Some problems in probabilistic modeling of germline and somatic evolutionary processes

William S. DeWitt

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Kelley Harris, Chair

Frederick A. Matsen, Chair

Joseph Felsenstein

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Some problems in probabilistic modeling
of germline and somatic evolutionary processes

William S. DeWitt

Co-Chairs of the Supervisory Committee:
Kelley Harris
Department of Genome Sciences

Frederick A. Matsen
Department of Statistics

Evolutionary processes shape biological systems at all scales, and understanding evolutionary mechanisms requires quantitative frameworks that are matched in sophistication to modern experimental capabilities. This dissertation covers quantitative work along two biological threads—evolutionary genomics and adaptive immunology. I describe how complex dynamics of mutational activity in evolving populations can be recovered from population-level whole-genome sequencing data, and show results on mutation spectrum evolution over thousands of generations in humans. Next, I describe inference of evolutionary histories in a regime of dense single-cell sampling of cellular diversification, where identical genotypes from clonal subpopulations are sampled, and genotype abundance influences the mutational output of a clone because it is closely related to clonal population size. In particular, I address phylogenetic tree inference for B cells evolving improved antibodies. I conclude with an outlook for future research that synthesizes evolutionary genomics and adaptive immunology, and views the latter as a powerful evolutionary model system.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

In 2016 I met Erick Matsen at a cafe in Seattle to discuss working together on adaptive immunology. I had been working on immune repertoire sequencing technology in a biotech industry job, and was interested in moving to academic research and getting a PhD. Erick posed questions about immunology in subtle probabilistic and evolutionary language, revealing to me a larger universe of ideas that could be brought to bear in this corner of biology. I sent a computer code sample for vetting (which Erick judged as "good, if not quite tidy"), and joined the Matsen group as a programmer, eventually transitioning into PhD student. I have learned an incredible amount from Erick, and am greatly indebted to him for my academic growth and success during a nontraditionally-timed PhD.

As I began my PhD I became curious about population genetics and evolutionary genomics, initially for what the theoretical and computational approaches in these fields might have to offer for questions about immunology. My co-advisor Kelley Harris helped me explore this curiosity, and also led me to develop research directions in population genetics for its own sake. Thanks to Kelley's support, and the freedom she gave me to pursue my interests and grow collaborations, I now feel at home in the broader popgen research community.

In addition to Erick and Kelley, my committee members—Joe Felsenstein, Phil Green, and Armita Nourmohammad—provided essential feedback on my dissertation work and numerous talks, and support for my post-PhD aspirations.

I've benefited along the way from many members of both the Matsen group and Harris lab. Andy Magee introduced me to coalescent theory, working through problems in the Wakeley book at a blackboard one day. Later, Andy and I teamed up with Sarah Hilton

from the Bloom lab to write a sneaky paper without our PIs [132]. I had the good fortune of collaborating with Jean Feng on a project [60], and learned a lot. Jed Carlson showed me the way through a thorough literature review.

I've been lucky to work closely with outside collaborators who are some fantastic people, and from whom I've learned so much. Vladimir Minin generously shared his expertise in branching process theory. Phil Bradley showed me that immune repertoires are an endless goldmine of interesting biology if you just keep digging. Gabriel Victora hears the music of the germinal centers. Kameron Decker Harris makes sure all the math is certified fresh. Aaron Ragsdale: if you build it, he will come (and make it better). Yun Song has been generous with his time and advice. Jay Shendure was a generous advisor before, during, and after my rotation. Zach Montague, Armita Nourmohammad, and Jakub Otwinowski let me help in a small way in their COVID-19 work, and kept me engaged with physics perspectives on biology.

The local popgen community at UW has been an incredible source of wisdom, especially the lively discussion at *PopGenLunch*. I'm grateful to Peter Ralph and Andy Kern at UO for the engagement with their research community in Eugene, words of encouragement, and conference adventures.

The course sequence in Genome Sciences is tough, but was transformative for me. I benefited especially from Celeste Berg's and Mary Kuhner's teaching. My PhD student cohort is an amazing bunch, and I appreciate all the camaraderie in 1st year survival mode and beyond (and weekly virtual beer hours during long pandemic lockdown months). I co-organized Molecular Evolution Supergroup (*MolEvolve*) meetings with Katherine Xue, Damien Wilburn, and Allie Greaney, and enjoyed hearing about exciting new work from around Seattle. I'm grateful to Brian Giebel at UW Genome Sciences, and Melissa Alvendia at Fred Hutch Computational Biology, whose admin skills kept me on track.

Additional acknowledgments are due to advisors, mentors, and teachers from earlier periods of my life that helped me along this meandering path to a PhD. My high school english teacher Sue Bergeron and physics teacher Greg Renner helped me grow tremendously. At UVM, Junru Wu, Don Manley, Kevork Spartalian, and Ken Golden gave me opportunities and encouragement. Kelvin Chu advised me in my physics MS thesis at UVM, and also helped me navigate challenges during my BA. Mark Schneider taught me to apply science for engineering goals, and developed my independence. Yufeng Shen welcomed me to computational biology.

Lastly, some special people have helped keep me afloat and balanced during my time as a PhD student. Sarah Hilton has been incredibly supportive, and generally an amazing person to spend a lot of time with and learn from—our pizza-based collaborations have become highly successful! The Wales family—Tyler, Yuko, Caiden, Hana Hiro, and Totoro—have treated me like a family member. My brother John DeWitt MD PhD and step sister Joy DeWitt have stayed awesome. Armita Nourmohammad and Jakub Otwinowski have been essential pandemic pals and adventure buddies. Taylor Soja and Devin Saywers have been a source of history lessons, and baked good joy. Ben Feintzeig approaches powder days and C^* algebras with equal glee. Justin Ashworth, Tom Manos, Dave Ketchum, and Josh Payne brought wisdom, balance, and adventure. Some friends are gone; Freeman Robie, Nathan Shishido, and Aaron Powers all shared an unusual thirst for knowledge that drew me in, and I am lucky to have crossed their paths.

DEDICATION

For Freeman Robie.

Chapter 1

INTRODUCTION

VALENTINE: It's how you look at population changes in biology,... it can be written down as mathematics...
HANNAH: Does it work for grouse?
VALENTINE: I don't know yet... There's more noise with grouse.
HANNAH: Noise?
VALENTINE: Distortions. Interference. Real data is messy... Like a piano in the next room, it's playing your song, but unfortunately it's out of whack...
HANNAH: What do you do?
VALENTINE: You start guessing what the tune might be. You try to pick it out of the noise...
HANNAH: (soberly) Yes, I see. And then what?
VALENTINE: I publish.
HANNAH: Of course. Sorry. Jolly good.

Tom Stoppard, *Arcadia*

Evolutionary forces shape biological systems across scales, from germline genome variation over generations of individuals to somatic diversification of cell lineages within individuals. Understanding these forces requires quantitative approaches that are matched in sophistication to modern technologies for biological data acquisition. This dissertation presents quantitative research along two biological directions—evolutionary genomics and adaptive immunology—where experiments provide richly textured data with eccentricities that are important to learn how to model properly. Although seemingly disparate biological threads, a unifying perspective on this work is the power of tree-valued stochastic processes (to wit: Kingman's coalescent process and the Galton-Watson branching process) in understanding

diverse biological systems.

My work in evolutionary genomics (Chapter 2) combines the languages of inverse problems, sparse optimization, and coalescent theory to infer histories of complex mutational processes acting alongside demographic processes in evolving populations. Mutation is the source of genetic variation driving evolution, but receives a simplistic mathematical treatment in modern population genetic inference. Recent literature (reviewed in § 1.2) makes clear that mutagenesis is itself a complex evolving trait. I prove a theorem describing how mutation spectrum history shapes genomic data. Using modern sparse optimization, I develop software to infer mutation spectrum history, revealing mutation signatures varying over thousands of generations in humans, and patterns of global divergence in mutational processes. Additionally, in Chapter 3 I present a software utility for bioinformatic processing of mutation spectrum data from large population genomic data sets.

My work in adaptive immunology (Chapter 4) develops an approach to infer evolutionary trees for single B cell lineages evolving high-affinity antibodies. B cells undergo rapid evolution in microanatomical sites in lymph nodes called *germinal centers* (reviewed in § 1.3), where they diversity under selection for improved pathogen recognition. Intricate experimental techniques are now able to isolate these evolving lineages at single-cell resolution, and provide detailed physiological, molecular, and sequence-level data. I developed branching process theory and a novel dynamic programming algorithm to formulate tree inference in a regime of dense sampling of such cellular diversification.

1.1 Previous publication and co-authorship of dissertation content

This dissertation contains content previously published as follows, and includes text and materials contributed by all authors [43, 31, 40, 41]:

- William S. DeWitt, Luka Mesin, Gabriel D. Victora, Vladimir N. Minin, and Frederick A. Matsen IV, Using Genotype Abundance to Improve Phylogenetic Inference. *Molecular Biology and Evolution*, 35(5):1253–1265, 02 2018.

- Jedidiah Carlson, William S. DeWitt, and Kelley Harris. Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation. *Current Opinion in Genetics & Development*, 62:50–57, 2020. Genetics of Human Origin

- William S. DeWitt. mutyper: assigning and summarizing mutation types for analyzing germline mutation spectra. *bioRxiv*, 2020.

- William S. DeWitt, Kameron Decker Harris, Aaron P. Ragsdale, and Kelley Harris. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, 118(21):e2013798118, 2021

1.2 Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation

There are many possible failure points in the transmission of genetic information that can produce heritable germline mutations. Once a mutation has been passed from parents to offspring for several generations, it can be difficult or impossible to identify its root cause; however, sometimes the nature of the ancestral and derived DNA sequences can provide mechanistic clues about a genetic change that happened hundreds or thousands of generations ago. Here, we review evidence that the sequence context "spectrum" of germline mutagenesis has been evolving surprisingly rapidly over the history of humans and other species. We go on to discuss possible causal factors that might underlie rapid mutation spectrum evolution.

Like all other complex traits, the germline mutations present in an individual's genome are ultimately governed by heritable genetic factors, environmental influences, and interactions thereof. Commonly studied properties of germline mutations include their rate (i.e., mutations per site per generation), spectrum (i.e., relative abundances of different mutation types), or spatial clustering throughout the genome. Because these phenotypic outcomes of mutation are embedded as variation in the genome, the evolutionary pressures acting on mutation phenotypes are intrinsically more complex than the forces that drive the evolution of other phenotypes. The theoretical complexities of this phenomenon have fascinated population geneticists for decades. In 1937, A.H. Sturtevant was the first to show that alleles that

modify the mutation rate become linked over time to mutations that they beget in nearby genomic regions, leading to time-delayed selection on mutation rate modifiers that is ultimately dependent on the fitness effects of much younger genetic variants [203]. Thirty years after Sturtevant's seminal paper, population geneticists remained daunted by the higher-order complexity of mutation rate evolution: Motoo Kimura wrote in 1967 that "the whole question of the evolutionary modification of the spontaneous mutation rate is quite puzzling and more evidence is needed to clarify the problem" [107].

In the fifty-three years since Kimura's landmark theoretical work, large amounts of genetic data have been produced, proffering new empirical evidence of past and ongoing mutation rate evolution. It is now feasible to directly measure variability in mutation rates by generating mutation accumulation lines and/or sequencing parent/offspring trios, then test hypotheses about the genetic and environmental causes of ongoing within-species germline mutation rate variability. Comparative genomics approaches have shown that, over long timescales, mutation rates evolve toward a level that is approximately proportional to effective population size divided by the size of the coding fraction of the genome [204].

However, between the extremes of inter-species mutation rate evolution and within-species mutation rate variation, there has been much less documentation of historical germline mutation rate variation within populations or between closely related populations, a fact that stands in the way of studying the evolution of this trait using standard techniques from quantitative genetics. To make matters even more complicated, environmental mutagens might also affect mutation rates, as do life history features such as the ages at which individuals tend to become parents. In a neutral model setting, disentangling historical mutation rate evolution from the effects of demographic history is a serious challenge for population genetic inference. Indeed, the rate of genetic drift influencing genomic variation is jointly determined by mutation rate and effective population size.

An indirect hint that mutational modifiers do segregate within populations is the exis-

tence of variation between populations in their mutation spectra. Mutations can be coarsely classified into 6 types according to the ancestral and derived alleles (A→C, A→G, A→T, C→A, C→G, and C→T are all distinct mutation types, but T→G, for example, is not distinct from this set if we ignore strand-specific effects because it is the strand complement of A→C). There is considerable variation between microorganism species (and even strains of the same species) in the relative abundances of these variant types [120], but to ascertain differences between closely related eukaryotes that have larger genomes and more stable mutation rates, it is useful to partition these mutation types further by trinucleotide context (yielding 96 unique types such as AAA→ACA, etc.) or even by extended sequence contexts as large as 7-mers [5, 33]. Trinucleotide mutation spectrum variation has been extensively catalogued across cancer types, an effort that has led to the discovery of dozens of "mutational signatures" that appear to be associated with specific exogenous and endogenous DNA damage agents [9, 8]. More recently, comparisons of mutation spectra between populations of humans, great apes, and laboratory mice has revealed that mutation spectrum phenotypes often vary so predictably between populations that they can be used to identify an individual's population of origin [82, 134, 83, 135, 147, 47, 205, 72].

The most striking mutation spectrum difference discovered thus far between closely related human populations is an excess of TCC→TTC mutations that exists in Europeans relative to East Asians and Africans. In the 1000 Genomes Phase I dataset, where this difference was first reported [82], private European variation is enriched for TCC→TTC mutations by about 50% compared to private African or East Asian variation. South Asian genomes contain an intermediate amount of TCC→TTC mutations, perhaps because this population was founded by admixture of a European-related Ancestral North Indian population into an East-Asian-related Ancestral South Indian population [172, 146]. Based on the allele frequencies of the excess TCC→TTC mutations in Europe, it has been estimated that the rate of this mutation increased in Europe between 15,000 to 30,000 years ago [83, 199]

(or perhaps even earlier, depending on the details of ancient European demographic history). It is also clear from the mutation spectrum of rare variants that the TCC→TTC mutation rate decreased again around 1,000 to 2,000 years ago. In direct sequencing data from parent child trios, the rate of this mutation type is indistinguishable between Europeans and other populations, suggesting that the pulse of excess mutagenesis either subsided entirely or is confined to a small subpopulation that has not been included in any trio sequencing efforts.

A few similar mutation types, including TCT→TTT, ACC→ATC, and CCC→CTC, are enriched in Europeans with the same marginal allele frequency distribution as TCC→TTC, suggesting that these are minor components of the same mutational signature [83, 135]. However, other mutation types show differentiation between populations that becomes strictly stronger with decreasing allele frequency. The first principal component of human mutation spectrum variation is not defined by TCC→TTC mutation abundance at all, but instead separates Africans from Eurasians. African mutations are slightly more likely than Eurasian mutations to have G/C ancestral alleles compared to A/T ancestral alleles, a trend that is not very dependent on sequence context and does not dramatically affect the frequency distribution of any specific trinucleotide mutation type. In addition, certain complex mutations that appear linked at adjacent sites show evidence of even more population differentiation than simple SNPs do [167].

An apparent separate pulse of NAC→NCC mutations in the history of the Japanese population was also reported by Harris and Pritchard; however, this mutation type appears to be strongly influenced by a cell line artifact or other technical issue that caused context specific errors in 1000 Genomes variant calls [12]. An overall excess of NAC→NCC mutations in East Asians compared to Europeans has been confirmed in two independent datasets [83], but it is not clear how this relates to the presence of similar high-frequency artifactual mutations in Japanese genomes that were sequenced using an older technological platform.

At present, evidence of the causal mechanisms underlying germline mutation rate evolu-

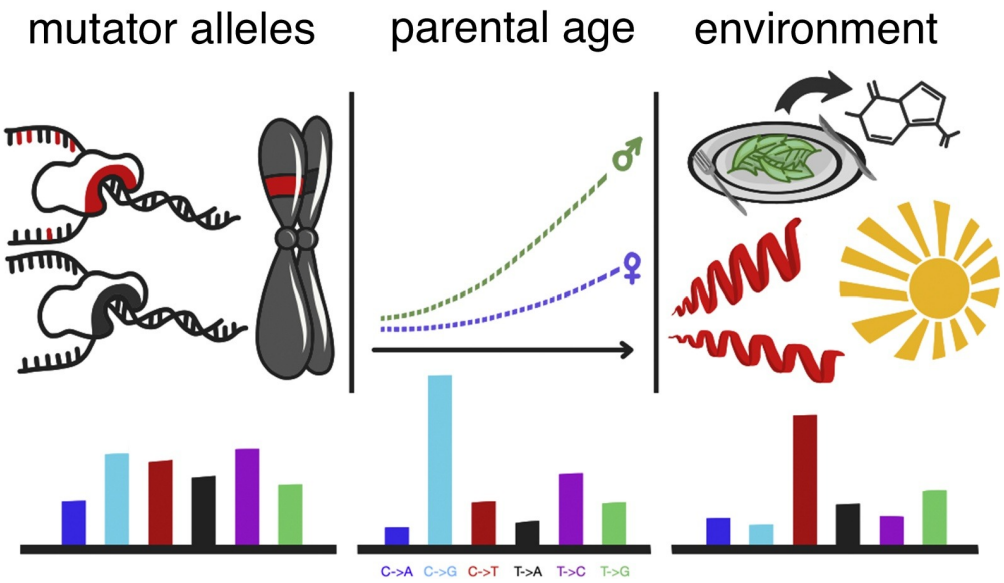

Figure 1.1: Three hypotheses have been proposed to explain why mutation spectra appear to evolve rapidly within populations: mutator alleles, life history trait evolution, and environmental mutagenesis. Firstly, mutator alleles are genetic variants in DNA polymerases, repair enzymes, or other genes that can raise or lower the mutation rate by altering DNA replication or repair fidelity in a heritable way. Secondly, changes in lifespan and reproductive age have the potential to change mutation spectra because of age-related changes in the rate and spectrum of mutagenesis in the male and female germlines. Finally, exposure to radiation, dietary mutagens, and toxic air particulates are all possible hazards of past and present existence on earth. These have the potential to modify mutation rate in a way that is dependent on time rather than genetic background. (Image copyright Natalie Telis)

tion in human populations is virtually nonexistent. In this review, we aim to summarize the recent proliferation of evidence for mutation spectrum variation at the levels of individuals, populations, and species, and explore the range of hypotheses about what may have caused these curious patterns of genetic variation.

1.2.1 The mutator allele hypothesis

DNA replication and repair are among the most essential of biological processes and are necessary for maintaining the fidelity of each species' genome from generation to generation. Sturtevant and Kimura posited that the mechanisms responsible for replication and repair, if mutated themselves, would alter the efficiency of these essential processes. We now know that there are dozens, if not hundreds, of genes involved in DNA replication and repair [34], most of which are highly conserved across eukaryotic organisms [207, 129]. Even so, the processes of replication and repair have evolved in each species' lineage, and the sequence and function of genes involved in highly conserved processes can diverge dramatically between species (an example is base excision repair [BER], a repair process that is operative in both yeast and mammals, but yeast BER pathways are far less efficient at repairing oxidative damage [104]). Inter-species variation in such genes is thus thought to be directly linked to inter-species variation in the mutation rate [204].

The inter-species variation in DNA replication and repair machinery and mutation rates leads us to question whether within-species variation of these essential genes can also explain within-species variation in mutation rates, both between individuals and looking back in time. Harris and Pritchard, who were the first to show evidence for a temporal pulse of elevated mutation rate at a specific triplet context in one population (TCC→TTC in Europeans), hypothesized that these dynamics are driven by the appearance and subsequent drift of mutator alleles that modify the germline mutation rate [83].

This hypothesis is bolstered by evidence that high mutation rates sometimes emerge spontaneously in experimentally evolving populations of microorganisms [121, 196, 75, 126]. A recent landmark experimental evolution study found that mutator alleles arose spontaneously in six of twelve experimental *Escherichia coli* populations that were maintained in stable, identical environments for over 60,000 generations [75]. In some cases, mutator alleles in these populations even reached fixation, and in others, the arrival of a mutator allele was

followed by the emergence of an "antimutator" allele that moderated the accelerated rate of mutation induced by the initial mutator [75, 215]. Importantly, mutator alleles are not exclusive to rapidly evolving bacteria or yeast. Mutator alleles are one of the hallmarks of human cancers [128], and viable strains of mice possessing germline mutator alleles have been bred in the lab (albeit with smaller litter sizes and lower rates of reproduction, pregnancy, and survival) [216].

The existence of mutator alleles in microorganisms, cancer cell populations, and lab-engineered mice has fueled tantalizing speculation that mutator alleles have been (and continue to be) a ubiquitous feature of every species' genome and evolutionary history. If mutator alleles are segregating nearly neutrally in human populations today, some will be predicted to fix over time. Some evidence for past mutator allele fixation can be found in great ape genomic diversity: closely related ape species like humans and chimps have more similar mutation spectra than distantly related species like humans and orangutans, despite the fact that chimps and orangutans inhabit much more similar environments today [72]. Proving that mutator alleles have arisen in and left a mark on the human genome, however, is an exceptionally challenging task. Since humans are a sexually-reproducing species, any mutator alleles arising in the genome will quickly become decoupled from any beneficial mutations they induce because of recombination, so, depending on the distribution of mutational fitness effects, most mutator alleles are expected to segregate neutrally or rapidly drop in frequency due to selection acting against the deleterious mutations that they create on an extended linked haplotype [107]. Consequently, the only hope of identifying mutator alleles from population sequencing data in sexually-reproducing organisms is to search for haplotypes carrying an excess of derived alleles and identify candidate mutator alleles on the same haplotype that have remained in linkage disequilibrium with the mutations they have induced [188].

If the TCC→TTC pulse was the result of a mutator allele, that mutator has likely been

lost from the population, so evidence of its genomic location and effect size may be lost to time. (Another possibility is that a mutator allele may remain in the population after attenuation by the arrival of an antimutator allele, which could explain why the TCC→TTC rate in present populations does not appear to have fallen back to its pre-pulse rate). However, it is possible that a TCC→TTC mutator still exists at extremely low frequency, or that other mutator alleles contribute to background levels of germline mutational signature differentiation. A mutator with a strong effect size could be identified by sequencing a carrier parent-child trio, and a mutator with a weaker effect size but high frequency could be identified by looking for a signal of elevated mutagenesis along an extended shared haplotype. However, attempting to detect this signal of spatially-clustered low-frequency variants may be extremely difficult because similar signals of genomic variation can be caused by sequencing artifacts [12], spatial variation in mutation rates across the genome [33, 205], natural selection [62], and demographic history [162, 30]. Milligan et al. develop an analytical and simulation-based account for how mutator loci could modulate mutation rate dynamics in parameter regimes relevant to the human TCC→TTC pulse, adding quantitative texture to the possible role of mutator alleles in shaping mutation spectrum histories [140].

1.2.2 The generation time hypothesis

A second line of reasoning proposes that temporal shifts in the mutation spectrum evident from population sequencing data might reflect changes in life history traits over human history. There is abundant archaeological [71], anthropological [89], and genetic [159] evidence that life history traits have changed over the course of human history, and fundamental principles of biology dictate that life history traits—specifically generation time—are directly related to changes in the germline mutation rate [141]. In species that are genetically predisposed to have very short generation times and lifespans, this generation time hardwiring may relax selection against mutators due to reduced opportunity for deleterious mutations

to occur over a shorter reproductive lifespan. This effect has been observed in the mitochondrial DNA of short-lived African killifish, which has the shortest known lifespan of any vertebrate [35].

The relationship between generation time and per-generation mutation rate in humans has been proven conclusively in sequencing studies of parent-offspring trios and larger pedigrees: the number of *de novo* point mutations increases roughly linearly by approximately 1 additional mutation per year of the father's age and 1 additional mutation per 3-4 years of the mother's age [222, 168, 74, 101, 100, 66, 179]. Recent evidence also suggests that the maternal mutation rate may actually increase exponentially with age rather than linearly [66]. Importantly, these effects of parental age are not universally true for all families: analyses of sequencing data from families with multiple children have found that paternal age effects can vary by an order of magnitude between families, indicating the mechanisms underlying this effect may themselves be under genetic and/or environmental control [168, 179].

Humans have longer reproductive life cycles than their closest great ape relatives, and one recent study calculated that this delayed onset of puberty and parenthood was sufficient to explain the difference in mutation rate between human and owl monkey pedigrees [212]. However, a study of great ape pedigrees found that chimpanzees, orangutans, and gorillas accumulate about 50% more mutations per year of paternal age than humans do, [24] suggesting that the dependence of mutation rate on life history has been evolving in concert with life history itself.

Not only have studies found that the mutation rate increases with parental age, but also that the mutation spectrum changes. As fathers age, the rate of CpG→TpG mutations increases more rapidly than other types, and as mothers age, C→G mutations increase at a faster rate [101, 66] and accumulate preferentially in specific hotspots [222, 101, 73], but to a lesser extent mark the entirety of the genome and inject a generation time signature into the mutation spectrum [189]. Hypothetically, if a population sustains older generation

times over a period of many generations, the spectrum of variation for mutations that arose during that period could look quite different from that of mutations that arose in previous generations [142]. In a recent study of Neanderthal introgression in the Icelandic population, Skov et al. [194] carried out a similar analysis and found that introgressed loci carried more C→G and fewer T→C and CpG→TpG variants compared to the non-introgressed regions of the genome. Given that these mutation types track maternal and paternal age, the authors concluded that, relative to humans, Neanderthals that interbred with humans may have had older mothers and younger fathers on average. Intriguingly, TCC→TTC is a mutation type that is significantly associated with maternal origin over paternal origin, though it is not clear how this association is related to the enrichment of TCC→TTC mutations in Europe and South Asia [4]. One study used allele age estimation to conclude that most apparent mutation spectrum evolution in humans is due to historical changes in generation time [220].

Although generation time variation can leave a predictable imprint on the mutation spectrum, this is unlikely to be a major driving cause of the historic mutation spectrum differences that exist between populations. Assuming that generation time has the same effect on mutagenesis in every genetic background, it can only explain a single dimension of mutation spectrum drift and will tend to explain variation within populations better than variation between populations. A caveat is the existence of hints that generation time has different effects on mutation rate in different genetic backgrounds, which might imply that it is capable of explaining more than one dimension of mutation spectrum variation [179]. In any case, exploring this dimension of variation could prove highly rewarding—if also highly challenging—because of its multiplicity of links to cultural and environmental changes as well as genetic determinants of the ages of puberty and menopause. Even if generation time alone cannot perfectly explain mutation spectrum drift over time, principled estimation of the generation time is a major source of uncertainty for demographic inference [141], so using mutation spectra to estimate generation time distributions could increase the accuracy of

demographic inference, provided the mutation spectra are not hopelessly confounded with biased gene conversion [166].

1.2.3 The environmental mutagen hypothesis

A third possibility is that the temporal variation in human mutation rate and spectra might be explained by some form of environmental exposure. We have long known that certain environmental mutagens can leave distinctive mutation signatures on the genome (e.g., mutagens in tobacco smoke are associated with G→T transversion mutations [160]), and in a recent study of the mutagenic effects of 79 environmental agents applied to somatic cell populations, over half were found to confer elevated mutation rates, often with distinctive spectra [117]. The mutation rate in healthy somatic tissues is highly variable, with the highest rates observed in tissues with greater exposure to environmental mutagens, including the skin, lung, blood and esophagus [221, 68]. Fewer studies have investigated the effects of environmental mutagens on mutation rates specifically in the germline, but those that have have demonstrated that many of the environmental mutagens known to affect somatic cells have similar effects in the germline. Mouse models have successfully determined that parental exposure to benzo(a)pyrene [17] and ionizing radiation [2, 180] lead to elevated mutation rates in offspring. In human trio sequencing studies, it has been shown that offspring of fathers exposed to ionizing radiation also carry an increase in multi-site *de novo* mutations [90] and offspring of fathers exposed to dioxin are enriched for A→T mutations [214]. Another study that compared *de novo* mutation patterns ascertained in trios from diverse populations even found a significant reduction in A→G mutations in the offspring of Amish parents and suggested this might be due to the unique environmental conditions of this population [105].

Regarding our focal example of the historic TCC→TTC pulse in European populations, we also know that the environment of the European continent was vastly different 15,000

years ago—the global climate was rapidly warming after the Last Glacial Maximum and during this period, massive flora and fauna extinction events took place, potentially due to direct and indirect activity of expanding hominin populations [195] and/or climate change [29]. There is also evidence of large-scale burning of biomass that occurred sometime during this period, possibly the result of a disintegrating comet that may have struck the Earth approximately 12,800 years ago [161, 63]. These environmental disruptions could potentially have impacted the genomes of archaic human populations in multiple ways: particulate matter from burning of biomass has recently been linked to various forms of DNA damage both *in vivo* and *in vitro* [149, 39], and extinction of primary food sources might have dramatically altered the diets of human populations, thus exposing them to new mutagens that humans had not yet evolved to avoid or metabolize [227].

During the late Pleistocene (between 11-19kya), human populations in Europe also acquired adaptive mutations in multiple genes that conferred lighter skin pigmentation, which rapidly swept to high frequency throughout most of Europe [21]. In the initial discovery of the TCC→TTC mutation pulse in Europeans, Harris hypothesized that this novel skin pigmentation phenotype might be indirectly responsible for elevated mutation rates, reasoning that light skin is much more sensitive to ultraviolet radiation, and UV radiation is known to degrade folate, which in turn is linked to endogenous mechanisms of DNA damage [82]. (Harris also rejected the possibility that direct exposure to UV radiation explains the TCC→TTC pulse, because UV radiation generates an excess of pyrimidine dimers leading to CC→TT tandem mutations—no studies that have replicated the TCC→TTC pulse finding have detected that this was accompanied by an excess of CC→TT mutations in the germline [83, 135, 82]). However, excess TCC→TTC mutations cannot be an obligatory consequence of adaptation to higher latitudes, since it is not seen in light-skinned East Asian populations. Based on evidence that the mutational signature of the TCC→TTC pulse is similar to that of cancer genomes that were treated with chemotherapy drugs which included alkylating

agents (which induce guanine methylation and subsequent G→A mutations) [135], Mathieson and Reich speculated that the ultimate environmental cause was not radiation-related but instead traceable to some unknown mutagenic alkylating agent that may have once been present in the human diet.

Another study found that a transmissible venereal tumor found in canines (CTVT) had a distinctive GTCCA→GTTCA pentamer mutation signature that was active roughly 1.9-8.5kya [13]. The fact that this pentamer is one of 16 possible NTCCN→NTTCN subtypes comprising TCC→TTC mutations, along with the closely-matched timing of the TCC→TTC pulse as estimated in humans [83], presents the rather bizarre possibility that a shared environmental mutagen might have played a role in both the human germline mutation process and the somatic mutation process in CTVT. Supplementary data from a manuscript by Aikens et al. shows that the GTCCA→GTTCA pentamer subtype that dominates the CTVT signature was indeed enriched in Europeans during the period of the TCC→TTC pulse in the human germline [6]. However, relative to other NTCCN→NTTCN subtypes, the p-value for enrichment of the GTCCA→GTTCA subtype in Europeans is only ranked 13 out of the 16 possible TCC-related pentamer mutations. This suggests that the dog tumor and the human TCC→TTC pulse may not be affected by the same mutational signature after all, despite their rapid appearance around the same time, unless the root cause was a shared mutagen whose sequence specificity exhibited subtle dependence on the genetic background of the cells being mutated.

1.2.4 mushi

As populations boom and bust, the accumulation of genetic diversity is modulated, encoding histories of living populations in present-day variation. Many methods exist to decode these histories, and all must make strong model assumptions. It is typical to assume that mutations accumulate uniformly across the genome at a constant rate that does not vary

between closely related populations. However, recent work shows that mutational processes in human and great ape populations vary across genomic regions and evolve over time. This perturbs the *mutation spectrum*: the relative mutation rates in different local nucleotide contexts. In Chapter 2, I develop theoretical tools in the framework of Kingman's coalescent to accommodate mutation spectrum dynamics. I present `mushi`, a method to perform nonparametric inference of demographic and mutation spectrum histories from allele frequency data. I use `mushi` to reconstruct trajectories of effective population size and mutation spectrum divergence between human populations, identify mutation signatures and their dynamics in different human populations, and calibrate the timing of a previously-reported mutational pulse in the ancestors of Europeans. I show that mutation spectrum histories can be placed in a well-studied theoretical setting and rigorously inferred from genomic variation data, like other features of evolutionary history.

1.2.5 `mutyper`

Characterization of germline mutation spectrum variation from population genomics data has shed light on the biological complexity of the mutation process, and its evolution within and between species. This analysis augments available population SNP data with estimates of local ancestral genomic context to assign mutation types and aggregate summary statistics thereof, and is increasingly common. There is a need for standardized computational tools to extract mutation spectrum information from sequencing data. In Chapter 3 I describe `mutyper`, a command-line utility and Python package that uses an ancestral genome estimate to assign mutation types to SNP data, compute mutation spectra for individuals, and compute sample frequency spectra resolved by mutation type for population genetic inference.

1.3 Evolutionary dynamics in adaptive immune systems

To defend against rapidly evolving pathogens, jawed vertebrates have specialized cells that evolve during each individual's lifetime. Lymphocytes (T cells and B cells) mount adaptive immune responses to protean pathogenic threats that would outpace defenses encoded in the host germline. They detect foreign antigens by maintaining a massive diversity of randomized genes that encode antigen-binding receptors, and mechanisms to proliferate and modify these receptors upon pathogen recognition. Immunity is essential to our survival; prediction and control of this complex adaptive system would have far-reaching impact on human health.

An immune receptor repertoire is a decentralized adaptive information processing system: it senses an unpredictable space of molecular targets, stores memories of sensory events, and encodes each individual's personal history of pathogen exposures in the statistics of receptor sequences within their repertoire. Underpinning these capabilities is the ability of immune receptors to undergo somatic evolution in step with evolving pathogens. This rapid evolutionary capability is itself an immune strategy shaped by evolutionary forces acting over deeper time within populations and species.

Complex systems researchers have long recognized that adaptive immune systems are a natural model system in which to study decentralized information processing [156, 157, 88]. However, accessible technologies to read the DNA sequences of receptors in real immune repertoires emerged and matured only over the last decade [193, 127]. Immune repertoire sequencing can read millions of receptor sequences from a single repertoire, and is revolutionizing translational and basic immunology [173, 91, 226]. This technology has birthed a new kind of quantitative immunology centered on the statistics of DNA and protein sequences of receptors sampled from individuals and populations [10, 77]. There is now a need to synthesize this new quantitative immunology with powerful models of biological sequence evolution that are used to understand the evolutionary forces that shape populations and species.

1.3.1 Germinal centers: evolutionary crucibles for antibody affinity maturation

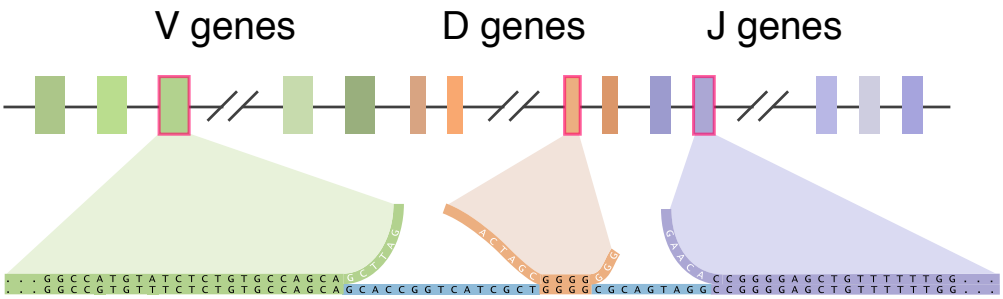

Figure 1.2: In VDJ recombination, germline-encoded V (green), D (orange), and J (violet) gene segments are randomly chosen, their ends resected a random amount, and joined together with random nontemplated nucleotide insertions (blue). Figure adapted from Murugan et al. [143]

The memories stored in repertoires are emergent features of somatic receptor evolution that occurs rapidly in response to each pathogen exposure. The parameters governing receptor sequence evolution are encoded in the germline: proliferation, mutation, and selection are carried out by complex cell signaling and enzymatic activity, and are presumably adapted for effective immune function. B cells—lymphocytes that make antibodies—bind antigen with the B cell receptor (BCR, the membrane-bound form of an antibody). In developing B cells, BCRs arise via *VDJ recombination*, a random DNA recombination process that can produce a vast number of possible receptor sequences (Figure 1.2). Upon encountering antigen, a *naive* B cell undergoes a Darwinian process of selection for improved antigen binding called *affinity maturation*. In microanatomical structures called *germinal centers*, they proliferate while actively mutating the BCR and locally competing against hundreds of other cells to bind antigen molecules [138] (Figure 1.3).

Germinal centers are complex structures that serve as evolutionary crucibles orchestrating B cell diversification, competition, and eventual memory development [192]. However, the details of how this cellular competition is orchestrated as a population process remain

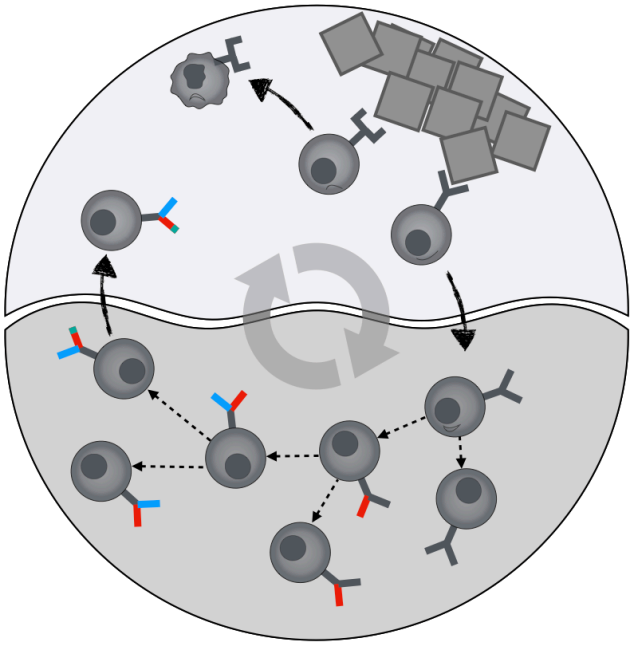

Figure 1.3: Germinal center B cells compete for antigen binding in the light zone (top) and proliferate with mutation in the dark zone (bottom).

an open question, and recent experimental results raise questions about the mode of evolutionary dynamics in germinal centers. Although classically understood by immunologists to impose selection for increased antigen binding affinity, recent studies are conflicted regarding germinal center evolution as adaptive toward high affinity [206, 118, 144]. Tas et al., using fate mapping and single germinal center B cell isolation, found that two germinal centers seeded with the same BCR and responding to the same antigen showed inconsistent outcomes for high-affinity variants, suggesting strong stochasticity [206]. Kuraoka et al. developed a single B cell culture system to show that germinal centers tolerate a wide range of BCR affinities, and that substantial fractions of germinal center B cells do not detectably bind antigen [118]. Murugan et al., studying repeated immunization with complex antigen, reported that most surviving mutations during affinity maturation were affinity-neutral [144].

These observations raise questions that will be at the center of my research directions. Why has nature evolved germinal centers—intricate evolutionary engines for drafting and revising BCRs—only to apply lax selection? What are the evolutionary parameters and molecular phenotypes governing germinal center evolution? Can we understand the strong stochasticity in germinal center evolution as an immune strategy that maintains robustness to pathogen evolution?

From a theoretical perspective, sequence evolution in germinal centers has a number of exotic features that challenge stochastic models of population genetic processes. Cells cycle between the light zone where selection acts, and the dark zone where they proliferate and mutate neutrally. This resembles population genetic models of islands exchanging migrants, but has the unusual property that evolutionary forces themselves are partitioned between the islands (selection and death on one, birth and mutation on the other), and have sequential dependence. The mutation process only occurs in the dark zone on cells that have previously been selected and migrated from the light zone. Mutations are focused on *mutational hotspots* defined by nucleotide context. As a lineage mutates (in a tree-like exploration of sequence space, starting from the naive receptor at the root), these hotspots are modified and ablated, so mutation slows down [217].

1.3.2 `GCtree`

Modern biological techniques enable very dense genetic sampling of unfolding evolutionary histories, and thus frequently sample some genotypes multiple times. This motivates strategies to incorporate genotype abundance information in phylogenetic inference. In Chapter 4, I synthesize a stochastic process model with standard sequence-based phylogenetic optimality, and show that tree estimation is substantially improved by doing so. Our method is validated with extensive simulations and an experimental single-cell lineage tracing study of germinal center B cell receptor affinity maturation.

Chapter 2

NONPARAMETRIC COALESCENT INFERENCE OF MUTATION SPECTRUM HISTORY AND DEMOGRAPHY

> the thing I came for:
> the wreck and not the story of the wreck
> the thing itself and not the myth
>
> Adrienne Rich, *Diving into the Wreck*

Over the past decade, population geneticists have developed many sophisticated methods for inferring population demography, and have consistently found that simple isolated populations of constant size are far from the norm (reviewed in [165, 185, 19]). Population expansions and founder events, as well as migration between species and geographic regions, have been inferred from virtually all high resolution genetic datasets. We now recognize that inferring these non-equilibrium demographies is often essential for understanding histories of adaptation and global migration. Population genetics has uncovered many features of human history that were once virtually unknowable by other means, revealing a complex series of migrations, population replacements, and admixture networks among human groups and extinct hominoids.

Although demographic inference methods can model complex population histories, the germline mutation process that creates variation has long received a comparatively simple treatment. A single parameter, μ, is used to represent the mutation rate per generation at all loci, in all individuals, and at all times. In humans, μ is estimated from parent-child trio sequencing studies, and modest variation in μ can have major effects on the interpretation of inferred parameters such as times of admixture and population divergence. In other

organisms, for which trio sequence data is usually unavailable, μ is estimated from sequence divergence between species with a fossil-calibrated divergence time, and these estimates come with still higher uncertainty.

A growing body of evidence indicates that simple, constant mutation rate models may not adequately describe how variation accumulates on either inter- or intraspecific timescales [76, 183, 82, 83]. Germline mutation rates appear to have evolved during the speciation of great apes and the divergence of modern human populations (reviewed in [31]). Much of this evolution might be caused by nearly neutral drift [130], but a contributing factor could be selection on traits like life history and chromatin structure that indirectly affect mutation accumulation. Because mutation is intimately tied to the basic housekeeping process of cell division, gamete production, and embryonic development, the accumulation of mutations is likely to be complexly coupled to other biological processes [187, 168, 67].

It is difficult to disentangle past changes in mutation rate from past changes in effective population size, which modulate levels of polymorphism even when mutation rate stays constant. However, evolution of the mutation process can be indirectly detected by measuring its effects on the *mutation spectrum*: the relative mutation rates among different local nucleotide contexts. Hwang and Green [95] modeled the triplet context-dependence of the substitution process in a mammalian phylogeny, finding varying contributions from replication errors, cytosine deamination, and biased gene conversion and showing that the relative rates of these processes varied between different mammalian lineages. Many cancers also exhibit somatic hypermutability of certain triplet motifs due to different DNA damage agents and failure points in the DNA repair process [8, 87]. Harris and Pritchard [82, 83] examined the variation of triplet spectra between closely related populations, counting single nucleotide variants in each triplet mutation type as a proxy for mutational input. They found that human triplet spectra distinctly cluster by continental ancestry group, and that historical pulses in mutation activity influence the distribution of allele frequencies in certain

mutation types. The divergence of mutation spectra among human continental groups has been replicated in independently generated datasets [135, 83], and similar patterns have been observed in other species, including great apes [72], mice [48], and yeast [99]. Some of the mutation spectrum divergence between mice and yeast lineages has been mapped to mutator alleles [99, 178].

Emerging from the literature is a picture of a mutation process evolving within and between populations, anchored to genomic features and accented by spectra of local nucleotide context. If probabilistic models of population genetic processes are to keep pace with these empirical findings, mutation deserves a richer treatment in state-of-the-art inference tools. In this chapter, I build on classical theoretical tools to introduce fast nonparametric inference of population-level *mutation spectrum history* (MuSH)—the relative mutation rate in different local nucleotide contexts across time—alongside inference of demographic history. Whereas previous work has uncovered mutation spectrum evolution using summary statistics of standing variation, we shift perspective to focus on inference of the MuSH, which we model on the same footing as demography.

Demographic inference requires us to invert the map that takes population history to the patterns of genetic diversity observable today. This task is often simplified by first compressing these genetic diversity data into a summary statistic such as the *sample frequency spectrum* (SFS), the distribution of derived allele frequencies among sampled haplotypes. The SFS is a well-studied population genetic summary statistic that is sensitive to demographic history. Inverting the map from demographic history to SFS is a notoriously ill-posed problem, in that many different population histories can have identical expected SFS [145, 25, 210, 14, 175]. One way to deal with the ill-posedness of demographic inference is to specify a parametric model of population size change, usually piecewise-linear or piecewise-exponential. An alternative, which generalizes to other inverse problems, is to allow a more general space of solutions, but to *regularize* by penalizing histories that contain biologically

unrealistic features (e.g. high frequency population size oscillations). Both approaches shrink the set of feasible solutions to the inverse problem so that it becomes well-posed, and can be thought of as leveraging prior knowledge. In particular, regularization constrains the population size from changing on arbitrarily small timescales, since significant population size change usually takes at least a few generations.

In this chapter, I extend a coalescent framework for demographic inference to accommodate inference of the MuSH from a SFS that is resolved into different local k-mer nucleotide contexts. This is a richer summary statistic that we call the k-SFS, where e.g. $k = 3$ means triplet context. We show using coalescent theory that the k-SFS is related to the MuSH by a linear transformation, while depending non-linearly on the demographic history. We infer both demographic history and MuSH by optimizing a cost that balances a data fitting term, using the forward map from coalescent theory, along with regularization terms that favor solutions with low complexity. Our open-source software `mushi` (mutation spectrum history inference) is available at `https://harrispopgen.github.io/mushi` as a Python package with extensive documentation. Using default settings and modest hardware, `mushi` takes only a few seconds to infer histories from population-scale sample frequency data.

The recovered MuSH is a rich object that illuminates new dimensions of population history, and addresses biological questions about evolution of the mutation process. After validating with data simulated under known histories, we use `mushi` to independently infer histories for each of the 26 populations (from 5 super-populations defined by continental ancestry) from the 1000 Genomes Project (1KG) [1], using recent high-coverage sequencing data. We demonstrate that `mushi` is a powerful tool for demographic inference that has several advantages over existing demographic inference methods, then go on to describe the newly illuminated features of human MuSH.

We recover demographic features that are robust to regularization parameter choices, including the out-of-Africa event (OOA) and the more recent bottleneck in the ancestors of

modern Finns, and we find that effective population sizes converge ancestrally within each super-population, despite being inferred independently. Decomposing human MuSH into mutation signatures varying through time in each population, we see global divergence in the mutation process that impacts many mutation types, and reflects population and super-population relatedness. Finally, we revisit the timing of a previously reported ancient pulse of elevated TCC→TTC mutation rate, active primarily in the ancestors of Europeans, and absent in East Asians [82, 83, 199, 198]. We find that the extent of the pulse into the ancient past is sensitive to the choice of demographic history model, but that all demographic models that fit the k-SFS yield a pulse timing that is significantly older than previously thought, seemingly arising near the divergence time of East Asians and Europeans.

With `mushi` we can quickly reconstruct demographic history and MuSH without strong model specification requirements. This adds a new approach to the toolbox for researchers interested only in demographic inference. For researchers studying the mutation spectrum, demographic history is necessary for time calibration of events in mutation history, so we expect that jointly modeling demography and MuSH will be important for studying mutational spectrum evolution in population genetics.

2.1 Model Summary

2.1.1 Augmenting the SFS with nucleotide context information

The sample frequency spectrum (SFS) is a summary statistic of population variation that counts variants partitioned by the number of sampled individuals who carry the derived allele. Since rare variants tend to be younger than common variants, this summary preserves considerable information about the distribution of allele age, which can reflect temporal variation in either the mutation rate or the intensity of genetic drift. To disentangle these two causal factors, we leverage the fact that genetic drift affects all mutations uniformly, whereas mutation rate changes may depend on genomic sequence context.

By default, we classify mutation types by their derived allele and ancestral k-mer nucleotide context, with k odd and the variant centered. There are $K = 2 \times 3 \times 4^{k-1}$ mutation types after collapsing by strand symmetry; e.g. considering C>T mutations and their complementary G>A mutations to be identical. When $k = 3$ there are $K = 96$ triplet mutation types, of which TCC→TTC is one. For a sample of n genomes, the standard SFS is an $(n-1)$-dimensional vector of the number of variants present in exactly i genomes, with i ranging from 1 to $n-1$. The k-SFS is an $(n-1) \times K$-dimensional matrix, where the (i, j)-th entry is the number of variants of mutation type j that are present in exactly i individuals.

Our goal is to sequentially infer demographic history and then MuSH by inferring histories that optimize a composite likelihood of an observed k-SFS data matrix $\mathbf{X}$. This requires computing $\boldsymbol{\Xi} \equiv \mathbb{E}[\mathbf{X}]$, the expected k-SFS as a function of effective population size and context-dependent mutation rate over time. Our main theoretical result, Theorem 1 in Materials and Methods, shows that $\boldsymbol{\Xi}$ is a linear functional of the K-element MuSH $\boldsymbol{\mu}(t)$ given the haploid effective population size history $\eta(t)$ (where $\eta(t) = 2N(t)$ for diploid populations): $\boldsymbol{\Xi} = \mathcal{L}(\eta)\boldsymbol{\mu}^{\intercal}$. Figure 2.1 sketches the generation of a sampled k-SFS matrix $\mathbf{X}$ in a toy setting of $n = 4$ sampled haplotypes, 3 mutation types, and a fixed genealogy.

The linear operator $\mathcal{L}(\eta)$ transforms the unknown MuSH into a matrix of observed allele frequencies across mutation types.

2.1.2 Regularization to select parsimonious population histories

Demographic inference—the recovery of population size history $\eta(t)$ from SFS summary data—is complicated by the fact that different population size histories can have identical expected sample frequency spectra. This non-identifiability problem has been extensively explored in the literature [145, 25, 210, 14, 175]. Although many different population size histories can optimally fit a SFS, it has been proven that uniquely good (identifiable) fits are available when excluding biologically unrealistic histories that contain rapid oscillations.

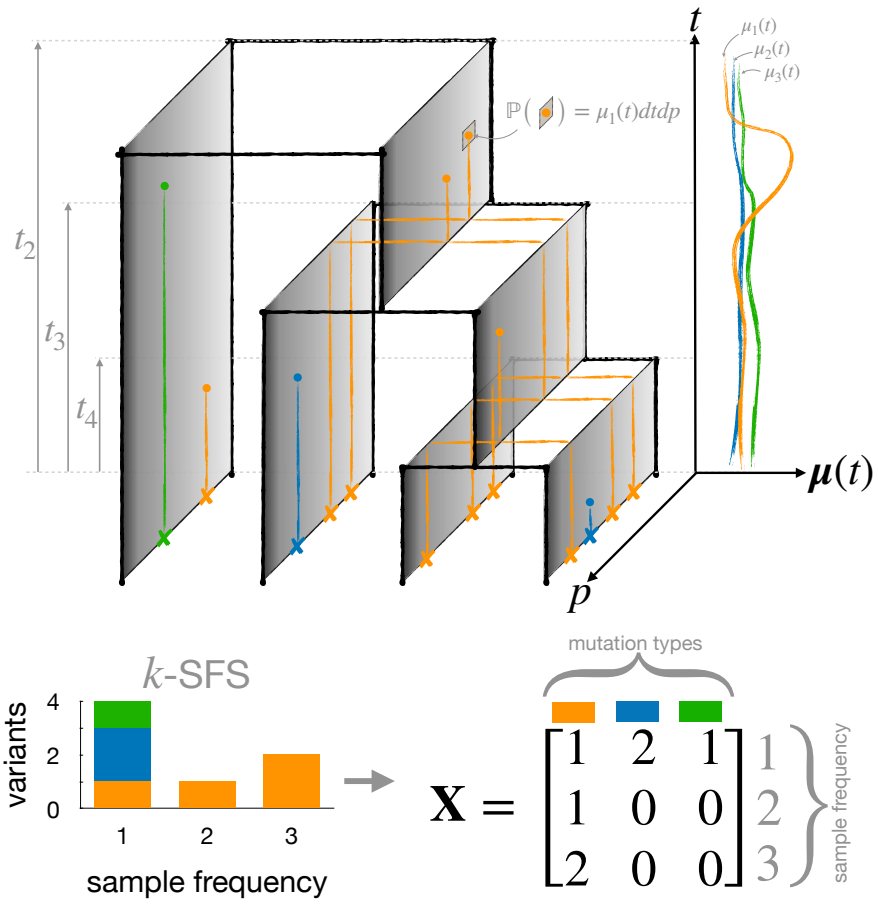

Figure 2.1: Mutation spectrum history and demography are encoded in the k-SFS as joint inverse problems. A schematic of a marked Poisson process with $n = 4$ sampled haplotypes is conditioned on coalescent times t_4, t_3, t_2. This simple example mutation spectrum history $\boldsymbol{\mu}(t) = [\mu_1(t)\ \mu_2(t)\ \mu_3(t)]^\intercal$ shows just three mutation types, marked by different colors to designate different sequence contexts or distinct mechanistic origins. Dots indicate mutation events placed by time t, genomic position p, and coalescent line (which are depicted as extruded in the genomic coordinate axis, grey sheets). The probability that a mutation of type i occurs in a differential time interval dt and genomic interval dp on a given line is proportional to the instantaneous mutation rate $\mu_i(t)$. The crosses on the sampled haplotypes indicate segregating variants of each mutation type. The sampled k-SFS data is shown as a stacked histogram (top right), and in matrix form (bottom right).

Here, we introduce a mathematical framework that expresses histories nonparametrically (approximating infinite dimensional functions), but prefers sparse solutions that consist of simple pieces, and disfavors histories that fit the k-SFS equally well with more erratic features.

Inference of the MuSH introduces a second identifiability problem of a different nature. The effective population size $\eta(t)$ and the mutation rate $\mu(t)$ are mutually non-identifiable for all t, meaning that the expected SFS $\boldsymbol{\xi}$ is invariant under a modification of $\eta(t)$ so long as a compensatory modification is made in $\mu(t)$. The non-identifiability of η and μ can be understood intuitively by considering two histories that can be tuned to have the same expected SFS: one where the mutation rate increases over an interval of time in the past, while the effective population size stays constant, and the other with a constant mutation rate where population size increases, dilating coalescence times on the same branches affected in the first scenario. "‘

While the total mutation rate is confounded with demography, the *composition* of the mutation spectrum—the relative mutation rate of each mutation type—reveals itself in the k-SFS. This can also be understood intuitively: an excess of variants of a given frequency in only a single mutation type (one column of the k-SFS) cannot be explained by an historical population boom, because all mutation types are associated to the same demographic history. In this case, we would infer a period of increased relative mutation rate for this mutation type. We cannot discern changes in total mutation rate, so `mushi` assumes a constant total rate μ_0 and time variation in the rate of drift is modeled only in $\eta(t)$. We handle this constraint using a transformation technique from the field of compositional data analysis (Materials and Methods).

Even with this compositional constraint on the total mutation rate, many different population histories may be equally consistent with an empirical k-SFS. As mentioned before, we overcome this with regularization methods to select simple demographies and MuSHs.

We penalize the model for three different types of irregularity. One penalty is motivated by the demographic inference literature: histories that feature rapid oscillations over time are disallowed in favor of similarly likely histories that change less rapidly and less often, and are composed of a few simple (low-order polynomial) trends. The second penalty favors models in which the mutation spectrum history $\boldsymbol{\mu}(t)$ is composed of a few mutation signatures that vary in their intensity over time for each mutation type, and represent a sparse vocabulary of mutagenic processes. The third regularization penalty is a classical ridge (or Tikhonov) penalty, which speeds up convergence of the optimization without significantly affecting the solution. Detailed formulation of our optimization problems and regularization strategies are in Materials and Methods.

The strength of all three regularizations can be tuned by changing the values of user-specified hyperparameters. Stronger regularization yields simpler histories, but eventually this will result in a poor fit to k-SFS data. Users should tune the regularization parameters to select histories that are simple while still fitting well, perhaps considering prior knowledge about the natural history of their study population. This process is designed to be flexible, and more straightforward than specifying an explicit parametric model.

2.1.3 Quantifying goodness of fit to the data

The probability distribution of an empirical SFS given an expected SFS is often specified using a Poisson random field (PRF) approximation [181], which stipulates that, neglecting linkage, the observed number of sites with derived allele count i is Poisson-distributed around the expected number of sites of this frequency. This PRF approximation is easily generalizable to k-SFS data. Recall that $\mathbf{X}$ is the observed k-SFS matrix, so the SFS is $\mathbf{x} \equiv \mathbf{X1}$ (row sums over mutation types). In Materials and Methods we show that the generalized PRF factorizes as $\mathbb{P}(\mathbf{X} \mid \eta, \boldsymbol{\mu}) = \mathbb{P}(\mathbf{x} \mid \eta)\ \mathbb{P}(\mathbf{X} \mid \mathbf{x}, \eta, \boldsymbol{\mu})$, with the first factor given by a Poisson distribution and the second by a multinomial distribution. We also show that the SFS $\mathbf{x}$ is

a sufficient statistic for the demographic history η with respect to the k-SFS $\mathbf{X}$. This means that estimation of η can be done by fitting the total SFS, which maximizes the first factor as a likelihood for η. Then the MuSH can be estimated by fitting the k-SFS, maximizing the second factor as a likelihood for $\boldsymbol{\mu}$, conditioned on the η estimate.

2.2 Results

2.2.1 Reconstructing simulated histories

We investigated `mushi`'s ability to recover histories in simulations where known histories are used to generate k-SFS data. Instead of simulating under the `mushi` forward model itself, we used `msprime` [103] to simulate a *tree sequence* describing the genealogy for 200 haplotypes of human chromosome 1 across all loci. This is a more difficult test, as it introduces linkage disequilibrium that violates our model assumptions.

We used the human chromosome 1 model implemented in the `stdpopsim` package [3], which includes a realistic recombination map [96]. We used a difficult demography consisting of a series of exponential crashes and expansions, variously referred to as the "sawtooth", "oscillating", or "zigzag" history. This pathological history has been widely used to evaluate demographic inference methods [184, 209, 211, 199], and is available in the `stdpopsim` package as the `Zigzag_1S14` model. Our simulated tree sequence contained about 250 thousand marginal trees.

We defined a MuSH with 96 mutation types, two of which are dynamic: one undergoing a pulse, and the other a monotonic increase. Since estimates of mutation spectra in real data are often confounded by misidentification of some ancestral alleles as derived, we modeled an ancestral misidentification rate of 0.01, with the two dynamic mutation types as misidentification partners. The total mutation rate varies slightly over time due to these two components—introducing another model misspecification, since inference assumes a constant total mutation rate. We placed mutations on the simulated tree sequence according to the

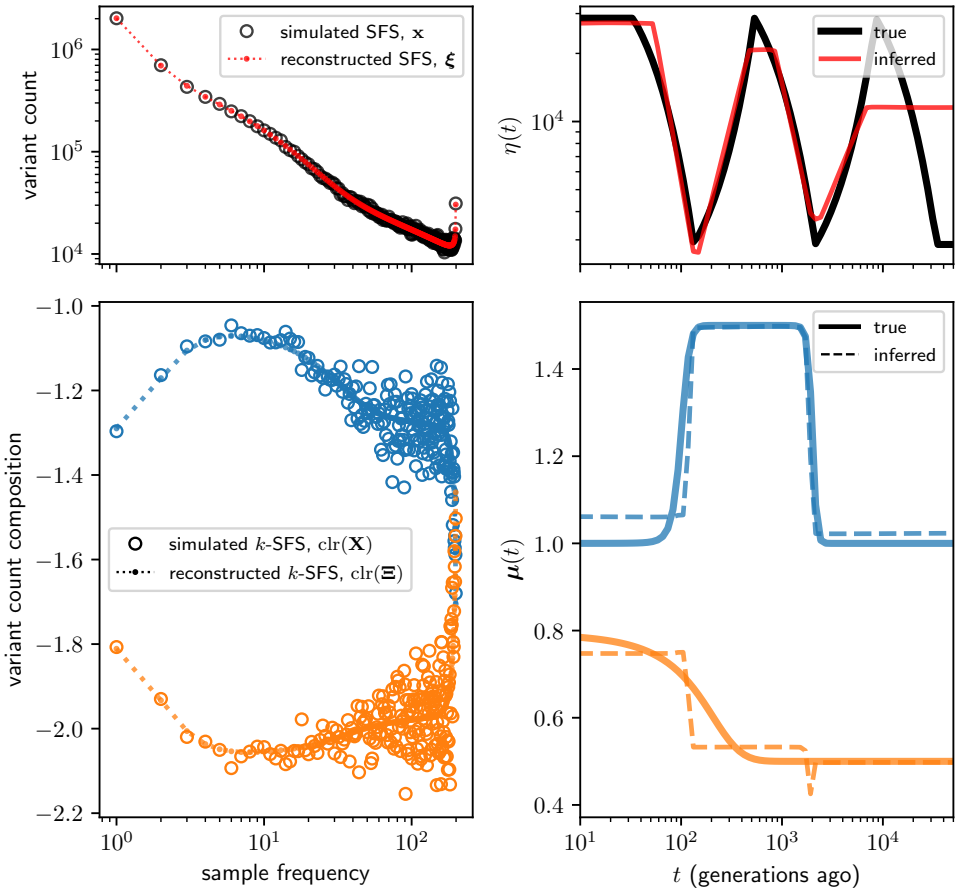

Figure 2.2: Simulation study of `mushi` performance. The sawtooth demography (top right) and a MuSH with 96 mutation types (bottom right, with two non-constant components shown) were used to simulate 3-SFS data for $n = 200$ sampled haplotypes. The MuSH has a total mutation rate of about $\mu_0 = 95.8$, generating about 9.5 million segregating variants. The top left panel shows the SFS, and the bottom left shows the two variable components of the k-SFS as a composition among mutation types at each allele frequency. Time was discretized with a logarithmic grid of 100 points. Inference was performed using a mixture of order 0 and order 1 trends for both demography, and 0 order trend for the MuSH (see Materials and Methods, and Figures B.1-B.2).

historical relative rate function for each mutation type, and computed the k-SFS.

Figure 2.2 depicts inference results for this simulation. We find that `mushi` accurately recovers the difficult sawtooth demography for most of its history, but over-smooths beyond the third population bottleneck because little information about this time survives in the SFS. The MuSH is accurately reconstructed as well, with both the pulse and ramp signatures recovered. The timing of the features in the MuSH also appears accurate, despite demographic misspecification that has the potential to distort the diffusion timescale. In Figures B.1 and B.2 we explore various hyperparameter choices and how they impact inferences of the demographic history, and of the different trends in the two variable mutation types. We find that demographic model selection does not significantly impact MuSH inference, and that different MuSH inference penalties hyperparameters recover the two distinct components of the MuSH with varying fidelity. The folded SFS—which uses minor allele frequencies instead of attempting ancestral state polarization—can also be used for demographic inference in `mushi`, and in Figure B.3 we show inference results are similar.

One noteworthy feature of our fit to the sawtooth demography is the tendency of `mushi` to smooth older demographic oscillations without smoothing younger oscillations as aggressively. In contrast to methods such as the pairwise sequential Markov coalescent (PSMC) [125] that tend to infer runaway population sizes in the ancient past, `mushi`'s history flattens in the limit of the ancient past. The same constraint underlies both PSMC's ancient oscillations and `mushi`'s ancient flattening: genomic data sampled from modern individuals cannot contain information about history older than the time to most recent common ancestor (TMRCA) of the sample, since mutations that occurred before then will be fixed, rather than segregating, in the sample. For example, we expect that population bottlenecks erase information about history, since they accelerate the fixation of variant sites that predate the bottleneck. While this information loss intuition holds for very general coalescent processes [200], the linearity in Theorem 1 enables us to make these statements precise for mutation

rate history via spectral analysis of the operator $\mathcal{L}(\eta)$. This is explored in detail for the case of a simple bottleneck demography in Appendix A.6 and Figure B.9.

2.2.2 Reconstructing the histories of human populations

We next inferred the histories of human populations from large publicly-available resequencing data. We computed a k-SFS for each of the 26 human populations from 5 continental ancestries sequenced in the 1000 Genomes Project (1KG) [1], using an unphased variant call set (mapped to hg38) from the recent high coverage (30x) resequencing data of 1KG samples from the New York Genome Center. Our bioinformatic pipeline for computing the k-SFS for each 1KG population is detailed in Materials and Methods. Briefly, we augment autosomal biallelic SNPs by adding triplet mutation type ($k = 3$) annotations, masking for strict callability and ancestral triplet identifiability. Across 1KG populations the resulting number of segregating variants ranged from ∼7.5 million (population FIN) to ∼15 million (population LWK). We also computed the genomic target sizes for each ancestral triplet, resulting in a total ascertained genome size of ∼2.0 Gb.

We use a de novo mutation rate estimate of $\mu_0 = 1.25 \times 10^{-8}$ per site per generation [182], which corresponds to ∼24.9 mutations per ∼2.0Gb masked haploid genome per generation. For time calibration, we assume a generation time of 29 years [61]. To discretize the time axis for our numerical implementation, we use a logarithmically-spaced grid of 200 points, with the most recent at 1 generation ago, and the oldest at 200 thousand generations (5.8 million years) ago.

Human demographic history

We used `mushi` to infer demographic history $\eta(t)$ independently for each 1KG population. Figure 2.3 shows results grouped by super-population: African, Amerindian, East Asian, European, and South Asian. Broadly, we recover many previously-known features of hu-

man demographic history that are highly robust to regularization parameters: a ∼100 kya (thousand years ago) out-of-Africa bottleneck in non-Africans, a second contraction ∼10 kya due to a founder event in Finland (FIN), and recent expansion of all populations. Histories ancestrally converge within each super-population. Figure B.4 (top panels) shows similar histories inferred using the folded SFS.

Human mutation spectrum history

An estimated demographic history induces a mapping of allele frequency onto a distribution of allele ages. With these distributions encoded in our model, we next used `mushi` to infer time-calibrated MuSHs for each population. First, to highlight the time calibration capabilities of `mushi`, we focus on the specific triplet mutation type TCC→TTC, which was previously reported to have undergone an ancient pulse of activity in the ancestors of Europeans, and is absent in East Asians [82, 83, 199, 198]. To produce sharp estimates of the timing of this TCC pulse, we used regularization that prefer histories with a minimum number of change points (see Materials and Methods). Figure 2.4a shows our fit to this component of the k-SFS for each European population, and Figure 2.4b shows the corresponding estimated component of the MuSH. With the consistent joint estimation performed by `mushi`, we find that the TCC pulse is much older than previously reported, beginning ∼80 kya.

It is also possible to run `mushi` without estimating a new demographic history from the input data, but instead assuming a pre-specified demography. When we use the Tennessen, et al. history [208], which was assumed by Harris and Pritchard [83] in their estimate of the timing of the TCC pulse, we recover a pulse that reaches a maximum around ∼20-30 kya, similar to that initial estimate (Figure B.5, top panels). However, this demography fits the SFS poorly, indicating that demographic misspecification may be distorting `mushi`'s time calibration. Indeed, a global scale shift in the SFS arises from inconsistency in the

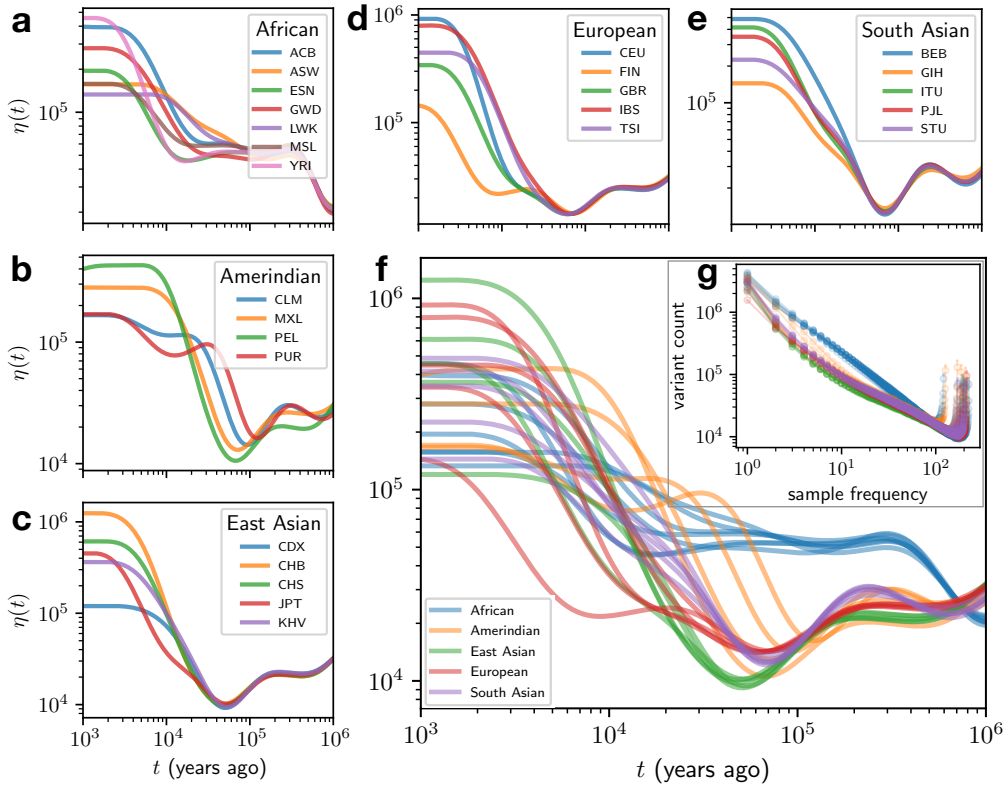

Figure 2.3: Effective population size histories estimated from high coverage 1000 Genomes Project data. **a-e.** Estimated demographic history $\eta(t)$ for all populations grouped by continental ancestry. Inference was performed using a mixture of order 0 and order 3 trends (see Materials and Methods and Figure B.6). African populations: African Caribbeans in Barbados (ACB); Americans of African Ancestry from the Southwestern US (ASW); Esan in Nigeria (ESN); Gambian in Western Divisions in the Gambia (GWD); Luhya in Webuye, Kenya (LWK); Mende in Sierra Leone (MSL); Yoruba in Ibadan, Nigeria (YRI). Amerindian populations: Colombians from Medellin (CLM); Mexican Ancestry from Los Angeles (MXL); Peruvians from Lima (PEL); Puerto Ricans (PUR). East Asian populations: Chinese Dai in Xishuangbanna (CDX); Han Chinese from Beijing (CHB); Han Chinese from Shanghai (CHS); Japanese from Tokyo (JPT); Kinh in Ho Chi Minh City(KHV). European populations: Utah Residents (CEPH) with Northern and Western European Ancestry (CEU); Finnish in Finland (FIN); British in England and Scotland (GBR); Iberian population in Spain (IBS); Toscani in Italia (TSI). South Asian populations: Bengali from Bangladesh (BEB); Gujarati Indian from Houston (GIH); Indian Telugu from the UK (ITU); Punjabi from Lahore (PJL); Sri Lankan Tamil from the UK (STU). **f.** The same $\eta(t)$ estimates on common axes, to allow comparison of super-populations. **g.** SFS data (open circles) for all populations grouped by continental ancestry, as well as fits based on the expected SFS from the estimated demographic history (points connected by lines).

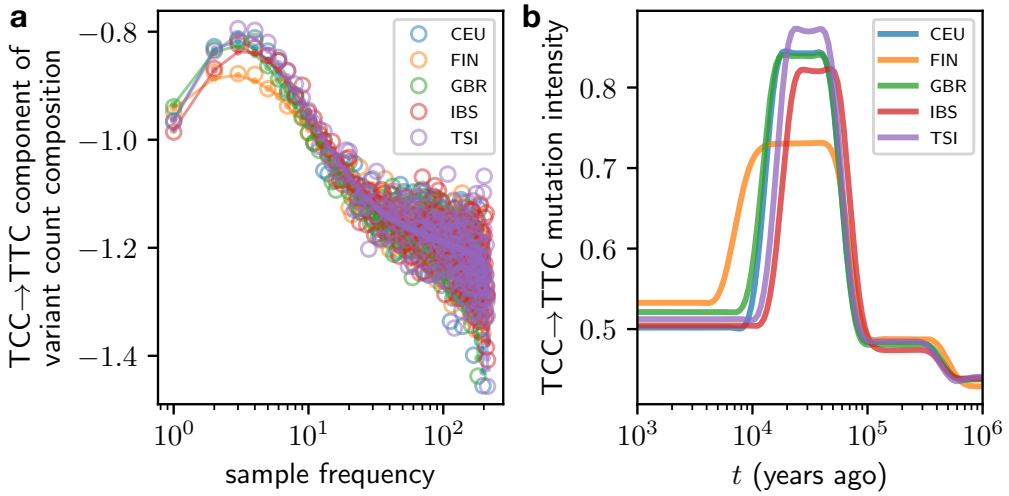

Figure 2.4: Timing of TCC→TTC pulse in Europeans. **a.** The relative composition of TCC→TTC variants (centered log ratio transform, see Materials and Methods) in each frequency class for each European population (open circles) shows an excess at intermediate frequencies. The expected values fit using the inferred MuSH are shown as points connected by lines. **b.** The inferred TCC→TTC mutation rate histories (in units of mutations per ascertained genome per generation). MuSH inference was performed using a mixture of order 0 and order 3 trends (see Materials and Methods, and Figures B.6, B.7, and B.8) for both demography and MuSH.

phylogenetically calibrated mutation rate used by Tennessen et al. (2.36×10^{-8}) and the more recent de novo rate used in `mushi` (1.25×10^{-8}). This inconsistency also distorted the estimate reported by Harris and Pritchard, since their Monte Carlo procedure used the more recent de novo mutation rate. To resolve this, we next rescaled the Tennessen demography to the de novo mutation rate (as done by Amorim et al. [11]), and inferred the TCC pulse with `mushi` again. This resulted in better fit to the SFS, and a clear shift to an older TCC pulse (Figure B.5, dotted lines in top panels), consistent with the pulse inferred using the `mushi` demographies.

We estimated another set of TCC pulses in Europeans conditioned on demographic histories that were inferred using the method Relate [199], which used the phase 3 1KG data to infer demographic histories for each population by first pruning the population genealogy from an inferred whole genome genealogy of all 1KG samples, and then independently inferring a coalescence rate history for each extracted genealogy. Conditioning on the Relate demographies yields younger estimates of the TCC pulse timing, similar to the estimate under the inconsistent Tennessen model (Figure B.5, bottom panels). The Relate demographic histories for each 1KG population are shown in Figure B.4 (bottom panels), with SFS fits.

Figure B.6 shows that our inference of the TCC pulse is highly robust to demographic model selection among demographic histories that fit the SFS. Figure B.7 shows that TCC pulse timing is robust to regularization strength. Figure B.8 indicates the stability of our history estimates under bootstrap resampling of the variant data (but we caution this does not provide confidence bounds on histories, since our penalized likelihood approach is strongly biased toward simple solutions).

After our focused study of the TCC pulse, we aimed to more broadly characterize how human MuSH decomposes into mutational signatures varying through time in each population. This is inspired by the use of nonnegative matrix factorization to infer mutational signatures associated with mutagenic processes in cancer genomes, which represents a set of tumor mu-

tational spectra as mixtures of a small set of mutational signatures, though our problem is more complex due to the time dimension. To capture this additional dimension, we designed a novel mutational signature extraction method that factorizes a three-dimensional tensor of mutation spectrum histories for all populations, rather than a 2-dimensional matrix of mutation spectra from static samples.

We first ran `mushi` on all 1KG populations using stronger order 3 (cubic) trend penalties that favor smoother variation over time compared with the discontinuous jumps of order 0 penalties that were needed to fit the TCC pulse (see Materials and Methods). This resulted in an estimated MuSH for each population of the 26 populations in the 1KG data. We then normalized each MuSH by the genomic target size for each triplet mutation type, so that mutation rate is rendered site-wise, and stacked the population-wise MusHs to form an order 3 tensor. This tensor is a 3D numerical array with dimensions (num. populations) $\times$ (num. time points) $\times$ (num. mutation types) $= 26 \times 200 \times 96$. When we slice the array along the time axis, we obtain a series of matrices whose rows are the inferred mutation spectra of each 1KG population at a past time t. The numerical value of an entry in the tensor indicates the mutation rate (in units of mutations per site per generation) in a given population, at a given time, and for a given mutation type.

We then sought to extract low-dimensional factors in the population, time, and mutation type domains, analogous to extracting mutation signatures that form a sparse vocabulary for explaining the mutation spectrum variation between tumor mutational profiles, but here including the dimension of time variation. To do this, we used non-negative canonical polyadic tensor factorization (NNCP, reviewed in [115]), which generalizes non-negative matrix factorization to tensors of arbitrary order. The addition of the time dimension means that each mutational signature is associated with a dosage that can jointly increase or decrease over the histories of all populations.

Briefly, we hypothesize that the MuSH tensor can be approximated by a sum of a few

rank-1 tensors, implying that most evolving mutational processes are shared across multiple populations, possibly with different relative intensities over time. A tensor of rank 5, which describes a set of 5 mutation signatures, accurately represents the 1KG MuSH tensor (Figure 2.5a inset). The NNCP decomposition results in 26×5, 200×5, and 96×5 factor matrices for population, time, and mutation type, respectively. Figure 2.5c,d projects population and mutation type factors from 5 dimensions to 2 principal components for visualization. The population factors clearly cluster by super-population. The mutation type factors show a number of mutation types with distinct outlier behavior, including TCC→TTC, as expected.

We next recast the MuSH for each population in terms of the 5 mutation signatures that comprise the tensor factors, capturing covariation among the set of 96 triplet mutation types with the smaller set of signatures. This allows us to characterize and biologically interpret the time dynamics of each mutation signature in each population. Figure 2.5a shows the 5 mutation signatures as loadings in each triplet mutation type. Figure 2.5b shows how each of these 5 signatures varies through time in each 1KG population (computed by projecting 96-dimensional spectra to the 5 mutational signatures in each population at each time). Signature 4 fits the profile of the TCC pulse that affects Europeans, South Asians, and European-admixed Amerindians, containing the previously reported minor component ACC→ATC. It does not, however, contain the minor component CCC→CTC, which was previously inferred from the low coverage 1000 Genomes data to be one of the mutation types associated with the TCC pulse. Signatures 1 and 3 are dominated by C→T mutations at CpG sites, the signature of error-prone repair of deaminated methylcytosines. These signatures are consistently enriched in rare (young) variants across populations. Some of this frequency bias is likely caused by purifying selection against mutations that disrupt the gene-regulatory function of methylated CpG sites. Another contributing factor is likely biased gene conversion, which disfavors the increase in frequency of C/G→A/T mutations (also called strong-to-weak mutations). Signature 2 is enriched for common (old) variants,

and have high loadings of A→G, which is consistent with the action of biased gene conversion to select for the retention of weak-to-strong mutations.

Although the time profiles of these 5 signatures appear to be modulated by biased gene conversion, they also vary between populations at recent times and cannot be explained by a selective force acting uniformly on all non-GC-conservative mutations. We note that we do not see evidence the profile of a signature reported to be enriched specifically in the Japanese population [83]. This signature was thought to stem from a subtle cell line artifact affecting the Japanese HapMap samples [12], and apparently is not a prominent feature of the new high coverage 1KG data, whose genotypes were called without imputation. Signature 5, which is dominated by C→T transitions, is notably depleted in East Asians.

Finally, we used uniform manifold approximation and projection (UMAP) [136] to compute a 2D embedding of mutation *signature* histories (after initially decomposing the MuSHs into 5 mutation signatures as described) of each 1KG population at each time point. Figure 2.5e shows this embedding with the time coordinate added as a third dimension. Despite performing independent inferences for each population's MuSH, we see tree structure that reflects population and super-population ancestry, and convergence toward an ancestral MuSH in the distant past.

2.3 Discussion

It is becoming clear that mutation spectrum variation is a common feature of genetically diverse populations. Initial reports on the existence of such variation were mostly qualitative in nature, focused on enumerating which populations exhibit robust variation along this newly characterized dimension and putting bounds on the possible contributions of bioinformatic error. Here, we have introduced a novel quantitative framework for inferring how this variation arose over time, utilizing variation of all ages from unphased whole genome data to resolve a time-varying portrait of germline mutagenesis. Our method `mushi` can decompose

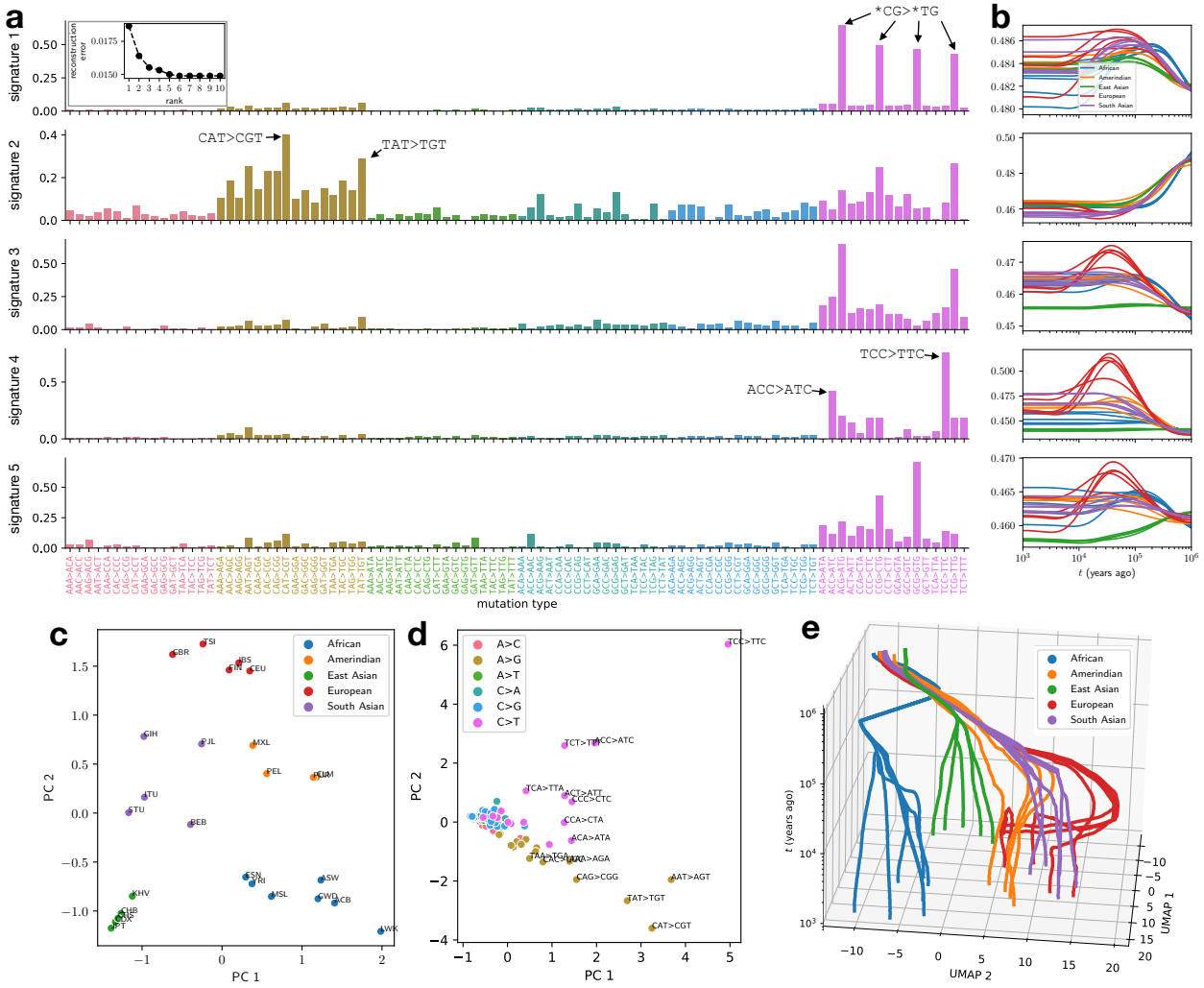

Figure 2.5: Decomposing mutation spectrum histories for 1000 Genomes Project populations into mutation signatures varying through time and between populations. **a.** Triplet mutation signatures, shown as loading into triplet mutation types for each signature (rows). Inset shows tensor reconstruction error over a range of ranks for NNCP decomposition, indicating rank 5 as a good approximation. **b.** Historical dynamics of each mutation signature's intensity in each 1KG population (panels correspond with rows in a). Figure B.10 shows the same intensity histories seperately for each super-population **c.** 5-dimensional population factors projected to first 2 principal components. **d.** 5-dimensional mutation signature factors projected to first 2 principal components. **e.** UMAP embedding of mutation signature histories, initialized using the first two PCs of the time-domain factors, and then performed with default parameters.

context-augmented sample frequency spectra into time-varying mutational signatures, regardless of whether those signatures are sparse and obvious like the European TCC pulse or represent more subtle concerted perturbations of mutation rates in many sequence contexts. Previous estimates of the timescale of mutation spectrum change were restricted to pulse-like signatures that are more obvious but less ubiquitous than diffuse signatures appear to be [83, 199].

Not all of the temporal structure unveiled by `mushi` can be interpreted as time variation in the germline mutational processes. Some time variation in signature dosage is consistent with biased gene conversion, and signatures may also be affected by cell line artifacts [12]. The strengths of `mushi` are to automate the visualization of deviations from mutation spectrum uniformity and localize them to particular populations, frequency ranges, and time periods. It is possible that profiles of germline signatures we report here will need to be revised as higher quality human datasets are published and inference methods are refined. Because `mushi` currently models each population independently, it will be import for future work to accommodate (or jointly model) structured populations that share recent ancestry, such as can be inferred through multipopulation SFS data [80, 102].

Although `mushi`'s most novel feature is the ability to infer mutation spectrum variation over time, it includes a demographic inference subroutine with some advantages over existing methods. We infer population size changes non-parametrically from SFS data with state-of-the-art regularization methods that yield population size histories with some more desirable properties than other methods. The method fastNeutrino [26] uses a piecewise exponential parameterization to infer demographic histories and locus-wise mutation rate from SFS data, and does not use regularization. The method SMC++ [209] uses smoothing spline regularization for demographic inference in a model that combines the efficiency of SFS models with a coalescent HMM. The method CubSFS [219] uses cubic smoothing spline regularization to infer demographic history from the SFS. The sparse trend filtering used in

`mushi` has been shown to have superior local adaptivity properties over the related spline methods [213].

The use of sample allele frequencies rather than phased whole genomes should make `mushi` broadly useful to researchers working on non-model organisms, which are still beyond the scope of many state-of-the-art methods that require long sequence scaffolds and phased data. The software is also very fast, returning results in seconds to minutes on a modest computer, and is designed for researchers familiar with scripting in Python.

The `mushi` model calibrates the times at which mutational signatures wax and wane using a demographic model inferred from the same input allele frequency data from which the signatures themselves are extracted. We estimated a surprisingly old start time to the TCC pulse, around 80 kya, which is older than any estimates of European/East Asian divergence times, and is robust to demographic models that maintain good fit to the SFS. However, `mushi` can also calibrate its timescale using a user-specified demographic history, which reveals that the timing of transient events like the TCC pulse in Europe are sensitive to underlying assumptions about effective population size that fit the SFS poorly. When we input the demographic history used in the initial report of the TCC pulse [83], we similarly find that the TCC pulse began 20-30 kya, comfortably later than Europeans' divergence from East Asians, who were not affected by the TCC pulse. However, it became apparent that the initially reported timing of the TCC pulse was distorted by a scaling issue between recent human de novo mutation rate, and the older phylogenetically calibrated mutation rate used for inferring the demographic history that was used [208]. Rescaling this demography to the de novo rate resulted in a strikingly older TCC pulse, matching the estimate that was obtained using the self-consistent demographic inference from `mushi`.

We also inferred TCC pulse timing using demographic histories that were inferred with the Relate method from whole genome genealogies [199] instead of allele frequency data, and found a younger TCC pulse, matching the initially reported timing that was obtained

with inconsistent demographic history scaling. These demographic histories also yield poor fits to the 1KG SFS data, with more deviation at lower frequencies. However, we note that inferred demographic histories are notoriously poor at predicting the distributions of genomic summary statistics other than the ones that were used to fit the models [20], and of course `mushi` would be unable to recapitulate haplotype structure, for example. We cannot rule out the possibility that another MuSH with a similar SFS, different haplotype structure, and more recent TCC pulse might fit the data better than the MuSH we infer.

If the older TCC pulse timing is correct, complex patterns of ancient gene flow are likely essential for reconciling it with other knowledge about human population history. Ancient DNA evidence suggests that the divergence of East and West Eurasians occurred gradually over a period that began more than 40,000 years ago [225], possibly beginning with the divergence of a basal Eurasian population before the interbreeding of other Eurasians with Neanderthals around 50,000 years ago [119]. Speidel, et al. recently discovered that the proportion of TCC→TTC mutations is highly correlated across populations with the proportion of ancestry from Neolithic Anatolia [198], a finding that underscores the need for future work modeling mutation spectrum evolution jointly with more complex demographic history involving substructure and migration between populations. It also points to the tantalizing possibility that the distribution of mutational signatures could provide extra information about hard-to-resolve substructure and gene flow between populations that lived in the distant past.

Although powerful new methods for inferring ancestral recombination graphs (ARGs) ultimately have the potential to estimate more accurate histories than can be accomplished by fitting compressed SFS data, these methods are still in a relatively early stage of development. In the method Relate [199], mutation rate history is approximately inferred from an ARG using independent marginal estimates for each epoch in a piecewise-constant history. This avoids joint inference over all epochs—which can also be formulated as a linear inverse

problem—by ignoring mutation rate variation within branches.

Until further developments make it possible to infer histories that fit both haplotype structure and site frequency spectra, our results underscore the importance of using more compressed summary statistics to validate inference results. The differences between our SFS-inferred histories and Relate-inferred histories imply that none of these histories yet capture the joint distribution of allele age and allele frequency, which could affect claims about the timing of gene flow and selection in addition to the claims about the timing of the TCC pulse that I focus on in this chapter. Until demographic inference methods are able to infer histories compatible with all features of modern datasets, it will be important for researchers to infer histories from different data summaries, including classical compressed statistics like the SFS, in order to understand the sensitivity of various biological and historical claims to methodological eccentricities.

2.4 Materials and methods

2.4.1 The expected SFS is a linear transform of the mutation intensity history

We work in the setting of Kingman's coalescent [109, 110, 111, 108], with all the usual niceties: neutrality, infinite sites, linkage equilibrium between segregating sites, and panmixia [218, 53]. In Appendix A.1 we retrace the derivation by Griffiths and Tavaré [78] of the frequency distribution of a derived allele conditioned on the demographic history, while generalizing to a time inhomogeneous mutation process. We make use of the results of Polanski et al. [163, 164] to facilitate computation. We use the time discretization of Rosen et al. [175], and adopt their notation. Detailed proofs can be found in the Appendix.

With n denoting the number of sampled haplotypes, denote the expected SFS column vector $\boldsymbol{\xi} = [\xi_1 \ \ldots \ \xi_{n-1}]^\intercal$, where ξ_i is the expected number of variants segregating in i out of n haplotypes. Let $\eta(t)$ denote the haploid effective population size history, with time measured retrospectively from the present in Wright-Fisher generations. Note that $\eta(t) = 2N(t)$ for

diploid populations. Let $\mu(t)$ denote the mutation intensity history, in units of mutations per ascertained genome per generation, understood to apply uniformly across individuals in the population at any given time. Under these model assumptions, we obtain the following theorem.

Theorem 1. *Fix the number of sampled haplotypes n. Then, for all bounded functions $\eta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ and $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, the expected SFS is $\boldsymbol{\xi} = \mathcal{L}(\eta)\mu$, where $\mathcal{L}(\eta)$ is a finite-rank bounded linear operator parameterized by η that maps mutation intensity histories μ to $(n-1)$-dimensional SFS vectors $\boldsymbol{\xi}$. Viewed as a nonlinear operator on η, $\mathcal{L}(\eta)$ is also bounded. In particular, $\mathcal{L}(\eta)\mu \equiv \mathbf{C}\,\mathbf{d}(\eta, \mu)$, where $\mathbf{C}$ is an $(n-1) \times (n-1)$ constant matrix with elements that can be computed recursively, and $\mathbf{d}(\eta, \mu)$ is an $(n-1)$-vector with elements*

$$d_j(\eta, \mu) \equiv \int_0^\infty \exp\left[-\binom{j}{2} \int_0^t \frac{dt'}{\eta(t')}\right] \mu(t)dt, \text{ for } j = 1, \ldots, n-1, \tag{2.1}$$

which is linear in μ and nonlinear in η.

Theorem 1 is proved in Appendix A.1. Recursions for computing $\mathbf{C}$ can be procedurally generated as described in Appendix A.2).

In order to partition the expected SFS $\boldsymbol{\xi}$ by k-mer mutation type, we promote the $(n-1)$-element expected SFS vector $\boldsymbol{\xi}$ to the $(n-1) \times K$ expected k-SFS matrix $\boldsymbol{\Xi}$. Similarly, the mutation intensity history function $\mu(t)$ is promoted to the K-element mutation spectrum history $\boldsymbol{\mu}(t)$, a column vector with each element giving the mutation intensity history function for one mutation type. Then Theorem 1 generalizes to

$$\boldsymbol{\Xi} = \mathcal{L}(\eta)\boldsymbol{\mu}^\intercal. \tag{2.2}$$

As in Theorem 1, the time coordinate is integrated over by the action of the operator $\mathcal{L}$.

Empirical SFS data contain a characteristic "smile" at high frequencies. As detailed in Appendix A.7 , we account for this by modeling ancestral state misidentification rates for each mutation type, and inferring them jointly with the history functions $\eta(t)$ and $\boldsymbol{\mu}(t)$.

We use the notation $\mathbf{X}$ to denote a sampled k-SFS matrix, i.e. the $(n-1) \times K$ matrix containing the sample counts for each mutation type. By construction, $\boldsymbol{\Xi} \equiv \mathbb{E}[\mathbf{X}]$.

2.4.2 Compositional modeling leads to identifiable mutation spectrum histories

As mentioned in the summary methods, the effective population size $\eta(t)$ and the mutation intensity $\mu(t)$ are non-identifiable for all t, meaning that the expected SFS $\boldsymbol{\xi}$ is invariant under a modification of η so long as a compensatory modification is made in μ. We now demonstrate this formally by introducing a change of variables that measures time in expected number of coalescent events since the present, i.e. the diffusion timescale [145, 175]. Let $R_\eta(t) \equiv \int_0^t \frac{dt'}{\eta(t')}$, and substitute $\tau \equiv R_\eta(t)$ in (2.1) to give

$$d_j(\eta, \mu) = \int_0^\infty \exp\left[-\binom{j}{2}\tau\right] \tilde{\eta}(\tau)\tilde{\mu}(\tau)d\tau, \tag{2.3}$$

where $\tilde{\eta}(\tau) \equiv \eta(R^{-1}(\tau))$ and $\tilde{\mu}(\tau) \equiv \mu(R^{-1}(\tau))$. In this timescale, we see η and μ appear as a product on the right of (2.3). This means we cannot jointly infer η and μ, since only their product influences the data. This non-identifiability is similarly manifest by a change of variables to measure time in expected number of mutations.

Because we cannot discern changes in total mutation rate, we assume a constant total rate μ_0, so that time variation in the rate of drift is modeled only in $\eta(t)$. A MuSH with K mutation types can then be written as $\boldsymbol{\mu}(t) = \mu_0\boldsymbol{\nu}(t)$, where $\boldsymbol{\nu}(t) \in \mathcal{S}^K$ for all t, and $\mathcal{S}^K \equiv \{\mathbf{x} \in \mathbb{R}^K_{>0} : \sum_{j=1}^K x_j = 1\}$ denotes the standard simplex. We call the relative mutation spectrum history $\boldsymbol{\nu}(t)$ a *composition*, and employ techniques from compositional data analysis [7, 51, 152].

To avoid difficulties arising from optimizing directly over the simplex, we represent compositions using Aitchison geometry. Briefly, analogs of vector-vector addition, scalar-vector multiplication, and an inner product are defined for compositions, and the simplex is closed under these operations. It is then possible to construct an orthonormal basis in the simplex

$\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{K-1}$ using the Gram-Schmidt orthogonalization. We first introduce the *centered log ratio transform* of some $\mathbf{x} \in \mathcal{S}^K$, defined as

$$\operatorname{clr}(\mathbf{x}) \equiv \left[\log \frac{x_1}{\bar{x}}, \ldots, \log \frac{x_K}{\bar{x}}\right]^\intercal, \tag{2.4}$$

where $\bar{x}$ denotes the geometric mean. The inverse transform clr^{-1} is the softmax function.

The *isometric log ratio transform* and its inverse allow us to transform back and forth between the simplex and a Euclidean space in which we will cast our optimization problem. The transform $\operatorname{ilr} : \mathcal{S}^K \to \mathbb{R}^{K-1}$ and its inverse are defined as

$$\operatorname{ilr}(\mathbf{x}) \equiv \boldsymbol{\Psi}^\intercal \operatorname{clr}(\mathbf{x}), \qquad \mathbf{x} \in \mathcal{S}^K \tag{2.5}$$

$$\operatorname{ilr}^{-1}(\mathbf{y}) \equiv \operatorname{clr}^{-1}(\boldsymbol{\Psi}\mathbf{y})), \qquad \mathbf{y} \in \mathbb{R}^{K-1} \tag{2.6}$$

where $\boldsymbol{\Psi} \equiv [\boldsymbol{\psi}_1 \ \ldots \ \boldsymbol{\psi}_{K-1}]$ is the $K \times (K-1)$ matrix of basis vectors. To build intuition about this transformation, which is an isometric isomorphism, we highlight the following behaviors: First, the center of the simplex maps to the origin in the Euclidean space. Second, approaching a corner of the simplex, i.e. with a component of the composition vanishing, corresponds to diverging to infinity in some direction the Euclidean space. Finally, a ball in the Euclidean space maps to a convex region in the simplex that is more distorted the further the ball is from the origin.

We use the convention that the clr and ilr act row-wise on matrices. Finally, we introduce the ilr-transformed MuSH: $\mathbf{z}(t) \equiv \operatorname{ilr}(\boldsymbol{\mu}(t))$ and write (2.2) as

$$\boldsymbol{\Xi} = \mu_0 \mathcal{L}(\eta) \ \operatorname{ilr}^{-1}(\mathbf{z})^\intercal. \tag{2.7}$$

Again, the time coordinate is integrated over by the action of the linear operator. Although the forward model is non-linear in $\mathbf{z}(t)$, it is convex given the convexity of the softmax function that appears in $\operatorname{ilr}^{-1}(\cdot)$.

2.4.3 Formulating and solving the inverse problem for population history given genomic variation data

The inverse problem (2.8) is ill-posed, meaning many very different and erratic histories can be equally consistent with the data [52]. We deal with this problem using regularization, seeking solutions that are constrained in their complexity without sacrificing data fit. We use optimization algorithms to find regularized demographies and MuSHs.

Time discretization

For numerical implementation, we need finite-dimensional representations of $\eta(t)$ and $\mathbf{z}(t)$. We use piecewise constant functions of time on m segments $[t_0, t_1), [t_1, t_2), \ldots, [t_{m-1}, t_m)$ where the grid $0 = t_0 < t_1 < \cdots < t_{m-1} < t_m = \infty$ is common to $\eta(t)$ and $\mathbf{z}(t)$. We take the boundaries of the segments as fixed parameters and, in practice, use a logarithmically-spaced dense grid of hundreds of segments to approximate infinite-dimensional histories. Let the m-vector $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$ denote the population size $\eta(t)$ during each segment, and define the $m \times (K-1)$ matrix $\mathbf{Z}$ as the constant ilr-transformed MuSH $\mathbf{z}(t)$ during each segment. In Appendix A.3 , we show that equation (2.7) discretizes to the following matrix equation

$$\mathbf{\Xi} = \mu_0 \mathbf{L}(\mathbf{y}) \ \mathrm{ilr}^{-1}(\mathbf{Z}), \tag{2.8}$$

where the $(n-1) \times m$ matrix $\mathbf{L}(\mathbf{y})$ is fixed given a fixed demographic history $\mathbf{y}$. The transformation $\mathrm{ilr}^{-1}(\mathbf{Z})$ is applied to each time point, i.e. row of $\mathbf{Z}$, independently.

Regularization

We implement three different regularization criteria: sparsity of trends in the solutions $\mathbf{y}$ and $\mathbf{Z}$ (hypothesizing that the time variation of $\eta(t)$ and $\mathbf{z}(t)$ is not excessively erratic), sparsity of the singular value spectrum of the matrix $\mathbf{Z}$ (hypothesizing that the number of independently evolving mutational signatures is much less than the number K of distinct

mutation types), and improved numerical conditioning of the problem. These goals are in some cases overlapping, but we add a regularization term for each one. Before computing the penalties on the demography $\mathbf{y}$, we apply a log transform, because variation over orders of magnitude is expected from population crashes and exponential expansions. This also has the benefit of enforcing non-negative solutions. We now explain the regularizations in detail.

Our first regularization imposes simplicity in the time domain by preferring solutions with a small number of piece-wise polynomial trends. This is achieved by penalizing the variation of $\log \eta(t)$ and $\mathbf{z}(t)$: the ℓ_1 norm of their time derivatives. Penalizing the $(\kappa+1)$-th order time derivative encourages piecewise κ-th order polynomial solutions, since the ℓ_1 norm favors sparse derivatives in time. For example, $\kappa = 0$ results in piecewise constant solutions, $\kappa = 1$ results in piecewise linear solutions, and so on. Penalties with different κ can be combined to obtain mixed trends, e.g. using $\kappa = 0$ and $\kappa = 3$ will allow solutions with both constant and cubic pieces. In the discretized model, the $(\kappa+1)$-th order derivative operator corresponds to a matrix $\boldsymbol{D}^{(\kappa+1)}$ of finite differences. This leads to the penalties $\left\|\boldsymbol{D}^{(\kappa+1)} \log \mathbf{y}\right\|_1$ and $\left\|\boldsymbol{D}^{(\kappa+1)}\mathbf{Z}\right\|_{1,1}$ (penalizing the MuSH column-wise). In the least-squares setting, this regularization is called *trend filtering* [106, 213], and is one of many generalizations of the Lasso method [85]. We later describe how we perform optimization with trend penalties in the setting of a more complex likelihood. Many demographic inference methods fit models composed of a small number of constant or exponential epochs that are motivated by prior knowledge about population histories. Although our histories are represented on a dense time grid, our regularization fuses the history at neighboring time points to discover epochs within which behavior is simple, while remaining flexible to capture more complicated behavior if the data justify it.

Second, because specific mutation processes may affect multiple mutation types, it is reasonable to assume that a small number of latent processes drive the majority of the variation across mutation types. We thus hypothesize that $\mathbf{Z}$ can be approximated by a

low-rank matrix and propose two regularizations to enforce this. Let $\boldsymbol{\sigma}$ be the vector of singular values of $\mathbf{Z} - \mathbf{Z}_{\text{ref}}$, where $\mathbf{Z}_{\text{ref}}$ is a reference, or baseline, MuSH taken to be the MLE constant solution by default. We use the nuclear norm $\|\mathbf{Z} - \mathbf{Z}_{\text{ref}}\|_* = \|\boldsymbol{\sigma}\|_1$ as a *soft* rank penalty, as it is the convex envelope of the rank function [54]. The soft rank penalty constrains the number of non-zero singular values, while also shrinking them toward zero. As an alternative to the soft rank penalty we also implement a *hard* rank penalty, which directly penalizes $\text{rank}(\mathbf{Z} - \mathbf{Z}_{\text{ref}}) = \|\boldsymbol{\sigma}\|_0$, equal to the number of nonzero singular values. The hard rank penalty results in a singular value thresholding step without shrinkage in the resulting algorithm, and it is not convex. Either of these rank regularizations assure that $\mathbf{Z}$ is a low-rank perturbation of the constant solution $\mathbf{Z}_{\text{ref}}$. Although the MuSH represents the history of each of K mutation types, this attempts to explain them using a smaller set of mutation signatures.

Finally, we include classical ℓ_2 (also called ridge or Tikhonov) penalties on both $\log \mathbf{y}$ and $\mathbf{Z}$. A small amount of this kind of regularization speeds up convergence without significantly influencing the solution. For the ridge penalty on the demography $\mathbf{y}$, we use a Tikhonov term $\|\log \mathbf{y} - \log \mathbf{y}_{\text{ref}}\|_2^2$ that shrinks toward a reference demography $\mathbf{y}_{\text{ref}}$. By default we use the MLE constant history for $\mathbf{y}_{\text{ref}}$ to speed the convergence of the $\mathbf{y}$ problem. Similarly, the ridge penalty on the MuSH is a Tikhonov term for each mutation type, the squared Frobenius norm $\|\mathbf{Z} - \mathbf{Z}_{\text{ref}}\|_{\text{F}}^2$.

Likelihood factorization: The SFS is a sufficient statistic for the demographic history with respect to the k-SFS

The Poisson Random Field (PRF) approximation neglects linkage disequilibrium to model the probability of the SFS $\mathbf{x}$ given the expected SFS $\boldsymbol{\xi}$ as independent Poisson random

variables for each sample frequency

$$\mathbb{P}(\mathbf{x} \mid \boldsymbol{\xi}) = \prod_{i=1}^{n-1} \mathbb{P}(x_i \mid \xi_i) = \prod_{i=1}^{n-1} \frac{e^{-\xi_i} \xi_i^{x_i}}{x_i!}. \tag{2.9}$$

We similarly model the k-SFS as generated by independent mutational targets for each mutation type, and then show that a constant total mutation rate allows us to factorize the joint likelihood for η and $\boldsymbol{\mu}$ into a sequential inference procedure for η then $\boldsymbol{\mu}$.

Proposition 1. *The PRF, when generalized to the 2D grid of sample frequency and mutation type, factorizes as* $\mathbb{P}(\mathbf{X} \mid \boldsymbol{\Xi}) = \mathbb{P}(\mathbf{x} \mid \boldsymbol{\xi})\ \mathbb{P}(\mathbf{X} \mid \mathbf{x}, \hat{\boldsymbol{\Xi}})$*, where* $\mathbb{P}(\mathbf{x} \mid \boldsymbol{\xi})$ *is the standard PRF* (2.9)*, and* $\mathbb{P}(\mathbf{X} \mid \mathbf{x}, \hat{\boldsymbol{\Xi}})$ *is independent multinomial for each sample frequency* i*, with multinomial parameter* $\hat{\Xi}_{i,j} \equiv \frac{\Xi_{i,j}}{\xi_i}$*.*

Proposition 1 is proved via a Poissonization argument in Appendix A.4.

Next we restore the η and $\boldsymbol{\mu}$ dependence of $\boldsymbol{\xi}$ and $\boldsymbol{\Xi}$ (with fixed total mutation rate μ_0) so Proposition 1 gives the factorization

$$\mathbb{P}(\mathbf{X} \mid \eta, \boldsymbol{\mu}) = \mathbb{P}(\mathbf{x} \mid \eta)\ \mathbb{P}(\mathbf{X} \mid \mathbf{x}, \eta, \boldsymbol{\mu}). \tag{2.10}$$

Lemma 1. *If the total mutation rate is a constant* $\boldsymbol{\mu}(t) = \mu_0 \in \mathbb{R}_{>0}$*, then the SFS* $\mathbf{x}$ *is a sufficient statistic for* η *with respect to the* k*-SFS* $\mathbf{X}$*.*

Lemma 1 is proved via a Poisson thinning argument in Appendix A.5. The result is intuitively obvious because information about historical coalescence rates recorded in the SFS does not change if we further specify how mutation counts are partitioned into different mutation types; this only adds information about relative mutation rates for alleles with a given age distribution. Although η appears in the second factor of (2.10), this only serves to map the MuSH rendered on the natural diffusion timescale $\tilde{\boldsymbol{\mu}}(\tau)$ to time measured in Wright-Fisher generations. Because this map is one-to-one, there is no statistical information about η in $\mathbf{X}$ not already present in $\mathbf{x}$. That is, $\mathbb{P}(\mathbf{X} \mid \mathbf{x}, \eta, \boldsymbol{\mu}) = \mathbb{P}(\mathbf{X} \mid \mathbf{x}, \tilde{\boldsymbol{\mu}})$.

This sufficiency is important from an inference perspective, because it means we can sequentially infer demography from the SFS, then infer the MuSH from the k-SFS with the demography fixed from the previous step. Sufficiency implies that the negative log-likelihood factors into the sum of two losses. We thus formulate two sequential optimization problems using negative log-likelihoods from the factors (2.10) as loss functions for assessing data fit. Recall that $\mathbf{y}$ and $\mathbf{Z}$ are the discrete forms of η and $\boldsymbol{\mu}$, respectively, $\boldsymbol{\Xi}$ is given by equation (2.8), and $\boldsymbol{\xi}$ is given by the row sums of $\boldsymbol{\Xi}$ and thus independent of $\mathbf{Z}$. Neglecting constant terms, the two loss functions are

$$\text{loss}_1(\log \mathbf{y}) = \sum_{i=1}^{n-1} (\xi_i - x_i \log \xi_i) \tag{2.11}$$

and

$$\text{loss}_2(\mathbf{Z}; \mathbf{y}) = -\sum_{i=1}^{n-1} \sum_{j=1}^{K} X_{ij} \log \Xi_{ij}. \tag{2.12}$$

As with regularization, we parameterize in terms of $\log \mathbf{y}$.

Optimization problems for `mushi`

We infer demography and MuSH by minimizing cost functions that combine the loss functions above, which measure error in fitting the data, with regularization. This may be considered a penalized likelihood method and, from a Bayesian perspective, may be interpreted as introducing a prior distribution over histories. Inference of $\log \mathbf{y}$ and $\mathbf{Z}$ is performed sequentially. We first initialize $\log \mathbf{y} = \mathbf{y}_{\text{ref}}$ using the elementary formula for the MLE constant demography $\frac{S}{2\mu_0 H_{n-1}}$ where $S \equiv \sum_{i=1}^{n-1} x_i$ is the number of segregating sites, and H_{n-1} denotes the n-th harmonic number. We then minimize

$$\begin{aligned} f_1(\log \mathbf{y}) = \text{loss}_1(\log \mathbf{y}) &+ \alpha_\kappa \left\| \boldsymbol{D}^{(\kappa+1)} \log \mathbf{y} \right\|_1 \\ &+ \frac{\alpha_{\text{ridge}}}{2} \left\| \log \mathbf{y} - \log \mathbf{y}_{\text{ref}} \right\|_2^2 \end{aligned} \tag{2.13}$$

over $\log \mathbf{y} \in \mathbb{R}^m$ to obtain the demographic history. Here, the α_κ hyperparameter controls the trend penalty strength, and determines the number of κ-th order polynomial pieces in the solution (a larger penalty results in fewer pieces). The α_{ridge} hyperparameter controls the strength of shrinkage toward $\mathbf{y}_{\text{ref}}$, and is intended to improve convergence without stongly biasing the solution.

Having fixed $\mathbf{y}$ from the previous step, we next infer $\mathbf{Z}$. We initialize $\mathbf{Z} = \mathbf{Z}_{\text{ref}}$ to the MLE constant MuSH: mutation type j has the constant rate $\mu_0 \frac{S_j}{S}$, where $S_j \equiv \sum_{i=1}^{n-1} X_{i,j}$ is the number of segregating sites in mutation type j. Using the default soft rank penalty, we then minimize

$$f_2(\mathbf{Z}) = \text{loss}_2(\mathbf{Z}; \mathbf{y}) + \beta_\kappa \left\| \boldsymbol{D}^{(\kappa+1)} \mathbf{Z} \right\|_{1,1} + \beta_{\text{rank}} \|\mathbf{Z} - \mathbf{Z}_{\text{ref}}\|_* + \frac{\beta_{\text{ridge}}}{2} \|\mathbf{Z} - \mathbf{Z}_{\text{ref}}\|_{\text{F}}^2 \tag{2.14}$$

over $\mathbf{Z} \in \mathbb{R}^{m \times (K-1)}$ to obtain the ilr-transformed MuSH. Using the hard rank penalty instead of the default soft rank penalty, we would replace the nuclear norm $\|\cdot\|_*$ with the rank function $\text{rank}(\cdot)$. The β_κ and β_{ridge} hyperparameters are analogous to α_κ and α_{ridge}. The β_{rank} hyperparameter controls the the rank of $\mathbf{Z}$ (a larger penalty results in smaller rank). We note that the trend order κ can be different for demography and MuSH inference, and each can use mixed trends, adding more terms if desired.

We now briefly cover the methods used for optimization. The cost function (2.13) is nonconvex due to the nonlinear dependence of $\boldsymbol{\xi}$ on $\mathbf{y}$, while the cost function (2.14) is convex. The trend penalties on both (2.13) and (2.14) are nonsmooth, as is the soft rank penalty on (2.14). If the hard rank penalty is used instead of the soft rank penalty, (2.14) is also nonconvex. Although we cannot guarantee convergence to the global minimum for the demographic history ($\mathbf{y}$) problem, we have found that proximal gradient methods rapidly converge to good solutions that are robust to initalization. Briefly, in proximal methods the cost is split into differentiable and non-differentiable parts, gradient descent steps are taken using the smooth part of the cost, then the *proximal operator* (or *prox*) of the non-differentiable piece

is applied. The prox projects to a nearby point which ensures that the nonsmooth part of the cost is small and can be computed for the trend filtering and hard or soft rank penalties. For the $\mathbf{y}$ problem, we use the Nesterov accelerated proximal gradient method with adaptive line search [148, 18]. For the MuSH ($\mathbf{Z}$) problem, we use a three operator splitting method to deal with the two nonsmooth terms [155]. We implemented a specialized ADMM trend filtering algorithm to compute proximal operators for our mixed trend penalties [171]. Our optimization algorithms are implemented very generally as a Python submodule in the `mushi` package: `https://harrispopgen.github.io/mushi/stubs/mushi.optimization.html`.

Hyperparameter tuning

Although `mushi` does not require a parametric model to be specified, it requires the user to tune a few key regularization hyperparameters to target reasonable solutions. Rather than treat the ridge penalties as adjustable hyperparameters, we fix them to $\alpha_{\text{ridge}} = \beta_{\text{ridge}} = 10^{-4}$ to improve convergence without noticably influencing solutions. This leaves the trend penalty α_κ (or penalties for mixed trends) for demographic inference. Inferring demography from SFS data requires strong priors on the simplicity of solutions, so there can be no general recipe for selecting optimal hyperparameters. It is generally advisable to explore a few trend orders κ and their strengths.

Small trend penalties give erratic, unregularized solutions. Increasing α_κ limits the number of κ-th order pieces in the solution, and can be set to produce solutions that are consistent with known features of population history. Over-regularization is indicated when the fit to the SFS becomes poor, and can be seen in an "elbow plot" of the loss with increasing penalization. Mixing a 0-th order term with higher-order models helps flatten the end points of the time domain, which may be desired.

We take a similar approach for the MuSH inference step. The two hyperparameters in this case are the trend hyperparameter β_κ and the rank hyperparameter β_{rank}. With $\kappa = 0$,

pulse-like histories can be recovered, while for higher orders (e.g. $\kappa = 3$) smoothly varying histories are recovered (but don't fit pulse components as well). Again, over-smoothing is indicated by poor fit to the k-SFS. We set β_{rank} to select a rank (number of latent histories) between 3 and 6. If β_{rank} is too large, the rank will be too small to fit all components of the k-SFS well. If it is too small, it is more is more difficult to find common features in different populations. By default we prefer the soft rank penalty for its convexity, but can choose the hard rank penalty if the former results in undesirable shrinkage.

2.5 Software implementation methods

2.5.1 The open-source `mushi` Python package

The `mushi` software is available as a Python 3 package at `https://harrispopgen.github.io/mushi` with extensive documentation. We use the `JAX` package [27] for automatic differentiation and just-in-time compilation of our optimization methods, and the `ProxTV` package [16] for fast computation of total variation proximal operators. We modified the compositional data analysis module in the `scikit-bio` package `http://scikit-bio.org` to allow `JAX` compatibility. Using default parameters, inferring the demography and MuSH for a population of hundreds of individuals takes a few seconds on a laptop with a modest hardware configuration.

2.5.2 Reproducible analysis

All of the analysis and figures for this paper can be reproduced using `Nextflow` pipelines [44] and `Jupyter` notebooks (`https://jupyter.org`) available at `https://github.com/harrispopgen/mushi-pipelines`. We used `msprime` [103] and `stdpopsim` [3] for simulations, `TensorLy` [116] for NNCP tensor decomposition, `umap-learn` [136] for UMAP embedding, and several other standard Python packages. We used the Mathematica package `fastZeil` [151] to procedurally generate recursion formulas for the combinatorial matrix $\mathbf{C}$

in Theorem 1 (see Appendix A.2).

We generated k-SFS data for each 1KG population using `mutyper` [40] (`https://harrispopgen.github.io/mutyper/`) and `BCFtools` [124] (`http://samtools.github.io/bcftools`). High coverage 1KG variant call data were accessed at `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/`, with sample manifest available at `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel`. Ancestral state estimates for hg38 were accessed at `ftp://ftp.ensembl.org/pub/release-100/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh38.tar.gz`, and the strict callability mask was accessed at `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38/StrictMask/20160622.allChr.mask.bed`. Relate coalescence rate histories were accessed at `https://zenodo.org/record/3234689`.

Chapter 3

MUTYPER: ASSIGNING AND SUMMARIZING MUTATION TYPES FOR ANALYZING GERMLINE MUTATION SPECTRA

Genomics studies of the germline mutation process seek to elucidate the mutational forces that drive genetic variation and provide the raw material for adaptive evolution. Germline mutations arise from spontaneous DNA damage caused by environmental mutagens, or errors in DNA replication. Populations and species may experience distinct histories of mutational input due to differences in environmental exposure, and heritable variation in the machinery controlling DNA replication fidelity.

Diverse mutational processes can reveal themselves in tell-tale mutation signatures left in the nucleotide sequence contexts in which they tend to be active. Population genomics has given increasing attention to nucleotide sequence context in the study of the germline mutation process (reviewed in [31]). SNPs can be assigned mutation context given the ancestral background they occurred on, which is typically estimated via parsimony with an outgroup species alignment or multispecies tree. It is standard to partition SNPs into *mutation types* according to ancestral and derived nucleotide states ("polarizing" the reference and alternative states), and a window of local nucleotide context in the ancestral background on which the SNP occurred. The *mutation spectrum* of an individual is defined as the composition of derived state SNPs over these mutation types.

Inter- and intra-specific germline mutation spectrum variation has revealed a dynamic and evolving germline mutation process shaping modern genomic diversity. Parsing mutation spectra temporally (i.e. according to allele frequency) and spatially (i.e. in different genomic compartments) has revealed the history and present of the mutational processes, and applying

such analysis to de novo mutation data may be clinically informative for rare or undiagnosed genetic diseases.

Despite many exciting findings to date, there is a lack of available software for general-purpose germline mutation type partitioning and mutation spectrum generation from population scale genomic variation data needed for larger bioinformatics pipelines or simple exploratory analysis. To address this need, I developed `mutyper`, an open-source command-line utility and Python package to:

- assign ancestral mutation types to SNP data in variant call format (VCF) files [37]
- summarize mutation types at the individual level with mutation spectra
- summarize mutation types at the population level with sample frequency spectra for each mutation type
- compute distributions of ancestral k-mer content in genomic features specified by a BED mask file, to standardize spectra across genomic feature sets.

The literature on cancer somatic mutation signatures includes several software tools (many implemented as R packages) for clustering and dimensionality reduction that are not directly amenable to population scale germline variation data, but the package `Helmsman` [32] enables more efficient interoperability with these tools. Complementing this integrative work, `mutyper` is a minimal, flexible, and extensible software package designed for population genomics researchers to generate the raw material needed to advance new analyses of germline mutation spectrum variation. `mutyper` can be installed using the `pip` package manager and is compatible with Python 3.6+. Documentation is provided at `https://harrispopgen.github.io/mutyper`; source code is available at `https://github.com/harrispopgen/mutyper`.

3.1 Implementation

3.1.1 CLI

`mutyper` is a Python package with a command-line interface (CLI) that implements the core functionality of assigning ancestral mutation types to SNPs that are input (or piped) in VCF/BCF format using the subcommand `mutyper variants`. Fast processing of VCF input is achieved with `cyvcf2` [153], and mutation types are assigned via the INFO field for each variant, e.g. by including a key-value pair such as `mutation_type=GAG>GTG`. A parameter `--k` allows the user to specify the k-mer context size desired (by default 3 for triplet mutation types and spectra). As in previous work, mutation types are, by default, collapsed by reverse complementation such that the ancestral state is either `A` or `C`. Alternatively, the argument `--strand_file` can be used to provide a BED defining the strand orientation for nucleotide context at each site (e.g. based on direction of replication or transcription).

To polarize ancestral and derived allelic states, and define ancestral k-mer backgrounds, an input FASTA defining the ancestral genome estimate is required. `mutyper` uses the package `pyfaidx` [191] for fast random access to ancestral genomic content, with minimal memory requirements. Ancestral genomes can be specified by various means. The ancestral FASTA sequence provided by the 1000 Genomes Project [1] was estimated from a multi-species alignment using `Ortheus` [150]. In such a case, the ancestral FASTA can be passed to `mutyper` directly. Alternatively, ancestral states can be simply estimated by polarizing SNPs using an outgroup genome aligned to the reference (e.g. the chimp genome liftover to the human genome).

Below is a simple example `Bash` command that uses ancestral states defined in FASTA `anc.fa` to assign mutation types to SNPs in VCF file `in.vcf`, and redirects the augmented VCF data to `out.vcf`.

```
$ mutyper variants anc.fa in.vcf > out.vcf
```

More complex analysis can be achieved by filtering the input SNPs (i.e. with `bcftools`, [124]), and piping to `mutyper` instead of passing `in.vcf` as a command-line argument. The `mutyper` command-line utility is fully compatible with standard command-line pipelines for filtering SNPs or samples, masking regions, and merging/concatenating VCFs.

In addition to this core functionality, the CLI includes several other subcommands that facilitate research that aims to characterize modern mutation spectrum variation, and infer its evolutionary history. The subcommand `mutyper spectra` generates mutation spectra for each individual in input VCF/BCF data. The subcommand `mutyper targets` computes the number of ancestral genomic targets for each mutation type, optionally restricting to genomic features specified in BED format with the argument `--bed`. This facilitates standardizing mutation spectra across genomic features with varying target content. The subcommand `mutyper ksfs` generates a site frequency spectrum (SFS) for each each mutation type. The SFS is a widely used population genetic summary statistic that informs both demographic history and mutation spectrum history when partitioned by mutation type [41].

3.1.2 Python API

While the `mutyper` CLI enables one to include mutation type analyses in the context of a bioinformatics pipeline, the `mutyper` Python API enables one to perform the functions above in an interactive notebook session, or to implement custom analyses of mutation type data by interfacing with the strong ecosystem of scientific computing packages available in Python. For example, dimensionality reduction (such as principal components analysis or non-negative matrix factorization) is often used to summarize mutation spectra, and the `scikit-learn` package [154] can be used in conjunction with the `mutyper` API for this purpose.

3.2 Conclusion

The literature has seen a recent and rapid expansion of studies of germline mutation spectrum variation and evolution. This motivates the development of computational tools to make these analyses more standardized, reproducible, and accessible; `mutyper` meets this need.

Chapter 4

USING GENOTYPE ABUNDANCE TO IMPROVE PHYLOGENETIC INFERENCE

> The flowers have been growing thorns for millions of years. For millions of years the sheep have been eating them just the same. And is it not a matter of consequence to try to understand why the flowers go to so much trouble to grow thorns which are never of any use to them? Is the warfare between the sheep and the flowers not important? Is this not of more consequence than a fat red-faced gentleman's sums?
>
> Antoine de Saint-Exupéry, *The Little Prince*

Although phylogenetic inference methods were originally designed to elucidate the relationships between groups of organisms separated by eons of diversification, the last several decades have seen new phylogenetic methods for populations that are evolving on the timescale of experimental sampling [46]. This development is being spurred by new experimental techniques that enable deep sequencing at single-cell resolution, some of which enable quantification of original abundance. For bulk sequencing, random barcodes can be used to quantify PCR template abundance [112, 97, 28]. More recently, cell isolation [190] or combinatorial techniques [36, 42, 92] have provided sequence data at single-cell resolution. With such data, a given unique genotype—among many in the data—is represented in a measured number of cells. The *abundance* of a genotype can be read out as the number of cells bearing that genotype. Here we demonstrate that incorporating genotype abundance improves phylogenetic inference for densely sampled evolutionary processes in which it is common to sample genotypes more than once.

We are motivated by the setting of B cell development in germinal centers. B cells are

the cells that make antibodies, or more generally *immunoglobulins*. Immunoglobulins are encoded by genes that undergo a stage of rapid Darwinian mutation and selection called *affinity maturation* [139]. During affinity maturation, immunoglobulin is in its membrane-bound form, known as the *B cell receptor* (BCR). The biological function of this process is to develop BCRs with high-affinity for a pathogen-associated *antigen* molecule, and later excrete large quantities of the associated antibody.

This affinity maturation process occurs in specialized sites called *germinal centers* in lymph nodes, which have specific cellular organization to enable B cells to compete among each other to bind a specific antigen (proliferating more readily if they do) while mutating their BCRs via a mechanism called *somatic hypermutation* (SHM). Using micro-dissection, researchers can extract germinal centers from model animals and sequence the genes encoding their BCR directly [206, 118]. Lymph node samples are also available through autopsy [202] or fine needle aspirates from living subjects [86]. Such samples provide a remarkable perspective on an ongoing evolutionary process.

Indeed, these samples contain a population of cells with BCRs that differentiated via SHM at various times and have various cellular abundances. Because the natural selection process in germinal centers appears permissive to a variety of BCR-antigen affinities [206, 118], earlier-appearing BCRs are present at the same time as later-appearing BCRs. The collection of descendants from a single founder cell in this process naturally form a phylogenetic tree. However, it is a tree in which each genotype is associated with a given abundance, and such that older ancestral genotypes are present along with more recent appearances. Reconstruction of phylogenetic trees from BCR data may benefit from methods designed to account for these distinctive features.

Standard sequence-based methods for inferring phylogenies fall into several classes according to their optimality criteria. *Maximum likelihood* methods posit a probabilistic substitution model on a phylogeny and find the tree that maximizes the probability of the

observed data under this model [57, 58, 55]. *Bayesian* methods augment likelihood with a prior distribution over trees, branch lengths, and substitution model parameters, and approximate the posterior distribution of all the above variables by Markov chain Monte Carlo (MCMC) [93, 45]. *Maximum parsimony* methods use combinatorial optimization to find the tree that minimizes the number of evolutionary events [50, 114, 64]. Parsimony methods often result in degenerate inference, in which multiple trees achieve the same minimal number of events (i.e. maximum parsimony) [131]. Additional approaches include *distance matrix* methods, which summarize the data by the distances between sequence pairs, and *phylogenetic invariants*, which select topologies based on the value of polynomials calculated on character state pattern frequencies. None of the above methods incorporate genotype abundance information, and it is standard for data with duplicated genotypes to be reduced to a list of *deduplicated* unique genotypes before a phylogeny is inferred.

In this chapter I show that genotype abundance is a rich source of information that can be productively integrated into phylogenetic inference, and we provide an open-source implementation to do so. We incorporate abundance via a stochastic branching process with infinitely many types for which likelihoods are tractable, and show that it can be used to resolve degeneracy in parsimony-based optimality. We first validate the procedure against simulations of germinal center BCR diversification. We also empirically validate our method using an experimental lineage tracing approach combining multiphoton microscopy and single cell BCR sequencing, allowing us to study individual germinal center B cell lineages from brainbow mice. Beyond the setting of BCR development, we foresee direct application to tumor phylogenetics in single-cell studies of cancer evolution (reviewed by Schwartz et al. [186]), and single-cell implementations of lineage tracing based on genome editing technology [137].

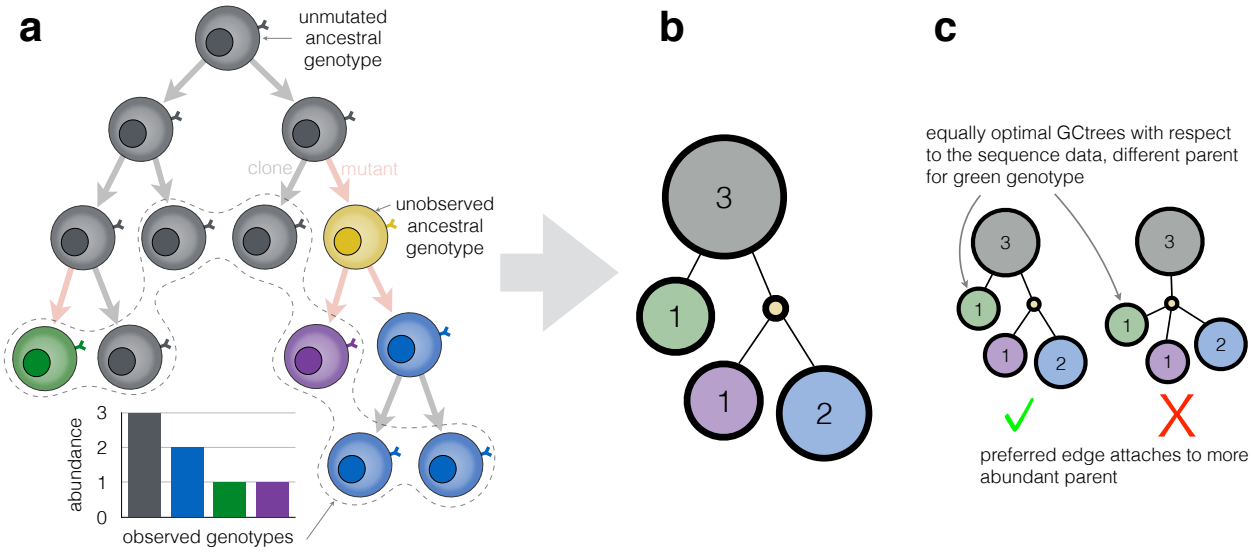

Figure 4.1: Genotype-collapsed trees. **(a.)** A diversifying B cell lineage is illustrated with distinct BCR genotypes colored. The final observed cells (enclosed by a dashed path) consist of genotypes at various abundances; note the yellow genotype is not observed. **(b.)** The corresponding genotype-collapsed tree (GCtree) describes the descent of distinct genotypes, and is our inferential goal. **(c.)** Genotype abundance informs topology inference. Two hypothetical GCtrees, equally optimal with respect to the sequence data, propose two possible parents of the green genotype—the gray and yellow genotypes (the yellow genotype was not sampled and thus has a small circle with no number inside). Intuitively, the abundance information indicates that the tree on the left is preferable because the more abundant parent is more likely to have generated mutant descendants.

4.1 New Approaches

4.1.1 Genotype-collapsed trees

Given sequence data obtained from a diversifying cellular *lineage tree* (Figure 4.1a), our goal is to infer the *genotype-collapsed tree* (GCtree) defining the lineage of distinct genotypes and their observed abundances (Figure 4.1b). The GCtree is constructed from the lineage tree by collapsing subtrees composed of cells with identical genotype to a single node annotated with its final cellular abundance. Our data consists of the genotypes sampled at least once

in the GCtree, along with their associated abundances. Under the *infinite types* assumption that every mutant daughter generates a novel genotype, each genotype can be identified with one subtree in the original lineage tree. We are not claiming any originality in the GCtree definition, but it is useful to have a word for this object.

We note that, unlike standard phylogenetic trees where only leaf nodes represent observed genotypes, GCtree internal nodes represent observed genotypes if they are annotated with non-zero abundance. Although not leaves *per se* in the GCtree, a nonzero abundance represents a clonal sub-lineage that resulted in a nonzero number of leaves of that genotype in the lineage tree. A node in the GCtree, along with its descending edges, summarizes the lineage outcome for a given genotype as its mutant offspring clades and the number of its clonal leaves. Because this summary does not completely specify the genotype's clonal lineage structure (Figure 4.2c), several branching structures may be consistent with a given node, and we have no information with which to distinguish between the various lineage trees consistent with a GCtree. Hence, our goal is to infer the GCtree topology.

4.1.2 Parsimony with a prior

BCR sequence data from a germinal center sample has the following characteristics from the perspective of phylogenetics: genotypes have abundances, there is a limited amount of mutation between genotypes, and ancestral genotypes are present along with later ones. The latter two features suggest maximum parsimony as a useful tool because of the limited amount of mutation and because ancestral genotypes can be assigned to internal nodes of the tree (although recent Bayesian methods can do such assignment as well [70, 69]). For these reasons, parsimony has been used extensively in B cell sequence analysis [15, 202]. Because having many duplicate sequences inhibits efficient tree space traversal, these studies have inferred trees using the unique genotypes (BCR sequences). This ignores the varying cellular abundances of the observed genotypes.

Here we wish to use a branching process model to rank trees that are equally optimal according to sequence-level optimality criteria. Indeed, maximum parsimony often results in degenerate inference: there are many trees that are maximally optimal [131]. We refer to these trees as a *parsimony forest*. In later sections we show, using *in silico* and empirical data, that parsimony degeneracy is common and often severe for BCR sequencing data, and that parsimony forests exhibit substantial variation in phylogenetic accuracy. It is common practice to arbitrarily select one tree in the parsimony forest at random, without regard for this variability in inference accuracy. Instead, we will rank trees in the parsimony forest with an auxiliary likelihood that incorporates abundance information, thereby resolving this degeneracy.

Genotype abundance is an additional source of information for phylogenetics, using the simple intuition that more abundant genotypes are more likely to have more mutant descendant genotypes. This intuition makes sense because relative sample abundance is a reasonable estimator of relative total historical abundance, and total historical abundance is related to the number of mutant offspring—i.e. genotypes with larger abundance are likely to have more mutant descendant genotypes simply because there are more individuals available to mutate. The number of mutant offspring genotypes is in turn related to the number of surviving mutant offspring sampled. Thus, given two equally parsimonious trees, this intuition would prefer the tree that has more mutant descendants of a frequently observed node (Figure 4.1c). We formalize this intuition using a stochastic process model for the phylogenetic development of germinal centers, and integrate this model with sequence-based tree optimality via empirical Bayes.

In this stochastic process model, a GCtree node i has a random number $T_i \in \mathbb{N}$ of mutant children (i.e. descending edges) and a random abundance $A_i \in \mathbb{N}$. We will index nodes in a "level order" as follows, which is well defined given an embedding of the tree into the plane. Index 1 refers to the root node, and 2 through $1 + T_1$ refer to the children of the root node.

The level-order continues in order through all tree nodes of the same level before nodes at the next level. Adopting this level-ordering convention, a GCtree containing N nodes is specified by integer-valued random vectors giving the (planar) topology $\mathbf{T} = (T_1, \ldots, T_N)$, and abundances $\mathbf{A} = (A_1, \ldots, A_N)$. We also have the observed genotype sequences associated with each node $\mathbf{G} = (G_1, \ldots, G_N)$.

A complete diversification model would give a joint distribution on $\mathbf{T}$, $\mathbf{G}$, and $\mathbf{A}$. As an approximation to such a model, facilitating use of existing sequence-based optimality methods, we propose a generative model containing conditional independences as follows (Figure 4.2a). First, we model abundances $\mathbf{A}$ and tree topology $\mathbf{T}$ as being drawn from a branching process likelihood, conditioned on parameters $\boldsymbol{\theta}$ (characterizing birth, death, and mutation rates in the underlying lineage tree): $\mathbb{P}(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta})$. This stochastic process likelihood will capture the intuition (described above) that more abundant genotypes are likely to have more mutant descendant genotypes. Next, we assume that genotype sequences $\mathbf{G}$ are generated by a mutation model conditioned on the fixed tree $\mathbf{T}$, independent of $\mathbf{A}$. This sequence-based optimality is captured by a distribution over $\mathbf{G}$ dependent only on $\mathbf{T}$: $\mathbb{P}(\mathbf{G} \mid \mathbf{T})$. The lack of direct dependence of $\mathbf{G}$ on $\mathbf{A}$ constitutes an approximation to a more realistic sequence-valued branching process. However, this formulation has the advantage that it allows us to leverage standard sequence-based phylogenetic optimality in the specification of $\mathbb{P}(\mathbf{G} \mid \mathbf{T})$. In a later section (*In silico* validation), we validate this approximation with simulations that do not assume this conditional independence.

In an empirical Bayes treatment (see Appendix C for details), a maximum likelihood estimate for the branching process parameters, $\hat{\boldsymbol{\theta}}$, can be obtained by marginalizing over $\mathbf{T}$, and this in turn can be used to approximate a posterior over $\mathbf{T}$ conditioned on the data $\mathbf{G}$ and $\mathbf{A}$ (as well as $\hat{\boldsymbol{\theta}}$). Using parsimony as our sequence-based optimality, one can rank trees in the parsimony forest (denoted $\mathcal{T}_{\mathbf{G}}$) according to the GCtree likelihood. We encode the parsimony criteria in $\mathbb{P}(\mathbf{G} \mid \mathbf{T})$ by assigning uniform weight to the trees in $\mathcal{T}_{\mathbf{G}}$, and zero to

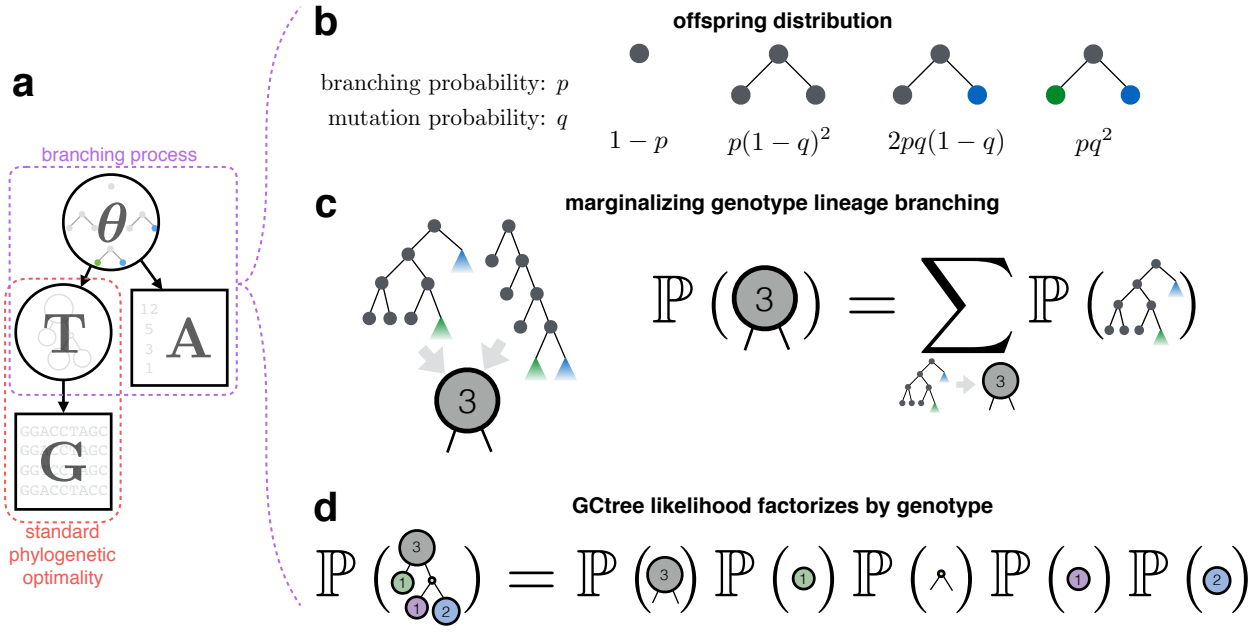

Figure 4.2: Modeling sequences equipped with abundances. **(a.)** Both genotype sequence data $\mathbf{G}$ and genotype abundance data $\mathbf{A}$ inform tree topology $\mathbf{T}$. As illustrated in this probabilistic graphical model, we assume independence between $\mathbf{G}$ and $\mathbf{A}$ conditioned on $\mathbf{T}$ rather than a fully joint model of $\mathbf{G}$, $\mathbf{A}$, and $\mathbf{T}$. This facilitates using standard sequence-based phylogenetic optimality for $\mathbf{G}$, augmented with a branching process (with parameters $\boldsymbol{\theta}$) for $\mathbf{A}$. **(b.)** For the binary infinite-type Galton-Watson process, $\boldsymbol{\theta} = (p, q)$. Four possible branching events characterize the offspring distribution common to all nodes. A node may bifurcate (with probability p) or terminate, and upon bifurcating its descendants each may be a mutant (with probability q). **(c.)** A GCtree node specifies a genotype's clonal leaf count and number of descendant genotypes, but not lineage details. The likelihood of a GCtree node marginalizes over consistent lineage branching outcomes. **(d.)** GCtree likelihood factorizes into the product of likelihoods for each genotype.

the other trees. This gives the following approximate maximum a posteriori tree:

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}\in\mathcal{T}_\mathbf{G}} \mathbb{P}\left(\mathbf{A}, \mathbf{T} \mid \hat{\boldsymbol{\theta}}\right), \tag{4.1}$$

where the point estimate $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{T}\in\mathcal{T}_\mathbf{G}} \mathbb{P}\left(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta}\right). \tag{4.2}$$

Next we turn to explicitly defining the GCtree likelihood $\mathbb{P}\left(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta}\right)$.

4.1.3 A stochastic process model of abundance

To compute likelihoods $\mathbb{P}\left(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta}\right)$ for GCtrees (Figure 4.1b), we model the lineage tree (Figure 4.1a) as a subcritical infinite-type binary Galton-Watson (branching) process [84] in which extinct leaf nodes correspond to observed cells. All mutations in an infinite-type process result in a novel genotype, embodying the assumption that each genotype can be identified with one subtree. Subcriticality ensures that the branching process terminates in finite time, so an explicit sampling time is not needed. The process is initiated with a single cell (a naive germinal center B cell before affinity maturation ensues), and runs to eventual extinction. This model is highly idealized and unable to capture many biological realisms of B cell affinity maturation and the sampling process. However, as we show in our validations, it is useful as a minimal model for leveraging genotype abundance information in a tractable likelihood.

The offspring distribution for our process, governing reproduction and mutation for all lineage tree nodes at all time steps, is specified by two parameters: the binary branching probability p, and the mutation probability q. Because the offspring distribution is independent of type, subcriticality simply requires that the expected number of offspring of any node is less than 1, in this case equivalent to $p < 0.5$. In this case a "mutation" is an event that causes the evolving lineage to change to a novel genotype (under the infinite-types assumption). Thus the corresponding offspring distribution supports four distinct branching

events (Figure 4.2b). Letting C and M denote the (random) number of clonal and mutant offspring of any given node in the lineage tree, respectively, the offspring distribution is

$$\mathbb{P}\left(C=c, M=m\right) = \begin{cases} 1-p & c=m=0, \\ p(1-q)^2 & c=2, m=0, \\ 2pq(1-q) & c=m=1, \\ pq^2 & c=0, m=2, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

We can compute the likelihood of a hypothetical binary lineage tree simply by evaluating (4.3) at each node in the tree and multiplying the results. The likelihood for a GCtree is then given by summing over all possible binary lineage trees that are consistent with that GCtree (i.e. that give the same GCtree when collapsing by genotype), thus marginalizing out the details of intra-genotype branching events that give rise to the same abundance. Here we show how to calculate the GCtree likelihood directly for the simple offspring distribution (4.3). Other work [23] has described how to calculate statistics of the infinite-type branching process with a general subcritical offspring distribution.

First consider the likelihood for an individual node in the GCtree, say the root node, in the context of the branching process described above. A GCtree node i is specified by its abundance A_i and the number of edges descending from it T_i (both random variables). There are, in general, multiple distinct branching process realizations for genotype i that result in $A_i = a$ clonal leaves and $T_i = \tau$ mutations off the genotype i lineage subtree (Figure 4.2c). Determining the likelihood of node i in the GCtree under this process, which we denote by $f_{a\tau}(p,q) = \mathbb{P}\left(A_i = a, T_i = \tau \mid \boldsymbol{\theta} = (p,q)\right)$, requires marginalizing over all such genotype lineage subtrees. In Appendix C we derive a recurrence for $f_{a\tau}(p,q)$ by marginalizing over the outcome of the branching event at the root of the lineage subtree for genotype i, and show that the GCtree node likelihood $f_{a\tau}(p,q)$ can be computed by dynamic programming.

A complete GCtree containing N nodes is specified by level-ordering the nodes as described above $\mathbf{T} = (T_1, \ldots, T_N)$, $\mathbf{A} = (A_1, \ldots, A_N)$. Because the same offspring distribution generates the lineage branching of each genotype subtree, the same recurrence can be applied to all GCtree nodes. Specifically, we show in Appendix C that the joint distribution over all nodes in a GCtree factorizes by genotype (Figure 4.2d):

$$\begin{aligned} &\mathbb{P}\left(\mathbf{T} = (\tau_1, \ldots, \tau_N), \mathbf{A} = (a_1, \ldots, a_N) \mid \boldsymbol{\theta} = (p, q)\right) \\ &\qquad = \prod_{i=1}^{N} f_{a_i \tau_i}(p, q)\,. \end{aligned} \tag{4.4}$$

Using dynamic programming and factorization by genotype, the computational complexity of the GCtree likelihood is $\mathcal{O}(\max(A)\max(T)+N)$. Ranking parsimony trees with `GCtree` requires a polynomial increase in runtime compared with finding the parsimony forest, which is itself NP-hard [65]. Figure D.1 depicts runtime from simulations of various size, and shows that, in practice, this increased runtime is negligible.

A computational implementation of the inference method above is available at `https://matsengrp.github.io/gctree`. The `GCtree` inference subprogram accepts sequence data in `FASTA` or `PHYLIP` format, determines a parsimony forest from the unique sequences using the `dnapars` program from the `PHYLIP` package [56], determines the genotype-collapsed form of these trees and outputs tree visualizations using the `ETE` package [94], and ranks them according to their GCtree likelihood using the sequence abundances. Bootstrap analysis is also implemented, providing confidence values of each split in the maximum likelihood GCtree. The GCtree maximizing the branching process likelihood (with optional bootstrap support) is the inference result. Next we show that resolving parsimony degeneracy using `GCtree` substantially increases both accuracy and precision of phylogenetic inference.

4.2 Results

4.2.1 In silico validation

To explore the accuracy and robustness of `GCtree` inference, we developed a simulation subprogram to generate random lineages starting with a naive BCR sequence. For simulated lineages, true trees can be compared against those inferred with the `GCtree` inference subprogram. The stochastic process model used in `GCtree` inference is intended as a minimal model (in terms of biological realism) that captures the intuition that genotype abundance is relevant to phylogenetic reconstruction. Experimental data need not obey our simplifying assumptions, thus we set out to test `GCtree`'s robustness to deviations of the data generating process from the inferential model.

A simulation process was implemented that includes biological realisms of B cells undergoing SHM (and violates inferential assumptions). These realisms of simulation—detailed in Materials and Methods—include: branching process multifurcations (controlled by a parameter λ, the expected number of children of a node in the cell lineage tree), sequence context sensitive mutations [49, 201] (with a baseline-line mutation rate λ_0, and a context-specific mutational model with 5mer mutabilities taken from [223]), explicit sampling time (t, or N representing the number of cells desired in the sampled generation), incomplete sampling (the number of cells to sample $n \leq N$), and repeated genotypes allowed (deviation from the infinite-type assumption). This constitutes a more challenging validation than simply simulating under the same assumptions that had been invoked for tractability of the inferential framework.

Our *in silico* validation workflow is demonstrated in Figure 4.3a for a small simulation that resulted in a parsimony forest with just two equally parsimonious trees. The output of the simulation software consists of `FASTA` data (sequences and their abundances), visualizations of the lineage tree and its GCtree equivalent, and a file containing the true GCtree

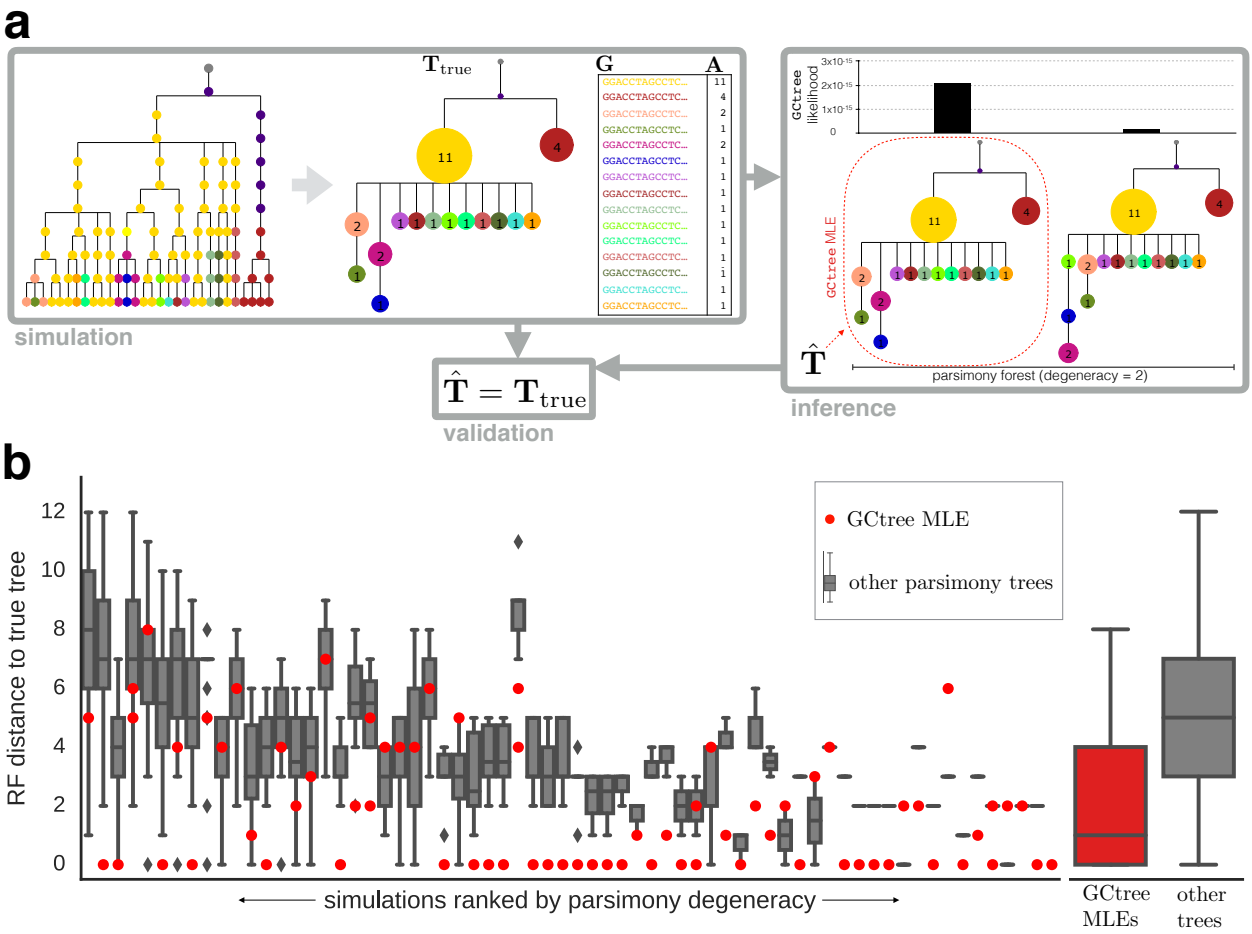

Figure 4.3: *In silico* validation of `GCtree` inference. **(a.)** Demonstrating the simulation–inference–validation workflow, a small simulation resulted in two equally maximally parsimonious trees, and the one inferred using `GCtree` was correct. The initial sequence was a naive BCR V gene from the experimental data described in Materials and Methods. Branch lengths in the cell lineage tree (left) correspond to simulation time steps, while those in collapsed trees correspond to sequence edit distance. **(b.)** 100 simulations were performed with parameters calibrated using the BCR sequencing data and summary statistics described in Materials and Methods. Of 100 simulations, 66 resulted in parsimony degeneracy, with an average degeneracy of 12 and a maximum degeneracy of 124. For each of these 66, we show the distribution of Robinson–Foulds (RF) distance of trees in the parsimony forest to the true tree. "RF" denotes a modified Robinson-Foulds distance: since nonzero abundance internal nodes in GCtrees represent observed taxa, RF distance was computed as if all such nodes had an additional descendant leaf representing that taxon. GCtree MLEs (red) tend to be better reconstructions of the true tree than other parsimony trees (gray boxes). Four simulations resulted in a tie for the GCtree MLE, and the two tied trees in these cases are both displayed in red. Aggregated data across all simulations are depicted on the right, clearly indicating superior reconstructions from `GCtree`.

structure. The `GCtree` inference subprogram can then be run on the `FASTA` data, and the resulting inferred GCtree compared to the true GCtree (in this case they were identical). To calibrate simulation parameters, we defined summary statistics on sequence data with abundance information, and tuned parameters to produce data similar to experimental BCR sequencing data under these statistics (see Materials and Methods).

Our validation shows that using abundance information via a branching process likelihood can substantially improve inference results (Figure 4.3b). For each simulation we ranked otherwise degenerately optimal parsimony trees using `GCtree`. For each parsimony forest, we compared the GCtrees in the forest to the true GCtree for that simulation using the Robinson–Foulds (RF) distance [174] as a measure of tree reconstruction accuracy. The maximum likelihood GCtree tends to be closer to the true tree than other equally parsimonious trees, which vary widely in accuracy, showing that GCtree is able to leverage abundance data to resolve parsimony degeneracy and improve the accuracy of tree reconstruction in this simulation regime.

4.2.2 Empirical validation

We next performed a biological validation by investigating if `GCtree` improves inference according to biological criteria using real germinal center BCR sequence data. The BCR is a heterodimer encoded by the immunoglobulin heavy chain (IgH) and immunoglobulin light chain (IgL) loci. Both loci undergo V(D)J recombination, and then evolve in tandem during affinity maturation. By obtaining matched sequences from both loci using single-cell isolation, we have two independent data sets to inform the same phylogeny of distinct cells (each of which is associated with a single IgH sequence and single IgL sequence). Performing separate and independent IgH and IgL tree inference, we can then validate `GCtree` by comparing the inferred IgH tree to the inferred IgL tree. If the GCtree likelihood (4.4) meaningfully ranks equally parsimonious trees, then the two MLE trees (IgH and IgL) would be expected

to be more correct reconstructions than the other parsimony trees. Thus, we are to expect that the two MLE trees are more similar to each other (in terms of the lineage of distinct cells) than other pairs of IgH and IgL parsimony trees (which, if they are more distorted phylogenies, should show less concordance in the partitioning of the distinct cells). Conversely, if the GCtree likelihood is not meaningfully ranking trees, we expect that the MLE IgH and IgL trees will not be significantly closer to each other than other pairs of IgH and IgL parsimony trees.

We used data from a previously reported experiment in which multiphoton microscopy and BCR sequencing were combined to resolve individual germinal center B cell lineages from mouse lymph nodes 20 days after subcutaneous immunization with alum-adsorbed chicken gamma globulin [206] (see Materials and Methods). *Brainbow* mice were used for multicolor cell fate mapping, enabling B cells and their progeny to be permanently tagged with different fluorescent proteins. In-situ photo-activation followed by fluorescence-activated cell sorting yielded B cells from a color-dominant germinal center (Figure 4.4a, left). BCR sequences were obtained for 48 cells in this lineage by single cell mRNA sequencing of the IgH and IgL loci, resulting in 32 distinct IgH and 26 distinct IgL genotypes due to SHM mutations acquired through affinity maturation. The unmutated naive IgH and IgL V(D)J rearranged sequences (not observed) were inferred with `partis` using each set of 48 sequences (IgH and IgL) as a clonal family using germline genetic information [169, 170]. These naive sequences were used as outgroups for rooting parsimony trees.

`GCtree` results are depicted in Figure 4.4b. Parsimony analysis resulted in degeneracy for both loci, with 13 equally parsimonious trees for IgH, and 9 for IgL. Empirical Bayes point estimation according to (4.2) yielded $\hat{p} = 0.495$, $\hat{q} = 0.388$ (IgH) and $\hat{p} = 0.495$, $\hat{q} = 0.304$ (IgL). GCtree likelihoods (4.4) were computed to rank the equally parsimonious trees, and the MLE trees are shown with support values among 100 bootstrap samples (see Materials and Methods). Because the binary Galton-Watson process assigns probability zero to a

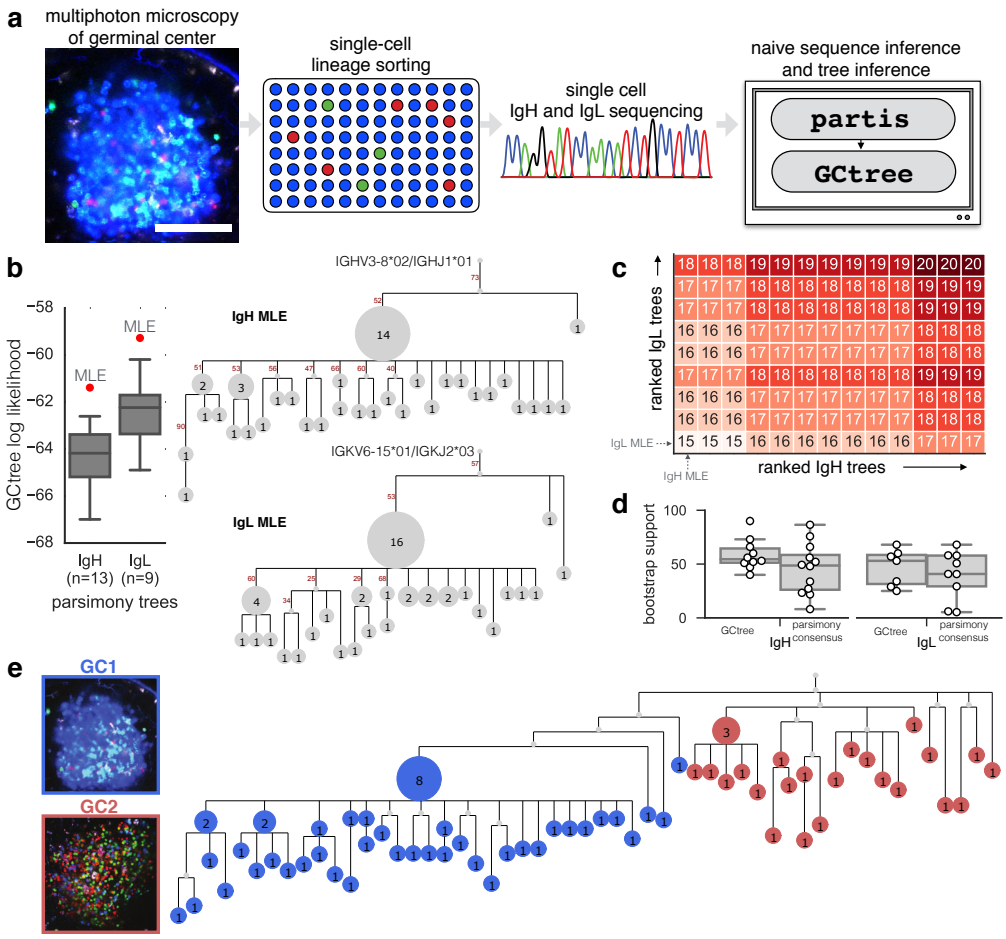

Figure 4.4: Empirical validation using lineage tracing and single cell germinal center BCR sequencing. **(a.)** A multiphoton image of a germinal center reveals a dominant blue lineage (scale bar 100μm). This lineage was sorted, and 48 cells sequenced to determine IgH and IgL genotypes of each. These sequences were analyzed with `partis` [169, 170] to infer naive (pre-affinity-maturation) ancestor sequences using germline genetic information, and trees were inferred with `GCtree`. **(b.)** `GCtree` inference was performed separately for IgH and IgL loci, resulting in parsimony degeneracies of 13 and 9, respectively. Maximum likelihood GCtrees for each locus are indicated in red and the GCtrees with annotated abundance are shown. Roots are labeled with the gene annotations of the naive state inferred using `partis`. Small unnumbered nodes indicate inferred unobserved ancestral genotypes. Numbered edges indicate support in 100 bootstrap samples. **(c.)** All possible pairings of IgH and IgL parsimony trees were compared in terms of the Robinson-Foulds distance between the IgH and IgL trees, labeled by cell identity. IgH and IgL parsimony trees are ordered by GCtree likelihood rank in columns and rows, respectively. Grid values show RF distance between each IgH/IgL pair. MLE trees result in more consistent cell lineage reconstructions between IgH and IgL (smaller RF values). **(d.)** For each locus, distributions of bootstrap support values are shown for the tree inferred by `GCtree` and for a majority rule consensus tree of all trees in the parsimony forest. The latter contain more partitions with very low support. **(e.)** Using additional data from a second germinal center from the same lymph node that had the same naive BCR sequence, `GCtree` correctly resolves the two germinal centers as distinct clades (as did other lower ranked parsimony trees).

GCtree node with frequency zero and one mutant descendant, the unobserved naive root node (which had one descendant after rerooting and collapsing identical genotypes in all parsimony trees) was given a unit pseudocount.

We then compared the concordance between pairs of heavy and light trees. Since both IgH and IgL loci have been recorded from the same set of 48 cells, the units of cell abundance in an IgH GCtree map to the units of cell abundance from an IgL GCtree (i.e. each cell identity among the 48 is associated with an IgH genotype and an IgL genotype). We can then consider the consistency of a given IgH tree and a given IgL tree in terms of the lineage of the 48 cell identities. For each possible pairing of an IgH parsimony tree with a IgL parsimony tree, we computed the RF distance [174] between the two trees using the cell identities (rather than the genotype sequences) to define splits. We observed that the GCtree MLE based on IgH sequences and GCtree MLE based on IgL sequences form the most concordant pair among all pairs of parsimony trees (Figure 4.4c). Moreover, pairs of parsimony trees that contained at least one GCtree MLE tree ranked consistently higher in terms of their similarity.

We assessed confidence in `GCtree` partitions by comparing bootstrap support values in `GCtree` trees to those from the majority-rule consensus parsimony trees made using the `consense` program from the `PHYLIP` package [56]. We observed the latter contained an excess of very low confidence partitions (Figure 4.4d, Figure D.4). These results demonstrate that parsimony reconstructions for real BCR data sets suffer from degeneracy, and that GCtree likelihood can correctly resolve this degeneracy by incorporating abundance information ignored by previously published methods.

Finally, using data collected from a second germinal center from the same lymph node, we tested `GCtree`'s ability to correctly group cells from each germinal center into separate clades when run on combined data from both germinal centers. The two germinal center sequence data sets appeared to have the same naive BCR sequence (IgH and IgL), indicating they were both seeded from the same B cell lineage. Concatenating the IgH and IgL sequences

for each cell in each germinal center, we used `GCtree` to infer a single tree for all cells from both germinal centers (Figure 4.4e). `GCtree` correctly resolved the two germinal centers as distinct clades (we note that all the parsimony trees had this feature, regardless of likelihood rank). This demonstrates the phylogenetic resolvability of germinal centers with the same naive BCR diversifying under selection for the same antigen specificity.

4.3 Discussion

We have shown that genotype abundance information can be productively incorporated in phylogenetic inference. By augmenting standard sequence-based phylogenetic optimality with a stochastic process likelihood, we were able to implement abundance-aware inference as a processing step downstream of results from an existing and widely used parsimony tree inference tool. We have shown that our method—implemented in the publicly available `GCtree` package—is useful for inferring B cell receptor affinity maturation lineages. Although branching processes have been used previously to infer parameters of BCR evolution [113, 133] and construct SHM lineage trees from error-prone bulk sequencing reads [197], to our knowledge we are the first to use branching processes to sharpen phylogenetic inference for BCRs sequenced at single-cell resolution from germinal centers.

We believe `GCtree` will find use in other settings where sequence data from dense quantitative sampling of diversifying loci are available. Studies of cancer evolution are increasingly performed with single-cell resolved sequencing, however most tumor phylogenetics approaches use standard phylogenetic methods (reviewed by Schwartz et al. [186]) that do not model genotype abundance. Exceptions include `OncoNEM` [176] and `SCITE` [98], both of which leverage single-cell data for tumor phylogenetic inference that is robust to genotyping errors and missing data, but do not aim to capture the intuition that genotype abundance and the number of direct mutant descendants are related. Single-cell implementations of lineage tracing based on genome editing technology [137] may also benefit from reconstruction

methods that model the abundance of observed editing target states, since cell types may vary widely in rates of proliferation.

Using parsimony as our sequence-based optimality resulted in particularly simple results, as the tree space necessary to explore is exactly the degenerate parsimony forest. However, our empirical Bayes formulation is agnostic to the particular choice of sequence-based optimality, so in the future we envision augmenting likelihood-based sequence optimality. This will require more computationally expensive tree space search and sampling schemes.

In contrast to `GCtree`, a fully Bayesian approach to incorporate genotype abundance could use the full set of sequences (without deduplication) in a Bayesian phylogenetics package—such as `BEAST` [45]—with a birth-death process prior. This would not enforce the infinite-type assumption, so a set of identical sequences could be placed in disjoint subtrees. However, such an approach will not scale well with many identical sequences: trees that only differ by exchange of identical sequences will create islands of constant posterior in tree space. Methods do not currently exist for tree space traversal that avoids moves within such islands. Even if such methods existed, they would need to be combined with algorithms to infer trees with sampled ancestors [70, 69] as well as multifurcations [122, 123]; even just this combination is not currently available.

Although our methods can be applied to other sequence-based optimality functions besides parsimony, it is important to recognize that `GCtree` (and indeed any tree inference procedure that deduplicates repeated sequences) contains an inherent weak parsimony assumption: that each unique genotype arose from mutation just once in the lineage and therefore corresponds to a single subtree in the lineage tree, and thus a single node in the GCtree. Thus it is important to continue to assess the impact of this weak parsimony assumption with simulation that does not make this assumption, as done here.

The `GCtree` framework can also be extended to non-neutral models. For example, one could consider a model in which each genotype obtains a random fitness encoded by branch-

ing process parameters $\boldsymbol{\theta}$ that are fixed within a given genotype but randomly drawn by the genotype founder cell upon mutation from its parent. This will likely necessitate modeling genotype birth time explicitly, rather than restricting to extinct subcritical processes, since a genotype with small abundance may be a result of low fitness or just young age. One might also consider extending the offspring distribution to separately model synonymous and nonsynonymous mutations. Synonymous mutations do not change fitness, while nonsynonymous mutations change fitness as described above. Another direction of extension is to incorporate mutation models specialized to the case of BCR evolution, such as the S5F model [223] used in our simulation study.

4.4 Materials and Methods

4.4.1 Simulation details.

To provide for a more challenging *in silico* validation study, several biological realisms were built into our simulation that defied simplifying assumptions in the `GCtree` inference methodology.

Arbitrary offspring distribution.

The recursion (C.7) used to compute GCtree likelihood components specifies a binary branching process, and such an approach would in general require an offspring distribution with bounded support on the natural numbers. Our simulation implements an arbitrary offspring distribution with no explicit bounding. In the results that follow, we used a Poisson distribution with parameter λ for the expected number of offspring of each node in the lineage tree.

Context sensitive mutation.

To generate mutant offspring, all offspring sequences (drawn from a Poisson as described above) were subjected to a sequence-dependent mutation process. The SHM process is known to introduce mutations in a sequence context-dependent manner, with certain hot-spot and cold-spot motifs [49, 201]. We used a previously published 5-mer context model S5F [223] to compute the mutabilities $\mu_1, \ldots, \mu_\ell$ of each position $1, \ldots, \ell$ within a sequence of length ℓ based on its local 5-mer context. This model also provided substitution preferences among alternative bases given the 5-mer context. To compute mutabilities for beginning and ending positions without a complete 5-mer context, we averaged over missing sequence context.

Although existing code can simulate a mutational process parameterized by S5F on branches of a fixed tree with a pre-specified number of mutations on each branch [79], in our simulations we wanted the number of mutations on the branches to be determined by the sequence mutability as it changes via mutation across the tree. For example, as an initial mutation hotspot motif acquires mutations down the tree, its mutability typically degrades as it diverges from the original motif. We defined the mutability of the sequence as a whole by the average over its positions $\mu_0 = \frac{1}{\ell}\sum_{i=1}^{\ell}\mu_i$. We defined a baseline mutation expectation parameter λ_0 as a simulation parameter, and the number of mutations m any given offspring sequence received was drawn from a Poisson distribution. The Poisson parameter was modulated by the sequence's mutability $m \sim \text{Pois}(\mu_0\lambda_0)$, so that more mutable sequences tended to receive more mutations. Given $m > 0$, the positions in the sequence to apply mutations were chosen sequentially as follows. A site j to apply the first mutation was drawn from a categorical distribution using the site-wise mutabilities to define relative probability of choosing each site $j \sim \text{Cat}(\mu_1, \ldots, \mu_\ell)$. We mutated the site using a categorical distribution over the three alternative bases parameterized by the substitution preferences defined by the site's context. We then updated mutabilities μ_0 and $\mu_1, \ldots, \mu_\ell$ as necessary to account for

contexts that had been altered by the mutation. This process was repeated m times.

Since the mutability of each node in the lineage tree will depend on the mutation outcome of its parent, the GCtree likelihood components will not factorize by genotype. Because the probability of mutation is sequence-dependent, the topology of the GCtree will be sequence-dependent. Therefore, the generative assumptions of the empirical Bayes inference do not hold in this simulation scheme, nor does the offspring distribution equivalence across lineage tree nodes specified by (4.3).

Sampling time.

Our inference model specifies a subcritical branching process run until extinction, and sampling of all terminated nodes (leaves). Our simulation more realistically assigns a discrete time of sampling parameter t (number of time steps from root), and thus does not need to constrain the offspring distribution to achieve subcriticality. At the specified time, extant nodes can be sampled, so all genotypes that terminated or mutated at a prior times are not observed. Alternatively, a parameter N specifying the desired number of simulated observed sequences may be passed, in which case the simulation runs until a time such that at least N sequences exist (unless terminated). Genotypes born at different times will be sampled under a process with different effective sampling times since birth. Thus this sampling time parameter also increases dependence between genotypes, further distancing the simulation model from the inferential model.

Incomplete sampling.

We introduce imperfect sampling efficiency with a parameter n for the number of simulated sequences that end up in the simulated sample data (`FASTA`), requiring $n \leq N$. This violates the inferential assumption of complete sampling, and renders the true genotype abundances latent variables (which a more complete likelihood approach might aim to marginalize out).

Repeated genotypes.

Our simulation is seeded with an initial naive BCR sequence, from which randomly mutated offspring are created. Because there is no built-in restriction that the same sequence cannot arise along different branches (or mutations could be reversed), the model assumption of infinite types—such that identical sequences can be associated with a single genotype subtree—does not necessarily hold. When this assumption is violated the tree must necessarily be incorrect.

4.4.2 Calibrating simulation parameters using summary statistics.

We defined several summary statistics on sequences equipped with abundances which were used to calibrate simulation parameters representative of a regime similar to experimental data. We chose these statistics to reflect information relevant to tree inference, but not actually require tree inference, so as to avoid circularity. Denote $g_0 \in \mathbf{G}$ as the naive BCR (root genotype) and $d_H(\cdot, \cdot)$ as the Hamming distance function between two sequences. Given simulation or experimental data $\mathbf{G}$ and $\mathbf{A}$, we characterize the degree of mutation (from naive BCR) in the lineage by the set of Hamming distances of the observed genotypes from the naive genotype: $\{d_H(g, g_o), g \in \mathbf{G}\}$. For a given genotype $g_i \in \mathbf{G}$, we can compute its number of Hamming neighbors in the data $\eta_i = |\{g_j \in \mathbf{G} : d_H(g_i, g_j) = 1\}|$.

A simulation is specified by parameters $(\lambda, \lambda_0, N(\text{or } t), n)$, a mutability model (here S5F [223]), and an initial sequence. We found parameters $(\lambda = 1.5, \lambda_0 = 0.25, N = 100, n = 65)$ produced simulations that were comparable to experimental data under these statistics. The experimental data used for comparison, consisting of 65 total BCR V gene sequences from a single germinal center lineage, is described in the following section. Figure D.2 depicts these summary statistics for 100 simulations, compared to experimental BCR data.

4.4.3 Germinal center BCR sequencing.

Germinal center B cell lineage tracing and B cell receptor sequencing was performed as previously described [206]. Full length IgH and IgL sequences from lymph node 2 germinal centers 1 and 2 from this reference were used for empirical validation results, while V gene sequences only (which are not dependent on `partis`-inferred naive sequences) were used for calibrating simulation parameters.

4.4.4 Bootstrap support.

We computed bootstrap support values for edges on a GCtree extending the standard approach [59]: we resampled columns from the alignment G 100 times with replacement, generating an inferred GCtree (maximum GCtree likelihood among equally parsimonious trees) for each. Each edge is equivalent to a bipartition of observed genotypes obtained by cutting the edge; such a bipartition is typically referred to as a *split*. We compute the number of bootstrapped trees that contain the same split, and annotate the edge with this number. Because resampling the alignment G can produce repeated genotypes, there can exist ambiguity about how to perform genotype collapse of a parsimony tree. We simply group genotypes in the bootstrap analysis that collapse to identical genotypes. For example, if two observed sister genotypes with resampled sequences are both identical in sequence to their mutual parent, both have a claim on collapsing into the parent. When collapsing this tree, both genotypes will be associated with this collapsed node, rather then just a single one.

4.4.5 Data availability.

Germinal center BCR sequence data can be found in Supplementary Database S1 of Tas et al. [206], lymph node 2 and germinal center 1.

4.4.6 Software availability.

The `GCtree` source code is available at `github.com/matsengrp/gctree` and extensive documentation is available at `https://matsengrp.github.io/gctree`. It is open-source software under the GPL v3.

Chapter 5

CONCLUSIONS

5.1 Germline mutation spectrum evolution

Mutation spectrum drift appears to accompany many cases of population divergence, as seen so far in humans, apes, and mice. Despite this intriguing empirical pattern, disentangling the interactions of mutator alleles, changes in life history traits, and environmental mutagens remains almost as challenging today as it was over 50 years ago when Kimura first expressed his exasperation at the task. Understanding the causality of mutation spectrum drift, however, is vital to many of the central questions in population and evolutionary genetics, such as calibrating estimates of demographic and phylogenetic history and understanding historical changes in mutagenic exposures.

Most of this drift consists of small shifts in the relative frequencies of mutations in many trinucleotide contexts (and likely higher order contexts as well), which seems compatible with the random drift of weak, non-specific mutator alleles, subtly tweaking the efficiencies of replication and repair genes. The TCC→TTC pulses in European populations and CTVT cell populations in canines, however, seem fundamentally different from the typical drift pattern in that they are "fast" changes that affected very few sequence contexts and reversed themselves after short periods of time. It is entirely possible that the genetic, environmental, and life history factors are all partially responsible for the evolution of mutation spectra. Further, these drivers of mutation rate evolution are not necessarily independent from one another; e.g., both environmental and genetic factors might influence life history traits or certain alleles might affect an organism's ability to metabolize mutagenic toxins in the environment.

The TCC→TTC pulse in European populations is the most striking example of mutation spectrum drift in humans known thus far, but it is possible that additional population-specific signatures may be found by studying more geographically diverse human populations, or that more pronounced signatures exist but are specific to certain genomic loci rather than genome-wide averages. Similarly, we speculate that exploring patterns of mutation spectrum variation within and between other species is likely to reveal novel signatures that reflect their unique evolutionary histories. As the field continues to move away from the outdated concept of mutation rate as a static constant and towards treating mutation rate like a dynamic, evolvable phenotype, there are many opportunities to incorporate new data sources and sophisticated simulation software [81], and develop new theoretical tools and inference methods to tackle these fundamental questions of genome evolution.

Expanding beyond the allele frequency-based approach in `mushi`, one possible direction to further elucidate the TCC→TTC pulse is to leverage inferred whole genome ancestral recombination graphs for humans and other species. Jointly modeling locus-wise ancestry and ancestry-conditioned mutation spectrum history, we may shed light on the etiology of the TCC→TTC pulse, which has evinced fine-grained ancestry specificity in an ancient DNA study [198].

5.2 Somatic evolutionary dynamics in adaptive immunity

A synthesis of quantitative immunology and evolutionary modeling is needed to clarify how immune repertoires are shaped by evolutionary processes. With the increasingly wide availability of massive immune repertoire data sets, and deep experimental characterization of variant functional effects for both pathogen proteins (like SARS-CoV-2 spike) and host immune receptors (antibodies and T cell receptors), I believe now is the time to be very ambitious on this front. We are extremely well-positioned with experimental characterizations, and there is now considerable scope for bringing a rigorous and modern population genet-

ics perspective to the fantastically complex evolutionary story playing out in the immune system.

Quantitative understanding of germinal center evolutionary dynamics has profound translational potential in advancing antibody engineering and vaccination strategies to elicit broadly neutralizing antibodies. These applications are usually focused on how receptor protein (sequence) determines antigen binding affinity (function), part of a rapidly expanding research area at the nexus of evolutionary biochemistry and deep learning of sequence $\rightarrow$ function maps [224]. However, engineering the evolutionary process in germinal centers will also require us to specify how binding affinity determines fitness and selection in germinal centers: a function $\rightarrow$ fitness map. We need an end-to-end map of sequence $\rightarrow$ function $\rightarrow$ fitness to realize a new generation of engineered immune responses.

Immune repertoires are a natural model system in which to study *function of evolving systems* [22]: their evolution is rapid, multi-scale, coupled to pathogen evolution, and can be densely sampled experimentally; their evolution is replicated, as the same gene segments are used to seed evolving lineages responding to repeated immune exposures; and their evolution can be tuned along an axis of complexity, ranging from simple lab-only model systems, to semi-natural, to fully natural repertoire complexity. Evolutionary modeling and quantitative approaches have a role to play alongside experimental model systems that and are tunable along an axis of complexity. Truly impactful advances in understanding immune repertoire dynamics will require a tight integration of computational, theoretical, and experimental expertise. This is a unique opportunity to develop new and powerful evolutionary inference tools that can advance both evolutionary biology and immunology.

Appendix A

MATHEMATICAL DETAILS FOR CHAPTER 2

A.1 Proof of Theorem 1 : the expected SFS given demographic and mutation intensity histories

Suppose n haplotypes are sampled in the present, and let random vector $\mathbf{T} = [T_2, \ldots, T_n]^\intercal$ denote the coalescent times measured retrospectively from the present, i.e. T_n is the most recent coalescent time, and T_2 is the TMRCA of the sample.

As in Section 3 of [78], we consider a marked Poisson process in which every mutation is assigned a random label drawn iid from the uniform distribution on $(0, 1)$. This is tantamount to the infinite sites assumption, with the unit interval representing the genome, and the random variate labels representing mutant sites. Further suppose that mutation intensity at time t (measured retrospectively from the present in units of Wright-Fisher generations) is a function of time $0 \leq \mu(t) < \infty$ (measured in mutations per genome per generation) applying equally to all lines in the coalescent tree. A given line in the coalescent tree then acquires mutations on a genomic subinterval $(p, p + dp) \subseteq (0, 1)$ at rate $\mu(t)dp$.

Let $\mathcal{E}_{dp,b}$ denote the event that a mutation present in $b \in \{1, 2, \ldots, n-1\}$ haplotypes in the sample occurred within a given genomic interval $(p, p + dp)$. Given the uniform labeling assumption, the probability of this event is independent of p, but the following argument can be generalized to allow the labelling distribution to be nonuniform over the unit interval without changing the result. Let I_k denote the kth intercoalescent time interval, i.e. $I_n = (0, T_n), I_{n-1} = (T_n, T_{n-1}), \ldots, I_2 = (T_3, T_2)$. Let $\mathcal{E}_{dp,b,k}$ denote the event that the mutation $\mathcal{E}_{dp,b}$ occurred during interval I_k (index k here is not to be confused for the k-mer

size in the main text). For small dp and finite $\mu(t)$ we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{dp,b} \mid \mathbf{T}) &= \sum_{k=2}^{n} \mathbb{P}(\mathcal{E}_{dp,b,k} \mid \mathbf{T}) \\ &= \sum_{k=2}^{n} p_{n,k}(b) \left(k\, dp \int_{t \in I_k} \mu(t) dt + O\left((dp)^2\right) \right),\end{aligned}$$

where

$$p_{n,k}(b) \equiv \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \tag{A.1}$$

is the probability that a mutant that arose when there were k ancestral lines of n sampled haplotypes will be present in b of them (see [78], eqn. 1.9). The quantity in parentheses is the probability that a mutation arose during the kth intercoalescent interval in a genomic interval of size dp. Marginalizing $\mathbf{T}$ gives

$$\mathbb{P}(\mathcal{E}_{dp,b}) = dp \sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}_{\mathbf{T}} \left[\int_{t \in I_k} \mu(t) dt \right] + O\left((dp)^2\right).$$

For small dp, each genomic interval $(p, p + dp)$ contains zero or one mutations. Therefore, taking the limit $dp \to 0$ and integrating over the genome, the expected number of mutations subtending b haplotypes (i.e. the bth component of the SFS) is

$$\xi_b = \int_0^1 \mathbb{P}(\mathcal{E}_{dp,b}) = \sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}_{\mathbf{T}} \left[\int_{t \in I_k} \mu(t) dt \right]$$

We now substitute in the bounds of every intercoalescent interval $I_k = (T_{k+1}, T_k)$, giving

$$\begin{aligned}\xi_b &= \sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}_{T_k} \left[\int_0^{T_k} \mu(t) dt \right] - \sum_{k=2}^{n-1} k p_{n,k}(b) \mathbb{E}_{T_{k+1}} \left[\int_0^{T_{k+1}} \mu(t) dt \right] \\ &= \sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}_{T_k} \left[\int_0^{T_k} \mu(t) dt \right] - \sum_{k=3}^{n} (k-1) p_{n,k-1}(b) \mathbb{E}_{T_k} \left[\int_0^{T_k} \mu(t) dt \right] \\ &= \sum_{k=2}^{n} B_{b,k} \mathbb{E}_{T_k} \left[\int_0^{T_k} \mu(t) dt \right],\end{aligned} \tag{A.2}$$

where

$$B_{b,k} \equiv \begin{cases} kp_{n,k}(b), & k = 2 \\ kp_{n,k}(b) - (k-1)p_{n,k-1}(b), & k > 2 \end{cases} \tag{A.3}$$

are combinatorial terms.

Polanski et al. [163], eqns. 5-8, give the marginal density for the coalescent time T_k as

$$\pi_k(t_k) = \sum_{j=k}^{n} A_{k,j} q_j(t_k), \tag{A.4}$$

for $k = 2, \ldots, n$, where $\mathbf{A}$ is an $(n-1) \times (n-1)$ matrix indexed from $2, \ldots, n$ with

$$A_{k,j} \equiv \begin{cases} 1, & k = j = n \\ 0, & j < k, \\ \frac{\prod_{l=k\neq j}^{n} \binom{l}{2}}{\prod_{l=k\neq j}^{n} \left(\binom{l}{2} - \binom{j}{2}\right)}, & \text{otherwise} \end{cases}$$

and

$$q_j(t) \equiv \frac{\binom{j}{2}}{\eta(t)} \exp\left[-\binom{j}{2} \int_0^t \frac{dt'}{\eta(t')}\right],$$

for $j = 2, \ldots, n$, and $\eta(t)$ is the haploid effective population size history. We assume that $0 < \eta(t) < \infty$. Note that $q_j(t)$ is the probability density of the time to the first coalescent event among any subset of j individuals in the present, with inhomogeneous Poisson intensity function $\binom{j}{2}/\eta(t)$.

The expectations in (A.2) can be expressed using (A.4) as

$$\begin{aligned} \mathbb{E}_{T_k}\left[\int_0^{T_k} \mu(t)dt\right] &= \int_0^{\infty} \pi_k(t_k) \int_0^{t_k} \mu(t)dt\, dt_k \\ &= \sum_{j=k}^{n} A_{k,j} \int_0^{\infty} q_j(t_k) \int_0^{t_k} \mu(t)dt\, dt_k \\ &= \sum_{j=k}^{n} A_{k,j} \int_0^{\infty} q_j(t_k) \int_0^{\infty} \mathbb{1}_{[0<t<t_k]} \mu(t)dt\, dt_k \\ &= \sum_{j=k}^{n} A_{k,j} \int_0^{\infty} r_j(t)\mu(t)dt \end{aligned} \tag{A.5}$$

where in the last line we exchange integration order (by Fubini's theorem) and define the inhomogeneous Poisson survival function

$$r_j(t) \equiv \int_0^\infty q_j(t')\mathbb{1}_{[0<t<t']}dt' = \exp\left[-\binom{j}{2}\int_0^t \frac{dt'}{\eta(t')}\right] \tag{A.6}$$

corresponding to density $q_j(t)$.

Using (A.5) in (A.2) gives

$$\begin{aligned}\xi_b &= \sum_{k=2}^n B_{b,k}\sum_{j=k}^n A_{k,j}\int_0^\infty r_j(t)\mu(t)dt \\ &= \sum_{j=2}^n\left(\sum_{k=2}^j B_{b,k}A_{k,j}\right)\int_0^\infty r_j(t)\mu(t)dt,\end{aligned} \tag{A.7}$$

exchanging summation order in the last line. We then have a linear expression for the expected SFS as a function of the mutation intensity history $\mu(t)$:

$$\boldsymbol{\xi} = \mathbf{C}\,\mathbf{d}(\eta,\mu), \tag{A.8}$$

where the $(n-1)\times(n-1)$ matrix $\mathbf{C} = \mathbf{BA}$ is constant in μ and η, and

$$d_j(\eta,\mu) \equiv \int_0^\infty r_j(t)\mu(t)dt = \int_0^\infty \exp\left[-\binom{j}{2}\int_0^t \frac{dt'}{\eta(t')}\right]\mu(t)dt, \tag{A.9}$$

for $j = 1,\ldots,n-1$, is a linear functional of μ and a nonlinear functional of η.

Given the boundedness assumptions that we have on η and μ, we now prove boundedness of the map from joint history functions (η,μ) to expected SFS vectors ξ. The vector $\mathbf{d}(\eta,\mu)$ may be viewed as a nonlinear operator $\mathbf{d} : L^\infty(\mathbb{R}_{\geq 0})\times L^\infty(\mathbb{R}_{\geq 0}) \to \ell^\infty_{n-1}$ of rank $n-1$, and is bounded element-wise. In the diffusion timescale (equation 3 of main text) , d_j is the Laplace transform of the bounded function $\tilde{\eta}\tilde{\mu}$ evaluated at $\binom{j}{2}$, and is thus bounded. In particular,

$$0 \leq d_j \leq \frac{\eta_{\max}\mu_{\max}}{\binom{j}{2}},\ \text{for } j = 1,\ldots,n-1, \tag{A.10}$$

where $\eta_{\max}$ and $\mu_{\max}$ are the respective bounds on η and μ. Boundedness of the full operator mapping (η,μ) to the expected SFS $\boldsymbol{\xi}$ follows from the fact that $\mathbf{C}$ is a matrix with bounded norm. This completes the proof of Theorem 1.

A.2 Computing the elements of C

We next develop an efficient recursive procedure for computing the matrix **C**. Using (A.3)

$$\begin{aligned} C_{b,j} &= \sum_{k=2}^{j} k p_{n,k}(b) A_{k,j} - \sum_{k=3}^{j} (k-1) p_{n,k-1}(b) A_{k,j} \\ &= W_{b,j}^{(1)} - W_{b,j}^{(2)}, \end{aligned}$$

where

$$W_{b,j}^{(1)} \equiv \sum_{k=2}^{j} k p_{n,k}(b) A_{k,j} \tag{A.11}$$

$$W_{b,j}^{(2)} \equiv \sum_{k=3}^{j} (k-1) p_{n,k-1}(b) A_{k,j}. \tag{A.12}$$

Polanski et al. [164], eqn. 11, show that the nonzero entries of **A** can be expressed as

$$A_{k,j} = \frac{n!(n-1)!}{(j+n-1)!(n-j)!} \cdot \frac{(2j-1)}{j(j-1)} \cdot \frac{(j+k-2)!}{(k-1)!(k-2)!(j-k)!} \cdot (-1)^{j-k}.$$

Given the form of $p_{n,k}(b)$ in (A.1), we see that (A.11) and (A.12) are definite sums over hypergeometric terms. We used Zeilberger's algorithm [158, 151], which finds polynomial recurrences for definite sums of hypergeometric terms, to procedurally generate the following second-order recursions in j:

$$\begin{aligned} W_{b,2}^{(1)} &= \frac{6}{(n+1)} \\ W_{b,3}^{(1)} &= \frac{10(5n-6b-4)}{(n+2)(n+1)} \\ W_{b,j+2}^{(1)} &= -\Big[(2j+3)\big(-(2j-1)W_{b,j+1}^{(1)}\big(2j(j+1)\big(b^2\big(j^2+j-2\big)-6b-j(j+1)-2\big) \\ &\quad - j(j+1)n\big(3b\big(j^2+j+2\big)+j^2+j-2\big)+\big(j(j+1)\big(j^2+j+6\big)+4\big)n^2+4n\big) \\ &\quad - (j-1)(j+1)^2(j-n)W_{b,j}^{(1)}(4(n+1)-j(j+2)(b-n-1))\big)\Big] \\ &\quad \Big/\Big[j^2(j+2)(2j-1)(j+n+1)\left(-bj^2+b+\left(j^2+3\right)(n+1)\right)\Big] \end{aligned}$$

and

$$W^{(2)}_{b,2} = 0$$
$$W^{(2)}_{b,3} = \frac{20(n-2)}{(n+1)(n+2)}$$
$$W^{(2)}_{b,j+2} = \frac{(2j+3)(j-n+1)}{j}\left(\frac{(j+1)}{(2j-1)(j+n)}W^{(2)}_{b,j} - \frac{(j(j+1)(2b-n+1)-2(n+1))}{(j-1)(j+2)(j-n)(j+n+1)}W^{(2)}_{b,j+1}\right).$$

These formulae are used to numerically compute the entries in $\mathbf{C}$.

A.3 Discretization of history functions and computation of $\mathbf{d}(\eta, \mu)$

We represent histories as piecewise constant functions of time on m pieces $[t_0, t_1), [t_1, t_2), \ldots, [t_{m-1}, t_m)$, where $0 = t_0 < t_1 < \cdots < t_{m-1} < t_m = \infty$. The grid is common to $\eta(t)$ and $\mu(t)$. We take the boundaries of the pieces as fixed parameters and in practice use a logarithmically-spaced dense grid of hundreds of pieces to approximate infinite-dimensional histories. Let column vector $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$ denote the constant population size $\eta(t)$ during each piece, and let $\mathbf{w} = [w_1, \ldots, w_m]^\intercal$ denote the constant mutation rate $\mu(t)$ during each piece.

With this we can follow the proof of Proposition 1 in [175], *mutatis mutandis*, with our (A.9) to arrive at

$$\mathbf{d} = \mathbf{M}(\mathbf{y})\mathbf{w} \tag{A.13}$$

where

$$\mathbf{M}(\mathbf{y}) \equiv \begin{bmatrix} 1 & & & \\ & \frac{1}{3} & & \\ & & \ddots & \\ & & & \frac{1}{\binom{n}{2}} \end{bmatrix} \begin{bmatrix} 1 & u_1 & \ldots & \prod_{i=1}^{m-1} u_i \\ 1 & u_1^3 & \ldots & \prod_{i=1}^{m-1} u_i^3 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & u_1^{\binom{n}{2}} & \ldots & \prod_{i=1}^{m-1} u_i^{\binom{n}{2}} \end{bmatrix} \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \operatorname{diag}(\mathbf{y}), \tag{A.14}$$

and $u_l \equiv \exp(-(t_l - t_{l-1})/y_l)$ for $l = 1, \ldots, m$. Note that the $(n-1) \times m$ matrix $\mathbf{M}(\mathbf{y})$ is a nonlinear function of the demographic history $\mathbf{y}$ because the u_l are nonlinear functions of $\mathbf{y}$.

This reflects the fact that it is a discretization of the nonlinear operator $\mathbf{d}(\cdot, \mu)$. Combining (A.13) with (A.8) gives the discretized forward model

$$\boldsymbol{\xi} = \mathbf{L}(\mathbf{y})\mathbf{w}, \tag{A.15}$$

where $\mathbf{L}(\mathbf{y}) \equiv \mathbf{CM}(\mathbf{y})$.

A.4 Proof of Proposition 1

The distribution of independent Poisson random variables factorizes into an aggregate Poisson random variable and a multinomial, a well-known result often called "Poissonization" [38]. Poissonization over the mutation type index j gives

$$\begin{aligned}\mathbb{P}(\mathbf{X} \mid \boldsymbol{\Xi}) &= \prod_{i=1}^{n-1}\prod_{j=1}^{K} \underbrace{\mathbb{P}(X_{i,j} \mid \Xi_{i,j})}_{\text{Poisson}} = \prod_{i=1}^{n-1} \underbrace{\mathbb{P}(x_i \mid \xi_i)}_{\text{Poisson}} \underbrace{\mathbb{P}\left([X_{i,1}, \ldots, X_{i,K}] \;\mid\; x_i, \left[\frac{\Xi_{i,1}}{\xi_i}, \ldots, \frac{\Xi_{i,K}}{\xi_i}\right]\right)}_{\text{multinomial}} \\ &= \underbrace{\mathbb{P}(\mathbf{x} \mid \boldsymbol{\xi})}_{\text{PRF}} \quad \underbrace{\mathbb{P}(\mathbf{X} \mid \mathbf{x}, \hat{\boldsymbol{\Xi}})}_{\substack{\text{multinomial random} \\ \text{field over } i}} \quad . \end{aligned} \tag{A.16}$$

This completes the proof of Proposition 1.

A.5 Proof of Lemma 1

Fix the mutation type i, and consider the multinomial over j

$$\mathbb{P}\left([X_{i,1}, \ldots, X_{i,K}] \;\mid\; x_i, \left[\frac{\Xi_{i,1}}{\xi_i}, \ldots, \frac{\Xi_{i,K}}{\xi_i}\right]\right).$$

We must show that any element of the multinomial vector

$$\hat{\Xi}_{i,j} \equiv \frac{\Xi_{i,j}}{\xi_i}$$

can be formulated without reference to η. From elementary properties of the multinomial, the conditional expection value of $X_{i,j}$ given x_i is

$$\mathbb{E}[X_{i,j} \mid x_i] = x_i \; \hat{\Xi}_{i,j}.$$

Now, since mutation events are independent we perform a thinning operation on each of the x_i mutation events

$$\mathbb{E}[X_{i,j} \mid x_i] = x_i \, \mathbb{P}\,(\text{a mutation of sample frequency } i \text{ is of type } j) \tag{A.17}$$

$$= x_i \int_0^\infty \frac{\tilde{\mu}_j(\tau)}{\mu_0} a_i(\tau) d\tau, \tag{A.18}$$

where $a_i(\tau)$ is the pdf of a mutation's age τ measured in expected coalescent events (diffusion time) conditioned on its sample frequency i. So

$$\hat{\Xi}_{i,j} = \int_0^\infty \frac{\tilde{\mu}_j(\tau)}{\mu_0} a_i(\tau) d\tau.$$

This is independent of η by definition of the diffusion time scale as the intensity measure of the coalescent process. This completes the proof of Lemma 1.

A.6 Tempora incognita: observability toward the coalescent horizon

The time-domain singular vectors of $\mathcal{L}(\eta)$ form an eigenbasis for solutions $\mu(t)$ that are possible, in principle, to reconstruct from the SFS. However, sampling noise about the expected SFS will corrupt information from singular vectors that are associated to smaller singular values. These corrupted components will be the directions in solution space associated with higher frequency and less smooth dynamics. Since the singular values of $\mathcal{L}(\eta)$ have a very large dynamic range (the condition number is large), the presence of noise will limit reconstruction to smoother, more slowly varying components that are least corrupted and erase information about more sudden events.

Figure B.9 depicts the observability of mutation rate history via spectral analysis of $\mathcal{L}(\eta)$ for a case with $\eta(t)$ a simple bottleneck history. From (A.4) and (A.6) in Appendix A.1, the CDF of the TMRCA can be computed given $\eta(t)$. We see only the top few components (ranked by singular value) persist at times older than the bottleneck, and all components vanish beyond the TMRCA of the sample. Higher frequency behavior becomes more difficult

to observe if it is older than the bottleneck, concretely illustrating how demographic events erase information about population history.

A.7 Modeling ancestral state misidentification

Computing the SFS and the k-SFS from variant data requires polarizing reference and alternative alleles to ancestral and derived alleles. Ancestral states are themselves inferred (usually by invoking parsimony criteria in a comparison to an outgroup reference genome, or a larger multi-species alignment), so ancestral state misidentification is expected at some fraction of sites. Misidentification results in allele frequency complementation: a variant at sample frequency i out of n sampled haplotypes will appear to have frequency $n - i$. Misidentification also results in mutation type reversion: e.g. a variant of triplet mutation type TCC→TTC will appear of mutation type GAA→GGA.

Under very general conditions, the expected SFS $\boldsymbol{\xi}$ is a non-increasing vector: $\xi_i \geq \xi_j \ \forall i < j$ [177]. This result covers all demographic histories $\eta(t)$. Given the pointwise nonidentifiability of η and μ (equation 3 of main text) , it also covers all mutation rate histories $\mu(t)$, so all columns of the expected k-SFS are non-increasing row-wise.

Ancestral state misidentification violates this non-increasing expectation result. The SFS for the subset of misidentified sites is reflected in frequency, so the SFS $\boldsymbol{\xi}$ becomes a sum of a non-increasing vector (for the correctly identified sites) and a non-*de*creasing vector (for the misidentified sites). This contributes to the so-called "smile" in empirical SFS data. Because this smile can't be explained by η and μ with a model that produces a non-increasing SFS, the misidentification rate r is identifiable as an additional parameter. Let ξ' denote the expected SFS with misidentification, so

$$\xi'_i = (1 - r)\xi_i + r\xi_{n-i}, \quad i = 1, \ldots, n - 1.$$

In matrix form this is

$$\boldsymbol{\xi}' = (1 - r)\boldsymbol{\xi} + r\boldsymbol{J}\boldsymbol{\xi},$$

where $\boldsymbol{J}$ denotes the $(n-1) \times (n-1)$ exchange matrix (with 1s on the anti-diagonal and 0s elsewhere).

For the k-SFS, misidentified sites contribute to counts in the reflected frequency row, and also in a different mutation type column, corresponding to the revertant mutation type. Let $\pi(j)$ denote the revertant partner of mutation type index j (π is a permutation of the mutation type columns). Let r_j denote the misidentification rate of mutation type j. Then the expected k-SFS with misidentification is

$$\Xi'_{i,j} = (1 - r_j)\Xi_{i,j} + r_j \Xi_{n-i,\pi(j)}, \quad i = 1, \ldots, n-1, \quad j = 1, \ldots, K.$$

In matrix form this is

$$\boldsymbol{\Xi}' = \boldsymbol{\Xi}\ (\boldsymbol{I} - \mathrm{diag}(\boldsymbol{r})) + \boldsymbol{J}\ \boldsymbol{\Xi}\ \boldsymbol{P}_\pi^\intercal\ \mathrm{diag}(\boldsymbol{r}),$$

where $\boldsymbol{P}_\pi$ is the permutation matrix corresponding to π.

We infer the total misidentification rate r jointly with η inference, then infer the rates for each mutation type $\boldsymbol{r}$ jointly with $\boldsymbol{\mu}$, constraining compositionally such that $\sum_{j=1}^{K} f_j r_j = r$ via the isometric log ratio transform, where f_j is the fraction of variants from each mutation type (the column sums of the k-SFS normalized by its total). These additional parameters allow us to obtain very good fits to empirical SFS and k-SFS data that include prominent high frequency "smiles".

Appendix B

SUPPLEMENTARY FIGURES FOR CHAPTER 2

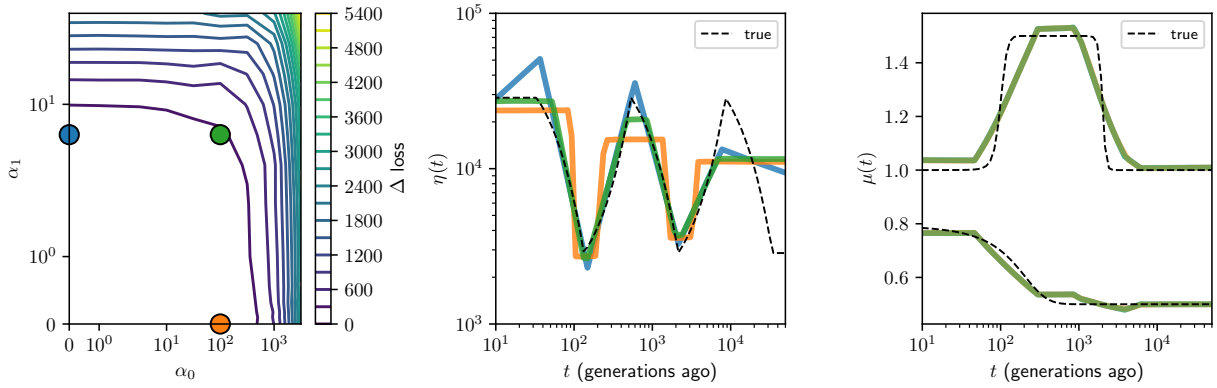

Figure B.1: The effect of demographic model selection on MuSH inference in our simulation study. Left panel shows change in loss (goodness of fit) as a function of 0-th order and 1st order trend penalties. Middle panel shows the inferred demography at each indicated penalty value (colors corresponding to points in left panel). Right panel shows the two variable components of subsequent MuSH inference for each demographic model, indicating they are very weakly effected by demographic model selection.

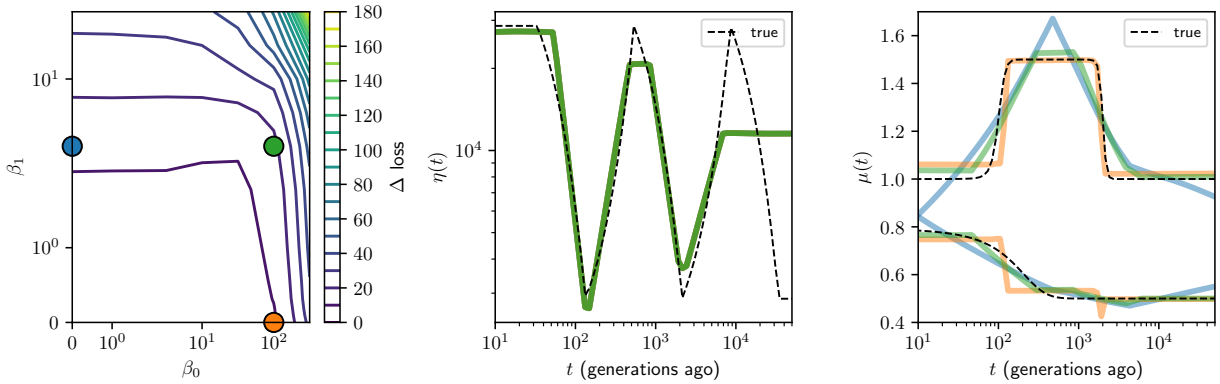

Figure B.2: The effect of MuSH regularization on MuSH inference in our simulation study. Left panel shows change in loss (goodness of fit) as a function of 0-th order and 1st order trend penalties. Middle panel shows the inferred demography (which is independent of the MuSH hyperparameters). Right panel shows the two variable components of the subsequent MuSH inference for each hyperparameter choice (colors corresponding to points in left panel).

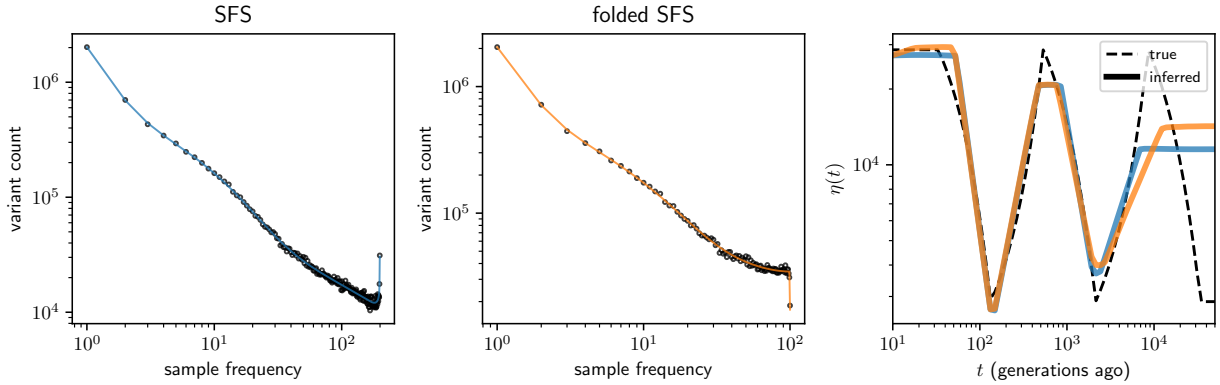

Figure B.3: Comparison of demographic inference using the folded Vs unfolded SFS in our simulation study. Using the same parameters as in Figure 2.2 , we fit the SFS (left), the folded SFS (middle), and show the resulting demographic histories are similar (right).

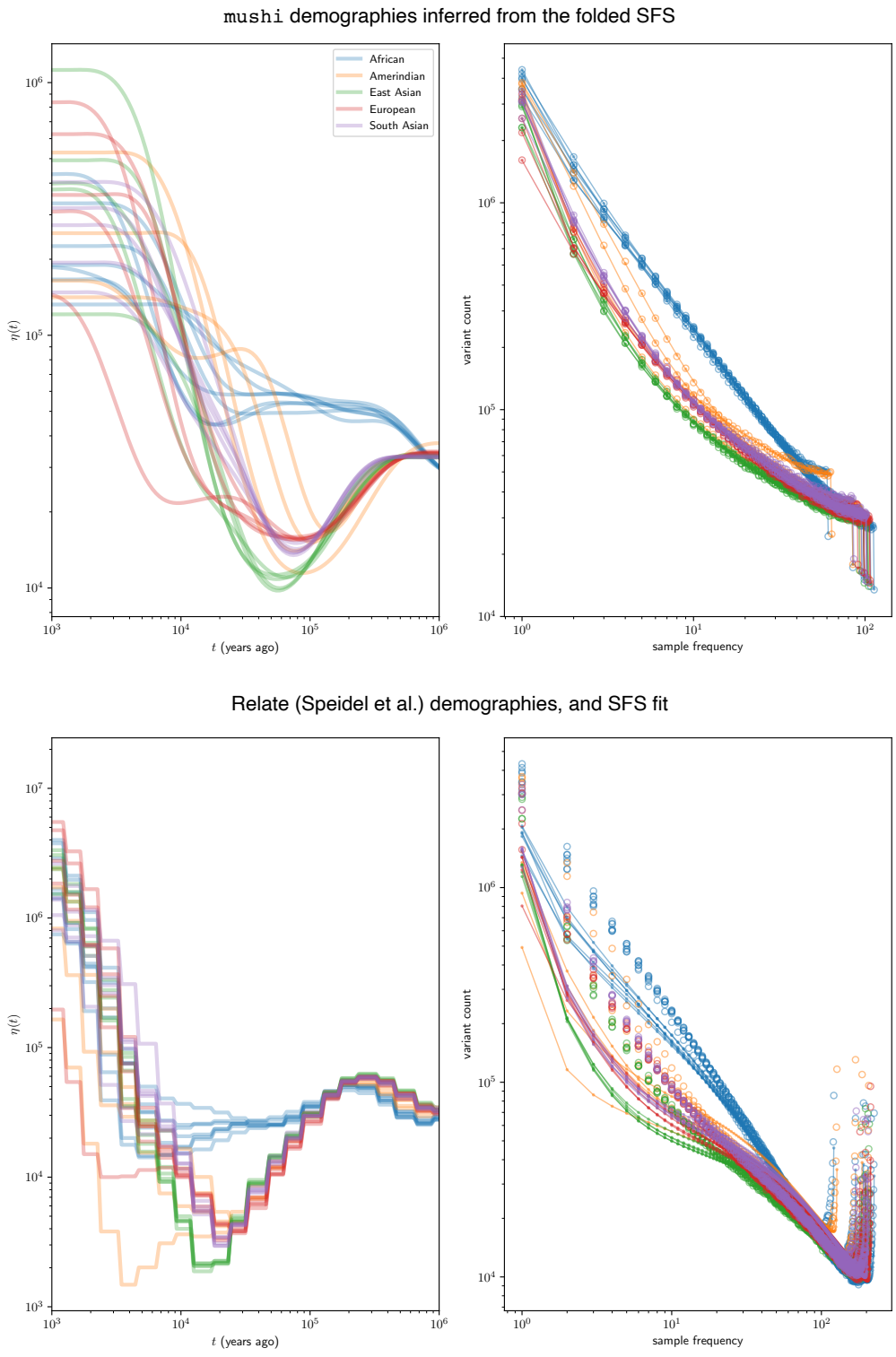

Figure B.4: Other effective population size histories for 1000 Genomes Project populations. Top panels show demographic histories estimated with `mushi` from high coverage data using the folded SFS, and the fit to the folded SFS, with no ancestral state polarization. Bottom panels show demographic histories inferred with Relate, as reported by Speidel et al. [199] (Fig S6 there), with the SFS fit displayed.

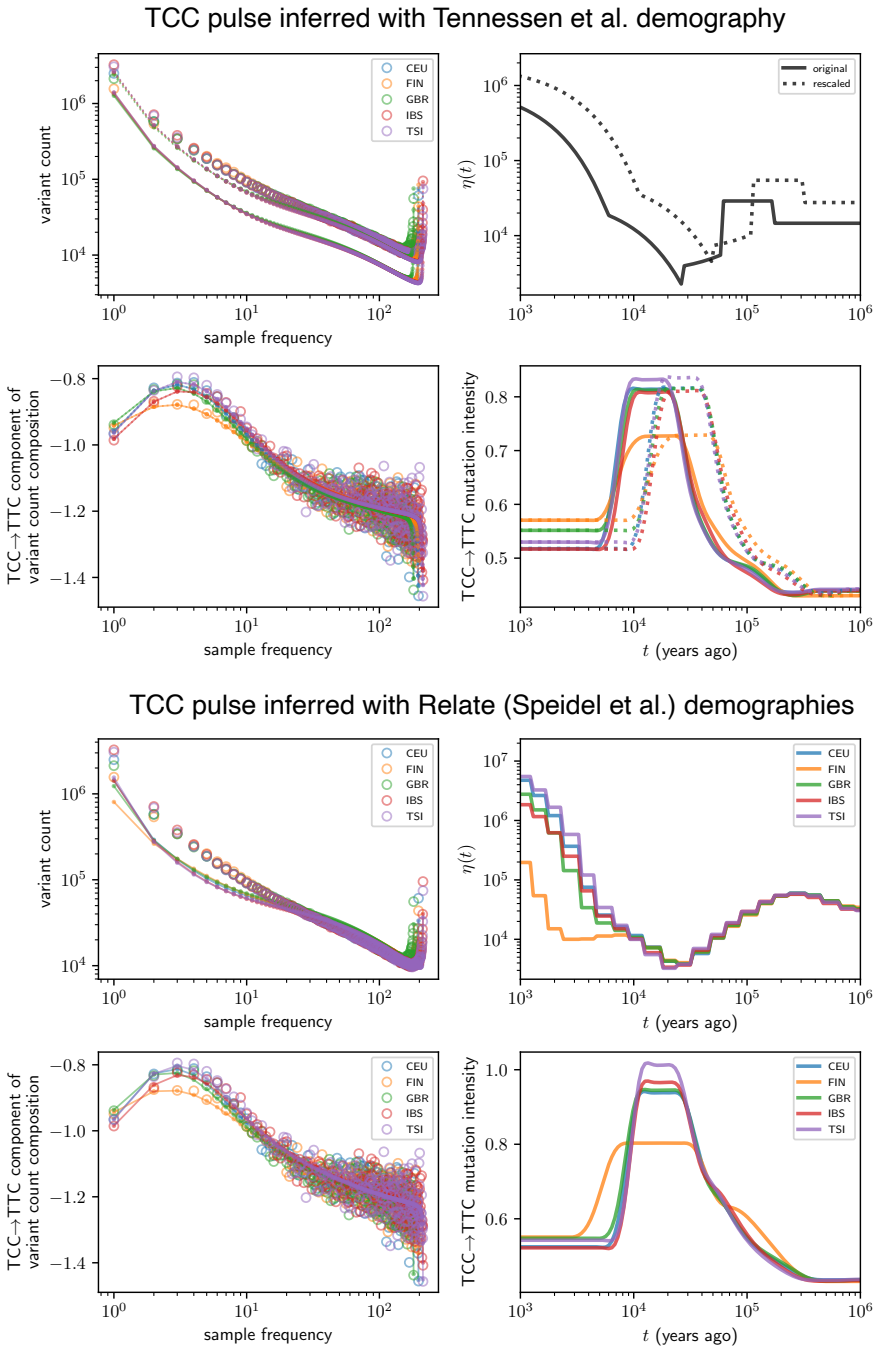

Figure B.5: Timing of TCC→TTC pulse in European populations conditioned on different demographies. Top panels show the SFS fit for each European population under the demographic history of Tennessen et al. [208], which was used by Harris and Pritchard to time the TCC→TTC pulse [83] (similar to Figure 3g), and the subsequent `mushi` fit to the TCC→TTC component of the k-SFS (similar to Figure 4). Dotted lines use a rescaled demography that accounts for the mutation rate difference between the Tennessen demographic inference and more recent de novo rate, resulting in TCC→TTC pulse that is shifted older, and better SFS fit. Bottom panels are similar, but use the European demographic histories inferred by Speidel et al. [199] using the Relate method.

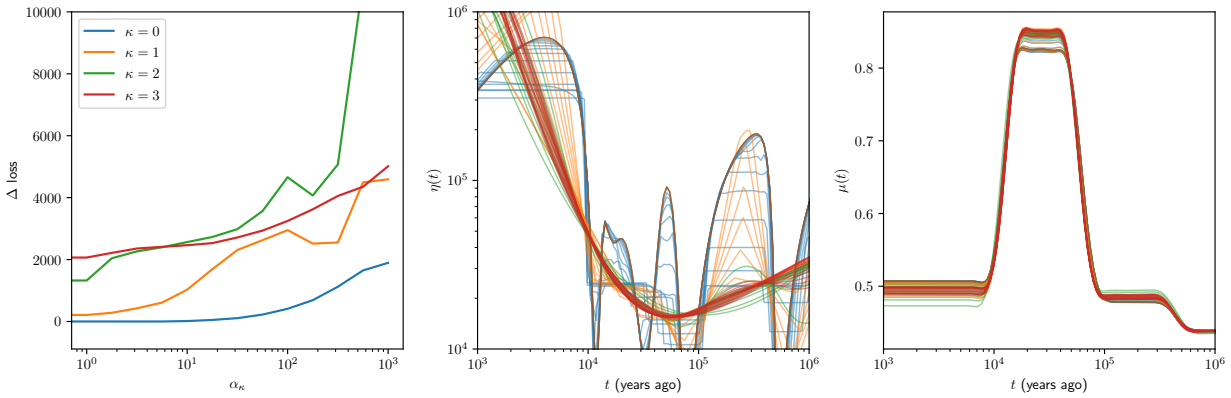

Figure B.6: The effect of demographic model selection on TCC→TTC pulse inference for CEU. Left panel shows change in loss (goodness of fit) as trend penalties of various order increase. Middle panel shows the inferred demography at each penalty value (colors corresponding to trend orders in left panel). Right panel shows the TCC→TTC component of subsequent MuSH inference for each demographic model.

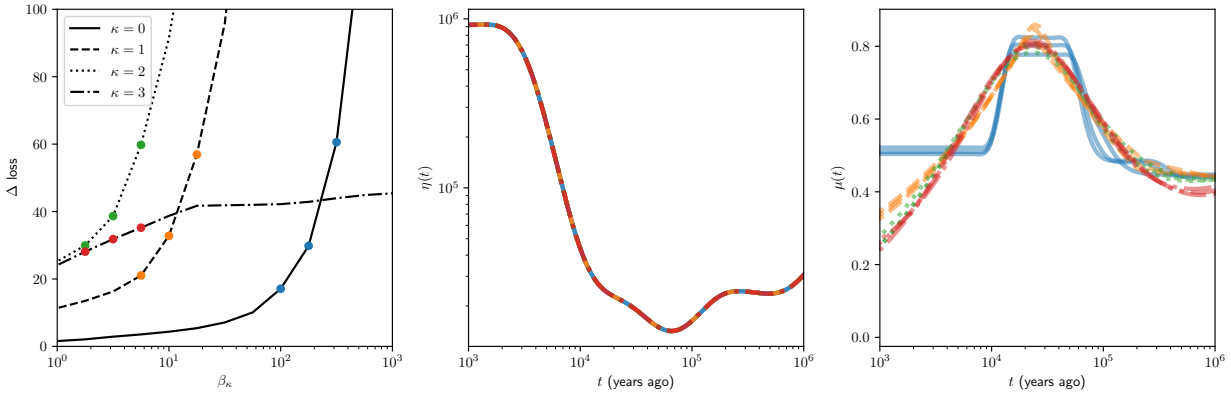

Figure B.7: The effect of MuSH regularization on TCC→TTC pulse inference for CEU. Left panel shows change in loss (goodness of fit) as the κ-th order trend penalty increases, for $\kappa = 0, 1, 2, 3$. Middle panel shows the inferred demography (which is independent of the MuSH hyperparameters). Right panel shows the TCC→TTC component of subsequent MuSH inference, with color corresponding to points in left panel.

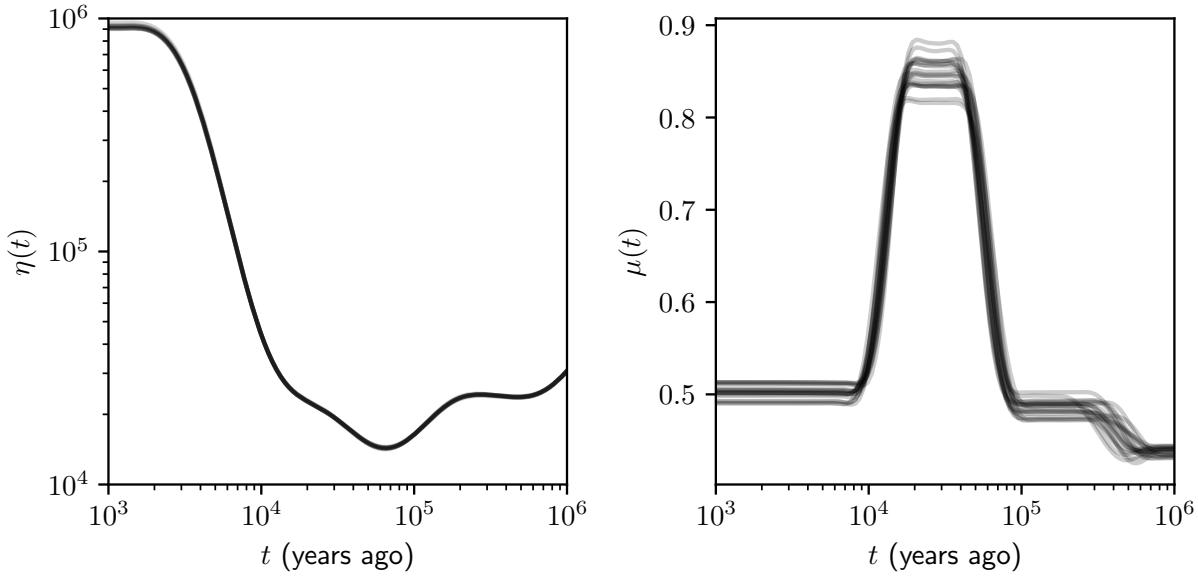

Figure B.8: Bootstrap for CEU demography and TCC→TTC pulse inference. This indicates the stability of inference under replicate data. Note that, because the penalized likelihood inference is strongly biased, this cannot be interpreted as providing confidence bounds of the histories.

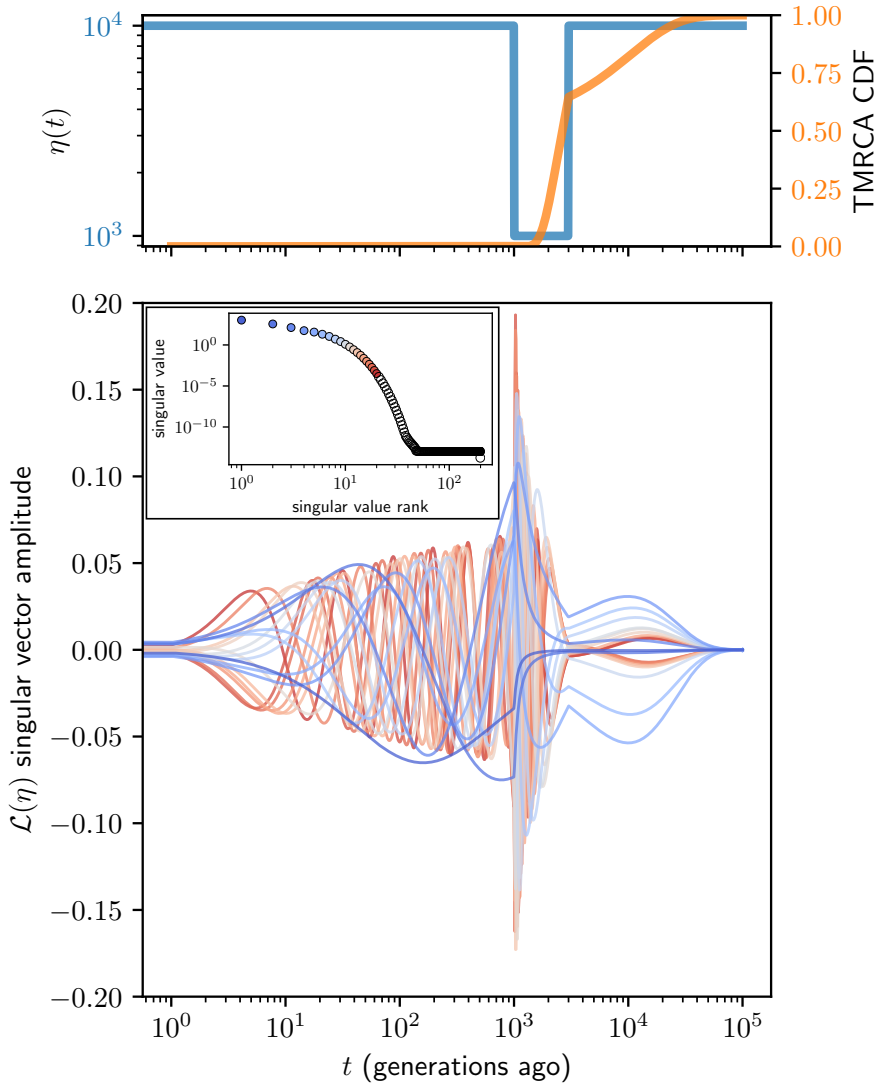

Figure B.9: Observability of mutation rate history via spectral analysis of $\mathcal{L}(\eta)$ for the case of a bottleneck history. The top panel plots demographic history with a bottleneck from about 3000 to 1000 generations ago (blue, left ordinate), and TMRCA CDF (orange, right ordinate). The bottom panel plots the top 20 time domain singular vectors, with the inset showing the corresponding ranked singular values. Time was discretized with a logarithmic grid of 1000 points, and $n = 200$ sampled haplotypes were assumed.

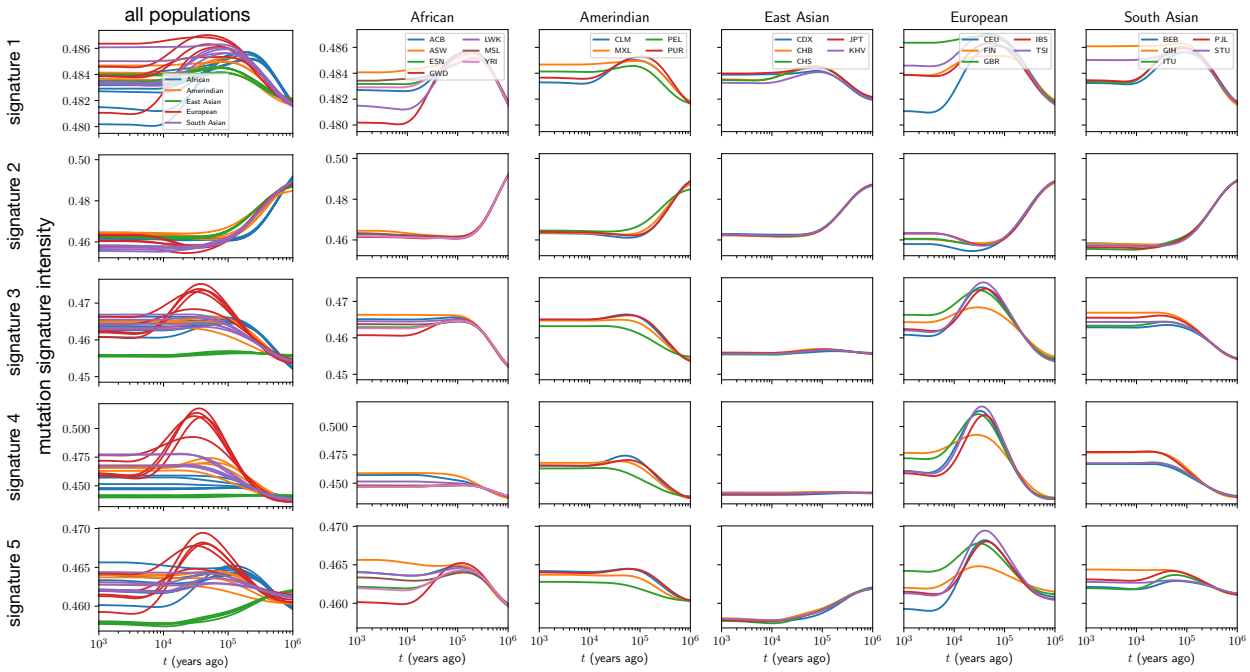

Figure B.10: Mutation signature history for each 1000 Genome Projection super-population. Same results as in main text, but plotted here separately for each super-population.

Appendix C

MATHEMATICAL DETAILS FOR CHAPTER 4

C.1 An empirical Bayes framework for incorporating genotype abundance in phylogenetic optimality.

Here we more fully develop the empirical Bayes perspective on our estimator for the model depicted in Figure 4.2a. This graphical model implies the factorization

$$\mathbb{P}(\mathbf{G}, \mathbf{A}, \mathbf{T}, \boldsymbol{\theta}) = \mathbb{P}(\mathbf{G} \mid \mathbf{T})\, \mathbb{P}(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta})\, \mathbb{P}(\boldsymbol{\theta}). \tag{C.1}$$

A hierarchical Bayes treatment would assign a prior $\mathbb{P}(\boldsymbol{\theta})$ (such as uniform over the unit square for the model $\boldsymbol{\theta} = (p, q)$) and compute the posterior over trees conditioned on the data, marginalizing over $\boldsymbol{\theta}$:

$$\begin{aligned}
\mathbb{P}(\mathbf{T} \mid \mathbf{G}, \mathbf{A}) &= \int d\boldsymbol{\theta}\ \mathbb{P}(\mathbf{T}, \boldsymbol{\theta} \mid \mathbf{G}, \mathbf{A}) \\
&= \int d\boldsymbol{\theta}\ \frac{\mathbb{P}(\mathbf{G}, \mathbf{A}, \mathbf{T}, \boldsymbol{\theta})}{\mathbb{P}(\mathbf{G}, \mathbf{A})} \\
&\propto \mathbb{P}(\mathbf{G} \mid \mathbf{T}) \int d\boldsymbol{\theta}\ \mathbb{P}(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta})\, \mathbb{P}(\boldsymbol{\theta}).
\end{aligned}$$

Rather then attempting this integral over $\mathbb{P}(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta})$, each evaluation of which requires dynamic programming, we first seek a maximum likelihood estimate for $\boldsymbol{\theta}$ marginalizing $\mathbf{T}$:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \mathbb{P}(\mathbf{G}, \mathbf{A} \mid \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{T}} \mathbb{P}(\mathbf{G}, \mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{T}} \mathbb{P}(\mathbf{G} \mid \mathbf{T})\, \mathbb{P}(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta}).
\end{aligned} \tag{C.2}$$

Using this point estimate, an approximate posterior over trees is

$$\mathbb{P}\left(\mathbf{T} \mid \mathbf{G}, \mathbf{A}, \hat{\boldsymbol{\theta}}\right) \propto \mathbb{P}\left(\mathbf{G} \mid \mathbf{T}\right) \mathbb{P}\left(\mathbf{A}, \mathbf{T} \mid \hat{\boldsymbol{\theta}}\right). \tag{C.3}$$

This formulation embodies an optimality over trees conditioned on both genotype sequence data $\mathbf{G}$ and genotype abundance data $\mathbf{A}$. Evaluation of $\hat{\boldsymbol{\theta}}$ with (C.2) in general requires summation over the space of all trees consistent with the data.

A simple application of this formalism is to augment parsimony-based tree optimality with abundance data. Let $\mathcal{T}_{\mathbf{G}}$ denote the degenerate set of maximally parsimonious trees given $\mathbf{G}$ (each of which has the same total genotype sequence distance over its edges). Encode parsimony optimality as a $\mathbb{P}\left(\mathbf{G} \mid \mathbf{T}\right)$ assigning uniform weight to each tree in $\mathcal{T}_{\mathbf{G}}$, and zero elsewhere. In this case, (4.2) becomes

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{T} \in \mathcal{T}_{\mathbf{G}}} \mathbb{P}\left(\mathbf{A}, \mathbf{T} \mid \boldsymbol{\theta}\right), \tag{C.4}$$

and (C.3) becomes

$$\mathbb{P}\left(\mathbf{T} \mid \mathbf{G}, \mathbf{A}, \hat{\boldsymbol{\theta}}\right) \propto \begin{cases} \mathbb{P}\left(\mathbf{A}, \mathbf{T} \mid \hat{\boldsymbol{\theta}}\right), & t \in \mathcal{T}_g \\ 0, & t \notin \mathcal{T}_g \end{cases}. \tag{C.5}$$

With (C.5), we have a framework using abundance information to distinguish among the otherwise equally optimal trees presented by a parsimony analysis. In our application, we use a subcritical infinite-type binary Galton-Watson branching process model for the lineage tree, and describe a recursion for computing GCtree likelihoods $\mathbb{P}\left(\mathbf{A}, \mathbf{T} \mid \hat{\boldsymbol{\theta}}\right)$ by dynamic programming to marginalize over compatible lineage trees.

C.2 Dynamic programming to marginalize lineage tree structure.

We derive a recurrence for $f_{a\tau}(p, q) = \mathbb{P}\left(A_i = a, T_i = \tau \mid \boldsymbol{\theta} = (p, q)\right)$ by marginalizing over the outcome $\{C, M\}$ of the branching event at the root of the lineage subtree for genotype i (the first cell of type i). We will use that a and τ are the sum over two iid processes for the

left and right clonal branches. We temporarily suppress the parameters $\boldsymbol{\theta} = (p, q)$, writing $f_{a\tau}$ for notational compactness. In the case $\{C = 2, M = 0\}$,

$$\mathbb{P}\left(A_i = a, T_i = \tau \mid C = 2, M = 0\right) = \sum_{a'=0}^{a} \sum_{\tau'=0}^{\tau} f_{a'\tau'} f_{a-a',\tau-\tau'}. \tag{C.6}$$

As this is the convolution of $f_{a\tau}$ with itself, we denote it as $f^{*2}_{a\tau}$. Marginalizing over all outcomes $\{C, M\}$, we have

$$\begin{aligned}
f_{a\tau} &= \sum_{(c,m)\in\mathbb{N}^2} \mathbb{P}\left(A_i = a, T_i = \tau \mid C = c, M = m\right) \mathbb{P}\left(C = c, M = m\right) \\
&= \delta_{a1}\delta_{\tau 0}(1-p) + f^{*2}_{a\tau} p(1-q)^2 + (1-\delta_{\tau 0}) f_{a,\tau-1} 2pq(1-q) + \delta_{a0}\delta_{\tau 2} pq^2 \\
&= \begin{cases}
0 & a = 0, \tau = 0, 1, \\
(1-p) & a = 1, \tau = 0, \\
pq^2 & a = 0, \tau = 2, \\
f^{*2}_{a0} p(1-q)^2 & a > 1, \tau = 0, \\
f_{a,\tau-1} 2pq(1-q) + f^{*2}_{a\tau} p(1-q)^2 & \text{otherwise},
\end{cases}
\end{aligned} \tag{C.7}$$

where $\delta_{..}$ denotes the Kronecker delta function. In light of the first case, the convolutional square may be written as

$$f^{*2}_{a\tau} = \sum_{(a',\tau')\notin\{(0,0),(a,\tau)\}} f_{a'\tau'} f_{a-a',\tau-\tau'},$$

showing that there are no terms containing $f_{a\tau}$ on the RHS of (C.7). The GCtree node likelihood $f_{a\tau}$ is thus amenable to computation by straightforward dynamic programming.

C.3 The GCtree likelihood factorizes by genotype.

We argue that the joint distribution over all nodes in a GCtree factorizes by genotype (Figure 4.2d):

$$\mathbb{P}\left(A_1 = a_1, T_1 = \tau_1, \ldots, A_N = a_N, T_N = \tau_N\right) = \prod_{i=1}^{N} f_{a_i\tau_i}. \tag{C.8}$$

Since τ_1 is the number of children of node 1 (the root node), the children of the root node are indexed in level order by $2, \ldots, 1+\tau_1$. Let Λ_i denote the set of indices of the nodes of the subtree rooted at node i, so $\Lambda_2, \ldots, \Lambda_{1+\tau_1}$ refer to sister subtrees rooted on each of the τ_1 children of the root. Using the definition of conditional probability, and since sister subtrees are independent, we have

$$\begin{aligned}\mathbb{P}(a_1, \tau_1, \ldots, a_N, \tau_N) &= \mathbb{P}(a_2, \tau_2, \ldots, a_{N,N} \mid a_1, \tau_1)\, \mathbb{P}(a_1, \tau_1) \\ &= f_{a_1\tau_1} \prod_{i=1}^{1+\tau_1} \mathbb{P}(\{(a_j, \tau_j) : j \in \Lambda_i\}),\end{aligned}$$

where random variable notation has been dropped for notational compactness. Now, within each subtree factor we may reindex in level order (that is, level order in that subtree) starting from 1. We then pull out factors $f_{a_2\tau_2}, \ldots, f_{a_{1+\tau_1}\tau_{1+\tau_1}}$ corresponding to the root nodes of the sister subtrees (children of the original root). We obtain (C.8) by applying this logic recursively. Restoring the offspring distribution parameters, we recognize this as the distribution needed in (4.1) and (4.2) to rank trees in a parsimony forest:

$$\mathbb{P}(\mathbf{T} = (\tau_1, \ldots, \tau_N), \mathbf{A} = (a_1, \ldots, a_N) \mid \boldsymbol{\theta} = (p, q)) = \prod_{i=1}^{N} f_{a_i\tau_i}(p, q), \tag{C.9}$$

where $f_{a_i\tau_i}(p, q)$ is computed by dynamic programming using (C.7).

Numerical validation of the GCtree likelihood is summarized in Figure D.3 using 10,000 Galton-Watson process simulations at each of several parameter values. The likelihood accurately recapitulates tree frequencies, and simulation parameters are recoverable by numerical maximum likelihood estimation.

Appendix D

SUPPLEMENTARY FIGURES FOR CHAPTER 4

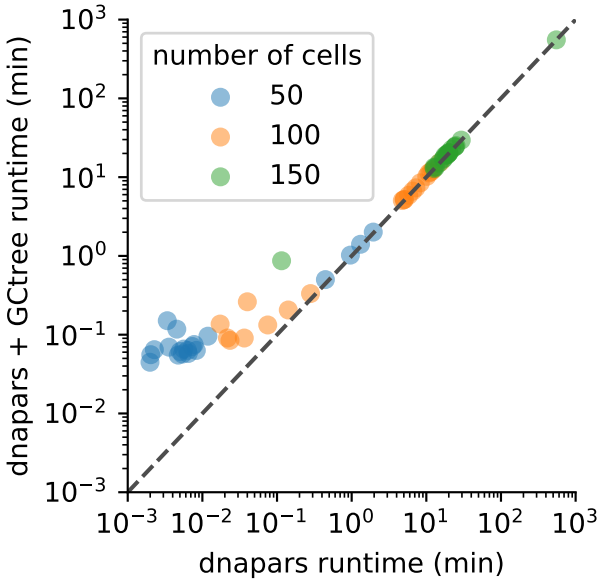

Figure D.1: Runtime experiments. Runtime for generating parsimony trees with `dnapars` and ranking using `GCtree` are shown. Fixed simulation parameters were $\lambda = 1.5$, $\lambda_0 = .25$, and 20 simulations were performed at each value for the number of cells ($N = 50$, $N = 100$, $N = 150$). These runtime experiments were performed on a laptop with a 2.9GHz CPU and 16GB RAM.

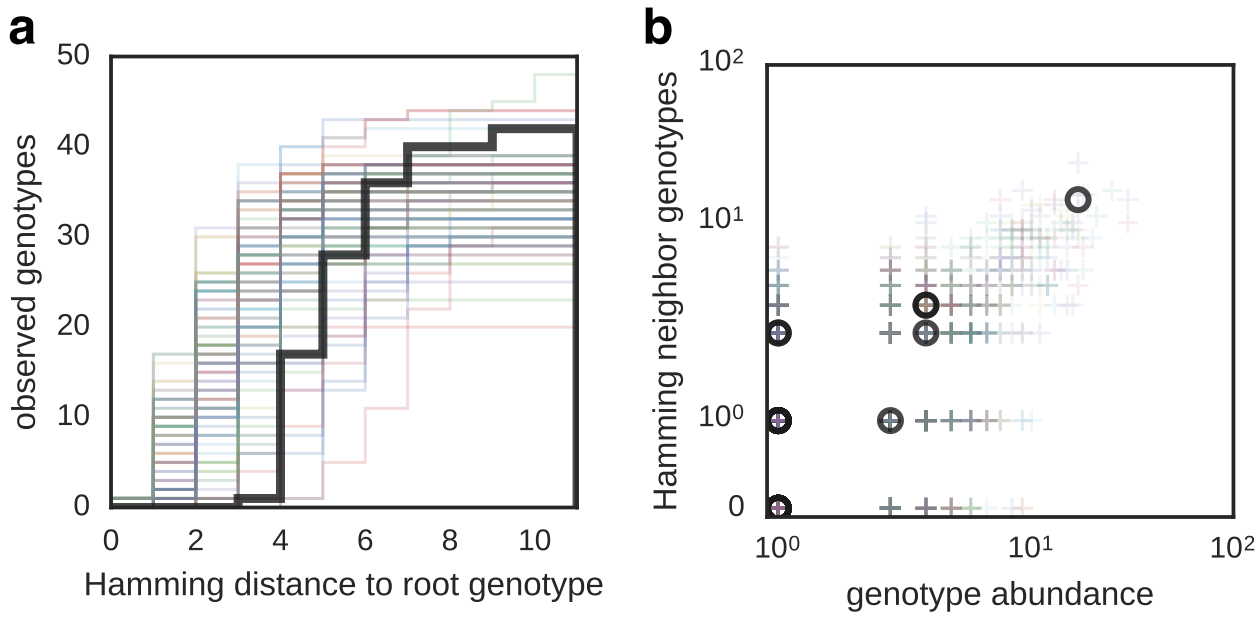

Figure D.2: Simulation summary statistics. simulation parameters: $\lambda = 1.5$, $\lambda_0 = .25$, $N = 100$, $n = 65$. **(a.)** The empirical CDF over genotypes of Hamming distance to the naive genotype for 100 simulations (colors) and germinal center BCR data (black). **(b.)** The distribution over genotypes of number of Hamming neighbors and genotype abundance for 100 simulations (colors) and germinal center BCR data (black).

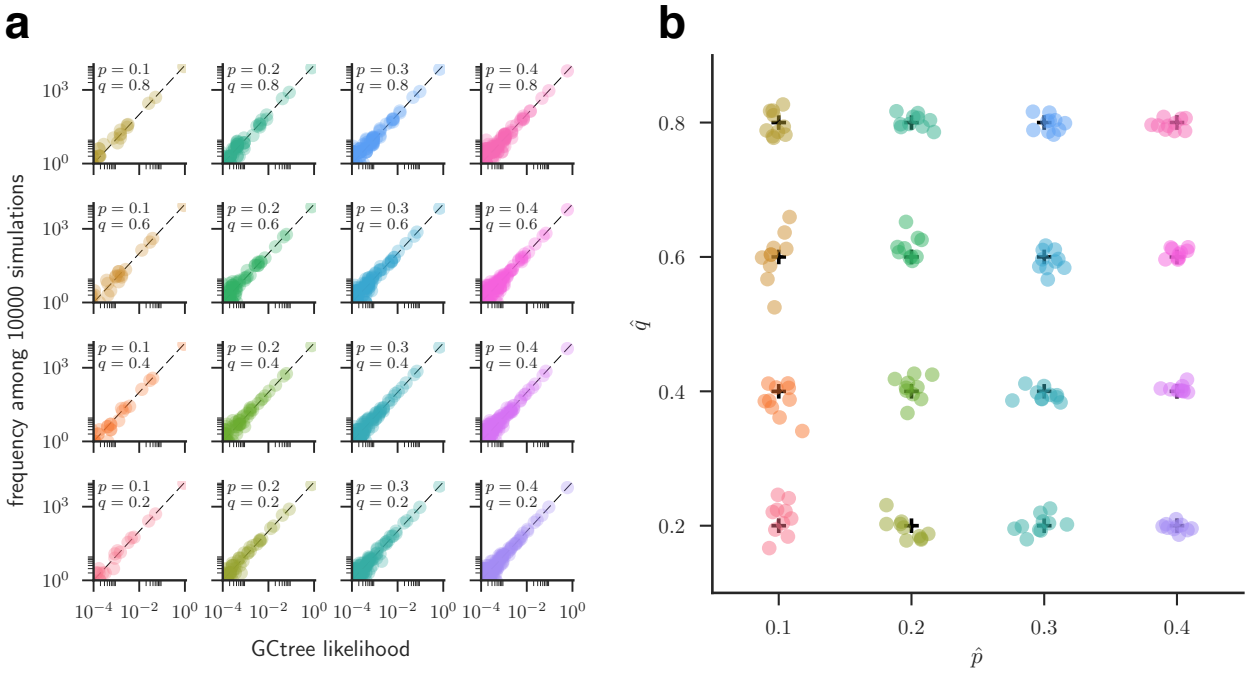

Figure D.3: Numerical validation of GCtree likelihood. Colors indicate simulation parameters. **(a.)** At each parameter value (p, q), 10,000 Galton Watson processes were simulated. For each distinct GCtree, the likelihood was computed according to (C.9), and the frequency of the tree (number of times this distinct tree occurs among the 10,000) was recorded. Dashed lines indicate the expected frequencies (likelihood multiplied by 10,000). **(b.)** Each set of 10,000 trees was partitioned into 10 groups of 1000, and maximum likelihood estimates ($\hat{p}$, $\hat{q}$) were computed for each set of 1000 by numerical maximization of (C.9).

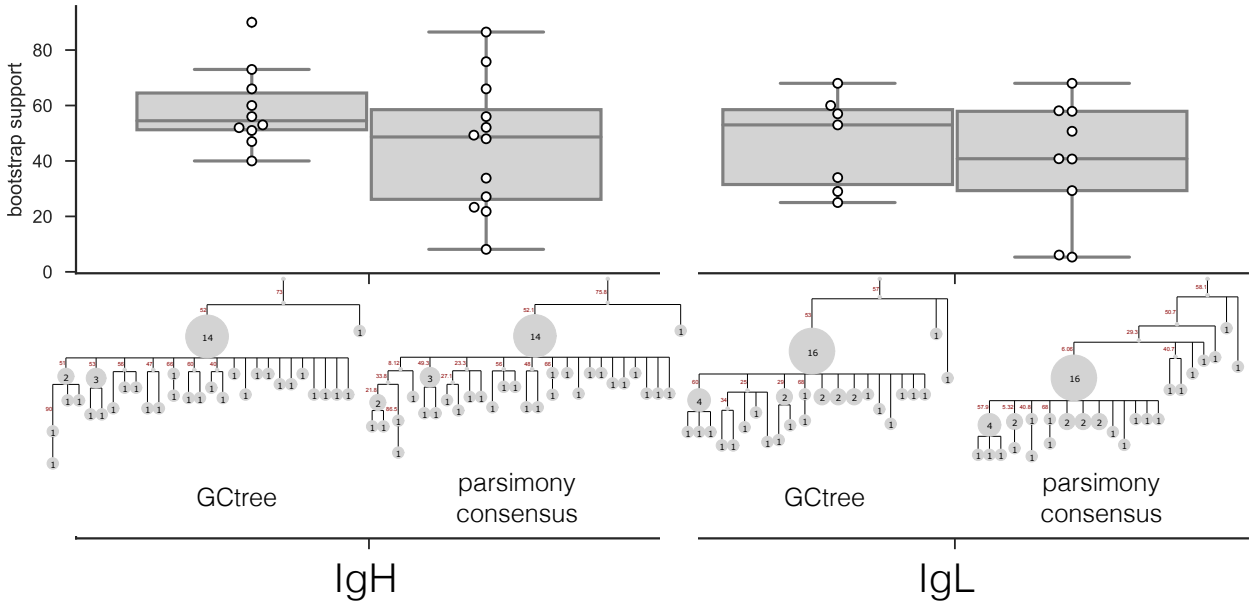

Figure D.4: Bootstrap support comparison. Support values among 100 bootstrap samples are shown for splits in the `GCtree` result and consensus parsimony tree for IgH and IgL sequence data from the same germinal center lineage.

BIBLIOGRAPHY

[1] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[2] Adeolu B Adewoye, Sarah J Lindsay, Yuri E Dubrova, and Matthew E Hurles. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nat. Commun.*, 6:6684, March 2015.

[3] Jeffrey R Adrion, Christopher B Cole, Noah Dukler, Jared G Galloway, Ariella L Gladstein, Graham Gower, Christopher C Kyriazis, Aaron P Ragsdale, Georgia Tsambos, Franz Baumdicker, et al. A community-maintained standard library of population genetic models. *Elife*, 9:e54967, 2020.

[4] Ipsita Agarwal and Molly Przeworski. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proc. Natl. Acad. Sci. U. S. A.*, 116(36):17916–17924, September 2019.

[5] Varun Aggarwala and Benjamin F Voight. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.*, 48(4):349–355, April 2016.

[6] Rachael C Aikens, Kelsey E Johnson, and Benjamin F Voight. Signals of variation in human mutation rate at multiple levels of sequence context. *Mol. Biol. Evol.*, 36(5):955–965, May 2019.

[7] J Aitchison. The statistical analysis of compositional data. *J. R. Stat. Soc. Series B Stat. Methodol.*, 44(2):139–160, January 1982.

[8] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd,

John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Marcin Imielinsk, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.

[9] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, 3(1):246–259, January 2013.

[10] Grégoire Altan-Bonnet, Thierry Mora, and Aleksandra M. Walczak. Quantitative immunology for physicists. *Physics Reports*, 849:1–83, 2020. Quantitative immunology for physicists.

[11] Carlos Eduardo G Amorim, Ziyue Gao, Zachary Baker, José Francisco Diesel, Yuval B Simons, Imran S Haque, Joseph Pickrell, and Molly Przeworski. The population genetics of human disease: The case of recessive, lethal mutations. *PLoS Genet.*, 13(9):e1006915, September 2017.

[12] Luke Anderson-Trocmé, Rick Farouni, Mathieu Bourgey, Yoichiro Kamatani, Koichiro Higasa, Jeong-Sun Seo, Changhoon Kim, Fumihiko Matsuda, and Simon Gravel. Legacy data confounds genomics studies. *Mol. Biol. Evol.*, August 2019.

[13] Adrian Baez-Ortega, Kevin Gori, Andrea Strakova, Janice L Allen, Karen M Allum, Leontine Bansse-Issa, Thinlay N Bhutia, Jocelyn L Bisson, Cristóbal Briceño, Artemio Castillo Domracheva, Anne M Corrigan, Hugh R Cran, Jane T Crawford, Eric Davis, Karina F de Castro, Andrigo B de Nardi, Anna P de Vos, Laura Delgadillo Keenan, Edward M Donelan, Adela R Espinoza Huerta, Ibikunle A Faramade, Mohammed Fazil, Eleni Fotopoulou, Skye N Fruean, Fanny Gallardo-Arrieta, Olga Glebova, Pagona G Gouletsou, Rodrigo F Häfelin Manrique, Joaquim J G P Henriques, Rodrigo S Horta, Natalia Ignatenko, Yaghouba Kane, Cathy King, Debbie Koenig, Ada

Krupa, Steven J Kruzeniski, Young-Mi Kwon, Marta Lanza-Perea, Mihran Lazyan, Adriana M Lopez Quintana, Thibault Losfelt, Gabriele Marino, Simón Martínez Castañeda, Mayra F Martínez-López, Michael Meyer, Edward J Migneco, Berna Nakanwagi, Karter B Neal, Winifred Neunzig, Máire Ní Leathlobhair, Sally J Nixon, Antonio Ortega-Pacheco, Francisco Pedraza-Ordoñez, Maria C Peleteiro, Katherine Polak, Ruth J Pye, John F Reece, Jose Rojas Gutierrez, Haleema Sadia, Sheila K Schmeling, Olga Shamanova, Alan G Sherlock, Maximilian Stammnitz, Audrey E Steenland-Smit, Alla Svitich, Lester J Tapia Martínez, Ismail Thoya Ngoka, Cristian G Torres, Elizabeth M Tudor, Mirjam G van der Wel, Bogdan A Viţălaru, Sevil A Vural, Oliver Walkinton, Jinhong Wang, Alvaro S Wehrle-Martinez, Sophie A E Widdowson, Michael R Stratton, Ludmil B Alexandrov, Iñigo Martincorena, and Elizabeth P Murchison. Somatic evolution and global expansion of an ancient transmissible cancer lineage. *Science*, 365(6452), August 2019.

[14] Soheil Baharian and Simon Gravel. On the decidability of population size histories from finite allele frequency spectra. *Theor. Popul. Biol.*, 120:42–51, March 2018.

[15] M Barak, NS Zuckerman, H Edelman, R Unger, and R Mehr. IgTree (c) : Creating immunoglobulin variable region gene lineage trees. *Journal of Immunological Methods*, 338(1-2):67–74, 2008.

[16] Alvaro Barbero and Suvrit Sra. Modular proximal optimization for multidimensional total-variation regularization. *The Journal of Machine Learning Research*, 19(1):2232–2313, 2018.

[17] Marc A Beal, Matthew J Meier, Andrew Williams, Andrea Rowan-Carroll, Rémi Gagné, Sarah J Lindsay, Tomas Fitzgerald, Matthew E Hurles, Francesco Marchetti, and Carole L Yauk. Paternal exposure to benzo(a)pyrene induces genome-wide mutations in mouse offspring. *Commun Biol*, 2:228, June 2019.

[18] Amir Beck and Marc Teboulle. A fast iterative Shrinkage-Thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, January 2009.

[19] Annabel C Beichman, Emilia Huerta-Sanchez, and Kirk E Lohmueller. Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, 2018.

[20] Annabel C Beichman, Tanya N Phung, and Kirk E Lohmueller. Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3*, 7(11):3605–3620, November 2017.

[21] Sandra Beleza, António M Santos, Brian McEvoy, Isabel Alves, Cláudia Martinho, Emily Cameron, Mark D Shriver, Esteban J Parra, and Jorge Rocha. The timing of pigmentation lightening in europeans. *Mol. Biol. Evol.*, 30(1):24–35, January 2013.

[22] Joy Bergelson, Martin Kreitman, Dmitri A Petrov, Alvaro Sanchez, and Mikhail Tikhonov. Functional biology in its natural context: A search for emergent simplicity. *eLife*, 10:e67646, jun 2021.

[23] Jean Bertoin. The structure of the allelic partition of the total population for Galton–Watson processes with neutral mutations. *Ann. Probab.*, 37(4):1502–1523, July 2009.

[24] Søren Besenbacher, Christina Hvilsom, Tomas Marques-Bonet, Thomas Mailund, and Mikkel Heide Schierup. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat Ecol Evol*, 3(2):286–292, February 2019.

[25] Anand Bhaskar and Yun S Song. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Stat.*, 42(6):2469–2493, December 2014.

[26] Anand Bhaskar, Y X Rachel Wang, and Yun S Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res.*, 25(2):268–279, February 2015.

[27] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, and Skye Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018.

[28] Johanna Brodin, Charlotte Hedskog, Alexander Heddini, Emmanuel Benard, Richard A Neher, Mattias Mild, and Jan Albert. Challenges with using primer IDs to improve accuracy of next generation sequencing. *PLOS One*, 10(3):e0119123, 5 March 2015.

[29] Jack M Broughton and Elic M Weitzel. Population reconstructions for humans and megafauna suggest mixed causes for north american pleistocene extinctions. *Nat. Commun.*, 9(1):5441, December 2018.

[30] Sharon R Browning, Brian L Browning, Ying Zhou, Serena Tucci, and Joshua M Akey. Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell*, 173(1):53–61.e9, March 2018.

[31] Jedidiah Carlson, William S DeWitt, and Kelley Harris. Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation. *Current Opinion in Genetics & Development*, 62:50–57, June 2020.

[32] Jedidiah Carlson, Jun Z Li, and Sebastian Zöllner. Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets. *BMC Genomics*, 19(1):845, November 2018.

[33] Jedidiah Carlson, Adam E Locke, Matthew Flickinger, Matthew Zawistowski, Shawn Levy, Richard M Myers, Michael Boehnke, Hyun Min Kang, Laura J Scott, Jun Z Li, Sebastian Zöllner, Devin Absher, Huda Akil, Gerome Breen, Margit Burmeister, Sarah Cohen-Woods, William G Iacono, James A Knowles, Lisa Legrand, Qing Lu, Matthew McGue, Melvin G McInnis, Carlos N Pato, Michele T Pato, Margarita Rivera, Janet L Sobell, John B Vincent, Stanley J Watson, and The BRIDGES Consortium. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.*, 9(1):3753, September 2018.

[34] Nimrat Chatterjee and Graham C Walker. Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.*, 58(5):235–263, June 2017.

[35] Rongfeng Cui, Tania Medeiros, David Willemsen, Leonardo N M Iasi, Glen E Collier, Martin Graef, Martin Reichard, and Dario Riccardo Valenzano. Relaxed selection limits lifespan by increasing mutation load. *Cell*, 178(2):385–399.e20, July 2019.

[36] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 22 May 2015.

[37] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011.

[38] Anirban DasGupta. *Probability for statistics and machine learning : fundamentals and advanced topics*. Springer texts in statistics. Springer, New York, 2011.

[39] Nilmara de Oliveira Alves, Alexandre Teixeira Vessoni, Annabel Quinet, Rodrigo Soares Fortunato, Gustavo Satoru Kajitani, Milena Simões Peixoto, Sandra de Souza Hacon, Paulo Artaxo, Paulo Saldiva, Carlos Frederico Martins Menck, and

Silvia Regina Batistuzzo de Medeiros. Biomass burning in the amazon region causes DNA damage and cell death in human lung cells. *Sci. Rep.*, 7(1):10937, September 2017.

[40] William S. DeWitt. mutyper: assigning and summarizing mutation types for analyzing germline mutation spectra. *bioRxiv*, 2020.

[41] William S. DeWitt, Kameron Decker Harris, Aaron P. Ragsdale, and Kelley Harris. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, 118(21):e2013798118, 2021.

[42] William S. DeWitt, Paul Lindau, Thomas M. Snyder, Anna M. Sherwood, Marissa Vignali, Christopher S. Carlson, Philip D. Greenberg, Natalie Duerkopp, Ryan O. Emerson, and Harlan S. Robins. A public database of memory and naive B-cell receptor sequences. *PLOS ONE*, 11(8):1–18, 08 2016.

[43] William S DeWitt, Luka Mesin, Gabriel D Victora, Vladimir N Minin, and IV Matsen, Frederick A. Using Genotype Abundance to Improve Phylogenetic Inference. *Molecular Biology and Evolution*, 35(5):1253–1265, 02 2018.

[44] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, 35(4):316–319, April 2017.

[45] Alexei J Drummond and Remco R Bouckaert. *Bayesian evolutionary analysis with BEAST*. Cambridge University Press, 2015.

[46] Alexei J Drummond, Oliver G Pybus, Andrew Rambaut, Roald Forsberg, and Allen G Rodrigo. Measurably evolving populations. *Trends Ecol. Evol.*, 18(9):481–488, September 2003.

[47] Beth L Dumont. Significant strain variation in the mutation spectra of inbred laboratory mice. *Mol. Biol. Evol.*, 36(5):865–874, May 2019.

[48] Beth L Dumont. Significant strain variation in the mutation spectra of inbred laboratory mice. *Mol. Biol. Evol.*, 36(5):865–874, May 2019.

[49] Deborah K. Dunn-Walters, Ahmet Dogan, Laurent Boursier, Connie M. MacDonald, and Jo Spencer. Base-specific sequences that bias somatic hypermutation deduced by analysis of out-of-frame human IgVH genes. *The Journal of Immunology*, 160(5):2360–2364, 1998.

[50] Richard V. Eck and Margaret O. Dayhoff. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science*, 152(3720):363–366, 1966.

[51] J J Egozcue, V Pawlowsky-Glahn, G Mateu-Figueras, and C Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Math. Geol.*, 35(3):279–300, April 2003.

[52] Charles L Epstein and John Schotland. The bad truth about laplace's transform. *SIAM Rev.*, 50(3):504–520, January 2008.

[53] Warren J Ewens. *Mathematical Population Genetics 1: Theoretical Introduction*. Springer Science & Business Media, October 2012.

[54] M Fazel, H Hindi, and S P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*, volume 6, pages 4734–4739 vol.6, June 2001.

[55] J. Felsenstein. *Inferring Phylogenies*. Sinauer, 2003.

[56] J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, 2005.

[57] Joseph Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240, 1973.

[58] Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[59] Joseph Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985.

[60] Jean Feng, William S DeWitt III, Aaron McKenna, Noah Simon, Amy D Willis, and Frederick A Matsen IV. Estimation of cell lineage trees by maximum-likelihood phylogenetics. *The Annals of Applied Statistics*, 15(1):343–362, 2021.

[61] Jack N Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, 128(2):415–423, 2005.

[62] Yair Field, Evan A Boyle, Natalie Telis, Ziyue Gao, Kyle J Gaulton, David Golan, Loic Yengo, Ghislain Rocheleau, Philippe Froguel, Mark I McCarthy, and Jonathan K Pritchard. Detection of human adaptation during the past 2000 years. *Science*, 354(6313):760–764, November 2016.

[63] R B Firestone, A West, J P Kennett, L Becker, T E Bunch, Z S Revay, P H Schultz, T Belgya, D J Kennett, J M Erlandson, O J Dickenson, A C Goodyear, R S Harris, G A Howard, J B Kloosterman, P Lechler, P A Mayewski, J Montgomery, R Poreda, T Darrah, S S Que Hee, A R Smith, A Stich, W Topping, J H Wittke, and W S Wolbach. Evidence for an extraterrestrial impact 12,900 years ago that contributed to the megafaunal extinctions and the younger dryas cooling. *Proc. Natl. Acad. Sci. U. S. A.*, 104(41):16016–16021, October 2007.

[64] Walter M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Biology*, 20(4):406, 1971.

[65] L.R. Foulds and R.L. Graham. The steiner problem in phylogeny is np-complete. *Advances in Applied Mathematics*, 3(1):43 – 49, 1982.

[66] Ziyue Gao, Priya Moorjani, Thomas A Sasani, Brent S Pedersen, Aaron R Quinlan, Lynn B Jorde, Guy Amster, and Molly Przeworski. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl. Acad. Sci. U. S. A.*, 116(19):9491–9500, May 2019.

[67] Ziyue Gao, Minyoung J Wyman, Guy Sella, and Molly Przeworski. Interpreting the dependence of mutation rates on age and time. *PLoS Biol.*, 14(1):e1002355, January 2016.

[68] Pablo E García-Nieto, Ashby J Morrison, and Hunter B Fraser. The somatic mutation landscape of the human body. *Genome Biol.*, 20(1):298, December 2019.

[69] Alexandra Gavryushkina, Tracy A Heath, Daniel T Ksepka, Tanja Stadler, David Welch, and Alexei J Drummond. Bayesian total evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, 66(1):57–73, 2017.

[70] Alexandra Gavryushkina, David Welch, Tanja Stadler, and Alexei J Drummond. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput. Biol.*, 10(12):e1003919, December 2014.

[71] R Gilchrist. Archaeological biographies: realizing human lifecycles, -courses and -histories. *World Archaeol.*, 31(3):325–328, February 2000.

[72] Michael E Goldberg and Kelley Harris. Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny. *Genome Biology and Evolution*, 14(1), 05 2021. evab104.

[73] Jakob M Goldmann, Vladimir B Seplyarskiy, Wendy S W Wong, Thierry Vilboux, Pieter B Neerincx, Dale L Bodian, Benjamin D Solomon, Joris A Veltman, John F Deeken, Christian Gilissen, and John E Niederhuber. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.*, March 2018.

[74] Jakob M Goldmann, Wendy S W Wong, Michele Pinelli, Terry Farrah, Dale Bodian, Anna B Stittrich, Gustavo Glusman, Lisenka E L M Vissers, Alexander Hoischen, Jared C Roach, Joseph G Vockley, Joris A Veltman, Benjamin D Solomon, Christian Gilissen, and John E Niederhuber. Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.*, June 2016.

[75] Benjamin H Good, Michael J McDonald, Jeffrey E Barrick, Richard E Lenski, and Michael M Desai. The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50, November 2017.

[76] M Goodman. Rates of molecular evolution: the hominoid slowdown. *Bioessays*, 3(1):9–14, July 1985.

[77] Victor Greiff, Enkelejda Miho, Ulrike Menzel, and Sai T. Reddy. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends in Immunology*, 36(11):738–749, 2015.

[78] R C Griffiths and S Tavaré. The age of a mutation in a general coalescent tree. *Stoch. Models*, 1998.

[79] Namita T. Gupta, Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H. Kleinstein. Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. *Bioinformatics*, 31(20):3356, 2015.

[80] Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, 5(10):e1000695, October 2009.

[81] Benjamin C Haller and Philipp W Messer. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Mol. Biol. Evol.*, 36(3):632–637, March 2019.

[82] Kelley Harris. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U. S. A.*, 112(11):3439–3444, March 2015.

[83] Kelley Harris and Jonathan K Pritchard. Rapid evolution of the human mutation spectrum. *Elife*, 6, April 2017.

[84] Theodore E Harris. *The Theory of Branching processes*. Courier Corporation, 2002.

[85] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, May 2015.

[86] Colin Havenar-Daughton, Diane G Carnathan, Alba Torrents de la Peña, Matthias Pauthner, Bryan Briney, Samantha M Reiss, Jennifer S Wood, Kirti Kaushik, Marit J van Gils, Sandy L Rosales, Patricia van der Woude, Michela Locci, Khoa M Le, Steven W de Taeye, Devin Sok, Ata Ur Rasheed Mohammed, Jessica Huang, Sanjeev Gumber, Anapatricia Garcia, Sudhir P Kasturi, Bali Pulendran, John P Moore, Rafi Ahmed, Grégory Seumois, Dennis R Burton, Rogier W Sanders, Guido Silvestri, and Shane Crotty. Direct probing of germinal center responses reveals immunological features and bottlenecks for neutralizing antibody responses to HIV env trimer. *Cell Rep.*, 17(9):2195–2209, 22 November 2016.

[87] Thomas Helleday, Saeed Eshtad, and Serena Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, 15(9):585–598, September 2014.

[88] Ron Hightower, Stephanie Forrest, and Alan S Perelson. The evolution of secondary organization in immune system gene libraries. Technical report, Los Alamos National Lab., NM (United States), 1993.

[89] K Hill and H Kaplan. Life history traits in humans: theory and empiricial studies. *Annu. Rev. Anthropol.*, 28:397–430, 1999.

[90] Manuel Holtgrewe, Alexej Knaus, Gabriele Hildebrand, Jean-Tori Pantel, Miguel Rodriguez de Los Santos, Kornelia Neveling, Jakob Goldmann, Max Schubach, Marten Jäger, Marie Coutelier, Stefan Mundlos, Dieter Beule, Karl Sperling, and Peter Michael Krawitz. Multisite de novo mutations in human offspring after paternal exposure to ionizing radiation. *Sci. Rep.*, 8(1):14611, October 2018.

[91] Dongni Hou, Cuicui Chen, Eric John Seely, Shujing Chen, and Yuanlin Song. High-throughput sequencing-based immune repertoire study during infectious disease. *Frontiers in Immunology*, 7:336, 2016.

[92] Bryan Howie, Anna M Sherwood, Ashley D Berkebile, Jan Berka, Ryan O Emerson, David W Williamson, Ilan Kirsch, Marissa Vignali, Mark J Rieder, Christopher S Carlson, and Harlan S Robins. High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.*, 7(301):301ra131, 19 August 2015.

[93] John P. Huelsenbeck, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314, 2001.

[94] Jaime Huerta-Cepas, François Serra, and Peer Bork. Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6):1635, 2016.

[95] Dick G Hwang and Phil Green. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 101(39):13994–14001, September 2004.

[96] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, October 2007.

[97] Cassandra B Jabara, Corbin D Jones, Jeffrey Roach, Jeffrey A Anderson, and Ronald Swanstrom. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID. *Proc. Natl. Acad. Sci. U. S. A.*, 108(50):20166–20171, 13 December 2011.

[98] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biol.*, 17:86, 5 May 2016.

[99] Pengyao Jiang, Anja R Ollodart, Vidha Sudhesh, Alan J Herr, Maitreya J Dunham, and Kelley Harris. A modified fluctuation assay reveals a natural mutator phenotype that drives mutation spectrum variation within *Saccharomyces cerevisiae*. *eLife*, 10:e68285, sep 2021.

[100] Hákon Jónsson, Patrick Sulem, Gudny A Arnadottir, Gunnar Pálsson, Hannes P Eggertsson, Snaedis Kristmundsdottir, Florian Zink, Birte Kehr, Kristjan E Hjorleifsson, Brynjar Ö Jensson, Ingileif Jonsdottir, Sigurdur Einar Marelsson, Sigurjon Axel Gudjonsson, Arnaldur Gylfason, Adalbjorg Jonasdottir, Aslaug Jonasdottir, Simon N

Stacey, Olafur Th Magnusson, Unnur Thorsteinsdottir, Gisli Masson, Augustine Kong, Bjarni V Halldorsson, Agnar Helgason, Daniel F Gudbjartsson, and Kari Stefansson. Multiple transmissions of de novo mutations in families. *Nat. Genet.*, 50(12):1674–1680, December 2018.

[101] Hákon Jónsson, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur Hjartarson, Marteinn T Hardarson, Kristjan E Hjorleifsson, Hannes P Eggertsson, Sigurjon Axel Gudjonsson, Lucas D Ward, Gudny A Arnadottir, Einar A Helgason, Hannes Helgason, Arnaldur Gylfason, Adalbjorg Jonasdottir, Aslaug Jonasdottir, Thorunn Rafnar, Mike Frigge, Simon N Stacey, Olafur Th. Magnusson, Unnur Thorsteinsdottir, Gisli Masson, Augustine Kong, Bjarni V Halldorsson, Agnar Helgason, Daniel F Gudbjartsson, and Kari Stefansson. Parental influence on human germline de novo mutations in 1,548 trios from iceland. *Nature*, 1:16027, September 2017.

[102] Jack Kamm, Jonathan Terhorst, Richard Durbin, and Yun S. Song. Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, 115(531):1472–1487, 2020.

[103] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*, 12(5):1–22, 05 2016.

[104] Mark R Kelley, Yoke W Kow, and David M Wilson, 3rd. Disparity between DNA base excision repair in yeast and mammals: translational implications. *Cancer Res.*, 63(3):549–554, February 2003.

[105] Michael D Kessler, Douglas P Loesch, James A Perry, Nancy L Heard-Costa, Daniel Taliun, Brian E Cade, Heming Wang, Michelle Daya, John Ziniti, Soma Datta, Juan C Celedón, Manuel E Soto-Quiros, Lydiana Avila, Scott T Weiss, Kathleen Barnes, Susan S Redline, Ramachandran S Vasan, Andrew D Johnson, Rasika A Mathias, Ryan Hernandez, James G Wilson, Deborah A Nickerson, Goncalo Abecasis, Sharon R Browning, Sebastian Zöllner, Jeffrey R O'Connell, Braxton D Mitchell, National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Population Genetics Working Group, and Timothy D O'Connor. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the amish founder population. *Proc. Natl. Acad. Sci. U. S. A.*, January 2020.

[106] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Rev. Soc. Ind. Appl. Math.*, 51(2):339–360, May 2009.

[107] Motoo Kimura. On the evolutionary adjustment of spontaneous mutation rates*. *Genet. Res.*, 9(1):23–34, February 1967.

[108] J F Kingman. Origins of the coalescent. 1974-1982. *Genetics*, 156(4):1461–1463, December 2000.

[109] J F C Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, September 1982.

[110] J F C Kingman. On the genealogy of large populations. *J. Appl. Probab.*, 19(A):27–43, 1982.

[111] JFC Kingman, G Koch, and F Spizzichino. Exchangeability and the evolution of large populations. *Exchangeability in probability and statistics*, 91:112, 1982.

[112] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9(1):72–74, 20 November 2011.

[113] Steven H Kleinstein, Yoram Louzoun, and Mark J Shlomchik. Estimating hypermutation rates from clonal tree data. *J. Immunol.*, 171(9):4639–4649, November 2003.

[114] Arnold G. Kluge and James S. Farris. Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, 18(1):1–32, 1969.

[115] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, August 2009.

[116] Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. Tensorly: Tensor learning in python. *Journal of Machine Learning Research*, 20(26):1–6, 2019.

[117] Jill E Kucab, Xueqing Zou, Sandro Morganella, Madeleine Joel, A Scott Nanda, Eszter Nagy, Celine Gomez, Andrea Degasperi, Rebecca Harris, Stephen P Jackson, Volker M Arlt, David H Phillips, and Serena Nik-Zainal. A compendium of mutational signatures of environmental agents. *Cell*, 177(4):821–836.e16, May 2019.

[118] Masayuki Kuraoka, Aaron G Schmidt, Takuya Nojima, Feng Feng, Akiko Watanabe, Daisuke Kitamura, Stephen C Harrison, Thomas B Kepler, and Garnett Kelsoe. Complex antigens drive permissive clonal selection in germinal centers. *Immunity*, 2 March 2016.

[119] Iosif Lazaridis, Dani Nadel, Gary Rollefson, Deborah C Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, Mario Novak, Beatriz Gamarra, Kendra Sirak, et al. Genomic insights into the origin of farming in the ancient near east. *Nature*, 536(7617):419–424, 2016.

[120] Heewook Lee, Ellen Popodi, Haixu Tang, and Patricia L Foster. Rate and molecular spectrum of spontaneous mutations in the bacterium escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 109(41):E2774–83, October 2012.

[121] Richard E Lenski, Michael R Rose, Suzanne C Simpson, and Scott C Tadler. Long-Term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *Am. Nat.*, 138(6):1315–1341, 1991.

[122] Paul O Lewis, Mark T Holder, and Kent E Holsinger. Polytomies and Bayesian phylogenetic inference. *Syst. Biol.*, 54(2):241–253, April 2005.

[123] Paul O Lewis, Mark T Holder, and David L Swofford. Phycas: Software for Bayesian phylogenetic analysis. *Syst. Biol.*, 9 January 2015.

[124] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, November 2011.

[125] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, July 2011.

[126] Haoxuan Liu and Jianzhi Zhang. Yeast spontaneous mutation rate and spectrum vary with environment. *Curr. Biol.*, 29(10):1584–1591.e3, May 2019.

[127] Xiao Liu and Jinghua Wu. History, applications, and challenges of immune repertoire research. *Cell biology and toxicology*, 34(6):441–457, 2018.

[128] Lawrence A Loeb. Human cancers express a mutator phenotype: Hypothesis, origin, and consequences. *Cancer Res.*, 76(8):2057–2059, April 2016.

[129] Scott A Lujan, Jessica S Williams, and Thomas A Kunkel. DNA polymerases divide the labor of genome replication. *Trends Cell Biol.*, 26(9):640–654, September 2016.

[130] Michael Lynch, Matthew S Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W Kelley Thomas, and Patricia L Foster. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.*, 17(11):704–714, October 2016.

[131] David R Maddison. The discovery and importance of multiple islands of Most-Parsimonious trees. *Syst. Zool.*, 40(3):315–328, 1 September 1991.

[132] Andrew F Magee, Sarah K Hilton, and William S DeWitt. Robustness of phylogenetic inference to model misspecification caused by pairwise epistasis. *Molecular biology and evolution*, 38(10):4603–4615, 2021.

[133] Reuma Magori-Cohen, Yoram Louzoun, and Steven H. Kleinstein. Mutation parameters from dna sequence data using graph theoretic measures on lineage trees. *Bioinformatics*, 22(14):e332–e340, 2006.

[134] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, Pontus Skoglund, Iosif Lazaridis, Sriram Sankararaman, Qiaomei Fu, Nadin Rohland, Gabriel Renaud, Yaniv Erlich, Thomas Willems, Carla Gallo, Jeffrey P Spence, Yun S Song, Giovanni Poletti, Francois Balloux, George van Driem, Peter de Knijff, Irene Gallego Romero, Aashish R Jha, Doron M Behar, Claudio M Bravi, Cristian Capelli, Tor Hervig, Andres Moreno-Estrada, Olga L Posukh, Elena Balanovska, Oleg Balanovsky, Sena Karachanak-Yankova, Hovhannes Sahakyan, Draga Toncheva, Levon Yepiskoposyan, Chris Tyler-Smith, Yali Xue, M Syafiq Abdullah, Andres Ruiz-Linares, Cynthia M Beall, Anna Di Rienzo, Choongwon Jeong, Elena B Starikovskaya, Ene Metspalu, Jüri Parik, Richard Villems, Brenna M Henn, Ugur Hodoglugil, Robert Mahley, Antti Sajantila, George Stamatoyannopoulos, Joseph T S Wee, Rita Khusainova, Elza Khusnutdinova, Sergey Litvinov, George Ayodo, David Comas, Michael F Hammer, Toomas Kivisild, William Klitz, Cheryl A Winkler, Damian Labuda, Michael Bamshad, Lynn B Jorde, Sarah A Tishkoff, W Scott Watkins, Mait Metspalu, Stanislav Dryomov, Rem Sukernik, Lalji Singh, Kumarasamy Thangaraj, Svante Pääbo, Janet Kelso, Nick Patterson, and David Reich. The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, October 2016.

[135] Iain Mathieson and David Reich. Differences in the rare variant spectrum among human populations. *PLoS Genet.*, 13(2):e1006581, February 2017.

[136] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[137] Aaron McKenna, Gregory M. Findlay, James A. Gagnon, Marshall S. Horwitz, Alexander F. Schier, and Jay Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298), 2016.

[138] Luka Mesin, Jonatan Ersching, and Gabriel D. Victora. Germinal center b cell dynamics. *Immunity*, 45(3):471–482, 2016.

[139] Luka Mesin, Jonatan Ersching, and Gabriel D Victora. Germinal center B cell dynamics. *Immunity*, 45(3):471–482, 20 September 2016.

[140] William R Milligan, Guy Amster, and Guy Sella. The impact of genetic modifiers on variation in germline mutation rates within and among human populations. *bioRxiv*, 2021.

[141] Priya Moorjani, Carlos Eduardo G Amorim, Peter F Arndt, and Molly Przeworski. Variation in the molecular clock of primates. *Proc. Natl. Acad. Sci. U. S. A.*, 113(38):10607–10612, September 2016.

[142] Priya Moorjani, Ziyue Gao, and Molly Przeworski. Human germline mutation and the erratic evolutionary clock. *PLoS Biol.*, 14(10):e2000744, October 2016.

[143] Anand Murugan, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan. Statistical inference of the generation probability of t-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, 2012.

[144] Rajagopal Murugan, Lisa Buchauer, Gianna Triller, Cornelia Kreschel, Giulia Costa, Gemma Pidelaserra Martí, Katharina Imkeller, Christian E Busse, Sumana Chakravarty, B Kim Lee Sim, Stephen L Hoffman, Elena A Levashina, Peter G Kremsner, Benjamin Mordmüller, Thomas Höfer, and Hedda Wardemann. Clonal selection drives protective memory B cell responses in controlled human malaria infection. *Sci Immunol*, 3(20), February 2018.

[145] Simon Myers, Charles Fefferman, and Nick Patterson. Can one learn history from the allelic spectrum? *Theor. Popul. Biol.*, 73(3):342–348, May 2008.

[146] Vagheesh M Narasimhan, Nick Patterson, Priya Moorjani, Nadin Rohland, Rebecca Bernardos, Swapan Mallick, Iosif Lazaridis, Nathan Nakatsuka, Iñigo Olalde, Mark Lipson, Alexander M Kim, Luca M Olivieri, Alfredo Coppa, Massimo Vidale, James Mallory, Vyacheslav Moiseyev, Egor Kitov, Janet Monge, Nicole Adamski, Neel Alex, Nasreen Broomandkhoshbacht, Francesca Candilio, Kimberly Callan, Olivia

Cheronet, Brendan J Culleton, Matthew Ferry, Daniel Fernandes, Suzanne Freilich, Beatriz Gamarra, Daniel Gaudio, Mateja Hajdinjak, Éadaoin Harney, Thomas K Harper, Denise Keating, Ann Marie Lawson, Matthew Mah, Kirsten Mandl, Megan Michel, Mario Novak, Jonas Oppenheimer, Niraj Rai, Kendra Sirak, Viviane Slon, Kristin Stewardson, Fatma Zalzala, Zhao Zhang, Gaziz Akhatov, Anatoly N Bagashev, Alessandra Bagnera, Bauryzhan Baitanayev, Julio Bendezu-Sarmiento, Arman A Bissembaev, Gian Luca Bonora, Temirlan T Chargynov, Tatiana Chikisheva, Petr K Dashkovskiy, Anatoly Derevianko, Miroslav Dobeš, Katerina Douka, Nadezhda Dubova, Meiram N Duisengali, Dmitry Enshin, Andrey Epimakhov, Alexey V Fribus, Dorian Fuller, Alexander Goryachev, Andrey Gromov, Sergey P Grushin, Bryan Hanks, Margaret Judd, Erlan Kazizov, Aleksander Khokhlov, Aleksander P Krygin, Elena Kupriyanova, Pavel Kuznetsov, Donata Luiselli, Farhod Maksudov, Aslan M Mamedov, Talgat B Mamirov, Christopher Meiklejohn, Deborah C Merrett, Roberto Micheli, Oleg Mochalov, Samariddin Mustafokulov, Ayushi Nayak, Davide Pettener, Richard Potts, Dmitry Razhev, Marina Rykun, Stefania Sarno, Tatyana M Savenkova, Kulyan Sikhymbaeva, Sergey M Slepchenko, Oroz A Soltobaev, Nadezhda Stepanova, Svetlana Svyatko, Kubatbek Tabaldiev, Maria Teschler-Nicola, Alexey A Tishkin, Vitaly V Tkachev, Sergey Vasilyev, Petr Velemínský, Dmitriy Voyakin, Antonina Yermolayeva, Muhammad Zahir, Valery S Zubkov, Alisa Zubova, Vasant S Shinde, Carles Lalueza-Fox, Matthias Meyer, David Anthony, Nicole Boivin, Kumarasamy Thangaraj, Douglas J Kennett, Michael Frachetti, Ron Pinhasi, and David Reich. The formation of human populations in south and central asia. *Science*, 365(6457), September 2019.

[147] Vagheesh M Narasimhan, Raheleh Rahbari, Aylwyn Scally, Arthur Wuster, Dan Mason, Yali Xue, John Wright, Richard C Trembath, Eamonn R Maher, David A van Heel, Adam Auton, Matthew E Hurles, Chris Tyler-Smith, and Richard Durbin. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat. Commun.*, 8(1):303, August 2017.

[148] YE Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[149] Chang Gyun Park, Hyun Ki Cho, Han Jae Shin, Ki Hong Park, and Heung Bin Lim. Comparison of mutagenic activities of various Ultra-Fine particles. *Toxicol. Res.*, 34(2):163–172, April 2018.

[150] Benedict Paten, Javier Herrero, Stephen Fitzgerald, Kathryn Beal, Paul Flicek, Ian Holmes, and Ewan Birney. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, 18(11):1829–1843, November 2008.

[151] Peter Paule and Markus Schorn. A mathematica version of zeilberger's algorithm for proving binomial coefficient identities. *Journal of symbolic computation*, 20(5-6):673–698, 1995.

[152] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, March 2015.

[153] Brent S Pedersen and Aaron R Quinlan. cyvcf2: fast, flexible variant analysis with python. *Bioinformatics*, 33(12):1867–1869, June 2017.

[154] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[155] Fabian Pedregosa and Gauthier Gidel. Adaptive three operator splitting. In *International Conference on Machine Learning*, pages 4085–4094. PMLR, 2018.

[156] Alan S. Perelson and Gérard Weisbuch. Immunology for physicists. *Rev. Mod. Phys.*, 69:1219–1268, Oct 1997.

[157] Alan S. Perelson and Frederik W. Wiegel. Some design principles for immune system recognition. *Complexity*, 4(5):29–37, 1999.

[158] Marko Petkovšek, Herbert S Wilf, and Doron Zeilberger. A= b, ak peters ltd. *Wellesley, MA*, 30, 1996.

[159] Jenni E Pettay, Loeske E B Kruuk, Jukka Jokela, and Virpi Lummaa. Heritability and genetic constraints of life-history trait evolution in preindustrial humans. *Proc. Natl. Acad. Sci. U. S. A.*, 102(8):2838–2843, February 2005.

[160] Gerd P Pfeifer, Mikhail F Denissenko, Magali Olivier, Natalia Tretyakova, Stephen S Hecht, and Pierre Hainaut. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*, 21(48):7435–7451, October 2002.

[161] Mario Pino, Ana M Abarzúa, Giselle Astorga, Alejandra Martel-Cea, Nathalie Cossio-Montecinos, R Ximena Navarro, Maria Paz Lira, Rafael Labarca, Malcolm A LeCompte, Victor Adedeji, Christopher R Moore, Ted E Bunch, Charles Mooney, Wendy S Wolbach, Allen West, and James P Kennett. Sedimentary record from patagonia, southern chile supports cosmic-impact triggering of biomass burning, climate change, and megafaunal extinctions at 12.8 ka. *Sci. Rep.*, 9(1):4413, March 2019.

[162] Vincent Plagnol and Jeffrey D Wall. Possible ancestral structure in human populations. *PLoS Genet.*, 2(7):e105, July 2006.

[163] A Polanski, A Bobrowski, and M Kimmel. A note on distributions of times to coalescence, under time-dependent population size. *Theor. Popul. Biol.*, 63(1):33–40, February 2003.

[164] A Polanski and M Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1):427–436, September 2003.

[165] John E Pool, Ines Hellmann, Jeffrey D Jensen, and Rasmus Nielsen. Population genetic inference from genomic sequence variation. *Genome research*, 20(3):291–300, 2010.

[166] Fanny Pouyet, Simon Aeschbacher, Alexandre Thiéry, and Laurent Excoffier. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife*, 7, August 2018.

[167] James G D Prendergast, Carys Pugh, Sarah E Harris, David A Hume, Ian J Deary, and Allan Beveridge. Linked Mutations at Adjacent Nucleotides Have Shaped Human Population Differentiation and Protein Evolution. *Genome Biology and Evolution*, 11(3):759–775, 01 2019.

[168] Raheleh Rahbari, Arthur Wuster, Sarah J Lindsay, Robert J Hardwick, Ludmil B Alexandrov, Saeed Al Turki, Anna Dominiczak, Andrew Morris, David Porteous, Blair Smith, Michael R Stratton, UK10K Consortium, and Matthew E Hurles. Timing, rates and spectra of human germline mutation. *Nat. Genet.*, 48(2):126–133, February 2016.

[169] Duncan K. Ralph and Frederick A. Matsen, IV. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLOS Computational Biology*, 12(1):1–25, 01 2016.

[170] Duncan K. Ralph and Frederick A. Matsen, IV. Likelihood-based inference of B cell clonal families. *PLOS Computational Biology*, 12(10):1–28, 10 2016.

[171] Aaditya Ramdas and Ryan J Tibshirani. Fast and flexible ADMM algorithms for trend filtering. *J. Comput. Graph. Stat.*, 25(3):839–858, July 2016.

[172] David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L Price, and Lalji Singh. Reconstructing indian population history. *Nature*, 461(7263):489–494, September 2009.

[173] Harlan Robins. Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, 25(5):646–652, 2013.

[174] D.F. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131 – 147, 1981.

[175] Zvi Rosen, Anand Bhaskar, Sebastien Roch, and Yun S Song. Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, page genetics.300733.2018, July 2018.

[176] Edith M Ross and Florian Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.*, 17:69, 15 April 2016.

[177] Ori Sargsyan and John Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor. Popul. Biol.*, 74(1):104–114, August 2008.

[178] Thomas A. Sasani, David G. Ashbrook, Lu Lu, Abraham A. Palmer, Robert W. Williams, Jonathan K. Pritchard, and Kelley Harris. A wild-derived antimutator drives germline mutation spectrum differences in a genetically diverse murine family. *bioRxiv*, 2021.

[179] Thomas A Sasani, Brent S Pedersen, Ziyue Gao, Lisa Baird, Molly Przeworski, Lynn B Jorde, and Aaron R Quinlan. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife*, 8, September 2019.

[180] Yasunari Satoh, Jun-Ichi Asakawa, Mayumi Nishimura, Tony Kuo, Norio Shinkai, Harry M Cullings, Yohei Minakuchi, Jun Sese, Atsushi Toyoda, Yoshiya Shimada, Nori Nakamura, and Arikuni Uchimura. Characteristics of induced mutations in offspring derived from irradiated mouse spermatogonia and mature oocytes. *Sci. Rep.*, 10(1):37, January 2020.

[181] S A Sawyer and D L Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, December 1992.

[182] Aylwyn Scally. The mutation rate in human evolution and demographic inference. *Curr. Opin. Genet. Dev.*, 41:36–43, December 2016.

[183] Aylwyn Scally and Richard Durbin. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.*, 13(10):745–753, October 2012.

[184] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

[185] Joshua G Schraiber and Joshua M Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, 2015.

[186] Russell Schwartz and Alejandro A. Schaffer. The evolution of tumour phylogenetics: principles and practice. *Nat Rev Genet*, 18(4):213–229, 04 2017.

[187] Laure Ségurel, Minyoung J Wyman, and Molly Przeworski. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.*, 15:47–70, June 2014.

[188] Cathal Seoighe and Aylwyn Scally. Inference of candidate germline mutator loci in humans from Genome-Wide haplotype data. *PLoS Genet.*, 13(1):e1006549, January 2017.

[189] Vladimir B Seplyarskiy, Ruslan A Soldatov, Evan Koch, Ryan J McGinty, Jakob M Goldmann, Ryan D Hernandez, Kathleen Barnes, Adolfo Correa, Esteban G Burchard, Patrick T Ellinor, et al. Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science*, 373(6558):1030–1035, 2021.

[190] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, 14(9):618–630, September 2013.

[191] Matthew D Shirley, Zhaorong Ma, Brent S Pedersen, and Sarah J Wheelan. Efficient "pythonic" access to FASTA files using pyfaidx. Technical Report e1196, PeerJ PrePrints, April 2015.

[192] Mark J. Shlomchik, Wei Luo, and Florian Weisel. Linking signaling and selection in the germinal center. *Immunological Reviews*, 288(1):49–63, 2019.

[193] Adrien Six, Encarnita Mariotti-Ferrandiz, Wahiba Chaara, Susana Magadan, Hang-Phuong Pham, Marie-Paule Lefranc, Thierry Mora, VÃ©ronique Thomas-Vaslin, Aleksandra Walczak, and Pierre Boudinot. The past, present, and future of immune repertoire biology—the rise of next-generation repertoire analysis. *Frontiers in Immunology*, 4:413, 2013.

[194] Laurits Skov, Moisès Coll Macià, Garðar Sveinbjörnsson, Fabrizio Mafessoni, Elise A Lucotte, Margret S Einarsdóttir, Hakon Jonsson, Bjarni Halldorsson, Daniel F Gudbjartsson, Agnar Helgason, Mikkel Heide Schierup, and Kari Stefansson. The nature of neanderthal introgression revealed by 27,566 icelandic genomes. *Nature*, April 2020.

[195] Felisa A Smith, Rosemary E Elliott Smith, S Kathleen Lyons, and Jonathan L Payne. Body size downgrading of mammals over the late quaternary. *Science*, 360(6386):310–313, April 2018.

[196] P D Sniegowski, P J Gerrish, and R E Lenski. Evolution of high mutation rates in experimental populations of e. coli. *Nature*, 387(6634):703–705, June 1997.

[197] Devin Sok, Uri Laserson, Jonathan Laserson, Yi Liu, Francois Vigneault, Jean-Philippe Julien, Bryan Briney, Alejandra Ramos, Karen F Saye, Khoa Le, Alison Mahan, Shenshen Wang, Mehran Kardar, Gur Yaari, Laura M Walker, Birgitte B Simen, Elizabeth P St John, Po-Ying Chan-Hui, Kristine Swiderek, Steven H Kleinstein, Stephen H Kleinstein, Galit Alter, Michael S Seaman, Arup K Chakraborty, Daphne Koller, Ian A Wilson, George M Church, Dennis R Burton, and Pascal Poignard. The effects of somatic hypermutation on neutralization and binding in the PGT121 family of broadly neutralizing HIV antibodies. *PLoS Pathog.*, 9(11):e1003754, November 2013.

[198] Leo Speidel, Lara Cassidy, Robert W Davies, Garrett Hellenthal, Pontus Skoglund, and Simon R Myers. Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. *Molecular Biology and Evolution*, 38(9):3497–3511, 06 2021.

[199] Leo Speidel, Marie Forest, Sinan Shi, and Simon R Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.*, 51(9):1321–1329, September 2019.

[200] Jeffrey P Spence, John A Kamm, and Yun S Song. The site frequency spectrum for general coalescents. *Genetics*, 202(4):1549–1561, April 2016.

[201] Jo Spencer and Deborah K. Dunn-Walters. Hypermutation at A-T base pairs: The a nucleotide replacement spectrum is affected by adjacent nucleotides and there is no reverse complementarity of sequences flanking mutated A and T nucleotides. *The Journal of Immunology*, 175(8):5170–5177, 2005.

[202] Joel N H Stern, Gur Yaari, Jason A Vander Heiden, George Church, William F Donahue, Rogier Q Hintzen, Anita J Huttner, Jon D Laman, Rashed M Nagra, Alyssa Nylander, David Pitt, Sriram Ramanan, Bilal A Siddiqui, Francois Vigneault, Steven H

Kleinstein, David A Hafler, and Kevin C O'Connor. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.*, 6(248):248ra107, 6 August 2014.

[203] A H Sturtevant. Essays on evolution. i. on the effects of selection on mutation rate. *Q. Rev. Biol.*, 12(4):464–467, 1937.

[204] Way Sung, Matthew S Ackerman, Samuel F Miller, Thomas G Doak, and Michael Lynch. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 109(45):18488–18492, November 2012.

[205] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[206] Jeroen M J Tas, Luka Mesin, Giulia Pasqual, Sasha Targ, Johanne T Jacobsen, Yasuko M Mano, Casie S Chen, Jean-Claude Weill, Claude-Agnès Reynaud, Edward P Browne, Michael Meyer-Hermann, and Gabriel D Victora. Visualizing antibody affinity maturation in germinal centers. *Science*, 351(6277):1048–1054, 4 March 2016.

[207] E M Taylor and A R Lehmann. Conservation of eukaryotic DNA repair mechanisms. *Int. J. Radiat. Biol.*, 74(3):277–286, September 1998.

[208] Jacob A Tennessen, Abigail W Bigham, Timothy D O'Connor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, Hyun Min Kang, Daniel Jordan, Suzanne M Leal, Stacey Gabriel, Mark J Rieder, Goncalo Abecasis, David Altshuler, Deborah A Nickerson, Eric Boerwinkle, Shamil Sunyaev, Carlos D Bustamante, Michael J Bamshad, Joshua M Akey, Broad GO, Seattle GO, and NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, July 2012.

[209] Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.*, 49(2):303–309, February 2017.

[210] Jonathan Terhorst and Yun S Song. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc. Natl. Acad. Sci. U. S. A.*, 112(25):7677–7682, June 2015.

[211] Jonathan G Terhorst. *Demographic Inference from Large Samples: Theory and Methods*. PhD thesis, UC Berkeley, 2017.

[212] Gregg W C Thomas, Richard J Wang, Arthi Puri, R Alan Harris, Muthuswamy Raveendran, Daniel S T Hughes, Shwetha C Murali, Lawrence E Williams, Harsha Doddapaneni, Donna M Muzny, Richard A Gibbs, Christian R Abee, Mary R Galinski, Kim C Worley, Jeffrey Rogers, Predrag Radivojac, and Matthew W Hahn. Reproductive longevity predicts mutation rates in primates. *Curr. Biol.*, 28(19):3193–3197.e5, October 2018.

[213] Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Ann. Stat.*, 42(1):285–323, February 2014.

[214] Nguyen Dang Ton, Hidewaki Nakagawa, Nguyen Hai Ha, Nguyen Thuy Duong, Vu Phuong Nhung, Le Thi Thu Hien, Huynh Thi Thu Hue, Nguyen Huy Hoang, Jing Hao Wong, Kaoru Nakano, Kazuhiro Maejima, Aya Sasaki-Oku, Tatsuhiko Tsunoda, Akihiro Fujimoto, and Nong Van Hai. Whole genome sequencing and mutation rate analysis of trios with paternal dioxin exposure. *Hum. Mutat.*, 39(10):1384–1392, October 2018.

[215] Maxwell A Tracy, Mitchell B Lee, Brady L Hearn, Ian T Dowsett, Luke C Thurber, Jason Loo, Anisha M Loeb, Kent Preston, Miles I Tuncel, Niloufar Ghodsian, Anna Bode, Thao T Tang, Andy R Chia, and Alan J Herr. Spontaneous Polyploids and Antimutators Compete During the Evolution of Saccharomyces cerevisiae Mutator Cells. *Genetics*, 215(4):959–974, 08 2020.

[216] Arikuni Uchimura, Mayumi Higuchi, Yohei Minakuchi, Mizuki Ohno, Atsushi Toyoda, Asao Fujiyama, Ikuo Miura, Shigeharu Wakana, Jo Nishino, and Takeshi Yagi. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res.*, 25(8):1125–1134, August 2015.

[217] Marcos C Vieira, Daniel Zinder, and Sarah Cobey. Selection and Neutral Mutations Drive Pervasive Mutability Losses in Long-Lived Anti-HIV B-Cell Lineages. *Molecular Biology and Evolution*, 35(5):1135–1146, 02 2018.

[218] John Wakeley. *Coalescent theory: an introduction*. 2009.

[219] Berit Lindum Waltoft and Asger Hobolth. Non-parametric estimation of population size changes from the site frequency spectrum. *Stat. Appl. Genet. Mol. Biol.*, 17(3), June 2018.

[220] Richard J Wang, Samer I Al-Saffar, Jeffrey Rogers, and Matthew W Hahn. Human generation times across the past 250,000 years. *bioRxiv*, 2021.

[221] Benjamin Werner and Andrea Sottoriva. Variation of mutational burden in healthy human tissues suggests non-random strand segregation and allows measuring somatic mutation rates. *PLoS Comput. Biol.*, 14(6):e1006233, June 2018.

[222] Wendy S W Wong, Benjamin D Solomon, Dale L Bodian, Prachi Kothiyal, Greg Eley, Kathi C Huddleston, Robin Baker, Dzung C Thach, Ramaswamy K Iyer, Joseph G Vockley, and John E Niederhuber. New observations on maternal age effect on germline de novo mutations. *Nat. Commun.*, 7:10486, January 2016.

[223] Gur Yaari, Jason A Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel N H Stern, Kevin C O'Connor, David A Hafler, Uri Laserson, Francois Vigneault, and Steven H Kleinstein. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.*, 4:358, 15 November 2013.

[224] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

[225] Melinda A Yang, Xing Gao, Christoph Theunert, Haowen Tong, Ayinuer Aximu-Petri, Birgit Nickel, Montgomery Slatkin, Matthias Meyer, Svante Pääbo, Janet Kelso, et al. 40,000-year-old individual from asia provides insight into early population structure in eurasia. *Current Biology*, 27(20):3202–3208, 2017.

[226] Baixin Ye, Daniel Smerin, Qingping Gao, Chunsheng Kang, and Xiaoxing Xiong. High-throughput sequencing of the immune repertoire in oncology: Applications for clinical diagnosis, monitoring, and immunotherapies. *Cancer Letters*, 416:42–56, 2018.

[227] Kaixiong Ye and Zhenglong Gu. Recent advances in understanding the role of nutrition in human genome evolution. *Adv. Nutr.*, 2(6):486–496, November 2011.