Estimating effects of local genomic features

We estimated the effect of 14 genomic features (data sources for these features are described in **Table A.6**) on the relative mutation rate of each 7-mer subtype using the following logistic regression framework. Let K be the index across all 7-mer subtypes with 20 or more observed singletons ($K \in \{1, \ldots, 24396\}$). Let j_K be the index across all sites that are centered at

the 7-mer motif that could produce a mutation of subtype K, and let $Z_{j_K} = 1$ if the site carries a singleton of subtype K and $Z_{j_K} = 0$ otherwise. We annotated each site of the considered subtype for 14 genomic features, generating predictors $F_{j_K,1}, \dots, F_{j_K,14}$. We treated 11 of these features as binary variables (seven histone marks, lamin-associated domains, CpG islands, DNase hypersensitive sites, exons), setting the predictor $F_{j_K,g} = 1, g \in \{1, \dots, 11\}$ if the central site of the motif was inside the specified regions and $F_{j_K,g} = 0$ otherwise. For the 3 continuous features (recombination rate, replication timing, surrounding GC content), we set the predictor $F_{j_K,g}, g \in \{12,13,14\}$ to the mean value of that feature in a 10kbp window centered at the site. Because the inferred effect of some features may be confounded by correlation with read depth and calling rates (e.g., GC content [92]), we included read depth at the central site of the 7-mer as covariate $F_{j_K,DP}$. For each 7-mer subtype K, we then evaluated the effect of the genomic predictors on the log odds of mutability for each site Z_{j_K} using the following logistic regression equation:

$$ln(\frac{Pr(Z_{j_K} = 1)}{Pr(Z_{j_K} = 0)}) = \beta_0^K + \beta_1^K F_{j_K,1} + \cdots + \beta_{14}^K F_{j_K,14} + \beta_{DP}^K F_{j_K,DP} \qquad (6)$$

where $(\beta_1^K, \dots, \beta_{14}^K)$ are effects of the 14 considered genomic features on the mutation rate of subtype *K*, and β_{DP}^K is the effect of the local sequencing depth. The intercept of this model, β_0^K, represents the feature-adjusted relative mutation rate for the considered 7-mer subtype. We performed this logistic regression and obtained parameter estimates in R v3.2.3 using the speedglm() function from the *speedglm* package. We performed this procedure for each of the $K \in \{1, \dots, 24396\}$ 7-mer subtypes, obtaining effect size estimates and standard errors for 16 x 24,396 parameters. Note that we did not consider estimating interaction effects between the 14 genomic features, as estimating all 2-way interactions would require an additional 14*(13-1)/2=91 parameters per subtype-specific regression, which would lead to overfitting concerns.

To generate a map of mutation rates across the genome, we used the estimated regression coefficients to predict the relative mutation rate (i.e., probability of observing a singleton) at each site $\boldsymbol{j}$ where a mutation of a given 7-mer subtype could occur:

$$Pr(Z_{j_K} = 1) = \frac{exp(\beta_0^K + \beta_1^K F_{j_K,1} + \cdots + \beta_{14}^K F_{j_K,14} + \beta_{DP}^K F_{j_K,DP})}{1 + exp(\beta_0^K + \beta_1^K F_{j_K,1} + \cdots + \beta_{14}^K F_{j_K,14} + \beta_{DP}^K F_{j_K,DP})} \qquad (7)$$

Because there are three possible mutations at every site, we predict 3 independent mutation probabilities (one for each possible alternative allele). For example, for a site centered at a ACGATTG motif, we predict probabilities for A>C, A>G, and A>T alleles, using the parameters estimated from those models. This prediction uses all estimated effects, not just the effects determined to be statistically significant. We note that we did not generate predictions for sites within 5Mbp of the start/end of a chromosome, because recombination rate data were not available for these regions [93].

To assess if inclusion of these genomic features improved upon the 7-mer mutation rate estimates in describing the true distribution of germline mutability, we again tested this model's ability to predict the known *de novo* mutations from the GoNL [14] and ITMI [42] studies. We annotated each of the $i = \{1, \ldots, 1046813\}$ testing sites with the predicted mutation rate, $Pr(Z_{i_K} = 1)$, and calculated the goodness-of-fit using equations 1-5 with this parameter as the predictor. Note that the GoNL/ITMI data included *de novo* mutations within the 5Mbp telomeric regions where we could not estimate effects of genomic features. Rather than excluding sites in these regions from our goodness-of-fit comparison, we simply assigned the marginal 7-mer relative mutation rate as the predicted value for these sites, to ensure models were compared using identical data.

Data and code availability

Predicted mutation rates based on sequence context and genomic features at each site have been formatted as a UCSC Genome Browser track, which can be accessed at http://mutation.sph.umich.edu.

All custom scripts used in downstream data processing and analyses are available at https://github.com/carjed/smaug-genetics. A web-based utility and command-line code for annotating a variant call format (VCF) file of genetic variants with estimated 7-mer mutation rates can be accessed at http://www.jedidiahcarlson.com/mr-eel/.

Chapter III.

Patterns and properties of multinucleotide mutations in the human germline

Introduction

Many methods in population and evolutionary genetics rely on the assumption that single-nucleotide variants (SNVs) are spatially independent from one another—that is, each observed SNV is assumed to have arisen through its own unique mutation event. Several recent studies, however, have estimated that approximately 1-5% of human germline single-nucleotide mutations are in fact the result of multinucleotide mutation events which simultaneously generate multiple point mutations, separated by relatively short distances ranging from 1 to 20,000 base pairs (bp) [13, 14, 17, 18, 21–23], violating the common assumption that SNVs are spatially independent from one another.

Failure to account for these multinuceotide mutations (MNMs) can cause serious confounding in efforts to identify regions of positive selection [20] or inference of population demographic history [21]. In addition, MNMs are important to consider as a distinct class of mutations for a variety of other reasons. First, MNMs are inherently characterized by properties that do not exist for single-nucleotide mutations, namely that they consist of multiple point

mutations and span multiple bases of mutated and non-mutated sequence. Collectively, these properties define what we refer to as the intrinsic mutation rate of MNMs. For simple MNMs consisting of just two constituent point mutations, the intrinsic mutation rate is effectively an inverse linear function of the inter-mutation distance (i.e., shorter inter-mutation distances correspond to higher intrinsic mutation rates). Because MNMs can result from a variety of distinct mechanisms, the intrinsic mutation rate is a crucially important property for understanding the biological processes underlying MNMs throughout the genome [17, 21–23, 94].

Due to the unique mechanistic origins of MNMs, they also exhibit distinctive mutation spectra and a general tendency to be enriched for transversion mutations [21–23]. Consequently, MNMs are more likely to have deleterious effects when occurring within coding regions and may play a significant role in the etiologies of various heritable diseases [18, 95]. Additionally, because MNMs inherently affect multiple bases in the genome, they have been hypothesized to contribute to accelerated evolution [19, 96]. Understanding the mechanistic origins of MNMs, their intrinsic properties, and their distribution throughout the genome therefore stands as an important task for explaining the demographic and evolutionary history of humans and deciphering the genetic architecture of complex traits.

Like most studies of single-nucleotide germline mutation patterns in humans, investigations into the patterns and properties of MNMs have largely relied on *de novo* mutations identified by whole-genome or whole-exome sequencing of families [14, 17, 18, 22, 23]. Although such datasets are widely considered to be the gold standard for characterizing mutation processes active in present-day human populations, they are not well-suited for understanding fine-scale variation of these mutation patterns [97]. This is particularly true for MNMs, which

are inherently rare events—on average, assuming a genome-wide multinucleotide mutation rate of 3.9 x 10^-9 [17], we expect each individual to carry only 1-2 *de novo* MNMs in their entire genome. Consequently, even the largest published trio sequencing studies to date have only been capable of ascertaining fewer than 2,000 *de novo* MNMs [22, 23].

An alternative strategy that has been used to study MNMs is to investigate spatially clustered common variants that occur in perfect linkage disequilibrium (LD) in population-based samples [19, 21]. These studies have ascertained tens of thousands of perfect-LD SNP pairs that likely arose via multinucleotide mutation events (compared to only hundreds or thousands of *de novo* MNMs identified through trio sequencing strategies), and thus are better poised to characterize granular properties of MNMs. However, due to the effects of recombination over many generations, this approach is limited in scope to MNMs whose constituent point mutations are very closely spaced (e.g., <100bp, though these studies often restrict their analyses to even shorter intervals [21]), and thus cannot adequately describe mutation processes which generate MNMs with lower intrinsic mutation rates.

We recently demonstrated that extremely rare variants (ERVs), ascertained from relatively large samples of unrelated individuals, are a powerful data source for studying germline mutation patterns [97]. ERVs represent mutations that have accumulated very recently in the population and, in sufficiently large samples, the vast majority of ERVs are young enough that their distribution has been virtually unaffected by natural selection and biased gene conversion. The putatively young age of ERVs also means that MNMs whose constituent mutations occur further apart are less likely to have been disrupted by recombination events. ERVs therefore may be a particularly useful resource to investigate fine-scale properties of MNMs.

In this study, we propose a simple statistical model to describe the patterns of spatial clustering observed among ERVs. We apply this model to a collection of 32,144,732 ERVs ascertained from 2,000 whole genomes sequenced as part of the Trans-Omics for Precision Medicine (TOPMed) study and show that the spatial distribution of ERVs throughout the genome can be explained accurately and parsimoniously as a mixture of four distinct processes. We find that two of these processes exhibit patterns in their intrinsic mutation rate and spectra that are indicative of two distinct multinucleotide mutation mechanisms known to be active in the human germline. Deeper analysis of these putative MNMs reveals that they are remarkably diverse in their intrinsic properties and genome-wide distribution and are influenced by various features of the genomic landscape, elucidating many previously unknown characteristics of MNMs. The patterns of variation among MNMs also appear to differ slightly between human populations, potentially indicating differences in endogenous or environmental mutation mechanisms. Finally, because each of these processes represents, to varying degrees, the joint influence of independent single-nucleotide mutations, multinucleotide mutations, and the demographic history of the sample, this model enables us to unambiguously quantify the extent to which each of these factors contributes to the observed clustering patterns of singletons.

Results

The TOPMed data

From the 11,759 unrelated individuals sequenced in freeze 3 of the TOPMed study, we selected the subset of 1,000 individuals with the highest proportion of European ancestry, and 1,000 with the highest proportion of African ancestry, with global ancestry inferred using

RFMIX [98] (**Methods**). We chose a sample size of N=1,000 for each subsample for two reasons: first, to restrict each subsample to individuals with relatively homogeneous ancestral backgrounds, and second, to limit the incidence of parallel mutations at hypermutable sites, which can alter the underlying spectra of rare variants in large samples [99]. In each subsample, we recalculated the minor allele counts of each SNV independently and identified 13,351,172 singleton variants in the European subsample, and 18,793,562 singletons in the African subsample.

Inter-singleton distances show evidence of mutational non-independence

We first summarized the spatial distribution of these singletons by calculating the successive distances between each set of singletons unique to each individual (**Methods**). The median inter-singleton distances in Africans and Europeans were 61,897bp and 102,952bp, respectively, reflecting the differences in the total number of singletons observed in each subsample. Next, we compared the empirical distribution of these inter-singleton distances to the distribution we would expect under a simple model of mutational independence, where all singletons are assumed to have arisen independently through a single random mutation process (**Methods**). Consistent with patterns of spatial clustering observed among both de novo mutations [17] and single nucleotide polymorphisms [21], we find that the empirical distribution of inter-singleton distances in the TOPMed data is heavily enriched for closely-spaced intervals that are relatively uncommon under the naive uniparametric models (**Fig. 3.1**).

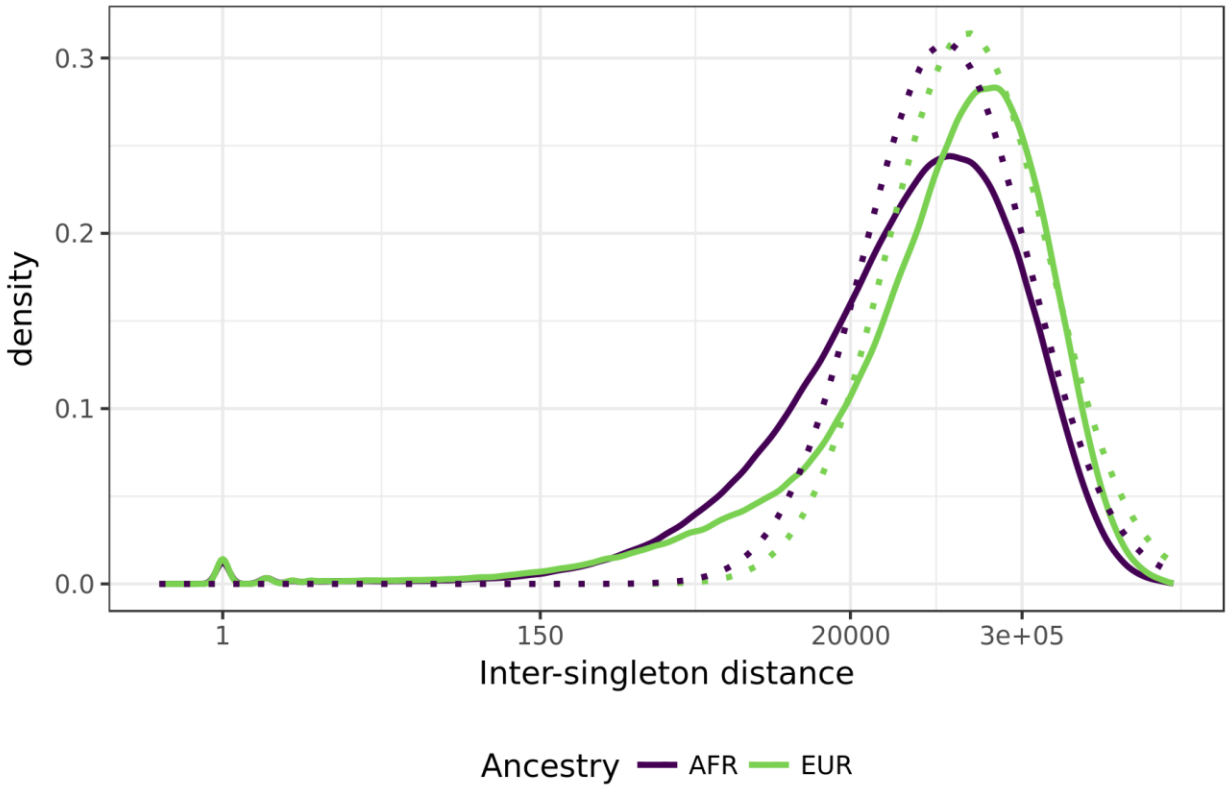


Figure 3.1 Comparison of observed and expected inter-singleton distance distributions. Observed distributions are shown in solid lines, and expected distributions under a single-parameter exponential model are shown in dotted lines. The rate parameters for the expected distributions are based on the median number of singletons per individual, as observed in the European and African ancestry subsamples. ($\theta_{EUR} = 4.5 \times 10^{-6}$; $\theta_{AFR} = 6.3 \times 10^{-6}$).

Modeling the spatial distribution of singletons as a mixture of exponential processes

Prior studies have established that MNMs are a non-trivial source of clustering among single-nucleotide variants [19–21, 100], so a logical explanation for the poor fit of the uniparametric model is that it simply does not account for singletons that arose through non-independent multinucleotide mutation events. It is widely understood that different mechanisms produce MNMs with distinct ranges of intrinsic mutation rates [16]. The inter-singleton distance distribution can therefore be considered as a mixture of a finite number of mutation processes,

where each process is presumed to generate some fraction of singletons occurring with a unique intrinsic mutation rate. Explicitly modeling the effects of these processes, however, would require prior knowledge of the number of operative multinucleotide mutation mechanisms, their intrinsic mutation rates, and what fraction of mutations are attributable to each underlying mechanism. Though these properties have been described for a limited number of multinucleotide mutation mechanisms [16, 17], their joint effects on the spatial distribution of rare variants have not been characterized.

We hypothesized that the empirical inter-singleton distance distribution could be used to estimate the parameters of such a mixture model, and that the properties of these inferred mixture components would reflect properties of the underlying multinucleotide mutation processes and their effects on the spatial distribution of singletons throughout the genome. To further explore this hypothesis, we iteratively fit mixture models to the observed distribution of inter-singleton distances in each of the 2,000 individuals in our sample, starting with two mixture components and increasing the number of inferred components until the goodness-of-fit plateaued (**Methods**). Each mixture component consists of two parameters: lambda, describing the relative contribution of this component to the overall number of inter-singleton intervals, and theta, describing the intrinsic mutation rate of singletons inferred to have been generated by this particular process. Using this strategy, we found that the majority (98%) of individual singleton distributions were best modeled by a four-component (eight-parameter) mixture, so we fixed the number of components per individual at four for our subsequent analyses. As predicted, this mixture model fit the observed data substantially better than the single-parameter exponential model (**Fig. B.1**), indicating these four components accurately explain the majority of variation in the observed inter-singleton distance distribution. Moreover, the estimated parameters of these

four mixture components were highly homogeneous across individuals, with virtually no overlap in the ranges of the four rate estimates, suggesting the same underlying processes were contributing to the inter-singleton distance distributions in all individuals (**Fig. 3.2**). We note that, with the exception of the component 1 rate parameter, all of the parameters for each of the four components differed significantly between the African and European ancestry subsamples (**Table B.1**).

Because we assume that the majority of singletons arise as independent mutations, we interpret the mixture component with the highest lambda value (i.e., that which contributes the most to the observed inter-singleton distance distribution) as corresponding to the independent point mutation process. Hence, we consider mixture component 4 to represent this independent point mutation process. We interpret the remaining three components as corresponding to processes which generate closely-spaced clustering patterns among the ERVs. Components 1 and 2 were of particular interest, as the intrinsic mutation rate estimates of these components were within the range typically used to define MNMs (i.e., <20,000bp) and appeared to directly implicate two distinct multinucleotide mutation mechanisms. Mixture component 1, which has a median intrinsic mutation rate of 1 singleton every 4bp, appears to reflect MNMs occurring through translesion synthesis (TLS), an error-prone process in which specialized polymerases bypass DNA lesions during replication [101]. TLS has been widely studied as a source of multinucleotide mutations, and most studies agree that MNMs with constituent point mutations separated by very short distances of 1-30bp are largely attributable to this mechanism [16, 17, 48, 101]. Mixture component 2 captures singletons occurring at a median intrinsic mutation rate of ~1 per 3,541bp. The intrinsic mutation rate of this component appears to correspond to that of MNMs resulting from hypermutability of single-stranded DNA intermediates that occur during

the repair of double-strand breaks (DSB), the constituent point mutations of which are typically separated by distances of 1,000-5,000bp [22, 23]. Unlike the other three mixture components, the parameters of mixture component 3 do not suggest an obvious biological interpretation. The presence of this component (with a median intrinsic mutation rate of ~1 singleton per 66,389bp) indicates that many singletons are spatially clustered at distances that cannot be explained by either independent mutation processes or known multinucleotide mutation processes. The potential sources of this cryptic clustering pattern are discussed later in this chapter.

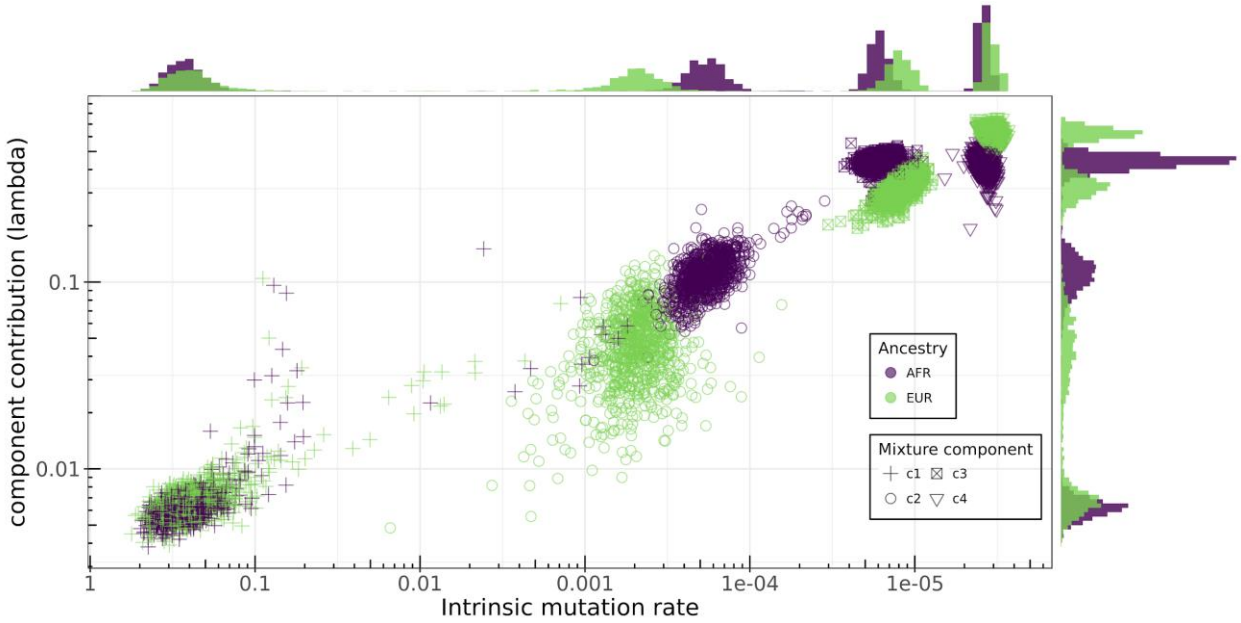


Figure 3.2. **Parameter estimates for exponential mixture models**. Each point represents one of the four components in one of the 2,000 individuals in the sample, colored by the majority ancestry of that individual. The rate of the component is shown across the x-axis, and the proportion that component contributes to the mixture on the y-axis (on a log-log scale). Marginal histograms show the distribution of the lambda and rate parameters for each component.

Properties of clustered singletons correspond to known MNM mechanisms

In addition to generating MNMs occurring at distinct ranges of intrinsic mutation rates, the processes of translesion synthesis and DSB repair are also known to exhibit unique mutation

spectra. Error-prone TLS is particularly prone to generating MNMs enriched for A>T and C>A transversions [17, 21], and MNMs resulting from DSB repair have been found to correspond to unusually high rates of C>G transversions [22, 23]. Therefore, if components 1 and 2 ascertained from our mixture model are genuinely associated with TLS and DSB mutation processes, we would expect that the known mutation spectra of these two mechanisms are reflected in the spectra of singletons we attribute to each of these components. To address this, we calculated the probability of component membership for each singleton (according to the mixture parameter estimates for the individual in which that singleton was observed) and assigned each singleton into one of the four components accordingly (**Methods**). We then tabulated the frequencies of the 6 basic mutation types (A>C, A>G, A>T, C>A, C>G, and C>T) in each of the 4 mixture components across the 2 ancestry subsamples and determined the mutation spectrum (i.e., the proportion of singletons of each basic mutation type) among each of these 8 groups (**Fig. 3.3**).

This analysis revealed that the spectrum of component 1 singletons was heavily enriched for transversions, particularly A>T and C>A, consistent with the typical spectrum of TLS-associated MNMs [17, 21] (**Fig. 3.3**). As described by Jonsson et al. [23] and Goldmann et al. [22], DSB-associated MNMs exhibit a strong enrichment of C>G transversions. If this process is contributing to the clustering patterns associated with mixture component 2 in our model, we should expect to see a similar enrichment of C>G transversions among the singletons assigned to this component. The observed spectrum of component 2 singletons was consistent with this expectation (**Fig. 3.3**)

Component 1 singletons in the African ancestry subsample appeared to have an even stronger transversion bias than those in the European ancestry subsample: A>T and C>A transversions respectively accounted for 13.2% and 15.8% of all component 1 singletons in

Africans, but only 11.5% and 14.1% of component 1 singletons in Europeans. Component 2 also showed slight differences between the two subsamples, with individuals of European ancestry tending to have a greater proportion of C>G transversions (10.9%) than individuals of African ancestry (10.3%). Testing for overall differences in the spectra between the two subsamples, we found that both the spectra of both components 1 and 2 differed significantly (5-df chi-squared tests;$P < 7.8 \times 10^{-188}$).

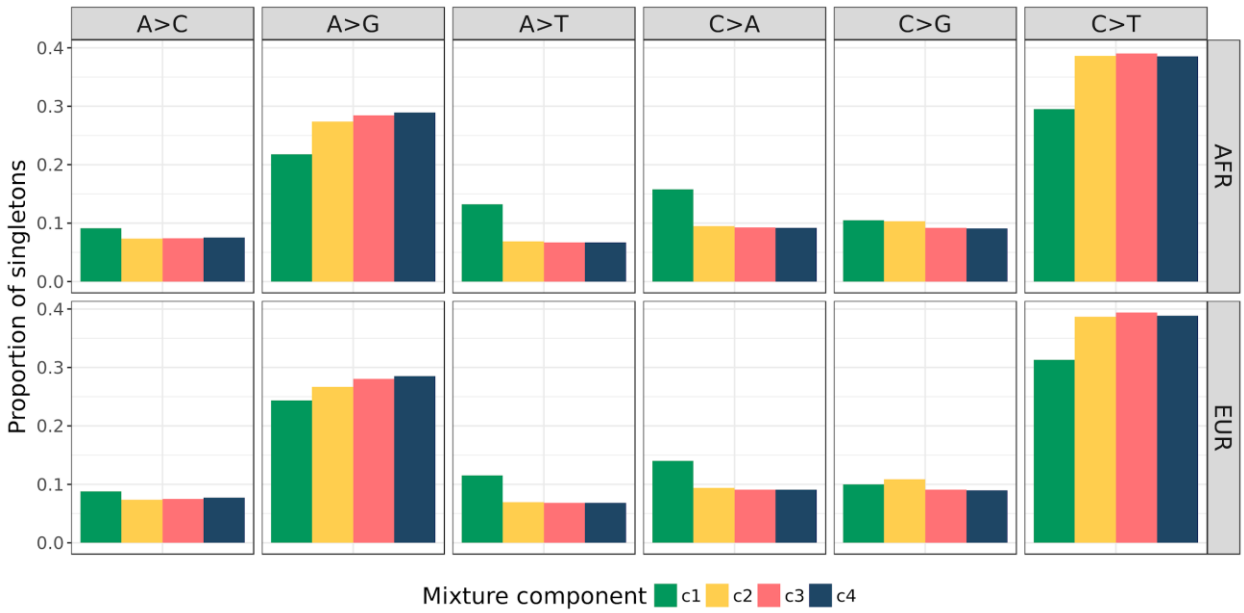


Figure 3.3. **Mutation spectra of mixture components.** For each of the four mutation components, we calculated the relative frequencies of the 6 basic mutation types in the African and European subsamples.

The multinucleotide mutation spectrum varies with the intrinsic mutation rate

Analyzing the spectra of each discrete mixture component provides a useful means of confirming that processes inferred by our model carry the signatures of known multinucleotide mutation mechanisms but may obscure more subtle patterns of variation in the multinucleotide mutation spectrum. For example, Besenbacher et al. found that when MNMs are binned into

discrete categories according to their intrinsic mutation rate, the mutation spectrum varies dramatically [17]. The authors concluded that the mutation spectrum of multinucleotide mutations is a function of the intrinsic mutation rate [17]. This study, however, classified MNMs into only six bins according to the inter-mutation distance of their constituent point mutations (1bp, 2-10bp, 11-100bp, 101-1,000bp, 1,001-5,000bp, and 5,001-20,000bp) [17]. Because the singletons in our data outnumber the *de novo* mutations analyzed by [17] by several orders of magnitude, we classified singletons into 1bp bins for inter-mutation distances <100bp, 100bp bins for distances between 100-20,000bp, and 1,000bp bins for distances >20,000bp, and investigated how the mutation spectra varied over these granular classifications.

This analysis showed that the mutation spectrum of putative MNMs undergoes several dramatic shifts as the intrinsic mutation rate decreases (**Fig. 3.4**). For putative tandem mutations (i.e., MNMs consisting of two immediately adjacent nucleotide changes), we observed the expected excess of A>T and C>A transversions, corresponding to the known signature of GA>TT and GC>AA mutations attributable to DNA polymerase zeta [21]. For putative MNMs with an inter-mutation distance of 2bp (e.g., CAC > GAG), the spectrum shifted immediately towards higher rates of A>G transitions and lower rates of C>A transversions. The spectrum continued to fluctuate as the intrinsic mutation rate decreases. For instance, the proportion of A>G transitions increased from 15% among tandem mutations to 30% in MNMs with an intrinsic mutation rate of 1/3-1/5bp, and A>T transversions decreased from the third most abundant type among tandem mutations (~18%) to the least common type among MNMs with an intrinsic mutation rate of 1/11bp (~9%).

The rank-order of relative frequencies for the 6 basic types eventually stabilized around an intrinsic mutation rate of 1/12bp (though still with high proportions of all 4 transversion

types), but MNMs with larger inter-mutation distances continued to exhibit relatively high rates of C>G transversions, consistent with the expected signature of DSB-associated MNMs [22, 23]. As the intrinsic mutation rate decreased, the high proportion of C>G transversion gradually tapered off until the intrinsic mutation rate reaches approximately 1/10,000bp, at which point the spectrum became nearly indistinguishable from that of unclustered singletons.

In the human genome, there are at least 5 DNA polymerases capable of translesion synthesis [101], so one plausible explanation for the extreme variation in spectra of TLS-associated MNMs is that different TLS polymerases, each of which has distinct error biases, tend to produce MNMs with distinct intrinsic mutation rates. Once the intrinsic mutation rate of putative MNMs drops below 1/100bp, however, the mutation spectra is quite stable, suggesting that DSB-associated MNMs exhibit the same C>G mutational bias, regardless of how closely the constituent point mutations occur. Though different mechanisms are typically assumed to correspond to distinct (and often non-overlapping) ranges of intrinsic mutation rates [16], we emphasize that these results do not preclude the possibility that the spectrum of MNMs with an intermediate range of intrinsic mutation rates (for example, between 1/10 and 1/100) might reflect MNMs that have been generated both by TLS-associated and DSB-associated mutation processes.

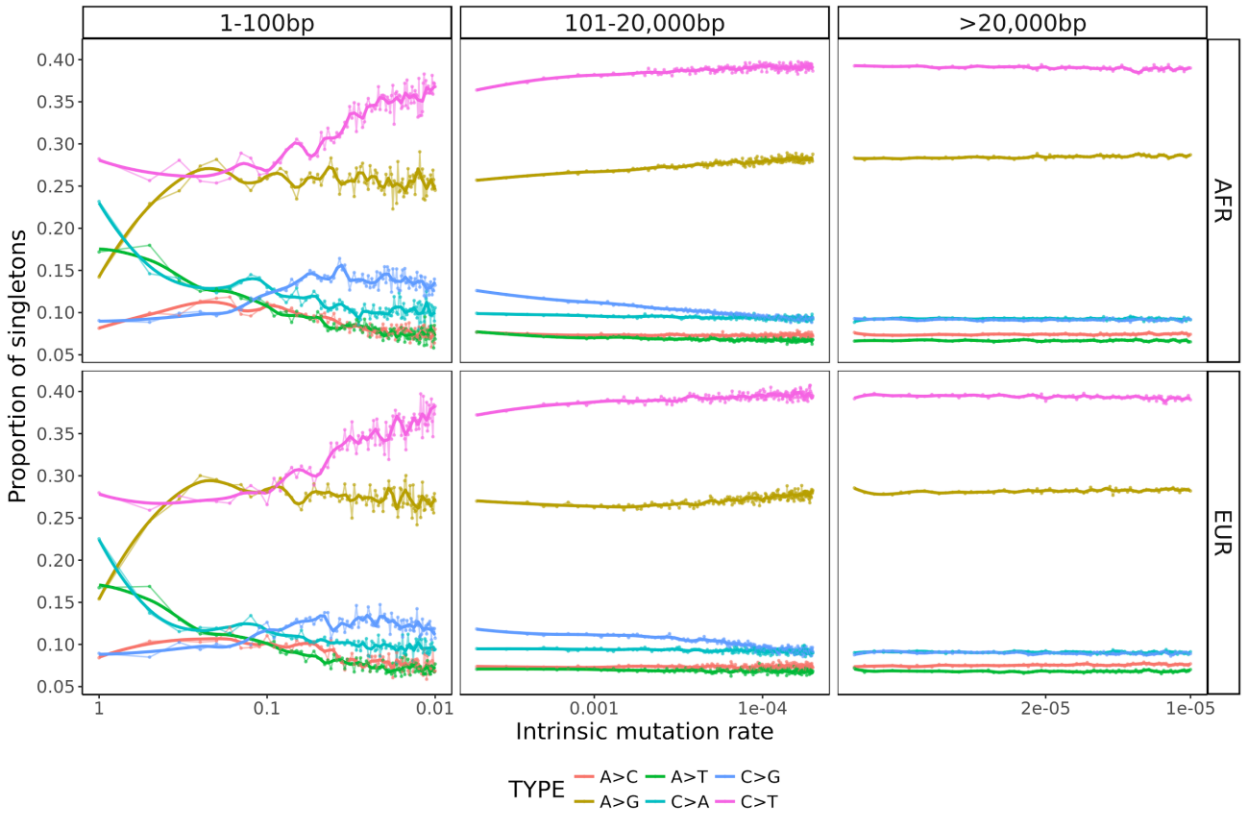


Figure 3.4. Variation in mutation spectra as a function of intrinsic mutation rate. We classified each singleton according to the minimum distance to the nearest singleton in the same individual. To ensure the spectra at each distance were supported by sufficiently many singletons, we classified singletons into 1bp bins for inter-mutation distances <100bp, 100bp bins for distances between 100-20,000bp, and 1,000bp bins for distances >20,000bp. For clarity, these three bin resolutions are plotted in separate panels from left to right. These spectra were calculated separately for the African (top panels) and European (bottom panels) subsamples.

Identification of multinucleotide mutation hotspots and associated genomic features

To characterize how these clustering patterns varied throughout the genome, we counted the number of singletons assigned to each component in non-overlapping 1 megabase pair (Mbp) windows and applied a 3-state Hidden Markov Model (HMM) to these frequency sequences to segment the genome into regions of cold, neutral, or hot spots for each component (**Methods**). Putative hotspots for component 2 singleton clusters showed a nearly perfect overlap with genomic loci previously found to be enriched for DSB-associated MNMs, specifically on chr2p,

chr8p, chr9p, and chr16p/q [22, 23] (**Fig. 3.5**). In addition, we identified several previously uncharacterized hotspots for these clusters, generally occurring in the subtelomeric regions of chr3p, chr4p, chr7p, chr9q, chr19p, chr20q, chr21q, and chr22q (**Fig. B.3**). These hotspots were consistently detected in both the African and European ancestry subsamples. The tendency for these hotspots to occur near the ends of chromosome arms reflects evidence from several studies that have shown subtelomeric regions are widely enriched for DSBs in eukaryotic genomes (as reviewed by [102]). The densities of component 1, 3 and 4 singletons tended to be relatively uniform throughout the genome (**Fig. B.2**, **Fig. B.4, Fig. B.5**), though we note that many of the same regions found to be enriched for component 2 singletons were classified as hotspots for component 3 singletons and cold spots for component 4 singletons.

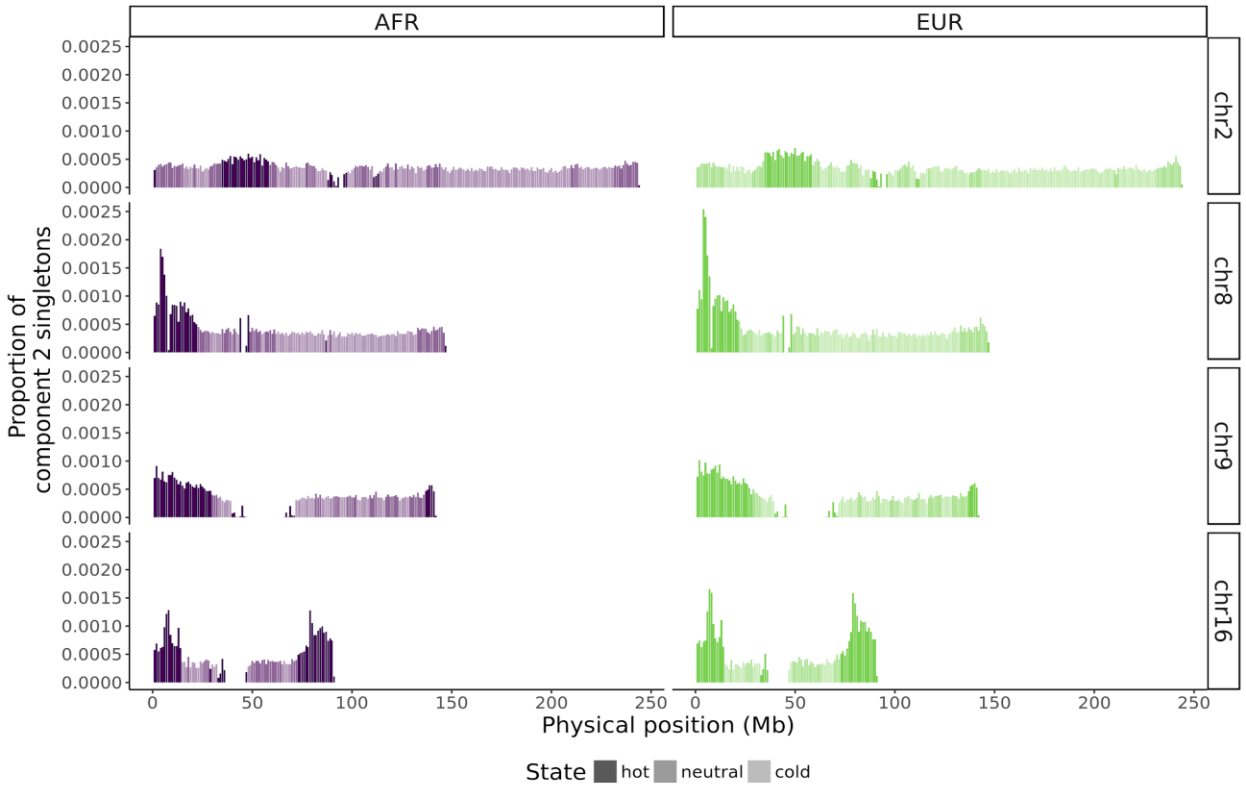


Figure 3.5. **Genomic hotspots of component 2 clustered singletons**. Each bar indicates a 1Mbp window on chromosomes 2, 8, 9, and 16. The height of the bar indicates the total proportion of component 2 singletons that occur in that window. The shade of each bar indicates the inferred state (hot, neutral, or cold) from the Hidden Markov Model applied over all chromosomes. Similar plots for all autosomal chromosomes are shown in Fig. B.3.

Next, we investigated whether the genome-wide variation in the density of each component was associated with various features of the genomic landscape. Using negative binomial regression models, we estimated the effects of 11 genomic features on the genome-wide density of each component (**Methods**). The results of these regression models are summarized in **Fig. 3.6**. The strongest predictor of component 1 density was the presence of CpG islands. Because CpG islands typically co-occur with gene promoters [103], this association suggests that promoter regions might be uniquely susceptible to MNMs generated by translesion synthesis. Moreover, because CpG islands are typically unmethylated [103], this result provides evidence against the possibility that component 1 singleton clusters are merely artifacts of CpG hypermutability. CpG islands also showed weaker positive associations with component 2 density and were negatively associated with the densities of components 3 and 4, indicating that CpG islands have a general tendency to be enriched for clustered singletons. Similarly, H3K4me1, a histone mark associated with decreased DNA methylation [104, 105], was positively associated with component 1 density, but negatively associated with the densities of components 2, 3, and 4. Components 2 and 3 both showed significant negative associations with exon density, potentially indicating a general depletion of mutations in actively-transcribed genes due to transcription-coupled repair processes [106], though we do not exclude the possibility that strong purifying selection has contributed to a depletion of singletons in gene-rich regions.

The remaining features we analyzed tended to have a significant effect in the same direction either for components 1, 2, and 3 (but not 4) or components 2, 3, and 4 (but not 1). H3K27me3 and H3K9me3, two histone marks associated with transcriptional silencing in sperm cells [107], were positively associated with components 1, 2, and 3. One of the strongest predictors of components 2, 3, and 4 was H3K4me3, a histone modification that is known to

correspond to an increased frequency of DSBs in the genomes of yeast, mice, and human cells [108–110], as well as decreased DNA methylation [67]. Four features (DNase hypersensitivity, lamin-associated domains, H3K27ac, and H3K9ac) also showed significant positive associations with the densities of components 2, 3, and 4. All 4 components showed a weak but significant positive relationship with histone mark H3K36me3. Collectively, these results demonstrate that different features of the genomic landscape likely play varying roles in shaping the genome-wide distribution of singleton clustering patterns.

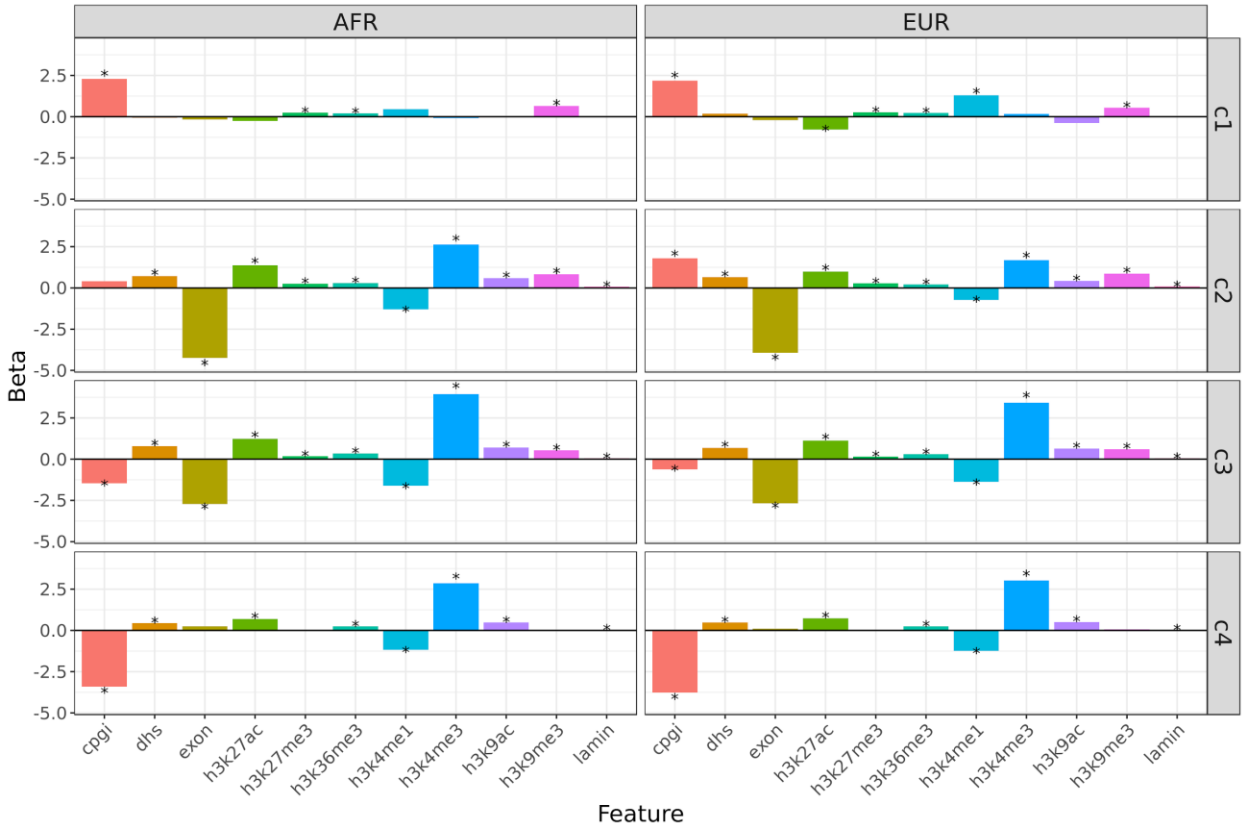


Figure 3.6. Estimated effects of genomic features on regional density of clustered singletons. The regression coefficients (Beta values) are based on negative binomial regression models. Statistically significant associations (P<0.05) are marked with a *.

Validation with de novo mutation data

To further validate these singleton clustering patterns as evidence of multinucleotide mutation processes, we applied our mixture deconvolution method to a collection of 71,769 *de novo* mutations, ascertained from 869 parent-offspring trios sequenced from freeze 5 of the TOPMed study (**Methods**). Because each individual possesses too few de novo mutations for us to perform this mixture deconvolution on a per-individual basis, we estimated the model parameters based on the distribution of inter-mutation distances aggregated across all individuals of a given ancestry. In both the African and European ancestry subsamples, the goodness-of-fit plateaued at 3 mixture components. Notably, the intrinsic mutation rate estimates of the first two components were within the same range as the rate estimates of components 1 and 2 from the singleton mixture model (**Table 3.1**).

Table 3.1 Mixture parameter estimates for de novo mutation distance distributions

	AFR		**EUR**	
	Intrinsic rate	**lambda**	**Intrinsic rate**	**lambda**
Component 1	1/4.89bp	0.010	1/4.04bp	0.008
Component 2	1/5.07kbp	0.014	1/5.83kbp	0.011
Component 3	1/25.6Mbp	0.976	1/23.2Mbp	0.981

These components also exhibited mutation spectra similar to the spectra of components 1 and 2 identified in the singleton mixture model, namely an enrichment for A>T and C>A transversions among component 1 de novo mutations, and an enrichment for C>G transversions among component 2 de novo mutations (**Fig. 3.7**). Testing for overall differences in the spectra

between the two subsamples, we found that neither component 1 nor component 2 differed significantly in spectra (5-df chi-squared tests;$P > 0.86$). As with component 4 inferred from the singleton mixture model, we interpret component 3 from this de novo mixture model to represent the background independent single-nucleotide mutation process.

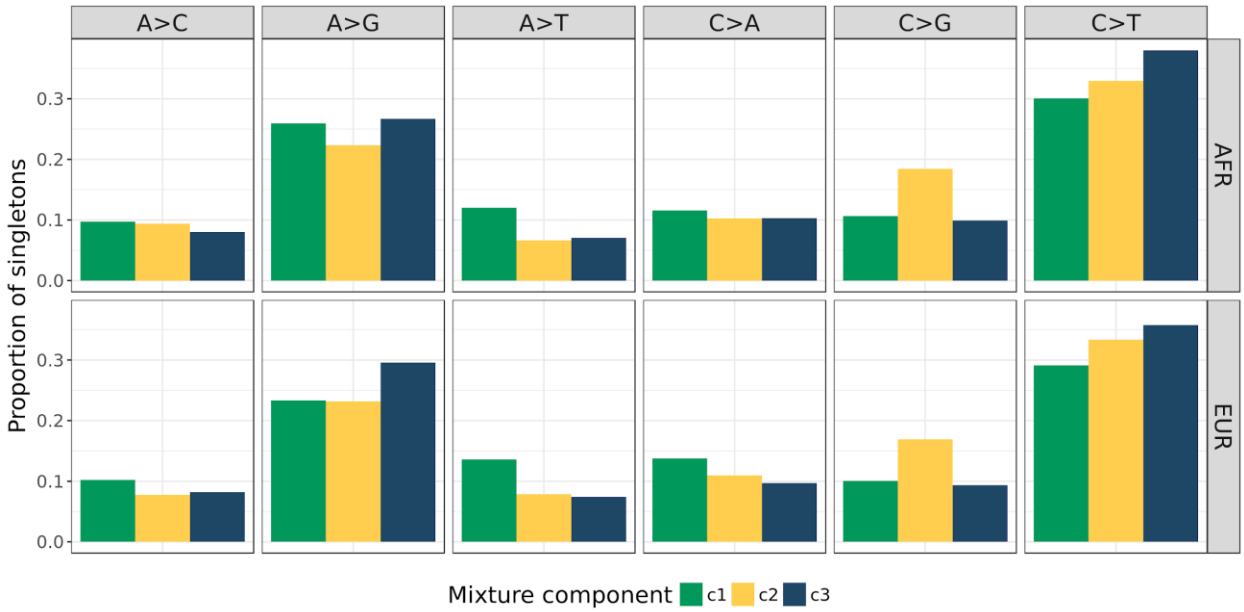


Figure 3.7. Mutation spectra of de novo mutation mixture components.

Other factors explaining the clustering patterns of singletons

Comparing the mixture parameters estimated from the singleton data with those estimated from de novo mutation data, we note two prominent discrepancies. First, the de novo mutations are accurately explained by just three mixture components, which we interpret as corresponding to the background independent mutation process (component 3), DSB-associated MNMs (component 2), and TLS-associated MNMs (component 1). Though the same three processes are implicated as unique mixture components contributing to the spatial variation of singletons in the TOPMed data, we consistently required an additional mixture component (with

an intrinsic mutation rate and lambda parameter intermediate to mixture components 2 and 4) to explain the observed distribution of inter-singleton distances.

In addition, the lambda parameters of mixture component 2 were notably lower among *de novo* mutations than among singletons, despite the similarities between intrinsic mutation rate parameters for these components. This discrepancy suggests that singletons are clustered at distances <20,000bp more often than we would expect if these are purely driven by multinucleotide mutation processes. Indeed, 31% and 42% of singletons in the TOPMed European and African ancestry subsamples occurred within 20,000bp of another singleton in the same individual, whereas past trio sequencing studies (and our present analysis of the TOPMed freeze 5 trios) suggest that only ~2-3% of *de novo* point mutations arise through multinucleotide mutation events [17, 22]. We note this discrepancy diminishes as the inter-singleton distance decreases: singletons occurring less than 100bp apart accounted for ~2% of all singletons, which is much closer to a prior estimate (1.8%) of the proportion of *de novo* point mutations expected to occur at such distances [21]. Tandem singletons account for 0.3% of singletons in the TOPMed data, which is within the range of previously-established tandem mutation rate estimates, typically around 0.2-0.4% [17, 111].

We propose two possible explanations for this enrichment of spatially-clustered singletons, both of which relate to the fact that singletons represent mutations that have accumulated in the population over the course of several past generations. First, we show that a substantial proportion of clustering can occur purely by chance due to the demographic history of the sample, even when all mutations are independent and generated by a single mutation process with uniform rate. Second, we demonstrate how these clustering patterns can arise stochastically through regional variation in the mutation rates.

By definition, a singleton variant occurs on an external branch of the coalescent tree describing the demographic history of a given sample of individuals. Assuming the mutation rate is constant over time, it follows that the number of singletons observed in each individual is a linear function of the external branch length (where the external branch length represents the amount of time in which new singleton variants, private to that individual, could have arisen). Individuals corresponding to longer external branches are therefore presumed to have accumulated more singletons, and hence are more likely to exhibit spatially-clustered independent singletons purely by chance. Intuitively, when the inter-singleton distances of all individuals in the sample are analyzed in aggregate, we would expect to see a general enrichment for spatially-clustered independent singletons.

To quantify the extent to which singleton clustering is attributable to the demographic history of the sample, we simulated the inter-singleton distance distribution as a mixture of 1,000 components, where the rate and lambda parameters of each component corresponded to the frequency of observed singletons in one of the individuals of a given ancestry from the TOPMed sample (**Methods**). Under this simulation, approximately 8.7% of singletons in the European ancestry subsample and 11.9% of singletons in the African ancestry subsample occurred within 20,000bp of another singleton in the same individual. Differences in sample demography would therefore appear to explain the discrepancy in the lambda parameters of singleton mixture components 2, 3, and 4, as shown in **Fig. 3.2**.

We also simulated the effect of external branch length heterogeneity on singleton clustering patterns under a general coalescent model, not specific to the observed singleton frequencies of individuals in the TOPMed sample. Using the msprime coalescent simulation library in Python [112], we simulated a sample of 1,000 diploid European genomes under a

feasible model of European demographic history [52] (**Methods**). This simulation generated 795,907 single-nucleotide variants, 432,597 (54.4%) of which were singletons (note that this abundance of singletons is to be expected for rapidly-expanding populations [52, 57]). Examining the inter-singleton distances of these simulated singletons, we found that 8.3% of singletons occurred within 20,000bp of another singleton in the same simulated individual. We attribute the slight discrepancy to uncertainty in the parameters implemented in this coalescent simulation. This simulation was particularly sensitive to the mutation rate we selected—varying the mutation rate from 1×10^{-8} to 2×10^{-8} increased the fraction of clustered singletons from 8.3% to 15.9%.

Assuming 3% of singletons are the result of multinucleotide mutation events and 8-10% occur at distances <20,000bp apart purely by chance due to heterogeneity of external branch lengths in the sample, this leaves more than half of the observed inter-singleton clusters in the TOPMed data (accounting for 31% of all singletons) unexplained. We hypothesized that the remaining clustering can be explained by regional heterogeneity in mutation rates. Consider, for example, a genomic region containing two singletons that arose independently in different generations. Based on the above coalescent simulations, if the underlying genome-wide average single-nucleotide mutation rate doubles from 1×10^{-8} to 2×10^{-8}, the probability that these singletons are separated by 20,000bp or less is expected to increase from 0.083 to 0.159.

As a proof of concept, we simulated the effects of mutation rate heterogeneity based on the per-individual empirical singleton frequencies from the TOPMed data as before, but here assumed that 10% of the genome is subject to a 2-fold increase in mutation rate. Under this simulation scenario, the fraction of singletons occurring <20,000bp from another singleton in the same individual increased from 8.7% to 9.5% in the European ancestry subsample and from

11.9% to 13.0% in the African ancestry subsample. If we instead assume 10% of the genome is subject to a 10-fold increase in mutation rate, the fraction of clustered singletons increases to 13.8% in the European ancestry subsample and 17.9% in the African ancestry subsample. These simulations demonstrate that regional heterogeneity in the rate of independent mutation processes can have a profound impact on the clustering patterns of SNVs observed throughout the genome.

Discussion

Multinucleotide mutations are a nontrivial source of genetic variation in the human genome, with unique mechanistic origins and a complex mutational footprint. In this study, we have investigated the spatial distribution of rare singleton SNVs throughout the genome with the goal of characterizing the fine-scale properties of MNMs, their variation throughout the genome, and how they influence the overall patterns of rare genetic variation. We showed that, in nearly all individuals, singleton densities can be parsimoniously represented as a mixture of just four basic processes. We provide extensive evidence that two of these processes appear to represent the effects of two particular multinucleotide mutation mechanisms: error-prone translesion synthesis (TLS), and errors induced during the repair of double-strand breaks (DSB).

There are several other properties of MNMs that we did not consider here, but that may prove useful in a deeper characterization of the causal mechanisms. First is the order in which the constituent point mutations occur within each MNM. Our analyses focused only on describing the basic mutation spectra (based on the 6 possible single-nucleotide mutation categories) of constituent point mutations, without taking into account their inherent ordering. Considering the ordered arrangements of the constituent point mutations of MNMs, however, it

is possible to classify MNMs into 6x6=36 possible categories, each of which could be independently analyzed for variation in their intrinsic mutation rate or genome-wide distribution. This granular classification strategy has already proven useful in confirming the distinctive tandem mutation signature of DNA polymerase zeta, which has a strong bias towards generating GA>TT and GC>AA mutations [17, 18, 21], but similar effects of other multinucleotide mutation processes have yet to be described. Furthermore, preliminary evidence from [94] and [18] has suggested that the composition and sequence of the non-mutated bases upstream, downstream, and within each MNM may carry additional information about the underlying mechanisms. We also note that, because our model is based on inter-singleton distance distributions and not discretely-defined mutation clusters, we do not attempt to describe the patterns and properties of MNMs consisting of more than 2 constituent point mutations. Similarly, a subset of MNMs have been shown to consist of complex combinations of point mutations and short insertions or deletions (indels) [17], potentially arising through other distinct mechanisms. Cataloging and analyzing these features of MNMs may help further elucidate relationship between different mutation signatures and the underlying multinucleotide mutation processes and provide insight into the extent to which each mechanism contributes to the overall burden of MNMs throughout the genome.

We have also investigated how singleton clusters ascribed to particular MNM processes vary regionally throughout the genome, confirming the presence of known DSB-associated mutation hotspots on chromosomes 2, 8, 9, and 16 [22, 23], as well as several novel hotspots with similar characteristics occurring in subtelomeric regions of other chromosomes (**Fig. 3.5**). These results fall into a broader body of evidence that telomeres are particularly prone to DSBs [102]. We found these regional patterns of clustering to be associated, to varying degrees, with

other features of the genomic landscape, such as CpG islands and exon density (**Fig. 3.6**), suggesting these features play a role in promoting or suppressing multinucleotide mutation mechanisms. We also observed that all 7 histone marks we considered were associated with different clustering patterns, consistent with earlier findings that chromatin structure affects the activity of both Y-family translesion polymerases in vitro [113] and DSB repair machinery [114]. Further experimental work is necessary to determine precisely how specific multinucleotide mutation mechanisms are impacted by the presence or absence of these features.

Our analysis of spatially non-independent mutations (and their distribution throughout the genome) might also be adapted to studying the patterns of temporally non-independent mutations, where mutations in one generation increase the likelihood of mutations in subsequent generations. Temporal non-independence of mutations can occur when damage and repair genes themselves acquire mutations that influence their efficacy or particular error biases. There is a growing body of evidence to suggest that the ongoing evolution of these genes has played a prominent role in shaping the observed patterns of human genetic variation [48, 80, 81, 115, 116]. These studies have proposed that, with sufficiently large samples of de novo mutations or rare variants, it may be possible to map loci responsible for variation in the rate or spectra of single-nucleotide variants [48, 116]. Given that MNMs are often attributed to very specific mutational pathways, implicating a handful of genes [16], we speculate that inter-individual variation in the rate and intrinsic properties of MNMs might be especially indicative of the presence of such mutator alleles.

Our use of unphased singletons in these analyses precluded our ability to confirm conclusively that any given cluster of singletons arose via a past multinucleotide mutation event, because unphased singletons, by nature, could also arise in different generations or on different

haplotypes. When we applied our statistical model to a new dataset of over 70,000 de novo mutations, however, we found strong evidence that two of the same processes responsible for singleton clustering patterns (i.e., components 1 and 2, which we interpreted as signatures of TLS-associated and DSB-associated multinucleotide mutations) were also responsible for de novo mutation clustering patterns. These similarities suggest that, even though many clustered singletons likely arose independently, our mixture modeling strategy successfully recovers the signals of past MNM events.

We concluded this study by demonstrating that patterns of clustering among singleton SNVs are not only the result of multinucleotide mutation processes, but also reflect the demographic history of the sample and variation in the local mutation rate. Through simulations, we estimated that approximately a third of inter-singleton intervals <20,000bp are purely attributable to variation in the demographic history of the sample. Assuming ~10% of inter-singleton intervals <20,000bp are the result of MNM events, this leaves over half of these intervals unexplained. We provided a proof-of-concept simulation to demonstrate that this remaining proportion of singleton clusters can be explained in part by stochastically-occurring clusters of independent singletons caused by variation in single-nucleotide mutation rates throughout the genome.

Other factors not considered here, such as selection, biased gene conversion, and evolving mutation rates over time are also expected to contribute to singleton clustering patterns. Because the signals of natural selection [50], biased gene conversion [117], local ancestry (and, more specifically, the external branch lengths) [118], single-nucleotide mutation rates [97] and multinucleotide mutation rates (**Fig. 3.5**) are continuously variable throughout the genome, we expect that in any given region there is a unique balance in the relative contributions of these

factors to the spatial distribution of singletons and other SNVs. Therefore, a deep understanding of how patterns of genetic variation are jointly influenced by regional heterogeneity and non-independence of mutation processes (both spatial and temporal) will play an important role in efforts to unambiguously discern signals of selection or demographic history.

Methods

Data

The TOPMed Freeze 3 dataset contains PCR-free, whole-genome sequencing data for 11,759 unrelated individuals. Global ancestry estimates for seven super-populations were obtained using RFMIX [98]. For our analyses, we selected two independent subsamples of 1,000 individuals European and African ancestry (>0.9), and recalculated the allele counts within each independent subsample. We chose to limit the subsample size to N=1000 for two reasons: first, to ensure each subsample shared homogenous ancestry, and second, to prevent biasing the singleton mutation spectra from recurrent mutations that occur in large samples [99].

Mixture Model Parameter Estimation

For each individual i, we collected the set of S singletons unique to that individual (with singleton status determined relative to other individuals from the same population subsample). Assuming singletons occur independently at a constant rate ϕ_i, we can model the probability of observing S singletons in individual i as a Poisson process:

$$f(s_i) = e^{-\phi_i G}(\phi_i G)^{S_i}/S_i!$$

where G is the size (in base pairs) of the mappable autosomal regions of the genome.

Then let D_i be a random variable describing the distance (in base pairs) between successive singletons in individual i. D_i follows an exponential distribution with rate $\theta_i = 1/\phi_i$:

$$f(d_i) = \theta_i e^{-\theta_i d_i}$$

Now suppose the set of S_i singletons are generated by $K > 1$ independent Poisson processes. For the subset of $S_{i,k}$ singletons resulting from process k, the successive distances D_k between these singletons follow an exponential distribution with rate $\theta_{i,k}$:

$$f_k(d_i;\ \theta_{i,k}) = \theta_{i,k} e^{-\theta_{i,k} d_i}$$

Then the distribution of inter-singleton distances across all S_i singletons is parameterized as a mixture of these K component distributions, given by:

$$f(d_i; \lambda_{\boldsymbol{i}}, \boldsymbol{\theta_i}) = \sum_{k=1}^{K} \lambda_{i,k}\, f_k(d_i; \theta_{i,k})$$

where $\theta_{i,1} < \theta_{i,2} < \ldots < \theta_{i,K}$ and $\lambda_{i,k} = S_{i,k}/S_i$ is the proportion of singletons resulting from process k, such that $\sum_{k=1}^{K} \lambda_{i,k} = 1$.

We estimate the parameters of this mixture $(\lambda_{i,1}, \ldots, \lambda_{i,K}, \theta_{i,1}, \ldots, \theta_{i,K})$ using the expectation-maximization (EM) algorithm as implemented in the mixtools R package [119]. To identify an optimal number of mixture components, we iteratively fit mixture models for increasing values of K and calculated the log-likelihood of observed data D given the parameter estimates $(\hat{\lambda}_{i,1}, \ldots, \hat{\lambda}_{i,K}, \hat{\theta}_{i,1}, \ldots, \hat{\theta}_{i,K})$, stopping at K components if the log-likelihood failed to increase by more than 10:

$$log(L(\hat{\lambda}_{\boldsymbol{i}}, \widehat{\boldsymbol{\theta}}_i \mid D_i, K+1)) - log(L(\hat{\lambda}_i, \hat{\theta}_i \mid D_i, K)) < 10$$

Classifying singletons by process-of-origin

Now let $k_{i,j}$indicate which of the four processes generated singleton j in individual i. We calculated the probability of being generated by process k as:

$$p(k_{i,j} = k \mid d_i^*;\ k \in \{1,\ldots,4\}) = \frac{p(d_i^*, k)}{p(d_i^*)} = \frac{\lambda_{i,k}\, f_k(d_i; \theta_{i,k})}{\sum_{k=1}^{4} \lambda_{i,k}\, f_k(d_i; \theta_{i,k})}$$

Where $d_i^* = d_{i,j} \wedge d_{i,j-1}$is the minimum of the distance either to the next singleton ($d_{i,j}$) or the previous singleton ($d_{i,j-1}$). We then classified the process-of-origin for each singleton per the following optimal decision rule:

$$\hat{k}_{i,j} = arg\ max_{k \in \{1,\ldots,4\}} p(k \mid d_{i,j}^*)$$

Identification of mixture component hotspots using Hidden Markov Models

Once each singleton was assigned to one of the four mixture components, we counted the number of singletons per individual per component in non-overlapping windows of length 1 million base pairs (Mpb) throughout the genome. In order to mitigate spurious signals caused by differences in local ancestry, in each window we omitted data from any individuals whose singleton frequencies were >2 standard deviations from the mean frequency across all individuals. We then assigned each window to one of three state (hot, neutral, or cold) with a 3-state Hidden Markov Model, as implemented in the depmixS4 package in R [120].

Modeling the relationship between MNM density and genomic features

In each 1Mbp window, we calculated the average signal for 11 genomic features (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3, exon density, DNase hypersensitivity, CpG island density, lamin-associated domain density), using the source datasets described in [97] (see **Table A.6**). For each mixture component, we then applied the following negative binomial regression model to estimate the effects of each feature on the density of that component in 1Mbp windows:

$$log(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{11} X_{11}$$

Where Y is the number of singletons of mixture component j and $X_1, \ldots, X_{11}$ are the signals of each of the 11 genomic features.

Note that although recombination rate and replication timing were analyzed in [97], we excluded these features from these regression models because the corresponding datasets typically exclude, at minimum, the first and last 5Mbp of each chromosome. Many of the hotspots identified by our HMM segregated near the ends of chromosome arms, so including these features in our regression models would force these hotspots to be excluded, potentially biasing the estimated effects of other genomic features.

De novo mutation calling

The TOPMed Freeze 5 data contained 1,675 parent-offspring trios. Among these trios, we considered de novo mutations to be any single-nucleotide variants that were exclusive to the offspring (i.e., not observed in either parent), provided the variant occurred at a site that was covered by an average read depth of 10 or higher and did not have a missing genotype in either

parent. We then restricted our analyses to autosomal de novo mutations ascertained in offspring determined to have >85% African or European ancestry.

Empirically-based simulations

To quantify the effects of external branch length heterogeneity on singleton clustering patterns, we simulated singletons under the following model. First, we can consider the N_i singletons in individual i to follow a Poisson(ϕ_i) distribution, where$\phi_i = \frac{N_i}{G}$ (here, G indicates the total number of mutable bases in the mappable autosomal regions of the reference genome). Consequently, the distances between successive singletons in individual i are expected to follow an exponential distribution with rate $\theta_i = 1/\phi_i$. For each individual i, we randomly drew N_i inter-singleton distances from the corresponding $exp(\theta_i)$ probability distribution.

We adapted this simulation strategy to simulate inter-singleton distances under a model of regional mutation rate heterogeneity. Here, inter-singleton distances were assumed to come from a 2-component mixture model. We randomly drew $0.9 \times N_i$ inter-singleton distances from an $exp(\theta_i)$ distribution, and $0.1 \times N_i$distances from an $exp(M \times \theta_i)$ distribution, where this second distribution represents 10% of singletons in individual i originating in genomic regions subject to a M-fold increase in single-nucleotide mutation rate.

Coalescent simulations

We used msprime [112] to simulate 2,000 European chromosomes (100Mbp in length) using a demographic model with parameter estimates reported by [52]. We performed simulations using a per-site, per generation mutation rate ranging from 1×10^{-8} to 2×10^{-8}. Because our aim was to compare these simulated singletons to unphased singletons in the

TOPMed data, we randomly assigned each of the 2,000 haploid samples into one of 1,000 diploid pairs, and recalculated the inter-singleton distances per diploid sample, ignoring the chromosome on which each simulated singleton originated.

Chapter IV.

Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets

Introduction[2]

The spectrum of somatic single-nucleotide variants (SNVs) in cancer genomes carries important information about the underlying mutation mechanisms, providing insight into the development, evolution, and etiology of the cancer cell populations [121]. Evaluating these patterns of variation, referred to as "mutational signatures," has become an important task in precision oncology, as mutational signatures can be used both to refine cancer diagnoses and identify effective targeted therapies [122].

Several software programs have been developed to identify and evaluate the mutational signatures present in cancer genomes [27–29]. Most methods consider 96 mutation subtypes, defined by the type of base change (C>A, C>G, C>T, T>A, T>C, T>G) and the trinucleotide sequence context (e.g., C[T>G]T, C[C>A]T, and so on) [25]. Mutation signature analysis methods express the observed mutation spectrum in each sample as a linear combination of K

[2]This chapter is published as a preprint at Carlson J, Li J, Zöllner S. Helmsman: fast and efficient generation of input matrices for mutation signature analysis. bioRxiv. 2018;:373076.

distinct mutational signatures, where the signatures are inferred directly from the input data, or taken from external sources such as the COSMIC mutational signature database [121]. These programs typically start with an input file, often in a standard format such as Variant Call Format (VCF) or Mutation Annotation Format (MAF), containing the genomic coordinates of each SNV and the sample(s) in which they occur. As a first step, these SNVs must be summarized into a NxS mutation spectra matrix, M, containing the frequencies of S different SNV subtypes in each of N unique samples (where the $M_{i,j}$ entry indicates the number of observed SNVs of subtype j in sample i). Most methods are implemented as R packages and must read the entire input file into memory prior to generating the mutation spectra matrix. For large input files, containing for example millions of SNVs and hundreds or thousands of samples, the memory required for this step can easily exceed the physical memory capacity of most servers, rendering such tools incapable of directly analyzing large datasets. To circumvent these computational bottlenecks, researchers must either limit their analyses to small samples, pool samples together, or develop new software to generate the mutation spectra matrix. Presently, the largest studies to perform mutation signature analysis have included millions of mutations in thousands of whole cancer genomes [26, 121], but these studies have pooled individual samples into ~30 distinct cancer types, potentially obscuring the presence of mutation signatures unique to individual cancer genomes or more granularly defined cancer types.

Implementation

To overcome the limitations of existing mutation signature analysis tools, we have developed a Python application, named *Helmsman*, for rapidly generating mutation spectra matrices and performing mutation signature analysis on arbitrarily large datasets. *Helmsman* accepts either VCF or MAF files as input.

For each SNV in a VCF file, *Helmsman* identifies the mutation type based on the reference and alternative alleles, then queries the corresponding reference genome for the trinucleotide context of the SNV, determining subtype j. The genotypes of the N samples for this SNV are represented as an integer array, with the number of alternative alleles per sample coded as 0, 1, or 2 according to the observed genotype [123]. *Helmsman* then updates the jth column of the mutation spectra matrix by vectorized addition of the genotype array (i.e., $M_{i,j}$ is incremented by 1 if individual i is heterozygous but does not change if individual i is homozygous for the reference allele). Consequently, *Helmsman's* processing time is independent of sample size and scales linearly with the number of SNVs. The only objects stored in memory are the array of N genotypes for the SNV being processed and the Nx96 mutation spectra matrix, so memory usage is independent of the number of SNVs and scales linearly with sample size.

Additional Features

In addition to being optimized for speed and low memory usage, *Helmsman* includes several features to accommodate various usage scenarios and minimize the amount of pre-processing necessary to analyze large mutation datasets. For example, if input data are spread across multiple files (e.g., by different sub-samples or genomic regions), *Helmsman* can process these files in parallel and aggregate them into a single mutation spectra matrix, providing additional performance improvements and avoiding the need to generate intermediate files. Similarly, in certain applications, it may be desirable to pool similar samples together (e.g., by tumor type) when generating the mutation spectra matrix. *Helmsman* can pool samples on-the-fly, without needing to pre-annotate or reshape the input file with the desired grouping variable.

Helmsman also includes basic functionality for extracting mutation signatures from the mutation spectra matrix using non-negative matrix factorization (NMF) or principal component

analysis (PCA) functions from the *nimfa* [124] and *scikit-learn* [125] Python libraries, respectively. Alternatively, *Helmsman* can generate an R script with all code necessary to load the output matrix into R and apply existing supervised and unsupervised mutation signature analysis packages (e.g., *SomaticSignatures* or *deconstructSigs*) without requiring users to perform the computationally expensive task of generating this matrix from within the R environment. All features are described in detail in the online documentation.

Results

We compared *Helmsman's* performance to that of three published R packages: *SomaticSignatures* [27], *deconstructSigs* [28], and *signeR* [29]. We also considered several other tools, and discuss their performance in **Appendix C**. For our tests, we generated a small VCF file (2.7MB compressed with bgzip) containing 15,971 germline SNVs on chromosome 22 from 2,504 samples sequenced in the 1000 Genomes Project phase 3 [55], and measured the runtime and memory usage necessary for each program to generate the mutation spectra matrix. We also attempted to run each program using the full chromosome 22 VCF file from the 1000 Genomes Project, containing 1,055,454 SNVs in 2,504 individuals. The number of SNVs in this VCF file is comparable to those of the large somatic SNV datasets analyzed in [121] and [26].

All programs generated the same mutation spectra matrices. *Helmsman* processed the small VCF file in 8 seconds, with a memory footprint of 140MB, and the full VCF file in 482 seconds (corresponding to a linear increase for ~60x more variants) with no increase in memory usage as the sample size remained the same. In contrast, to process the small VCF file, *SomaticSignatures* took 227 seconds with a memory footprint of 18GB, *deconstructSigs* took 2,376 seconds and 7.5GB of memory, and *signeR* took 1,740 seconds and 10.2GB of memory (**Fig. 4.1**). None of these R packages were able to load the full VCF file due to memory

allocation errors. All other tools we considered showed similar performance bottlenecks when compared to *Helmsman* (**Appendix C**; **Fig. C.1**).

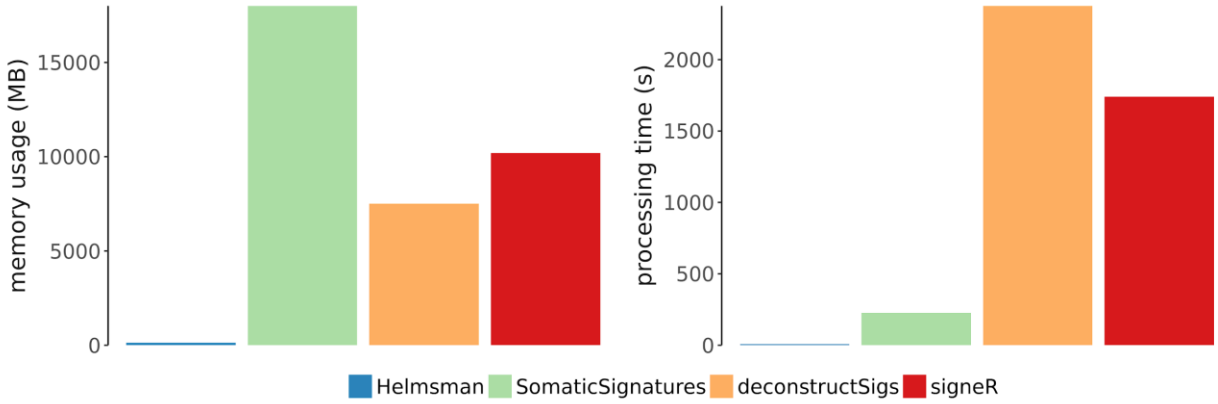


Figure 4.1. Performance comparison for generation of the mutation spectra matrix by different programs. For *Helmsman* and three other mutation signature analysis tools (*SomaticSignatures*, *deconstructSigs*, and *signeR*), we measured the maximum memory usage in megabytes (**a**) and processing time in seconds (**b**) required to generate the 2,504 x 96 mutation spectra matrix from a VCF file containing 15,971 SNVs in 2,504 samples from the 1000 Genomes project.

To further highlight the speed and efficiency of *Helmsman* for large datasets, we evaluated the entire set of 36,820,990 autosomal biallelic SNVs from the 1000 Genomes phase 3 dataset (14.4 GB when compressed with bgzip). Using 22 CPUs (one per chromosome VCF file), *Helmsman* generated the mutation spectra matrix in 64 minutes (approximately 1.5 seconds per sample), with each process requiring <200MB of memory.

Conclusions

As massive sequencing datasets become increasingly common in areas of cancer genomics and precision oncology, there is a growing need for software tools that scale accordingly and can be integrated into automated workflows. Our program, *Helmsman*, provides an efficient, standardized framework for performing mutation signature analysis on arbitrarily large, multi-sample VCF or MAF files. For small datasets, *Helmsman* performs this task up to

300 times faster than existing methods and is the only tool that can be directly applied to modern large sequencing datasets.

Availability of data and materials

The chromosome 22 VCF file from the 1000 Genomes Phase 3 study used in evaluating the software is available in the 1000 Genomes FTP repository at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. The small VCF file is available on the project home page at https://github.com/carjed/helmsman/tree/master/data. The MAF file used to compare performance of the Maftools and Mutation-Signatures software (described in the Supplementary Material) is available from The Cancer Genome Atlas data repository at https://portal.gdc.cancer.gov/legacy-archive/files/15ce66c6-0211-4f03-bd41-568d0818a044.

Chapter V.

Doomsayer: quality control for whole-genome sequencing data using mutation signature analysis

Background

Detection of genetic variants by whole-genome sequencing (WGS) is a complex task consisting of several experimental and computational processing steps [126]. Rigorous quality control (QC) must be performed at each stage of the sequencing pipeline to limit the influence of false positive variants and batch effects in downstream analyses [127]. While many applications are devoted to QC of raw sequencing data, mapped reads, and individual variants, methods for assessing the overall quality of each sample are a crucial, yet often overlooked, step to ensure that downstream analyses are performed on clean, unbiased data [128].

A common strategy for evaluating the quality of individual samples is to calculate various sample-level metrics which act as surrogates for sequencing quality, then remove samples which exhibit extreme values of one or more of these statistics based on guidelines established by previous studies [128]. One of the most widely-used quality metrics is the transition:transversion (Ti:Tv) ratio, calculated as the number of transition single nucleotide variants (SNVs) (purine ↔ purine or pyrimidine ↔ pyrimidine) divided by the number of transversion SNVs (purine ↔ pyrimidine) [129]. Prior studies have suggested that high-quality

genome-wide human sequencing data should have a Ti:Tv ratio of approximately 2.0-2.1 [54]. The Ti:Tv ratio is particularly useful for sample-level QC because it is often directly impacted by erroneous SNV calls—if we assume that errors occur at random with no biases towards transitions or transversions, then a sample whose SNV calls consist entirely of errors should have a Ti:Tv ratio near 0.5, because there are 8 possible transversions and only 4 possible transitions. So, a Ti:Tv ratio below 2.0 suggests that some of the included variants are random errors.

There are two major limitations to using the Ti:Tv ratio for sample QC, however. First, the assumption that sequencing errors are purely random is inaccurate. Certain technical artifacts are known to exhibit unique type-specific and motif-specific error profiles, such as a slightly elevated A>C error rate inherent to many Illumina sequencers and a tendency for higher miscall rates at bases preceded by a GGC motif [130, 131]. Because the Ti:Tv ratio does not differentiate by SNV type (e.g., C>G versus C>A transversions) or by sequence context, it lacks the sensitivity to distinguish these subtler patterns of error. Further, because the Ti:Tv ratio parses SNVs into only two categories, the Ti:Tv ratio can appear to indicate a high-quality sample even when the underlying data disagree. Consider an extreme case where a sample's SNV calls consist entirely of T>C transitions and T>G transversions. These SNVs could still have a Ti:Tv ratio within the acceptable range of 2.0-2.1, but the SNVs clearly do not represent what we would expect for a normal human genome.

Second, because sequencing errors often manifest as rare, low-frequency SNVs [132], sample-level Ti:Tv ratios are most informative when calculated exclusively on rare SNVs occurring at low frequencies in the population [33]. Moreover, the Ti:Tv ratio is sensitive to the number of samples sequenced [95], and particularly so when considering only rare SNVs [99].

Consequently, an acceptable range for the Ti:Tv ratio of rare SNVs is a moving target, and the consensus of 2.0-2.1 is almost certainly too conservative for most modern sequencing studies.

Filtering criteria for other sample-level QC metrics are similarly difficult to establish. Summary statistics which measure specific technical issues, such as GC-biased coverage and sample contamination, are influenced by the environment in which DNA samples are collected, stored, and sequenced [88, 133]. Measures of genotype concordance are another useful QC statistic, but can only be used in studies where samples are genotyped on multiple platforms, and vary according to the sequencing platform, variant calling algorithm, depth of coverage, and genotyping array [134]. Due to the vast range of factors that can vary between different whole-genome sequencing studies, guidelines suggested in one study may not translate well to future studies [128].

Here we present a novel computational method for performing unsupervised outlier detection in whole-genome sequencing studies. We have implemented this method in a software package called *Doomsayer* (Detection Of Outliers using Mutation Signature AnalYsis in Extremely Rare variants). Our method provides several benefits over traditional supervised outlier detection approaches, including: 1) Doomsayer is agnostic to sequencing platform, sample size, and other study conditions, 2) accounts for potential type- and motif-specific error biases, 3) focuses exclusively on rare SNVs, and 4) provides a means of interpreting potential error biases present among the rare SNV calls in a sample.

Results

Overview of the method

The outlier detection procedure we have designed and implemented in *Doomsayer* is inspired by a popular computational method in cancer genomics known as mutation signature analysis, used to characterize mutational processes unique to various cancer types [25, 121, 135]. This method starts by classifying SNVs into one of 96 3-mer subtypes, defined according to the type of base substitution (e.g., C>A, T>G, and so on) and the flanking nucleotides in the reference sequence (e.g., C<u>C</u>T>C<u>A</u>T, C<u>T</u>T>C<u>G</u>T, and so on). Note that for simplicity each subtype is referred to by the substitution occurring at the pyrimidine of the base pair, so C<u>C</u>T>C<u>A</u>T SNVs are counted together with G>T SNVs occurring at the reverse complement of the CCT motif (i.e., A<u>G</u>G> A<u>T</u>G and C<u>C</u>T>C<u>A</u>T are equivalent). The relative frequencies of these subtypes within each sample are referred to as the mutation spectrum. For our QC purposes, we specifically focus on singleton SNVs (i.e., SNVs unique to a given sample with a minor allele count of 1). The singleton SNV spectra for the N samples are compiled into a Nx96 matrix M, where the $M_{i,j}$ entry indicates the relative proportion of observed SNVs of subtype j in sample i. The next step is to approximate the observed singleton SNV spectrum of each sample as a linear combination of R mutation signatures, where R<<96. These signatures (and their relative contributions to the singleton SNV spectrum of each sample) are ascertained by applying various dimensionality reduction algorithms to the M matrix; the most commonly used algorithms for this task are principal component analysis (PCA) or nonnegative matrix factorization (NMF).

Just as we assume the Ti:Tv ratio of germline SNVs to be consistent between samples (and interpret extreme values of the Ti:Tv ratio as evidence of higher error rates), *Doomsayer* relies on the assumption that the 3-mer SNV spectra (and the contributions of the decomposed signatures thereof) are also consistent between samples. This assumption is well-supported by results from family-based sequencing studies, which have shown that there is little inter-sample heterogeneity in the 3-mer mutation spectra of *de novo* germline mutations [41]. Hence, when summarizing the observed singleton SNV spectra as a linear combination of R underlying signatures, if the contributions of these signatures in a given sample are heavily skewed, we interpret this as an indicator of technical artifacts that have resulted in error biases throughout that sample's sequence.

Therefore, the final step in our outlier detection process is to determine which samples exhibit signature contributions that are considered extreme relative to other samples in the dataset. To this end, we apply two unsupervised anomaly detection algorithms to the results of the signature deconvolution procedure and flag as outliers any samples classified as anomalous by these algorithms. Finally, we provide an option to generate a diagnostic report, containing interactive plots and summary information about the flagged outliers, to assist users in interpreting the underlying patterns in the outliers' singleton SNV spectra and compare these differences to the non-outlier samples. A graphical summary of the Doomsayer workflow is presented in **Fig. 5.1**, and each stage is described in greater detail in the **Methods**.

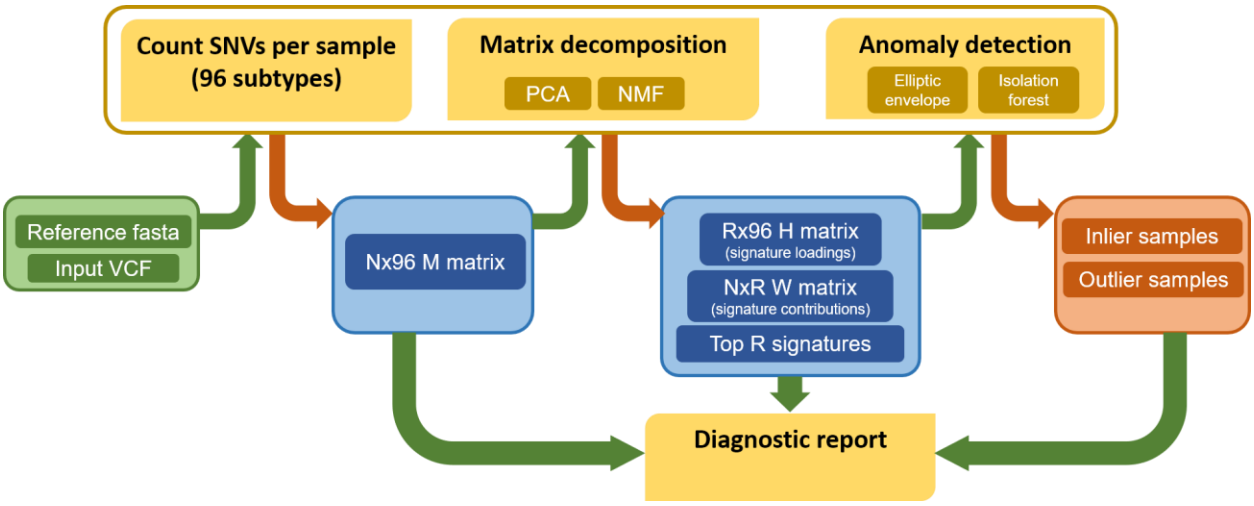


Figure 5.1. Summary of the Doomsayer workflow. Users must provide an input VCF file (or multiple files), along with a corresponding fasta-formatted reference genome file. Doomsayer then generates the singleton SNV spectra matrix, performs matrix decomposition to identify the common signatures in the data, then applies anomaly detection methods to identify samples with anomalous contributions of one or more signature. Finally, the intermediate output generated by this pipeline is passed to an R script which generates an HTML-formatted diagnostic report containing further information about the singleton SNV spectra of the flagged outliers.

Doomsayer identifies signatures of GC-biased coverage and oxidative damage in the BRIDGES dataset

To demonstrate the efficacy of our method, we applied *Doomsayer* to a whole-genome sequencing dataset of N=3,765 unrelated individuals of European ancestry, sequenced as part of the *Bipolar Research in Deep Genome and Epigenome Sequencing* (BRIDGES) study [97]. These samples had already passed through several standard QC procedures applied to the raw data and aligned reads (e.g., removal of duplicated reads and recalibration of base quality scores [33]) and individual variant calls (using the default hard filters of the GotCloud variant calling pipeline followed by a support vector machine filter trained on known SNVs from dbSNP as positive examples and SNVs that failed multiple hard filters as negative examples [33]). In addition, we performed preliminary sample-level filtering for population outliers, highly contaminated samples, improperly generated BAM files, and sequences with low coverage or

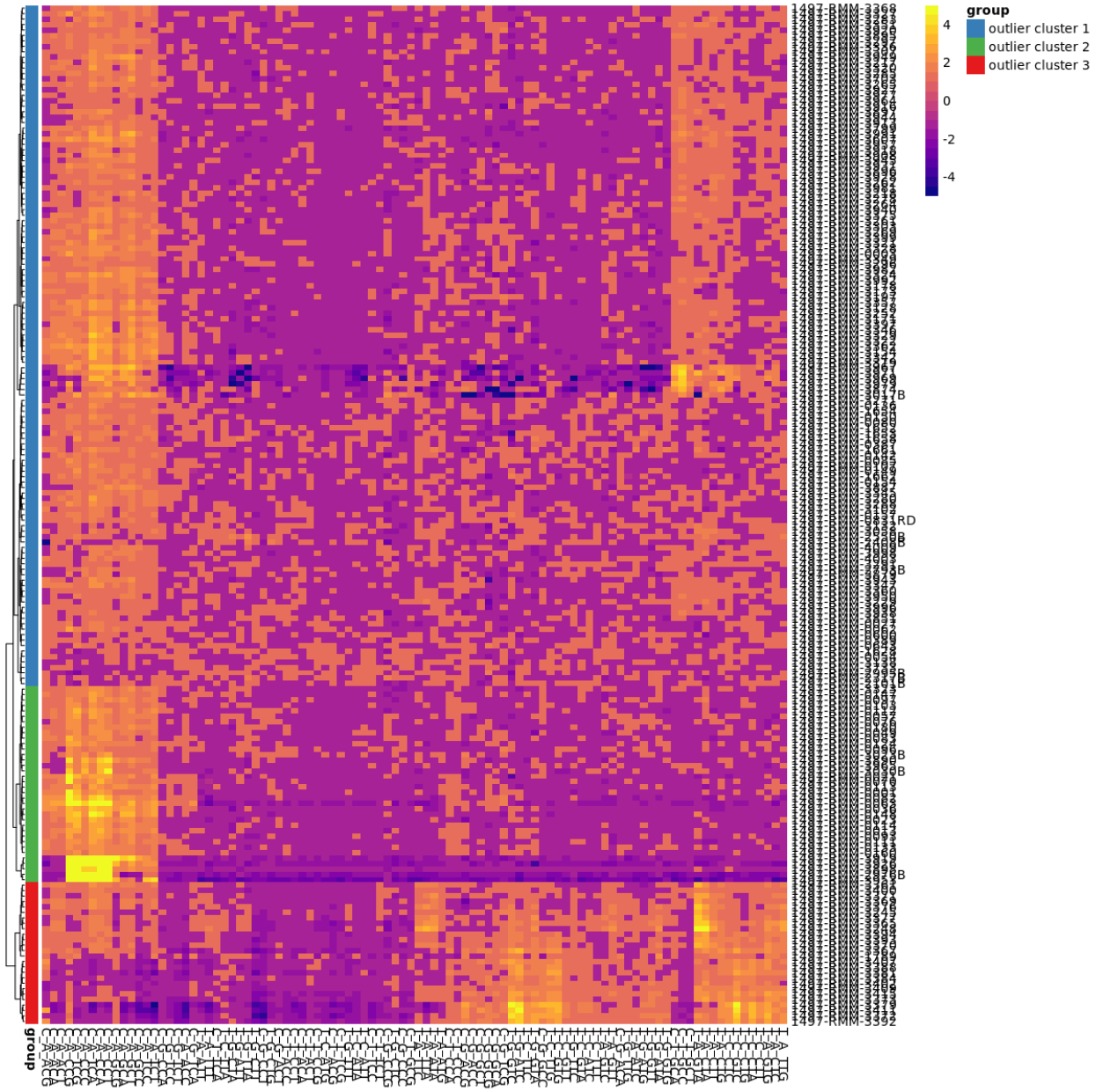
imbalanced read counts (as described in [97]). This dataset contained approximately 36 million singleton SNVs, or an average of ~10,000 singletons per sample. Note that the analyses presented here were based on this preliminary version of the BRIDGES data (N=3,765 samples) and the results contributed to determining which samples to include in the final release of the BRIDGES data (containing N=3,560 samples) [97].

Using the default parameters (stringency threshold tau=0.05; sample must be flagged by both anomaly detection algorithms to be considered an outlier), Doomsayer identified 157 outliers (**Fig. 5.2**). Hierarchical clustering of the flagged outliers (and their observed singleton SNV spectra) assigned these samples into one of three major clusters (**Fig. 5.2a**), with members of each cluster exhibiting similar patterns of enrichment or depletion for particular 3-mer subtypes. Outliers in the first cluster contain an excess of C>A singletons, a well-known artifact of oxidative DNA damage resulting from the buffering conditions during DNA sonication [136, 137], as well as CTN>CAN transversions and GCN>GTN transitions. The latter subtypes may be an indicator of an error phenomenon known as T-accumulation, in which adenine and cytosine tend to be miscalled at higher rates as thymine during later sequencing cycles [130]. A second cluster show a much stronger enrichment for C>A transversions, but no enrichment for T>A and C>T singletons as observed in the first cluster. A third cluster of outliers show a general tendency to contain more singletons at A:T base pairs than G:C base pairs, compared to the non-outlier samples.

Each of the three outlier clusters segregated along one or two of the top 3 principal component axes (**Fig. 5.2b**). Outlier clusters 1 and 2 had higher scores along principal component 1, whereas outlier cluster 3 had higher scores along principal component 2. Principal component 3 appeared to separate outlier clusters 1 and 2. We note that the set of outliers

determined by the first 3 principal components were highly concordant with those determined using a rank 3 non-negative matrix factorization, with 87.2% of PCA-detected outliers also being detected when using the NMF algorithm. Similarly, increasing the number of principal components to 4 or 5 had little effect on the set of outliers, with 92.4% and 88.6% overlap, respectively.

a.


group
outlier cluster 1
outlier cluster 2
outlier cluster 3
4
2
0
-2
-4
group

b.

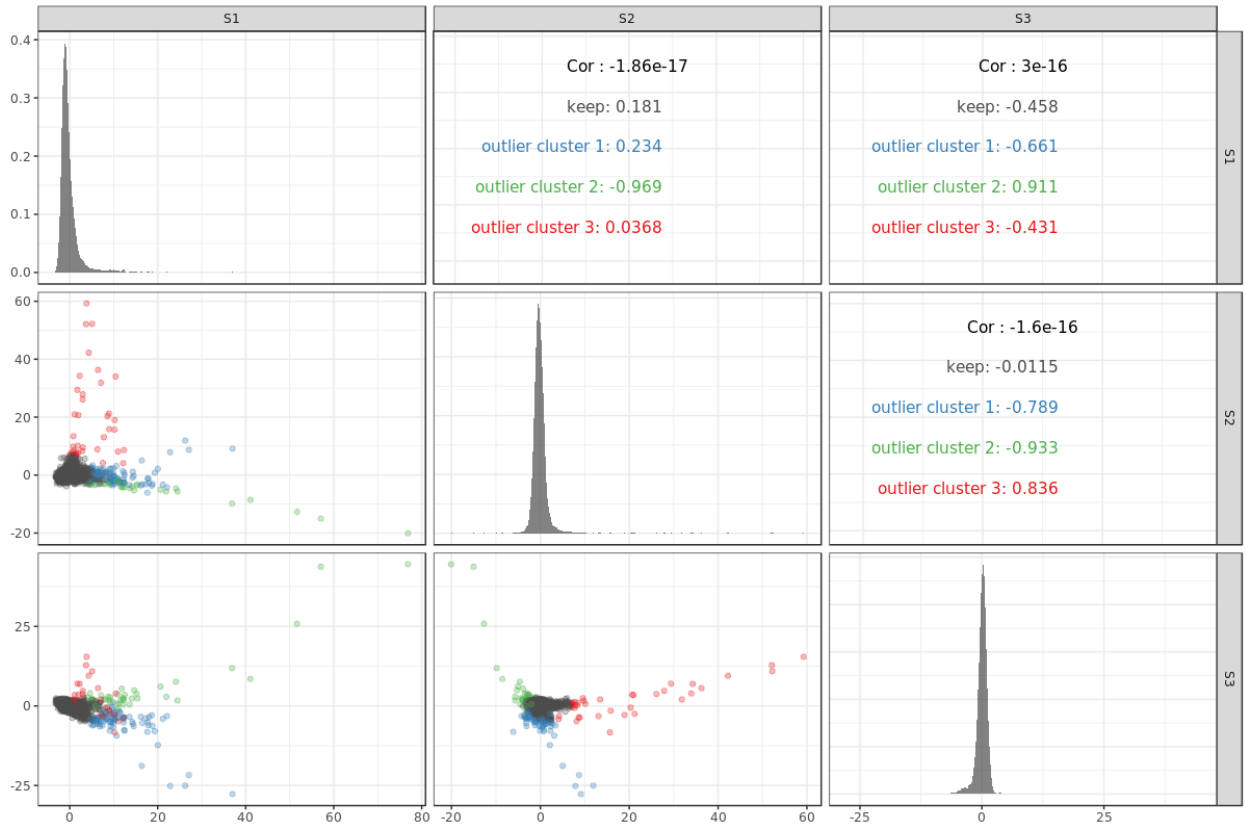


Figure 5.2. **Summary of outliers in the BRIDGES data**. **a.** Heatmap showing the relative enrichment or depletion for each of the 96 3-mer subtypes (columns) in each of the 174 flagged outliers (rows). The color in each cell indicates the fold-difference of the contribution of the subtype in that sample, calculated relative to the mean contribution of that subtype across all non-outlier samples. The scale is truncated to +/-5-fold difference. Outliers fall into one of three major clusters, indicated by the colored bar below the dendrogram to the right of the heatmap: high C>A transversions, GCN>GTN transitions, CTN>CAN transversions (green), high C>A transversions (blue), and high T>N transitions and transversions (red). **b.** Pairwise scatterplots of the first 3 principal components. Each dot indicates an individual, with non-outlier samples colored grey and outliers colored according to the cluster membership indicated in **a**. Marginal distributions for each component are shown in the diagonal panels.

BRIDGES outliers are supported by other QC statistics

We corroborated our findings by examining other sample-level summary statistics that are often used to differentiate low-quality samples. We examined nine such statistics: average depth-of-coverage, GC-bias scores (an indicator of whether coverage is systematically lower GC-rich regions [92]), median insert size, number of bases sequenced with base quality score >20 (q20bases), and percent of reads with a mapping quality score of 0 (zeromap), all obtained using the QPLOT software [138]; sample contamination, measured using the CHIPMIX software

[88]; and total singletons, heterozygosity, and singleton Ti:Tv ratio, all obtained using the vcfast submodule of the EPACTS software [139]. We hypothesized that if the samples flagged as outliers by our method were truly affected by quality issues, they would likely differ from non-outlier samples across these established sample-level summary statistics.

When we compared the distributions of these QC statistics between the non-outlier samples and the three major outlier clusters detected by Doomsayer, we found that each outlier cluster showed anomalous distributions of one or more of these statistics (**Fig. 5.3**). Samples in outlier cluster 1, characterized by a combination of modestly higher proportions of C>A transversions and certain C>T and T>A subtypes, showed lower singleton Ti:Tv ratios. Samples in outlier cluster 2, with very high proportions of C>A transversions, showed lower singleton Ti:Tv ratios, lower median insert sizes and slightly elevated contamination scores. Samples in outlier cluster 3, with high proportions of T>N singletons, tended to have higher GC bias scores, lower heterozygosity, and higher overall singleton counts.

These results are further supported by an orthogonal quality control analysis of the BRIDGES data, where we found that 27 of the 35 cluster 3 outliers also carried an unusually large number (>500) of artefactual copy number variants (CNVs) (Zawistowski et al., in preparation). These 27 samples originated from Prechter sub-study of the BRIDGES Consortium, and 25 of these were sequenced on the same plate (**Table D.1**), suggesting a systematic batch effect led to both the excess of A>N singletons and CNV artifacts.

Outlier clusters 1 and 2 also showed evidence of batch effects. In outlier cluster 1, three plates (two from the Prechter study and one from the USC study) accounted for 54 of the 75 outliers. We note that the two plates from the Prechter study (accounting for 30 of these 54 cluster 1 outliers) were the same as those which suggested batch effects among the cluster 3

outliers. In outlier cluster 2, two different plates containing samples from the Prechter study accounted for 24 of the 38 outliers.

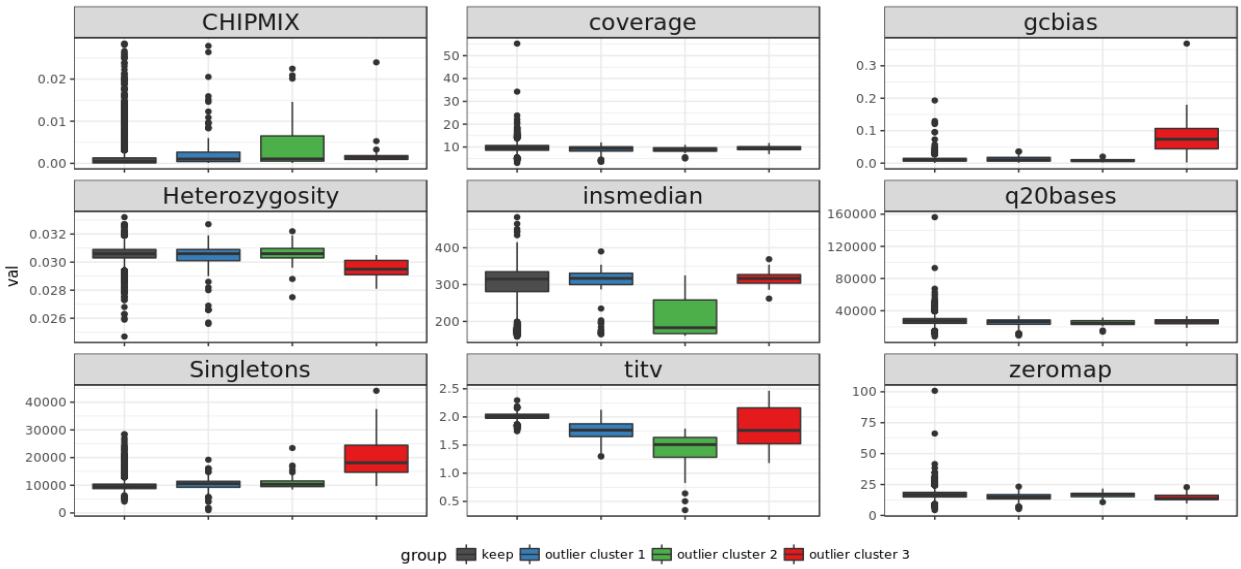


Figure 5.3. **Comparison of QC metrics for outliers**. Boxplots comparing the distributions of other sample-level summary statistics (contamination [measured by CHIPMIX], average coverage, number of doubletons, GC bias, median insert size, and number of singletons) of non-outlier samples and the three outlier clusters detected by our method.

Error signatures in the 1000 Genomes sample

Next, we applied *Doomsayer* to N=2,504 human genomes sequenced in phase 3 of the 1000 Genomes Project [55]. Because these data have gone through extensive cleaning and QC measures, we ran *Doomsayer* with more stringent criteria, setting the stringency threshold tau to be 0.01 (i.e., no more than 1%, or 25 samples, would be flagged as outliers). Under these conditions, *Doomsayer* identified 19 samples as outliers. These results are again based on the first 3 principal components.

A single sample, HG01149, occupied its own outlier cluster and contained a large excess of C>A transversions, suggesting the presence of the same oxidative damage signature implicated in outlier clusters 1 and 2 from the BRIDGES data. We note that this sample was also

flagged as an outlier in a previous analysis of rare variant signatures in 1000 Genomes Phase 3 data, based on f2 (i.e., doubletons, with an allele count of 2) and f3 (i.e., tripletons, with an allele count of 3) variants [81], though that study did not indicate the nature of the particular signature found to be overrepresented in HG01149. Three other outliers showed a similar enrichment for C>A singletons, suggesting they too might have been affected by oxidative damage, albeit to a much lower degree than what we observed in HG01149.

The fifteen remaining outliers showed an enrichment for T>G transversions, particularly at NTT trinucleotide motifs. Importantly, these outliers were unrelated and from diverse populations (**Table D.2**), and so did not appear to be recapitulating any of the population-specific rare variant signatures described by Mathieson and Reich [81]. Mathieson and Reich also performed a PCA-based analysis of singleton SNVs from 300 individuals sequenced by the Simons Genome Diversity Project and found that principal components 1 and 2 differentiated cell-line versus non-cell-line derived samples, suggesting that cell line artifacts may substantially influence the observed rare variant spectrum in a given sample [81]. We repeated this analysis using singleton SNVs from the 1000 Genomes samples and found that blood-derived and LCL-derived samples tended to separate along principal component 3 (**Fig. D.1**). Moreover, the outlier samples enriched for T>G singletons had some of the highest scores for principal component 3, supporting Mathieson and Reich's claim and providing further evidence that our outlier detection method is capable of capturing these cryptic technical artifacts.

Application to PCR-free whole-genome data

Several important advances in sequencing technology have occurred since the BRIDGES and 1000 Genomes datasets were generated, namely the use of PCR-free library preparation methods and increasing accessibility of high-coverage sequencing. Both of these advances have significantly reduced the incidence of technical artifacts in next-generation sequencing data, particularly in the identification and analysis of rare variants [140]. To assess how well our method performs on such data, we applied *Doomsayer* to high-coverage (~30x), PCR-free whole-genome sequencing data from N=3,286 individuals in the Framingham Heart Study, sequenced as part of the Trans-omics in Precision Medicine (TOPMed) consortium. We excluded 389 samples with fewer than 1,000 singleton SNVs. As with the 1000 Genomes dataset, we used more stringent outlier detection criteria, setting tau=0.01.

Our anomaly detection criteria identified a preliminary set of 27 outlier samples. Application of the secondary cluster-based filtering criteria (**Methods**) found only a single outlier cluster, containing seven samples, was significantly enriched for multiple 3-mer subtypes when compared to the non-outlier samples (**Fig. D.2**). Specifically, these outliers were enriched for T>A and T>G transversions at NTT and NTA motifs, with a particularly strong enrichment at the C[T>G]T subtype (**Fig. D.2**).

Coincidentally, all 7 of these outliers were parents within 7 different parent-offspring trios sequenced in the Framingham Heart Study. This enabled us to assess whether the singleton SNV spectra of these samples was the result of genuine biological differences or technical artifacts. We hypothesized that if these patterns of variation were biological in nature, the enriched T>A and T>G variants should be transmitted at normal Mendelian proportions. In that case, the spectra of singleton SNVs in the outlier parents (i.e., variants not transmitted to the

offspring) should be similar to that of the private doubleton SNVs shared between an outlier parent and their offspring (i.e., transmitted variants). Significant differences would violate the law of Mendelian transmission and indicate that the singleton SNV spectra in the outlier parents is driven by technical artifacts. Among the 7 parent-child dyads, we identified 25,550 private doubleton SNVs (compared to 36,045 singleton SNVs in the parents) and compared the spectrum to that of the outlier singleton spectrum using chi-squared tests for equal proportions across the 6 basic SNV types (**Table 5.1**). Results of this test were highly significant (P<5.1e-70), indicating the non-transmitted SNV spectra in the outlier parents were likely enriched for sequencing errors. These differences were even more significant when considering the full 96-subtype 3-mer spectra (P<9.1e-81). In contrast, when we compared the spectra of SNVs transmitted from outlier parents to their offspring in these 7 trios with that of the non-transmitted SNVs unique to the 7 non-outlier parents, we found no evidence of significant differences in either the 1-mer (chi-squared test; 5df; P=0.16; **Table D.3**) or 3-mer (chi-squared test; 95df; P=0.45). These additional tests confirmed that the spectra of SNVs shared by offspring with their outlier parents were of high quality and unlikely to contain other cryptic error signatures.

Table 5.1. 1-mer spectra of non-transmitted and transmitted SNVs in the FHS outliers

SNV Type	Frequency in non-transmitted SNVs (% of total)	Frequency in transmitted SNVs (% of total)
C>A	3920 (10.9%)	2495 (9.8%)
C>G	2979 (8.3%)	2328 (9.1%)
C>T	12572 (34.9%)	9888 (38.7%)
T>A	3118 (8.6%)	1718 (6.7%)
T>C	9592 (26.6%)	7193 (28.2%)
T>G	3864 (10.7%)	1928 (7.5%)
Total	36045	25550

Discussion

Doomsayer provides a new approach to quality control in next-generation sequencing studies that circumvents many of the limitations of standard sample-level QC strategies. By leveraging information about the spectra and sequence context of rare single nucleotide variants, *Doomsayer* effectively identifies outliers and batch effects that may go undetected by more simplistic supervised QC procedures.

In our analyses of the BRIDGES, 1000 Genomes, and Framingham Heart Study datasets, we found compelling evidence that *Doomsayer* identifies low-quality samples that went undetected by previously-applied QC checks. In both the BRIDGES and 1000 Genomes datasets, we identified samples with an abnormal excess of C>A singletons, presumed to stem from oxidative damage that occurred during sample preparation, suggesting this may be a common artifact in whole-genome sequencing data where PCR was used to amplify the DNA of each sample. Many of the outliers detected in the BRIDGES dataset were sequenced in the same plates, indicating that our method may be particularly powerful for identifying cryptic batch effects. Outliers in the 1000 Genomes data also included samples with an abnormal excess of A>N singletons, possibly indicating artifacts of sequencing DNA obtained from cell lines rather than fresh tissue. Our findings of such artifacts in the 1000 Genomes dataset may have broader implications for genomics research, as the 1000 Genomes data are widely used as a benchmark of known genetic variation for tasks ranging from genotype imputation to population genetics inference to development of pathogenicity scoring algorithms. Research involving rare SNVs from the 1000 Genomes data may be particularly susceptible to spurious signals or subtle biases caused by including these outliers.

In the Framingham Heart Study dataset, we identified 7 samples with significantly elevated rates of T>G singletons. Because the offspring of these samples were also sequenced, we were able to compare the spectra of transmitted and non-transmitted SNVs and confirm the presence of a cryptic error signature shared by these outlier samples. Intriguingly, the singleton SNV spectrum of these outliers was quite similar to that of the outliers in the 1000 Genomes dataset that we speculate were affected by cell-line artifacts. None of the Framingham Heart Study outliers, however, were sequenced from cell lines, so the origin of the error signature remains unclear. These findings demonstrate that our method is an effective means of quality control and sequencing error detection even when applied to high-coverage data prepared with PCR-free protocols.

There are a few limitations of our method that bear further discussion. First, as with all post hoc computational QC methods, our method is no replacement for careful study design and rigorous quality control throughout the earlier stages of the sequencing pipeline. Because unsupervised anomaly detection algorithms assume that anomalous samples account for a relatively small fraction of the data, severe quality issues, affecting a large fraction of samples, can be underestimated.

Related to this point, unsupervised anomaly detection also runs the risk of being overly conservative or liberal with flagging outliers. Though one of our main motivations in developing Doomsayer was to improve the objectivity of sample-level QC practices and provide deeper investigation of potential quality issues, the decision of which samples to exclude from downstream analyses remains inherently subjective. These decisions must often be balanced against competing considerations such as loss of power in downstream analyses due to sample

size reduction and the financial and practical feasibility of resequencing samples deemed to be low-quality.

Second, *Doomsayer* is unlikely to be an effective means of sample-level QC in whole-exome sequencing datasets. In studies containing a few thousand individuals, a single exome is expected to contain roughly 100 singleton SNVs. Because *Doomsayer* parses singleton SNVs into 96 distinct subtypes, a single exome will not carry enough observations per subtype to confidently summarize the 3-mer singleton SNV spectrum. As a potential means of circumventing this limitation, we have included a feature in Doomsayer to pool groups of samples together and perform outlier detection on a per-group rather than per-individual basis. With this feature, Doomsayer can still be used to detect systematic batch effects in whole-exome sequencing datasets.

Third, we acknowledge that several recent studies have found evidence that different human populations tend to have slightly different spectra among higher-frequency SNVs, suggesting rapidly evolving mutational processes [48, 81, 115]. Though these findings might appear to challenge our assumption that the singleton SNV spectra are consistent between individuals regardless of population, our analysis of the 1000 Genomes data show that the detected outliers come from diverse ancestral backgrounds. Nevertheless, in future studies that use *Doomsayer* for QC in cohorts with heterogeneous ancestries, it may be prudent to evaluate each population separately to avoid conflating genuine population differences in singleton SNV spectra with technical artifacts present among the singleton SNV calls.

Finally, we emphasize that *Doomsayer* is specifically designed for evaluating the quality of germline, not somatic, genomes. The genomes of somatic cells are exposed to a variety of mutagenic processes, so what might be considered an anomaly in a collection of diverse somatic

genomes may very likely be due to differing levels of exposure to different mutagenic processes. However, *Doomsayer* may still be useful in instances where the somatic genomes are homogeneous in cell type and exposure (for example, lung tumors collected from individuals who smoke tobacco). In such a case, we advise users carefully assess their data and apply other appropriate QC measures to evaluate whether outliers are due to technical artifacts, or whether they represent distinct etiological differences.

Conclusions

We here present Doomsayer, a novel method for performing unsupervised sample-level quality control in large-scale next-generation sequencing studies through the use of mutation signature analysis. We provide evidence that the spectrum of singleton SNVs in an individual genome is sensitive to a variety of systematic error processes. Our method to evaluate this spectrum thus allows not only for the identification of problematic samples, but also deeper investigation of how the single-nucleotide variant calls might be affected by cryptic technical artifacts. We anticipate that this method, when applied in conjunction with rigorous quality control throughout the sequencing and data cleaning process, will help improve the veracity of both old and new next-generation sequencing datasets, leading to more robust and replicable scientific results.

Methods

Analysis pipeline

Doomsayer is implemented in Python and accepts one or more variant call format (VCF) files containing the variants to analyze. Each variant site in the input VCF file(s) must contain columns indicating the genotypes for each individual; the VCF file(s) cannot be in a "sites only" format. Doomsayer assumes that the input VCF file has already gone through standard variant-level filtering provided by the variant caller, and by default will only analyze singleton SNVs (with an allele count of 1) which have passed prior variant-level filters (i.e., with a PASS value in the FILTER field of the VCF file), though each of these options can be adjusted by the user. Other non-SNV variant types (e.g., indels) are ignored. Users may also perform any desired preprocessing of the VCF file using other programs (e.g., bcftools [34]) and pipe the output directly to Doomsayer.

Each SNV in the VCF file that meets the specified criteria is annotated with two properties: 1) the substitution type, defined by the major and minor allele, and 2) the trinucleotide sequence context, defined by querying a fasta-formatted reference genome file and identifying the bases immediately upstream and downstream from the variant site. The substitution type and trinucleotide sequence context jointly define the 3-mer subtype for that SNV. As each qualifying SNV in the VCF file is processed, the entries of the Nx96 SNV spectra matrix, M, are incremented accordingly. Doomsayer relies on the cyvcf2 [123] and pyfaidx [141] Python libraries for parsing VCF and fasta files, respectively. Once all qualifying SNVs in the input VCF(s) have been counted, each row of the M matrix (containing the frequencies of 96

SNV subtypes for each individual) is scaled by its sum, producing a new matrix, $M'_{N \times 96}$, where the $M'_{i,j}$ entry indicates the fraction of singletons of subtype j in genome i.

Decomposing the singleton SNV spectra matrix

By default, Doomsayer will then perform principal component analysis (PCA) on the M' matrix, using functions from the scikit-learn Python library [125]. Users can optionally perform non-negative matrix factorization (NMF) instead, in which case Doomsayer will decompose the matrix into the approximate product of two smaller matrices, $W_{N \times R}$ and $H_{R \times 96}$, where R is the number of signatures, $W_{N \times R}$ indicates the relative contributions of each signature to the spectra of each sample, and $H_{R \times 96}$ indicates the loadings of the 96 3-mer subtypes within each of the R signatures, essentially as described by Alexandrov et al. [25]. The NMF decomposition in Doomsayer is implemented using functions from the nimfa Python library [124]. The code for processing VCF files and performing matrix decomposition was modified from our somatic mutation signature analysis software, *Helmsman* [142].

Unsupervised outlier detection

To identify individuals whose singleton SNV signatures appear abnormal and indicative of potential quality issues, *Doomsayer* applies two unsupervised anomaly detection algorithms to the data contained in the $W_{N \times R}$ signature matrix. A brief summary of these algorithms is provided below; a more detailed description of each can be found at [143]. In each case, the user must specify a tolerance threshold, tau, indicating the maximum fraction of samples to flag as potential outliers. By default, tau is set to 0.05, meaning no more than 5% of samples will be flagged as outliers by a given method. Unless specified by the user, Doomsayer applies both

anomaly detection algorithms and flags as outliers any samples that fail both decision criteria; users can optionally choose a more liberal filter, in which a sample flagged by either algorithm will be returned as outliers, or consider only those samples flagged by a specific algorithm.

The first anomaly detection algorithm is known as elliptic envelope filtering. If we assume the distribution of the R signature scores (i.e., the columns of the W matrix) follow a multivariate Gaussian distribution, the elliptic envelope method estimates the covariance structure of the data and fits an R-dimensional ellipsoid that covers (1-tau)*100% of individuals. Samples falling outside the boundary of this ellipsoid are flagged as outliers.

The second algorithm implemented is an isolation forest. This algorithm uses random forests to recursively partition the rows of the W matrix, forming a tree structure. Samples with very similar singleton SNV spectra will require many more partitions to isolate into individual leaf nodes than samples with more unique singleton SNV spectra, which can be isolated using relatively few partitions. The isolation forest method therefore flags as outliers the tau*100% of samples that are isolated with the shortest paths on this tree.

After determining which samples qualify as outliers, *Doomsayer* generates two text files, *doomsayer_keep.txt* and *doomsayer_drop.txt*, each containing a list of sample IDs that passed and failed the anomaly detection procedure, respectively. These files can be used in downstream analysis programs, such as bcftools [34] and PLINK [90].

Visualization and diagnostic reports

The optional diagnostic report generated by Doomsayer provides three interactive plots for users to better understand how the singleton SNV spectra of the flagged outliers differ from non-outliers. First, *Doomsayer* generates a heatmap showing the relative enrichment or depletion

of each of the 96 3-mer subtypes in each of the outlier samples (as in **Fig. 5.2a**). We apply hierarchical clustering to the rows and columns of this heatmap (using Ward's method [144]) to group together samples (rows) with similar singleton SNV spectra, and subtypes (columns) that tend to be enriched or depleted together. A dendrogram is displayed next to the rows of this heatmap to illustrate the structure of these sample clusters. Outliers are assigned into one of R major sample clusters (where R is the pre-specified number of SNV signatures analyzed), determined by identifying the R innermost nodes below the root node of the dendrogram.

Second, Doomsayer generates pairwise scatterplots of the first R principal components or NMF signature scores, depending on which dimensionality reduction algorithm was specified (as in **Fig. 5.2b**). Each sample is represented by a point in each subplot—outlier samples are colored according to the major cluster classifications determined by the hierarchical clustering procedure described above, and non-outlier samples are colored grey. This figure shows where the outliers fall in the R-dimensional space, and which of the R signatures appear to separate certain groups of outliers from the non-outlier samples.

Third, the diagnostic report will include a figure showing side-by-side barplots of the observed singleton SNV spectra for the non-outlier samples and each of the outlier clusters. For each outlier cluster and each of the 96 3-mer subtypes, Doomsayer will perform a t-test to determine if the mean proportion of SNVs of that particular subtype differs from that of the non-outliers. Subtypes that differ significantly after multiple testing ($P<0.05/[R*96]$) are highlighted in this figure to show which subtypes are most strongly under- or over-represented within each outlier cluster. Optionally, Doomsayer can use these results to generate a more stringent list of outliers, where only samples within outlier clusters exhibiting statistically significant differences for 4 or more of the 96 subtypes are considered to be low-quality.

Additional features

Doomsayer includes several additional features and options that were not applicable to the analyses presented in this manuscript. These features are described in detail in the online documentation, available at https://www.jedidiahcarlson.com/docs/doomsayer/.

Availability of data and materials

Source code for Doomsayer, along with detailed documentation, is freely available at https://github.com/carjed/doomsayer under the MIT license. Doomsayer is also available as a pre-built Docker image from https://hub.docker.com/r/carjed/doomsayer/, and an interactive cloud-based environment and tutorial can be accessed at https://mybinder.org/v2/gh/carjed/doomsayer/master.

Data availability for the BRIDGES dataset is described in [97]. Data for the 1000 Genomes Phase 3 sample were downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/. Data for the TOPMed Framingham Heart Study sample are available on the NCBI dbGaP database under accession number phs000974.v3.p2.

Chapter VI.

Discussion

The rate at which new mutations occur is of central importance to studies of genetic variation, heritable disease, and genome evolution. In this dissertation, we have extensively investigated fine-scale patterns of mutation rate heterogeneity in humans, providing a nuanced portrait of how single-nucleotide and multinucleotide mutation processes shape the patterns of variation throughout the genome. We also have developed computational methods to harness information about mutation rate heterogeneity to improve various aspects of genomics research. In this section, we review the key findings and insights of each study and discuss their broader significance and implications for future research.

In Chapter 2, we introduced a novel approach for investigating fine-scale germline mutation patterns, by studying the properties of extremely rare variants (ERVs) ascertained from large whole-genome sequencing datasets. This study relied on the intuition that ERVs are an unbiased representation of recent mutation events that have occurred in the population. This intuition, along with the fact that ERVs are abundant throughout the genome, enabled us to study patterns of mutation rate variation at an exceptionally fine resolution. Specifically, we used these ERVs to extensively quantify how mutation rates vary with respect to local sequence context and various other features of the genomic landscape. These patterns of variation led to several important findings concerning the biological mechanisms of mutation. In particular, we

discovered that the mutability of many sequence motifs appears to be modulated by the presence of particular genomic features. This result substantially enhances our understanding of how these features impact mutation rates--up to this point, studies had only attempted to describe whether a given feature tends to results in a uniform increase or decrease in mutability in a given region, with no regard for motif-specific effects [13]. In addition to improving our basic understanding of how different mutation processes are affected by genomic context, the map of estimated mutation rates that we generated in this study has immediate relevance for a range of applications, such as interpreting variants associated with heritable diseases, inferring signals of natural selection, calibrating variant detection algorithms, and improving simulations of genetic variation. We note that the estimates we have generated in this study have already been incorporated in a number of such applications [145–147].

Because the data for the genomic features considered in these analyses were largely assayed from somatic cell lines, our estimates are admittedly crude, and do not necessarily reflect the precise mutational landscape of male and female germ cells. Future investigation of the mutagenic effects of genomic features will certainly benefit from assaying these features directly in sperm and egg cells. Further, our use of ERVs precluded our ability to differentiate mutation patterns unique to the male or female germline or how these mutation patterns might be influenced by parental age—these effects can only be evaluated through de novo mutations where the parental origin is known [14, 23, 40–42]. We also note that our choice to focus on mutation rates defined by a 7-mer sequence context was driven by practical concerns, not necessarily biological insights (though a paper published during the development of our study did demonstrate that 7-mers capture important patterns of variation [12]). It is entirely possible

that an even broader sequence context is necessary to capture the full range of sequence-dependent variation in germline mutation rates.

In Chapter 3, we investigated the phenomenon of multinucleotide mutations in the human germline, again relying on singleton SNVs as a proxy for recent mutation events. Here, we introduced a novel statistical model to describe the inter-singleton distance distribution as a mixture of four exponential processes. We showed that two of these inferred processes are characteristic of two particular multinucleotide mutation mechanisms known to be active in the human germline, namely error-prone translesion synthesis (TLS) and repair of double-strand breaks (DSB).

We next analyzed how each of the four inferred processes contributing to the inter-singleton distance distribution varied regionally throughout the genome. This analysis confirmed the presence of all previously-reported hotspots for DSB-associated MNMs [22, 23] and identified several novel hotspots associated with this process, often occurring in the subtelomeric regions of many other chromosomes. We further explored how these genome-wide clustering patterns of germline mutations associate with other features of the genomic landscape. Similar to our findings in Chapter 2, the results of this analysis showed that various genomic features are likely to subtly influence the efficacy of multinucleotide mutation mechanisms (or subsequent repair processes). These results also evince potential similarities between somatic and germline mutation processes, such as our finding that CpG islands tend to be enriched for TLS-associated singleton clusters, reminiscent of a recent study that found promoter regions in various cancers are targeted by translesion polymerase eta [62]. More generally, the effects of genomic features will be an important consideration in future studies of mutation rate heterogeneity.

We concluded Chapter 3 by investigating how sample demography and variation in the single-nucleotide mutation rate also contribute to singleton clustering patterns throughout the genome. Through simulations, we estimate that approximately a third of clustered singletons (<20,000bp apart) are purely attributable to sample demography, and much of the remaining clustering may be due to regional heterogeneity in single-nucleotide mutation rates. We acknowledge that clustering patterns can also result from a variety of other processes, such as selection, biased gene conversion, and evolving mutation processes. This presents many opportunities for future research to dissect precisely how much each of these factors contributes to observed patterns of genetic variation at any given genomic region, ultimately improving the accuracy and precision of inference in population and evolutionary genetics applications.

One particularly important aspect of this study was our comparison of how multinucleotide mutation patterns differ between individuals with different ancestral backgrounds. The vast majority of studies of genetic variation published in the last 20 years have focused almost exclusively on individuals of European ancestry, which both perpetuates health disparities for underrepresented populations and causes researchers to miss important biological insights [148, 149]. Our understanding of human germline mutation patterns is no exception to this problematic trend: nearly all of the whole-genome trio sequencing studies published thus far have been restricted to individuals of European ancestry, with three of the largest samples coming from Iceland [23], Great Britain [41], and the Netherlands [14]. This lack of diversely-sampled populations calls into question the extent to which the results of these studies can be generalized to individuals with different ancestral backgrounds. The largest currently published whole-genome trio sequencing study to contain samples of diverse ancestry comes from the Inova Translational Medicine Institute [150]. This dataset has been used extensively to

characterize germline mutation patterns [22, 42, 151], but these studies have notably not attempted to describe how mutation patterns vary according to the ancestry of each individual. By deeper consideration for how mutation patterns vary between populations, future studies stand to gain substantial scientific insights into the causes and consequences of mutation processes while also ensuring that any subsequent applications for improving human health are equitable for minority populations and help reduce health disparities.

Another logical extension of the work presented in Chapters 2 and 3 is to investigate the regional variation and clustering patterns of other classes of mutations, such as short insertions or deletions (collectively referred to as indels) or larger structural variants (SVs). These classes of mutations, however, are often difficult to genotype and map with certainty using short-read sequencing technologies [152]. Emerging sequencing technologies, such as nanopore sequencing and other long-read platforms [153], have shown promising capability for detecting indels and SVs. We anticipate that widespread adoption of these technologies will enable exciting new discoveries about the processes underlying these mutations and their implications for understanding human health and evolution.

Though our analyses have primarily focused on describing germline mutation patterns averaged across many individuals, these mutation patterns do in fact differ from person to person. One of the most salient extensions of the work presented in Chapters 2 and 3 will be a deeper investigation of how and why germline mutation rates and patterns vary between individuals. Like other complex traits, the mutation patterns unique to each genome are the result of multiple genetic and environmental factors. By studying how these "mutational phenotypes" vary throughout the population, it may be possible to identify specific genetic variants that influence endogenous mutation processes and isolate the mutation signatures that result from

distinct environmental exposures. Such findings will be essential for unmasking how mutation processes have evolved over time and elucidate a deeper understanding of the etiologies of various diseases.

The first task of this endeavor is to define and quantify the mutational phenotype(s) of interest. Prior studies of inter-individual variation in mutation patterns have focused on the genome-wide average mutation rate or basic mutation spectrum per individual as the primary mutational phenotype (e.g., [40, 116]). In this dissertation, we have characterized many additional mutation patterns that can be phenotyped on an individual basis, such as the granular (e.g., 7-mer) mutation spectrum, the magnitude of mutagenic effects for various genomic features, clustering patterns, intrinsic mutation rates of MNMs, etc. These mutational phenotypes represent only a few of the many ways in which mutation data can be collected and quantified. For example, the de novo mutations observed in an individual arose over multiple cellular generations in the paternal and maternal germlines, so the mutational phenotype may therefore be defined more stringently by inferring the stage of germline development and the parent-of-origin in which the phenotype of interest emerged.

The concept of mutational phenotypes goes beyond simply cataloging the extent of variation and has obvious implications for studying how mutation mechanisms have evolved over time. If a germline mutation alters the efficacy of endogenous DNA damage or repair mechanisms (for example, affecting the catalytic domain of a mismatch repair gene), all individuals who inherit that allele will be subject to the resulting mutational outcome, be it an increased mutation rate, altered mutation spectrum, or emergence of new mutation hotspots. Over time, through the processes of genetic drift and natural selection, mutator alleles may rise to appreciable frequencies in the population, such that distantly related individuals share the

same distinct mutational phenotype. Recent work has demonstrated that past shifts in the mutation spectrum are indeed detectable, suggesting the presence of mutator alleles [48, 80, 81], but these studies have yet to isolate which genetic polymorphism(s) (if any) are associated with these shifted mutation spectra. Identifying the mutator alleles responsible for these ongoing evolutionary changes will be an important challenge for the field going forward.

As mentioned above, mutational phenotypes are jointly affected by both genetic and environmental factors. Unlike other complex traits, however, heritable mutational phenotypes can carry signatures of non-heritable environmental factors. Consider a scenario where an environmental mutagen causes an increase in C>T mutations in the germline. The offspring of individuals exposed to this mutagen will of course carry more C>T mutations than expected, so the offspring's offspring will also carry more C>T heterozygous sites, even if they were never exposed to the same mutagen that affected their grandparents' germ cells. This example demonstrates that the mutational phenotype can serve as a window into the past to identify not only mutator alleles affecting endogenous mutation processes, but also exogenous mutagenic processes active in the environment at specific times in history.

Our singleton-based approach to studying mutation patterns may prove particularly useful in future attempts to identify these historical mutation patterns. Because de novo mutations are so rare, it is difficult to collect a sample large enough to systematically test for mutator alleles that are active in the present generation. Singletons, however, represent mutations that have accumulated over several generations, so the effects of any mutator alleles will have compounded in this time, meaning that there is potentially a strong enough signal to establish an association (with the caveat that meiotic recombination is also acting during the time in which

singletons accumulate, so the mutational signal resulting from a given mutator allele will only be present on the same haplotype as the mutator allele itself).

Inter-individual variation in mutation patterns can also be an important indicator of disease risk. For example, germline mutation rates are known to increase as parents age [5], a phenomenon which has been implicated in multiple complex and highly heritable disorders, such as autism and schizophrenia [40]. Characterizing the host of genetic and environmental factors that influence the inter-individual variation in mutation patterns stands as an important task in the quest to understand the etiologies of complex diseases, with the potential for rapid adoption in the fields of genetic counseling and precision medicine.

A central thesis of this dissertation is that an understanding of germline mutation rate patterns is fundamental to nearly every aspect of genome analysis. In Chapters 4 and 5, we shifted our focus towards developing computational methods designed to incorporate a detailed understanding of mutation rate heterogeneity to improve common bioinformatics tasks in genomics research. One such method, known as mutation signature analysis, is arguably one of the most important computational methods to have emerged from the field of cancer genomics, and over a dozen software tools incorporating this method have been published in the last 5 years [27–31, 154–160]. When we attempted to apply these tools to very large datasets, however, it quickly became apparent that none of the existing software were suitable for analyzing datasets containing thousands of samples and millions of variants. This computational bottleneck led to the development of *Helmsman*, an extremely fast and memory-efficient program for performing mutation signature analysis, described in Chapter 4. Through the use of efficient programming practices and carefully optimized data processing, *Helmsman* achieves a level of performance that is orders of magnitude faster and more memory-efficient than other mutation signature

analysis programs. We anticipate that *Helmsman* will prove useful in cancer genomics and precision oncology, where increasingly massive datasets are leading to demand for fast, automated software.

Though *Helmsman* is extensively documented and intentionally designed to be accessible to users with a range of bioinformatics expertise and/or computational resources, some users may not be comfortable with using command-line programs or may not have access to the computational environment necessary to take full advantage of *Helmsman's* parallel processing capability. Many bioinformatics applications have moved towards a model of providing users with a graphical web interface that is linked to cloud-based servers to perform the computationally-intensive processing. Although *Helmsman* was not released specifically as a cloud-based application, we have already incorporated aspects of this strategy, making it available at a Docker container that can be deployed with minimal software/hardware dependencies on virtually any server, either locally or on the cloud. Once deployed, Helmsman can be run through an interactive Jupyter notebook directly in a web browser, without end users needing to touch a terminal. In the future, we hope to assess the evolving needs of users and introduce additional features and optimizations accordingly.

In Chapter 5, we presented another practical application of mutation rate heterogeneity. Here, we developed a novel method and software for performing quality control in whole-genome sequencing datasets, based on the analysis of the singleton SNV spectra across samples. The development of this method was guided by our intuition that, if underlying germline mutation processes are similar throughout the human population, the spectra of genuine biological variation should be similar between samples. This idea is conceptually similar to using the Ti:Tv ratio in each sample as an indicator of data quality—just as we expect the transition

and transversion SNVs in each sample to be distributed in a particular ratio, we extend this expectation to consider the distribution of SNVs across 96 3-mer subtypes.

Applying this method to the BRIDGES and 1000 Genomes Phase 3 datasets, we identified several noteworthy quality issues that went undetected by earlier QC procedures. Chief among these was our discovery that several samples in both datasets carried a signature of oxidative DNA damage, characterized by an excess of C>A transversions. These findings add to a growing body of evidence that errors resulting from oxidative damage are pervasive in sequencing datasets where DNA was amplified using PCR [137, 161]. In the BRIDGES dataset, we also found evidence of batch effects, wherein multiple samples prepared or sequenced at the same time carried similar error signatures. Although the careful experimental design of the BRIDGES study likely mitigated any spurious associations caused by these batch effects (which could occur if cases and controls were not balanced across each batch), this finding is a strong reminder that batch effects exclusively impacting cases or controls have the potential to cause serious (and in some cases, irreparable, as described in [162]) confounding of downstream analyses.

Our method also proved effective for identifying quality issues present in PCR-free, high-coverage sequencing data from the Framingham Heart Study. Though the extent of the quality issues in the FHS dataset was far less dramatic than what we observed in the BRIDGES and 1000 Genomes datasets (both of which used low-coverage sequencing of PCR-amplified DNA), these results demonstrate that our method's usefulness extends beyond retrospective analysis of older sequencing data, and is sensitive enough to detect cryptic error biases present among modern sequencing datasets with inherently lower error rates. Thus, we anticipate our method

will continue to be useful for quality control in ongoing and future studies that use the latest sequencing protocols.

We acknowledge the inherent limitations of using unsupervised anomaly detection algorithms in our method—because we make no prior assumptions about which samples are high-quality and which are affected by error biases, our outlier detection strategy may be ineffective in datasets where error signatures are diffused throughout the samples. In the future, it may be desirable to incorporate some form of supervised learning algorithm where we explicitly query the singleton SNV spectra of each sample for known error signatures.

In sum, this dissertation contributes new and nuanced insight into the underlying patterns of variation in human mutation processes. We show how different mutation mechanisms contribute to observed patterns of variation, highlighting a diverse range of potential applications which stand to benefit from this new-found knowledge. We anticipate that these studies will continue to enhance our ability to use whole-genome sequencing as a window into human history and a tool for improving human health.

Appendix A: Supplementary Material for Chapter II

Identification of outlier samples

For the 3,716 individuals that passed our initial sample-level filters, we summarized the per-sample distribution of extremely rare variants (ERVs) across 3-mer subtypes and used this information to flag individuals that showed abnormal patterns of variation indicative of systematic sequencing errors or batch effects. In brief, we adapted the non-negative matrix factorization (NMF) technique described by Alexandrov et al. [25] to deconvolute the 3-mer mutation spectra as a composite of 3 distinct "signatures." Assuming the population has been susceptible to the same mutation processes over the timespan in which ERVs have accumulated, we expect that the relative contribution of the 3 NMF signatures is stable across individuals. Applying this strategy, we identified 156 individuals where one or more signatures had a contribution >2 standard deviations away from the mean contribution of that signature (calculated across all individuals).

These outliers exhibited one of two distinct signatures indicative of error biases. The first signature, characterized by an unusually high proportion of C>A and G>T singletons, was overrepresented in 112 of these samples, consistent with patterns of oxidative damage that are known to occur during DNA shearing, likely due to the presence of reactive contaminants [137]. The second signature, characterized by depleted rates of C>N and G>N ERVs, was overrepresented in the remaining 44 samples. Further investigation of the samples carrying this

signature showed many had higher GC bias scores (i.e., systematically lower depth of coverage in GC-rich regions), likely resulting in lower calling rates for C>N and G>N types. Moreover, 24 of the 44 samples were sequenced in the same batch, and the remaining 20 samples were distributed across only 8 of the 48 other batches, indicating that these coverage biases and resulting error signatures clustered by batch. To limit the confounding effects of nonbiological variation present in the data, we excluded the 156 samples displaying either of these error signatures. Note that doubletons in the pre-filtered sample that would have become singletons in the post-filtered sample were not included in our analysis. Many of these variants are likely true doubletons in the BRIDGES sample and hence present in the population at a higher frequency (i.e., having arose further in the past) than the average singleton, so retaining these ambiguous variants might inadvertently affect the distribution of variants.

Estimation of false discovery rate by Ts/Tv statistics

We estimate the false discovery rate among BRIDGES ERVs using the following method.

(1) Let $TS_o = TS_t + TS_f$ be the number of observed transitions (23,733,766), consisting of both true positives (TS_t), and false positives (TS_f)

(2) Let $TV_o = TV_t + TV_f$ be the number of observed transversions (11,840,651).

(3) Based on findings from other large-scale sequencing studies, the true positive Ts/Tv ratio, $TSTV_T = \frac{TS_t}{TV_t}$ is expected to be between 2.0 and 2.1 [54].

(4) Because there are 8 possible transversions and 4 possible transitions, if errors have occurred at random, the Ts/Tv ratio for random false positive errors ($TSTV_\epsilon$) should be 0.5, that is, $\frac{TS_f}{TV_f} = 0.5$, assuming no systematic sequencing error biases.

Solving this system of four equations, it follows that $TV_f = \frac{TSTV_T \times TV_o - TS_o}{TSTV_T - 0.5}$ and $TS_f = 0.5 \times TV_f$, so the false discovery rate, $\frac{TS_f + TV_f}{TS_o + TV_o}$, can be estimated as:

$$\frac{TS_f + TV_f}{TS_o + TV_o} = \frac{0.5\left(\frac{TSTV_T \times TV_o - TS_o}{TSTV_T - 0.5}\right) + \frac{TSTV_T \times TV_o - TS_o}{TSTV_T - 0.5}}{TS_o + TV_o}$$

Assuming the true Ts/Tv ratio ($TSTV_T$) is between 2.0 and 2.1, by this calculation we estimate a false discovery rate of 0.1-2.9% among the BRIDGES ERVs.

Potential sources of bias among ERVs

Motif-specific error rates

Certain sequence motifs may be more susceptible to sequencing error, which could lead to a non-random distribution of false positive singleton calls and subsequently bias our analyses [131, 163]. Allhoff et al. [131] reported context-specific errors for the Illumina HiSeq platform, noting that the most common of these are strand-specific T>N errors at 5'-GGGT-3' motifs (i.e., there is no evidence of an excess of A>N errors at the reverse complement 5'-ACCC-3' motifs). We reason that if the BRIDGES ERVs are enriched for such context-specific errors, we should see significantly more T>N ERVs at the 5'-GGGT-3' motif than A>N ERVs at the 5'-ACCC-3' and motif. Of the 127,831 ERVs that occur at this motif, 63,861 were 5'-[A>N]CCC-3' variants, and 63,970 were 5'-GGG[T>N]-3' variants; this difference was not significant, indicating there is no evidence for an enrichment of T>N ERVs at this error-prone motif (exact binomial test; P=0.67). Allhoff et al. remark that the variants called at error-prone positions tended to have low base quality scores as well as significant strand bias, both of which are detectable with standard filtering protocols [131]. We therefore assume that most motif-specific errors are filtered by the default strand-bias and quality filters used in our variant calling pipeline, and any undetected

errors have a negligible impact on our calculation of relative mutation rates and downstream analyses.

Mapping error

We expect the majority of ERVs in our data are mapped with high confidence, as the pre-filtering steps in our variant calling pipeline remove sites occurring on reads with average phred-scaled mapping quality score (MQ) <20 and/or where more than 10% of reads were ambiguously mapped (MQ0>10). This filtering strategy is similar to the filters employed by other large-scale sequencing projects that have demonstrated well-controlled error rates among singleton calls [56, 95]. Because mapping errors are more likely to occur in highly-repetitive regions, such as centromeric and pericentromeric loci [164], including these regions in our analyses might bias our estimates of motif-specific mutation rates and/or the impact of genomic features. However, excluding these regions entirely might have detrimental side effects: dropping ERVs in these regions will reduce the precision of our estimates, and removing hard-to-map regions might preclude our ability to assess mutation patterns unique to these regions, as they may have many levels of heterogeneous overlap with genomic features.

To determine if excluding repeat-rich regions systematically influenced our inferred rates, we compared the 7-mer relative mutation rates estimated from the full, unfiltered set of ERVs with 7-mer rates estimated if we only count ERVs and reference motifs within the 1000 Genomes strict accessibility mask, which delineates the most uniquely mappable regions of the genome (covering ~72% of non-N bases). These two sets of estimates were very well-correlated: within-type correlations were >0.96, indicating the estimated rates were highly consistent regardless of whether hard-to-map regions were removed (**Fig. A.6a**). Moreover, subtypes with larger differences between the two estimates tended to have fewer ERVs (**Fig. A.6b**), suggesting

that most observed discrepancies might simply be an artifact of reduced precision among rare mutation classes.

When we applied the masked rates to predict the set of *de novo* mutations, we found these estimates had worse predictive performance than the unmasked estimates (**Table 2.1**). This result leads us to conclude that aggressively filtering for the highest-confidence call set comes at a cost of substantially reducing the precision of the relative mutation rate estimates, and potentially causing greater bias by ignoring the information captured by ERVs in the masked regions. Although we cannot entirely exclude the possibility of mapping error biases among the unmasked estimates, the benefits of having more numerous singletons across more contiguous genomic regions in the unmasked data outweigh the concerns about errors caused by poor mapping quality.

Mispolarization of ERVs

While most singletons in the BRIDGES sample are the true derived allele, population genetic theory suggests that <1/N=0.014% of singletons in a sample are the ancestral allele, and hence subject to the same evolutionary biases we wish to avoid. These mispolarized singletons may be hard to detect, as we expect ~0.25% of all singletons to carry the same allele in human and chimpanzee due to parallel mutations that have occurred since splitting from a common ancestor. Intuitively, these parallel mutations are especially likely to occur in hypermutable loci, so removing the 0.25% "ancestral" alleles created by parallel mutation may create a bigger bias than including the 0.015% truly ancestral alleles.

To understand the impact of removing all putatively ancestral alleles, we used an ancestral genome inferred by 6-way primate alignment [89] to annotate each allele with the putative ancestral state. We identified 363,705 singletons (~1% of all singletons) where the

alternative allele was the same as the ancestral allele, and recalculated 7-mer relative mutation rates after removing these putatively mispolarized singletons. We found that this polarization filter did not strongly affect estimated rates: across all types combined as well as within each type, the rates before and after removal of these sites were nearly perfectly correlated (Spearman's r>0.999). Further, we found that only 9 of the 24,576 7-mer rates differed significantly after applying this filter, and the re-estimated rates for these 9 subtypes differed from the original rates by no more than 10%. More importantly, 8 of these 9 subtypes were hypermutable CpG>TpG subtypes, consistent with our intuition that many putatively mispolarized sites are in fact parallel mutations in the human and chimpanzee lineages.

As a final analysis of the potential effects of mispolarization on our estimates, we applied these filtered rates to predict the GoNL/ITMI *de novo* mutations [14, 42] in the same logistic regression framework used to compare other estimation strategies. Goodness-of-fit statistics indicated that the filtered rates predicted *de novo* mutations better than 7-mer rates estimated without the polarization filter (ΔAIC=298). However, when comparing goodness-of-fit between type-specific models, these differences largely disappeared, with seven types showing negligible differences in AIC (ΔAIC < 7), and the unfiltered rates had lower AIC for three of these (non-CpG C>T, CpG>GpG, and CpG>ApG). Only two types had differences in AIC greater than 10: A>T types were predicted slightly better by the filtered rates (ΔAIC=16), but CpG>TpG types were predicted better by the unfiltered rates (ΔAIC=22), suggesting the accuracy of the filtered rates is particularly affected by parallel mutations at hypermutable CpG sites. Given this lack of consistent type-specific improvement when applying the polarization filter, we performed all subsequent analyses using the full set of 35.6 million ERVs.

Curation of MAC10+-derived mutation rate estimates

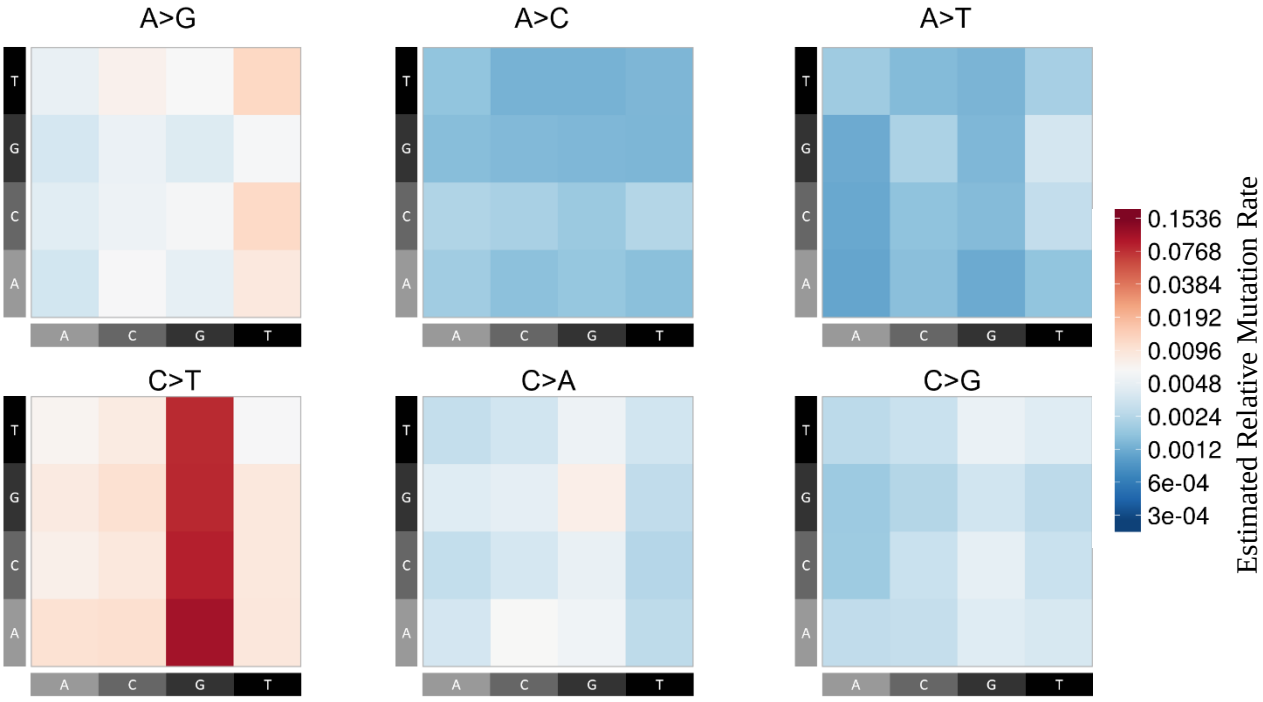
A potential concern with comparisons between our ERV-derived mutation rate estimates and Aggarwala and Voight's 1000G-based estimates [12] is that discrepancies might be partially attributable to technical differences between the two samples, not necessarily because the 1000G estimates are based on ancestrally older SNVs. For a more direct comparison, we curated a set of higher-frequency SNVs found in the BRIDGES data, removing the possibility that the dissimilar estimates are a result of differences in sequencing platform, variant calling, QC methods, and sampled individuals.

Aggarwala and Voight's mutation rate estimates are based on 7,051,667 intergenic variants observed in N=379 Europeans from the 1000 Genomes Phase I study [12]. Aggarwala and Voight do not state the exact site frequency spectrum for the European intergenic variants, but claim 26% of intergenic variants in the 1000G Phase I African sample are singletons or doubletons [12]. Thus, it is reasonable to assume that >80% of European intergenic SNVs in the 1000G data occur at a frequency greater than 1/(379*2)=0.0013 (i.e., the sample MAF of a singleton in the 1000G sample). To obtain SNVs in the BRIDGES sample in a frequency range comparable to this, we selected all SNVs with a minor allele count ≥10 (MAF≥0.0014). We identified 12,088,037 MAC10+ variants in our data, from which we estimated 7-mer relative mutation rates. We compared these estimates to 1) a set of ERV-derived 7-mer estimates calculated after randomly downsampling to an equivalent number (12,088,037 ERVs), and 2) the 1000G estimates. These comparisons show that the MAC10+ estimates are more closely correlated with the 1000G estimates (**Fig. A.3**) than with the downsampled ERV-derived estimates (**Fig. A.4**). We also used the MAC10+ estimates to predict the GoNL/ITMI *de novo*

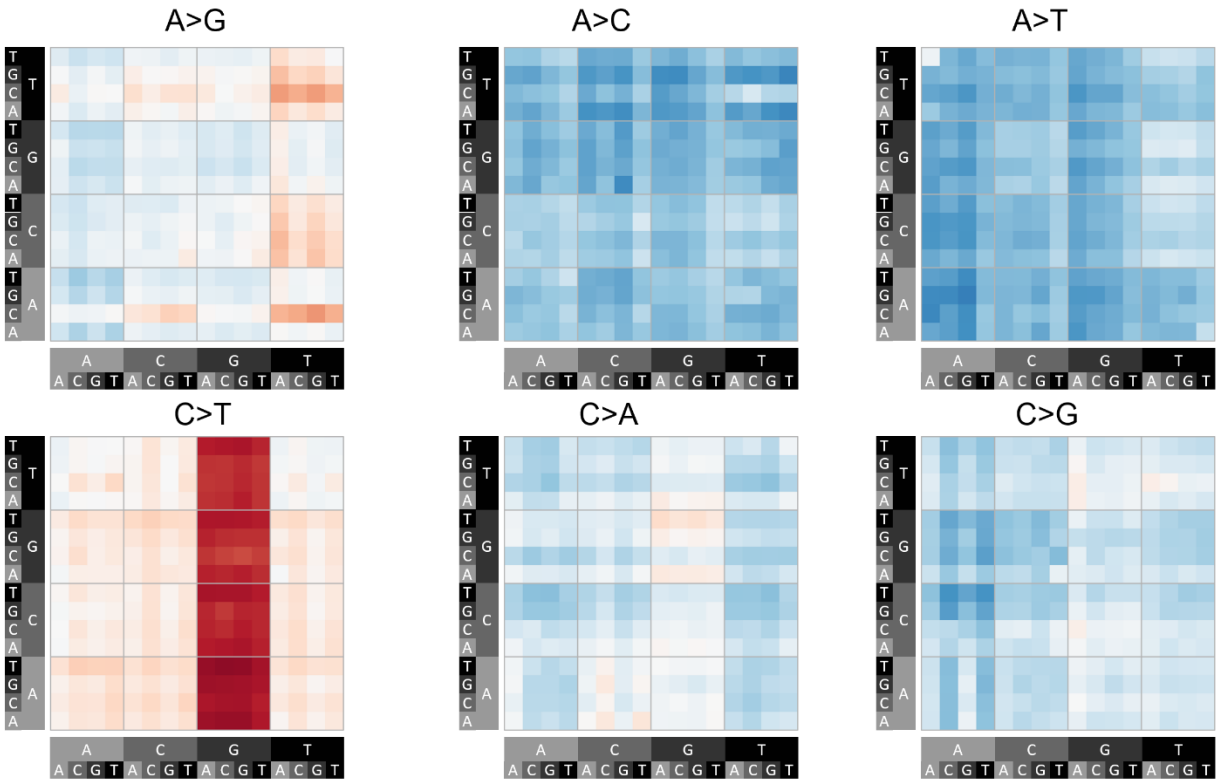
mutations and found that this model tended to perform comparably to the 1000G model (**Table A.5**).

Supplementary Figures

a.



b.



c.

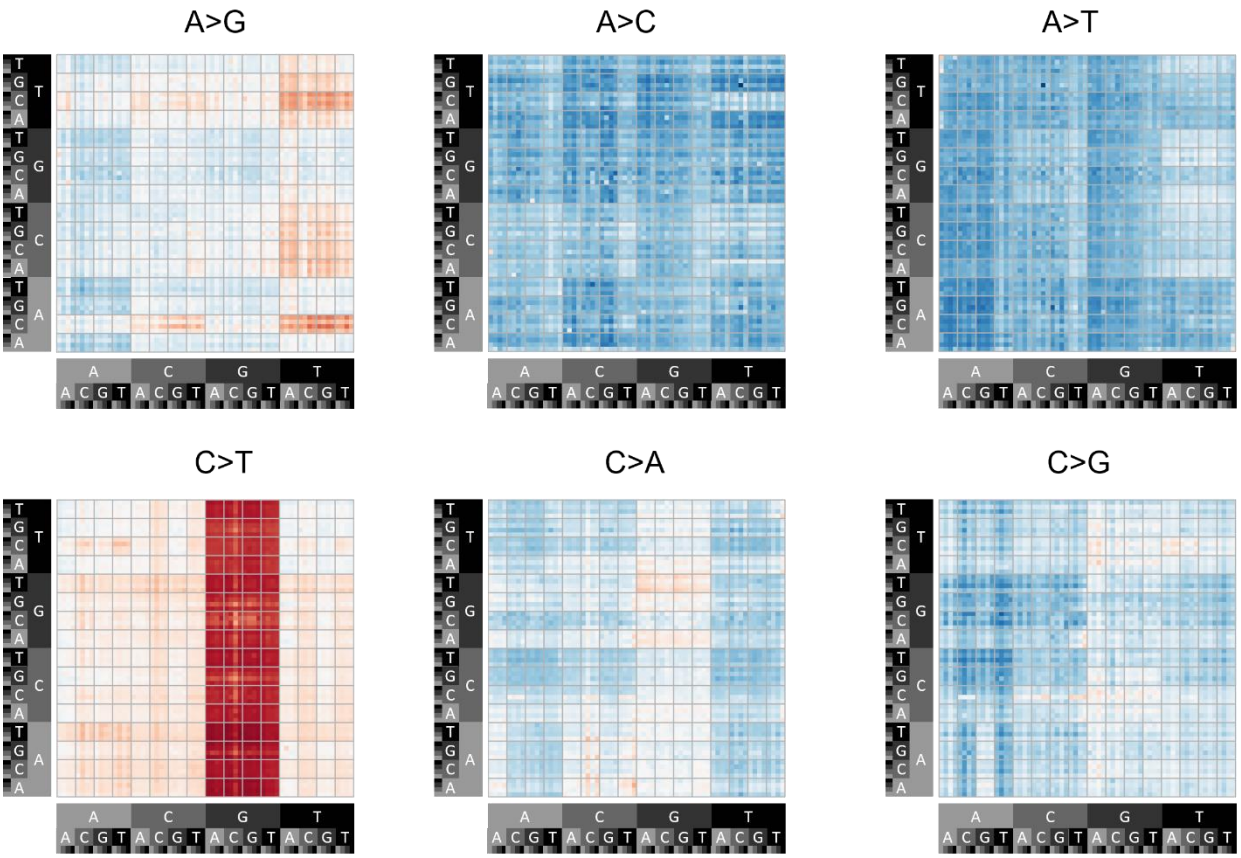


Figure A.1 High-resolution heatmaps of relative mutation rates for mutation subtypes up to a 7-mer resolution, estimated from the BRIDGES ERVs (a) estimates for 3-mer mutation subtypes. **(b)** estimates for 5-mer mutation subtypes. **(c)** estimates for 7-mer mutation subtypes. Each cell delineates a subtype defined by the upstream sequence (y-axis) and downstream sequence (x-axis) from the central (mutated) nucleotide.

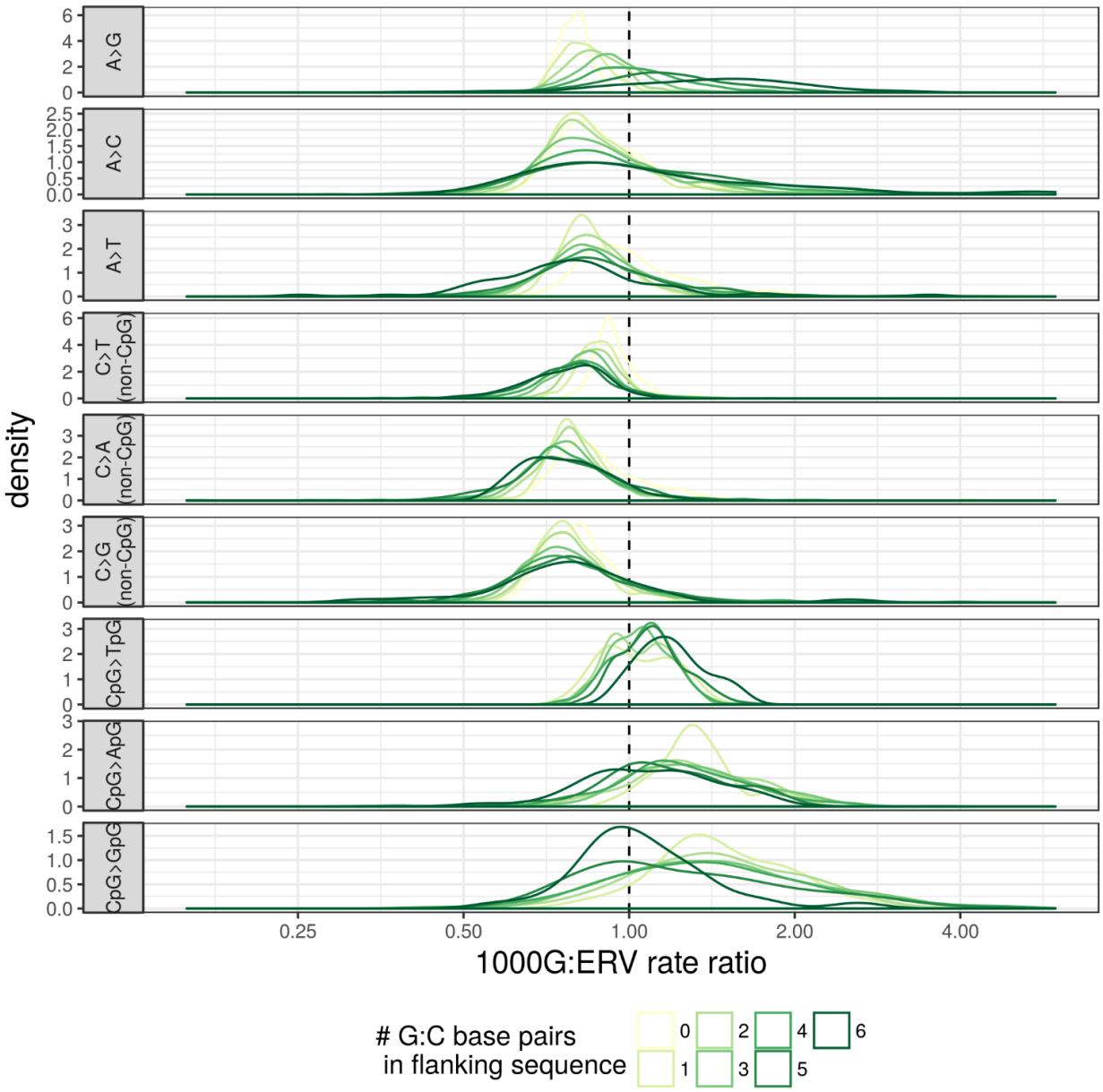


Figure A.2 Density plots comparing the distribution of ratios between the 1000G and ERV rate estimates
For each type, we grouped 7-mer subtypes by the number of G:C base pairs in the +/-3 flanking sequence, and plotted the distribution of ratios separately for each of these group. Mass to the right of the dashed line indicates estimated rates tend to be higher in the 1000G data, while mass to the left shows subtypes where estimated rates are higher in the BRIDGES ERV data.

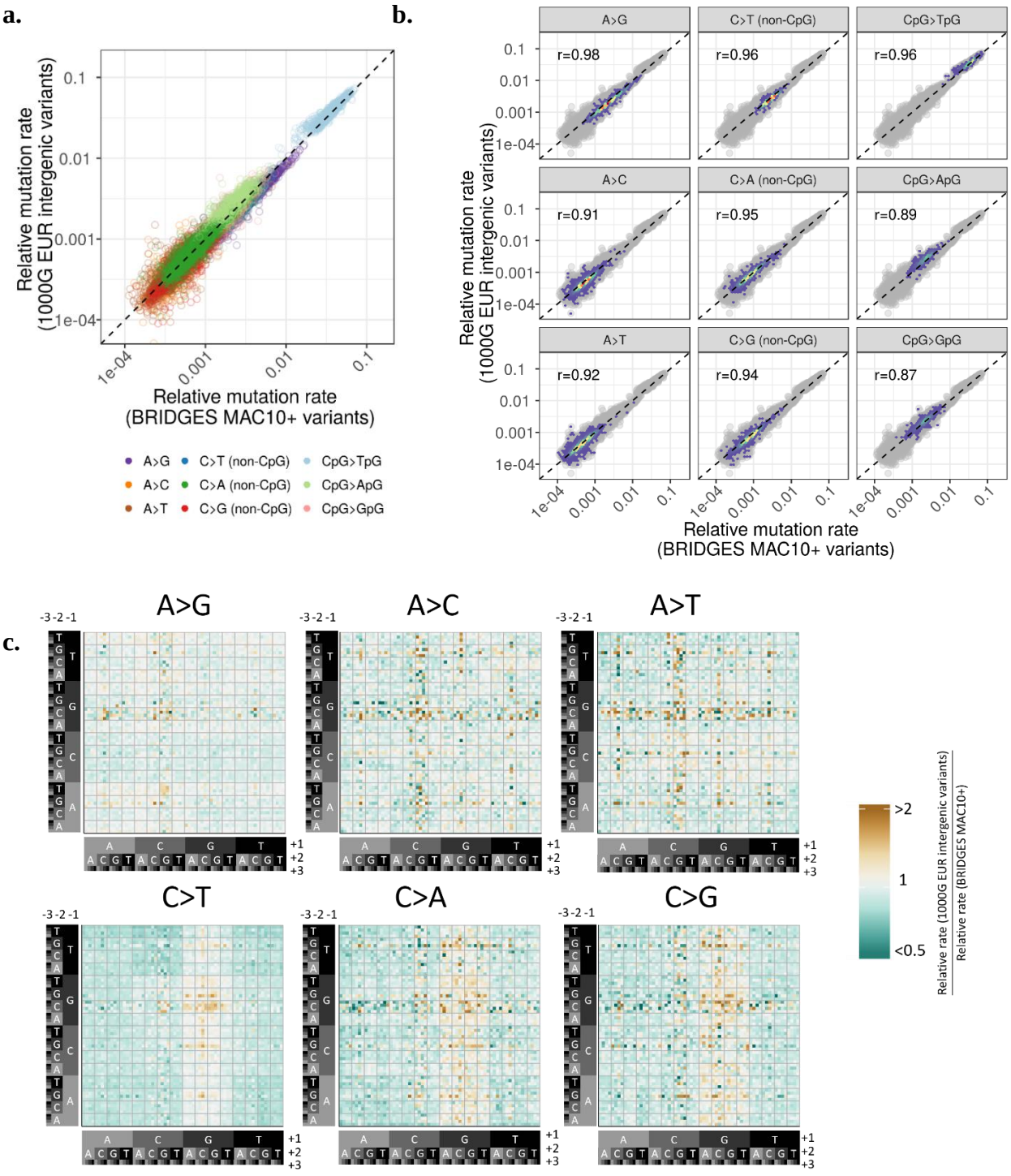


Figure A.3. Comparison of 7-mer relative mutation rates estimated from BRIDGES MAC10+ variants and 1000G Intergenic SNVs (a) Scatterplot of 7-mer subtype rates estimated from the BRIDGES MAC10+ data (x-axis), and 1000G intergenic SNV data (y-axis) **(b)** Type-specific 2D-density plots, as situated in the scatterplot of **a**. The dashed line indicates an expected least-squares regression line if there is no bias present. **(c)** Heatmap shows ratio between relative mutation rates calculated on MAC10+ variants and 1000G variants for each 7-mer mutation subtype. Subtypes with higher 1000G-derived rates relative to MAC10+-derived rates are shaded gold, and subtypes with lower 1000G-derived rates relative to MAC10+-derived rates are shaded green. 1000G-derived rates shown here are scaled relative to the MAC10+-derived rates.

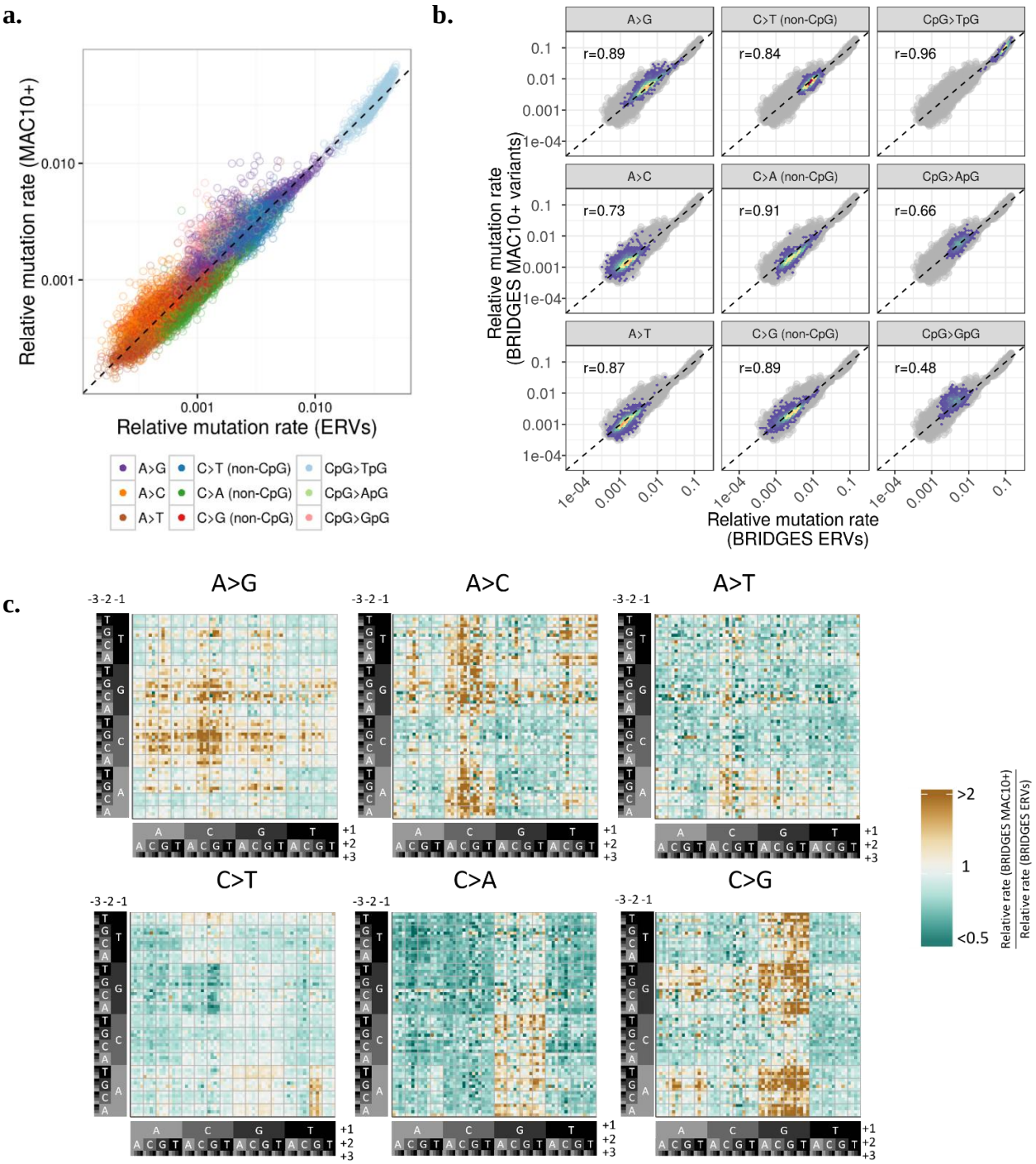


Figure A.4 Comparison of 7-mer relative mutation rates estimated from BRIDGES ERVs and BRIDGES MAC10+ variants (a) Scatterplot of 7-mer subtype rates estimated from the BRIDGES ERV data, after randomly downsampling the ERVs to 12,088,037 (x-axis) and the BRIDGES MAC10+ data (y-axis). **(b)** Type-specific 2D-density plots, as situated in the scatterplot of **a**. The dashed line indicates an expected least-squares regression line if there is no bias present. **(c)** Heatmap shows ratio between relative mutation rates calculated on MAC10+ variants and ERVs for each 7-mer mutation subtype. Subtypes with higher MAC10+-derived rates relative to ERV-derived rates are shaded gold, and subtypes with lower MAC10+-derived rates relative to ERV-derived rates are shaded green.

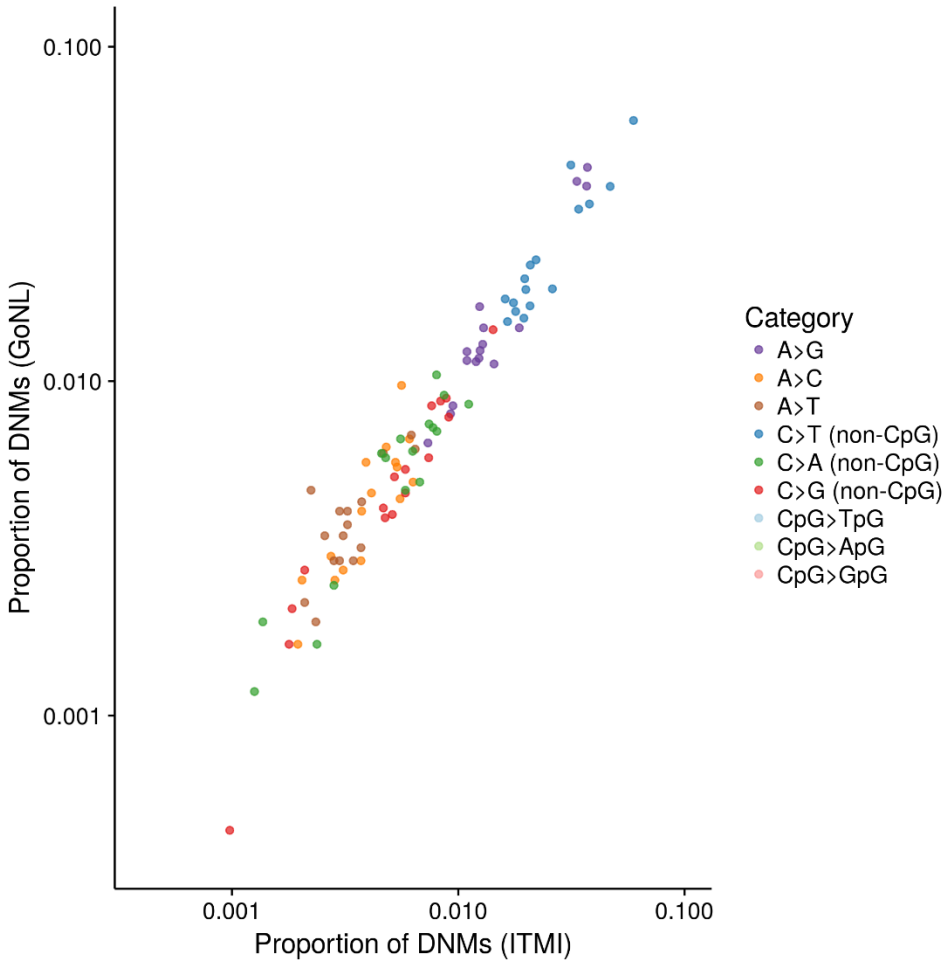
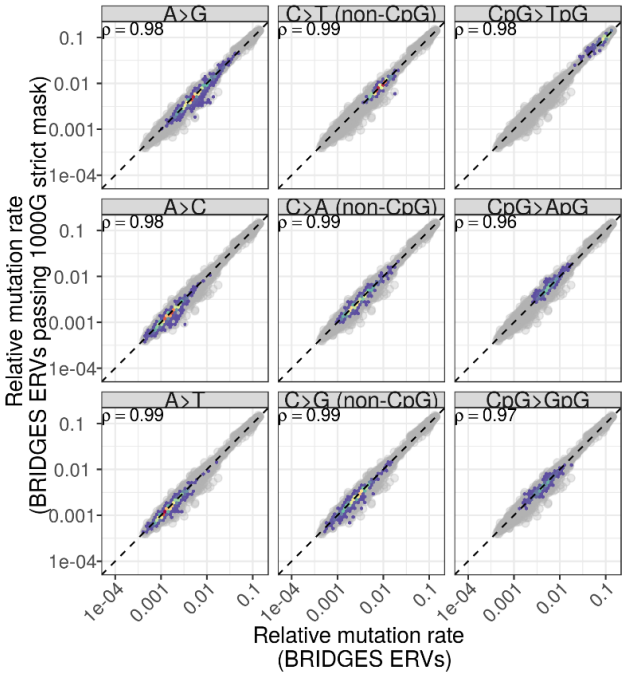


Figure A.5 Similar mutation spectra of the GoNL and ITMI data
Scatterplot shows the 3-mer mutational spectra (i.e., the proportion of all mutations falling within each of the 96 3-mer subtypes), calculated among *de novo* mutations from the ITMI (x-axis) GoNL (y-axis) trio sequencing studies.

a.



b.

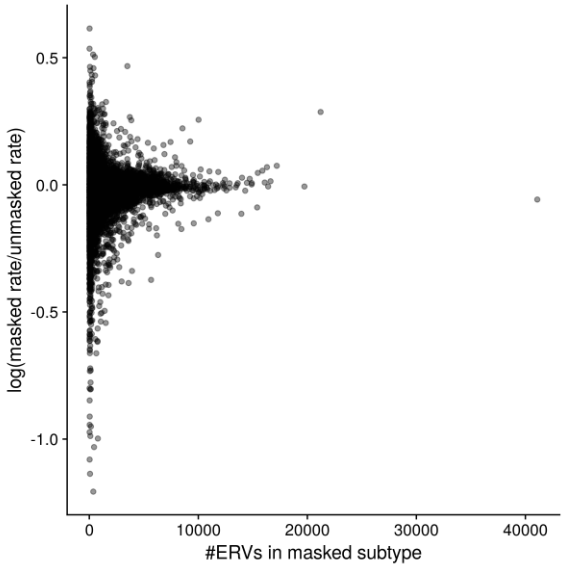


Figure A.6 Genome-wide estimates for ERV-based 7-mer subtypes are consistent with estimates from ERVs restricted to uniquely-mappable regions

(a) Relationship between masked and unmasked 7-mer relative mutation rate estimates, separated by type. **(b)** Relationship between number of ERVs per subtype (x axis) and discordance between the masked and unmasked rates, measured as the log ratio between the estimates (y axis).

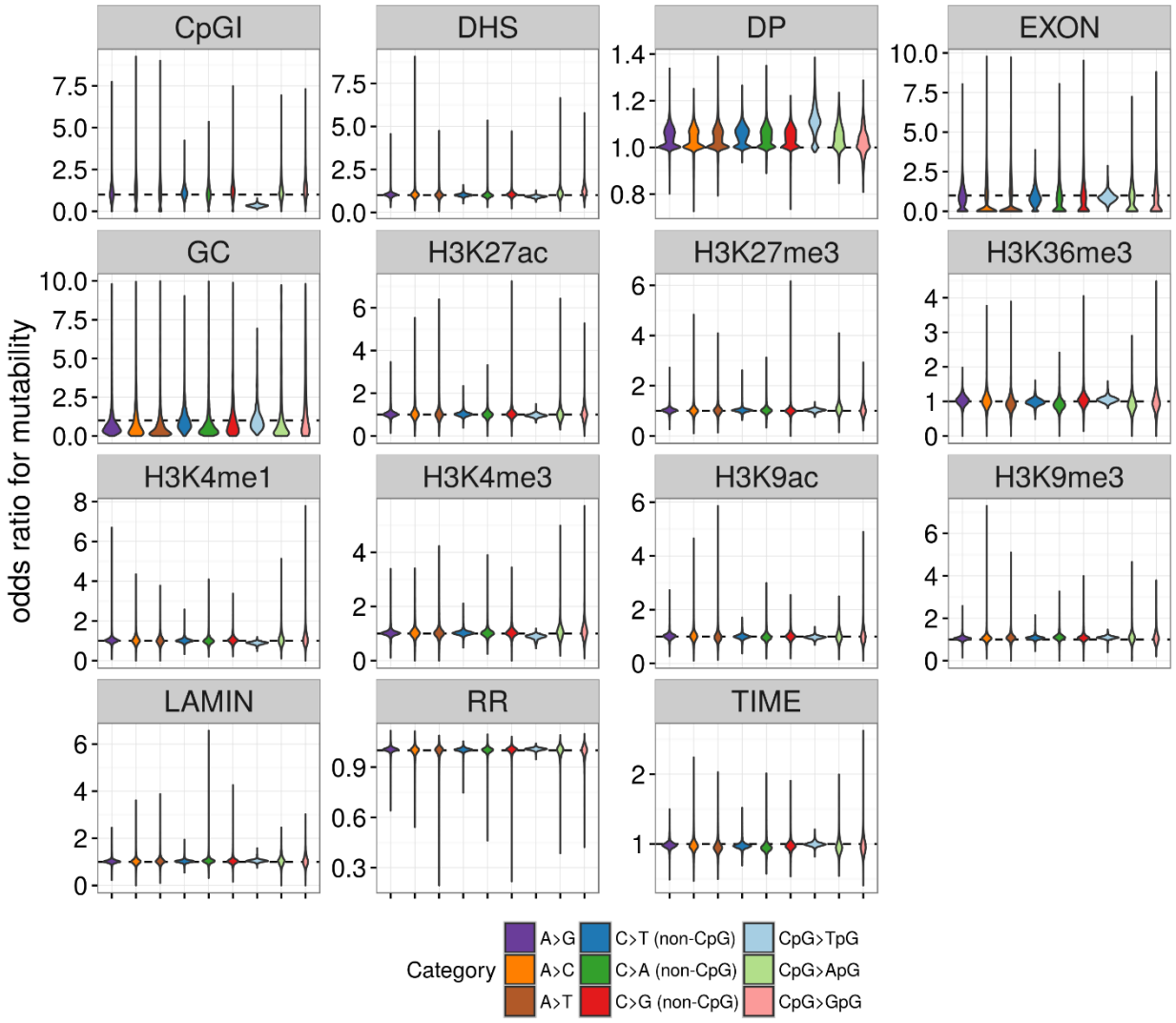


Figure A.7 Distributions of effect sizes on mutability for 14 genomic features and depth of sequencing. For each feature, we plotted the empirical distributions of the subtype-specific odds ratios for each basic mutation type, as estimated by our logistic regression models. *Replication timing is coded with negative values indicating later replicating regions, so an OR<1 means mutation rate increases in late-replicating regions.

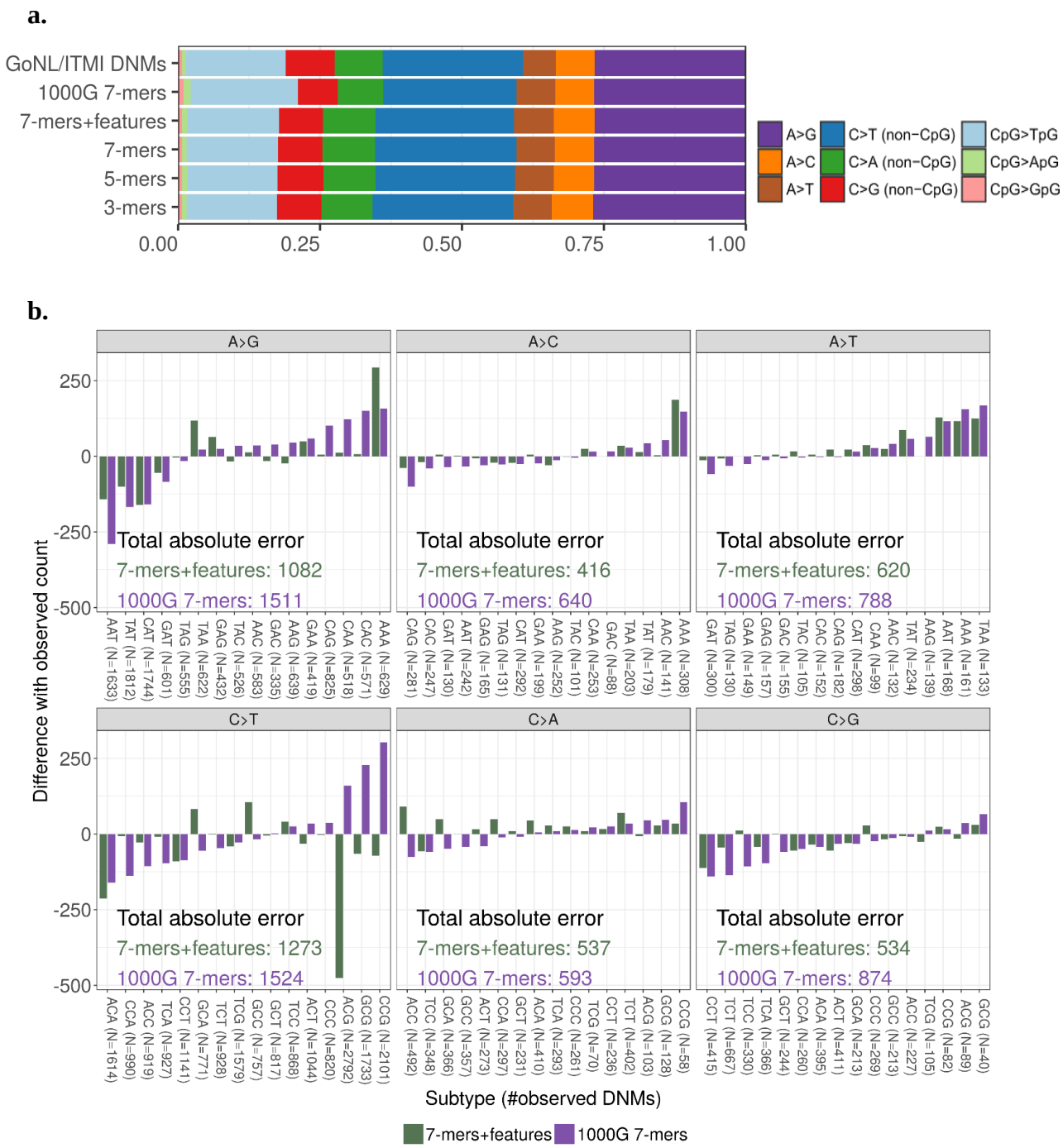


Figure A.8 Predicted mutation distributions under ERV-based models are more accurate than 1000G model (a) Distribution of the GoNL/ITMI *de novo* mutations across basic mutation types compared to the distributions predicted under the 1000G 7-mer model and each of the BRIDGES ERV-based models. **(b)** Difference between model-predicted and observed number of mutations per 3-mer subtype for the 7-mer+features model (green bars) and 1000G 7-mer model (purple bars). The number of observed mutations for each subtype is indicated along the x-axis. In each panel, subtypes are sorted in increasing order of differences under the 1000G 7-mers model.

Supplementary Tables

Table A.1 Quality comparison between filtered partitions of BRIDGES singletons

Partition	# Singletons	Ts/Tv ratio	%dbSNP (b142)	% of Full Set
Full Set	35,574,417	2.00	17.4	100
Filter 2 (MQ>56)	33,550,098	2.01	17.3	94
Filter 3 (passed 1000G strict mask)	26,810,791	1.97	17.5	75
All Filters (MQ>56, 1000G strict mask)	16,535,856	2.00	17.6	46

Table A.2. Rate estimates in GC-rich motifs are biased in 1000G data

Type	Mean 1000G/ERV ratio (≤3 C/G bases)	Mean 1000G/ERV ratio (≥4 C/G bases)	P-value
A>C	0.97	1.12	8.00e-30
A>G	1.00	1.28	2.37e-161
A>T	0.89	0.89	0.81
C>A (non-CpG)	0.76	0.72	2.61e-09
C>G (non-CpG)	0.89	0.93	2.98e-04
C>T (non-CpG)	0.93	0.85	1.75e-39
CpG>ApG	1.15	0.96	4.97e-22
CpG>GpG	1.46	1.33	2.80e-04
CpG>TpG	1.02	0.98	1.01e-09

For each mutation subtype, we calculated the ratio between 1000G-derived and ERV-derived relative mutation rates. Then, for each of the 9 basic types, we grouped 7-mer subtypes into low C/G subtypes (≤3 C/G bases in the +/-3 flanking positions) and high C/G subtypes (≥4 C/G bases in the +/-3 flanking positions) and performed t-tests for differences in the mean 1000G/ERV ratios of these two groups.

Table A.3 Comparison of observed and simulated goodness-of-fit for *de novo* prediction models under different sized non-mutated backgrounds

Model	Observed		Simulated		Background size
	AIC	R^2	AIC	R^2*	
1-mers	292542	.109	272925	.185	500,000
3-mers	284889	.139	241863	.299	
5-mers	282995	.146	239672	.307	
7-mers	282491	.148	238967	**.310**	
7-mers (BRIDGES MAC10+ SNVs)	283599	.144	240434	.304	
7-mers (1000G intergenic SNVs)	284764	.139	241724	.300	
1-mers	353896	.088	344108	.117	1,000,000
3-mers	343716	.118	317322	.197	
5-mers	341778	.124	315400	.202	
7-mers	341295	.126	314760	**.204**	
7-mers (BRIDGES MAC10+ SNVs)	342886	.121	316791	.198	
7-mers (1000G intergenic SNVs)	344003	.118	317953	.195	
1-mers	416998	.072	414016	.080	2,000,000
3-mers	404738	.102	392367	.132	
5-mers	402853	.107	390698	.136	
7-mers	402375	.108	390051	**.138**	
7-mers (BRIDGES MAC10+ SNVs)	404378	.103	392509	.132	
7-mers (1000G intergenic SNVs)	405523	.100	393741	.129	
1-mers	454267	.066	452950	.069	3,000,000
3-mers	441042	.095	434665	.109	
5-mers	439153	.099	433243	.112	
7-mers	438700	.100	432517	**.114**	
7-mers (BRIDGES MAC10+ SNVs)	441059	.095	435270	.108	
7-mers (1000G intergenic SNVs)	442181	.092	436443	.105	

*The simulated R^2 of the best possible model for each background size, indicated in bold, represents the optimal performance we can expect.

Table A.4 Comparison of model AIC specific to GoNL or ITMI de novo mutations

Model	GoNL DNMs (11,020 mutations)	ITMI DNMs (35,793 mutations)
1-mers	114945	288707
3-mers	111952	280025
5-mers	111507	278542
7-mers	111381	278201
7-mers (BRIDGES MAC10+ SNVs)	111913	279580
7-mers (1000G intergenic SNVs)	112185	280401

Models fitted to a background of 1 million non-mutated sites, as described previously. Note that the difference in AIC between the two datasets is due to the difference in number of DNMs, and is not comparable between the GoNL and ITMI studies. Goodness of fit statistics for both datasets have the same rank order.

Table A.5 Type-specific model fit statistics for mutation rate estimation strategies applied to the *de novo* testing data. Each type is shown in a sub-table, with the number of *de novo* mutations and non-mutated sites used in the partitioned testing data indicated in the subheading.

A>C (2920 *de novo* mutations; 198481 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.002	32831
5-mers	0.007	32701
7-mers	0.009	32641
7-mers+features	0.009	32636
7-mers (downsampled BRIDGES ERVs)	0.008	32670
7-mers (BRIDGES MAC10+ SNVs)	0.003	32809
7-mers (1000G intergenic SNVs)	0.004	32775

A>G (11400 *de novo* mutations; 198793 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.039	91474
5-mers	0.065	89455
7-mers	0.068	89212
7-mers+features	0.069	89111
7-mers (downsampled BRIDGES ERVs)	0.064	89505
7-mers (BRIDGES MAC10+ SNVs)	0.061	89732
7-mers (1000G intergenic SNVs)	0.061	89746

A>T (2455 *de novo* mutations; 198320 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.015	28130
5-mers	0.016	28114
7-mers	0.016	28106
7-mers+features	0.016	28105
7-mers (downsampled BRIDGES ERVs)	0.007	28350
7-mers (BRIDGES MAC10+ SNVs)	0.001	28498
7-mers (1000G intergenic SNVs)	0.003	28463

non-CpG C>A (3620 *de novo* mutations; 128765 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.012	35362
5-mers	0.022	35039
7-mers	0.03	34794
7-mers+features	0.032	34743
7-mers (downsampled BRIDGES ERVs)	0.029	34823
7-mers (BRIDGES MAC10+ SNVs)	0.024	35000
7-mers (1000G intergenic SNVs)	0.027	34892

non-CpG C>G (3561 *de novo* mutations; 128746 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.006	35889
5-mers	0.018	35490
7-mers	0.024	35321
7-mers+features	0.024	35321
7-mers (downsampled BRIDGES ERVs)	0.023	35350
7-mers (BRIDGES MAC10+ SNVs)	0.019	35480
7-mers (1000G intergenic SNVs)	0.018	35489

non-CpG C>T (10321 *de novo* mutations; 128774 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.005	79879
5-mers	0.012	79502
7-mers	0.014	79379
7-mers+features	0.014	79353
7-mers (downsampled BRIDGES ERVs)	0.013	79395
7-mers (BRIDGES MAC10+ SNVs)	0.012	79487
7-mers (1000G intergenic SNVs)	0.013	79434

CpG>ApG (304 *de novo* mutations; 6108 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.014	2788
5-mers	0.024	2767
7-mers	0.027	2763
7-mers+features	0.029	2761
7-mers (downsampled BRIDGES ERVs)	0.025	2763
7-mers (BRIDGES MAC10+ SNVs)	0.022	2771
7-mers (1000G intergenic SNVs)	0.025	2762

CpG>GpG (270 *de novo* mutations; 6292 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.013	2560
5-mers	0.015	2557
7-mers	0.022	2545
7-mers+features	0.026	2538
7-mers (downsampled BRIDGES ERVs)	0.015	2556
7-mers (BRIDGES MAC10+ SNVs)	0.015	2556
7-mers (1000G intergenic SNVs)	0.011	2564

CpG>TpG (6960 *de novo* mutations; 6289 non-mutated sites)

Model	Nagelkerke's R^2	AIC
3-mers	0.011	20321
5-mers	0.02	20232
7-mers	0.025	20173
7-mers+features	0.06	19777
7-mers (downsampled BRIDGES ERVs)	0.024	20182
7-mers (BRIDGES MAC10+ SNVs)	0.027	20151
7-mers (1000G intergenic SNVs)	0.027	20148

Table A.6 Genomic features used in mutation models

Feature	Source	Cell Type	Resolution
H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3	Roadmap Epigenomics Project [68]	Peripheral Blood Mononuclear Primary Cells	1bp (inside vs. outside of broad peak)
Replication timing	Koren et al., 2012 [165]	Lymphoblastoid	1kbp window
Recombination rate	Kong et al., 2010 [93] (deCODE sex-averaged recombination rate map)	—	10kbp window
Lamin B1 domains	Guelen et al., 2008 [166]	Tig3ET normal human embryonic lung fibroblasts	1bp (inside vs. outside of LAD)
DNase hypersensitivity sites	ENCODE [167]	multiple	1bp (inside vs. outside of DHS region)
Exonic site	RefSeq gene database	—	1bp (inside vs. outside of exon)
CpG island	Wu et al., 2010 [168]	—	1bp (inside vs. outside of CpG island)
% GC content	Calculated from reference genome	—	10kbp

A script to download the exact external data files used in this paper is available at https://github.com/carjed/smaug-genetics

Table A.7 Chi-squared tests for enrichment or depletion of *de novo* mutations occurring in feature-associated subtypes

Feature	Expected direction of effect	*de novo* relative mutation rate		p-value
		[a]Inside feature	[b]Outside feature	
H3K9me3[†]	Increased	1.98E-05	1.73E-05	**4.87E-05**
High Recombination rate (> 2)	Increased	3.66E-05	3.43E-05	0.18
H3K27me3[†]	Decreased	5.44E-06	3.14E-06	0.99
H3K27ac	Decreased	1.22E-04	1.23E-04	0.50
Exons	Decreased	1.20E-04	8.66E-05	0.99
H3K4me1	Decreased	1.10E-04	1.40E-04	**1.84E-10**
H3K4me3[†]	Decreased	1.00E-04	1.50E-04	**4.92E-23**
H3K9ac[†]	Decreased	1.49E-05	7.49E-06	0.99
Lamin-associated domains	Increased	6.91E-05	7.46E-05	0.75
High GC content (> 0.55)	Decreased	1.23E-05	9.74E-06	0.82
	Increased	1.14E-05	4.65E-06	**6.61E-04**
H3K36me3	Decreased	4.73E-06	6.14E-06	**2.59E-03**
	Increased	1.99E-05	1.51E-05	**5.50E-10**
CpG Islands	Decreased	3.68E-05	1.60E-04	**5.00E-117**
	Increased	5.39E-06	6.69E-06	0.79
Late replication timing (< -1.25)*	Increased	6.18E-06	5.48E-06	**0.026**
Early replication timing (> 1.25)*	Increased	1.55E-05	8.06E-06	**2.25E-02**
DHS	Decreased	5.03E-05	3.08E-05	0.99
	Increased	1.75E-05	1.21E-05	**4.92E-04**

Significant differences that are consistent with the expected direction of effect are indicated by a one-sided p-value in bold. [†]Four features had associations in the opposite direction, but these predicted effects could not be tested due to a lack of *de novo* mutations observed within the associated subtypes. *Some subtypes showed a significant *negative* association with replication timing, such that the mutation rate would be higher in *early*- rather than late-replicating regions, so we tested these subtypes separately.

Appendix B: Supplementary Material for Chapter III

Table B.1 T-tests for population differences in parameter estimates from mixture models

	rate			lambda		
Mixture component	**AFR mean rate**	**EUR mean rate**	**p.value**	**AFR mean lambda**	**EUR mean lambda**	**p.value**
1	1/27	1/17	1.06E-01	0.0075	0.0092	2.87E-04
2	1/5716	1/2202	6.07E-295	0.1128	0.0595	1.40E-206
3	1/60869	1/78123	7.23E-158	0.4485	0.3150	0.00E+00
4	1/260723	1/282923	1.50E-78	0.4312	0.6164	0.00E+00

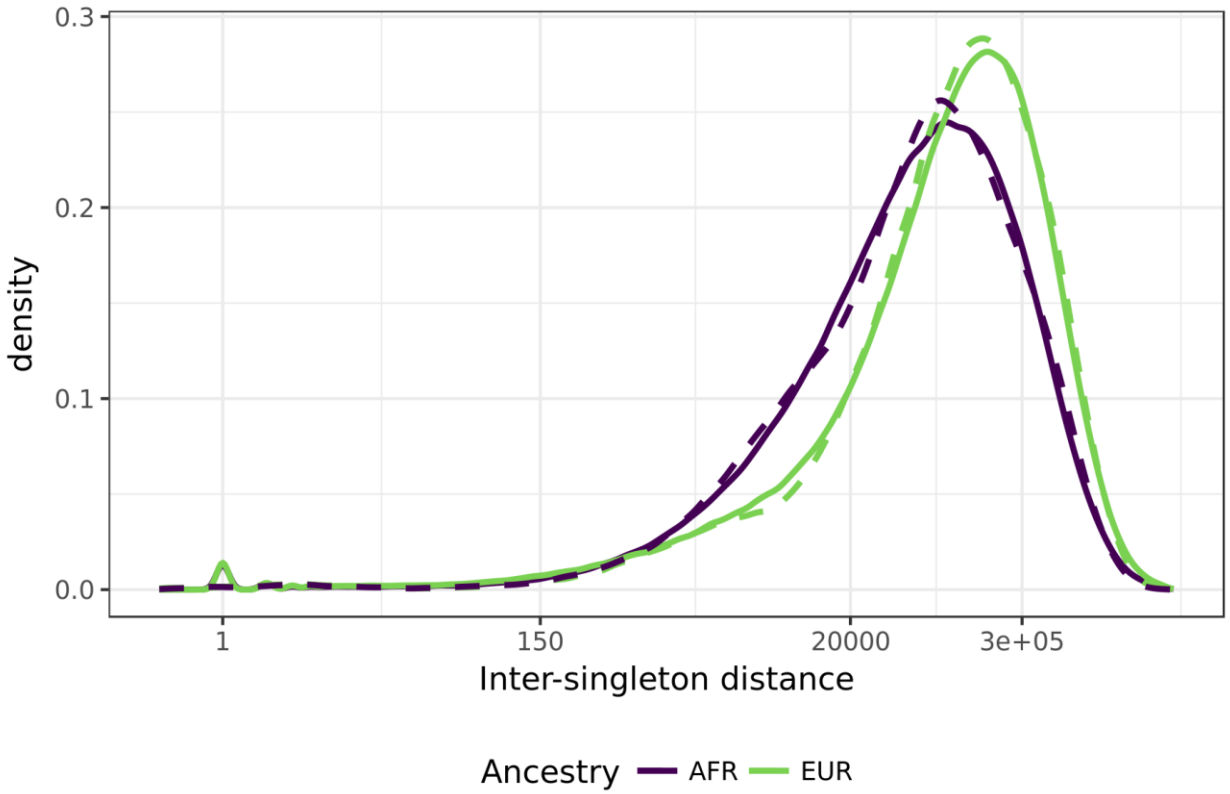


Figure B.1 Mixture models accurately describe observed inter-singleton distance distribution. Observed distributions are shown in solid lines, and expected distributions under a 4-component exponential mixture model are shown in dashed lines. Parameters for the 4-component mixture model were taken to be the median rate and lambda from each ancestry subsample.

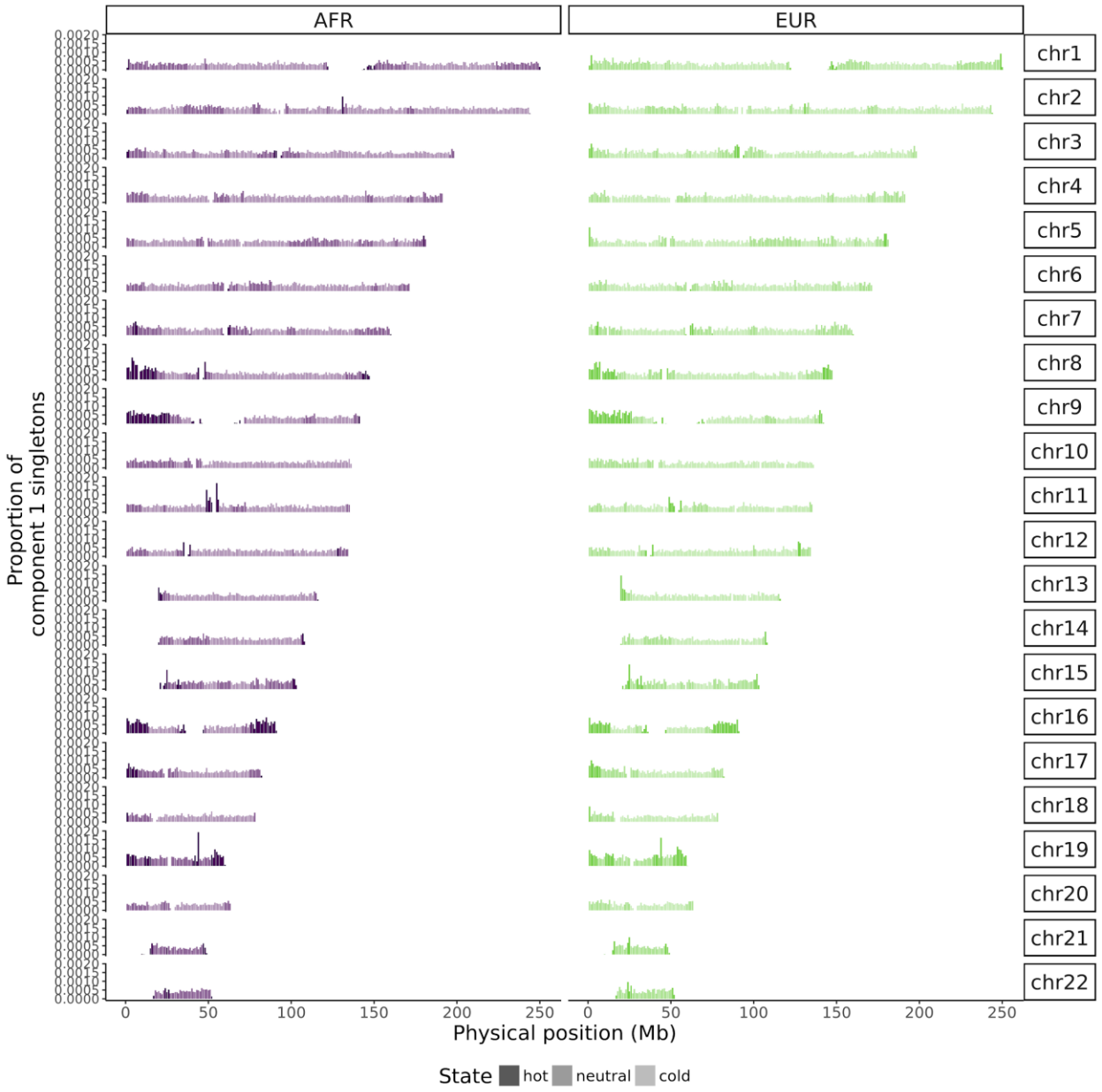


Figure B.2 Genome-wide distribution of component 1 singleton density, in 1Mbp windows. Windows are shaded according to the inferred state (hot, neutral, cold) from a 3-state HMM.

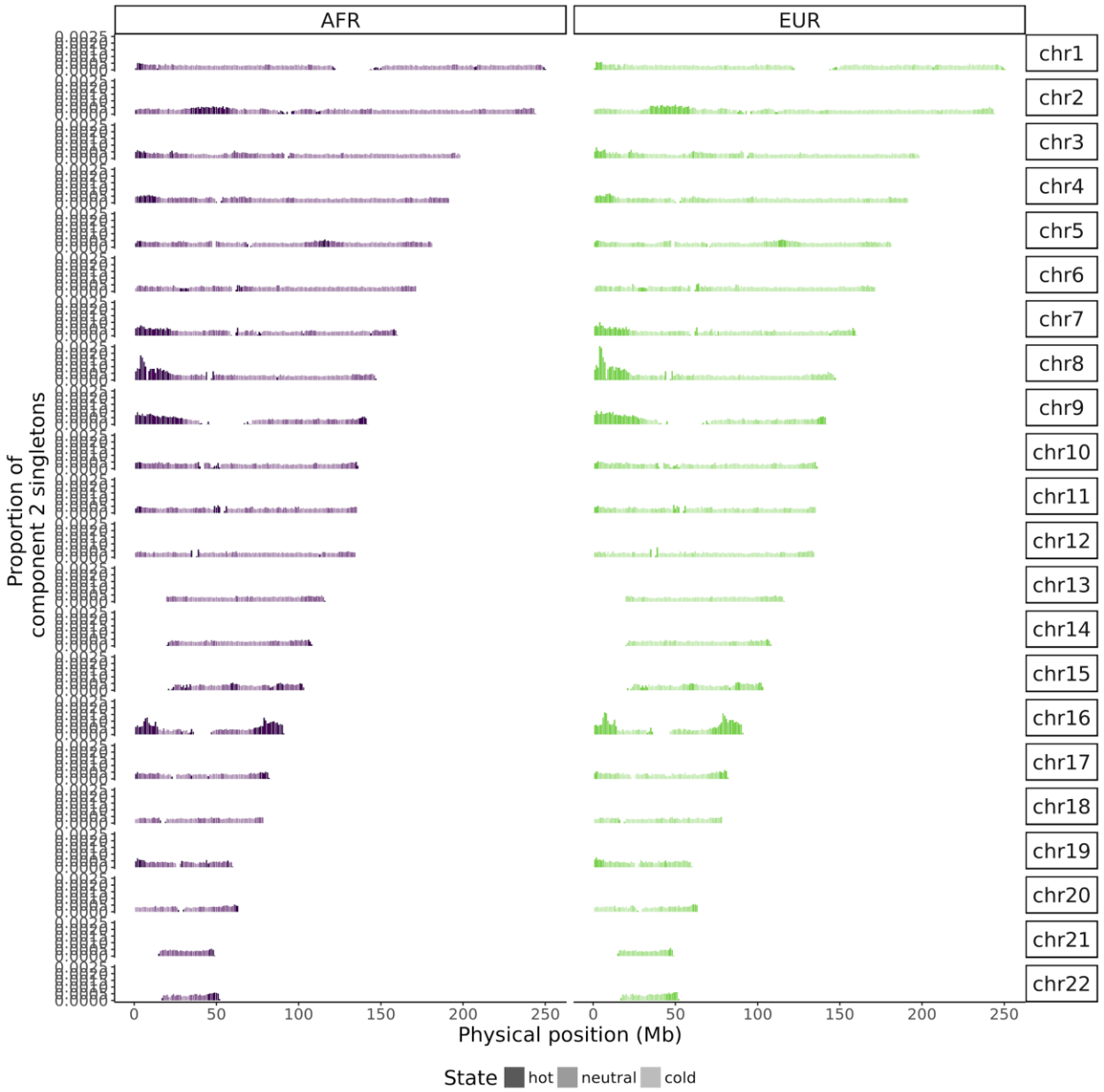


Figure B.3 Genome-wide distribution of component 2 singleton density, in 1Mb windows. Windows are shaded according to the inferred state (hot, neutral, cold) from a 3-state HMM.

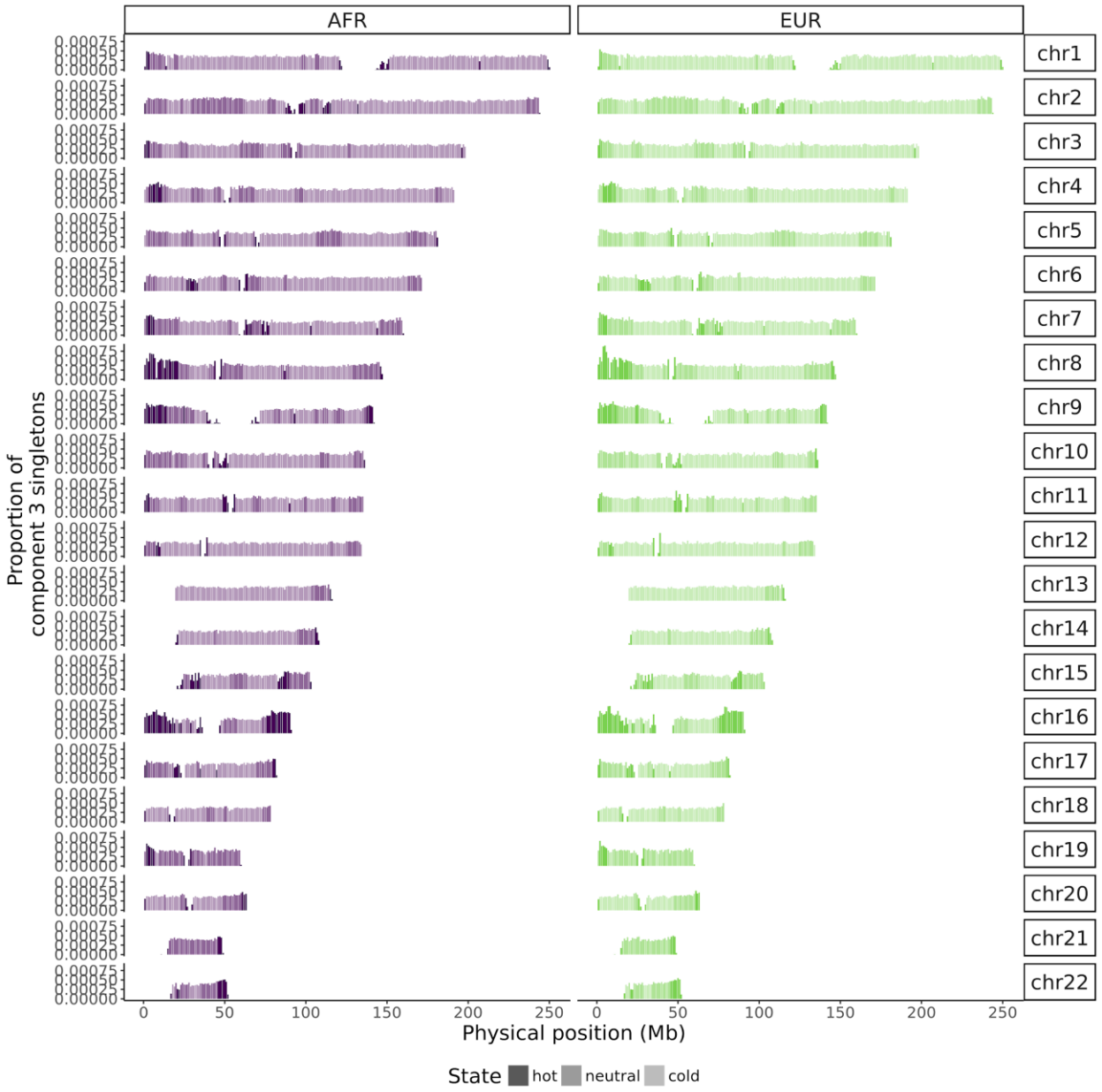


Figure B.4 Genome-wide distribution of component 3 singleton density, in 1Mbp windows. Windows are shaded according to the inferred state (hot, neutral, cold) from a 3-state HMM.

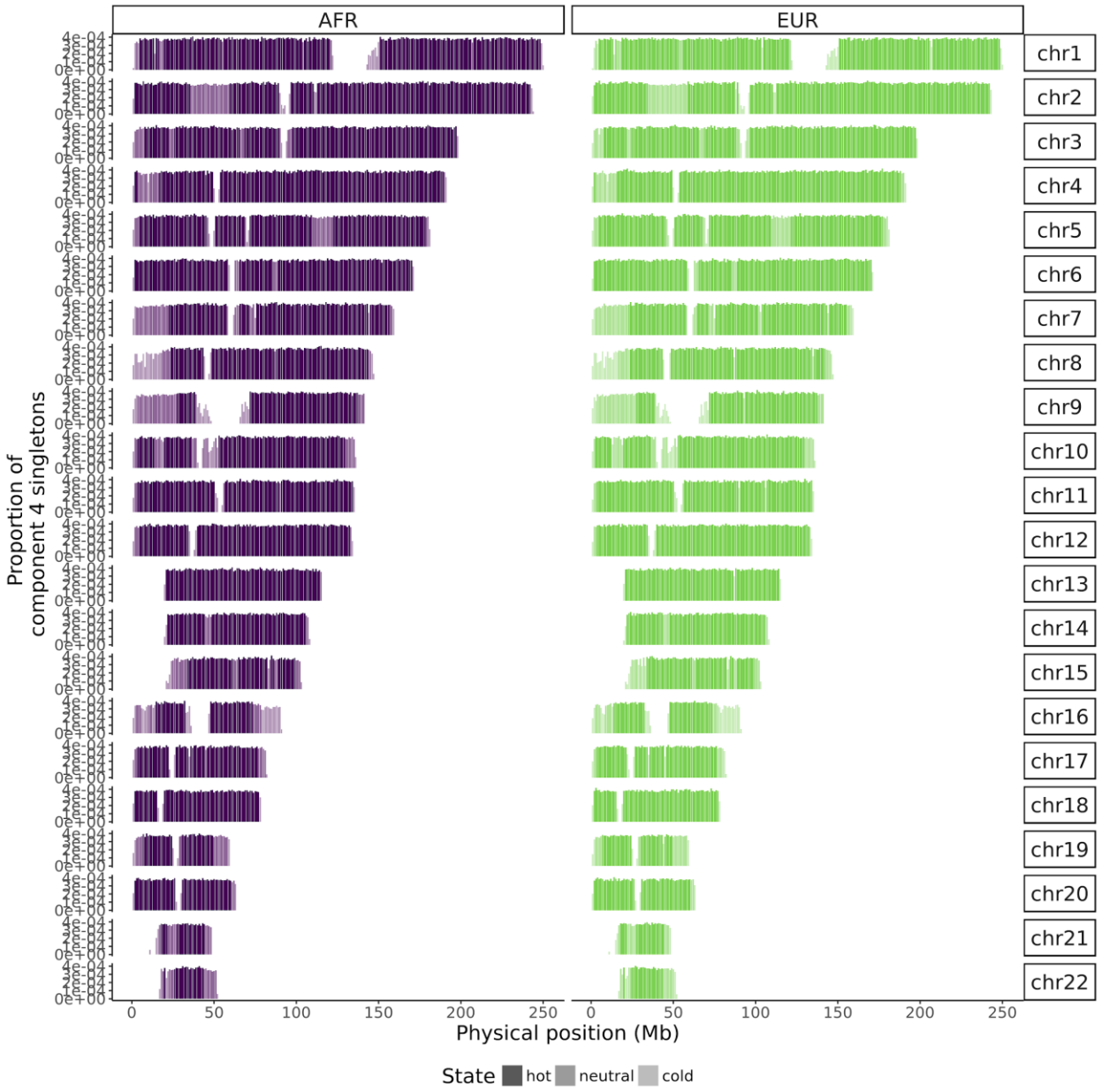


Figure B.5 Genome-wide distribution of component 4 singleton density, in 1Mbp windows. Windows are shaded according to the inferred state (hot, neutral, cold) from a 3-state HMM.

Appendix C: Supplementary Material for Chapter IV

Performance of other mutation signature analysis tools

In addition to the R packages we evaluated in the main paper, we considered several other mutation signature analysis tools which provide functions for generating mutation spectra matrices from VCF or Mutation Annotation Format (MAF) files. Where applicable, we used the small and full 1000 Genomes chromosome 22 VCF files described in the main paper. In our tests, each of these tools showed substantial performance bottlenecks compared to *Helmsman* or were subject to other limitations that made them infeasible for applying to our test datasets.

Mutagene and Mutalisk

Mutagene [30] and *Mutalisk* [31] are implemented as web servers and provide graphical interfaces for users to upload their data, with all data processing performed on the server end. We successfully uploaded the 158.6MB uncompressed small VCF file to *Mutalisk*, and the uploading and processing took approximately 60 seconds (compared to 8 seconds with *Helmsman*). *Mutagene* would not accept the small VCF file in either compressed or uncompressed format. Neither tool would accept the full VCF file when we attempted to upload it. Moreover, although *Mutalisk* at first appeared to offer reasonably fast performance for the small VCF file, we found that it did not properly parse the data into 2,504 unique samples as expected, and incorrectly assumed the SNVs were all from a single sample. *Mutalisk* does allow users to upload multiple

single-sample VCF files, but limits input to 300 files, and is therefore only feasible for relatively small sample sizes.

MutSpec

MutSpec is implemented as a Galaxy toolbox, enabling users with limited programming expertise to perform mutation signature analysis with a graphical interface [155]. Though we did not have a Galaxy server available to directly evaluate *MutSpec's* performance on our test datasets, the authors reported that it takes ~7 minutes to annotate a VCF file containing 100,000 variants (in an unstated sample size) using 24 CPUs, and 4 hours using a single CPU. Assuming *MutSpec's* runtime scales linearly with the number of SNVs, we estimate that it would take at least 60 seconds using 24 CPUs and over 40 minutes using a single CPU to parse the 15,971 SNVs in the small VCF file, compared to 8 seconds on a single CPU when using *Helmsman*. Similarly, to parse the 1,055,454 SNVs contained in the full chromosome 22 VCF file used in our tests, we estimate that *MutSpec* would take over an hour when using 24 CPUs, and over 40 hours on a single CPU, compared to 8 minutes on a single CPU when using *Helmsman*.

We note that these performance estimates for *MutSpec* are based on the reported runtime only for the annotation step of the *MutSpec* pipeline, which generates an intermediate tab-delimited file containing functional and structural annotations for each SNV in the input VCF file. Our estimates did not take into account the additional processing time required to parse this intermediate file into the Nx96 mutation spectra matrix, so our estimates represent a lower bound for the runtime necessary to generate the mutation spectra matrix using *MutSpec*.

Maftools and Mutation-Signatures

Somatic mutation data are sometimes represented in Mutation Annotation Format (MAF) files, a tab-delimited format with one variant per row, and several dozen additional annotation columns (described in detail at https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/). Unlike VCF files, MAF files do not indicate the genotypes of each individual in the sample, so a variant present in two or more individuals must be indicated as multiple rows. We considered two programs designed specifically for applying mutation signature analysis to MAF files: *maftools*, an R package [159], and *Mutation-Signatures*, an unpublished collection of Python scripts developed by researchers at Memorial Sloan Kettering Cancer Center in New York, NY (https://github.com/mskcc/mutation-signatures).

We evaluated the performance of these MAF-specific tools using a MAF file with data from 377 Liver Hepatocellular Carcinoma (LIHC) samples, available from The Cancer Genome Atlas at https://portal.gdc.cancer.gov/legacy-archive/files/15ce66c6-0211-4f03-bd41-568d0818a044. This file was 1.4GB in size and contained 60,691 somatic SNVs (interspersed with 1,415,224 non-SNV variants that are not considered in this type of analysis).

Helmsman generated the mutation spectra matrix from this MAF file in 96 seconds and required less than 130MB of memory (**Fig. C.1**). Both *Mutation-Signatures* and *maftools* generated output identical to that of *Helmsman*. *Mutation-Signatures* performs this task in two steps, first creating an intermediate MAF file with each SNV annotated with the surrounding trinucleotide context, then parsing this file to generate the mutation spectra matrix and, in the same step, performing supervised decomposition of each sample into 30 pre-specified signatures. *Mutation-Signatures* took a total of 402 seconds to run these scripts (145 seconds to generate the intermediate MAF file, and 377 seconds to generate the mutation spectra matrix and perform the

signature decomposition), with a maximum memory footprint of 6.5GB (memory usage peaked when generating the intermediate MAF file) (**Fig. C.1**). *Maftools* took 207 seconds and required 9.2GB of memory to read the same input MAF file and generate the mutation spectra matrix, using the functions read.maf and trinucleotideMatrix, respectively (**Fig. C.1**). Like the VCF-specific R packages we evaluated in the main paper, we note that *maftools* is memory-intensive, even for relatively small input files, and susceptible to memory bottlenecks as the input file size increases. *Maftools*' high memory usage and longer processing time is largely attributable to the inherently high dimensionality of MAF files: although only five columns (Chromosome, Position, Reference Allele, Alternative Allele, and Sample ID) are necessary to generate the mutation spectra matrix, *Maftools* requires the input MAF file to contain many additional mandatory columns.

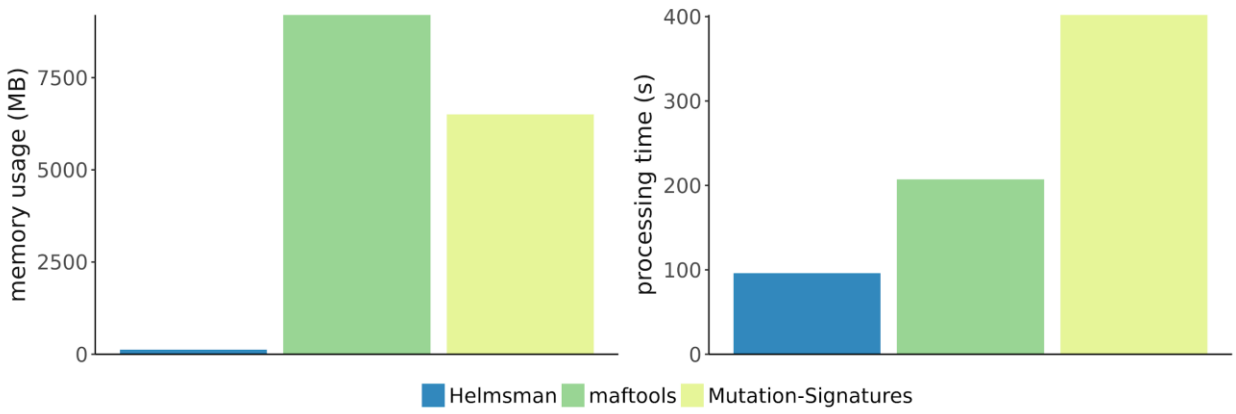


Figure C.1. Performance comparison for generation of the mutation spectra matrix from a MAF file. The MAF file used contained 60,691 SNVs (in addition to 1,415,224 non-SNV variants that were present in the file but not analyzed) in 377 samples.

Appendix D: Supplementary Material for Chapter V

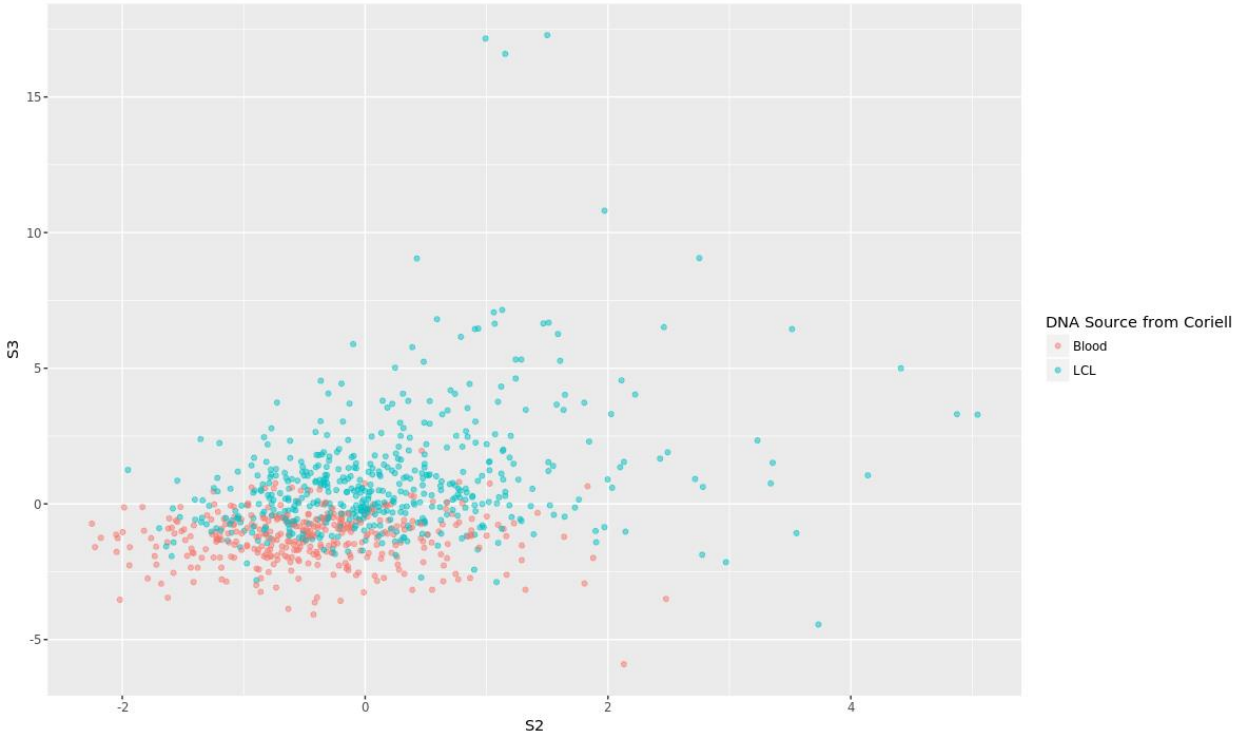


Figure D.1. Singleton SNV spectra differ according to DNA source in 1000 Genomes data. We obtained data for the DNA source (either fresh blood or LCL cell lines) of 870 of the 2,504 1000 Genomes Phase 3 samples from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130606_sample_info/20130606_sample_info.txt. Each point in this figure represents a sample, colored according to the DNA source, and the x and y axes correspond to principal components 2 and 3, determined from applying PCA to the 2,504x96 3-mer singleton SNV spectra matrix. The remaining 1634 samples had missing information for their DNA source, so are excluded from this figure.

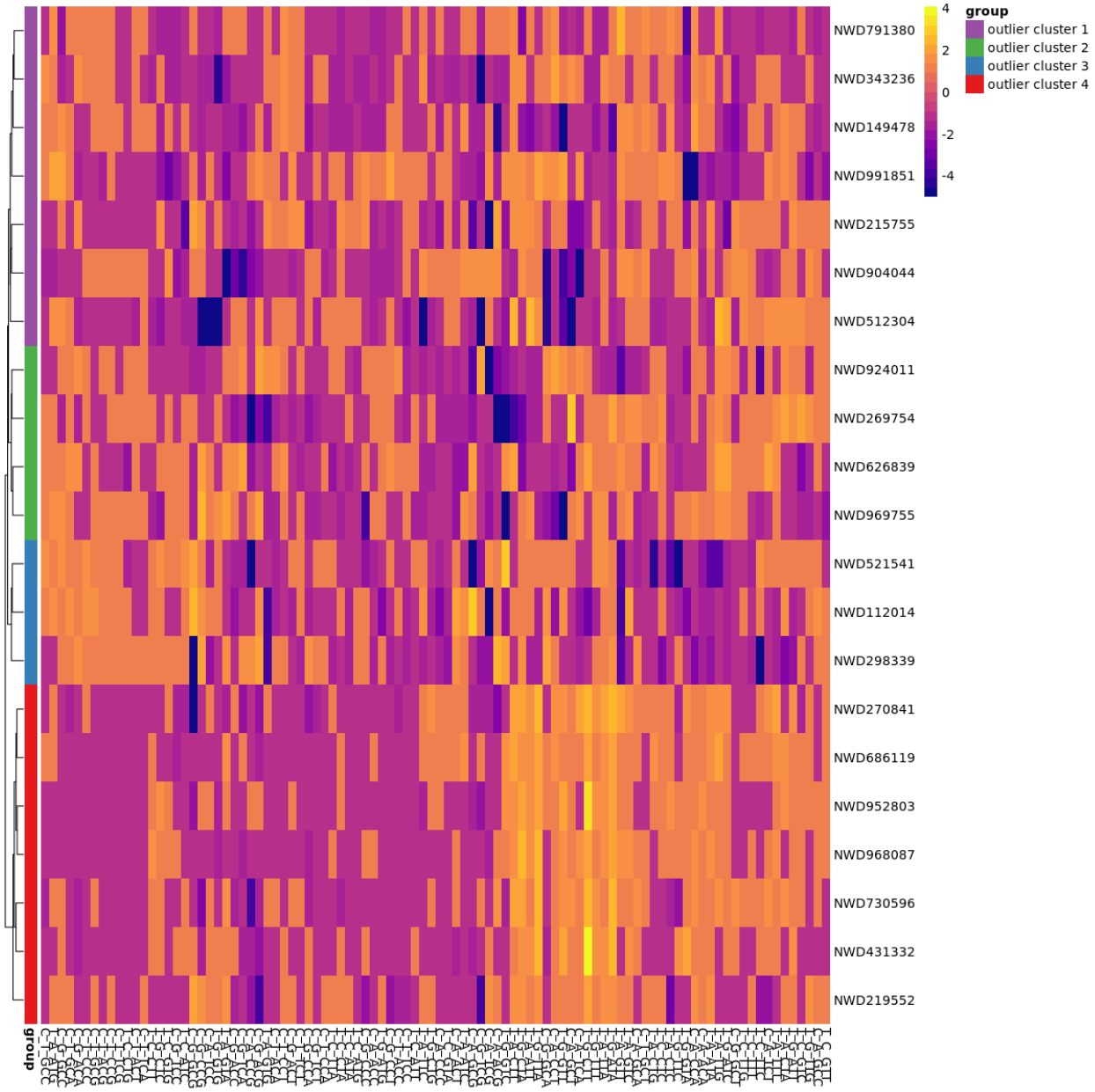


Figure D.2. Summary of outliers detected in the Framingham Heart Study data. Heatmap showing the relative enrichment or depletion for each of the 96 3-mer subtypes (columns) in each of the 27 flagged outliers (rows). The color in each cell indicates the fold-difference of the contribution of the subtype in that sample, calculated relative to the mean contribution of that subtype across all non-outlier samples. The scale is truncated to +/-5-fold difference. Only outlier cluster 4 (indicated by the red bar to the left of the heatmap), containing 7 samples, was found to have multiple statistically significant differences in spectra when compared to the non-outlier samples, as implemented in the secondary cluster-based filtering. These outliers show higher-than-expected rates of T>A and T>G transversions at NTT and NTA motifs, with a particularly strong enrichment at the C[T>G]T subtype.

Table D.1. Frequencies of BRIDGES outliers by sequencing plate

Outlier cluster	Plate	# Samples
outlier cluster 1	plate33B	1
outlier cluster 1	plate35	7
outlier cluster 1	plate36	11
outlier cluster 1	plate37	19
outlier cluster 1	plate38	2
outlier cluster 1	plate42	1
outlier cluster 1	plate43	24
outlier cluster 1	plate44	9
outlier cluster 1	<NA>	1
outlier cluster 2	plate01	14
outlier cluster 2	plate02	10
outlier cluster 2	plate19	1
outlier cluster 2	plate33B	2
outlier cluster 2	plate37	1
outlier cluster 2	plate42	2
outlier cluster 2	plate43	5
outlier cluster 2	<NA>	3
outlier cluster 3	plate05	1
outlier cluster 3	plate06	3
outlier cluster 3	plate08	1
outlier cluster 3	plate09	1
outlier cluster 3	plate10	2
outlier cluster 3	plate36	2
outlier cluster 3	plate37	25

Table D.2. Summary of outliers detected in 1000 Genomes Phase 3 dataset

ID	Family ID	Population
HG00182	HG00182	FIN
HG00186	HG00186	FIN
HG00272	HG00272	FIN
HG00373	HG00373	FIN
HG01149	CLM12	CLM
HG01377	CLM38	CLM
HG02582	GB23	GWD
HG02645	GB38	GWD
HG02839	GB74	GWD
HG03127	NG35	ESN
HG04131	BD44	BEB
NA12340	*1330**	*CEU*
NA12341	*1330**	*CEU*
NA12342	*1330**	*CEU*
NA12748	1444	CEU
NA12890	1463	CEU
NA18498	Y003	YRI
NA18608	NA18608	CHB
NA19175	Y044	YRI

*Three CEU samples, indicated in italics, were members of a single parent-offspring trio (Family ID 1330).

Table D.3. 1-mer spectra of non-transmitted SNVs in non-outlier parents compared to SNVs transmitted from outlier parents to offspring in the FHS dataset

SNV Type	Frequency in non-transmitted SNVs in non-outlier parents (% of total)	Frequency in SNVs transmitted from outlier parents to offspring (% of total)
C>A	3183 (9.8%)	2495 (9.8%)
C>G	2968 (9.2%)	2328 (9.1%)
C>T	12860 (39.7%)	9888 (38.7%)
T>A	2124 (6.6%)	1718 (6.7%)
T>C	8869 (27.4%)	7193 (28.2%)
T>G	2389 (7.4%)	1928 (7.5%)
Total	32393	25550

BIBLIOGRAPHY

1. Dobzhansky T. Chance and Creativity in Evolution. In: Ayala FJ, Dobzhansky T, editors. Studies in the Philosophy of Biology: Reduction and Related Problems. London: Macmillan Education UK; 1974. p. 307–38.

2. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187:367–83.

3. Schneider S, Excoffier L. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics. 1999;152:1079–89.

4. Przeworski M. The signature of positive selection at randomly chosen loci. Genetics. 2002;160:1179–89.

5. Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet. 2014;15:47–70.

6. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. Molecular basis of base substitution hotspots in Escherichia coli. Nature. 1978;274:775–80.

7. Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature. 1980;287:560–1.

8. Shibutani S, Takeshita M, Grollman AP. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. Nature. 1991;349:431–4.

9. Pfeifer GP, You Y-H, Besaratinia A. Mutations induced by ultraviolet light. Mutat Res. 2005;571:19–31.

10. Echols H, Goodman MF. Fidelity mechanisms in DNA replication. Annu Rev Biochem. 1991;60:477–511.

11. Panchin AY, Mitrofanov SI, Alexeevski AV, Spirin SA, Panchin YV. New words in human mutagenesis. BMC Bioinformatics. 2011;12:268.

12. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet. 2016;48:349–55.

13. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012;151:1431–42.

14. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet. 2015;47:822–6.

15. Messer PW. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. Genetics. 2009;182:1219–32.

16. Chan K, Gordenin DA. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. Annu Rev Genet. 2015;49:243–67.

17. Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, et al. Multi-nucleotide de novo Mutations in Humans. PLoS Genet. 2016;12:e1006315.

18. Kaplanis J, Akawi N, Gallone G, McRae JF, Prigmore E, Wright CF, et al. Mutational origins and pathogenic consequences of multinucleotide mutations in 6,688 trios with developmental disorders. bioRxiv. 2018;:258723. doi:10.1101/258723.

19. Schrider DR, Hourmozdi JN, Hahn MW. Pervasive multinucleotide mutational events in eukaryotes. Curr Biol. 2011;21:1051–4.

20. Venkat A, Hahn MW, Thornton JW. Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nature Ecology & Evolution. 2018;:1.

21. Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res. 2014;24:1445–54.

22. Goldmann JM, Seplyarskiy VB, Wong WSW, Vilboux T, Neerincx PB, Bodian DL, et al. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. Nat Genet. 2018. doi:10.1038/s41588-018-0071-6.

23. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017;1:16027.

24. Qin Y, Yalamanchili HK, Qin J, Yan B, Wang J. The Current Status and Challenges in Computational Analysis of Genomic Big Data. Big Data Research. 2015;2:12–8.

25. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013;3:246–59.

26. Alexandrov L, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, et al. The Repertoire of Mutational Signatures in Human Cancer. bioRxiv. 2018;:322859. doi:10.1101/322859.

27. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics. 2015;31:3673–5.

28. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016;17:31.

29. Rosales RA, Drummond RD, Valieris R, Dias-Neto E, da Silva IT. signeR: an empirical Bayesian approach to mutational signature discovery. Bioinformatics. 2017;33:8–16.

30. Goncearenco A, Rager SL, Li M, Sang Q-X, Rogozin IB, Panchenko AR. Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic Acids Res. 2017;45:W514–22.

31. Lee J, Lee AJ, Lee J-K, Park J, Kwon Y, Park S, et al. Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. Nucleic Acids Res. 2018. doi:10.1093/nar/gky406.

32. Lo YY. Statistical Methods, Analyses and Applications for Next-Generation Sequencing Studies. University of Michigan; 2015. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/116761/yancylo_1.pdf?sequence=1&isAllowed=y.

33. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. 2015;25:918–25.

34. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.

35. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6.

36. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15:1566–75.

37. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508:469–76.

38. Zhang W, Bouffard GG, Wallace SS, Bond JP, NISC Comparative Sequencing Program. Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. J Mol Evol. 2007;65:207–14.

39. Lercher MJ, Hurst LD. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. 2002;18:337–40.

40. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488:471–5.

41. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, et al. Timing,

rates and spectra of human germline mutation. Nat Genet. 2016;48:126–33.

42. Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, et al. Parent-of-origin-specific signatures of de novo mutations. Nat Genet. 2016. doi:10.1038/ng.3597.

43. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000;156:297–304.

44. Jiang C, Zhao Z. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics. 2006;88:527–34.

45. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437:69–87.

46. Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, St Jean PL, et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome Res. 2013;23:1974–84.

47. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311.

48. Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. Elife. 2017;6. doi:10.7554/eLife.24284.

49. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. Localizing recent adaptive evolution in the human genome. PLoS Genet. 2007;3:e90.

50. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet. 2007;8:857–68.

51. Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 2009;5:e1000336.

52. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493:216–20.

53. Rashkin S, Jun G, Chen S, Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), Abecasis GR. Optimal sequencing strategies for identifying disease-associated singletons. PLoS Genet. 2017;13:e1006811.

54. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

55. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

56. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K

project identifies rare variants in health and disease. Nature. 2015;526:82–90.

57. Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science. 2012;337:100–4.

58. Meunier J, Duret L. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol. 2004;21:984–90.

59. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488:504–7.

60. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015;521:81–4.

61. Li F, Mao G, Tong D, Huang J, Gu L, Yang W, et al. The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. Cell. 2013;153:590–600.

62. Supek F, Lehner B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell. 2017;170:534–47.e23.

63. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nat Biotechnol. 2014;32:71–5.

64. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature. 2016;532:264–7.

65. Fryxell KJ, Moon W-J. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol. 2005;22:650–8.

66. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet. 2011;44:40–6.

67. Balasubramanian D, Akhtar-Zaidi B, Song L, Bartels CF, Veigl M, Beard L, et al. H3K4me3 inversely correlates with DNA methylation at a large class of non-CpG-island-containing start sites. Genome Med. 2012;4:47.

68. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

69. Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer Science & Business Media; 2003.

70. Matsuda T, Bebenek K, Masutani C, Hanaoka F, Kunkel TA. Low fidelity DNA synthesis by human DNA polymerase-eta. Nature. 2000;404:1011–3.

71. Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JWH. Differential DNA repair

underlies mutation hotspots at active promoters in cancer genomes. Nature. 2016;532:259–63.

72. Mantovani R. A survey of 178 NF-Y binding CCAAT boxes. Nucleic Acids Res. 1998;26:1135–43.

73. Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A. 1997;94:1872–7.

74. Servant G, Streva VA, Derbes RS, Wijetunge MI, Neeland M, White TB, et al. The Nucleotide Excision Repair Pathway Limits L1 Retrotransposition. Genetics. 2017;205:139–53.

75. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ. Understanding nucleotide excision repair and its roles in cancer and ageing. Nat Rev Mol Cell Biol. 2014;15:465–81.

76. Strauss BS. The "A" rule revisited: polymerases as determinants of mutational specificity. DNA Repair . 2002;1:125–35.

77. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.

78. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002;18:337–8.

79. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46:944–50.

80. Harris K. Evidence for recent, population-specific evolution of the human mutation rate. Proc Natl Acad Sci U S A. 2015;112:3439–44.

81. Mathieson I, Reich D. Differences in the rare variant spectrum among human populations. PLoS Genet. 2017;13:e1006581.

82. Scott LJ, Muglia P, Kong XQ, Guan W, Flickinger M, Upmanyu R, et al. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc Natl Acad Sci U S A. 2009;106:7501–6.

83. Pato MT, Sobell JL, Medeiros H, Abbott C, Sklar BM, Buckley PF, et al. The genomic psychiatry cohort: partners in discovery. Am J Med Genet B Neuropsychiatr Genet. 2013;162B:306–12.

84. Langenecker SA, Saunders EFH, Kade AM, Ransom MT, McInnis MG. Intermediate: cognitive phenotypes in bipolar disorder. J Affect Disord. 2010;122:285–93.

85. Sklar P, Smoller JW, Fan J, Ferreira MAR, Perlis RH, Chambert K, et al. Whole-genome association study of bipolar disorder. Mol Psychiatry. 2008;13:558–69.

86. Miller MB, Basu S, Cunningham J, Eskin E, Malone SM, Oetting WS, et al. The Minnesota Center for Twin and Family Research genome-wide association study. Twin Res Hum Genet.

2012;15:767–74.

87. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9.

88. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012;91:839–48.

89. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.

90. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

91. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.

92. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.

93. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467:1099–103.

94. Pinto Y, Gabay O, Arbiza L, Sams AJ, Keinan A, Levanon EY. Clustered mutations in hominid genome evolution are consistent with APOBEC3G enzymatic activity. Genome Res. 2016;26:579–87.

95. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.

96. Schrider DR, Houle D, Lynch M, Hahn MW. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. Genetics. 2013;194:937–54.

97. Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nat Commun. 2018;9:3753.

98. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93:278–88.

99. Harpak A, Bhaskar A, Pritchard JK. Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans. PLoS Genet. 2016;12:e1006489.

100. Amos W. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? Proc Biol Sci. 2010;277:1443–9.

101. Waters LS, Minesinger BK, Wiltrout ME, D'Souza S, Woodruff RV, Walker GC. Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. Microbiol Mol Biol Rev. 2009;73:134–54.

102. Murnane JP. Telomere dysfunction and chromosome instability. Mutat Res. 2012;730:28–36.

103. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25:1010–22.

104. Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet. 2009;10:295–304.

105. Fernández AF, Bayón GF, Urdinguio RG, Toraño EG, García MG, Carella A, et al. H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. Genome Res. 2014. doi:10.1101/gr.169011.113.

106. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol. 2008;9:958–70.

107. Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, Cairns BR. Distinctive chromatin in human sperm packages genes for embryo development. Nature. 2009;460:473–8.

108. Borde V, Robine N, Lin W, Bonfils S, Géli V, Nicolas A. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. EMBO J. 2009;28:99–111.

109. Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. Genetic recombination is directed away from functional genomic elements in mice. Nature. 2012;485:642–5.

110. Tchurikov NA, Kretova OV, Fedoseeva DM, Chechetkin VR, Gorbacheva MA, Snezhkina AV, et al. Genome-wide mapping of hot spots of DNA double-strand breaks in human cells as a tool for epigenetic studies and cancer genomics. Genom Data. 2015;5:89–93.

111. Chen J-M, Cooper DN, Férec C. A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: ~0.4% of the single-nucleotide substitution mutation rate. Hum Mutat. 2014;35:392–4.

112. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput Biol. 2016;12:e1004842.

113. Sabbioneda S, Gourdin AM, Green CM, Zotter A, Giglia-Mari G, Houtsmuller A, et al. Effect of proliferating cell nuclear antigen ubiquitination and chromatin structure on the dynamic properties of the Y-family DNA polymerases. Mol Biol Cell. 2008;19:5193–202.

114. Falk M, Lukasova E, Kozubek S. Higher-order chromatin structure in DSB induction, repair and misrepair. Mutat Res. 2010;704:88–100.

115. Narasimhan VM, Rahbari R, Scally A, Wuster A, Mason D, Xue Y, et al. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. Nat Commun. 2017;8:303.

116. Seoighe C, Scally A. Inference of Candidate Germline Mutator Loci in Humans from Genome-Wide Haplotype Data. PLoS Genet. 2017;13:e1006549.

117. Dutta R, Saha-Mandal A, Cheng X, Qiu S, Serpen J, Fedorova L, et al. 1000 human genomes carry widespread signatures of GC biased gene conversion. BMC Genomics. 2018;19:256.

118. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics. 2004;1:274–86.

119. Benaglia T, Chauveau D, Hunter D, Young D. mixtools: An R Package for Analyzing Mixture Models. Journal of Statistical Software, Articles. 2009;32:1–29.

120. Visser I, Speekenbrink M. depmixS4: An R Package for Hidden Markov Models. Journal of Statistical Software, Articles. 2010;36:1–21.

121. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

122. Kumar-Sinha C, Chinnaiyan AM. Precision oncology in the age of integrative genomics. Nat Biotechnol. 2018;36:46–60.

123. Pedersen BS, Quinlan AR. cyvcf2: fast, flexible variant analysis with Python. Bioinformatics. 2017. doi:10.1093/bioinformatics/btx057.

124. Žitnik M, Zupan B. NIMFA : A Python Library for Nonnegative Matrix Factorization. J Mach Learn Res. 2012;13 Mar:849–53.

125. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12 Oct:2825–30.

126. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135–45.

127. Guo Y, Zhao S, Sheng Q, Ye F, Li J, Lehmann B, et al. Multi-perspective quality control of Illumina exome sequencing data using QC3. Genomics. 2014;103:323–8.

128. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. 2014;15:879–89.

129. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics. 2015;31:318–23.

130. Kircher M, Stenzel U, Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol. 2009;10:R83.

131. Allhoff M, Schönhuth A, Martin M, Costa IG, Rahmann S, Marschall T. Discovering motifs that induce sequencing errors. BMC Bioinformatics. 2013;14 Suppl 5:S1.

132. Johnston HR, Hu Y, Cutler DJ. Population genetics identifies challenges in analyzing rare variants. Genet Epidemiol. 2015;39:145–8.

133. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14:R51.

134. Cheng AY, Teo Y-Y, Ong RT-H. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics. 2014;30:1707–13.

135. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149:979–93.

136. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. 2012;109:14508–13.

137. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013;41:e67.

138. Li B, Zhan X, Wing M-K, Anderson P, Kang HM, Abecasis GR. QPLOT: a quality assessment tool for next generation sequencing data. Biomed Res Int. 2013;2013:865181.

139. Kang HM. Efficient and Parallelizable Association Container Toolbox (EPACTS). 2016. https://github.com/statgen/EPACTS. Accessed 12 Jun 2018.

140. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun. 2015;6:8018.

141. Shirley MD, Ma Z, Pedersen BS, Wheelan SJ. Efficient “pythonic” access to FASTA files using pyfaidx. PeerJ PrePrints; 2015. doi:10.7287/peerj.preprints.970v1.

142. Carlson J, Li J, Zöllner S. Helmsman: fast and efficient generation of input matrices for mutation signature analysis. bioRxiv. 2018;:373076. doi:10.1101/373076.

143. Pimentel MAF, Clifton DA, Clifton L, Tarassenko L. A review of novelty detection. Signal Processing. 2014;99:215–49.

144. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc. 1963;58:236–44.

145. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. bioRxiv. 2017;:220814. doi:10.1101/220814.

146. Liu Y, Liang Y, Cicek AE, Li Z, Li J, Muhle RA, et al. A Statistical Framework for Mapping Risk Genes from De Novo Mutations in Whole-Genome-Sequencing Studies. Am J

Hum Genet. 2018;102:1031–47.

147. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature. 2018. doi:10.1038/nature25983.

148. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538:161–4.

149. Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: why the uneven progress? J Community Genet. 2017;8:255–66.

150. Rappoport N, Toung J, Hadley D, Wong RJ, Fujioka K, Reuter J, et al. A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth. Sci Rep. 2018;8:226.

151. Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, et al. New observations on maternal age effect on germline de novo mutations. Nat Commun. 2016;7:10486.

152. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14 Suppl 11:S1.

153. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.

154. Fischer A, Illingworth CJR, Campbell PJ, Mustonen V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol. 2013;14:R39.

155. Ardin M, Cahais V, Castells X, Bouaoun L, Byrnes G, Herceg Z, et al. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC Bioinformatics. 2016;17:170.

156. Gori K, Baez-Ortega A. sigfit: flexible Bayesian inference of mutational signatures. bioRxiv. 2018;:372896. doi:10.1101/372896.

157. Díaz-Gay M, Vila-Casadesús M, Franch-Expósito S, Hernández-Illán E, Lozano JJ, Castellví-Bel S. Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples. BMC Bioinformatics. 2018;19:224.

158. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 2018;10:33.

159. Mayakonda A, Phillip Koeffler H. Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. bioRxiv. 2016;:052662. doi:10.1101/052662.

160. Shiraishi Y, Tremmel G, Miyano S, Stephens M. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. PLoS Genet. 2015;11:e1005657.

161. Chen L, Liu P, Evans TC Jr, Ettwiller LM. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science. 2017;355:752–6.

162. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.

163. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 2011;12:R112.

164. Horvath JE, Viggiano L, Loftus BJ, Adams MD, Archidiacono N, Rocchi M, et al. Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. Hum Mol Genet. 2000;9:113–23.

165. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. Am J Hum Genet. 2012;91:1033–40.

166. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008;453:948–51.

167. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.

168. Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP. Redefining CpG islands using hidden Markov models. Biostatistics. 2010;11:499–514.