January 2020

Inference Of Natural Selection In Human Populations And Cancers: Testing, Extending, And Complementing Dn/ds-Like Approaches

William Meyerson

Inference of Natural Selection in Human Populations and Cancers: Testing, Extending, and Complementing dN/dS-like Approaches

A Thesis Submitted to the Yale University School of Medicine in Partial Fulfillment of the Requirements for the Degree of Doctor of Medicine

by
William Ulysses Meyerson

2020

Abstract

Heritable traits tend to rise or fall in prevalence over time in accordance with their effect on survival and reproduction; this is the law of natural selection, the driving force behind speciation. Natural selection is both a *consequence* and (in cancer) a *cause* of disease. The new abundance of sequencing data has spurred the development of computational techniques to infer the strength of selection across a genome. One technique, dN/dS, compares mutation rates at mutation-tolerant synonymous sites with those at nonsynonymous sites to infer selection. This dissertation tests, extends, and complements dN/dS for inferring selection from sequencing data. First, I test whether the genomic community's understanding of mutational processes is sufficient to use synonymous mutations to set expectations for nonsynonymous mutations. Second, I extend a dN/dS-like approach to the noncoding genome, where dN/dS is otherwise undefined, using conservation data among mammals. Third, I use evolutionary theory to co-develop a new technique for inferring selection within an individual patient's tumor. Overall, this work advances our ability to infer selection pressure, prioritize disease-related genomic elements, and ultimately identify new therapeutic targets for patients suffering from a broad range of genetically-influenced diseases.

Acknowledgments

This dissertation would not be possible without the support of more individuals than can be named here.

The responsiveness of the Molecular Biophysics & Biochemistry MD Thesis Chair Dr. Alan Garen and members of the Office of Student Research, including Donna Corranzo and Dr. Erica Herzog and the support of the Office of Student Research at large, also including Reagin Carney, Dr. Sarwat Chaudhry, and Kelly Jo Carlson helped ensure a successful thesis review and submission process.

I thank my collaborators and colleagues: I thank the following lead authors for including me in their research: Matt Bailey, Sushant Kumar, Patrick McGillivray, Leonidas Salichos, Bo Wang, Daifeng Wang, and Jing Zhang. I also thank Declan Clarke, Li Ding, Donghoon Lee, Shantao Li, Jason Liu, Lucas Lochovsky, Inigo Martinocoreña, Joel Rozowsky, Michael Rutenberg-Schoenberg, and Jonathan Warrell for helpful discussions and all current and former lab members for building a welcoming, intellectual community at work.

I thank my funders and support: I thank Lori Ianicelli, whose administrative support has enabled me to participate in scientific conferences and get access to data sets. I thank Mihali Felipe for informatics support, sharing his expertise in troubleshooting, and together with Jason Liu, procuring the latest equipment to power my research. I thank Lisa Sobel, Hongyu Zhou, and Mark Gerstein for their leadership of my graduate program in Computational Biology & Bioinformatics. I thank Yale's MD-PhD Office including Cheryl DeFilippo, Reiko Fitzsimonds, Fred Gorelick, the late Jim Jamieson, Barbara Kazmierczak, and Susan Sansone for providing a home for us aspiring physician-scientists. I thank Yale's High Performance Computing center for maintaining the infrastructure for my research. I thank the National Institute of General Medical Sciences for salary and tuition support through the NIH/NIGMS T32 GM007205 training grant.

Most of all, I thank my advisor Mark Gerstein: Thank you, Mark, for your mentored guidance all these years, for including me in interesting group research projects, for connecting me with scientists, data, and questions in our field, for involving us all in lab management discussions, and for paying my way. Your unwavering support has made for an outstanding training experience, and I look forward to continued collaborations.

Table of Contents

Introduction

Content Attribution

This thesis describes work that first appeared in the following publications and preprints:

Meyerson W and Gerstein MB. Genomic variants concurrently listed in a somatic and a germline mutation database have implications for disease-variant discovery and genomic privacy. *bioRxiv* [**preprint**]. 2018.

Kumar S, Warrell J, Li S, McGillivray P, **Meyerson W**, *et al*. Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell*. 2020;180(5):915-927.e16.

Salichos L, **Meyerson W**, Warrell J, Gerstein M. Estimating growth patterns and driver effects in tumor evolution from individual samples. *Nat Commun*. 2020;11(1):732.

And in places builds on ideas developed in the following publications:

Bailey M, **Meyerson W**, *et al.* Comparison of exome variant calls in 746 cancer samples with joint whole exome and whole genome sequencing. *Nat Comm. Accepted.*

Zhang J, Lee D, Dhiman V, Jiang P, Xu J, McGillivray P, Yang H, Liu J, **Meyerson W**, *et al.* An integrative ENCODE resource for cancer genomics. *Nat Comm. Accepted*

Wang B, Yan C, Lou S, Emani P, Li B, Xu M, Kong X, **Meyerson W**, Yang YT, Lee D, Gerstein MB. Integrating genetic and structural features: building a hybrid physical-statistical classifier for variants related to protein-drug interactions. *Structure*. (2019).

Navarro F, Mohsen H, Yan C, Li S, Gu M, **Meyerson W**, Gerstein MB. Genomics and data science: an application within an umbrella. *Genome Biology* (2019). 20:109.

McGillivray P, Clarke D, **Meyerson W**, Zhang J, Lee D, Gu M, Kumar S, Zhou H, Gerstein MB. Network Analysis as a Grand Unifier in Biomedical Data Science. (2018). *Annual Review of Biomedical Data Science* Vol. 1.

Wang D, Yan KK, Sisu C, Cheng C, Rozowsky J, **Meyerson W**, Gerstein MB. Loregic: a method to characterize the cooperative logic of regulatory factors. (2015). *PLoS Comput Biol* 11: e1004132.

Natural selection is a fundamental biological force

What does it mean to be alive? A humanist might say that to be alive is to feel, think, and make choices. Certainly, this is our experience of what it is like to be alive. But our lived experience as cognitively complex creatures and the prehistorical process that got us here depend on a more basic set of functions studied by biologists: the ability of various assemblies of organic matter to create local order from external energy, to self-reproduce, and to limitlessly adapt. This final defining property is the most unique to life, for even crystals can focus order and self-reproduce. We humans are a dramatic consequence of adaptation, a far cry from the single-celled organism from which we evolved.

We now know that our evolutionary history proceeded as a series of errors in the replication of some primordial instruction manual for life –a primordial genome. These errors (mutations) were initially random, but the mutations that endured and spread tended to be those that assisted in the survival and reproduction of the creatures who inherited them, which included humans, all plants and animals, stranger things, and single-celled creatures more efficient or specialized than their ancestors. This the spread of fitness-enhancing mutations is termed positive selection; the disappearance of fitness-reducing mutations is termed negative selection; and the random fixation of neutral mutations is neutral drift.

Over evolutionary time scales, natural selection drives speciation, but over shorter time scales, natural selection is most relevant to humans in its role as both a consequence and cause of disease. Natural selection is a *consequence* of disease when individuals with heritable disease do not survive to adulthood; the mutations in those individuals are not inherited by the next generation, making those mutations therefore subject to negative selection. Natural selection is a *cause* of disease when runaway positive selection drives unchecked growth of cancer cells at the host's expense. In either case, by identifying and quantifying selection in the human genome, we implicate mutations in disease and identify the genes in which these mutations matter most. Identifying disease genes guides the rational development and deployment of targeted therapies.

Computational genomics can be used to quantitatively infer natural selection in action

Natural selection takes place over generations, so it is most straightforwardly studied when time has elapsed. However, achieving sufficient time elapse is not always possible. Prospective field studies can watch selection in action by charting the rise and fall of subgroups within a population across generations in response to an environmental change, but these studies are laborious and long, are only practical in short-lived species, and are susceptible to confounding by population structure and secondary environmental disturbances.[i] A more rigorous form of prospective study is the controlled experiment, such as with CRISPR gene-editing to induce selection,[ii] but these studies are ethically contentious if performed in the human germline[iii], and even in model organisms and cultured human cells they are expensive, have been argued to suffer from off-target effects[iv], and may not be faithful to *in vivo* human biology. Retrospective studies can use archaeological evidence and ancient DNA to reconstruct the historical action of natural selection.[v] These retrospective studies can effectively acquire very long follow-up times in human populations, but ancient samples are a scarce resource; moreover, environmental confounders across millennia of human history are especially

problematic because they are many and unknown. For these and other logistic and ethical reasons, the bulk of human genomic data available today for the inference of selection comes from large observational cohorts that can be taken to represent a single point in time.

The central challenge of inferring selection from a *snapshot* of the human gene pool is that the mutations present today represent the combined effects of mutation generation and natural selection, and we wish to disentangle the two. One popular analytic approach to distinguishing these processes is to compare the ratio of the observed number of nonsynonymous mutations in a gene to expectations derived from the patterns in synonymous mutations in the same sample set, the *dN/dS* approach.[vi,vii]

To appreciate how dN/dS separates natural selection from mutation generation, we need to understand the workings of cells and the consequences of mutation. A cell, the smallest self-sustaining unit of life today, is a city of proteins dissolved in water, contained within a tiny sac. The thousands of kinds of proteins each have their own job to support the cell in energy processing, movement, reproduction, and other tasks. These proteins are in turn made of chains of 20 different kinds of amino acid building blocks, whose pattern determines the protein's chemical properties and therefore function. The sequence of these amino acids is defined by the sequence of nucleotides in the corresponding genes of the cell's genome, but critically for dN/dS, there is redundancy in the genetic code: most amino acids can be equivalently coded for by any of multiple nucleotide sequences. This leads to the distinction between synonymous mutations, which change a nucleotide while preserving the coded amino acid, and nonsynonymous mutations, which change both a nucleotide and the coded amino acid – and therefore the chemical properties and possibly function of the resulting protein.

In dN/dS approaches, synonymous mutations are assumed to be fitness-neutral, and used to isolate the mutation generation effect in the absence of selection. The estimated mutation generation rate at synonymous sites is then extrapolated to nonsynonymous sites. Differences between observed mutation rates at nonsynonymous sites and the neutral expectations for those sites extrapolated from the synonymous baseline are interpreted as selection-driven changes in the fixation rate of the nonsynonymous variants. These differences can be aggregated by gene to estimate the extent of selection acting on each gene in the genome.

In more advanced versions of the dN/dS technique, nonsynonymous mutations are only compared against synonymous mutations bearing the same nucleotide context (involved and sometimes neighboring bases) because nucleotide context is known to powerfully impact mutation generation rates and systematically differs between synonymous and nonsynonymous mutations.[viii,ix] Other advanced modifications of the technique involve estimating neutral mutation rates separately for each genomic region or gene and smoothing neutral expectations across genes with similar epigenetic features.

Part I: The assumptions of dN/dS

dN/dS is able to make causal claims from observational data by making assumptions about the relations within those data. Specifically, dN/dS assumes that 1) nonsynonymous mutations are more likely to have fitness effects than are synonymous mutations; 2) selection primarily acts on mutations affecting fitness; 3) input data are accurate; 4) the mutation generation rate at synonymous sites is comparable to that at nonsynonymous sites; and 5)

mutations affecting the same gene have fitness effects that tend in the same direction. These premises are biologically-motivated but hold to varying and uncertain extents. In this Introduction to Part I, I comment on the first three of these assumptions and set the stage for my investigation of the fourth. The fifth assumption is addressed in Part III.

The fitness effects of synonymous mutations are dwarfed by those of nonsynonymous mutations

The premise that nonsynonymous mutations have a much higher average fitness impact than synonymous mutations has been well-established. ClinVar is a compendium of assertions about the clinical significance of human mutations, and it is endorsed by the American College of Medical Genetics and Genomics.[x] At time of submission of this dissertation, ClinVar lists 19,743 pathogenic nonsynonymous variants but only 307 pathogenic synonymous variants, a 30-fold enrichment after taking into account the larger number of nonsynonymous than synonymous sites in the genome. To be clear, the number of pathogenic synonymous variants is greater than zero, and research articles describing the functional relevance of synonymous variants receive attention in the scientific literature, and have included roles in splicing, transcript factor and microRNA binding, protein translation efficiency, and mRNA folding, but these are exceptions that prove the rule.[xi] Of these roles, the most important in humans is splicing variants, so state-of-the-art dN/dS excludes splice-site synonymous variants from the synonymous sites used in analysis.[xii] There is better evidence for the fitness importance of minute changes in protein translation efficiency in prokaryotes (which lack introns) than in eukaryotes.[xiii]

Selection has indirect effects on fitness-neutral sites

The premise that selection primarily acts on fitness-affecting sites is generally but incompletely true. In both the germline and somatic settings, the evolutionary fates of fitness-affecting sites are linked to those of some sites that do not affect fitness, albeit with different patterns. In the germline, subjects who inherit a fitness-impacting variant are very likely to inherit all the other variants in a neighboring genomic region that were present in the ancestor in whom that fitness-impacting variant first arose; this is due to the fact that genome chunks are inherited together in the germline. This linkage disequilibrium can, for example, cause positive selection to indirectly increase the population frequency of a fitness-neutral variant located near a fitness-positive variant. Nonetheless, this linkage disequilibrium decays over evolutionary time, and the co-inherited blocks become progressively smaller though crossing over during meiotic recombination. Moreover, in the somatic setting, linkage between a fitness-affecting variant and concurrent hitchhiker mutations is complete, because crossing over does not occur in the somatic setting. Nonetheless, in the somatic setting, the set of variants that hitchhike with a fitness-affecting variant are effectively random and will differ from patient to patient, such that in aggregate analyses, the signal of selection on neutral sites will tend to average out while the selection on truly fitness-impacting sites will be amplified.

Sequencing is a mature technology that falters in special cases

Next generation sequencing technologies and associated variant callers support research-grade variant calls, but these calls have errors. For example, oxidation of sequencing products leads to mutations in the sequencing samples that were not present before sequencing, although recent computational approaches can help to identify and remove these artefactual mutations.[xiv] Moreover, the short reads of sequencing lead to mapping errors, especially in repetitive regions of the genome.[xv] I have conducted original research with Dr. Bailey *et al*. that shows that Illumina Hi-Seq sequencing of somatic tissue is highly reproducible except for mutations that are present in a small fraction of the cells in the somatic sample.[xvi] I have also conducted original research that shows that about 1% of somatic variants called in a premier somatic sequencing project most likely represent misclassification of variants that are actually derived from the patient's germline.[13] Moreover, different variant-calling strategies yield slightly different calls: Illumina short read sequencing of germline samples followed by variant calling with three different platforms resulted in SNV calls that are 92% concordant.[xvii]

Uncertainty about mutation generation rates would, if large, invalidate dN/dS

The comparison at the heart of dN/dS -that of the nonsynonymous mutation against the synonymous- presupposes that these mutation types are compar*able*. Specifically, this comparison depends on synonymous and nonsynonymous sites being similarly prone to mutation generation; or, as this is demonstrably false[xviii], that the differences between mutation generation rates at synonymous and nonsynonymous sites can be adjusted for. In turn, proper adjustment depends on a knowledge of the causal determinants of mutation generation rates, lest we create new confounders by conditioning on colliders.[xix] The steady hum of new publications describing novel determinants of mutation rates sets the expectation that undiscovered determinants of mutation rates abound, making our knowledge of what to adjust for incomplete.[xx,xxi] Does this mean that dN/dS and related methods are invalid? The answer depends on how important undiscovered determinants of mutation rates are relative to discovered determinants, which in turn seems at first to be an unknowable quantity: who are we to say what future scientists may or may not discover? In this Part, I set out to answer this seemingly unknowable question.

Part II: Selection in the noncoding genome

The portion of the human genome that directly codes for protein represents a mere 1% of genomic real estate. The other 99% of the genome contains promoters, enhancers, and noncoding RNAs that regulate the expression of protein coding genes; pseudogenes which are the genomic fossils of inactivated genes and potent raw materials for new genes; telomeres and centromeres which provide structural support for chromosomes; and vast stretches of transposable "junk" DNA that selfishly replicates.

The size of the noncoding genome varies tremendously among eukaryotes. The human noncoding genome is over 3 billion base pairs. *Utricularia gibba* is a carnivorous plant that lives in phosphate-starved waters where DNA is expensive: its noncoding genome is merely 3% of base pairs; apparently in appropriately adapted eukaryotic organisms, much of the noncoding genome is dispensable.[xxii] Meanwhile, the Australian lungfish, a close relative to the direct

ancestor of all terrestrial animals, has a noncoding genome 15 times larger than the human noncoding genome without marked change in the size of the coding genome.[xxiii]

The coding genome has always been the main focus of the scientific community, for good reason: the average coding base pair has a much greater functional impact than the average noncoding base pair, the coding genome is more interpretable, and exome sequencing which focuses on the coding genome and related portions has been much cheaper than whole genome sequencing. However, as the cost of sequencing falls, whole genome sequencing is becoming more cost effective. Regardless, to push the scientific frontier further, we must wrestle with the dark matter of the genome.

The noncoding genome's unique features and sheer size motivates the development of new computational tools tailored for this region. In particular, dN/dS is only defined in protein coding regions of the genome because synonymous and nonsynonymous mutations refer to the direct impact of the mutation on the coded protein. I sought to develop an analogue of dN/dS that is suitable for exploring selection pressure in the noncoding genome.

Part III: dN/dS is insufficient for Personalized Medicine

While dN/dS plays an important role in prioritizing the development of new targeted therapeutic agents, the personalized deployment of these agents in the clinical setting requires new approaches.

Genewise dN/dS provides one selection coefficient per gene. Genewise dN/dS does not distinguish between genes that have weak, diffuse selection throughout, vs genes in which most variants are neutral but a few variants are highly subject to selection, nor in the latter case does genewise dN/dS indicate which variants are those subject to strong selection. In theory, these limitations could be partly resolved by dividing genes into sub-gene regions up to the limit of available sample sizes. However, even with sub-gene dN/dS, due to incomplete penetrance and epistatic and environmental interactions, a variant strongly selected in one patient could be weakly selected in another patient.

In the clinical setting, the choice of targeted therapeutic agent depends on which mutations are causally contributing to the condition of the patient at hand. In the present day, these targeted therapies are most common in cancer patients. When a cancer patient has a variant of unknown significance in a gene for which targeted therapies exist, the clinician is uncertain whether to start the patient on the targeted therapy or to default to chemotherapies with broader cancer coverage but greater side effects. In this situation, genewise dN/dS scores are of limited further utility. For these situations, which will become increasingly common as the repertoire of targeted agents expands, it will be a clinical priority to estimate the selection pressure on individual variants in individual patients using data types that can be readily produced in the clinical setting.

Natural selection would seem to not to be inferable in a single individual

However, natural selection cannot be meaningfully observed acting in any one individual, because chance and the environment play a large role in the survival and reproduction of any one individual. Therefore, natural selection must be inferred from populations, presenting a conceptual obstacle in the inference of selection for personalized

medicine. An important exception is that, for somatic diseases like cancer, one patient contains a whole population of cells, which offers the necessary sample sizes for inferring selection. In Part III, I describe a new analytic procedure I developed with colleagues Dr. Leonidas Salichos *et al*. that exploits the multicellular nature of tumors to perform patient-level, variant-level inferences of selection pressure from bulk somatic DNA sequencing samples

The billions of cells present in a single tumor allow natural selection to be inferred in a single tumor

A single tumor is a vast, evolving population. A tumor is composed of cancer cells, potentially trillions in number, that descend from a single, transformed ancestral tumor cell. Untransformed stromal and vascular tissue may also infiltrate the tumor; these have a different ancestor. Each cell division within the tumor involves the error-prone recopying of an existing cell's genome, which creates on average three new SNVs in the daughter cells, which otherwise inherit all the mutations present in their parent cell. Because of this process, mutations that arise in the earliest descendants of the ancestral cell will be inherited by a large fraction of the tumor, whereas mutations that arise late will spread to a small fraction of the tumor, assuming no cell death. When there is cell death, it becomes possible for mutations that arise at any point in the tumor's history to fixate throughout the whole tumor or fall to extinction. Mutations that increase tumor cell fitness will tend towards fixation, and mutations that decrease tumor cell fitness will have a (slight) tendency towards extinction. The large population size and evolutionary dynamics of individual tumor cells within a tumor allow for natural selection to run its course and - in theory, be detected - within a single cancer patient.

Our ability to probe the evolutionary substructure of a tumor comes from the sampling properties of next generation sequencing technologies. Standard Illumina next generation sequencing of somatic tissue involves the digestion of tumor DNA from one million cells into fragments, and the sequencing of these fragments to an average read depth of 60-100. This means that a given site on the genome will tend to have 60-100 sequenced reads that overlap that site. Critically these 60-100 sequenced reads will come from different cells in the tumor. In this way, the fraction of reads overlapping a site that bear a given mutation at that site (the *variant allele frequency*), relates to the fraction of tumor cells that carry the mutation. Because the human genome is diploid (two copies of each chromosome), the fraction of tumor cells that bear a given mutation is approximately twice the variant allele frequency of the mutation. Sampling noise, tumor impurities, and copy number alterations distort this relationship. When these sources of error can be controlled or ignored, variant allele frequency gives us a window into the intra-tumor dynamics on which selection acts.

Dr. Salichos demonstrates that under the assumption of exponential growth, a tumor lacking intra-tumor selection will exhibit a characteristic variant allele frequency distribution. Specifically, a mutation arising in a given tumor cell will tend to reach a variant allele frequency half that of a mutation arising in the cell's direct parent. This is because, after cell-division, the parent cell becomes two daughter cells of stipulated equal fitness, and the same process holds across all cancer cells of a given generation, effectively doubling the tumor's size and halving the share any one mutation has of all extant cells in the tumor. By corollary, a mutation in a tumor cell will tend to reach a variant allele frequency a quarter of that of one arising in the

cell's grandparent, and so on. These patterns, *the hitchhiker equations*, form the expectations for neutrality in an exponentially growing tumor. In contrast, a mutation subject to positive selection will tend to achieve a variant allele frequency more than half that of its parent, because its descendants will tend to outgrow its sister and cousin cells. Similarly, a mutation subject to negative selection will tend to achieve a variant allele frequency less than half that of its parent. These departures are predicted to be asymmetrical: a mutation arising in the daughter cell of a fitness-enhancing mutation will be expected to achieve half the VAF of its parent because it will have inherited the parent's fitness enhancing mutation. In this way, Dr. Salichos argues that a discontinuity in the VAF slope around a mutation can be a signal of that mutation's fitness impact.

The hitchhiker equations theoretically relate the VAF of hitchhiker mutations to the fitness effect (k) of the subclonal driver with which they are hitchhiking, mediated by various growth parameters. The fitness effect of subclonal drivers is not directly observable but is of primary biomedical importance. The VAF of hitchhiker mutations is directly observable, and we wanted to use these VAFs to estimate the fitness effect of subclonal drivers. Our approach is to fit the known VAFs of the hitchhiker mutations to the hitchhiker equations to estimate the fitness effect of subclonal drivers. This requires simultaneous estimation of the various other growth parameters, which we perform using non-linear least-squares optimization. To address the fact that real tumors differ from idealized behavior, we make use of sliding windows in the parameter estimation to prevent departures from idealized behavior in one part of the VAF spectrum from interfering with parameter estimation in other parts of the VAF spectrum.

Simulations are a well-established tool in bioinformatics generally and tumor evolution specifically

Simulation has a long history in the study of tumor evolution. Tumors lend themselves to simulation because they are composed of discrete cells whose essential behavior can be approximated by simple rules. The simplest models of tumor evolution allow tumor cells to divide and die. These events can happen one at a time or in large synchronous waves. In the simplest models, the simulations need only keep track of how many cells of each type exist at each time. Simpler models are more tractable, can be grown to large tumor sizes, are less prone to software bugs, and are easier for other scientists to understand and reproduce.

More complex models incorporate numerous other factors that modify or enrich the birth and death dynamics of simulated tumors. Complex models may, for example, incorporate effects from new mutations, spatial structure, vascularization, migration, differentiation, and even game theoretical interactions between tumor cells. The incorporation of these effects may require the storage and processing of information about each cell in the tumor, which is vastly more computationally intensive than merely keeping track of cell counts and precludes the simulation of large tumors. More complex models are more flexible, and allow more biological questions to be asked. In principle, more complex models can better represent the complexity of real tumors but only insofar as the relevant biological parameters are known, which is often not the case.

For our simulations, we chose to use the simplest model that captures the essential evolutionary nature of real tumors and that allows us to ask our target question. In this

philosophy, we follow in the tradition of Williams et al. and McFarland et al., whose examples are instructive:

Williams et al. 2016 sought to examine whether neutral evolution could account for the VAF distribution observed in real tumors. Williams et al. simulated neutrally evolving tumors as a branching process beginning with a single tumor cell where in each generation, each cell divides into two daughter cells until the tumor reaches 1,024 cells. The number of new mutations that arise in each daughter cell is simulated as a random draw from a Poisson distribution with fixed mean. The authors showed that the VAF distributions that result from these neutral simulations are sufficiently similar to observed VAF distributions from many tumors, which they used to controversially argue that tumors are frequently subject to neutral evolution.

McFarland et al. 2013 simulated growing tumors to assess the impact of deleterious passenger mutations on tumor growth dynamics. The authors employed a logistic growth model wherein the death rate is proportional to the population size and the birth rate per cell increases exponentially with the number of driver mutations and decreases exponentially with the number of deleterious passenger mutations present in the cell. Birth and death events were then sampled using the Gillespie algorithm applied to these rates. The rate of occurrence of deleterious passenger mutations was set at 500,000 * 10^-8 per cell division and for driver mutations was 700 * 10^-8 per cell division. Driver strength was chosen to increase growth by 10%, and deleterious passenger strength was chosen to decrease growth by 0.1%. Tumors were grown until reaching 1 million cells. The authors find that this model leads to fixation of deleterious passengers and tumors that regress unless the population reaches a critical size, which they argue resembles the clinical behavior of real tumors.

Statement of Purpose

The overall purpose of this work is to develop our understanding of and ability to infer evolutionary selection pressure from human genomic sequencing data.

Part I: Estimate the uncertainty in our knowledge of human mutation generation rates using variants shared between the germline and somatic settings.

The inference of natural selection from human genomic sequencing data depends on our ability to precisely estimate mutation generation rates to distinguish which mutation patterns are due to generation effects vs selective effects. Part I assesses how confident the scientific community should be in our estimates of mutation generation rates from the perspective of accounting for variants shared between two disparate contexts of human mutation.

Part II: Devise a new measure to quantify evolutionary selection pressure in noncoding regions.

A popular measure of selection pressure compares the observed rate of nonsynonymous mutations to synonymous mutations, and to attribute the corrected difference to selective effects. However, this approach is only possible in coding regions of the genome, which make up less than 2% of genomic real estate. Part II extends this general approach to the noncoding genome using a ratio of the mutation rate at conserved vs unconserved sites in the mammalian genome.

Part III: Benchmark a framework for estimating the fitness effects of individual mutations from individual tumors.

Claims regarding natural selection are usually only possible to apply to averages of large cohorts of individuals, which is insufficient for personalized medicine. Tumors represent a special case where the billions of cells within a single tumor provide the sample sizes necessary to make claims about evolutionary forces within a single individual human subject. Part III conducts simulations to benchmark a colleague's method of applying evolutionary theory to populations of cancer cells to identify cancer driving mutations.

Methods

Part I

Overall approach to estimate our uncertainty of mutation generation rates

This Part develops and applies an approach to quantify our uncertainty about mutation generation rates. The overall approach I take is to estimate the total implied heterogeneity in mutation rates and then back out the portion of that heterogeneity that can be explained by known factors. To estimate the total implied heterogeneity in mutation rates, I adapt a form of recurrence analysis pioneered by Hodgkinson et al. I test the power of this approach using simulations. To back out the portion of heterogeneity that can be explained by known factors, I develop a new statistical measure that requires separating a data set into partitions defined by levels of a categorical variable. I apply this approach to two large, carefully filtered genomic databases. A number of supporting analyses were conducted as part of this study but receive only passing mention in this dissertation.

The number of mutations that arise independently in multiple samples can be used to estimate the total implied heterogeneity of mutation generation rates

A certain number of mutations are expected to arise independently in multiple samples by chance. Between two genomic samples (or sets of genomic samples), there is an expectation that a certain number of variants will occur in both. If the genomic locations of these mutations in these samples were drawn statistically independently, then the expected fraction of mutations that co-occur in both samples is the fraction of sites that are mutated in one sample times the fraction of sites that are mutated in the other.

When the actual number of recurrent mutations significantly exceeds these expectations, this indicates heterogeneity in the mutation rate. If the fraction of co-occurring mutations were significantly higher than the expectation, this would negate the hypothesis that two samples are statistically independent and would instead indicate that the same sites tend to be mutated in the two samples. In turn, this implies a certain amount of mutation rate heterogeneity, with some sites receiving multiple mutations and other sites receiving no mutations.

Recurrence of single nucleotide variants (SNVs) captures the aggregate effect of both discovered and undiscovered determinants of mutation rates. An SNV is the smallest unit of mutation that can co-occur across samples. The possible SNV of Chr 1, position 1,000,000, from A-> C will approximately share the following features between two samples: identities of involved and adjacent bases, distance from telomeres, distance from any nucleotide motifs, local curvature of DNA, relative order within a codon, distance from coding regions, both being on Chromosome 1, and local DNA melting temperature, to name nine. An SNV being mutated in multiple samples could arise because any of these attributes might be impacting its mutation rate.

Under special circumstances, heterogeneity in the mutation rate can be approximately equated with heterogeneity in the mutation generation rate. Heterogeneity in mutation rate can signify multiple things: that mutation generation rates across genomic sites are correlated between the samples; that mutation selection pressure is correlated between the samples; or that some mutations in the samples are identical by common descent. I set up my two comparison sets for the recurrence analysis as the somatic and germline settings to minimize the relevance of the

latter two explanations (mutations in germline will not be identical by descent with true somatic variants, and mutations positively selected in the soma will generally not be positively selected in the germline), to focus on the portion of excess recurrence that is explained by correlated mutation rates across genomic sites between the two sets. When an SNV is concurrently mutated in the germline genome of one subject and the somatic genome of another, I term this a cSNV.
When these circumstances are met, this estimate is a lower bound. Determinants of mutation rates that depend on tissue exposures, unique epigenetic features, or particular mutation repair deficiencies will not reproducibly affect the same sites in different samples and tissues; thus, recurrence analysis across samples is incompletely sensitive.

Simulations were used to benchmark this approach

Simulations were conducted to test the power of cSNV enrichment for detecting mutational processes shared between the somatic and germline genomes. In these simulations, at baseline, each genomic site of the human reference genome had the same fixed probability of being mutated in a simulated germline database and a different but homogenous probability of being mutated in a simulated somatic database. On top of this baseline, we simulated one arbitrary mutational process per run. These mutational processes affected an arbitrary subset of the genome and had the effect of increasing or decreasing the probability of mutation at affected sites by the same constant amount. We enforced these mutational processes to be completely shared between the soma and germline by ensuring that the same genomic sites were affected in the simulated soma and germline and that the mutation rate multiplier was identical in the somatic and germline draws of a given run.
Fixed parameters were chosen to resemble the actual eligible genome size and number of somatic and germline variants; that is: 50,639,535 potential SNVs, 3,339,715 germline SNVs, and 1,309,369 somatic SNVs, implying 86,354 expected cSNVs. Across different runs of the simulation, two additional parameters were allowed to vary: the fraction of sites affected by the mutation process, and the impact that the mutation process had on the mutation rates of affected bases. To efficiently sweep through a broad range of possibilities, the fraction of sites affected by the mutation process took on the following values in different runs: ½^10, ½^9, ½^8, …, ½^1, 1-½^2, 1-½^3, …, 1-½^10. The mutation rate multiplier of affected sites took on the following values in different runs: 2^-7, 2^-6, 2^-5, … 2^0, 2^1, … 2^7, and all combinations of these parameters were tested. After each run, the cSNV enrichment was calculated as defined in the preceding section.

A new statistic to quantify the portion of explainable heterogeneity

I developed a statistical framework that estimates the portion of excess cSNVs that may be accounted for by categorical variables associated with both somatic and germline mutation rates. This measure, the partitional dependence of F on categorical variable v (such as nucleotide context), is given by

$$F_p(v) = 1 - \frac{P(v) - 1}{F - 1}$$

where P(v), the partition-conditioned Forbes coefficient is given by

$$P(v) = \frac{c}{\sum_{i=1}^{m} e_i}$$

where c is the number of cSNVs, m is the number of factor levels of categorical variable v and

$$e_i = {}^{g_i}/_{n_i} * {}^{s_i}/_{n_i} * n_i$$

and where g_i, s_i, and n_i are the number of G events, S events, and total elements of the ith partition of the full domain.

When $F_p(v)$ = 0, variable v does not account for any observed excess or shortage of shared variants. $F_p(v)$ = 1, variable v can account for all of the observed excess or shortage of shared variants. Although the current application is for cSNVs, this framework is valid much more generally to co-occurrence of any kind of event between two settings.

I applied this framework to well-chosen subsets of two large variant databases

This approach requires dense sets of SNVs. In the original Hodgkinson et al. study, these dense SNV sets were achieved by comparing divergences that arose in the chimp reference genome with respect to the human-chimp ancestor with germline polymorphisms in human populations. This interspecies comparison came at the cost of alignment errors and concerns that some minor alleles in human populations preceded the human-chimp split. My use of human-only variant databases avoids these problems.

The main databases used were large, high-quality public databases of somatic and germline variants

The Cancer Genome Atlas (TCGA) is a database of variants called in over 10,000 somatic whole exomes from cancer patients. Mutations in these samples were recently uniformly recalled by a pancancer mutation reanalysis project in Spring 2018. The genome aggregation consortium (gnomAD) is a database of variants called in 125,748 germline whole exomes from healthy subjects and adult patients without severe childhood disease. For corroboration, a database of de novo germline variants (denovo-db) and a panel of healthy somatic tissues were also analyzed, but their results are not discussed here.

The eligible variants for analysis were filtered down to remove technical artifacts

Two kinds of filters were used: filtration of sites and filtration of variants. Filtered sites are removed from both the numerator and denominator of mutation rate calculations. Filtered variants are removed from just the numerator of mutation rate calculations. Sites of common human polymorphisms were excluded as were somatic variants with low matched normal read depth, to reduce the impact of germline contamination. Non-uniquely mappable sites and sites listed in the Duke and DAC Encode blacklists were excluded to avoid mapping errors. Sites falling outside of the Broad and GATK exome interval lists were removed to minimize the impact of differences in exome capture. Sites with low read depth in either database were removed to reduce the impact of coverage bias.

Variants were partitioned into nucleotide contexts and genomic regions to apply my statistical framework

The statistical plan was to quantify the total heterogeneity in mutation rates, and then back out the portion of that heterogeneity that could be attributed to discovered determinants of mutation rates. This required assigning each potential variant of the eligible reference genome to one of several levels of a categorical variable. In this analysis there were two model choices to make: 1) what genomic feature to use as the categorical variable, and 2) how many levels of the categorical variable to consider.

There is a long list of features known to affect mutation rates, but I wanted to prioritize which ones to test first. For this, I chose nucleotide context as the first determinant of mutation rates to back out, and genomic region as the second, because these were the features upheld by the premier genomic consortium PanCancer Analysis of Whole Genomes as the first most important and second most important determinants of mutation rates, and the only determinants they adjust for in their models of neutral mutation. In a third set of runs, I create a combined categorical variable of nucleotide context-genomic region pairs.

Instead of pre-specifying how many levels to include for each categorical variable upfront, I opted to separately test the impact of using fewer or more levels of each categorical variable. When testing nucleotide context with the fewest levels, I consider the reference and alternate allele only. In a nucleotide context test with more levels, I considered also whether the potential variant involved a CpG->T transition, since these have an especially high mutation rate in mammals. With more levels still, I include the immediate 5' and 3' flanking bases; then two flanking bases on either side; and finally three flanking bases on either side, which is the precision limit of what has been tested in high profile genomic studies. Each eligible site of the human exome was annotated with nucleotide context using the GenomicRanges R package.

For genomic region, I considered chromosome only as the fewest number of levels; and going up from there in number of levels: 10 MB regions, 1 MB regions, 100 kb region, and finally 10 kb regions as most precise.

Part II

Developing an analogue of dN/dS for the noncoding genome

The distinction between synonymous and nonsynonymous sites amounts to a prior assessment about that fitness impact of variants of those sites. To extend a dN/dS-like approach, to the noncoding genome, I sought an alternative prior for fitness impact that is defined in both the coding and noncoding genomes. I chose to use the Genome Evolutionary Rate Profiling (GERP) score from Sidow et al. because it is consistently and simply defined across both coding and noncoding regions of the genome. The GERP score is site-specific measure of evolutionary conservation among mammals. The GERP framework looks at each site of the genome is an ancestral shrew-like mammal, and compares how many descendant mammal species have diverged from the ancestral allele vs how many were expected from a neutral mutation model. If more mammals retain the ancestral allele than expected, then the site is considered subject to purifying selection and the GERP score is positive. In theory a negative GERP score could mean that divergences from the mammalian ancestor are positively selected, but the GERP

authors argue convincingly that a GERP score less than 0 should be treated as a neutrally evolving site.
From the GERP score, I formulated a dC/dU statistic, analogous to dN/dS. In dC/dU, the ratio of conserved mutations is compared against expectations for the number of conserved mutations based on the number of unconserved mutations and trinucleotide adjustment. Sites are considered conserved if the GERP score was greater than 0. In theory sites with GERP score less than 0 should be positively selected, but the GERP authors argue that sites with GERP score less than 0 should be treated as unconserved.

Applying dC/dU to detect negative selection in the noncoding genome in cancer

As part of an analysis Kumar et al. for the PCAWG consortium, I applied dC/dU to more than 2500 somatic whole genomes from cancer patients. In this work, we only considered variants from PCAWG samples which satisfy Pan Cancer Analysis of Whole Genomes-(PCAWG)wide quality control (QC) criteria and were included in the final release. Additionally, certain cohorts such as prostate adenocarcinoma and lymphomas had replicates for certain patients. Among patients with multiple replicate samples, only one sample with the best overall QC metric was included per patient. Moreover, 38 hyper-mutated melanoma and lymphoma samples were excluded in this work. In total, we used variant calls from 2548 PCAWG samples for our analysis. Variants were annotated with a set of features that included GERP scores by Dr. Kumar. Sites were considered conserved if GERP was greater than 0, otherwise conserved. Coding regions were excluded to focus on the noncoding genome. The noncoding regulators of known or suspected cancer driver genes were removed to avoid canceling out the signal of negative selection with positive selection.

Part III

Simulations for benchmarking a new evolutionary tool

Dr. Salichos's insight into the theoretical VAF consequences of fitness-impacting mutations is an important theoretical development. This theory is based on a deterministic, idealized model of tumor growth. Real tumors, however, grow stochastically and otherwise depart from idealized behavior. Dr. Salichos developed a software package EvoTum that is intended to infer the identities and fitness effects of subclonal drivers, but it was unclear whether the stochastic and non-ideal behavior of tumors would lend themselves to analysis with software designed from a deterministic, idealized model. I recognized the need to benchmark this software using simulations to test how robust EvoTum is to stochasticity and departures from idealized behavior.

Tumors were simulated as a stochastic, time-branching process

Dr. Salichos's software takes as input the sorted VAFs of mutations within a tumor, and estimates the VAF decay within sliding windows. Then the software fits the hitchhiker equations to these VAF distributions to assign a most probable driver mutation and estimate the strength of selection on that hitchhiker, using a maximum likelihood approach.

To benchmark this software, I simulated evolving tumors. The simulations start with a single cancer cell at time 0. This cell represents the ancestor of the "clonal" lineage. All of its descendants are considered "clonal" until a second, simulated driver mutation arises; at that point, the cell in which that second driver mutation arises and all its descendants are termed "subclonal" cells. Each time step, three quantities are sampled: the kind of event that occurs, which origin cell the event applies to, and the time elapse since the previous event. There are seven kinds of events: 1) cell-division of a clonal cell, 2) spawning of a subclonal driver from a clonal cell, 3) cell-division of a cell bearing the subclonal driver, 4) the death of a clonal cell, 5) the death of a subclonal cell, 6) quiescence of a clonal cell, and 7) quiescence of a subclonal cell. Quiescent cells endure and contribute DNA to the final sequenced cohort but lose the capacity to divide, which is a frequent cell state in biological tumors. The design maxim of the simulations is for them to be no more complicated than they need to be, but it turns out that without the quiescent cells, mutations collapse too frequently onto the same VAFs as each other, which is not biologically valid and not optimal for Dr. Salichos's software. The rates at which these seven events occur are governed by logistic growth equations.

Each event has a particular kind of cell that must serve as the "parental" cell. The cell-division of a clonal cell, spawning of a subclonal driver, death of a clonal cell, and quiescence of a clonal cell all occur to clonal cells. Cell division, death, and quiescence of a subclonal cell, must occur to a subclonal cell. Once the event for a time-step is chosen, the particular parental cell is randomly sampled from the pool of available cells of the relevant parental class.

Each time-step also requires the determination of how much time has elapsed since the previous time-step. If time-steps are too large, the size of the sub-populations of cells used to calculate reaction rates are no longer valid, because the sub-populations should have changed in size within the large time step. In contrast, if time-steps are too small, then the simulation proceeds too slowly to be computationally tractable. The key is to balance the time-steps to be small enough to yield valid approximations while being large enough to make forward progress. No single time-step size will be sufficient over the course of a growing tumor's life, because the rates of events will tend to increase as the tumor increases in size. Therefore, preferred time-step sizes should be adaptively selected based off of the total reaction rates in the tumor at a given time. The Gillespie equation naturally handles this adaptive time-step size selection. Within the Gillespie equation, the expectation of the sampled time-elapse is inversely proportional to the sum of the rates of the 7 possible events. Hence time-steps in the simulations are sized to sample times that it would actually take for the next event to occur.

In different runs of the simulation, the net balance of birth over death is chosen to follow either exponential growth –where the theory is better developed – or logistic growth – which is more realistic. When a simulated cell reproduces, its two daughter cells each acquire one new mutation, and inherit all those of their mother. I keep track of which mutations are in which cells with a tree data structure. The mutations are notional and have no specified genomic site.

One mutation per simulation is randomly chosen as a positively selected variant (the "subclonal driver"), which increases the balance of birth over death in cells that inherit that mutation to some random extent.

The simulation stops at some random point after the subclonal driver mutation has expanded to a double-digit percentage of the tumor but before it has fixated to the whole tumor, with

typically 1,000 – 100,000 variants. In idealized simulations, the true variant allele frequencies of each mutation are calculated from the tree structure. In addition, more realistic mock-observed variant allele frequencies are simulated from a binomial distribution with probability equal to the true VAF and number of trials equal to various simulated read depths. The resulting idealized and realistic VAF distributions are used as input to Dr. Salichos's software, with the identity and fitness effects of the subclonal driver withheld. The goal is for the software to correctly identify the subclonal driver from among the neutral variants, and to quantitatively estimate the fitness impact of the subclonal driver.
Simulations were run in replicate with a range of parameters. Overall, I simulated more than 4,000 tumors, with varying growth families, death rates, background mutation rates, and selection coefficients. In the main simulations, there were three styles of simulation: memoryless exponential, cell-cycle aware exponential, and memoryless logistic. In memoryless simulations, event rates depend only on the current state of the tumor, whereas in cell-cycle aware simulations, cells that have recently divided cannot divide again until they progress in the cell cycle. In exponential simulations, death rates are a constant fraction and in logistic simulations, death rates linearly approach the birth rate as the population reaches carrying capacity.

Results

Part I

Simulations indicate that the approach is well-powered

Excess recurrence was detectable either when the fraction of affected bases was moderately high or when the mutational process had a large effect on mutation rates. From these simulations, we calculated that a mutation process that affects about 1% of the genome must increase the mutation rate of the affected bases about 4-fold to increase the number of recurrent variants by 5% above expectations. Our power analysis indicates that with current sample sizes, we are theoretically powered to detect a 0.6% excess of recurrent variants; once somatic databases grow to the size of current germline databases, a 0.35% excess recurrence could be detected. In contrast, if we were to rely on publicly available de novo variants, the minimal detectable excess of cSNVs would need to be 12% above expectations, due to their smaller sample size. Both mutation-promoting processes and mutation-inhibiting processes led to excess recurrence. These simulations show that recurrence analysis is theoretically well-powered to detect the impact of a broad range of shared mutational processes that might be active in our somatic and germline data sets.

Mutation rates are sufficiently heterogeneous to result in three times as many recurrent variants than expected by chance

16,879,845 genomic sites pass all filters, implying a universe of 50,639,535 potential SNVs. Of these, 3,339,715 are observed in the germline database (gnomAD) and 1,309,369 are observed in the somatic database (TCGA). Under statistical independence, it was expected that 86,354 unique SNVs would be simultaneously present in gnomAD and TCGA; instead, we observed

268,250 concurrent variants, a 3-fold enrichment (Forbes coefficient 3.106, binomial p-value << 5e-324) in our maximally-filtered set. This overall enrichment was not sensitive to filtering strategy: with minimal filtering, this coefficient is 2.95. The calculated cSNV Forbes coefficient on the filtered set represents our first-pass estimate of the total similarity between somatic and germline mutation patterns. Our simulations indicate that there are many ways this enrichment could arise (green and orange grid elements of Figure 1); one example is if the same quarter of the genome was 16-fold more mutable than the rest of the genome in both the germline and soma. We observed that an elevated cSNV rate is a pervasive phenomenon, not confined to a few outlier somatic samples.

Nucleotide context is a major determinant of variants shared between the soma and germline

We initially observed that 67% of all cSNVs occur at N[C->T]G contexts, which are known to frequently mutate in the germline and soma. Our Forbes dependence metric estimated that 92% of the cSNV enrichment may be attributed to the high rate of N[C->T]G mutations. Our local Forbes coefficient test estimated that N[C->T]G cSNVs and non- N[C->T]G cSNVs occur 4% and 55% more frequently than expected, respectively, after conditioning on the base rates of these particular kinds of mutation in the germline and soma. Further partitioning potential SNVs into seven types of context-related variants (see Table 1 legend) explains 97.2% of cSNV enrichment, leaving about 3% of the implied heterogeneity in mutation rates unexplained. Extended nucleotide contexts added minimal explanatory value overall, but offered a moderate boost in the ability to explain cSNVs outside of N[C->T]G contexts (from 82.5% using seven types of nucleotide contexts to 88.6% by treating each of 24,576 heptamers separately).

Genomic region is a minor determinant of variants shared between the soma and germline

The effects of genomic region were minor. Similarities in the somatic and germline mutation rate by megabase explains only 0.4% of excess cSNVs. (Table 2)

Nonetheless, combining regional features with nucleotide context features explained a greater share of excess cSNVs (97.9%) than did nucleotide context alone (97.2%) (Table 3). The increased explanatory power of the combined model was not an artifact of the greater number of partitions in the combined model, because a dummy combined model, which randomized the megabase membership of SNVs, did not lead to any change in explanatory power compared to the nucleotide context-only model (97.2% in both cases).

Part II

Trace levels of negative selection in the cancer noncoding genome generally

Using the dC/dU framework on these 2500 noncoding cancer samples, I found 1.8% fewer conserved mutations than expected based off the distribution of unconserved mutations. This provides a point estimate of approximately 50 noncoding mutations per tumor removed by negative selection. However, the evidence in favor of this point estimate is not strong, because the effect size is not large enough to get us past the range of uncertainty we established in Part I about our models of mutation rate generation.

Higher signals of negative selection in the most critical regions of the noncoding genome

I next tested whether negative selection was stronger at genomic regions of higher prior functional relevance. I repeated the dC/dU analysis on the promoters of genes essential in cancer (as determined by CRISPR knockout experiments)2 in haploid samples/regions. Haploid regions were called by the PCAWG consortium. Here the point estimate for the effect size of negative selection in the noncoding cancer genome was much larger. Based on the distribution of unconserved mutations, there were 32% fewer conserved mutations than expected.

Part III

Overall, our approach was able to infer information about the identity and fitness impact of subclonal drivers in simulated tumors

During simulated growth, we assigned a "driver" mutation with additional propagating effects from nearly neutral to high (k=1.1, 2, 3, and 4), thus leading to faster growth for the respective subpopulation that contains the specific mutation. For each simulation we calculated each mutation's frequency in the total population and ordered them based on that frequency. Then, by applying our method we obtained the distance (as a number of ordered mutations) between the true driver and our predicted driver (growth peak), as well as the driver's effect k. Overall, across the different simulation models of tumor growth, we were able to approximate the driver's time point and the driver's effect (Figure 2). Similar results were obtained in alternative simulations.

Challenges & Troubleshooting

The probability of detecting a read carrying a true mutation from a sequencing sample is proportional to the fraction of cells in the tumor carrying that mutation. However, detecting a read with a mutation is not sufficient to call a mutation at that site. This is because sequencing errors are relatively common on a per-read basis in next generation sequencing technologies, which is why variant callers typically require 2 or even 3 unique reads to carry a somatic mutation before calling it as a valid mutation. Thus, the probability of calling a true variant in a tumor is approximately proportional to the square or even cube of the fraction of cells in the tumor carrying a mutation, depending on the variant caller. The implication is that mutations present at extremely low variant allele frequencies are vanishingly unlikely to be called in a sequencing sample. Therefore, the mutations from simulations of greatest practical important are those present with sufficient prevalence in the tumor, such as with a true VAF greater than 0.05. On the other hand, mutations with 100% fixation in the tumor are of little relevance for the EvoTum software, which obtains its signal from the differences in VAF between incompletely fixated mutations.

The most persistent challenge in coding the simulations was in producing enough mutations with distinct VAFs within the VAF region of interest. Real biological tumors have 100s or 1000s of mutations with a VAF between 0.05 and 0.5, which provides the sample sizes believed necessary to support inference of selection. It is very easy to simulate a tumor with a large number of mutations by simply growing the tumor to huge sizes or by letting mutations accumulate over cycles of tumor cell birth and death. However, in the former case, the

resulting VAFs will be too small to be detected in sequencing, and in the latter case, the bulk of the mutations will tend to fixate to 100% of the tumor.
Theory predicts that in a well-mixed, neutrally, exponentially growing tumor without cell death, a mutation that arises at the 2-cell stage will be present in half the tumor, a mutation that arises at the 4-cell stage will be present in a quarter of the tumor, and so on. This means that mutations that reach a VAF corresponding to 1/40th of the tumor will tend to have occurred by the 32-cell stage or early (the 64-cell stage being already too late), which limits the number of unique mutations in the desired range that can be produced by such simulations. No amount of playing with parameters seemed to get beyond this theoretical limitation. This was puzzling because biological tumors routinely carry 100s to 1000s of mutations in the desired VAF range. One way to increase the number of mutations that pushed into the desired VAF range was to introduce a subclonal driver mutation. The faster growth rate of the subclonal driver cell and its descendants buoyed up the VAF of the subclonal driver mutation and all its ancestors. The subclonal driver could be made to arise arbitrarily deep in the tumor's history so long as there was enough time provided for it to catch up in prevalence to some of the earliest mutations. If the subclonal driver had 100 ancestral mutations, and the subclonal driver itself was allowed to grow in VAF to > 0.05 without full fixation, then nearly 100 additional mutations would be added to the desired VAF range. However, even then, VAF collapse would tend to occur, where the gradual die-off of non-driver descendants of the subclonal driver's ancestors collapsed many of the sublconal driver's ancestor mutations to have identical VAFs. Because EvoTum relies on VAF differences between hitchhiker mutations, VAF collapse in the simulations would be predicted to frustrate EvoTum. VAF collapse is not observed in biological tumors, even in very deeply sequenced tumors. To prevent VAF collapse in the simulations, I introduced a new kind of cell, the "quiescent" cell, which endures and contributes DNA to sequencing but ceases to reproduce and is no longer at risk of death.

Discussion

Part I

These findings indicate that the known determinants of mutation rates explain a much larger portion of mutation rates that do undiscovered determinants. This is good news for dN/dS because it argues in favor of our ability to adjust the mutation rates at nonsynonymous and synonymous sites to make them comparable.
Nonetheless, our ability to explain mutation rate heterogeneity was found to be incomplete. There is still 3% of the total excess recurrence that cannot be explained with nucleotide context or genomic region. This matters because it means that when we identify selected genes or variants, we need a signal stronger than the theoretical minimum. Otherwise, we risk calling genes as subject to selection when really we have just failed to correctly model their neutral mutation rates. My results do not allow us to assign a number as to how much stronger signal must be than the theoretical minimum because this study was looking at global effects across the genome, but neutral departures from current expectations might pile up at particular genes. Until a more rigorous analysis of these local effects is conducted, a reasonable starting point might be to be suspicious of assertions about selection when the number of observed

mutations does not depart from expectations by more than 3%, regardless of the sample size and p-value.

There are a number of important caveats and limitations to this study. One limitation is that it only considers determinants of mutation rates that consistently act at the same bases between the soma and germline. It could very well be that we are better at explaining the portion mutation rate heterogeneity across bases that is conserved across disparate settings. This aspect of the study design could lead to underestimation of the uncertainty in mutation rates. On the other hand, the study did not exhaustively back out all known determinants of mutation rates -only nucleotide context and genomic region-, which tends to lead to overestimation of the uncertainty in mutation rates in a theoretical sense. (In another sense, since best-in-class models include nucleotide context and genomic region but not many other determinants of mutation rates, this "limitation" of the study makes the warnings of the previous paragraph more practical).

Here, the partitional dependence of the Forbes coefficient was used to separate explained and unexplained portions of mutation rate variation. The statistical framework itself is more general, and can apply to any setting in which the elements of two binary vectors of equal length have concordant levels of a categorical variable.

For example, consider applying this framework to the study of clinical co-morbidity between two diseases, say, major depressive disorder and reduced ejection fraction heart failure. In this case, one binary vector would represent whether a cohort of patients had a diagnosis of major depressive disorder, and the other binary vector would represent whether the same cohort of patients had a diagnosis of reduced ejection fraction heart failure. The classic Forbes coefficient applied to these vectors would give us a measure of the total comorbidity between these diseases. We could then use zip code of residence to partition vector elements to calculate the partitional dependence of the Forbes coefficient. This statistic would tell us the portion of co-morbidity between depression and heart failure in this cohort that could not be explained by zip code.

I do not want to give the impression that we have "solved" the question of what affects mutation generation rates, for these findings leave much space for discovery. For example, even determinants of mutation rates that are counted as known do not have well-understood mechanisms for influencing mutation rates. Nonetheless, the small uncertainty estimates in this study hint that the field of the basic science of genomics is maturing, and the time is ripe for working to translate this basic science into clinical practice.

Part II

This analysis was performed as one part of a larger paper about functional and fitness effects of mutations in noncoding cancer genomes. A separate paper could be written that benchmarks and extends this technique in particular. A comparative analysis with Agarwalla et al.35, could be used as a benchmark. Results could be reported for individual gene regulatory elements and/or pathways of gene regulatory elements. The analysis could be repeated using a different GERP threshold as a cutoff or using a different measure of functional impact, such as the CADD score, to see how sensitive results are to modeling decisions. Moreover, the technique could be applied to germline noncoding data.

An important limitation of this study is that GERP (and other measures) of prior functional impact has a special kind of bias that affects the interpretability of results. The calculation of the GERP scores themselves depends on comparing the observed vs expected mutation rate among mammals at each site. The problem with this is that a misspecification in the expected mutation rate at the phase of GERP score calculation is likely to occur in just the same genomic sites as a misspecification of expected mutation rates at the phase of mutation recurrence across human samples. The use of a distinction between nonsynonymous and synonymous sites in classic dN/dS has the advantage of not ever relying on the estimation of neutral mutation rates to decide which sites are synonymous and which sites are nonsynonymous.

An unmet need in the analysis of noncoding data is to find a prior measure of functional impact that does not depend on or correlate with neutral mutation rates. While it might be tempting to use the degree of disturbance of transcription factor binding sites in noncoding regions as a prior for fitness effects, even this criterion is problematic because transcription factor binding has been shown to interfere with DNA repair enzyme access and therefore neutral mutation rates. My best candidate going forward for a noncoding prior measure of functional impact without systematic relation to neutral mutation rates is the degree of disturbance of RNA binding protein sites, because this binding occurs on RNA molecules away from the chromosomes and should therefore not interfere with DNA repair.

Why is the signal of negative selection so weak in cancer compared with the germline?

One possibility for the weaker signal of selection in cancer vs germline is if fewer mutations in cancer than in the germline have a fitness impact. After all more genes (and their noncoding regulatory elements) are required for organismal survival than for cellular survival Cells support an organism, and the organism supports its cells, but the relationship is asymmetrical: cells are more fundamental. This asymmetry is illustrated with the following point: The existence of cancer cell cultures indicates that, in the right conditions, human stem cells can survive and reproduce indefinitely outside the body; in contrast, a human body without cells is a skeleton and puddle of fluid.

A mutation in the germline affects cells of all lineages, so if the affected gene is important in any lineage, germline mutations that hurt the activity of the gene will be harmful to the organism. In contrast, somatic mutations (of differentiated cells) only affect the cell lineage in which they arise. Therefore, if the somatic mutation affects a gene that is not expressed in the lineage anyway, then that somatic mutation is of no fitness relevance.

Similarly, some genes are necessary for organismal survival during development and are therefore deleterious when mutated in the germline, but are no longer necessary in an adult and can be neutral when mutated in the soma.

Moreover, in a healthy body, cells specialize and cooperate to serve each other's needs. All human cells need oxygen, but only some human cells need to directly contribute to ventilation and perfusion for all to benefit. Cancer cells in a tumor can stop its body-supporting functions and survive so long as other cells can pick up the slack. Effectively, cooperation between cells in the body is a Prisoner's Dilemma, in which cancer cells are not harmed and may even be helped by defecting from cooperating with other cells. (Eventually, when tumor burden becomes too high, this defection strategy catches up and the organism perishes along with the cancer cells.)

In addition, perhaps negative selection is less efficient in cancer than in the germline. One important consideration for the relative efficiency of selection in cancer vs the germline is the

corresponding amounts of evolutionary time. Homo sapiens have existed for about 15,000 organismal generations, which is approximately 300,000 cellular generations when considering that about 20 cell divisions separate a new zygote from a typical gamete descended from that zygote. In contrast, a single tumor survives for only 100s to low 1000s of cellular generations. This means that selection has fewer opportunities to act on tumors than on the human germline.

Moreover, selection is more efficient in the germline due to greater likelihood of heterozygous variants becoming homozygous. Suppose there were a mutation that would be deleterious if homozygous, but neutral if heterozygous. This kind of recessive variant, when heterozygous in the soma, will not impose an appreciable fitness cost or be negatively selected so long as the affected region remains diploid. In contrast, a heterozygous truly recessive germline mutation, when common in the population or in situations of inbreeding, increases the risk of the organism's descendants of becoming homozygous for the deleterious allele. This process would tend to decrease the prevalence of the deleterious allele in the gene pool, such that a signal of negative selection could be observed in population sequencing.

Furthermore, perhaps some signals of negative selection in the germline are spurious. For example, consider the fact that linkage disequilibrium is unique to the germline. In the germlines of different human subjects, the same variants will be inherited together due to incomplete crossing over during meiosis, which leads to the co-segregation of neighboring alleles. When neutral alleles co-segregate with fitness-affecting alleles, it can cause the fitness-neutral allele to be subject to negative selection in the germline. In contrast, in the somatic genomes of different human subjects, somatic mutations separately arose in each patient, so there is no clear pattern of particular mutations being inherited in blocks in the soma.

Similarly, the risk of misinterpreting an environmental signal as genetic is greatest in the germline. Genome wide association studies and their successor, the polygenic risk score, have linked genotypes to a wide range of phenotypes. One criticism of these studies is that any environmental traits are correlated with ancestry or family structure can be misread as a genetic signal in these observational genotype-to-phenotype association studies. Perhaps some portion of the discrepancy between germline and somatic estimates of the degree of negative selection in the human genome comes from germline overestimation of the fitness importance of genetic variants.

Part III

These simulations lend themselves to a range of possible extensions in the future. Instead of enforcing a single mutation to occur at each cell division, a Poisson model could be used to sample 0 to several mutations arising at a single cell division. Instead of treating the tumor as a well-mixed population, the spatial structure could be modeled by placing the cells along a grid, and having dividing cells form adjacent to their parents, displacing or replacing pre-existing neighbor cells. Instead of testing the effects of VAF noise only, we could also introduce sequencing errors. Instead of enforcing logistic growth dynamics, we could also allow for Gompertzian growth, which is more generalized than logistic growth and has been shown to more closely mimic biological tumors.

A particularly interesting alternative approach was suggested by Dr. Warrell: integration of single cell and bulk sequencing data from a single tumor. As currently designed, the EvoTum

framework is best used for assessing dominant tumor subclones, whose prevalence exceeds that of all other subclones. This is because only when the subclone of interest is dominant can we ascertain that a mutation belongs to the subclone of interest, and not another subclone with similar prevalence. However, the use of single-cell sequencing would allow us to more confidently assign subclonal memberships to mutations. It would be interesting to see how this sort of information improves the accuracy of EvoTum, and the benchmarking of this approach would benefit from specialized simulations.

These simulations could be easily extended to application settings other than cancer. We could analyze mutations on the Y chromosome or mitochondrial chromosome throughout the global human population, using population allele frequencies instead of VAFs. Fitness effects of mutations in asexual yeast and bacteria could in principle be uncovered through this approach.

The EvoTum approach will be most powerful when ultra-deep sequencing is performed, as this will give us the most precise VAF information. Moreover, ultra-deep sequencing is well within our technological capabilities today but its extra cost has been considered unnecessary. EvoTum gives the community a reason to pay this extra cost in select samples.

Could simulations win a chess match against cancer?

The main bottleneck in curing patients of cancer is the emergence of resistant subclones. The probability and timing of the emergence of resistance depends on patient-specific evolutionary trajectory of their tumor. The Patient-derived xenograft (PDX) model for cancer is to have a particular patient's tumor grow out in mice in order to predict the evolutionary sequence and empirical response to therapy of the patient's tumor. The PDX approach specifically was met with mixed success, in part due to the fact that the cancer's evolutionary trajectory changes when it is transferred from human to mouse. I propose instead a Patient Derived Simulation (PDS) – a simulated tumor fitted to the mutations and growth parameters of a patient's tumor – to serve the same function, but avoid interspecies transitions and spare the mouse. While much more fundamental work must be done prior to making simulations true to life, this would be an ultimate endpoint in the development of tumor simulations.

References

[i] Grant PR, Grant BR. Unpredictable evolution in a 30-year study of Darwin's finches. Science. 2002;296(5568):707-11.

[ii] Hart T, Chandrashekhar M, Aregger M, et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015;163(6):1515-26.

[iii] Lanphier E, Urnov F, Haecker SE, Werner M, Smolenski J (2015) Don't edit the human germ line. Nature 519:410–411

[iv] Fu Y, Foden JA, Khayter C, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013;31(9):822-6.

[v] Mathieson I, Lazaridis I, Rohland N, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528(7583):499-503.

[vi] Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275–276.

[vii] Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.

[viii] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586-91.

[ix] Alexandrov L.B., Nik-Zainal S., Wedge D.C., Aparicio S.A., Behjati S., Biankin A.V., Bignell G.R., Bolli N., Borg A., Børresen-Dale A.L., Australian Pancreatic Cancer Genome Initiative. ICGC Breast Cancer Consortium. ICGC MMML-Seq Consortium. ICGC PedBrain Signatures of mutational processes in human cancer. Nature. 2013;500:415–421.

[x] Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980-5.

[xi] Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 2006;7(2):98-108.

[xii] Martincorena I, Raine KM, Gerstung M, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171(5):1029-1041.e21

[xiii] Shao ZQ, Zhang YM, Feng XY, Wang B, Chen JQ. Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency. PLoS ONE. 2012;7(3):e33547.

[xiv] Costello M, Pugh TJ, Fennell TJ, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013;41(6):e67.

[xv] Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, Ribeca P. Fast computation and applications of genome mappability. PLoS One. 2012;7(1):e30377.

[xvi] Bailey M, **Meyerson W**, *et al*. Comparison of exome variant calls in 746 cancer samples with joint whole exome and whole genome sequencing. *In submission.*

[xvii] Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875.

[xviii] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586-91.

[xix] Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39(2):417-20.

[xx] Sabarinathan R, Mularoni L, Deu-pons J, Gonzalez-perez A, López-bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature. 2016;532(7598):264-7.

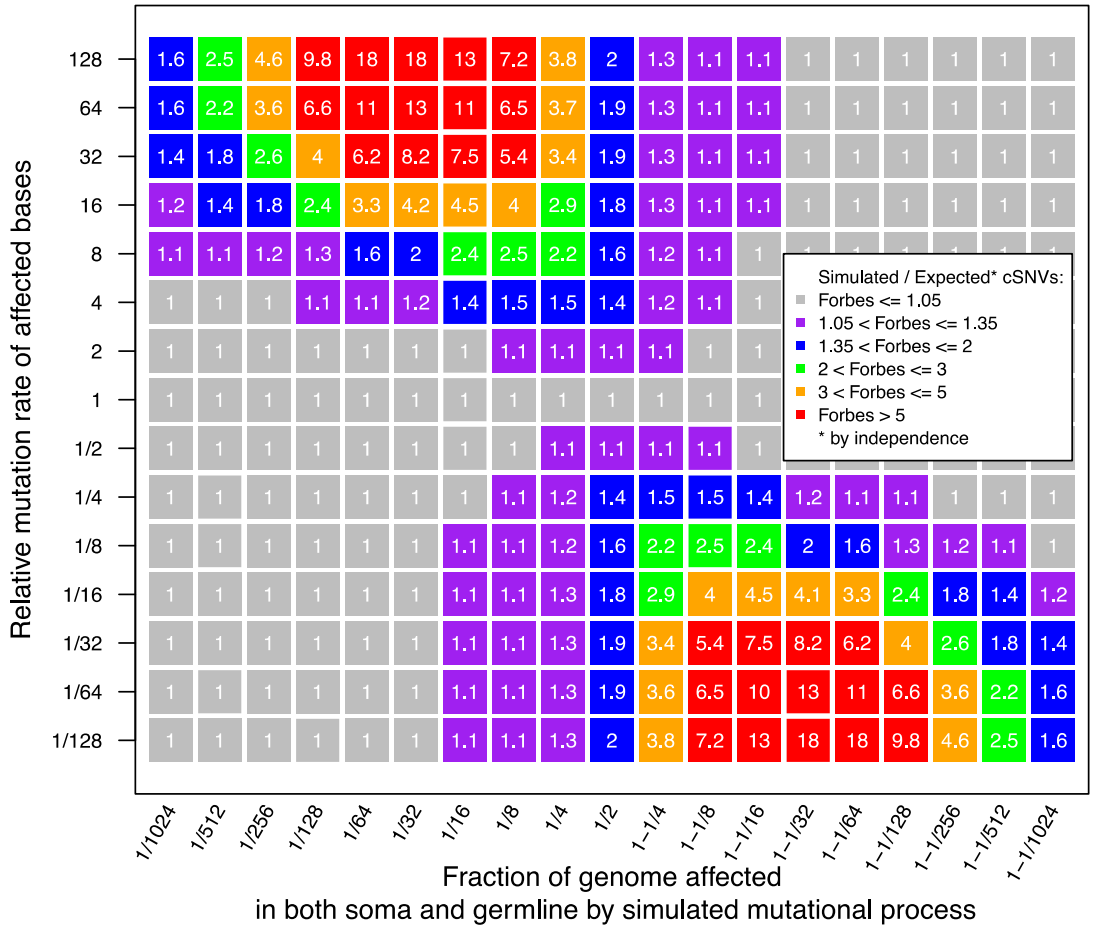
[xxi] Pich O, Muiños F, Sabarinathan R, Reyes-salazar I, Gonzalez-perez A, Lopez-bigas N. Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. Cell. 2018;175(4):1074-1087.e18.

[xxii] Ibarra-laclette E, Lyons E, Hernández-guzmán G, et al. Architecture and evolution of a minute plant genome. Nature. 2013;498(7452):94-8.

[xxiii] Metcalfe CJ, Filée J, Germon I, Joss J, Casane D. Evolution of the Australian lungfish (Neoceratodus forsteri) genome: a major role for CR1 and L2 LINE elements. Mol Biol Evol. 2012;29(11):3529-39.

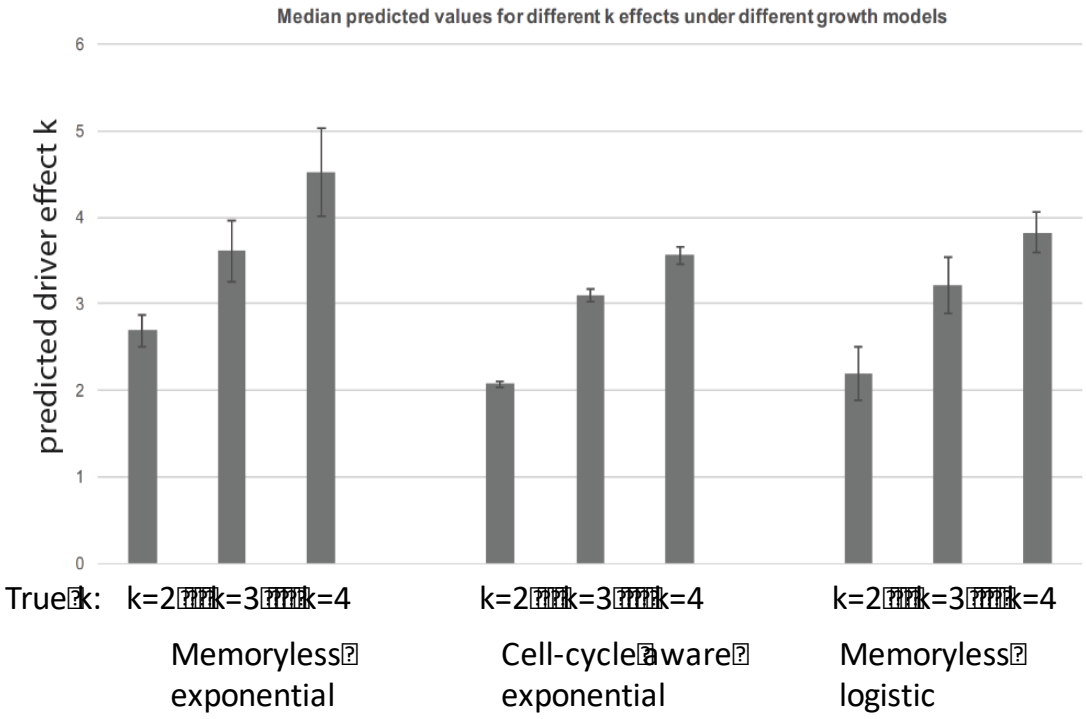
Figures

Figure 1: Simulated mutational processes generate a detectable excess of recurrent variants.



Unknown, hypothetical mutational processes were simulated to act in a coordinated manner in the soma and germline. Each simulated mutational process equally affects the same effectively random subset of genomic sites in the soma and germline. This leads to an excess of sites that are concurrently mutated in the soma and germline, over expectations by independence (the Forbes coefficient).[xxiii] *The magnitude of resulting Forbes coefficients (numbered grid squares) depends on the fraction of genomic sites subject to the mutational process (x-axis) and the mutation rate multiplier (y-axis) of the affected bases relative to unaffected bases. Symmetry arises because a mutation-promoting process affecting 25% of the genome is equivalent to a mutation-inhibiting process affecting 75% of the genome.*

Figure 2. Fitness parameters of simulations are inferable from different model types.



Tables

Table 1: Excess recurrence is statistically explainable by nucleotide context

Nucleotide context	Number of partitions	Partition-conditioned score	Partition dependence
Unpartitioned	1	3.106	NA
N[C->T]G vs all others	2	1.168	92.0%
Seven Type	7	1.059	97.2%

Trinucleotide	96	1.064	97.0%
Pentanucleotide	1536	1.054	97.4%
Heptanucleotide	24,576	1.051	97.6%

Seven Type refers to the 7 basic types of variant from Arndt *et al*. (C->A, C->G, C->T, N[C->T]G , T->A, T->C, and T->G after collapsing purine-centered contexts onto their central pyrimidine).[xxiii] The partition-conditioned score gives the excess recurrence over a form of expectations that incorporate the fact that different nucleotide contexts have different mutation rates. The partition dependence gives the percent of excess recurrent variants that can be statistically explained by the different mutation rates of the various partitions.

Table 2: Genomic region explains a small fraction of excess cSNVs.

Region	Number of partitions	Partition-conditioned score	Partition dependence
Whole Genome	1	3.106	NA
Chromosome	22	3.107	0.0%
Megabase	2394	3.097	0.4%
100kb	14,214	3.084	1.0%
10kb	54,889	3.076	1.4%

Table 3: A combined model with region and nucleotide context explains excess cSNVs slightly better than nucleotide context-only model

Bin	Number of partitions	Partition-conditioned score	Partition dependence
Whole Genome	1	3.106	NA
Seven Type	7	1.059	97.2%
Megabase	2394	3.097	0.4%
100kb	14,214	3.084	1.0%
1MB x Seven Type	16,737	1.054	97.4%
100kb x Seven Type	98,960	1.044	97.9%
Dummy.100kb x Seven Type	99,250	1.058	97.2%