Mutation-biased adaptation

MUTATION-BIASED ADAPTATION

Alejandro V Cano

Diss. No. 28508

DISS. ETH NO. 28508

MUTATION-BIASED ADAPTATION

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

ALEJANDRO V. CANO
MSc, Universidad de Los Andes, Mérida

born on 20.12.1991
citizen of Venezuela

accepted on the recommendation of

Prof. Dr. Joshua L. Payne, ETH Zurich, examiner
Prof. Dr. Alex R. Hall, ETH Zurich, co-examiner
Prof. Dr. Ard A. Louis, Oxford University, co-examiner

2022

DOI: 10.3929/ethz-a-28508

...the gold puts the focus on the cracks, on the failures, which are treated as unique and endearing aspects of the object that tell a story...
— Arlin Stoltzfus

SUMMARY

Mutations are usually thought of as random due to their inherent unpredictability. However, such randomness is not uniform because some molecular changes in DNA sequences are more likely to occur than others. Here we refer to adaptation as *mutation-biased* when the mutations that are more likely to occur are also more likely to contribute to adaptation. In this thesis, we have integrated empirical data with computational and theoretical approaches to study the conditions for, and consequences of, mutation-biased adaptation.

Early theoretical work used synthetic genotype-phenotype landscapes (i.e., a mapping from genotypes to a quantitative phenotype) to study the conditions for mutation-biased adaptation. In Chapter 2, we characterised 746 empirical genotype-phenotype landscapes of transcription factor binding affinities to study the influence of mutation bias on adaptive evolution of increased binding affinity. We found that such empirical landscapes exhibit composition bias, namely the enrichment of a particular type of mutation in the adaptive mutational trajectories of the landscape. We aggregated mutations into two classes: transitions (purine-purine or pyrimidine-pyrimidine changes) and transversions (purine-pyrimidine or vice-versa), and we quantified composition bias relative to the null expectation that there is one transition possible per every two transversions. Our results uncover composition bias among the accessible mutational trajectories towards adaptive peaks, and show that such composition bias can interact with mutation bias to influence the rate of adaptation, the evolution of genetic diversity and mutational robustness, as well as the predictability of evolution.

To what extent does mutation bias influence the process of adaptation? To address this question, in Chapter 3, we developed a statistical framework to quantify the influence of mutation bias on the spectrum of missense mutations underlying adaptive evolution. More specifically, we used negative binomial regression to model the observed frequencies of adaptive codon-to-amino acid substitutions by incorporating empirical estimates of codon frequencies and neutral per-nucleotide mutation rates. We separately applied this approach to three large data sets of adaptive changes in *Saccharomyces cerevisiae*, *Escherichia coli*, and *Mycobacterium tuberculosis*. In all three cases, we found that mutation bias has a proportional influence on the spectra of adaptive substitutions. Moreover, evolutionary simulations revealed that the influence of mutation bias on adaptive evolution is modulated by different factors, such as mutation supply and the breadth and heterogeneity of the mutational target for adaptation.

A central tenet of evolutionary theory is that mutation rates are uncorrelated with fitness effects. That is, a mutation with a large fitness effect is expected to arise at the same rate as a mutation with a small fitness effect. Technological advances, such as deep-sequencing and mutation-accumulation approaches, are now making it possible to characterise the actual associations between these mutation rates and fitness effects. However, such associations may differ depending on whether they are characterised before or after adaptation takes place. This is because adaptation is jointly conditioned on what is mutationally likely and on what is selectively favored, and such joint conditioning can induce non-causal associations between the causal variables (e.g., mutation rates and selection coefficients), a phenomenon known as Berkson's paradox. In Chapter 4, we studied the possible non-causal associations between mutation rates and selection coefficients that adaptation can induce combining theoretical and computational approaches, as well as analyses of large data sets of adaptive changes. Our results showed that such non-causal associations can emerge under a variety of evolutionary conditions and in different biological systems, including viruses and cancer.

In Chapter 5, we highlight the major conclusions of this thesis, and provide potential directions for future research.

RÉSUMÉ

Les mutations sont généralement considérées comme aléatoires en raison de leur inhérente imprévisibilité. Ce caractère aléatoire n'est cependant pas uniforme car certains changements moléculaires dans les séquences d'ADN sont plus susceptibles de se produire que d'autres. Nous qualifions ici l'adaptation de biaisée par la mutation lorsque les mutations qui ont le plus de chances de se produire sont également plus probables de contribuer à l'adaptation. Dans cette thèse, nous avons intégré des données empiriques à des approches computationnelles et théoriques afin d'étudier les conditions et les conséquences de l'adaptation par mutation.

Les travaux théoriques précédents utilisaient des paysages génotype-phénotype synthétiques (c'est-à-dire une correspondance entre les génotypes et un phénotype quantitatif) pour étudier les conditions de l'adaptation biaisée par la mutation. Dans le Chapitre 2, nous avons caractérisé 746 paysages génotype-phénotype empiriques d'affinités de liaison de facteurs de transcription pour étudier l'influence du biais de mutation sur l'évolution adaptative d'une affinité de liaison plus forte. Nous avons constaté que ces paysages empiriques présentent un biais de composition, à savoir l'enrichissement d'un type particulier de mutation dans les trajectoires mutationnelles adaptatives du paysage. Pour cela, nous avons regroupé les mutations en transitions (changements purine-purine ou pyrimidine-pyrimidine) et en transversions (purine-pyrimidine ou vice-versa), et nous avons quantifié le biais de composition par rapport à l'hypothèse nulle selon laquelle il se produit une transition pour deux transversions. Nous avons trouvé un biais de composition parmi les trajectoires mutationnelles accessibles vers des pics adaptatifs, et nous avons montré que ce biais de composition peut interagir avec le biais de mutation pour influencer le taux d'adaptation, l'évolution de la diversité génétique et la robustesse des mutations, ainsi que la prédictibilité de l'évolution.

Dans quelle mesure les biais mutationnels peuvent-ils influencer le processus d'adaptation ? Pour répondre à cette question, nous avons développé, dans le Chapitre 3, un cadre statistique pour quantifier l'influence du biais de mutation sur le spectre des mutations faux-sens qui sous-tendent l'adaptation. Plus précisément, nous avons utilisé une régression binomiale négative pour modéliser les fréquences observées des substitutions adaptatives de codons en acides aminés en incorporant des estimations empiriques des fréquences de codons et des taux de mutation neutres par nucléotide. Nous avons appliqué cette approche séparément à trois grands jeux de données de changements adaptatifs chez *Saccharomyces cerevisiae*, *Escherichia coli* et *Mycobacterium tuberculosis*. Dans chacun des cas, nous avons constaté que le biais de mutation a une influence proportionnelle sur les spectres de substitutions adaptatives. Cependant, les simulations évolutives ont montré que cette influence peut être modulée par différents facteurs, tels que le nombre de mutations par génération et l'ampleur et l'hétérogénéité de la cible mutationnelle pour l'adaptation.

L'un des principes centraux de la théorie de l'évolution est que les taux de mutation ne sont pas corrélés aux effets de la fitness. En d'autres termes, on s'attend à ce qu'une mutation ayant un effet important sur la fitness se produise au même rythme qu'une mutation avec un faible effet sur la fitness. Les progrès technologiques, tels que les approches de séquençage profond et d'accumulation de mutations, permettent aujourd'hui de caractériser les associations réelles entre ces taux de mutation et les effets de la fitness. Toutefois, ces associations peuvent être différentes selon qu'elles sont caractérisées avant ou après l'adaptation. En effet, l'adaptation est conditionnée conjointement par ce qui est susceptible de muter et ce qui est sélectivement favorisé, et ce conditionnement conjoint peut induire des associations non causales entre les variables causales (c'est-à-dire les taux de mutation et les coefficients de sélection), un phénomène connu sous le nom de paradoxe de Berkson. Dans le Chapitre 4, nous avons étudié les associations non causales possibles entre les taux de mutation et les coefficients de sélection que l'adaptation peut induire, en combinant des approches théoriques et computationnelles ainsi que des analyses

de jeux de données sur les changements adaptatifs. Nos résultats ont montré que de telles associations non causales peuvent émerger dans de nombreuses conditions évolutives et dans différents systèmes biologiques, y compris les virus et le cancer.

Dans le Chapitre 5, nous soulignons les principales conclusions de cette thèse et proposons de potentielles directions pour les recherches futures.

CONTENTS

1

INTRODUCTION

1.1 MUTATION BIAS

Natural selection acts on phenotypic variation to cause adaptive evolutionary change. Such phenotypic variation is often caused by mutational processes that generate heritable genetic variation. Mutation is therefore integral to adaptation, and understanding how mutation causes adaptive phenotypic variation is central to understanding evolution.

There are different types of genetic mutations, which can be characterised by their underlying molecular change. Single point mutations, for instance, consist of the exchange of one nucleotide for another, but also more than one nucleotide can be subject to exchanges (i.e., so-called multinucleotide mutations), and insertions and deletions of genetic information of variable length can occur as well. There are further levels of aggregation within mutation types, for example in the case of single point mutations, it is common to discriminate between transitions (purine-to-purine or pyrimidine-to-pyrimidine changes) and transversions (purine-to-pyrimidine or pyrimidine-to-purine changes) (Fig. 1.1a). Moreover, mutation types can also be characterised by their molecular change in a given genetic context, like an C>T point mutation given the surrounding DNA sequence (e.g., C>T in the context TCC in Fig. 1.1b). A large body of empirical work has explored the distribution of mutational types in diverse organisms, uncovering pervasive heterogeneities in the rate of occurrence of the different mutation types [1]–[10].

Mutation bias is the general term that captures a wide range of such heterogeneities in the rates at which genetic changes occur. Mutation bias may refer to asymmetries in mutation rates for different types of mutations or to heterogeneity in the rates of genetic changes at different specific genomic regions [11]–[15]. Such biases have been widely reported in studies that experimentally characterised mutation patterns in neutral evolution [8]–[10], [16]. For example, *Mycobacterium smegmatis*, a non-pathogenic bacterium often used to experimentally study other Mycobacteria species, exhibits an unusual A:T>C:G mutation bias in mutation accumulation experiments [10]. However, such biases are not only observed among neutral mutations but have further been found in mutation patterns linked to adaptation in a variety of biological systems [17]–[26]. For instance, in a dataset of adaptive mutations that increase hemoglobin affinity to oxygen in high-altitude birds, an enrichment was found for amino acid replacements associated to mutations in CpG dinucleotide regions [22]. These genomic regions are of particular interest for mutation bias research because they are hotspots of nucleotide point mutations due to the effect of cytosine methylation on DNA damage and repair, and exhibit mutation rates several-fold higher than other regions [27], [28].

Mutation-biased adaptation occurs when the mutations that are more likely to happen are also more likely to contribute to adaptation. A further challenge in identifying mutation-biased adaptation from empirical data is to determine whether an enrichment for some type of mutation relates to its higher mutation rate or rather to an inherent selective advantage for that particular mutation type. For example, transition mutations are considered to be less disruptive than

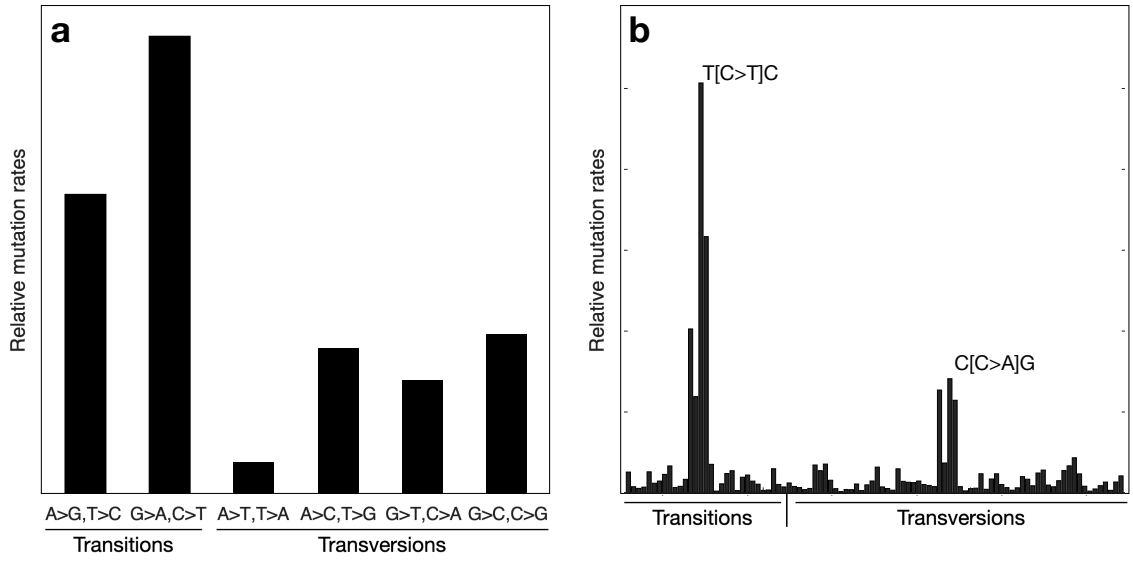

FIGURE 1.1: **Schematic representation of two mutation spectra. a**. Bar plot of a mutation spectrum describing the relative rates of the six possible nucleotide changes. **b**. Bar plot of a mutation spectrum describing the relative rates of the six possible nucleotide changes specified by the identity of the bases that immediately flank the mutated base. In this example, C>T mutations in the the context of TCC are the type of transition that occurs most often, whereas C>A mutations in the context of CCG are the type of transversion that occurs most often.

transversions for two reasons. The first one is that the standard genetic code is more robust to transition mutations, because these mutations are more likely to preserve the biochemical properties of amino acids [29], [30]. The second reason is that, because purines and pyrimidines differ in size, transition mutations are less likely to cause structural changes to the DNA double helix [31]. Nevertheless, evidence for significant selective differences across different mutation types remains inconclusive [32]–[34].

1.2 EMPIRICAL CHARACTERISATION OF MUTATION BIAS

To comprehensively characterise the spontaneous rate of newly arisen mutations has proven challenging for a variety of reasons. Perhaps the most evident one is the fact that mutations often lead to changes in phenotype that have fitness consequences ([35], [36]) that could potentially bias the estimates [15]. Other reasons include the rather low magnitudes of mutation rates (especially in eukaryotes [37], [38]), the stochastic behavior of mutation processes, and its dependence both on genetic background and on environmental and physiological conditions [39]–[42]. Initial efforts to estimate mutation rates relied upon strategies such as the study of polymorphisms in natural populations, or parent-offspring genotype comparisons —whenever the mutation rates were sufficiently high to count several mutations in a single generation (for an extensive review see [43]). For instance, a study of single-nucleotide polymorphisms within a variety of bacterial species showed that mutation is AT-biased in every case they analysed, with C/G to T/A transitions being the most frequent mutation [4]. Studies like this one, however, rely on several assumptions (e.g. constant population sizes, the absence of epistasis or uniform selective pressure across sites). More recently, the combination of mutation accumulation (MA) experiments with

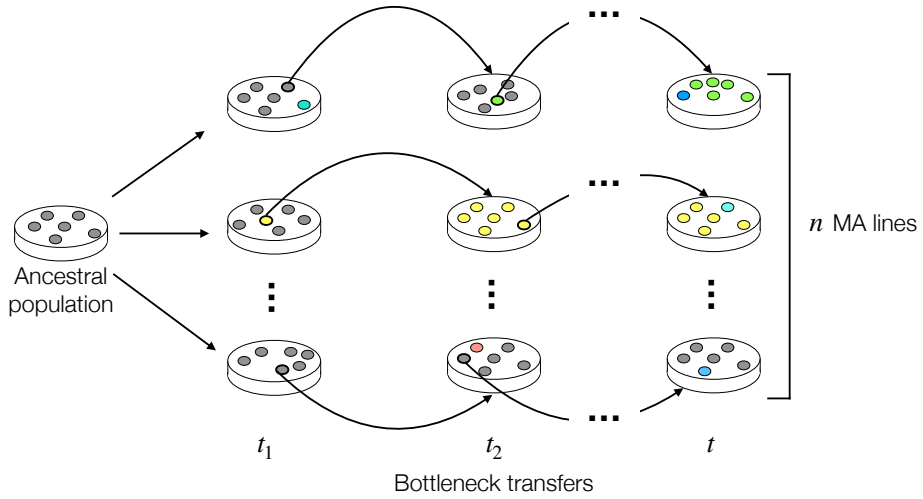

FIGURE 1.2: **Schematic representation of a mutation accumulation experiment.** In each of the n MA replicate lines, a clonal ancestral population is plated using a single colony. Each colony results from growth and replication of a single cell. At each of the t bottleneck transfers, one randomly chosen colony is propagated (indicated by the arrows). If the founding cell in this colony had a mutation (indicated by alternative colors), all of its progeny colonies will also contain that mutation. Adapted from [48].

high-throughput sequencing methods has become a widely employed procedure to empirically estimate mutation rates that overcome such challenges [44].

In such MA experiments (Fig. 1.2), new mutations are allowed to accumulate for several generations in independent replicate evolving lines. These lines are derived from an inbred ancestral population under recurrent extreme bottlenecks each generation. Such bottlenecks enhance evolutionary divergence thanks to the accumulation of mutations by genetic drift (with the exception of the lethal mutations), so that the strength of selection is drastically reduced. Hence, MA studies provide a robust approach to empirically estimate the occurrence of mutations in virtually neutral conditions. The further integration of high-throughput sequencing methods allow for the precise characterisation of the spontaneous mutation rates for different mutation classes across the whole genome, uncovering mutation biases in a variety of contexts [5], [6], [9], [10], [45]–[47].

1.3 THE ROLE OF MUTATION BIAS IN ADAPTIVE EVOLUTION

How do such mutational biases influence the process of adaptation? Historically, adaptation was principally depicted as a process consisting of the redistribution of abundant pre-existing variation [49]–[52]. In such a depiction, novel variation, for example in the form of new genetic mutations, was not absolutely necessary for adaptation because selection could act upon the abundant variation already present in the population. Thus, selection was considered the sole creative agent of adaptive evolutionary change, and therefore its driving force. There were two main arguments against mutation's potential to drive adaptive evolutionary change: First, mutation rates are small in comparison to selection coefficients. Thus, for mutation to influence the course of adaptation in this view, the mutation rates should be unrealistically high [53], [54]. Second, mutation, like genetic drift, was considered a mere diffusive factor due to its stochastic

nature. Mutations were therefore considered undirected relative to the direction of adaptive evolution [55].

An alternative depiction of adaptation emerged with the development of molecular biology, where evolutionary change was considered as a sequence of different amino acid substitutions [56]–[58], each determined by a particular mutation. In such models of sequential fixation events, or origin-fixation models, evolutionary change consists of events in which a variant is introduced into a population by mutation, and then is accepted or rejected based on its individual fitness effects [59]–[61]. This alternative view is by no means neglecting the importance of selection, since after all, selection preserves its discriminatory nature, and beneficial alleles are still the ones contributing to adaptation. However, directionality is no longer an attribute restricted to selection, because out of the set of possible favorable mutations, evolving populations will follow the adaptive trajectories directed by newly arisen molecular changes. Thus, heterogeneities in the rates of emergence of these different beneficial variants are potentially capable of directing adaptive outcomes.

1.4 GENOTYPE-PHENOTYPE LANDSCAPES

Understanding how mutation, which acts at the level of genotypes, can cause phenotypic variation, is a fundamental task in biology research, with relevant implications for the understanding of development, disease and adaptive processes [62]–[64]. A genotype-phenotype landscape is an abstract mapping that portrays the associations between genotype and a quantitative phenotype. In such a representation, genotype sequences are discrete points in genotype space, and they are connected by the underlying mutation that changes one genotype into another. The quantitative phenotype is then assigned to each genotype, for example, it could describe the binding energy of a DNA sequence to a specific protein. Most of our knowledge about such landscapes comes from computational models that predict the mapping of genotypes onto phenotypes in a variety of biological systems [65]–[67]. Understanding the structure of genotype-phenotype landscapes has important evolutionary consequences [68]–[71], for example, for the evolution of genetic diversity [68].

A fitness landscape is a particular type of genotype-phenotype landscape in which the phenotype associated to each genotype is its fitness (i.e., a quantification of reproduction and/or survival rate) [72], [73]. Adaptive evolution in such landscapes can be seen as a hill-climbing process that leads evolving populations towards adaptive peaks. Here, each adaptive step is characterised by the necessary mutation that connects the two genotypes, and the corresponding change in "elevation" that reflects the fitness change. The topographical properties of such landscapes have important evolutionary consequences [61], [74], [75], more specifically for the evolution of sex [76], [77], speciation [78] and the predictability of evolution [79], [80]. For example, let us consider the simple case of two genetic changes that are separately deleterious, but jointly they confer a selective advantage. Such a non-additive relationship in the fitness effects of these mutations, known as reciprocal sign epistasis [81], implies a local valley that would impede the population from reaching that genotype with higher fitness, trapping the population in a sub-optimal peak. The increase of epistatic interactions in the landscapes, and the concomitant increase in local peaks, will diminish the likelihood that the population will eventually reach the optimal peak [61],

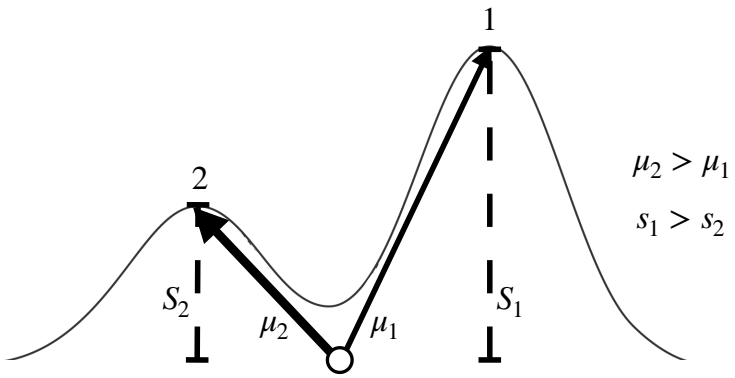

FIGURE 1.3: **Schematic representation of a simple two-locus model with two adaptive peaks.** In this model, peak 1 has a higher selection coefficient than peak 2 ($S_1 > S_2$), but peak 2 has a higher mutation rate ($\mu_2 > \mu_1$).

[75], [82], [83]. Overall, our understanding of adaptive evolution can greatly benefit from the implementation of genotype-phenotype landscape models.

1.5 MUTATION-BIASED ADAPTATION ON SYNTHETIC LANDSCAPES

What evolutionary conditions allow for mutation-biased adaptation? Early theoretical work used simple synthetic landscapes to investigate this question [84], [85]. For example, in a simple two-locus model consisting of two adaptive peaks with different selection coefficients and mutation rates (Fig. 1.3), mutation bias can increase the probability with which an evolving population converges on the sub-optimal, but mutationally-favored peak [84]. The extent to which it does, however, depends upon population genetic conditions. If the dynamics of the adaptive process depend on events that introduce novel variants, mutation bias will then influence which type of genetic change goes to fixation. The reason is that, when the mutation supply is low, mutations are rarely introduced in the population, thus the first adaptive mutation to occur and reach a substantial frequency is likely to go to fixation on a first-come-first-served basis. Therefore, under such conditions, mutationally-favored changes have a higher chance to go to fixation [84]. In contrast, when mutations occur more frequently in the population, there is an interplay between the relative difference between mutation rates and selection coefficients [84]. Recent analytical work found the exact expressions for the conditions needed for mutation-biased adaptation (i.e., the fixation of mutationally-favored variants) for a single-locus synthetic landscape, showing the specific threshold that the differences in mutation rates between mutational classes must exceed with respect to those in selection coefficients [86]. Both studies concur on the fact that, as mutation supply increases, therefore increasing clonal interference, the influence of mutation bias on adaptation will decrease, because late-arising variants with larger selection coefficients prevent the fixation of the early-arising variants that are favored by mutation [84], [86].

This growing body of theoretical work [84]–[86] sheds light both on the conditions in which mutation bias can act as a dispositional force in adaptive evolution, and on the evolutionary consequences of such mutational patterns on different aspects of adaptive processes.

1.6 EMPIRICAL CHARACTERISATION OF GENOTYPE-PHENOTYPE LANDSCAPES

There is one main limitation of using synthetic landscapes to study evolutionary processes such as mutation-biased adaptation. Namely, one must make strong assumptions about their structural properties. However, the empirical characterisation of landscapes skirt this issue. How do the structural properties of empirical landscapes influence mutation-biased adaptation? This is an open question. A rigorous approach to construct such empirical landscapes would require an exhaustive assignment of phenotypes or fitness measures to all possible genotype sequences of a given length, an extremely challenging task. For example, for a DNA alphabet composed of four nucleotides, ACTG, genotype space grows exponentially as 4^L, where L is the genotype's length. Thus, genotype space is usually referred to as hyper-astronomically large: all possible DNA sequences of even a short length would possess more mass than the whole observable universe [87]. However, over the last decade we have witnessed the development of novel high-throughput experimental techniques that allow for the empirical measurement of quantitative phenotypes or fitness for a relatively large number of genotypes in a parallel manner [88]–[91]. This number of genotypes is only a fraction of a much larger landscape, they nevertheless improve our understanding of the structures of landscapes in different biological domains [92]–[94]. For example, transfer RNAs are relatively short RNA molecules that carry amino acids to ribosomes for protein synthesis, and mutations in such molecules are associated with several human diseases, such as deafness and heart deficiency [95]. One study used deep sequencing to quantify the fitness effects of 207 point mutations in a transfer RNA gene, under high-temperature stress in the species model *Saccharomyces cerevisiae*. Three aspects of their results are worth highlighting: First, in correspondence with the neutral evolution theory, they found that most mutations are either neutral or deleterious, whereas only around 1% were beneficial. Second, they uncovered pervasive epistatic interactions between mutations. And third, the fitness measurements were correlated with the correct folding of the secondary structure of the transfer RNA molecules [96]. Overall, studies like this one can both provide more realistic landscape depictions and uncover the biophysical basis of fitness landscapes.

Transcription factor-DNA interactions

Given the immensity of sequence space, a sensible approach to empirically characterise and analyse complete genotype-phenotype landscapes is to focus on short sequences [97], [98]. Eukaryotic transcription factor DNA binding sites are usually around 10 nucleotides long [99], which makes them ideal candidates for such characterisation [100], [101]. Transcription factors control when and where transcriptional processes, namely the conversion of DNA to RNA, occur. They do so, along with other proteins, by attaching to DNA binding sites, to further recruit or block the recruitment of RNA polymerase binding to DNA. That is why mutations in DNA binding sites are highly relevant to understand transcription factor-DNA interactions: they can change the identity of the transcription factor that is able to engage in binding, as well as alter the affinity with which the site is bound [102]. Such binding interactions and their regulatory consequences are both crucial for organismal development and function [103], as well as for the evolution of phenotypic diversity [104], [105].

The benefit of understanding transcription factor-DNA interactions in this context is two-fold. It can reveal the structural properties of empirical genotype-phenotype landscapes [75], [106], while providing valuable information about the evolution of regulatory elements [75], [107], [108].

1.7 ASSOCIATIONS BETWEEN MUTATION RATES AND SELECTION COEFFICIENTS

Recent mutation accumulation experiments allow for the accurate characterisation of species-specific patterns of mutation bias (for an extensive review see [109]). Such a picture of mutation bias, in addition to the study of empirical fitness landscapes, could serve as valuable information to assess the underlying associations between mutation rates and selection coefficients for a large amount of adaptive mutations. The actual associations between these quantitative traits remains an open question that is of particular importance because one of the main axioms of evolutionary theory establishes an independence between the rates of occurrence of random mutations and their corresponding effects [110]. This means that the probability with which a particular mutation occurs is unrelated to its phenotypic, and potentially selective outcomes. This proposition is central to most models of adaptive evolution, including origin-fixation models: first a random mutation occurs, and then selection increases or decreases its frequency. They are two independent processes. The integration of empirically characterised mutation rates and selection coefficients to evolutionary models could provide important information about these fundamental factors of adaptive processes, reshaping our expectations about the correspondence between what is selectively beneficial and what is mutationally likely.

1.8 EMPIRICAL APPLICATIONS OF MUTATION BIAS

What are the implications of mutation-biased adaptation beyond evolutionary theory? The incorporation of mutation bias has improved the accuracy of a variety of evolutionary models [25], [111], [112]. For example, models of protein evolution were able to recapitulate patterns of empirical amino acid substitution, by integrating mutation bias, the structure of the genetic code and selection for protein thermodynamic stability [113]. Moreover, mutation bias has been observed in the evolution of resistance to multiple human-produced toxins, including insecticides, anti-parasitic and anti-viral drugs [21]. This can improve our understanding of such adaptive processes, allowing for the limitation of the proliferation of dangerous biological agents such as microbial pathogens and parasites.

One of the most threatening biological issues for our society are pandemics. A variety of SARS-CoV-2 genome investigations have uncovered mutation biases strongly favoring uracil content [114], while selection seems to go against uracil content [115], [116]. This has potential implications for vaccine development, because increasing uracil content could function as a strategy for the production of attenuated versions of viruses [115]. In addition, given the repeated evolution patterns of SARS-CoV-2 variants, studies characterising the role of mutation bias on virus adaptation may shed light on the forecasting of vaccine resistance, potentially leading to the development of vaccines that can confer longer-term immunity.

Cancer is another evolutionary process that affects human health. The evolutionary processes behind the transition from healthy somatic cells to cancer are being increasingly explored in recent

years [117]–[119]. For example, APOBEC (Apolipoprotein B mRNA Editing Catalytic Polypeptide-like) is a family of evolutionarily conserved proteins that bind RNA and single-stranded DNA, and are associated with DNA hypermutation and promiscuous RNA editing when there is loss of cellular control in APOBEC activity. A study of APOBEC-induced mutations in carcinoma cells found that the relative selective advantage of mutations for the cancer phenotype often differed from their prevalence. This means that variants with low selection coefficient but high mutation rates were recurrently observed, whereas some variants barely occurred despite having high selection coefficients due to their low mutation rates [118].

Overall, the incorporation of mutation bias in models of adaptive evolution on a variety of research domains can both expand our understanding of adaptive processes and provide more accurate descriptions of evolutionary outcomes.

1.9 THESIS OUTLINE

This thesis focuses on the evolutionary conditions for mutation-biased adaptation, and its potential consequences in different facets of evolution, from the evolution of eukaryotic gene regulation to the predictability of microbial evolution. We employ a combination of empirical data with computational and theoretical approaches to tackle the previously stated open questions. The thesis consists of two studies that have been published (Chapter 2 and Chapter 3), and a third study that is currently in an advanced stage of preparation (Chapter 4).

In Chapter 2, we construct empirical genotype-phenotype landscapes of transcription factor binding affinities for 746 transcription factors from 129 eukaryotic species. In such landscapes, we uncover the presence of composition bias, namely the prevalence of a particular type of mutation in the mutational paths of the landscape. The type of mutation we consider was a transition mutation, and we measure its prevalence relative to the null expectation that one transition occurs for every two transversions. We find that composition bias is common among accessible mutational paths to a landscape's global adaptive peak. We also show that such composition bias can interact with mutation bias to influence both the probability that the population will reach the optimal peak, and the predictability of the evolutionary process.

In Chapter 3, we study to what extent mutation bias shapes the spectrum of missense mutations underlying adaptation. For this, we use negative binomial regression to model observed numbers of adaptive codon-to-amino acid substitutions as a function of codon frequencies and neutral per-nucleotide mutation rates. We separately apply this approach to three large data sets of missense changes associated with adaptation in *Saccharomyces cerevisiae*, *Escherichia coli*, and *Mycobacterium tuberculosis*. We find that, in all three cases, mutation bias has a strong and roughly proportional influence on the spectra of mutations associated to adaptation. We additionally perform population genetic simulations and find that the predictive power of our framework depends on multiple factors, including mutation supply and the breadth and heterogeneity of the mutational target for adaptation.

In Chapter 4, we perform an initial exploration of the non-causal associations between mutation rates and selection coefficients induced by adaptive processes. For this, we combine theoretical and computational approaches, as well as analyses of large data sets of adaptive changes. Our results show that non-causal associations between mutation rate and selection coefficients can emerge under a variety of evolutionary conditions. This suggest that such associations may be different depending on whether they are determined from the subset of mutations that reached fixation, or from the complete set of adaptive mutations.

In Chapter 5, we discuss the main conclusions of this thesis, and highlight future research directions.

REFERENCES

[1] L. B. Alexandrov, J. Kim, N. J. Haradhvala, *et al.*, "The repertoire of mutational signatures in human cancer", *Nature*, vol. 578, no. 7793, pp. 94–101, 2020.

[2] R. M. Schaaper and R. L. Dunn, "Spectra of spontaneous mutations in *Escherichia coli* strains defective in mismatch correction: The nature of *in vivo* DNA replication errors", *Proceedings of the National Academy of Sciences*, vol. 84, pp. 6220–6224, 1987.

[3] Z. Zhang and M. Gerstein, "Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes", *Nucleic Acids Research*, vol. 31, pp. 5338–48, 2003.

[4] R. Hershberg and D. A. Petrov, "Evidence that mutation is universally biased towards AT in bacteria", *PLoS Genetics*, 2010.

[5] H. Lee, E. Popodi, H. Tang, and P. L. Foster, "Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing", *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[6] Y. O. Zhu, M. L. Siegal, D. W. Hall, and D. A. Petrov, "Precise estimates of mutation rate and spectrum in yeast", *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[7] M. D. Pauly, M. C. Procario, and A. S. Lauring, "A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses", *eLife*, vol. 6, e26437, 2017.

[8] S. Ossowski, K. Schneeberger, J. Lucas-Lledó, *et al.*, "The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*", *Science*, vol. 327, pp. 92–4, 2010.

[9] P. Keightley, U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. Blaxter, "Analysis of the genome sequences of 3 *Drosophila melanogaster* spontaneous mutation accumulation lines", *Genome research*, vol. 19, pp. 1195–201, 2009.

[10] S. Kucukyildirim, H. Long, W. Sung, S. Miller, T. Doak, and M. Lynch, "The rate and spectrum of spontaneous mutations in *Mycobacterium smegmatis*, a bacterium naturally devoid of the post-replicative mismatch repair pathway", *G3*, vol. 6, pp. 2157–2163, 2016.

[11] J. W. Schroeder, W. G. Hirst, G. A. Szewczyk, and L. A. Simmons, "The effect of local sequence context on mutational bias of genes encoded on the leading and lagging strands", *Curr Biol*, vol. 26, no. 5, pp. 692–7, 2016.

[12] M.-L. Weng, C. Becker, J. Hildebrandt, *et al.*, "Fine-grained analysis of spontaneous mutation spectrum and frequency in arabidopsis thaliana", *Genetics*, vol. 211, no. 2, pp. 703–714, 2019.

[13] X. Chen, Z. Chen, H. Chen, *et al.*, "Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes", *Science*, vol. 335, no. 6073, pp. 1235–1238, 2012.

[14] A. Mira and H. Ochman, "Gene location and bacterial sequence divergence", *Molecular biology and evolution*, vol. 19, no. 8, pp. 1350–1358, 2002.

[15] J Monroe, T. Srikant, P. Carbonell-Bejerano, *et al.*, "Mutation bias reflects natural selection in *Arabidopsis thaliana*", *Nature*, pp. 1–5, 2022.

[16] M. M. Dillon, W. Sung, M. Lynch, and V. S. Cooper, "The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*", *Genetics*, vol. 200, no. 3, pp. 935–946, 2015.

[17] D. Rokyta, P. Joyce, S. Caudle, and H. Wichman, "An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus", *Nature Genetics*, vol. 37, pp. 441–444, 2005.

[18] C. Maclean, G. Perron, and A. Gardner, "Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for Rifampicin resistance in *Pseudomonas aeruginosa*", *Genetics*, vol. 186, pp. 1345–54, 2010.

[19] A. Couce, A. Rodríguez-Rojas, and J. Blazquez, "Bypass of genetic constraints during mutator evolution to antibiotic resistance", *Proceedings of the Royal Society London B*, vol. 282, p. 20 142 698, 2015.

[20] A. M. Sackman, L. W. McGee, A. J. Morrison, *et al.*, "Mutation-driven parallel evolution during viral adaptation", *Molecular Biology and Evolution*, vol. 34, no. 12, pp. 3243–3253, 2017.

[21] A. Stoltzfus and D. M. McCandlish, "Mutational biases influence parallel adaptation", *Molecular Biology and Evolution*, vol. 34, no. 9, pp. 2163–2172, 2017.

[22] J. F. Storz, C. Natarajan, A. V. Signore, C. C. Witt, D. M. McCandlish, and A. Stoltzfus, "The role of mutation bias in adaptive molecular evolution: Insights from convergent changes in protein function", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, no. 1777, p. 20 180 238, 2019.

[23] J. L. Payne, F. Menardo, A. Trauner, *et al.*, "Transition bias influences the evolution of antibiotic resistance in Mycobacterium tuberculosis", *PLoS Biology*, vol. 17, no. 5, 2019.

[24] F. Bertels, C. Leemann, K. J. Metzner, and R. R. Regoes, "Parallel evolution of HIV-1 in a long-term experiment", *Molecular Biology and Evolution*, vol. 36, no. 11, pp. 2400–2414, 2019.

[25] S. Leighow, C. Liu, H. Inam, B. Zhao, and J. Pritchard, "Multi-scale predictions of drug resistance epidemiology identify design principles for rational drug design", *Cell Reports*, vol. 30, pp. 3951–3963, 2020.

[26] S. Katz, S. Avrani, M. Yavneh, H. S., J. Gross, and H. R., "Dynamics of adaptation during three years of evolution under long-term stationary phase", *Molecular Biology and Evolution*, 2021, In press.

[27] T. Smith, G. Ho, J. Christodoulou, *et al.*, "Extensive variation in the mutation rate between and within human genes associated with mendelian disease", *Human mutation*, vol. 37, no. 5, pp. 488–494, 2016.

[28] A. Siepel and D. Haussler, "Phylogenetic estimation of context-dependent substitution rates by maximum likelihood", *Molecular biology and evolution*, vol. 21, no. 3, pp. 468–488, 2004.

[29] T. Maeshiro and M. Kimura, "The role of robustness and changeability on the origin and evolution of genetic codes", *Proceedings of the National Academy of Sciences*, vol. 95, no. 9, pp. 5088–5093, 1998.

[30] J. Zhang, "Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes", *Journal of molecular evolution*, vol. 50, no. 1, pp. 56–68, 2000.

[31] C. Guo, I. C. McDowell, M. Nodzenski, *et al.*, "Transversions have larger regulatory effects than transitions", *BMC genomics*, vol. 18, no. 1, pp. 1–9, 2017.

[32] A. Stoltzfus and R. W. Norris, "On the causes of evolutionary transition:transversion bias", *Molecular Biology and Evolution*, vol. 33, no. 3, pp. 595–602, 2015.

[33] D. M. Lyons and A. S. Lauring, "Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses", *Molecular biology and evolution*, vol. 34, no. 12, pp. 3205–3215, 2017.

[34] Z. Zou and J. Zhang, "Are nonsynonymous transversions generally more deleterious than nonsynonymous transitions?", *Molecular biology and evolution*, vol. 38, no. 1, pp. 181–191, 2021.

[35] P. D. Keightley and A. Eyre-Walker, "Terumi mukai and the riddle of deleterious mutation rates", *Genetics*, vol. 153, no. 2, pp. 515–523, 1999.

[36] J. W. Drake, "Chaos and order in spontaneous mutation", *Genetics*, vol. 173, no. 1, pp. 1–8, 2006.

[37] M. Lynch, "The origins of eukaryotic gene structure", *Molecular biology and evolution*, vol. 23, no. 2, pp. 450–468, 2006.

[38] M. Lynch, B. Koskella, and S. Schaack, "Mutation pressure and the evolution of organelle genomic architecture", *science*, vol. 311, no. 5768, pp. 1727–1730, 2006.

[39] R. Holliday and G. Grigg, "Dna methylation and mutation", *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 285, no. 1, pp. 61–67, 1993.

[40] R. Z. Chen, U. Pettersson, C. Beard, L. Jackson-Grusby, and R. Jaenisch, "Dna hypomethylation leads to elevated mutation rates", *Nature*, vol. 395, no. 6697, pp. 89–93, 1998.

[41] I. B. Rogozin and Y. I. Pavlov, "Theoretical analysis of mutation hotspots and their dna sequence context specificity", *Mutation Research/Reviews in Mutation Research*, vol. 544, no. 1, pp. 65–85, 2003.

[42] J. E. Kucab, X. Zou, S. Morganella, *et al.*, "A compendium of mutational signatures of environmental agents", *Cell*, vol. 177, no. 4, pp. 821–836, 2019.

[43] F. A. Kondrashov and A. S. Kondrashov, "Measurements of spontaneous rates of mutations in the recent past and the near future", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1544, pp. 1169–1176, 2010.

[44] V. Katju and U. Bergthorsson, "Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches", *Genome Biology and Evolution*, vol. 11, no. 1, pp. 136–165, 2019.

[45] P. A. Lind and D. I. Andersson, "Whole-genome mutational biases in bacteria", *Proceedings of the National Academy of Sciences*, vol. 105, no. 46, pp. 17 878–17 883, 2008.

[46] A. Farlow, H. Long, S. Arnoux, *et al.*, "The spontaneous mutation rate in the fission yeast schizosaccharomyces pombe", *Genetics*, vol. 201, no. 2, pp. 737–744, 2015.

[47] D. R. Schrider, D. Houle, M. Lynch, and M. W. Hahn, "Rates and genomic consequences of spontaneous mutational events in drosophila melanogaster", *Genetics*, vol. 194, no. 4, pp. 937–954, 2013.

[48] S. Singhal, *Digest: Unpacking fitness effects of spontaneous mutations*, 2017.

[49] G. Stebbins, *Processes of Organic Evolution*, ser. Concepts of modern biology series. Prentice-Hall, 1966.

[50] G. G. Simpson, "Organisms and molecules in evolution", *Science*, vol. 146, no. 3651, pp. 1535–1538, 1964.

[51] T. Dobzhansky, *Evolution*. Frreeman, 1977.

[52] E. Mayr, "Animal species and evolution", Belknap Press of Harvard University Press, Tech. Rep., 1963.

[53] R. Fisher, *The Genetical Theory of Natural Selection*. London: Oxford University Press, 1930.

[54] J. Haldane, *The Causes of Evolution*. New York: Longmans, Green and Co., 1932, p. 235.

[55] M. Pagel, *Encyclopedia of Evolution*, ser. Encyclopedia of Evolution vol. 1. Oxford University Press, 2002.

[56] E. Margoliash, "Primary structure and evolution of cytochrome c", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 50, no. 4, p. 672, 1963.

[57] E. Zuckerkandl and L. Pauling, *Molecular Disease, Evolution, and Genic heterogeneity*. Academic Press, 1962.

[58] E. Zuckerkandl and L. Pauling, "Evolutionary divergence and convergence in proteins", in *Evolving Genes and Proteins*, Academic Press, 1965, pp. 97–166.

[59] J. H. Gillespie, "A simple stochastic gene substitution model", *Theoretical Population Biology*, vol. 23, no. 2, pp. 202–15, 1983.

[60] J. H. Gillespie, "Molecular evolution over the mutational landscape", *Evolution*, pp. 1116–1129, 1984.

[61] S. Kauffman and S. Levin, "Towards a general theory of adaptive walks on rugged landscapes", *Journal of Theoretical Biology*, vol. 128, no. 1, pp. 11–45, 1987.

[62] B. Lehner, "Genotype to phenotype: Lessons from model organisms for human genetics", *Nature Reviews Genetics*, vol. 14, no. 3, pp. 168–178, 2013.

[63] P. Alberch, "From genes to phenotype: dynamical systems and evolvability", *Genetica*, vol. 84, no. 1, pp. 5–11, 1991.

[64] G. P. Wagner and J. Zhang, "The pleiotropic structure of the genotype–phenotype map: The evolvability of complex organisms", *Nature Reviews Genetics*, vol. 12, no. 3, pp. 204–213, 2011.

[65] D. J. Lipman and W. J. Wilbur, "Modelling neutral and selective evolution of protein folding", *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 245, no. 1312, pp. 7–11, 1991.

[66] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker, "From sequences to shapes and back: A case study in RNA secondary structures", *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 255, no. 1344, pp. 279–284, 1994.

[67] S. Ciliberti, O. C. Martin, and A. Wagner, "Innovation and robustness in complex regulatory gene networks", *Proceedings of the National Academy of Sciences*, vol. 104, no. 34, pp. 13 591–13 596, 2007.

[68] E. Van Nimwegen, J. P. Crutchfield, and M. Huynen, "Neutral evolution of mutational robustness", *Proceedings of the National Academy of Sciences*, vol. 96, no. 17, pp. 9716–9720, 1999.

[69] J. A. Draghi, T. L. Parsons, G. P. Wagner, and J. B. Plotkin, "Mutational robustness can facilitate adaptation", *Nature*, vol. 463, no. 7279, pp. 353–355, 2010.

[70] S. Manrubia and J. A. Cuesta, "Evolution on neutral networks accelerates the ticking rate of the molecular clock", *Journal of The Royal Society Interface*, vol. 12, no. 102, p. 20 141 010, 2015.

[71] S. Schaper and A. A. Louis, "The arrival of the frequent: How bias in genotype-phenotype maps can steer populations to local optima", *PloS ONE*, vol. 9, no. 2, e86635, 2014.

[72] H. A. Orr, "Fitness and its role in evolutionary genetics", *Nature Reviews Genetics*, vol. 10, no. 8, pp. 531–539, 2009.

[73] T. F. Hansen, "On the definition and measurement of fitness in finite populations", *Journal of Theoretical Biology*, vol. 419, pp. 36–43, 2017.

[74] S. Manrubia, J. A. Cuesta, J. Aguirre, *et al.*, "From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics", *Physics of Life Reviews*, vol. 38, pp. 55–106, 2021.

[75] J. Aguilar-Rodríguez, J. L. Payne, and A. Wagner, "A thousand empirical adaptive landscapes and their navigability", *Nature Ecology and Evolution*, vol. 1, no. 2, 2017.

[76] F. A. Kondrashov and A. S. Kondrashov, "Multidimensional epistasis and the disadvantage of sex", *Proceedings of the National Academy of Sciences*, vol. 98, no. 21, pp. 12 089–12 092, 2001.

[77] J. A. G. de Visser, S.-C. Park, and J. Krug, "Exploring the effect of sex on empirical fitness landscapes", *The American Naturalist*, vol. 174, no. S1, S15–S30, 2009.

[78] S. Gavrilets, *Fitness Landscapes and the Origin of Species*. Princeton University Press, 2004.

[79] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, "Predictability of evolutionary trajectories in fitness landscapes", *PLoS Computational Biology*, vol. 7, no. 12, e1002302, 2011.

[80] J. De Visser and J. Krug, "Empirical fitness landscapes and the predictability of evolution", *Nature Reviews Genetics*, vol. 15, no. 7, pp. 480–490, 2014.

[81] D. M. Weinreich, R. A. Watson, and L. Chao, "Perspective: Sign epistasis and genetic costraint on evolutionary trajectories", *Evolution*, vol. 59, no. 6, pp. 1165–1174, 2005.

[82] F. J. Poelwijk, S. Tănase-Nicola, D. J. Kiviet, and S. J. Tans, "Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes", *Journal of Theoretical Biology*, vol. 272, no. 1, pp. 141–144, 2011.

[83] C. Bank, "Epistasis and adaptation on fitness landscapes", *arXiv preprint arXiv:2204.13321*, 2022.

[84] L. Y. Yampolsky and A. Stoltzfus, "Bias in the introduction of variation as an orienting factor in evolution", *Evolution & Development*, vol. 3, no. 2, pp. 73–83, 2001.

[85] A. Stoltzfus, "Mutation-biased adaptation in a protein NK model", *Molecular Biology & Evolution*, vol. 23, pp. 1852–1862, 2006.

[86] A. d. A. Soares, L. Wardil, L. B. Klaczko, and R. Dickman, "Hidden role of mutations in the evolutionary process", *Physical Review E*, vol. 104, p. 044 413, 4 2021.

[87] A. A. Louis, "Contingency, convergence and hyper-astronomical numbers in biological evolution", *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 58, pp. 107–116, 2016.

[88] J. Domingo, P. Baeza-Centurion, and B. Lehner, "The causes and consequences of genetic interactions (epistasis)", *Annual Review of Genomics and Human Genetics*, vol. 20, pp. 433–460, 2019.

[89] J. B. Kinney and D. M. McCandlish, "Massively parallel assays and quantitative sequence–function relationships", *Annual Review of Genomics and Human Genetics*, vol. 20, pp. 99–127, 2019.

[90] I. Fragata, A. Blanckaert, M. A. D. Louro, D. A. Liberles, and C. Bank, "Evolution in the light of fitness landscape theory", *Trends in Ecology & Evolution*, vol. 34, no. 1, pp. 69–82, 2019.

[91] P. T. Dolan, S. Taguwa, M. A. Rangel, *et al.*, "Principles of dengue virus evolvability derived from genotype-fitness maps in human and mosquito cells", *eLife*, vol. 10, e61921, 2021.

[92] P. Julien, B. Miñana, P. Baeza-Centurion, J. Valcárcel, and B. Lehner, "The complete local genotype–phenotype landscape for the alternative splicing of a human exon", *Nature Communications*, vol. 7, no. 1, pp. 1–8, 2016.

[93] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, *et al.*, "Local fitness landscape of the green fluorescent protein", *Nature*, vol. 533, no. 7603, pp. 397–401, 2016.

[94] G. Diss and B. Lehner, "The genetic landscape of a physical interaction", *eLife*, vol. 7, e32472, 2018.

[95] J. A. Abbott, C. S. Francklyn, and S. M. Robey-Bond, "Transfer RNA and human disease", *Frontiers in Genetics*, vol. 5, p. 158, 2014.

[96] C. Li, W. Qian, C. J. Maclean, and J. Zhang, "The fitness landscape of a tRNA gene", *Science*, vol. 352, no. 6287, pp. 837–840, 2016.

[97] D. Ray, H. Kazan, E. T. Chan, *et al.*, "Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins", *Nature biotechnology*, vol. 27, no. 7, pp. 667–670, 2009.

[98] G. Badis, M. F. Berger, A. A. Philippakis, *et al.*, "Diversity and complexity in DNA recognition by transcription factors", *Science*, vol. 324, no. 5935, pp. 1720–1723, 2009.

[99] A. J. Stewart and J. B. Plotkin, "Why transcription factor binding sites are ten nucleotides long", *Genetics*, vol. 192, no. 3, pp. 973–985, 2012.

[100] D. E. Newburger and M. L. Bulyk, "UniPROBE: An online database of protein binding microarray data on protein-DNA interactions", *Nucleic Acids Research*, vol. 37, no. SUPPL. 1, 2009.

[101] M. T. Weirauch, A. Yang, M. Albu, *et al.*, "Determination and inference of eukaryotic transcription factor sequence specificity", *Cell*, vol. 158, no. 6, pp. 1431–1443, 2014.

[102] J. C. Kwasnieski, I. Mogno, C. A. Myers, J. C. Corbo, and B. A. Cohen, "Complex effects of nucleotide variants in a mammalian cis-regulatory element", *Proceedings of the National Academy of Sciences*, vol. 109, no. 47, pp. 19 498–19 503, 2012.

[103] I. Guerreiro, A. Nunes, J. M. Woltering, *et al.*, "Role of a polymorphism in a hox/pax-responsive enhancer in the evolution of the vertebrate spine", *Proceedings of the National Academy of Sciences*, vol. 110, no. 26, pp. 10 682–10 686, 2013.

[104] G. A. Wray, "The evolutionary significance of cis-regulatory mutations", *Nature Reviews Genetics*, vol. 8, no. 3, pp. 206–216, 2007.

[105] N. Gompel, B. Prud'homme, P. J. Wittkopp, V. A. Kassner, and S. B. Carroll, "Chance caught on the wing: Cis-regulatory evolution and the origin of pigment patterns in drosophila", *Nature*, vol. 433, no. 7025, pp. 481–487, 2005.

[106] J. Aguilar-Rodríguez, L. Peel, M. Stella, A. Wagner, and J. L. Payne, "The architecture of an empirical genotype-phenotype map", *Evolution*, vol. 72, no. 6, pp. 1242–1260, 2018.

[107] J. L. Payne, F. Khalid, and A. Wagner, "RNA-mediated gene regulation is less evolvable than transcriptional regulation", *Proceedings of the National Academy of Sciences*, vol. 115, no. 15, E3481–E3490, 2018.

[108] M. Srivastava and J. L. Payne, "The transformability of genotype-phenotype landscapes", *bioRxiv*, 2022.

[109] V. Katju and U. Bergthorsson, "Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches", *Genome Biology and Evolution*, vol. 11, no. 1, pp. 136–165, 2018.

[110] D. Futuyma, *Evolutionary Biology*. Sinauer Associates, 1986, ISBN: 9780878931880.

[111] H. A. Orr, "The population genetics of adaptation: the adaptation of DNA sequences", *Evolution*, vol. 56, no. 7, pp. 1317–1330, 2002.

[112] H. A. Orr, "The distribution of fitness effects among beneficial mutations", *Genetics*, vol. 163, no. 4, pp. 1519–1526, 2003.

[113] C. Norn, B. I. Wicky, D. Juergens, *et al.*, "Protein sequence design by conformational landscape optimization", *Proceedings of the National Academy of Sciences*, vol. 118, no. 11, e2017228118, 2021.

[114] D. J. Hamelin, D. Fournelle, J.-C. Grenier, *et al.*, "The mutational landscape of SARS-CoV-2 variants diversifies T cell targets in an HLA-supertype-dependent manner", *Cell Systems*, vol. 13, no. 2, 143–157.e3, 2022.

[115] A. M. Rice, A. Castillo Morales, A. T. Ho, *et al.*, "Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: Implications for vaccine design", *Molecular Biology and Evolution*, vol. 38, no. 1, pp. 67–83, 2020.

[116] M. Kosuge, E. Furusawa-Nishii, K. Ito, Y. Saito, and K. Ogasawara, "Point mutation bias in SARS-CoV-2 variants results in increased ability to stimulate inflammatory responses", *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.

[117] A. O. Giacomelli, X. Yang, R. E. Lintner, *et al.*, "Mutational processes shape the landscape of tp53 mutations in human cancer", *Nature Genetics*, vol. 50, no. 10, pp. 1381–1387, 2018.

[118] V. L. Cannataro, S. G. Gaffney, T. Sasaki, *et al.*, "APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma", *Oncogene*, vol. 38, no. 18, pp. 3475–3487, 2019.

[119] V. Cannataro, J. Mandell, and J. Townsend, "Attribution of cancer origins to endogenous, exogenous, and actionable mutational processes", *bioRxiv*, pp. 10.1101/2020.10.24.352989, 2020.

2

MUTATION BIAS INTERACTS WITH COMPOSITION BIAS TO INFLUENCE ADAPTIVE EVOLUTION

Published as: Alejandro V. Cano and Joshua L. Payne (2020) Mutation bias interacts with composition bias to influence adaptive evolution. *PLoS Computational Biology* 16(9): e1008296. https://doi.org/10.1371/journal.pcbi.1008296

Author's contributions: A.V.C. and J.L.P. designed research; A.V.C. performed research; A.V.C. and J.L.P. analyzed data; and A.V.C. and J.L.P. wrote the paper.

2.1 ABSTRACT

Mutation is a biased stochastic process, with some types of mutations occurring more frequently than others. Previous work has used synthetic genotype-phenotype landscapes to study how such mutation bias affects adaptive evolution. Here, we consider 746 empirical genotype-phenotype landscapes, each of which describes the binding affinity of target DNA sequences to a transcription factor, to study the influence of mutation bias on adaptive evolution of increased binding affinity. By using empirical genotype-phenotype landscapes, we need to make only few assumptions about landscape topography and about the DNA sequences that each landscape contains. The latter is particularly important because the set of sequences that a landscape contains determines the types of mutations that can occur along a mutational path to an adaptive peak. That is, landscapes can exhibit a composition bias — a statistical enrichment of a particular type of mutation relative to a null expectation, throughout an entire landscape or along particular mutational paths — that is independent of any bias in the mutation process. Our results reveal the way in which composition bias interacts with biases in the mutation process under different population genetic conditions, and how such interaction impacts fundamental properties of adaptive evolution, such as its predictability, as well as the evolution of genetic diversity and mutational robustness.

AUTHOR SUMMARY

Mutation is often depicted as a random process due its unpredictable nature. However, such randomness does not imply uniformly distributed outcomes, because some DNA sequence changes happen more frequently than others. Such mutation bias can be an orienting factor in adaptive evolution, influencing the mutational trajectories populations follow toward higher-fitness genotypes. Because these trajectories are typically just a small subset of all possible mutational trajectories, they can exhibit composition bias - an enrichment of a particular kind of DNA sequence change, such as transition or transversion mutations. Here, we use empirical data from eukaryotic transcriptional regulation to study how mutation bias and composition bias interact to influence adaptive evolution.

2.2 INTRODUCTION

Mutation exhibits many forms of bias, both in genomic location and toward particular DNA sequence changes [1]. For instance, a bias toward transitions (mutations that change a purine to a purine, or a pyrimidine to a pyrimidine), relative to transversions (mutations that change a purine to a pyrimidine, or vice versa; Fig. 2.1a), has been widely observed in studies of mutation spectra, such as those based on reversion assays [2], [3], mutation accumulation experiments [4]–[6], sequence comparisons of closely related species [7]–[10], and analyses of putatively neutral polymorphisms in natural populations [11]. Because mutation provides the raw material of evolution, mutation bias may influence adaptive evolutionary change [12]–[14]. Indeed, transition bias has influenced the adaptive evolution of phenotypes as different as antibiotic resistance in *Mycobacterium tuberculosis* [15] and increased hemoglobin-oxygen affinity in high-altitude birds [16].

Adaptive evolution is often conceptualized as a hill-climbing process in a genotype-phenotype landscape, in which each location or coordinate corresponds to a genotype in an abstract genotype space, and the elevation of each location corresponds to fitness or some related quantitative phenotype [17], [18]. The topography of a genotype-phenotype landscape influences a wide range of evolutionary phenomena [19]–[21], including the evolution of genetic diversity, mutational robustness, and evolvability, as well as the predictability of the evolutionary process itself [19]. It also influences landscape navigability — the ability of an evolving population to reach the global adaptive peak via DNA mutation and natural selection [22]. Smooth, single-peaked landscapes are highly navigable, whereas rugged landscapes are not, because evolving populations can become trapped on local peaks [23], which frustrates further adaptive change [19], [21], [24].

Previous work has studied the interplay of mutation bias and landscape topography, and its influence on adaptive evolution, using synthetic genotype-phenotype landscapes [25], [26]. These studies have revealed that in single-peaked landscapes, mutation bias does not influence navigability, although it can influence the adaptive trajectory to the global peak, such that the average composition of the sequences in the trajectory reflects the sequence bias of the mutation process [26]. In this sense, mutation bias can be thought of as an "orienting factor" in evolution [25], which may affect predictability by making some mutational trajectories more likely than others. For example, in a simple two-locus model with two adaptive peaks of different heights, mutation bias increases the probability with which an evolving population converges on the suboptimal, but mutationally-favored peak [25]. Mutation bias is therefore capable of influencing the navigability of rugged landscapes. The extent to which it does, however, depends upon population genetic conditions [25]. Specifically, when the mutation supply is low, the first adaptive mutation to reach a substantial frequency is likely to go to fixation ("first-come-first-served"), and a bias in mutation supply will therefore influence which genetic changes drive adaptation. In contrast, when the mutation supply is high, the fittest adaptive mutation is likely to go to fixation ("pick-the-winner"), and a bias in mutation supply will have less of an effect on adaptation.

The study of genotype-phenotype landscapes is currently being transformed by methodological advances, including those in genome editing and in massively parallel assays such as deep mutational scanning [15], [20], [24], [27]. These facilitate the assignment of phenotypes to a large number of genotypes, and thus allow for the construction of empirical genotype-phenotype landscapes directly from experimental data. Examples include the "splicing-in" of exons [28], the binding preferences and enzymatic activities of macromolecules [29]–[32], the gene expression patterns of regulatory circuits [33], and the carbon utilization profiles of metabolic pathways [34]. In nearly all of these examples, the genotype-phenotype landscape is necessarily incomplete, representing only a small fraction of a much larger landscape, which cannot be constructed in its entirety due to the hyper-astronomical size of the corresponding genotype space [18]. Complete landscapes — those in which a phenotype is assigned to all possible genotypes — can only be constructed for systems with sufficiently small genotype spaces.

Transcription factor-DNA interactions are one such system [22], [35]–[37]. Transcription factors are regulatory proteins that bind DNA to induce or inhibit gene expression [38]. The DNA sequences they bind are typically short (6-12nt) [39], which makes it possible to exhaustively characterize transcription factor binding preferences and thus to construct complete genotype-

phenotype landscapes of transcription factor binding affinities [22]. In such landscapes, genotypes are short DNA sequences (transcription factor binding sites), and the phenotype of a sequence is its relative binding affinity for a transcription factor. Understanding how DNA mutations affect transcription factor binding affinity is important because such mutations are commonly implicated in disease [40], [41], as well as in evolutionary adaptations and innovations [42], [43]. Previous work on these landscapes has revealed that they are highly navigable [22]. They tend to contain few adaptive peaks, which comprise binding sites that are both mutationally robust and accessible, meaning that it is typically possible to reach these peaks via a series of mutations that only move "uphill." These peaks often comprise multiple sequences, which facilitates the evolution of genetic diversity in high-affinity binding sites [22], [36]. Additionally, these landscapes often interface and overlap with one another, which has implications for the evolvability of transcription factor binding sites [35], [36].

There are two features that differentiate these landscapes from other empirical landscapes. First, they are complete. They comprise a measure of relative binding affinity for all possible DNA sequences of length eight. Second, data are available for many such landscapes, which facilitates statistical analyses of how landscape properties influence adaptive evolution. Moreover, in contrast to synthetic landscapes, these empirical landscapes make very few assumptions about topography and about the DNA sequences each landscape contains. The latter is particularly important in the context of mutation bias, because the set of sequences a landscape contains determines the kinds of mutations that are present in adaptive mutational trajectories. For example, the TATA-binding protein binds sequences enriched for thymines and adenines, which means that most of the mutations present in this protein's genotype-phenotype landscape are transversions (A $>$ T or T $>$ A). Such composition bias — an enrichment of a particular type of mutation relative to a null expectation, throughout an entire landscape or along particular mutational paths — likely interacts with mutation bias to influence various aspects of adaptive evolution, such as landscape navigability, enhancing it when mutation is biased toward transversions, and hindering it when mutation is biased toward transitions. However, to our knowledge, the interaction between mutation bias and composition bias, and its influence on adaptive evolution, has not been studied, neither in the context of synthetic nor empirical genotype-phenotype landscapes.

Here, we study this interaction and its influence on adaptive evolution using 746 empirical genotype-phenotype landscapes of transcription factor binding affinities, under the assumption of selection for increased binding affinity (although selection does not always act to increase binding affinity [44], [45]). We find that when the mutation supply is low, mutation bias can increase or decrease landscape navigability as well as the predictability of evolution, depending upon whether mutation bias is aligned with composition bias in the adaptive trajectories to the global peak (i.e., both forms of bias are toward the same type of mutation — transitions or transversions). When the mutation supply is high, mutation bias does not influence navigability, as expected on theoretical grounds [25], but it can influence how a population is distributed throughout the landscape, which has implications for the evolution of genetic diversity, mutational robustness, and evolvability. Taken together, our results show that mutation bias and composition bias interact to influence adaptive evolution under a broad range of population genetic conditions.

2.3 RESULTS

Genotype-phenotype landscapes exhibit composition bias in accessible mutational paths

We used protein-binding microarray data [46], [47] to construct empirical genotype-phenotype landscapes of transcription factor binding affinities for 746 transcription factors from 129 eukaryotic species, representing 48 distinct DNA binding domain structural classes (Methods; Table S1). For each transcription factor, these data include an enrichment score (*E*-score) — a proxy for relative binding affinity — for all possible 32,896 DNA sequences of length eight. We constructed one landscape per transcription factor, using only those DNA sequences that specifically bound the transcription factor, as indicated by an *E*-score exceeding 0.35 [22], [35], [36]. We represented each landscape as a genotype network, in which nodes are transcription factor binding sites and edges connect nodes if their corresponding sequences differ by a single point mutation (Fig. 2.1b) [35]. For some transcription factors ($\sim$37%), the genotype network fragmented into several disconnected components. When this occurred, we only considered the largest component, which always comprised more than 100 bound sequences. We refer to this as the dominant genotype network. We discarded non-dominant components because they usually comprised few sequences and rarely met our size requirement of 100 sequences. (Fig. S2.1). Each dominant genotype network formed the substrate of a genotype-phenotype landscape, whose surface was defined by the relative affinities (*E*-scores) of the network's constituent binding sites [22]. We accounted for noise in the protein binding microarray data using a noise threshold δ, which allowed us to determine whether two DNA sequences differed in their binding affinity [22] (Methods). We refer to a mutation as accessible if it increases binding affinity more than δ, and we refer to a series of accessible mutations as an accessible mutational path (Fig. 2.1c).

We developed a measure of composition bias on the unit interval, which we applied to entire landscapes and to accessible mutational paths (Methods). It is based on the null expectation that one transition occurs per every two transversions. When this measure equals 0.5, there is no composition bias. Values below 0.5 imply a composition bias toward transversions, whereas values above 0.5 imply a composition bias toward transitions. Fig. 2.2a shows the distribution of composition bias across all 746 landscapes. When considering all mutations in a landscape, we observed variation in composition bias ranging from 0.19 to 0.75, but as a whole, this distribution did not differ from our null expectation of one transition per every two transversions (black distribution in Fig. 2.2a, one-sample t test, $t_{745} = -0.4310$, $p = 0.66$). In contrast, when we considered accessible mutational paths to the global peak, we found significant composition bias toward transversions (white distribution in Fig. 2.2a, one-sample t test, $t_{745} = -14.3941$, $p < 10^{-40}$). This bias was sometimes extreme. For example, for 68 transcription factors, fewer than 1 in 7 mutations on an accessible mutational path were transitions, more than a three-fold decrease relative to our null expectation. Such composition bias varied among transcription factors from different DNA binding domain structural classes (Fig. 2.2b). For example, AT hook transcription factors had a strong bias toward transversions, because their binding sites are enriched for adenine and thymine bases, and thus have a very low GC content. At the opposite extreme were transcription factors with a basic helix-loop-helix domain (bHLH), which had a bias toward transitions, because the sequences they bind are more neutral in terms of GC content (average

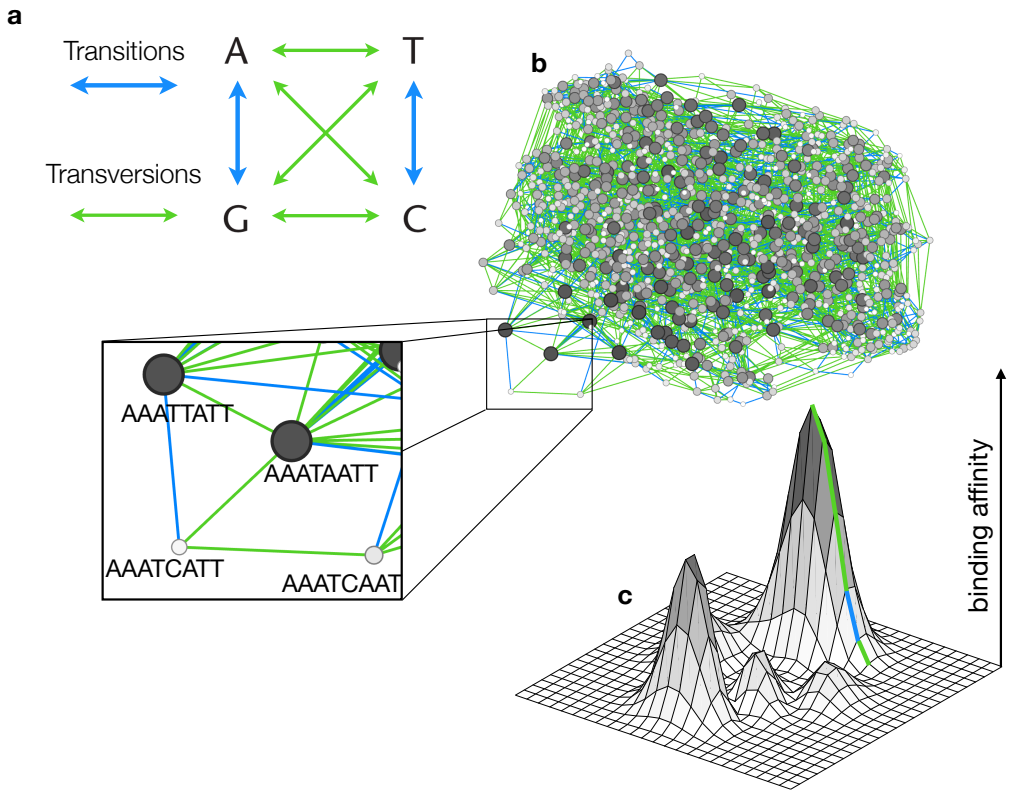

FIGURE 2.1: **Genotype-phenotype landscapes of transcription factor binding affinities and their composition bias. (a)** Transitions are mutations from a purine to a purine, or from a pyrimidine to a pyrimidine. Transversions are mutations from a purine to a pyrimidine, or vice versa. **(b)** The dominant genotype network for the yeast transcription factor Sum1. Each node corresponds to a DNA sequence that binds Sum1 with an E-score > 0.35. Node size is proportional to number of connections (bigger = more) and color to binding affinity (darker = higher). Two nodes are connected by an edge if their corresponding sequences differ by a single point mutation (e.g., see inset), either a transition (blue edges) or a transversion (green edges). **(c)** Schematic representation of a genotype-phenotype landscape, and an accessible mutational path to the global peak involving four transversions and one transition.

GC content = 0.61). Landscapes with either many high or many low GC content sequences were more likely to exhibit composition bias toward transversion mutations (Fig. S2.2). Overall, we observed more extreme bias toward transversions than toward transitions in accessible mutational paths (Fig. 2.2a), a bias that became even more pronounced as we increased the noise parameter δ (Fig. S2.3). This is because transversions tend to cause larger changes in binding affinity [48], and thus have larger regulatory effects [49] — a phenomenon we observed mainly near the global peak (Fig. S2.4). Because these distributions deviate so strongly from our null expectation, we focus on the composition bias of accessible mutational paths in all subsequent analyses, and we refer to this simply as composition bias for brevity.

To test the sensitivity of these observations to our use of E-score as a proxy for relative binding affinity, we considered an alternative proxy — the median signal intensity Z-score — which is available for 713 of the 746 transcription factors in our dataset. We found a strong correlation between the composition bias calculated from the two scores (Pearson's correlation coefficient $r = 0.7212$, $p < 10^{-10}$) (Fig. S2.5), and although composition bias often varied quantitatively among the two scores, it did not often vary qualitatively. That is, it was uncommon for landscapes to switch from exhibiting a composition bias toward transversions to a composition bias toward transitions, or vice versa. There were only 83 transcription factors (12%) for which such switching occurred, and of these roughly half ($\sim$48%) exhibited little to no composition bias when using E-scores (in the range (0.45, 0.55)). We therefore used E-scores for the rest of our analyses. In addition, we tested the sensitivity of our results to an alternative, less conservative definition of an accessible mutational path, which included any mutational steps that did not decrease binding affinity beyond some threshold (δ - our noise parameter). Fig. S2.6 shows that this alternative definition of accessibility results in composition biases that are highly similar to those observed under our initial more conservative definition, which we therefore use for the rest of our analyses, unless otherwise noted.

Mutation bias interacts with composition bias to influence landscape navigability

We first explored how mutation bias and composition bias interact to influence adaptive evolution when the mutation supply is low and selection is strong. Under these population genetic conditions, only one mutation is present in the population at any time [50], which makes the process amenable to modeling as a random walk in genotype space (Methods). In this framework, each time step corresponds to the number of generations needed for a mutation to go to fixation, and the fixation probability of a mutant is proportional to its binding affinity and to the likelihood of that particular type of mutation occurring. The latter was determined by a mutation bias parameter α, which is defined similarly to our measure of composition bias (Methods). Specifically, when $\alpha = 0.5$ there is no bias in mutation supply, values of α below 0.5 mean that the mutation supply is biased toward transversions, whereas values above 0.5 mean that the mutation supply is biased toward transitions.

As a measure of landscape navigability, we calculated the probability of reaching the global peak (Methods). Fig. 2.3 shows how mutation bias and composition bias interact to influence landscape navigability. We found that mutation bias could either enhance or diminish landscape navigability, relative to an unbiased mutation supply, depending upon whether the bias in

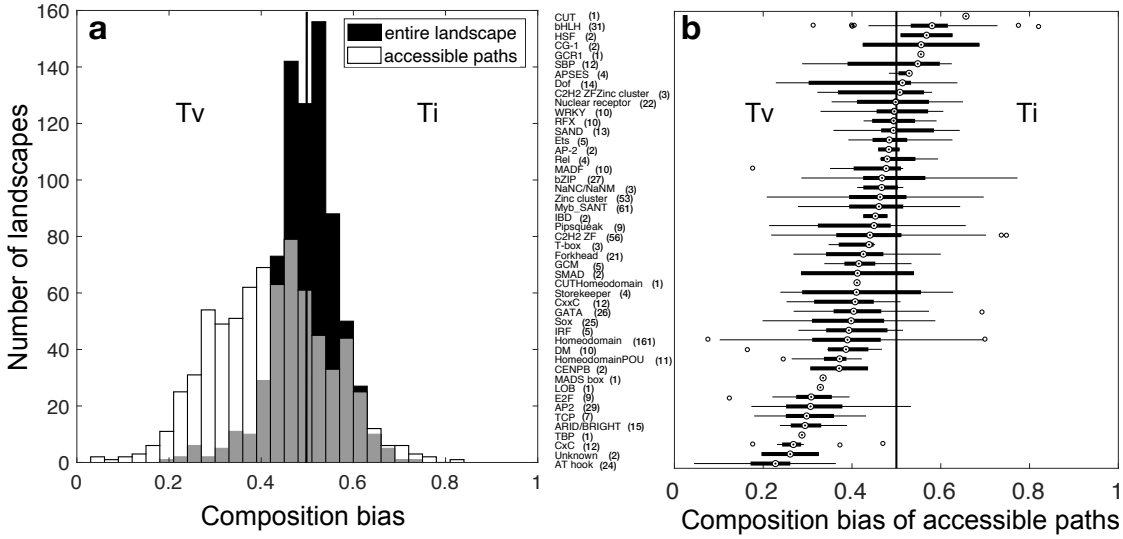

FIGURE 2.2: **Accessible mutational paths exhibit composition bias toward transversions. (a)** The black bars show the distribution of composition bias across entire landscapes. The white bars show the distribution of composition bias across accessible mutational paths to the global peak, starting from the 10% of binding sites with the lowest affinities in each landscape. Gray indicates the overlap in the distributions. Data pertain to all 746 landscapes. **(b)** Composition bias of accessible mutational paths, with landscapes grouped by DNA binding domain structural class. Numbers in parentheses indicate the number of transcription factors per class in our dataset.

mutation supply aligned with the composition bias of the accessible mutational paths in the landscape (i.e., the two forms of bias were toward the same kind of mutations — transitions or transversions). This effect is clearly seen when comparing the value of mutation bias that maximizes the probability of reaching the global peak across the five panels of Fig. 2.3 (dashed vertical lines), which group landscapes according to their composition bias. Fig. S2.7 shows the correlation between the mutation bias that maximizes P_{peak} and composition bias, measured using two different definitions of accessible mutational paths. Landscapes with extreme composition bias exhibited increased sensitivity to mutation bias, relative to landscapes without composition bias. For example, for landscapes with a strong composition bias toward transversions (Fig. 2.3a), the probability of reaching the global peak increased from 0.12 with a strong mutation bias toward transitions to 0.31 with a strong mutation bias toward transversions — a 2.6-fold increase. In contrast, for landscapes without composition bias (Fig. 2.3c), only a 1.7-fold increase in the probability of reaching the global peak was obtained by varying the bias in mutation supply.

Overall, landscapes with strong composition bias were more navigable than those with intermediate or no bias, in that they had higher probabilities of reaching the global peak. We reasoned this is because landscapes with strong composition bias tend to contain fewer binding sites than those with intermediate or no bias (Fig. S2.8), and the probability of evolving to a landscape's global peak decreases with the number of binding sites in the landscape (as shown in Fig. S2.9 for an unbiased mutation supply). In addition, our finding that navigability will be enhanced when mutation bias and composition bias are aligned and diminished otherwise is insensitive to whether selection acts to increase or decrease binding affinity (Fig. S2.10). This is important

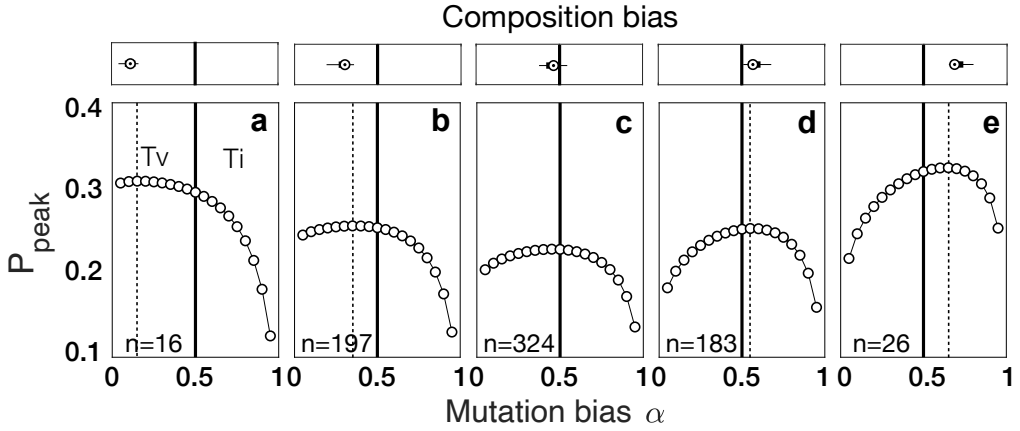

FIGURE 2.3: **Mutation bias interacts with composition bias to influence landscape navigability.** The probability P_{peak} of reaching the global peak is shown for 19 different values of the mutation bias parameter α. The solid vertical lines indicate no bias in mutation supply ($\alpha = 0.5$) and the dashed vertical lines indicate the value of α that maximizes P_{peak}. Landscapes are grouped based on their composition bias and the distribution of composition bias per panel is shown on top of each panel. The number of landscapes per panel is indicated is the bottom left corner.

because low-affinity binding sites often drive gene expression, for example in developmental enhancers [44], [45].

Mutation bias interacts with composition bias to influence the predictability of evolution

When the mutation supply is low, a bias in mutation supply can influence an evolving population's adaptive trajectory through a genotype-phenotype landscape, making some mutational paths more likely than others [25], [26]. Mutation bias may therefore influence the predictability of evolution. We next explored how mutation bias interacts with composition bias to influence the predictability of evolution, quantifying predictability using a measure of path entropy [51] (Methods). This measure takes on low values when an evolving population tends to take few mutational paths to the global peak, each with high probability. It takes on high values when an evolving population tends to take many mutational paths to the global peak, each with low probability. It is therefore inversely related to the predictability of evolution. Fig. S2.11 shows the relationship between path entropy and the mutation bias parameter α for 50 randomly chosen landscapes.

As expected, we found that mutation bias is considerably more likely to increase the predictability of evolution than to decrease it, in line with the notion of mutation bias as an orienting factor in evolution [25]. Across all landscapes, there was at least one mutation bias value that decreased path entropy relative to when there was no mutation bias (Fig. 2.4a), and on average 73% of mutation bias values decreased path entropy relative to when there was no mutation bias (Fig. S2.12). Moreover, a mutation bias toward transitions minimized path entropy more

often than one toward transversions (Fig. 2.4a). Mutation bias therefore readily increases the predictability of evolution. Fig. 2.4b shows how the misalignment between mutation bias and composition bias usually leads to the minimization of path entropy, thus increasing predictability. For landscapes with little to no composition bias, a strong mutation bias toward transitions was more likely to minimize path entropy than a strong mutation bias toward transversions. This is because in landscapes with no composition bias there are twice as many transversions than transitions, so a mutation bias toward transitions represents a greater evolutionary constraint in such cases. To determine the extent to which entropy changes in response to mutation bias, we calculated the ratio of the entropy observed in the absence of mutation bias to the minimum entropy caused by mutation bias ($\text{entropy}_{\text{no bias}}/\text{entropy}_{\text{min}}$), for each landscape. Path entropy decreased by an average of approximately 2-fold and 3-fold in the most extreme mutation biases toward transversions and transitions, respectively (Fig. 2.4c).

Perhaps counter-intuitively, we found that mutation bias can also decrease the predictability of evolution. For 693 landscapes (93%), there was at least one mutation bias value that increased path entropy relative to when there was no mutation bias (Fig. 2.4d), and on average across all 746 landscapes, 27% of mutation bias values increased path entropy relative to when there was no mutation bias (Fig. S2.12). To explore this further, we determined the mutation bias parameter α that maximized path entropy for each landscape. Fig. 2.4d shows the distribution of this parameter, which varied across the full range of α values considered. We hypothesized that the value of α that maximizes path entropy will be positively correlated with composition bias, such that path entropy will be maximized when these biases are aligned. Fig. 2.4e reveals this positive correlation, which is statistically significant, but weak in explanatory power (Spearman's rank correlation coefficient $\rho = 0.13$, $p < 10^{-4}$). The reason is that our measure of composition bias does not capture how mutations are distributed across accessible mutational paths, which strongly influences the value of α that maximizes path entropy (Fig. S2.13). In addition, Fig. 2.4f shows the relative change in entropy ($\text{entropy}_{\text{max}}/\text{entropy}_{\text{no bias}}$) in relation to the mutation bias parameter α that maximized path entropy. It reveals that path entropy increased by up to 7-fold, and that the largest increases were associated with the most extreme biases in mutation supply, either toward transitions or toward transversions. In these cases, evolution became less predictable because an evolving population could traverse a greater diversity of accessible mutational paths. In sum, these analyses reveal that mutation bias and composition bias interact to influence the predictability of evolution, in most cases increasing predictability, but in many others decreasing it. Thus, whether mutation bias acts as an orienting or dispersive factor in evolution depends upon the prevalence and type of composition bias in the landscape.

Mutation bias influences the distribution of polymorphic populations in genotype-phenotype landscapes

We next explored how mutation bias influences adaptive evolution when the mutation supply is high and selection is strong. Under these population genetic conditions, multiple mutations coexist in the population and compete for fixation. Because this process is challenging to model analytically, we used computer simulations of a Wright-Fisher model of evolutionary dynamics (Methods). As in our previous analyses, we used the probability of reaching the global peak as a measure of landscape navigability. We found that mutation bias has no effect on navigability

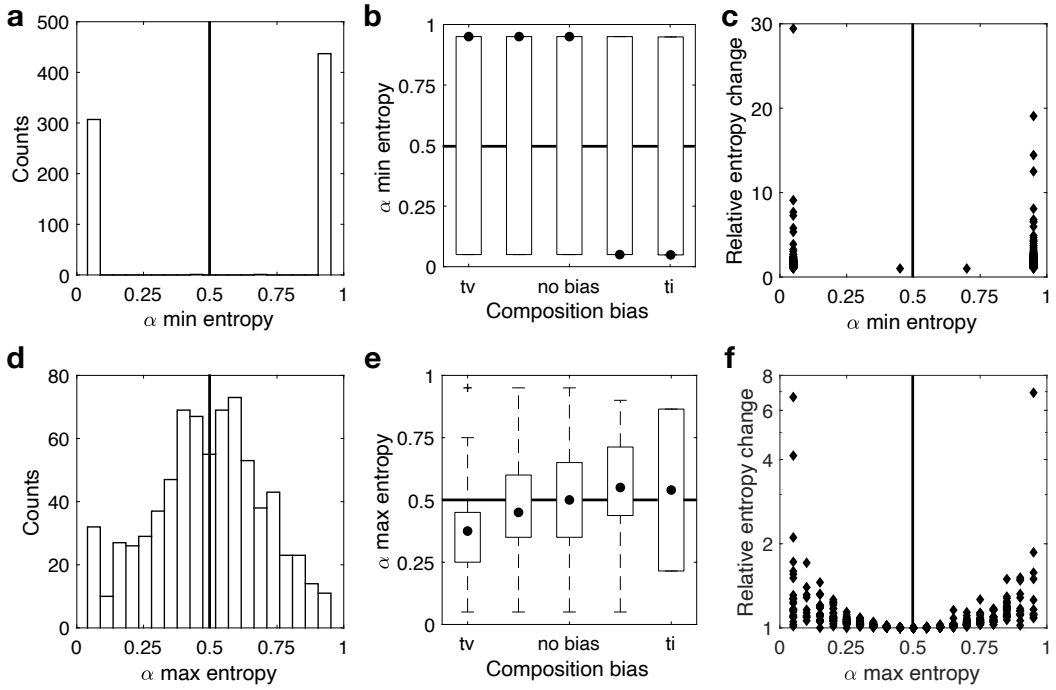

FIGURE 2.4: **Mutation bias interacts with composition bias to impact the predictability of evolution.** **(a)** Distribution of the mutation bias parameter α that minimizes path entropy for each landscape. **(b)** Mutation bias parameter that minimizes path entropy, shown in relation to composition bias. **(c)** Relative entropy change ($\text{entropy}_{\text{no bias}}/\text{entropy}_{\text{min}}$), shown in relation to the mutation bias parameter α that minimizes path entropy. **(d)** Distribution of the mutation bias parameter α that maximizes path entropy for each landscape. **(e)** Mutation bias parameter that maximizes path entropy, shown in relation to composition bias. **(f)** Relative entropy change ($\text{entropy}_{\text{max}}/\text{entropy}_{\text{no bias}}$), shown in relation to the mutation bias parameter α that maximizes path entropy. Data in all panels pertain to all 746 landscapes.

in this "pick-the-winner" regime (Fig. S2.14), as expected on theoretical grounds [25]. However, we reasoned that mutation bias may still influence other properties of an evolving population, specifically those that depend upon the population's distribution in the landscape. To explore this possibility, we used a measure called an overlap coefficient, which quantifies the similarity of two populations as the proportion of individuals that are common to both (Methods). This coefficient takes on its minimum value of 0 when there are no individuals in common between two populations; it takes on its maximum value of 1 when both populations are identical, having the same individuals in the same proportions. We applied this measure to pairs of populations after they had evolved for 1000 generations, reaching steady state (Fig. S2.15). As a baseline for comparison, we first calculated the overlap coefficient for pairs of replicate populations. That is, pairs of populations with identical initial conditions, but with different random number generator seeds (Fig. 2.5a). This allowed us to assess how different we expect two evolved populations to be at steady state, due solely to the stochastic nature of the evolutionary simulations. For replicate populations, the overlap coefficient ranged from 0.912 to 1, with a median of 0.976 and a 75th percentile of 0.817 (Fig. 2.5b). This indicates that while the stochastic nature of the evolutionary simulation can cause large changes in a polymorphic population's distribution in a landscape, it usually does not. Replicate populations tend to converge on highly similar distributions. In contrast, when we calculated the overlap coefficient for pairs of evolved populations with identical initial conditions (including identical random number generator seeds), but different values of the mutation bias parameter, we observed far less overlap (two-sample Kolmogorov-Smirnov test, $D = 0.2695$, $p < 10^{-31}$). Specifically, the overlap coefficient ranged from 0.692 to 1, with a median of 0.908 and a 75th percentile of 0.274 (Fig. 2.5b). This indicates that mutation bias often has a strong influence on an evolved population's distribution in a genotype-phenotype landscape. The strength of this influence depends upon how different the mutation bias values are in the populations being compared (Fig. 2.5c), with larger differences corresponding to larger changes in population distribution, a trend that also holds for infinitely large populations (Fig. S2.16).

Mutation bias and composition bias interact to influence the evolution of genetic diversity and mutational robustness in polymorphic populations

We next asked how mutation bias interacts with composition bias to influence the evolution of genetic diversity. We reasoned that when mutation bias is aligned with composition bias, evolving populations will be less constrained in their exploration of the landscape and will therefore accumulate greater genetic diversity. To explore this possibility, we measured the genetic diversity of populations at steady state using Shannon's diversity index (Methods). This measure takes on its maximum value of 1 when the population comprises all possible individuals in equal proportions. For DNA sequences, this means that all four bases are equally likely to appear at all positions in the sequence. The measure takes on its minimum value of 0 when the population comprises N copies of just a single individual. Fig. 2.6a-e shows that mutation bias can either increase or decrease genetic diversity, relative to an unbiased mutation supply, and depending on whether mutation bias aligns with composition bias. This effect was most pronounced in landscapes with strong composition bias, either toward transversions (Fig. 2.6a) or transitions (Fig. 2.6e), and when the mutation supply was high. For example, in landscapes

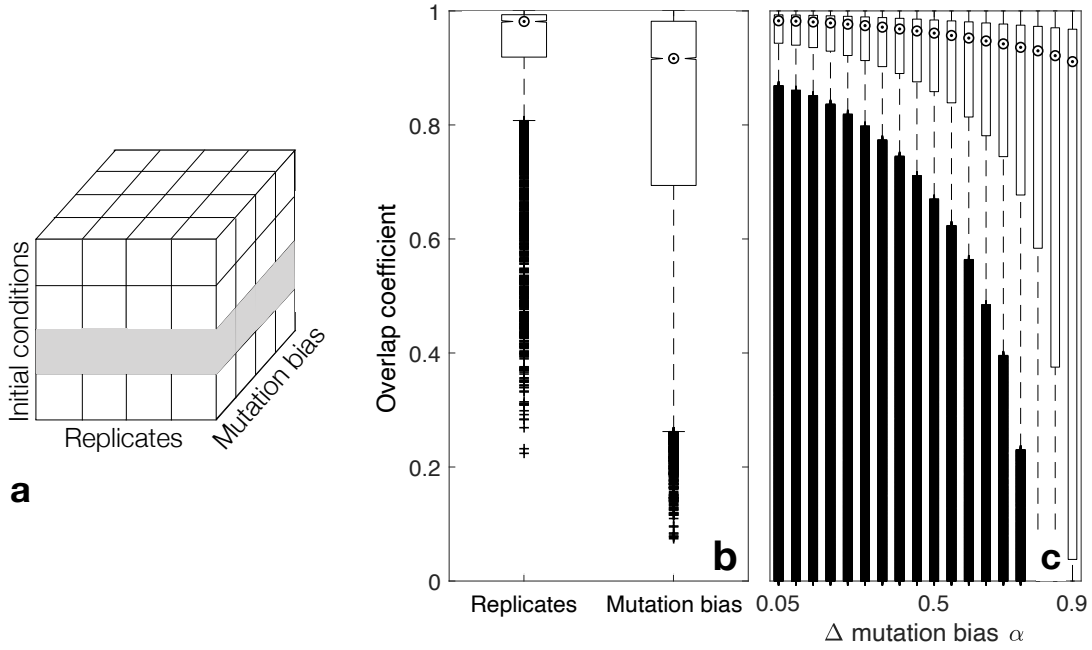

FIGURE 2.5: **Evolving polymorphic populations are more sensitive to changes in mutation bias than to the stochastic nature of the evolutionary simulations. (a)** Schematic figure of our experimental design. For each landscape and combination of population size and mutation rate ($N\mu$), we considered 10 replicates for each of 10 different initial conditions and 19 values of the mutation bias parameter α. Importantly, we used the replicate number to seed the random number generator of each evolutionary simulation, facilitating the comparison of variation across replicates versus across the mutation bias parameter α. For example, the matrix elements indicated in gray contain the information necessary to compare the effects of the mutation bias parameter with the stochasticity of the evolutionary simulations, for one initial condition. **(b)** Overlap coefficient for pairs of evolved populations that differ in random number generator seed ("Replicates") or in mutation bias parameter ("Mutation bias"). Notches indicate medians, whiskers indicate the 25th and 75th percentiles, and cross symbols indicate outliers. **(c)** Overlap coefficient for pairs of evolved populations, shown in relation to the difference in their mutation bias parameters.

with a strong composition bias toward transversions (Fig. 2.6a), a bias in mutation supply could change genetic diversity 1.9-fold when $N\mu = 50$, but had almost no effect when $N\mu = 5$. In these cases, genetic diversity could reach levels higher than those observed on landscapes with little to no bias (Fig. 2.6c). Calculating genetic diversity using nucleotide diversity π results in similar patterns (Fig. S2.17). Specifically, diversity increases when mutation bias aligns with composition bias and decreases otherwise. To further illustrate how mutation bias and composition bias interact to influence population diversity, we show in Fig. S2.18 the allele frequency spectra of populations evolved on two different landscapes, which we chose as illustrative examples because of their strong composition bias toward transversions (Fig. S2.18a) and toward transitions (Fig. S2.18b). As expected, the allele frequency spectra reflect the sequence bias of the mutation process, with more transition polymorphisms present when mutation is biased toward transitions and more transversion polymorphisms present when mutation is biased toward transversions. These examples also illustrate two different ways in which the interaction between mutation bias and composition bias influence the frequency of mutations in the population. In Fig. S2.18a, polymorphisms are present, but they are rare, and the sequence that evolves to the highest frequency is the same as in the absence of mutation bias. In contrast, in Fig. S2.18b some transition polymorphisms come to dominate the population, such that the sequence that evolves to the highest frequency is not the same as in the absence of mutation bias.

We then explored the potential consequences of these changes in genetic diversity. First, we characterized how mutation bias and composition bias interact to influence the mutational robustness of binding sites in evolved populations at steady state. We quantified the mutational robustness of an individual binding site as the fraction of all possible mutations to that site that created another site that was also part of the landscape [35], [37]. The mutational robustness of a population of binding sites was then simply calculated as the average mutational robustness of its constituent sites. Fig. 2.6f-j shows that mutation bias can increase or decrease the mutational robustness of an evolved population, relative to an unbiased mutation supply, especially in landscapes with strong composition bias, either toward transversions (Fig. 2.6f) or transitions (Fig. 2.6j). In contrast, for landscapes with little to no composition bias (Fig. 2.6h), only the most extreme values of mutation bias influenced mutational robustness. For landscapes with strong composition bias, the changes in mutational robustness caused by mutation bias mirrored the changes observed for genetic diversity (compare Fig. 2.6a and e to Fig. 2.6f and j). In these landscapes, therefore, a decrease in genetic diversity was associated with a decrease in mutational robustness. Moreover, we observed that populations evolving on landscapes with composition bias tended to be less mutationally robust at steady state. The reason is that landscapes with composition bias tended to comprise genotypes with fewer mutational neighbors, relative to landscapes without composition bias (Fig. S2.19). When mutation bias aligned with composition bias, those genotypes with more mutational neighbors were more likely to evolve, because the mutation bias oriented the population toward those genotypes.

Mutation bias influences the evolvability of polymorphic populations

Because mutational robustness is a cause of evolvability [21], [52], [53], we explored whether mutation bias and composition bias interact to influence evolvability — defined in this context

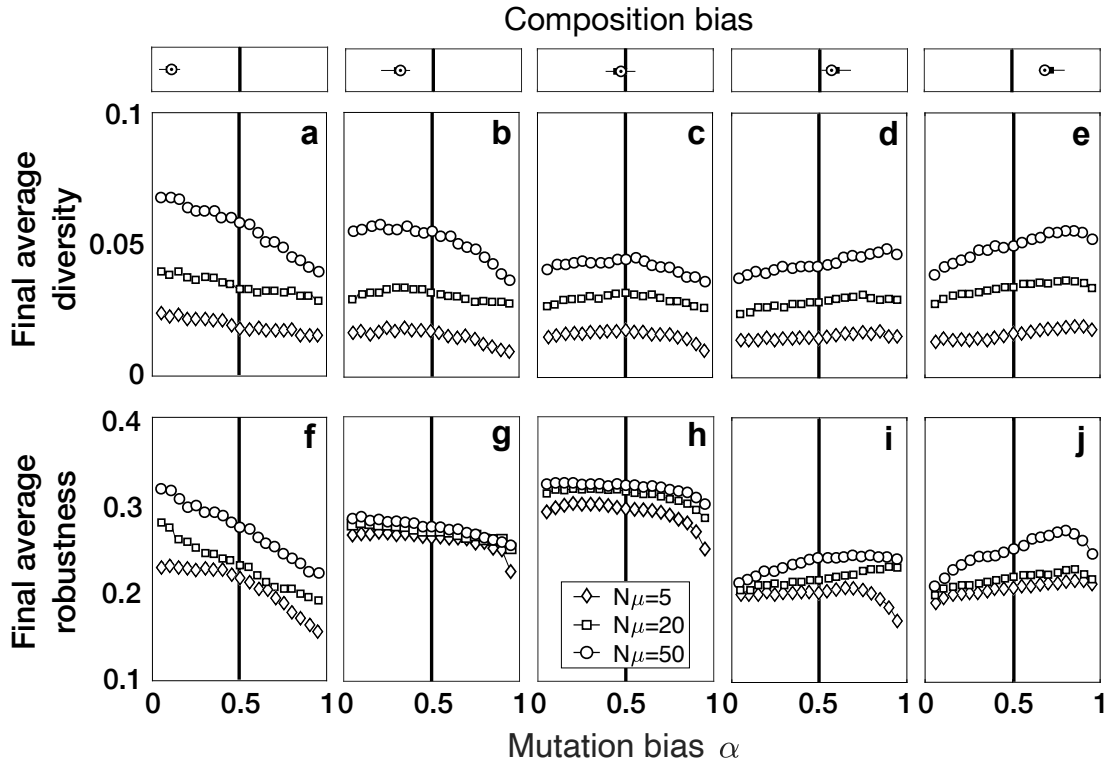

FIGURE 2.6: **Mutation bias interacts with composition bias to influence the evolution of genetic diversity and mutational robustness. (a-e)** The average genetic diversity and **(f-j)** mutational robustness of evolved populations at steady state is shown for 19 different values of the mutation bias parameter α, and for each of three different values of mutation supply $N\mu$ (see legend). The solid vertical lines indicate no bias in mutation supply ($\alpha = 0.5$). Landscapes are grouped based on their composition bias and the distribution of composition bias per panel is shown on top of each panel as in Fig. 2.3.

as the ability of mutation to bring forth new binding phenotypes (Methods) [35]. To test this, we calculated the average number of transcription factors that bind the one-mutant neighbors of any individual in the population at steady state, and computed the difference between these averages for populations with a strong mutation bias toward transversions ($\alpha = 0.05$), relative to an unbiased mutation supply ($\alpha = 0.5$), as well as toward transitions ($\alpha = 0.95$), relative to an unbiased mutation supply. Fig. S2.20 shows how a bias in mutation supply affected the evolvability of polymorphic populations for 128 transcription factors from *Arabidopsis thaliana* and 128 transcription factors from *Mus musculus*, the two species with the most transcription factors in our dataset. In approximately 70% of the landscapes, mutation bias had no influence on evolvability. However, in the remaining 30% of landscapes, mutation bias could increase or decrease the number of transcription factors accessible via point mutation. In *A. thaliana*, this ranged from plus or minus 6 transcription factors ($\sim 5\%$), whereas in *M. musculus* this ranged from plus 12 to minus 7 transcription factors (between $\sim 5\%$ and $\sim 9\%$). These observations suggest that mutation bias is capable of orienting evolving populations both toward and away from more evolvable regions of genotype-phenotype landscapes. However, this effect is apparently independent of composition bias. This may seem counterintuitive at first glance, because the interaction between mutation bias and composition bias influences both genetic diversity and mutational robustness, two properties that facilitate evolvability [53]. However, the global peaks of the landscapes considered here are sufficiently narrow that any increase of diversity or robustness within them is not sufficient to cause a change in the number of phenotypes accessible via mutation. Said differently, the fraction of sequence space covered by these peaks is too small to facilitate mutational access to the landscapes of other transcription factors. We speculate that in landscapes with broader, mesa-like peaks, mutation bias and composition bias could interact to influence evolvability.

2.4 DISCUSSION

The mapping of genotype to phenotype influences how mutation brings forth phenotypic variation [54]. Biases in this map can therefore influence evolution [55]. For example, so-called phenotypic bias – that most phenotypes are realized by few genotypes, but a few phenotypes are realized by many genotypes – can cause a phenomenon known as the "arrival of the frequent", wherein phenotypes evolve not because they are the most fit, but rather because they are the most common [56], [57]. Moreover, phenotypes can exhibit a bias in their mutational connectivity to other phenotypes, such that phenotype-changing mutations are more likely to lead to common phenotypes than to uncommon phenotypes, thus reducing the ability of mutation to bring forth phenotypic variation [58].

Here, we used empirical genotype-phenotype landscapes of transcription factor binding affinities to study a different form of bias, namely composition bias. This describes the prevalence of a particular kind of mutation in a landscape, or in a subset of mutational paths in a landscape, relative to a null expectation. The kind of mutation we considered was a transition mutation, and we measured its prevalence relative to the null expectation that one transition occurs for every two transversions. We found that composition bias is common among accessible mutational paths to a landscape's global adaptive peak, and that for a large diversity of transcription factor families, this

bias is toward transversions. We showed that such composition bias can interact with mutation bias to influence the navigability of genotype-phenotype landscapes and the predictability of evolution, as well as the evolution of genetic diversity and mutational robustness.

Such interaction was most pronounced in landscapes with strong composition bias, and when the bias in mutation supply was either aligned or in opposition with composition bias. Estimates of base-substitution mutation spectra are available for at least five of the species studied here [59]: *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae*. Transforming the reported transition:transversion ratios to our mutation bias parameter α results in values that range from roughly the null expectation of one transition per two transversions ($\alpha = 0.47$ for *C. elegans*) to a strong bias toward transitions (0.83 in *Arabidopsis thaliana* — a 4.8-fold increase over the null expectation). Our results therefore suggest that biologically-realistic values of mutation bias can influence the evolution of transcription factor binding sites, specifically for transcription factors whose genotype-phenotype landscapes exhibit composition bias, either toward transitions or transversions.

Ideally, we could detect the influence of such an interaction on the evolution of transcription factor binding sites *in vivo*. Previous work used single nucleotide polymorphism and functional genomics data to show how the topology and topography of a genotype-phenotype landscape influences the evolution of binding site diversity in *A. thaliana* and *S. cerevisiae* [22], [36]. For example, transcription factors with broad global peaks were found to exhibit more diversity in their high-affinity binding sites than transcription factors with narrow global peaks [22]. It would be desirable to use such data to study how mutation bias and composition bias interact to influence the evolution of transcription factor binding sites *in vivo*. For example, one could ask whether the binding sites of transcription factors with a strong composition bias toward transversions exhibit less diversity in genomic regions prone to transition mutations than in other genomic regions, as our results suggest they would. However, such an analysis is complicated because one would need to identify functional binding sites for the same transcription factor in genomic regions that differ in mutation bias, but not in mutation rate. While some regulatory regions are more prone to transition mutations than others, such as CpG-rich promoters, which are susceptible to C>T transitions due to the spontaneous deamination of 5-methylcytosines, these regions also exhibit elevated mutation rates [60].

It may be possible to overcome this challenge with experiments. For example, a single low-affinity binding site for an activating transcription factor could be used to seed two separate populations of binding sites, where the two populations differ in their mutation bias. This could be achieved by introducing mutations with error-prone PCR, using enzymes that differ in their mutation spectra, but have similar mutation rates. The mutated binding sites could then be cloned into plasmids upstream of a reporter gene, such as yellow fluorescence protein, transformed into bacterial cells, and exposed to selection for increased fluorescence using flow-activated cell sorting. This process of mutation and selection could then be repeated for several rounds, cloning the mutated binding sites back into the ancestral plasmid backbone and transforming the plasmids into fresh bacterial cells before each round of selection to ensure that increased fluorescence is driven by mutations in the binding site, rather than by mutations in the protein or elsewhere in the bacterial genome. By comparing replicate experiments for different binding site seeds and for transcription factors that vary in their composition bias, it may be possible to determine if and

how mutation bias and composition bias interact to influence the *in vivo* evolution of increased affinity in transcription factor binding sites.

Our analysis makes two key assumptions, the caveats of which are worth highlighting. The first is our assumption of selection for increased binding affinity. Low-affinity binding sites are also commonly employed to regulate gene expression, particularly for the auto-regulation of high-copy number transcription factors in bacteria [61] and during the development of multicellular organisms [44], [45], [62]. The second assumption is that of a linear relationship between the selective advantage conferred by a mutation and the change in binding affinity that the mutation causes. In reality, this relationship is likely non-linear, site-specific, and dependent upon local transcription factor concentrations. Relaxing these two assumptions will transform the topographies of the landscapes studied here, and will alter the composition biases of their accessible mutational paths. Such transformations are therefore likely to affect landscape navigability. However, we do not anticipate they will affect the way in which mutation bias and composition bias interact to influence landscape navigability. Regardless of the particular topographical properties of the landscape under investigation, navigability will be enhanced when mutation bias and composition bias are aligned, and diminished otherwise.

While we focused our study on transcription factor-DNA interactions, composition bias is likely to exist in other genotype-phenotype landscapes as well. For example, many RNA binding proteins target sequences enriched for guanine and uracil [63], and their binding affinity landscapes will therefore exhibit a composition bias toward transversions. Additionally, composition bias is not limited to genotype-phenotype landscapes of intermolecular interactions as studied here, but it may also be present in the landscapes of some macromolecules. Finally, other forms of composition bias and mutation bias may interact to influence adaptive evolution, including GC:AT bias and deletion bias. The mutation bias signatures of various cancers [64], which influence the *de novo* evolution of transcription factor binding sites [65], may also interact with composition bias to influence landscape navigability. As the scope and scale of genome editing and deep mutational scanning studies continues to expand, we will gain a better understanding of the prevalence of composition bias in empirical genotype-phenotype landscapes and its potential to interact with mutation bias in shaping adaptive mutational trajectories.

2.5 METHODS

Data

We constructed genotype-phenotype landscapes using data from protein binding microarrays, which we downloaded from the UniPROBE [46] and CIS-BP [47] databases. These data include proxies for the relative binding affinity of a transcription factor to all possible $(4^8 - 4^4)/2 + 4^4 = 32,896$ eight-nucleotide, double-stranded DNA sequences (transcription factor binding sites). These proxies include an enrichment score (E-score), which for each sequence is a function of the fluorescence intensities of a subset of probes that contain the sequence and a subset of probes that do not contain that sequence [66], and a Z-score, which for each sequence is the difference between the logarithm of the median fluorescence intensities of a subset of probes that contain the sequence and the logarithm of the median fluorescence intensities of all probes, reported in

units of standard deviation. We considered a sequence as specifically bound by a transcription factor if its *E*-score exceeded 0.35, following previous work [22], [35], [37], [67]. We included a landscape in our dataset if its dominant genotype network comprised at least 100 bound sequences. According to these criteria, our dataset included 746 transcription factors, representing 129 eukaryotic species, and 48 DNA-binding domain structural classes. Details are provided in Table S1.

Constructing and analyzing genotype-phenotype landscapes

For each transcription factor, we used the Genonets Server [68] to construct a genotype-phenotype landscape from the set of sequences that specifically bound the factor (i.e., with an E-score > 0.35). We represented each such sequence as a node in a genotype network, and connected nodes with edges if their corresponding sequences differed in a single point mutation. Note that we did not consider indels in our definition of a mutation, like we did in our previous work [22], [35]–[37]. The reason is that we were interested in understanding the influence of a form of composition bias defined by point mutations (transitions vs. transversions), and we were concerned that the inclusion of indel-based edges would confound our analyses. For genotype networks that were fragmented into multiple components, we only considered the largest component, which we call the dominant genotype network.

Each dominant genotype network served as the substrate of a genotype-phenotype landscape, the surface of which was defined by relative binding affinity. For our main analyses, this was captured by the *E*-score, whereas for some of our sensitivity analyses, this was captured by the *Z*-score. We studied accessible mutational paths to the global peaks of these landscapes. An accessible mutational path comprises edges that each confer an increase in binding affinity greater than the noise threshold parameter δ [22]. For each transcription factor, we calculated δ as the residual standard error of a linear regression between the affinity values of all bound sequences from the two replicate protein binding microarrays [22]. Thus, each transcription factor had its own δ, which reflects the noise in the replicated measurements for that particular transcription factor.

Mutation bias and composition bias

We report both mutation bias and composition bias relative to the null expectation that one transition occurs for every two transversions. Letting Ti and Tv represent mutation rates of transitions and transversions, respectively, we define mutation bias as

$$\alpha = \frac{Ti}{Ti + Tv/2}. \tag{2.1}$$

A mutation bias of $\alpha = 0.5$ corresponds to the null expectation of one transition per two transversions. Values below 0.5 mean there are more transversions than expected under the null, whereas values above 0.5 mean there are more transitions than expected under the null.

Composition bias was measured in the same way, except with Ti and Tv representing the number of transitions and transversions in a landscape, or in an accessible mutational path to the global peak of a landscape.

Origin-fixation model of evolutionary dynamics

We used an origin-fixation model to study evolutionary dynamics when the mutation supply is low ($N\mu \ll 1$). This was implemented using Markov chains, a memoryless process that gives the jumping probability from one genotype i to another genotype j using the matrix

$$P_{i,j} = \frac{\phi_{i,j} f_{i,j}}{\sum_{\forall k} \phi_{i,k} f_{i,k}} \tag{2.2}$$

where $f_{i,j}$ is the relative difference in binding affinity b

$$f_{i,j} = \begin{cases} b_j / b_i - 1 & \text{if } b_j > b_i \\ 0 & \text{if otherwise.} \end{cases} \tag{2.3}$$

and

$$\phi_{i,j} = \begin{cases} \alpha & \text{if edge } (i,j) \text{ is a transition} \\ (1-\alpha) & \text{if edge } (i,j) \text{ is a transversion.} \end{cases} \tag{2.4}$$

Then in general, the probability of going from any state to another state in a Markov chain given by the matrix P (Eq (2.2)) after t steps is

$$(P^t)_{i,j}. \tag{2.5}$$

Wright-Fisher model of evolutionary dynamics

We carried out simulations of a Wright-Fisher model to study evolutionary dynamics when the mutation supply is high ($N\mu > 1$). Each simulation was initialized with a monomorphic population comprising N copies of the same sequence, chosen from the bottom 10% of binding affinity values in the landscape. In each generation t, N sequences were chosen from the population at generation $t-1$ with replacement and with a probability that was linearly proportional to binding affinity. Mutations were introduced to these sequences at a rate μ per sequence per generation with a mutation bias α. For each of the 746 landscapes, we performed 15 replicate simulations for each of initial conditions, 19 linearly spaced mutation bias values between 0.05 and 0.95, and 3 mutation supply values ($N\mu \in \{5, 20, 50\}$). Each simulation ran for 1000 generations, which was sufficient to ensure that the population had reached steady state.

Landscape navigability

As a measure of landscape navigability, we calculated the probability of reaching the global peak, starting from the 10% of sequences in the dominant genotype network with the lowest binding affinity. For low mutation supply, this was calculated as the average probability of going from the initial sequences to the global peak using Eq(2.5) after $k = 1000$ steps. For high mutation supply, this was calculated as the fraction of simulations per landscape in which at least 50% of the population reached the global peak.

Path entropy

PathMAN (Path Matrix Algorithm for Networks), is a publicly available Python script that efficiently calculates path statistics of a given Markov process [51]. We employed PathMAN to calculate the Shannon's entropy of the path distribution, which accounts for the predictability of the process. Low entropy means that few paths with large probability dominate the process, while large entropy means that several low-probability paths contribute. We calculated the path entropy for 19 different values of the mutation bias parameter within the range [0.5,0.95], in order to find the mutation bias parameters $\alpha_{\min}$ and $\alpha_{\max}$ that minimize and maximize the path entropy for each landscape.

Overlap coefficient

The overlap coefficient between two different polymorphic populations A and B was calculated as

$$O_{A,B} = \frac{|C|}{\min(|A|, |B|)}, \tag{2.6}$$

where A and B are multisets — sets that permit multiple instances of an element. The cardinality of such multisets is defined as

$$|A| = \sum_{x \in A} m_a(x) \quad \text{and} \quad |B| = \sum_{x \in B} m_B(x), \tag{2.7}$$

where the number of occurrences of the element x in the multiset is indicated by the multiplicity function $m(x)$.

Then C is the multiset defined as $C = A \cap B$, with multiplicity function

$$m_C(x) = \min(m_A(x), m_B(x)) \quad \forall x \in A \cup B. \tag{2.8}$$

For example, if $A = \{1, 1, 2, 2, 2, 3\}$ and $B = \{1, 2, 2, 4\}$, then $C = \{1, 2, 2\}$ and the overlap coefficient is $O_{A,B} = 0.75$.

Quasispecies dynamics of infinite populations

Since the matrix in Eq(2.2) is non-negative and connected, Perron-Frobenius theorem for non-negative matrices applies [69]. Hence, the steady-state distribution of an infinite size population on a genotype network is determined by the eigenvector associated to the largest eigenvalue of the matrix in Eq(2.2). Per each landscape, the eigenvectors were computed numerically for 19 different values of mutation bias parameter α within the range [0.5,0.95].

Genetic diversity

We measured the diversity of a population as the average Shannon's diversity over all genotypes, normalized by the maximum diversity per landscape $\log_2(n)$:

$$H = \frac{-\sum_i p_i \log_2 p_i}{\log_2(n)} \tag{2.9}$$

where n is the number of sequences in the landscape, p_i is the fraction of the steady-state population that is at sequence i.

Evolvability

We quantified the evolvability of a transcription factor's binding sites as follows. First, we determined the set of binding sites that have evolved at steady state in our Wright-Fisher simulations. Then we enumerated the set of DNA sequences that differ by one mutation from any of these binding sites, but are not themselves part of the focal transcription factor's landscape. Finally, we determined the fraction of transcription factors in our dataset that these one-mutant neighbors bind. This fraction was our measure of evolvability.

2.6 ACKNOWLEDGEMENTS

This research was supported by Swiss National Science Foundation Grant PP00P3_170604.

REFERENCES

[1] L. Y. Yampolsky and A. Stoltzfus, "Mutational biases", *eLS*, 2008.

[2] R. E. Hudson, U. Bergthorsson, and H. Ochman, "Transcription increases multiple spontaneous point mutations in Salmonella enterica", *Nucleic Acids Research*, vol. 31, no. 15, pp. 4517–4522, 2003.

[3] M. D. Pauly, M. C. Procario, and A. S. Lauring, "A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses", *eLife*, vol. 6, 2017.

[4] P. D. Keightley, U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. L. Blaxter, "Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines", *Genome Research*, vol. 19, no. 7, pp. 1195–1201, 2009.

[5] S. Ossowski, K. Schneeberger, J. I. Lucas-Lledó, *et al.*, "The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*", *Science*, vol. 327, no. 5961, pp. 92–94, 2010.

[6] S. Kucukyildirim, H. Long, W. Sung, S. F. Miller, T. G. Doak, and M. Lynch, "The rate and spectrum of spontaneous mutations in Mycobacterium smegmatis, a bacterium naturally devoid of the postreplicative mismatch repair pathway", *G3: Genes, Genomes, Genetics*, vol. 6, no. 7, pp. 2157–2163, 2016.

[7] M. P. Francino, L. Chao, M. A. Riley, and H. Ochman, "Asymmetries generated by transcription-coupled repair in enterobacterial genes", *Science*, vol. 272, no. 5258, pp. 107–109, 1996.

[8] T. Gojobori, W. H. Li, and D. Graur, "Patterns of nucleotide substitution in pseudogenes and functional genes", *Journal of Molecular Evolution*, vol. 18, no. 5, pp. 360–369, 1982.

[9] D. A. Petrov and D. L. Hartl, "Patterns of nucleotide substitution in *Drosophila* and mammalian genomes", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 4, pp. 1475–1479, 1999.

[10] Z. Zhang and M. Gerstein, "Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes", *Nucleic Acids Research*, vol. 31, no. 18, pp. 5338–5348, 2003.

[11] R. Hershberg and D. A. Petrov, "Evidence that mutation is universally biased towards AT in bacteria", *PLoS Genetics*, vol. 6, no. 9, 2010.

[12] A. Stoltzfus and L. Y. Yampolsky, "Climbing mount probable: Mutation as a cause of nonrandomness in evolution", *Journal of Heredity*, vol. 100, no. 5, pp. 637–647, 2009.

[13] E. R. Lozovsky, T. Chookajorn, K. M. Brown, *et al.*, "Stepwise acquisition of pyrimethamine resistance in the malaria parasite", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 29, pp. 12 025–12 030, 2009.

[14] D. R. Rokyta, P. Joyce, S. B. Caudle, and H. A. Wichman, "An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus", *Nature Genetics*, vol. 37, no. 4, pp. 441–444, 2005.

[15] J. L. Payne, F. Menardo, A. Trauner, *et al.*, "Transition bias influences the evolution of antibiotic resistance in Mycobacterium tuberculosis", *PLoS Biology*, vol. 17, no. 5, 2019.

[16] J. F. Storz, C. Natarajan, A. V. Signore, C. C. Witt, D. M. McCandlish, and A. Stoltzfus, "The role of mutation bias in adaptive molecular evolution: Insights from convergent changes in protein function", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, no. 1777, 2019.

[17] S. Wright, "The roles of mutation, inbreeding, crossbreeding and selection in evolution.", *Sixth International Congress on Genetics*, vol. 1, no. 6, pp. 356–366, 1932.

[18] J. M. Smith, "Natural selection and the concept of a protein space.", *Nature*, vol. 225, no. 5232, pp. 563–4, 1970.

[19] J. A. G. M. De Visser and J. Krug, *Empirical fitness landscapes and the predictability of evolution*, 2014.

[20] I. Fragata, A. Blanckaert, M. A. Dias Louro, D. A. Liberles, and C. Bank, *Evolution in the light of fitness landscape theory*, 2019.

[21] J. L. Payne and A. Wagner, "The causes of evolvability and their evolution", *Nature Reviews Genetics*, vol. 20, no. 1, pp. 24–38, 2019.

[22] J. Aguilar-Rodríguez, J. L. Payne, and A. Wagner, "A thousand empirical adaptive landscapes and their navigability", *Nature Ecology and Evolution*, vol. 1, no. 2, 2017.

[23] S. Kauffman and S. Levin, "Towards a general theory of adaptive walks on rugged landscapes", *Journal of Theoretical Biology*, vol. 128, no. 1, pp. 11–45, 1987.

[24] J. Domingo, P. Baeza-Centurion, and B. Lehner, "The Causes and Consequences of Genetic Interactions (Epistasis)", *Annual Review of Genomics and Human Genetics*, vol. 20, no. 1, pp. 433–460, 2019.

[25] L. Y. Yampolsky and A. Stoltzfus, "Bias in the introduction of variation as an orienting factor in evolution", *Evolution and Development*, vol. 3, no. 2, pp. 73–83, 2001.

[26] A. Stoltzfus, "Mutation-biased adaptation in a protein NK model", *Molecular Biology and Evolution*, vol. 23, no. 10, pp. 1852–1862, 2006.

[27] J. B. Kinney and D. M. McCandlish, "Massively Parallel Assays and Quantitative Sequence–Function Relationships", *Annual Review of Genomics and Human Genetics*, vol. 20, no. 1, pp. 99–127, 2019.

[28] P. Julien, B. Miñana, P. Baeza-Centurion, J. Valcárcel, and B. Lehner, "The complete local genotype-phenotype landscape for the alternative splicing of a human exon", *Nature Communications*, vol. 7, 2016.

[29] D. W. Anderson, A. N. McKeown, and J. W. Thornton, "Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites", *eLife*, vol. 4, 2015.

[30] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, *et al.*, "Local fitness landscape of the green fluorescent protein", *Nature*, vol. 533, pp. 397–401, 2016.

[31] C. Qiu, O. C. Erinne, J. M. Dave, *et al.*, "High-Resolution Phenotypic Landscape of the RNA Polymerase II Trigger Loop", *PLoS Genetics*, vol. 12, no. 11, 2016.

[32] G. Diss and B. Lehner, "The genetic landscape of a physical interaction", *eLife*, vol. 7, 2018.

[33] Y. Schaerli, A. Jiménez, J. M. Duarte, *et al.*, "Synthetic circuits reveal how mechanisms of gene regulatory networks constrain evolution", *Molecular Systems Biology*, vol. 14, no. 9, 2018.

[34] M. C. Bassalo, A. D. Garst, A. Choudhury, *et al.*, " Deep scanning lysine metabolism in Escherichia coli ", *Molecular Systems Biology*, vol. 14, no. 11, 2018.

[35] J. L. Payne and A. Wagner, "The robustness and evolvability of transcription factor binding sites", *Science*, vol. 343, no. 6173, pp. 875–877, 2014.

[36] J. Aguilar-Rodríguez, L. Peel, M. Stella, A. Wagner, and J. L. Payne, "The architecture of an empirical genotype-phenotype map", *Evolution*, vol. 72, no. 6, pp. 1242–1260, 2018.

[37] J. L. Payne, F. Khalid, and A. Wagner, "RNA-mediated gene regulation is less evolvable than transcriptional regulation", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 15, E3481–E3490, 2018.

[38] F. Spitz and E. E. Furlong, "Transcription factors: From enhancer binding to developmental control", *Nature Reviews Genetics*, vol. 13, no. 9, pp. 613–626, 2012.

[39] A. J. Stewart and J. B. Plotkin, "Why transcription factor binding sites are ten nucleotides long", *Genetics*, vol. 192, no. 3, pp. 973–985, 2012.

[40] M. T. Maurano, R. Humbert, E. Rynes, *et al.*, "Systematic localization of common disease-associated variation in regulatory DNA", *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[41] E. Khurana, Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin, and M. Gerstein, "Role of non-coding sequence variants in cancer", *Nature Reviews Genetics*, vol. 17, no. 2, pp. 93–108, 2016.

[42] G. A. Wray, "The evolutionary significance of cis-regulatory mutations", *Nature Reviews Genetics*, vol. 8, no. 3, pp. 206–216, 2007.

[43] P. J. Wittkopp and G. Kalay, "Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence", *Nature Reviews Genetics*, vol. 13, no. 1, pp. 59–69, 2012.

[44] J. Crocker, N. Abe, L. Rinaldi, *et al.*, "Low affinity binding site clusters confer HOX specificity and regulatory robustness", *Cell*, vol. 160, no. 1-2, pp. 191–203, 2015.

[45] E. K. Farley, K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar, and M. S. Levine, "Suboptimization of developmental enhancers", *Science*, vol. 350, no. 6258, pp. 325–328, 2015.

[46] D. E. Newburger and M. L. Bulyk, "UniPROBE: An online database of protein binding microarray data on protein-DNA interactions", *Nucleic Acids Research*, vol. 37, no. SUPPL. 1, 2009.

[47] M. T. Weirauch, A. Yang, M. Albu, *et al.*, "Determination and inference of eukaryotic transcription factor sequence specificity.", *Cell*, vol. 158, no. 6, pp. 1431–1443, 2014.

[48] C. Guo, I. C. McDowell, M. Nodzenski, *et al.*, "Transversions have larger regulatory effects than transitions", *BMC Genomics*, vol. 18, no. 1, p. 1, 2017.

[49] M. Kircher, C. Xiong, B. Martin, *et al.*, "Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution", *Nature Communications*, vol. 10, no. 1, 2019.

[50] D. M. McCandlish and A. Stoltzfus, "Modeling evolution using the probability of fixation: History and implications", *Quarterly Review of Biology*, vol. 89, no. 3, pp. 225–252, 2014.

[51] M. Manhart, W. Kion-Crosby, and A. V. Morozov, "Path statistics, memory, and coarse-graining of continuous-time random walks on networks", *Journal of Chemical Physics*, vol. 143, no. 21, 2015.

[52] J. A. G. M. de Visser, J. Hermisson, G. P. Wagner, *et al.*, "Perspective:Evolution and Detection of Genetic Robustness", *Evolution*, vol. 57, no. 9, p. 1959, 2003.

[53] A. Wagner, "Robustness and evolvability: A paradox resolved", *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, no. 1630, pp. 91–100, 2008.

[54] P. Alberch, "From genes to phenotype: dynamical systems and evolvability", *Genetica*, vol. 84, no. 1, pp. 5–11, 1991.

[55] S. E. Ahnert, "Structural properties of genotype-phenotype maps", *Journal of the Royal Society Interface*, vol. 14, no. 132, 2017.

[56] S. Schaper and A. A. Louis, "The arrival of the frequent: How bias in genotype-phenotype maps can steer populations to local optima", *PLoS ONE*, vol. 9, no. 2, 2014.

[57] J. A. García-Martín, P. Catalán, S. Manrubia, and J. A. Cuesta, "Statistical theory of phenotype abundance distributions: A test through exact enumeration of genotype spaces", *Epl*, vol. 123, no. 2, 2018.

[58] M. C. Cowperthwaite, E. P. Economo, W. R. Harcombe, E. L. Miller, and L. A. Meyers, "The ascent of the abundant: How mutational networks constrain evolution", *PLoS Computational Biology*, vol. 4, no. 7, 2008.

[59] M. Lynch, "Rate, molecular spectrum, and consequences of human mutation", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 3, pp. 961–968, 2010.

[60] B. Vernot, A. B. Stergachis, M. T. Maurano, *et al.*, "Personal and population genomics of human regulatory variation", *Genome Research*, vol. 22, no. 9, pp. 1689–1697, 2012.

[61] A. Grönlund, P. Lötstedt, and J. Elf, "Transcription factor binding kinetics constrain noise suppression via negative feedback", *Nature Communications*, vol. 4, 2013.

[62] M. Hajheidari, Y. Wang, N. Bhatia, P. Huijser, X. Gan, and M. Tsiantis Correspondence, "Autoregulation of RCO by Low-Affinity Binding Modulates Cytokinin Action and Shapes Leaf Diversity", *Current Biology*, vol. 29, pp. 1–10, 2019.

[63] D. Ray, H. Kazan, K. B. Cook, *et al.*, "A compendium of RNA-binding motifs for decoding gene regulation", *Nature*, vol. 499, no. 7457, pp. 172–177, 2013.

[64] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, *et al.*, "Signatures of mutational processes in human cancer", *Nature*, vol. 500, no. 7463, pp. 415–421, 2013.

[65] C. W. Yiu Chan, Z. Gu, M. Bieg, R. Eils, and C. Herrmann, "Impact of cancer mutational signatures on transcription factor motifs in the human genome", *BMC Medical Genomics*, vol. 12, no. 1, 2019.

[66] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk, "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities", *Nature Biotechnology*, vol. 24, no. 11, pp. 1429–1435, 2006.

[67] G. Badis, M. F. Berger, A. A. Philippakis, *et al.*, "Diversity and complexity in DNA recognition by transcription factors", *Science*, vol. 324, no. 5935, pp. 1720–1723, 2009.

[68] F. Khalid, J. Aguilar-Rodríguez, A. Wagner, and J. L. Payne, "Genonets server-a web server for the construction, analysis and visualization of genotype networks", *Nucleic acids research*, vol. 44, no. W1, W70–W76, 2016.

[69] J. Aguirre, P. Catalán, J. A. Cuesta, and S. Manrubia, "On the networked architecture of genotype spaces and its critical effects on molecular evolution", *Open Biology*, 2018.

2.7 SUPPLEMENTARY MATERIAL

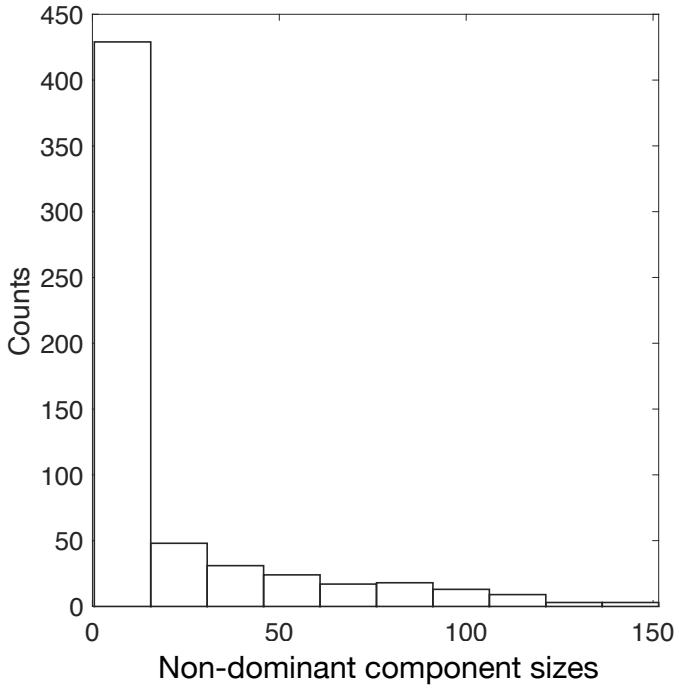

FIGURE S2.1: **Non-dominant genotype network components usually comprise few sequences.** Histogram of non-dominant component sizes. In total, 46% are singletons and 97% are not large enough to satisfy our inclusion criterion of containing 100 sequences.

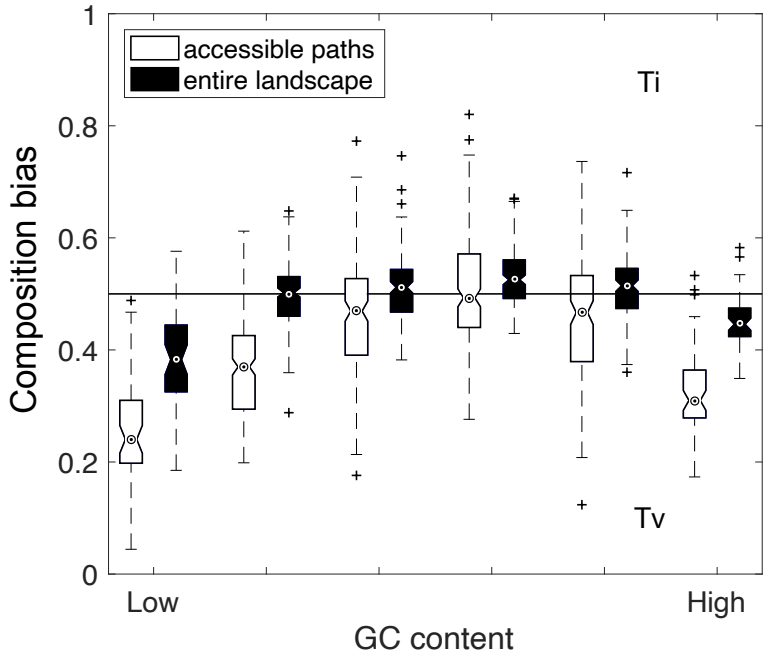

FIGURE S2.2: **Composition bias is most pronounced in landscapes with low or high GC content.** The composition bias in entire landscapes, and in accessible mutational paths connecting the 10% of binding sites with the lowest affinity to the global peak, is shown in relation to the average GC content of the sequences in the landscape. Data pertain to all 746 landscapes. Notches indicate medians, whiskers indicate the 25th and 75th percentiles, and cross symbols indicate outliers. The horizontal line indicates no composition bias (0.5).

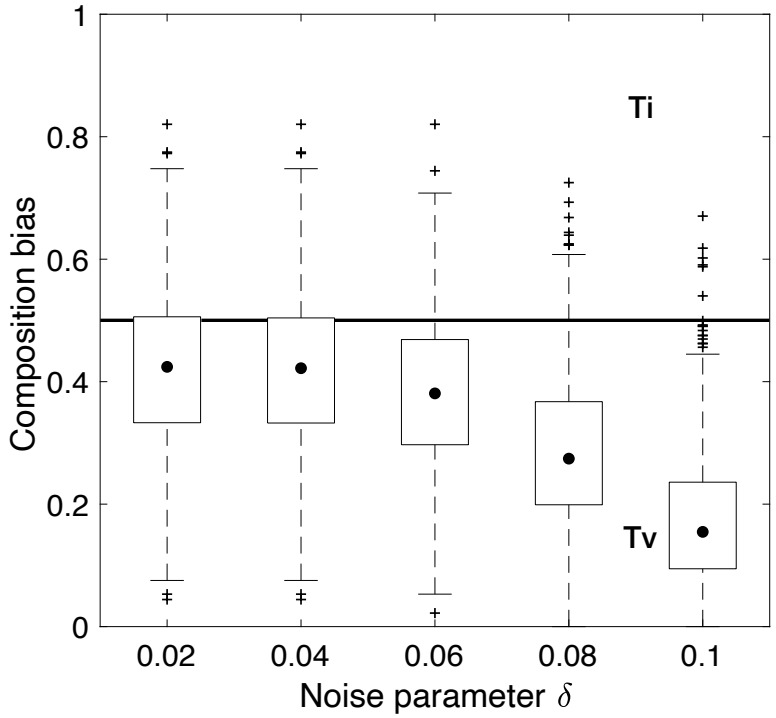

FIGURE S2.3: **Composition bias becomes more pronounced as the noise threshold parameter δ increases.** For each landscape and for five different values of the noise threshold δ, we calculated the composition bias along the accessible mutational paths connecting the 10% of binding sites with the lowest affinity to the global peak. Data pertain to all 746 landscapes. Black dots indicate medians, whiskers indicate the 25th and 75th percentiles, and cross symbols indicate outliers.

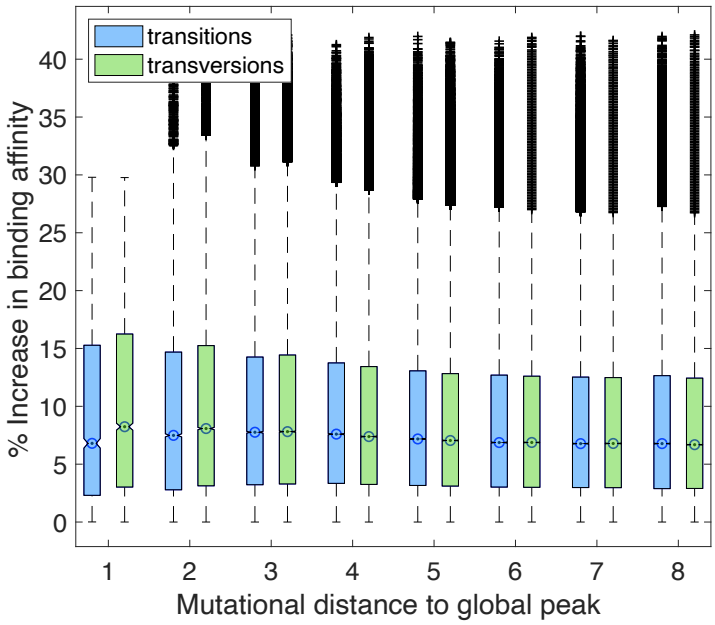

FIGURE S2.4: **Transversions cause larger changes in binding affinity than transitions, but only near the global peak.** The % increase in binding affinity conferred by transition and transversion mutations along accessible mutational paths is shown in relation to the mutational distance of a binding site to the global peak. For each binding site in each accessible path at each mutational distance d, we calculated the increase in affinity as the percentage change at mutational distance $d-1$ along the path, relative to the affinity of the binding site at distance d. Notches indicate medians, whiskers indicate the 25th and 75th percentiles, and cross symbols indicate outliers. Mutational distances 1 and 2 exhibit statistically significant differences in the increase in binding affinity conferred by transitions and transversions (Bonferroni corrected two-sample t test, $q < 10^{-3}$ and $q < 0.05$, respectively).

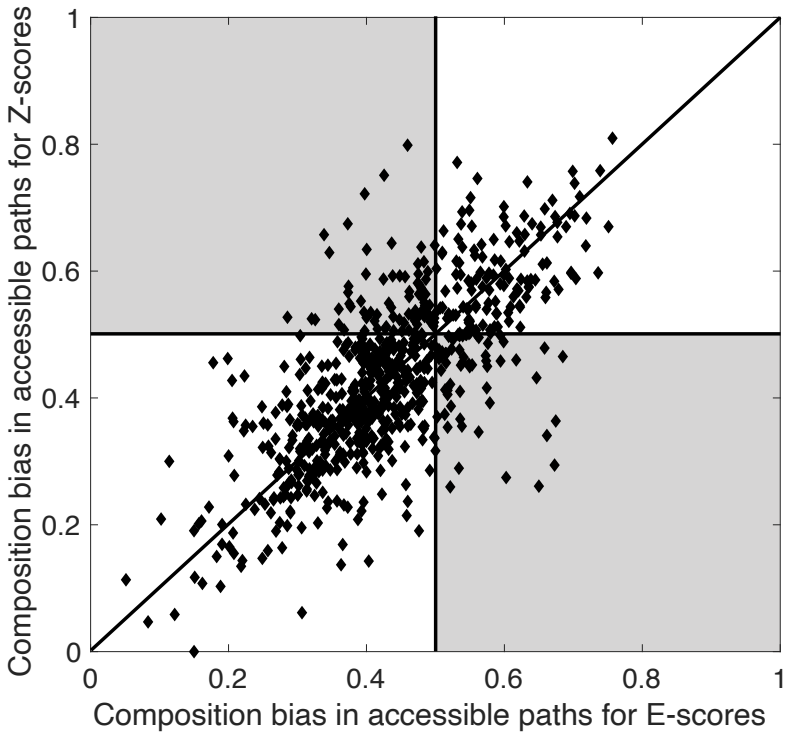

FIGURE S2.5: **Constructing genotype-phenotype landscapes with E-scores or Z-scores results in highly similar composition biases among accessible mutational paths.** Pearson's correlation coefficient $r = 0.7212$, $p < 10^{-10}$. Data pertain to all 746 landscapes. The shaded gray regions highlight the 83 landscapes that switch from exhibiting a composition bias toward transversions to a composition bias toward transitions (or vice versa).

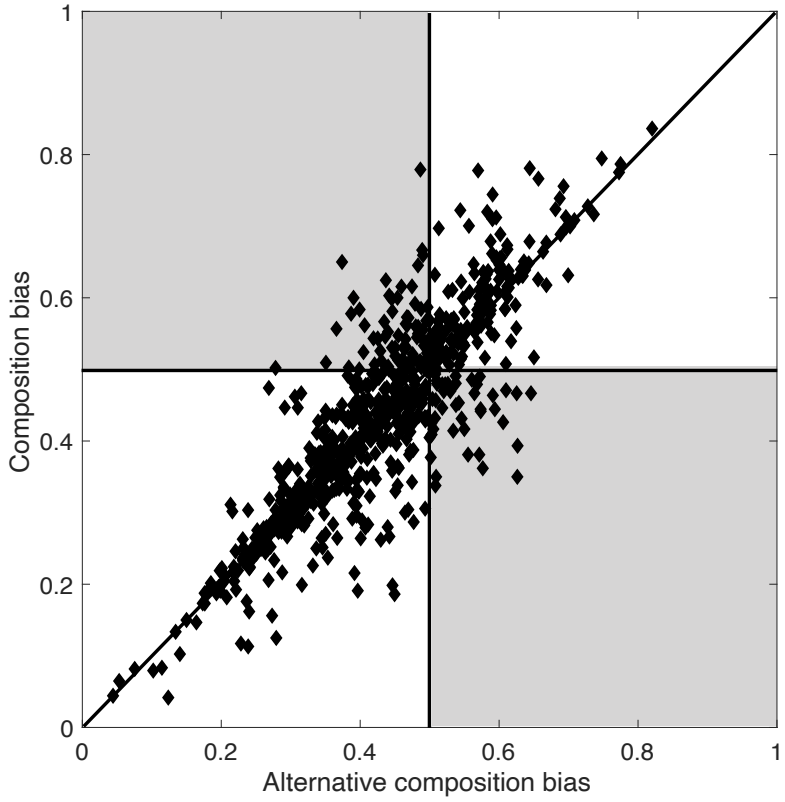

FIGURE S2.6: **Quantifying composition bias using an alternative accessibility criteria results in very similar composition biases.** Data pertain to all 746 landscapes, Pearson's correlation coefficient($r = 0.8541$, $p < 10^{-12}$). The shaded gray regions highlight the 77 landscapes that switch from exhibiting a composition bias toward transversions to a composition bias toward transitions (or vice versa).

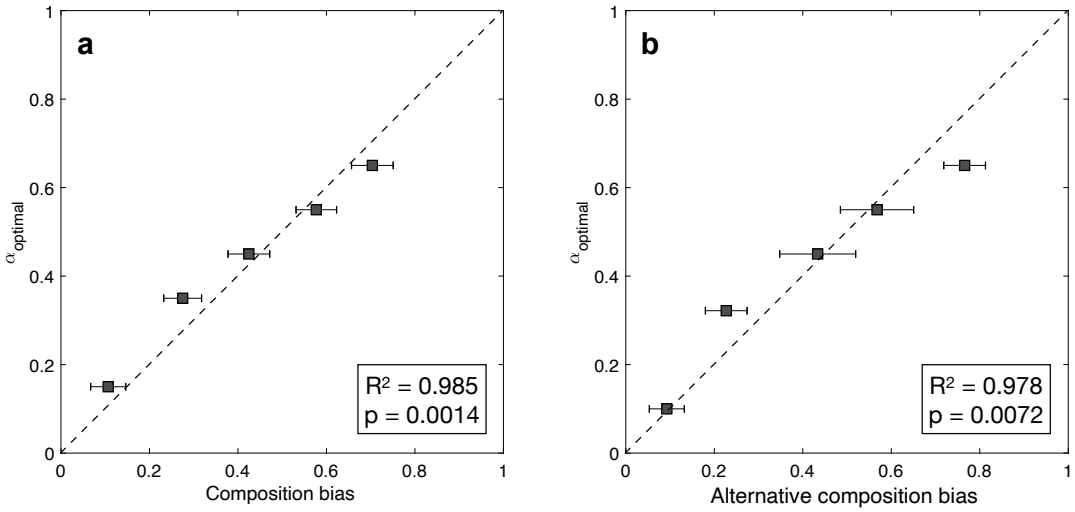

FIGURE S2.7: **The mutation bias that maximizes P_{peak} correlates strongly with composition bias, measured using two definitions of accessible mutational paths.** In **(a)**, each step in an accessible mutational path increases binding affinity by at least δ. In **(b)**, each step in an accessible mutational path does not decrease binding affinity more than δ. Data pertain to all 746 landscapes, each of which has its own noise threshold δ.

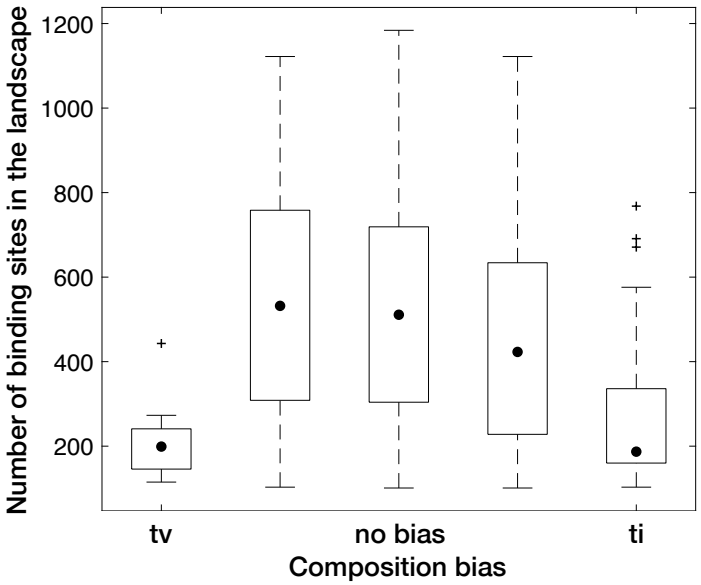

FIGURE S2.8: **Landscapes with strong composition bias comprise fewer binding sites than landscapes with little or no composition bias.** The number of binding sites per landscape is shown in relation to composition bias. Landscapes are grouped as in Fig. 2.3. Data pertain to all 746 landscapes. Black dots indicate medians, whiskers indicate the 25th and 75th percentiles, and cross symbols indicate outliers.

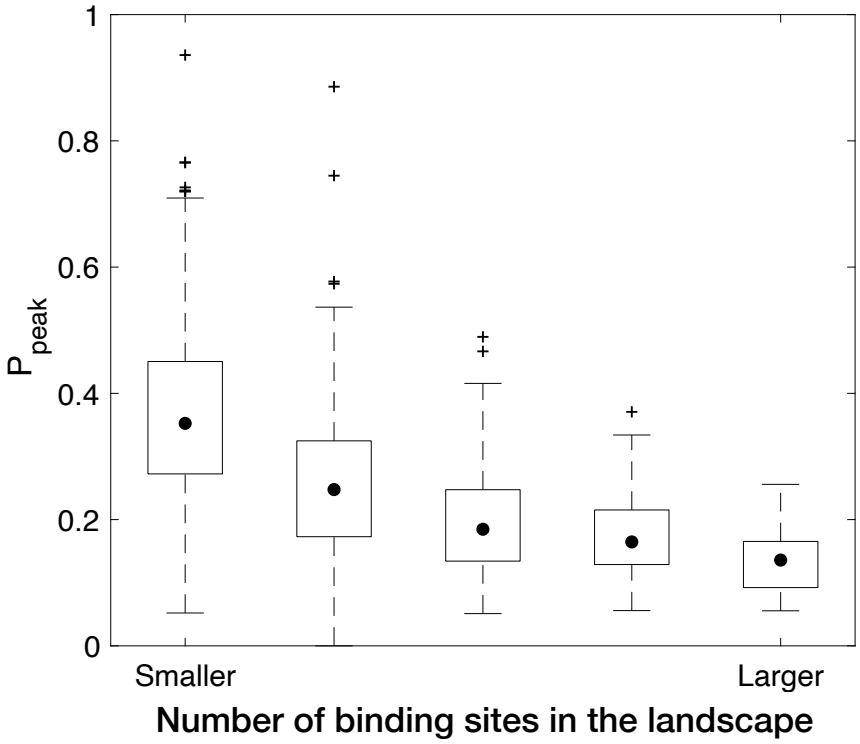

FIGURE S2.9: **Smaller landscapes are more navigable.** Landscapes are grouped based on the number of binding sites they comprise, with the average number of binding sites per landscape per group ranging from 165.07 to 1037.60 in five linearly spaced increments. The probability of evolving to the global peak in the absence of mutation bias ($\alpha = 0.5$) is shown in relation to landscape size. Data pertain to all 746 landscapes. Black dots indicate medians, whiskers indicate the 25th and 75th percentiles, and cross symbols indicate outliers.

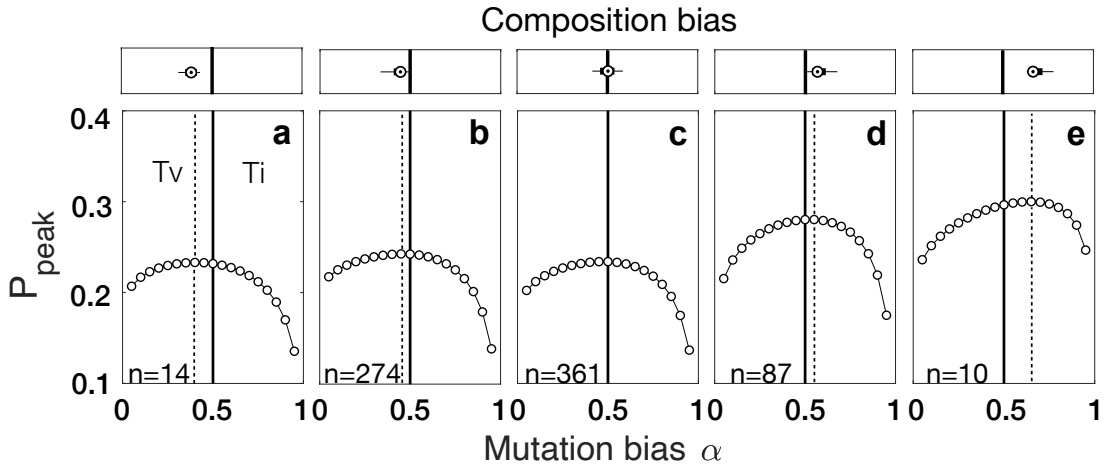

FIGURE S2.10: **Mutation bias interacts with composition bias to influence landscape navigability, regardless of whether selection favors low or high affinity binding sites.** The probability P_{peak} of reaching the global peak is shown for 19 different values of the mutation bias parameter α. The solid vertical lines indicate no bias in mutation supply ($\alpha = 0.5$) and the dashed vertical lines indicate the value of α that maximizes P_{peak}. Landscapes are grouped based on their composition bias and the distribution of composition bias per panel is shown on top of each panel. The number of landscapes per panel is indicated is the bottom left corner. Fitness is a function of binding affinity (E-score) using the Gaussian function $exp(-((E - E_{\text{opt}})/\sigma)^2)$, where E is the E-score of a binding site, E_{opt} is the optimal E-score, and σ is the variance parameter. Here, $E_{\text{opt}} = 0.35$ (the lowest E-score in our landscapes) and $\sigma = 0.1$.

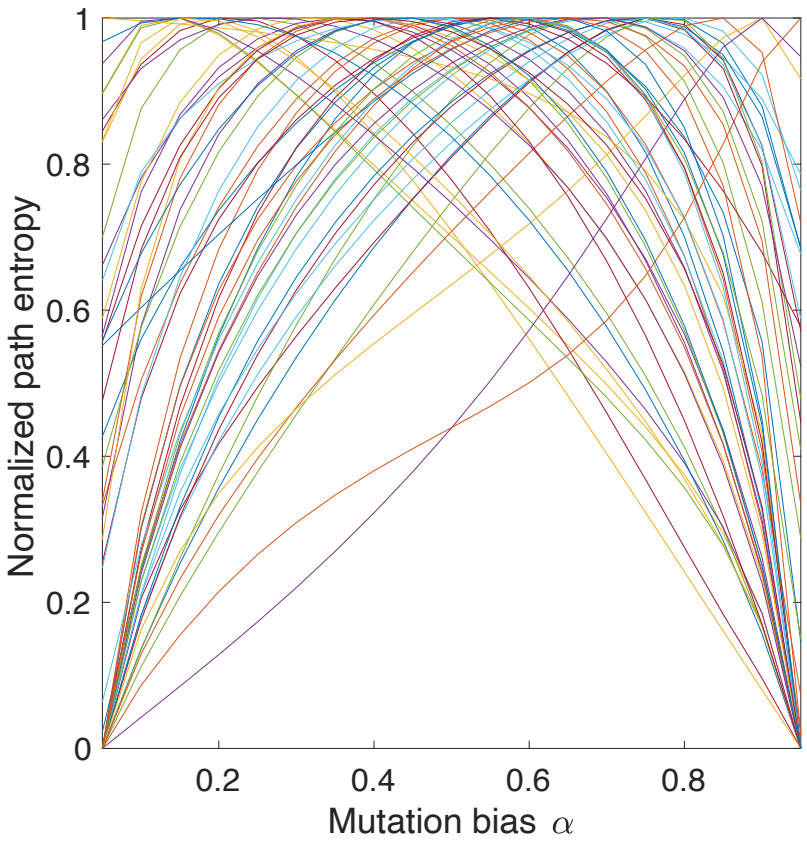

FIGURE S2.11: **Path entropy as a function of mutation bias.** Data pertain to 50 randomly chosen landscapes for 19 different values of the mutation bias parameter α.

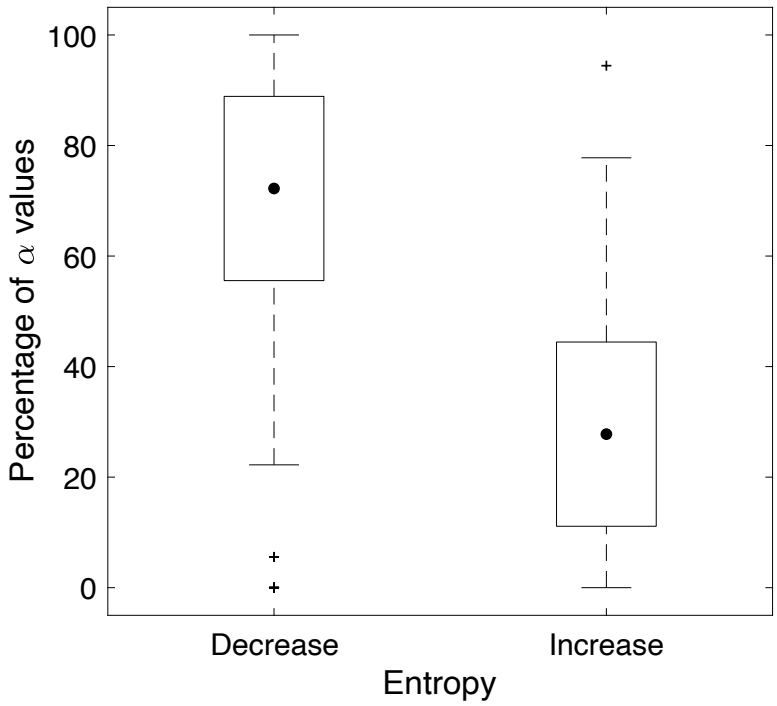

FIGURE S2.12: **Mutation bias typically increases, but sometimes decreases, the predictability of evolution.** Shown are the percentage of mutation bias values α that increase or decrease path entropy (which is inversely related to the predictability of evolution), relative to when there is no mutation bias ($\alpha = 0.5$). Data pertain to all 19 values of the mutation bias parameter α on each of the 746 landscapes. Black dots indicate medians, whiskers indicate the 25th and 75th percentiles, and cross symbols indicate outliers.

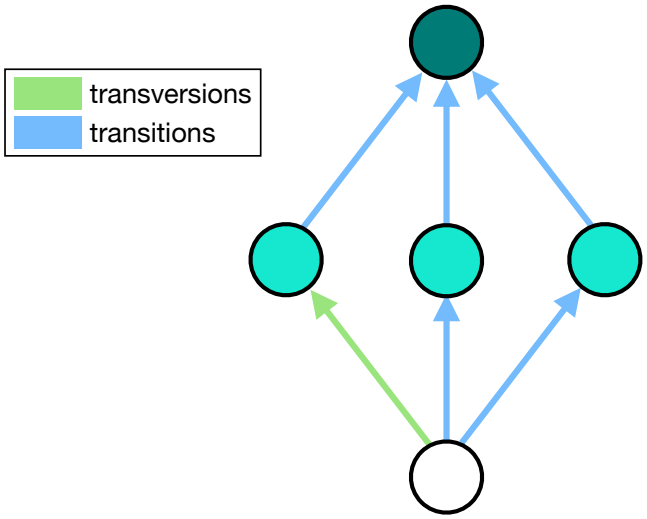

FIGURE S2.13: **Composition bias is a weak predictor of the mutation bias α that maximizes path entropy.** Nodes represent sequences in a landscape, and directed edges represent accessible mutations between sequences. Edge colors represent mutation type and node colors represent binding affinity (darker = higher). This landscape exhibits a strong composition bias toward transitions. Path entropy is therefore minimized by a strong mutation bias toward transversions, because an evolving population will utilize only one of the three accessible mutational paths. Conversely, one might expect path entropy to be maximized by a strong mutation bias toward transitions. However, this is not the case, because an evolving population will only utilize two of the three accessible mutational paths. The mutation bias that maximizes path entropy is actually the one that makes the three first-step mutations equiprobable. In more complex scenarios, with more and longer paths that include a greater diversity of binding affinities and more heterogeneous distributions of mutation types, a single summary statistic like composition bias is unlikely to accurately predict the value of mutation bias that maximizes path entropy.

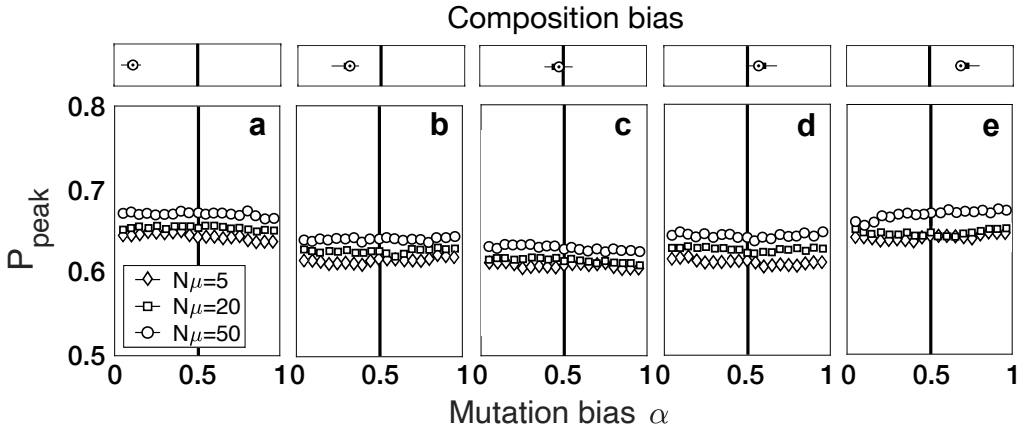

FIGURE S2.14: **Mutation bias has little to no effect on landscape navigability when the mutation supply is high.** The probability P_{peak} of reaching the global peak is shown for 19 different values of the mutation bias parameter α. This probability is calculated as the proportion of simulations in which at least half of the population evolves to the global peak. The solid vertical lines indicate no bias in mutation supply ($\alpha = 0.5$). Landscapes are grouped based on their composition bias and the distribution of composition bias per panel is shown on top of each panel as in Fig. 2.3.

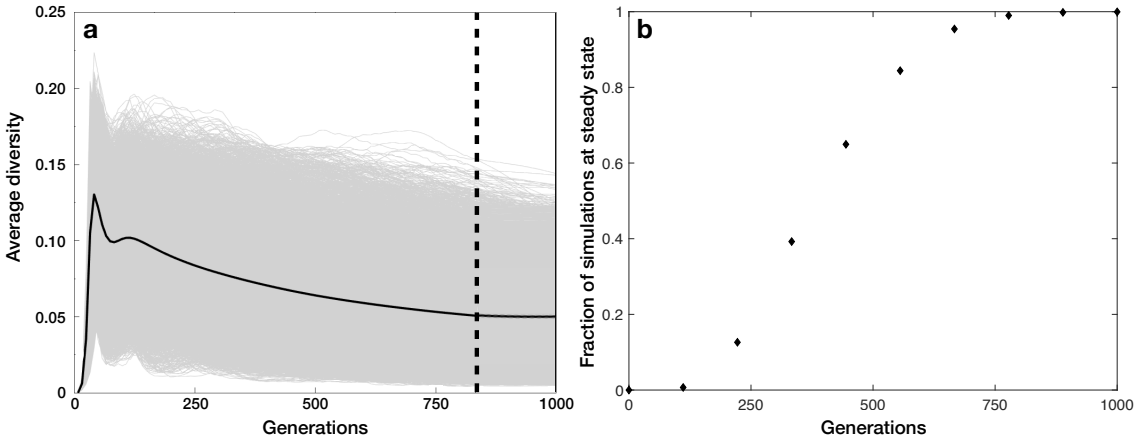

FIGURE S2.15: **Simulations of the Wright-Fisher model typically reach steady state within 1,000 generations. (a)** The average Shannon's diversity is shown in relation to the number of generations. After 827 generations, more than 99.9% of the 4,252,200 simulations reached steady state diversity levels within a tolerance of 0.01% of the final diversity level. The black line shows the average across all simulations. **(b)** The fraction of simulations that have reached steady state diversity levels is shown in relation to generation number.

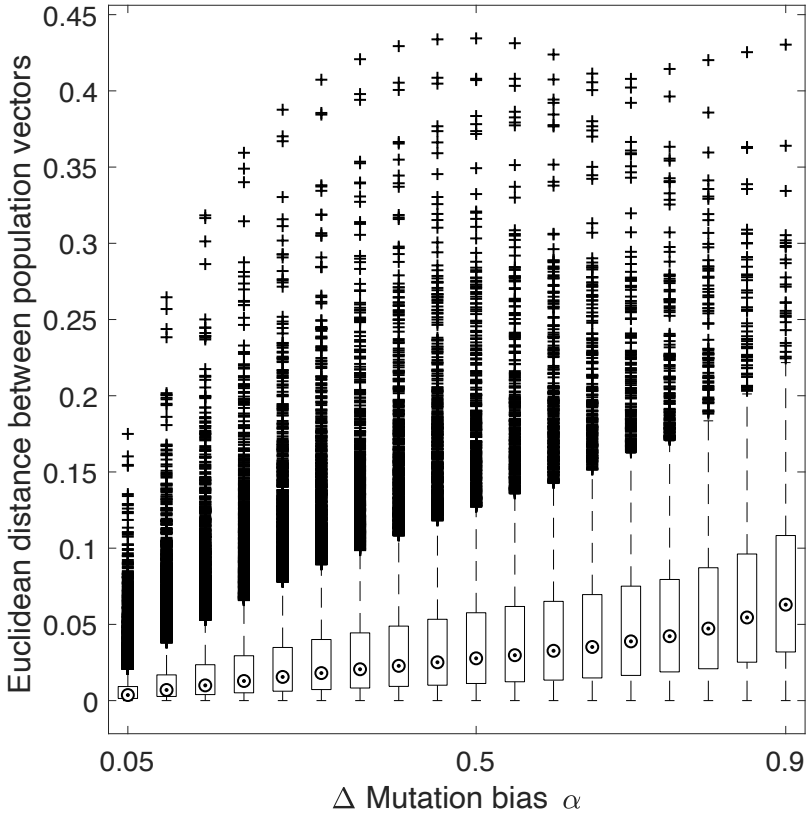

FIGURE S2.16: **Mutation bias influences the distribution of infinite populations on a genotype-phenotype landscape.** We characterized the steady state distribution of infinite populations as the eigenvector that corresponds to the largest eigenvalue of the matrix P (Methods). For any pair of such populations, we measured their overlap as the Euclidean distance between these eigenvectors - the shorter the distance, the higher the overlap - This panel shows this distance for pairs of populations in relation to the difference in their mutation bias parameters.

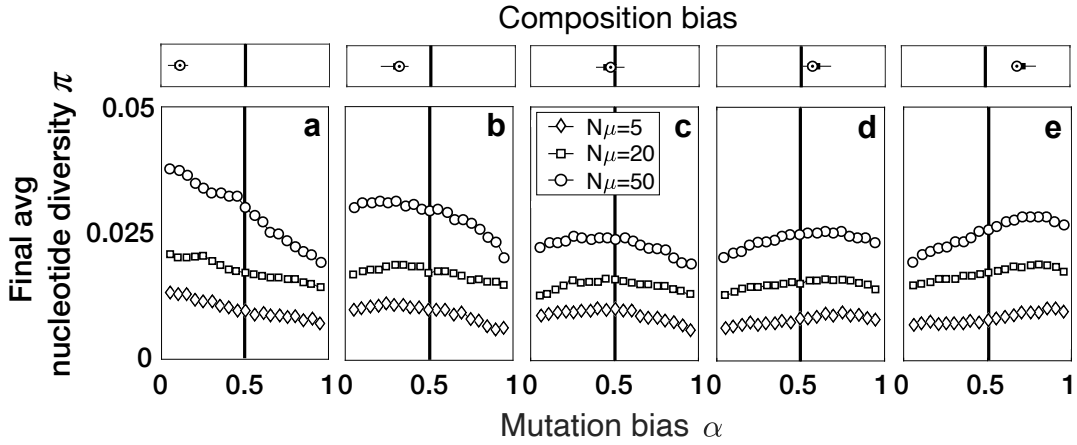

FIGURE S2.17: **Mutation bias interacts with composition bias to influence the evolution of nucleotide diversity π. (a-e)** The final average nucleotide diversity of evolved populations at steady state is shown for 19 different values of the mutation bias parameter α, and for each of three different values of mutation supply $N\mu$ (see legend). The solid vertical lines indicate no bias in mutation supply (α= 0.5). Landscapes are grouped based on their composition bias and the distribution of composition bias per panel is shown on top of each panel, as in Fig. 2.3.

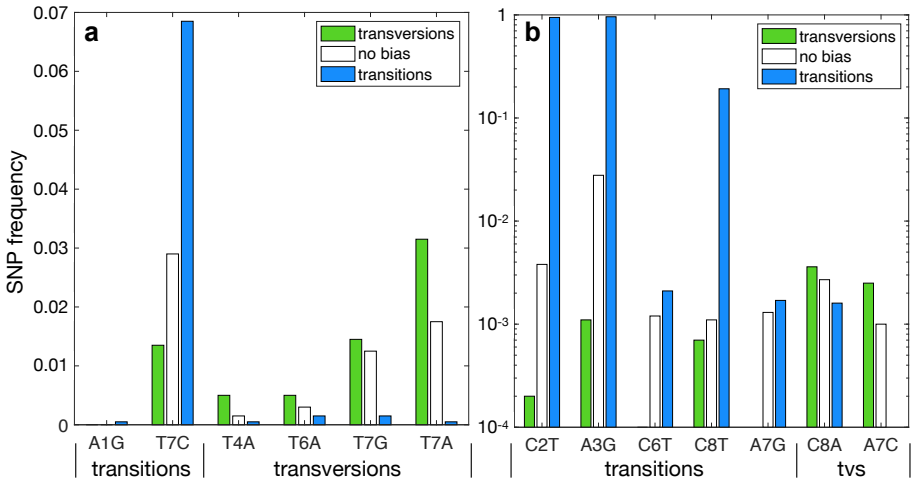

FIGURE S2.18: **Mutation bias and composition bias interact to influence allele frequency spectra.** The bars correspond to polymorphisms in the population at steady state, relative to the sequence that evolved to the highest frequency in the absence of mutation bias. Bar colors correspond to three different values of the mutation bias parameter α: green for transversions (α=0.05), white for no bias (α=0.5) and blue for transitions (α=0.95). The panels correspond to two *Mus musculus* landscapes **(a)** Arid5a (composition bias toward transversions) and **(b)** Gm397 (composition bias toward transitions).

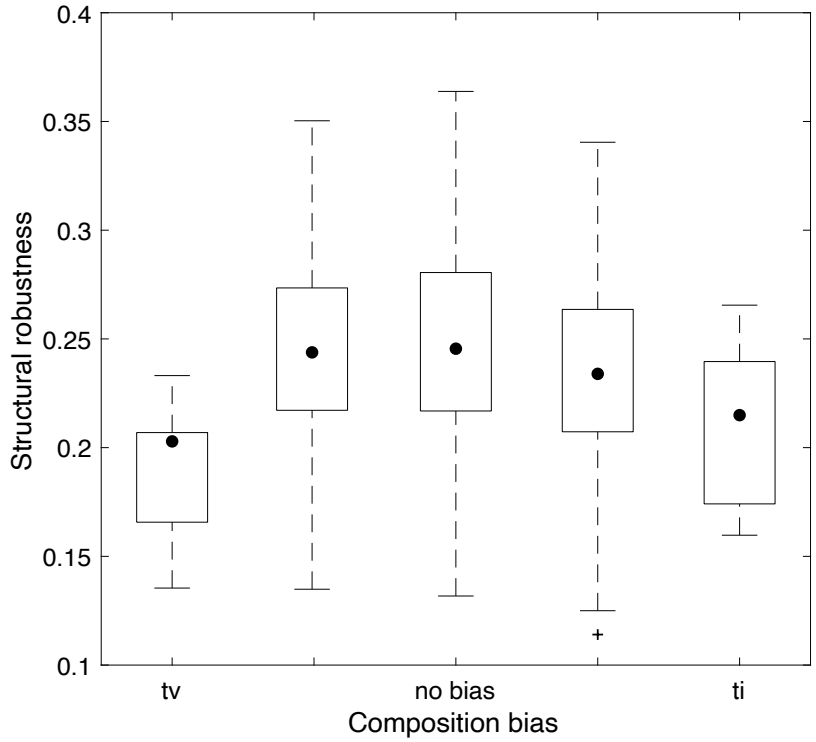

FIGURE S2.19: **Genotypes in landscapes with strong composition bias are less robust than genotypes in landscapes without composition bias.** The y-axis shows the average mutational robustness of all genotypes in each landscape. The x-axis shows the composition bias. Landscapes are grouped as in Fig. 2.3. Data pertain to all 746 landscapes. Black dots indicate medians, whiskers indicate the 25th and 75th percentiles, and cross symbols indicate outliers.

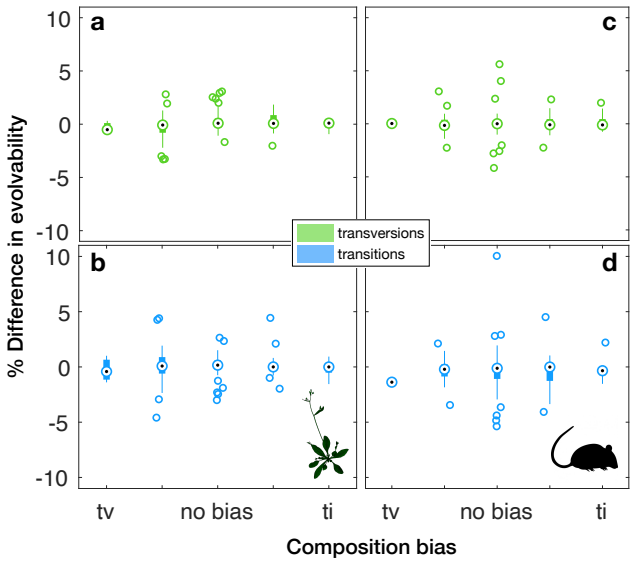

FIGURE S2.20: **Mutation bias influences the evolvability of polymorphic populations.** The y-axis shows the difference in evolvability, which is calculated as the difference between the evolvability of a population at steady state when there is no mutation bias and when there is a strong bias toward transversions ($\alpha = 0.05$) or transitions ($\alpha = 0.95$). Data pertain to 128 transcription factors from **(a,b)** *Arabidopsis thaliana*, and **(c,d)** 128 transcription factors from *Mus musculus*. Landscapes are grouped according to their composition bias as in previous figures. Parameters: $N = 10^4$, $N\mu = 50$. As an example, for a given transcription factor, a 10% change in evolvability under strong transition bias could mean that the one-mutant neighbors of the sequences evolved at steady state bind 11 transcription factors, whereas without mutation bias, the one-mutant neighbors of the sequences evolved at steady state bind 10 transcription factors.

3

MUTATION BIAS SHAPES THE SPECTRUM OF ADAPTIVE SUBSTITUTIONS

Published as: Alejandro V. Cano, Hana Rozhoňová, Arlin Stoltzfus, David M. McCandlish, & Joshua L. Payne (2022). Mutation bias shapes the spectrum of adaptive substitutions. *Proceedings of the National Academy of Sciences*, 119(7), e2119720119. https://doi.org/10.1073/pnas.2119720119

Author's contributions: A.V.C., H.R., A.S., D.M.M., and J.L.P. designed research; A.V.C. and H.R. performed research; A.V.C., H.R., A.S., D.M.M., and J.L.P. analyzed data; and A.V.C.,H.R., A.S., D.M.M., and J.L.P. wrote the paper.

3.1 ABSTRACT

Evolutionary adaptation often occurs by the fixation of beneficial mutations. This mode of adaptation can be characterized quantitatively by a spectrum of adaptive substitutions, i.e., a distribution for types of changes fixed in adaptation. Recent work establishes that the changes involved in adaptation reflect common types of mutations, raising the question of how strongly the mutation spectrum shapes the spectrum of adaptive substitutions. We address this question with a codon-based model for the spectrum of adaptive amino acid substitutions, applied to three large data sets covering thousands of amino acid changes identified in natural and experimental adaptation in *S. cerevisiae*, *E. coli*, and *M. tuberculosis*. Using species-specific mutation spectra based on prior knowledge, we find that the mutation spectrum has a proportional influence on the spectrum of adaptive substitutions in all three species. Indeed, we find that by inferring the mutation rates that best explain the spectrum of adaptive substitutions, we can accurately recover the species-specific mutation spectra. However, we also find that the predictive power of the model differs substantially between the three species. To better understand these differences, we use population simulations to explore the factors that influence how closely the spectrum of adaptive substitutions mirrors the mutation spectrum. The results show that the influence of the mutation spectrum decreases with increasing mutational supply ($N\mu$), and that predictive power is strongly affected by the number and diversity of beneficial mutations.

SIGNIFICANCE STATEMENT

How do mutational biases influence the process of adaptation? A common assumption is that selection alone determines the course of adaptation from abundant pre-existing variation. Yet, theoretical work shows broad conditions under which the mutation rate to a given type of variant strongly influences its probability of contributing to adaptation. Here we introduce a statistical approach to analyzing how mutation shapes protein sequence adaptation. Using large data sets from three different species, we show that the mutation spectrum has a proportional influence on the types of changes fixed in adaptation. We also show via computer simulations that a variety of factors can influence how closely the spectrum of adaptive substitutions reflects the spectrum of variants introduced by mutation.

3.2 INTRODUCTION

The spectrum of adaptive substitutions is a distribution of types of changes fixed in adaptation. A systematic empirical picture of the spectrum of adaptive substitutions is beginning to emerge from methods of identifying and verifying individual adaptive changes at the molecular level. The most familiar method is the retrospective analysis of adaptive species differences, often in cases where multiple substitutions target the same protein, e.g., changes to photoreceptors involved in spectral tuning [1], changes to ATPase involved in cardiac glycoside resistance [2], or changes to hemoglobin involved in altitude adaptation [3]. Other retrospective analyses focus on cases of recent local adaptation, such as the repeated emergence of antibiotic-resistant bacteria [4], [5] or herbicide-resistant plants [6]. In addition, experimental studies of adaptation in the laboratory

provide large and systematic sets of data on the spectrum of adaptive substitutions [7], [8]. While the first two types of studies tend to focus on specific target genes, the third approach, combined with genome sequencing, casts a much broader net, covering the entire genome. Such data were rare just 15 years ago, but they are now sufficiently abundant—cataloging thousands of adaptive events—that accounting for the species-specific spectrum of adaptive substitutions represents an important challenge.

One aspect of this challenge is to understand the role of mutation in shaping the spectrum of adaptive substitutions. Systematic studies of the distribution of mutational types in diverse organisms [9]–[17] have demonstrated the presence of a variety of biases, including transition bias and GC:AT bias, as well as CpG bias and other context effects (for review, see [18]). At the same time, multiple studies have now shown that adaptive substitutions are enriched for mutationally likely changes [5], [19]–[27]. For instance, the influence of a mutational bias favoring transitions is evident in the evolution of antibiotic resistance in *Mycobacterium tuberculosis* [5]. Likewise, the evolution of increased oxygen-affinity in hemoglobins of high-altitude birds shows a tendency to occur at CpG hotspots [24].

Such studies have shown effects of specific types of mutation bias using statistical tests for asymmetry, i.e., tests for a significant excess of a mutationally favored type, relative to a null expectation of parity. A more general question is how strongly the entire mutation spectrum shapes the spectrum of adaptive substitutions. That is, the entire mutation spectrum reflects (simultaneously) all relevant mutation biases, because it describes the relative rates of the different mutation types. Mutation spectra have been experimentally characterized in a diversity of species [9]–[17], and these universally reveal some form of mutation bias in that the different mutation types do not occur with the same relative rates. Such biased mutation spectra shape the spectra of adaptive substitutions to some degree that is, in principle, quantifiable and measurable.

Here, we provide an approach to answer this more general question, based on modeling the spectrum of missense mutations underlying adaptation as a function of the nucleotide mutation spectrum. More specifically, we use negative binomial regression to model observed numbers of adaptive codon-to-amino acid substitutions as a function of codon frequencies and per-nucleotide mutation rates, which we estimate from published data on mutation frequencies. This modeling framework allows us to measure the influence of mutation bias on adaptive evolution in terms of the regression coefficient associated with the mutation spectrum.

We separately apply this approach to three large data sets of missense changes associated with adaptation in *Saccharomyces cerevisiae*, *Escherichia coli*, and *Mycobacterium tuberculosis*. We find that, in each case, the regression on the mutation spectrum is significant, with a regression coefficient close to 1 (proportional effect) and significantly different from zero (no effect). This indicates that mutational biases play an important role in determining which mutations, among those that are beneficial, underlie molecular adaptation. Whereas the ability to predict the spectrum of adaptive substitutions differs substantially amongst the three species, in each case we find that experimentally determined mutation spectra provide better model fits than the vast majority of randomized mutation spectra, confirming the relevance of empirical mutation spectra outside of the controlled conditions in which they are typically measured. Moreover, we show that by inferring the optimal mutational spectrum based on the spectrum of adaptive substitutions, we can accurately recover species-specific patterns of mutational bias previously documented via

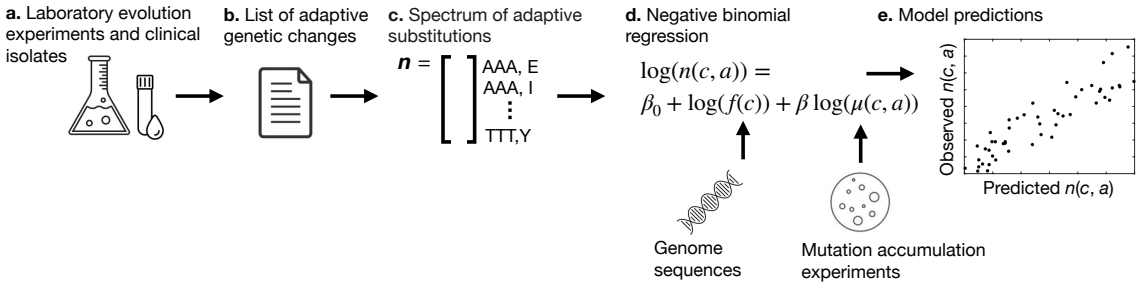

FIGURE 3.1: **Workflow.** (a) We use data from laboratory evolution experiments (*E. coli* and *S. cerevisiae*) and clinical isolates (*M. tuberculosis*) to curate (b) a list of genetic changes associated with adaptation for each species. (c) From each list of adaptive changes, we construct the spectrum of adaptive substitutions $\mathbf{n}$. Each element in this spectrum $\mathbf{n}(c, a)$ corresponds to one of the 354 distinct changes from codon c to amino acid a that can be produced by a single nucleotide mutation under the standard genetic code, and tallies the number of adaptive events per codon-to-amino acid change. (d) We perform negative binomial regression to model the influence of mutation bias on the spectrum of adaptive events, using codon frequencies derived from genome sequences and experimentally characterized mutation spectra. (e) We use the fitted model to predict the spectrum of adaptive events.

mutation accumulation experiments or patterns of neutral diversity. Finally, we use simulations of a population model to explore the possible reasons for differences in predictability of the spectrum of adaptive substitutions. As expected, the impact of the mutation spectrum decreases as the total mutation supply ($N\mu$) increases. However, other factors are important, such as the size and heterogeneity (in adaptive value) of the set of adaptive mutations.

3.3 RESULTS

Data and model.

We curated a list of previously-reported missense substitutions associated with adaptation for each of three species: *S. cerevisiae*, *E. coli*, and *M. tuberculosis* (Fig. 3.1a,b; Methods). Note that "substitution" here refers to an evolutionary change, whereas we restrict the term "mutation" to mutational changes or categories, following the definitions provided in the *SI Appendix*. For *S. cerevisiae*, the substitutions were associated with adaptation to high salinity [28], low glucose [28], and rich media [29], as well as the genetic stress of gene knockout [30]; for *E. coli*, the substitutions were associated with adaptation to temperature stress during laboratory evolution [8]; for *M. tuberculosis*, the substitutions were identified in clinical isolates resistant to one or more of eleven antibiotics or antibiotic classes [5]. Whereas the *M. tuberculosis* data set is composed entirely, or almost entirely, of *bona fide* adaptive changes that have been experimentally verified to confer antibiotic resistance [5], the data sets for *S. cerevisiae* and *E. coli* are likely contaminated with hitchhikers, i.e., mutations that are not drivers of adaptation, but which reached a high frequency due to linkage with a driver. Below, we first present our results under the assumption that substitutions in each data set are exclusively adaptive, and then use simulations to assess the robustness of our conclusions to various degrees of contamination.

Each data set consists of a list of events of putatively adaptive missense substitution, each of which can be defined by a specific initial and final genomic state. For example, the substitution

defined by a G→C transversion in the second position of codon 315 of KatG in *M. tuberculosis*, which changes Ser (AGC) to Thr (ACC), confers resistance to the antibiotic isoniazid [31]. In our data set, we observe 445 independent instances of adaptation via this specific genomic alteration; for the sake of brevity we describe this as observing 445 "events" corresponding to this specific adaptive "path". Here, to define a spectrum of adaptive substitutions, we further aggregate these adaptive missense substitutions into types of changes. Possible types of changes include nucleotide-to-nucleotide, codon-to-codon, codon-to-amino acid, and amino acid-to-amino acid changes, each of which results in a different level of aggregation of the mutational events. We focus on codon-to-amino acid changes, which we track only by the initial codon and final amino acid of the substitution, without regard to the specific gene or amino acid position where the substitution occurred. Given that there are 354 such types of codon-to-amino-acid changes allowed by the standard genetic code, the spectrum of adaptive substitutions for each species is a 354-element vector $\mathbf{n}$, where each element $\mathbf{n}(c, a)$ is a count of the number of events of single-nucleotide changes from codon c to amino acid a (Fig. 3.1c; Methods). Table 3.1 reports the total number of mutational paths and events, as well as the number of non-zero elements in the spectrum of adaptive substitutions (out of 354) for each data set.

Our goal is to quantify how strongly the mutation spectrum shapes the spectrum of adaptive substitutions. To do so, we specify a phenomenological model that treats each element in the spectrum of adaptive substitutions as the product of the starting codon frequency and the relevant mutation rate, raised to an exponent β representing the degree of mutational influence, e.g., $\beta = 0$ would indicate no influence. More specifically, we model the expected number $\mathbb{E}[\mathbf{n}(c, a)]$ of adaptive substitutions from codon c to amino acid a as being directly proportional to the genomic frequency $f(c)$ of codon c (i.e., $f(c)$ is the number of times codon c appears in protein-coding regions of the genome divided by the total number of codons in protein-coding regions of the genome) and the total mutation rate $\mu(c, a)$ of codon c to codons for amino acid a raised to the power of β, as follows:

$$\mathbb{E}[\mathbf{n}(c, a)] \propto f(c)\mu(c, a)^{\beta}. \tag{3.1}$$

Taking the logarithm of this equation gives

$$\log \mathbb{E}[\mathbf{n}(c, a)] = \beta_0 + \log f(c) + \beta \log \mu(c, a) \tag{3.2}$$

where β_0 is the logarithm of the constant of proportionality (see Methods and *SI Appendix*). This formulation allows us to estimate β_0 and β from our observed data sets using negative binomial regression, which is appropriate for counts data that are over-dispersed [32], as is the case for the observed spectra of adaptive substitutions.

Given the form of this regression, β represents a coefficient of mutational influence, capturing the effect of the entire mutation spectrum on the entire spectrum of adaptive substitutions. An inferred value of $\beta = 0$ indicates that $\mathbb{E}[\mathbf{n}(c, a)]$ does not depend on $\mu(c, a)$, implying that the mutation spectrum has no influence on the spectrum of adaptive substitutions; when $\beta = 1$, $\mathbb{E}[\mathbf{n}(c, a)]$ is directly proportional to $\mu(c, a)$, indicating a strong influence of the mutation spectrum on the spectrum of adaptive substitutions; values of β between 0 and 1 indicate an intermediate influence.

	Data		Influence of mutation spectrum	
Species	Paths	Events	β	p_β
S. cerevisiae	534	713	1.05 ± 0.08	$< 10^{-16}$
E. coli	492	602	0.98 ± 0.14	$< 10^{-11}$
M. tuberculosis	283	4413	0.85 ± 0.23	$< 10^{-3}$

TABLE 3.1: **Data and negative binomial regression.** Shown are the observed numbers of paths and events for adaptive changes in the three data sets, along with calculated values for the mutation coefficient β (with standard error) and its p-value.

Population-genetic theory and prior simulation studies suggest a variety of factors likely to influence β, including population size, absolute mutation rates, fitness landscape architecture, and whether adaptation is short-term or long-term [33]–[36]. In particular, prior results suggest that the supply of beneficial mutations will often influence β.

When new mutations are sufficiently rare, beneficial mutations sweep through the population one at a time, resulting in the so-called origin-fixation [37] or strong-selection-weak-mutation (SSWM) [38] regime. In this regime, the substitution rate is directly proportional to the mutation rate, implying $\beta \approx 1$ [33], [37]. When the beneficial mutation supply is high, multiple adaptive mutations may compete against each other, resulting in "clonal interference" [39]. Due to clonal interference, late-arising mutant alleles with larger selection coefficients may prevent the fixation of early-arising alleles favored by mutation, decreasing the influence of mutation bias [33], [35], and leading to an expected reduction in β.

Mutation bias strongly influences adaptation in three distinct species.

To what extent does the mutation spectrum influence the outcome of adaptive evolution? To answer this question, we used empirical mutation spectra generated in prior studies from mutation accumulation experiments or patterns of neutral diversity. These prior studies were carried out independently of the studies used to characterize the spectrum of adaptive substitutions. The three species differ substantially in their mutation spectra (*SI Appendix*, Fig. S3.1a). *M. tuberculosis* shows the greatest heterogeneity, with a 14.7-fold range of rates, whereas *S. cerevisiae* and *E. coli* have smaller ranges of 5.6-fold and 4.7-fold, respectively. The species also differ substantially in the rates of individual types of nucleotide mutations, e.g., the rate of G→C transversion is 2.1-fold higher in *S. cerevisiae* than in *E. coli* (*SI Appendix*, Fig. S3.1b), whereas the rate of A→T transversions is 2.5-fold higher in *S. cerevisiae* (*SI Appendix*, Fig. S3.1c) and 2.9-fold higher in *E. coli* (*SI Appendix*, Fig. S3.1d) than in *M. tuberculosis*.

Our first observation is that, when we reduce the adaptive missense substitutions to the six types of underlying nucleotide mutations, the distribution closely follows the mutation spectrum for each species (Fig. 3.2a-c). Specifically, the correlation coefficients between the mutation rates of the six mutation types and their frequencies among adaptive substitutions are 0.83 ($p = 0.041$),

0.91 ($p = 0.012$), and 0.93 ($p = 0.008$) for *S. cerevisiae*, *E. coli* and *M. tuberculosis*, respectively. However, this naive comparison ignores potentially confounding effects of the genetic code and codon usage, where in particular the three species differ substantially in their patterns of codon usage (*SI Appendix*, Fig. S3.1e-g). For example GAA (Glu) is the most frequent codon in *S. cerevisiae* (frequency 0.045) and the 2nd most frequent codon in *E. coli* (frequency 0.039), but it appears less frequently in *M. tuberculosis* (frequency 0.016). Thus, we might expect adaptive GAA→AAA (Glu→Lys) changes to occur more frequently in *S. cerevisiae* and *E. coli* than in *M. tuberculosis*, merely by merit of the greater frequency of GAA in the former two species. To account for this type of influence, we apply negative binomial regression to the codon-based model described above (Eqn. 3.2). The results, shown in Table 3.1, reveal a strong and statistically significant influence of mutation bias in all three species, with each of the 95 % confidence intervals containing $\beta = 1$ (proportional effect), and excluding $\beta = 0$ (no effect). Specifically, for *S. cerevisiae*, $\beta = 1.05$ (95 % CI, 0.89 to 1.21), for *E. coli*, $\beta = 0.98$ (95 % CI, 0.71 to 1.25), and for *M. tuberculosis*, $\beta = 0.85$ (95 %, 0.31 to 1.37), so that in all three species, differences in mutation rates produce approximately proportional changes in the spectrum of adaptive substitutions. Whereas such strong mutational effects are typically associated with neutral evolution, theory [33], [35], [36], [40], prior evidence [5], [19]–[27], and our simulations (below) indicate that such effects are possible even when all fixations are selective. What this suggests about the roles of mutation and selection is addressed further in the Discussion.

Prior work has uncovered an enrichment of transition mutations in the *M. tuberculosis* data set, which was attributed to the high transition-transversion ratio in the mutation spectrum of this species [5]. We therefore wondered whether the entire mutation spectrum provides a better model fit than just the transition-transversion ratio. To find out, we used a likelihood ratio test to compare two nested models that differ in the mutation term: a model that only uses the transition-transversion ratio, and a model that uses both the transition-transversion ratio and the rest of the mutation spectrum (Methods). For all three species, we find that the model using both the transition-transversion ratio and the rest of the mutation spectrum provides significantly better fits, and that $\beta \approx 1$ on both terms of the regression (*SI appendix*, Table S3.2).

Having seen the influence of the mutation spectrum on the spectrum of adaptive substitutions, we can also ask to what extent the mutation spectrum, pattern of codon usage, and the structure of the standard genetic code are jointly sufficient to explain the spectrum of adaptive substitutions observed in each species. Figure 3.2d-f shows the observed frequency of each type of codon-to-amino acid change in relation to its predicted frequency under our fitted models. We observe from this figure that despite the mutation spectrum having its maximum theoretically predicted influence ($\beta \approx 1$) in each species, the predictive power of our model nonetheless differs substantially among the three species, with the correlation between predicted and observed frequencies dropping dramatically from 0.68 in *S. cerevisiae*, to 0.41 in *E. coli*, to only 0.16 in *M. tuberculosis*. While all of these correlations are statistically significant (Table 3.2), it is clear that the predictive power of a model depending only on mutation rates, codon frequencies, and the structure of the standard genetic code differs between these three species, an observation that we will return to shortly.

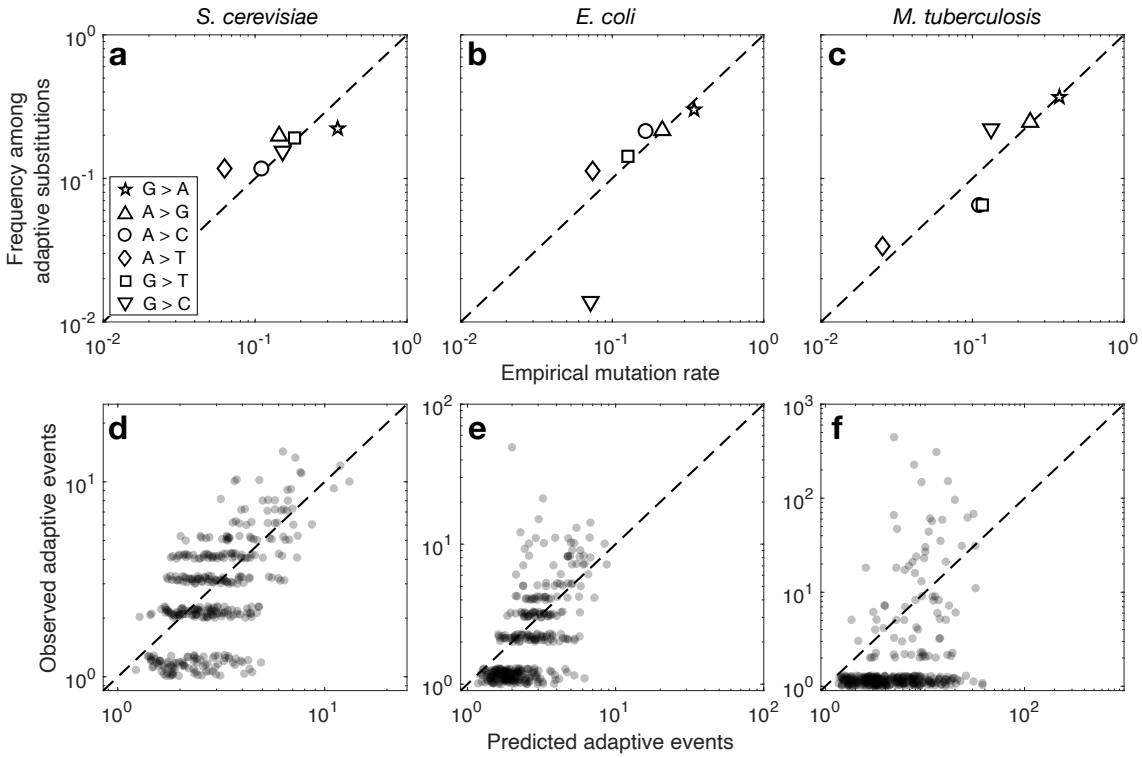

FIGURE 3.2: **Predicted and observed substitutions at the nucleotide and codon-to-amino acid levels.** (a-c) The frequency of nucleotide changes among adaptive substitutions is plotted as a function of the empirical mutation rate for (a) *S. cerevisiae*, (b) *E. coli*, and (c) *M. tuberculosis*. The symbols correspond to the six different types of point mutations (inset in panel a). (d-f) The predicted spectra of adaptive substitutions are shown in relation to the observed spectra of adaptive substitutions for (d) *S. cerevisiae*, (e) *E. coli*, and (f) *M. tuberculosis*. See *SI Appendix*, Table S3.3 for model predictions using codon frequencies alone. For visualisation purposes, a pseudo count of 1 event and a jitter of range [0,0.3] were added to both the observed and predicted numbers of events in panels (d-f).

	Prediction model		Spectrum elements	
Species	Correlation [CI]	p_{corr}	Non-zero	Entropy
S. cerevisiae	0.68 [0.62, 0.73]	$< 10^{-16}$	265	0.91
E. coli	0.41 [0.31, 0.49]	$< 10^{-14}$	176	0.80
M. tuberculosis	0.16 [0.05, 0.26]	0.003	111	0.53

TABLE 3.2: **Model predictions.** Shown are the Pearson correlations between observed and predicted spectra of adaptive substitutions, their 95% confidence intervals and p-values, the number of non-zero elements in the spectrum of adaptive substitutions (out of 354), and the entropy of the spectrum of adaptive substitutions normalized so that uniformity corresponds to an entropy of 1.

Randomization tests confirm the relevance of empirical mutation spectra for adaptive evolution.

The species-specific mutation spectra employed above reflect either (1) mutation-accumulation experiments under laboratory conditions in the absence of selection (*S. cerevisiae*, *E. coli*), or (2) the frequencies of putatively neutral single-nucleotide polymorphisms in natural populations (*M. tuberculosis*). The observation that the 95 % confidence interval for the inferred values of the coefficient of mutational influence β includes one in all three species highlights the relevance of these species-specific mutation spectra for adaptive evolution.

To explore the relevance of precise estimates of the mutation spectrum more thoroughly, we repeated our regression above 10^6 times for each species, each time using a randomized mutation spectrum instead of the empirical spectrum (each randomized spectrum was generated by drawing 6 random uniform numbers, then normalizing the sum to 1). We then calculated the difference between the log-likelihood of the model fit with the randomized mutation spectrum and the log-likelihood of the model fit with the empirical mutation spectrum. When this difference is positive, the fit using the randomized mutation spectrum explains the spectrum of adaptive substitutions better than the fit using the empirical mutation spectrum, and when this difference is negative the empirical mutation spectrum provides the better explanation. Fig. 3.3a-c shows that the empirical mutation spectra almost always explain the spectra of adaptive substitutions better: randomly generated spectra outperform the observed spectrum with frequency 0.002 for *S. cerevisiae*, 0.037 for *E. coli*, and 0.042 for *M. tuberculosis*. While so far we have attempted to predict the spectrum of adaptive substitutions based on experimentally characterized mutation spectra, the strong relationship between the mutational and adaptive spectra in these three species suggests that it might also be possible to estimate the mutation spectrum from the spectrum of adaptive substitutions. To do this, we again fitted a negative binomial model but treated the rates of the six possible types of single nucleotide mutations as free parameters, which we estimated using maximum likelihood. Fig. 3.3d-f shows that the inferred mutation spectra strongly resemble the experimentally characterized mutation spectra, with Pearson correlation coefficients of 0.945 ($p = 0.004$) for *S. cerevisiae*, 0.960 ($p = 0.002$) for *E. coli*, and 0.837 ($p = 0.038$) for *M. tuberculosis*. Thus, it is possible to accurately recover species-specific mutation spectra directly from species-specific spectra of adaptive substitutions.

What factors influence the predictive power of the model?

Although the analysis above reveals a statistically significant and approximately directly proportional contribution of mutational biases to the spectrum of adaptive substitutions for all three data sets, there is considerable variation in the strength of the correlation between the predicted and observed spectra of adaptive substitutions, with this correlation being strongest and most significant for *S. cerevisiae*, and weakest and least significant for *M. tuberculosis* (Table 3.2; Fig. 3.2d-f).

One immediate hypothesis is that this variation in predictive power is driven by differences in the completeness of our estimates of the spectrum of adaptive substitutions. Even though our data sets include hundreds to thousands of adaptive events per species, a substantial fraction of the 354 possible types of codon-to-amino acid changes are missing from the spectrum for each

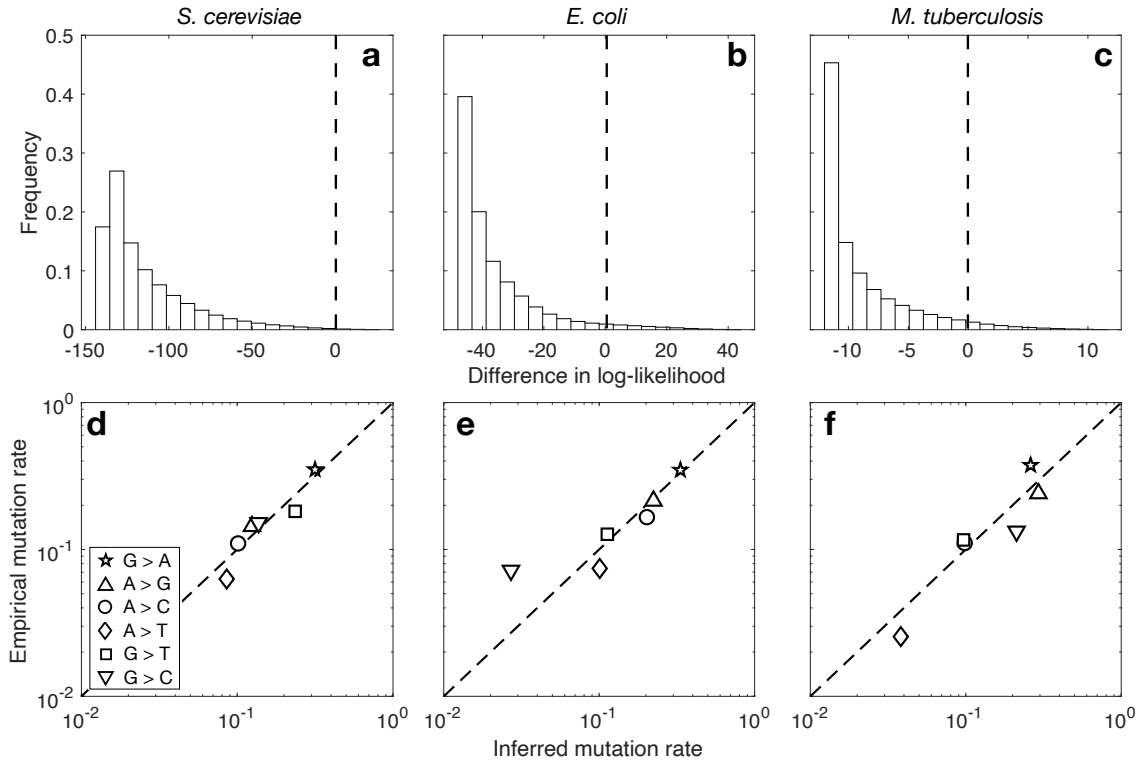

FIGURE 3.3: **Empirical mutation rates explain the spectrum of adaptive substitutions better than randomized rates.** In the upper panels, the white bars show the distribution of log-likelihood differences for randomized vs. empirical mutation rates for (a) *S. cerevisiae*, (b) *E. coli*, and (c) *M. tuberculosis*. A value of 0 (dashed vertical line) means that a randomized rate performs as well as the empirical mutation rate. The fraction of randomized rates providing a better model fit than the empirical rates (i.e., right of 0) is 0.2 %, 3.7 %, 4.2 % for panels a, b and c, respectively. Data based on 10^6 randomized rates. Note that the three panels have different limits on their horizontal axes. In the lower panels, the empirical mutation rate is shown in relation to the inferred mutation rate on a double logarithmic scale for (d) *S. cerevisiae*, (e) *E. coli*, and (f) *M. tuberculosis*. Symbol types correspond to inset in (d). The dashed diagonal line indicates $y = x$.

species (Table 3.2), a situation that likely arises both due to finite sample size effects and the limited diversity of distinct adaptive paths under a specific ecological circumstance (e.g., only a limited number of mutations confer resistance to any given antibiotic). Indeed, we note that at a qualitative level, the smaller the number of missing codon-to-amino acid changes, the stronger the correlation between predicted and observed spectra of adaptive substitutions (Table 3.2). Moreover, when we aggregate the adaptive substitutions into just six types of distinct nucleotide changes, all six types are well represented and there is a strong correlation with the mutation spectrum for all three species (Fig. 3.2a-c).

To evaluate the influence of this kind of sampling effect on the predictive power of our model, we first simulated random data under the codon model assuming $\beta = 1$, sampling adaptive events according to their expected frequencies, based on the empirical codon frequencies and mutation spectrum of each species, but restricting the sampled events to the observed set of non-zero elements for each species-specific spectrum of adaptive substitutions. We then used negative binomial regression to fit this simulated spectrum of adaptive substitutions and measured the correlation between the simulated spectrum of adaptive substitutions and the spectrum of adaptive substitutions predicted by the fitted model. We repeated this process 10^3 times for each species to obtain a distribution of correlations. These distributions are shown in *SI Appendix*, Fig. S3.2. On average, the correlations decreased from *S. cerevisiae* (0.76) to *E. coli* (0.75) to *M. tuberculosis* (0.61), suggesting that sampling effects are partly responsible for differences in model fits between the three species. However, *SI Appendix*, Fig. S3.2 also shows that the correlations for these simulated data sets are considerably stronger than those obtained with models fit to the observed spectra of adaptive substitutions, and the decrease is far less dramatic than the drop from 0.68 to 0.41 to 0.16 noted above (triangles in *SI Appendix*, Fig. S3.2). This suggests that factors other than sampling effects also modulate the predictive power of our modeling framework.

To address a combination of additional factors, we turned to population-genetic simulations of evolution in a haploid genome, with variable parameters for population size N, mutation rate μ, and fraction of beneficial mutations B. The model genome consists of 500 codons subject to neutral synonymous mutations and non-neutral missense mutations, where a fraction B of missense mutational paths are beneficial, with a positive selection coefficient drawn from an exponential distribution, and other missense paths are deleterious, with effects drawn from a reflected gamma distribution (Methods). Note that the inclusion of both advantageous and deleterious mutations allows our simulations to capture both the effects of interference between multiple advantageous mutations (clonal interference) [39], [41] and the effects of selection against linked deleterious alleles (i.e., background selection) [42]. We implemented the simulations in SLiM v3.4 [43]. For each run of the simulation, we recorded the identity of all adaptive mutations on the first sequence to reach fixation, repeating this process 1000 times to produce a simulated spectrum of adaptive substitutions similar in size to our empirical data sets. For each combination of N, μ and B, we simulated 50 data sets and analyzed them using negative binomial regression (Methods).

Previous theoretical work suggests that mutational supply $N\mu$ will modulate the influence of mutational biases on the spectrum of adaptive substitutions [33]–[36], [44]. In particular, the simplest effect of increasing $N\mu$ is that multiple beneficial mutations are typically simultaneously present in the population, competing with each other, so that the adaptive mutation that ultimately fixes in the population is determined more by selective differences between these segregating

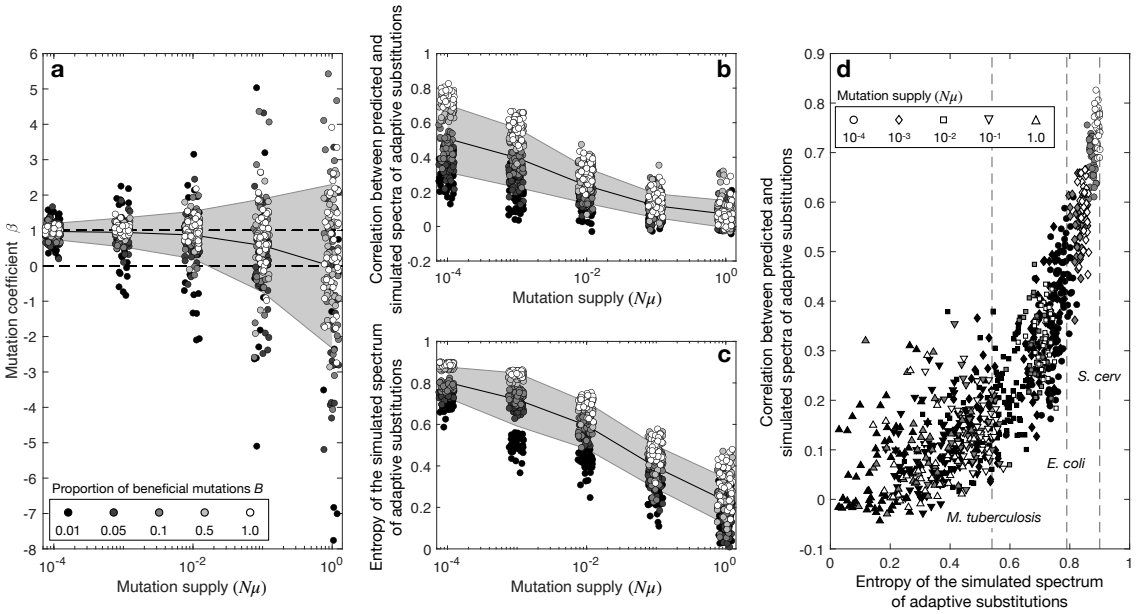

FIGURE 3.4: **Evolutionary simulations show mutation supply and mutational target size jointly modulate the predictive power of our model.** (a) The inferred mutation coefficient β as a function of $N\mu$ for five different values of B, the fraction of beneficial mutations (the same color scheme for B is used in all panels). Dashed horizontal lines are drawn at $\beta = 0$ and $\beta = 1$ to indicate no influence and proportional influence of the mutation spectrum on the spectrum of adaptive substitutions, respectively. (b) Pearson's correlation coefficient between predicted and simulated spectra of adaptive substitutions as a function of $N\mu$ for five different values of B, and (c) entropy of simulated spectra of adaptive substitutions as a function of $N\mu$ for five different values of B. In (a-c), the black lines show the mean and the gray areas show the standard deviation. (d) The Pearson's correlation coefficient between predicted and simulated spectra of adaptive substitutions is shown in relation to the entropy of the simulated spectra of adaptive substitutions for different levels of mutation supply. The dashed vertical lines show the entropy of the spectrum of adaptive substitutions for each of our three study species.

mutations than by which beneficial mutation becomes established in the population first. This expectation is confirmed by Fig. 3.4a, which shows the inferred values of β relative to $N\mu$ for different proportions of beneficial mutations B. At the lowest mutation supply, β is approximately one, reflecting the direct proportionality expected for the origin-fixation regime [33], [37]. As the mutation supply increases, β tends toward zero, reflecting a diminished influence of the mutation spectrum on adaptation. At the same time, the distribution of estimates for β becomes more dispersed (Fig. 3.4a), and the individual estimates become both less significant and less certain, as indicated by increasing average p-values and increasingly large confidence intervals (*SI Appendix*, Fig. S3.3). Similarly, the predictive power of the model decreases with increasing mutation supply, as measured by a decreasing average correlation between the predicted and observed spectra of adaptive substitutions (Fig. 3.4b).

The fraction of beneficial mutations B also influences the predictive power of the fitted models, but in a somewhat more surprising manner. Intuitively, one might think that increasing the proportion of beneficial mutations would decrease predictive power, as increasing B effectively increases the beneficial mutational supply, allowing increased competition between simultaneously segregating beneficial mutations. However, Fig. 3.4a and b show the opposite pattern. At low and intermediate levels of mutation supply, the largest values of B (white dots) yield the best

correlations, the lowest values of B (black dots) yield the worst correlations, and intermediate values of B (grey dots) are intermediate. At high mutation supply, all of the correlations are poor regardless of B.

A potential explanation for this unexpected effect of B relates to the way that biases in nucleotide mutations have relatively broad effects, in the sense that changing a single nucleotide mutation rate will affect the rates of $\sim$60 codon-to-amino acid changes. Because nucleotide mutational biases thus enrich broad classes of codon-to-amino acid changes, they will tend to perform poorly in predicting distributions of adaptive events when those distributions are highly concentrated on a small set of codon-to-amino acid changes. Increasing B expands the set of possible beneficial mutations to cover more diverse types of changes at various genomic sites, and this effect may be expected to improve the correlation of predicted and observed changes. Indeed, weak correlations due to this effect might arise, not only from having relatively few available adaptive paths in a given selective environment (small B), but also from limited sampling density, or even from a broad and well sampled distribution of adaptive substitutions that is nonetheless heavily skewed toward a small number of strongly favored changes.

To quantify both the breadth of the adaptive spectrum (i.e., the distribution of events across the non-zero elements of the spectrum of adaptive substitutions) and its effects on the predictive power of our model, we calculated the entropy of observed and simulated spectra of adaptive substitutions, normalized so that the entropy has a minimum value of 0 when all adaptive events correspond to a single codon-to-amino acid change, and a maximum value of 1 when the adaptive events are uniformly distributed across all possible codon-to-amino acid changes (Methods). Fig. 3.4c shows that the entropy decreases as mutation supply increases, and that for any level of mutation supply, a lower proportion of beneficial mutations likewise decreases the entropy. To determine whether these patterns of decreasing entropy are sufficient to explain differences in the predictive power of our model across the range of model parameters, we plotted the correlation between predicted and simulated spectra of adaptive substitutions against the entropy of the simulated spectrum of adaptive substitutions (Fig. 3.4d). We see that increasing entropy, either via a decreased mutation supply or an increased proportion of beneficial mutations, increases the correlation between simulated and predicted spectra of adaptive substitutions. These observations from the evolutionary simulations are qualitatively similar to our empirical observation that as the entropy of the spectrum of adaptive substitutions increases from *M. tuberculosis* to *E. coli* to *S. cerevisiae*, there is a corresponding increase in the correlation between predicted and observed spectra of adaptive substitutions (Table 3.1). Indeed, the correlations for our three empirical data sets are well within the range of what we would expect from our simulations given their respective entropies (Fig. 3.4d).

To summarize, the results from our evolutionary simulations show that the predictive power of our model is strongest when the mutation supply is low and the mutational target size is large. However, we note that predictive power might also be influenced by other factors not included in our simulations, e.g., heterogeneity in the mutation rate across the genome, such as that caused by local sequence context [15], [45]–[50].

Assessing possible effects of contamination.

A key assumption of the analysis above is that the events used to populate the spectrum of adaptive codon-to-amino acid changes represent adaptive substitutions. While this is likely the case for the *M. tuberculosis* data set, because these mutations have been shown experimentally to confer antibiotic resistance [5], we now consider the possibility that some fraction of observations in the *S. cerevisiae* and *E. coli* data sets represent contamination such as hitchhikers. If contaminants reflect the mutation spectrum more than genuine adaptive changes, this will exaggerate the correspondence with mutational predictions. Using the method of Tenaillon, et al. [8], based on the observed dN/dS ratio (Methods), we estimate these proportions to be $\sim$24% and $\sim$13% for *S. cerevisiae* and *E. coli*, respectively.

To assess the influence of contamination up to, and even beyond, these estimated levels, we randomly remove a fraction q of events, sampled according to the species-specific empirical mutation spectrum. This procedure simulates the removal of a hypothetical contaminant fraction of size q under the worst-case scenario in which the nucleotide changes in the contaminant fraction mirror the mutation spectrum. As shown in *SI Appendix*, Fig. S3.4, even under the assumption that 40 % of the events are contaminants, we observe a strong and statistically significant influence of mutation bias on adaptive evolution. In fact, we estimate that for *S. cerevisiae* and *E. coli*, levels of contamination of $\sim$65 % and $\sim$44 %, respectively, would be required to increase the p-value of β to the point where the influence of mutation bias would no longer be significant.

3.4 DISCUSSION

A growing body of evidence suggests that specific mutation biases influence the types of genetic changes involved in adaptation [5], [19]–[27], consistent with a small body of theoretical work on how biases in the introduction of variation — both low-level mutational biases and higher-level systemic biases — are expected to influence adaptive evolution [33], [35], [36], [40]. Yet a general approach for quantifying this influence was missing. Here, we have developed and applied such a general approach to assess how the entire mutation spectrum shapes the spectrum of adaptive substitutions. It uses negative binomial regression to model the spectrum of adaptive substitutions as a function of codon frequencies and the mutation spectrum, measuring the influence of mutation in terms of a single statistic — the coefficient of mutational influence β.

This statistic takes on a value of zero when the mutation spectrum has no influence, a value of one for a proportional influence, and intermediate values for intermediate degrees of influence. Applying this framework to large data sets from *Saccharomyces cerevisiae*, *Escherichia coli*, and *Mycobacterium tuberculosis*, we find a clear signal that the mutation spectrum strongly shapes the spectrum of adaptive substitutions. Specifically, the inferred values of β are not significantly different from one in any species. This result holds even when we account for the contamination by hitchikers that is likely present in the data sets for *S. cerevisiae* and *E. coli*.

Our approach also illustrates how the spectrum of adaptive substitutions may be interrogated to reveal clues about the genetic basis of adaptation. We used our fitted models to predict the spectrum of adaptive substitutions in each species, and uncovered variation in their predictive capacity, decreasing from *S. cerevisiae* to *E. coli* to *M. tuberculosis*. Using evolutionary simulations,

we uncovered multiple potential sources of this variation. Specifically, we found that the degree to which the mutation spectrum is a good predictor of the spectrum of adaptive substitutions depends on how the adaptive events are distributed among all possible codon-to-amino acid changes, with reduced predictive capacity associated with distributions concentrated on a small number of codon-to-amino acid changes. Factors that affect the degree of concentration include data set size, population genetic conditions, diversity of selective environments, and the genetic architecture of adaptive traits. Importantly, population genetic conditions that modulate the influence of mutation bias on adaptation, such as mutation supply, and non-population genetic conditions, such as the diversity of environmental conditions included in the data set, can affect the predictive capacity of our model in similar ways.

While additional work is needed to disambiguate these various causes of differing model fits between species, our results are consistent with known facts concerning the population-genetic conditions, as well as the environmental conditions and mutational target sizes for adaptive mutations for the three species studied here. *M. tuberculosis* has one of the lowest mutation supplies of all bacteria [51], a small population size upon infection [52], and the 11 antibiotics considered here target specific gene products [5]. For example, Rifampicin targets the beta subunit of bacterial RNA polymerase, and only a small handful of mutations to the *rpoB* gene that encodes this subunit cause resistance [53]. Thus, while the population genetic conditions of *M. tuberculosis* are more likely similar to origin-fixation dynamics than clonal interference dynamics, and the set of observations is large, the mutational target size for antibiotic resistance is small. In contrast, *E. coli* experiences clonal interference due to a relatively higher mutation supply [54], but adaptation to temperature stress involves a larger mutational target [8], [55]. Similarly, *S. cerevisiae* experiences clonal interference due to a high mutation supply [29], but because the data we study include adaptation to several environmental conditions, the mutational target size is large. Thus, the inferred influence of mutation bias on adaptation in these three species, increasing from *M. tuberculosis* to *E. coli* to *S. cerevisiae*, is consistent with our findings from evolutionary simulations that mutation supply and mutational target size modulate the influence of mutation bias on adaptation. However, it may also be the case that the diminished influence of mutation bias in *M. tuberculosis*, relative to *E. coli* and *S. cerevisiae*, results from differences in the way the data were collected (clinical isolates vs. laboratory evolution experiments).

The three species studied here also share several important features that suggest a need for similar studies across a greater diversity of population-genetic conditions. For example, all of the data analyzed here were obtained either from clonally reproducing experimental populations (*E. coli* and *S. cerevisiae*) or, in the case of *M. tuberculosis*, from natural populations with little or no recombination [52], [56], [57]. This absence of recombination amplifies both the role of background selection [42] and the degree of interference between selected alleles [41], and it remains an open question whether mutational biases in practice play as large a role in sexual populations. Another important population-genetic commonality across the datasets studied here is the low degree of genetic diversity prior to the onset of selection, so that adaptation likely proceeds in all three systems from new mutations rather than standing genetic variation. This low initial diversity is the result of either the experimental setup in the case of *E. coli* and *S. cerevisiae*, or the low world-wide nucleotide diversity empirically observed for *M. tuberculosis* [56], which is likely due to repeated bottlenecks at transmission events as well as other factors [58].

The discovery that mutation biases strongly shape the spectrum of adaptive substitutions has implications for several other related issues in evolutionary genetics. First, it has implications for the predictability of evolution [59]–[61], because it shows that mutationally favored types of changes are more likely to contribute to evolutionary adaptation, an effect that is both large and readily predictable from prior data on the relevant mutation spectrum. When the spectrum of adaptive substitutions is compared to the mutation spectrum, we see a significant correlation (of variable strength) for the spectrum of codon-to-amino-acid changes, and a consistently strong correlation for the six types of nucleotide changes. This can be understood as an effect of aggregation. Many previous studies based on laboratory evolution experiments show that aggregating distinct genomic paths of adaptation by functional criteria (e.g., shared gene, operon, or functional category) highlight predictable effects that are presumably effects of selection [8], although effects of mutation are also evident at the gene level [44]. The extreme aggregation of distinct genomic changes into just six types of nucleotide changes also reveals a highly predictable effect, but it is an effect of mutation rather than selection, because the criterion of aggregation is the mutational type. At the opposite extreme of aggregation — particular nucleotide changes at specific genomic coordinates — mutation bias is unlikely to be predictive of the genetic changes that cause adaptation.

Second, the discovery of a direct influence of mutation bias on evolutionary adaptation parallels recent reports that driver mutations in cancer reflect the underlying biases of cancer-associated mutational processes, including exogenous effects of UV light and tobacco exposure, and endogenous effects of DNA mismatch repair and APOBEC activity [62]–[64]. The increased predictability of such changes, due to mutational effects, can inform rational drug design, as has been suggested for drugs for leukemia, prostate cancer, breast cancer, and gastrointestinal stromal tumors [26]. The same may be true for designing antibiotic treatments for mycobacteria, which evolve multi-drug resistance via a sequence of mutations, several of which interact epistatically, such that only a subset of possible mutational trajectories to multi-drug resistance are possible [65].

Finally, the broadest context for the present work is a debate about the role of so-called "internal" causes in shaping the course of evolution. Arguments dating back to the origins of theoretical population genetics emphasize selection as the sole directional force in evolution, with mutation treated as a weak and ineffectual pressure due to the smallness of mutation rates [66]–[68]. Haldane concluded that mutation can influence the course of evolution only under neutral evolution, or when mutation rates are unusually high [66]. Accordingly, strong effects of mutation bias have been historically associated with neutral evolution (see [69]). However, more recent theoretical work has shown that this classic way of thinking depends on the assumption that evolution begins with abundant standing genetic variation, so that mutation acts only as a frequency-shifting force and not as a source of genetic novelty [33]. When the dynamics of an evolutionary process depend on events that introduce novel variants, biases in the introduction process, such as toward particular nucleotide changes, systematically influence which type of genetic changes are involved in adaptation [33], [70].

A variety of statistical frameworks assume a proportional influence of the mutation spectrum on the spectrum of adaptive substitutions, including those for quantifying selection pressures on proteins. For example, the ratio of non-synonymous to synonymous mutations (dN/dS) - a commonly used statistical test to detect proteins undergoing adaptation - is often corrected to

account for the mutation spectrum [54], [71]. Implicit in this accounting is the assumption that the mutation spectrum influences neutral and adaptive mutations in the same way. Our finding that the mutation spectrum can be directly inferred from the spectrum of adaptive substitutions provides empirical support for this assumption, at least for the species and evolutionary conditions considered here.

Some have responded to the theory of mutation-biased adaptation by arguing that such an influence is unlikely, on the grounds of requiring sign epistasis or unusually small population sizes [72]. However, modeling here and in other work shows that mutation bias can influence adaptation across a range of conditions, including in the absence of sign epistasis and when conditions induce clonal interference among concurrent mutations [35]. More broadly, while theoretical arguments are surely helpful for sharpening our understanding, ultimately the prevalence and magnitude of the mutational influence on adaptation is an empirical question, and the impact of mutational biases on adaptation has now been shown for several different types of mutations, in a range of systems from bacteriophage to birds to somatic evolution in human cancers [5], [19]–[27].

This growing body of work on mutation-biased adaptation provides a basis to reconsider certain long-standing claims about how variational properties influence the evolutionary process. For instance, evo-devo arguments about bias or constraint relate evolutionary patterns to tendencies of developmental variation, but the causal nature of this link, in terms of population-genetic principles, is typically unspecified (e.g., [73], [74]). Likewise, a significant body of neo-structuralist work on "findability" or "self-organization", going back at least to Kauffman [75], emphasizes the tendency of evolution to prefer structures common in abstract state-spaces, e.g., in regard to RNA folds [76] or regulatory circuit motifs [77], without linking this effect to a population-genetic cause. Recent work on mutation-biased adaptation provides a rigorous body of theory and evidence establishing how tendencies of variation may act as dispositional causes in evolution, suggesting a previously missing population-genetic basis for these long-standing claims. Our results contribute to the empirical case that mutational biases, which are more accessible to study at the level of population genetics, have a strong and measurable impact on adaptive evolution.

3.5 METHODS

Data.

Our modeling framework is built around three key quantities, which are specific to each species: A spectrum of adaptive substitutions $\mathbf{n}$, a table of codon frequencies f, and a mutation spectrum μ. These are all constructed using empirical data, as described below.

Spectrum of adaptive substitutions.

We curated a list of missense mutations associated with adaptation from the published literature for each of three species: *S. cerevisiae*, *E. coli*, and *M. tuberculosis*. For each mutation, these lists specify a genomic coordinate, nucleotide change, amino acid substitution, and literature reference (Datasets S1-S7). We refer to each unique combination of genomic coordinate and nucleotide change as a mutational path and each instance of adaptive change along a mutational path as an

adaptive event. The number of adaptive events per mutational path are also reported in Datasets S1-S7.

For *S. cerevisiae*, the adaptive events were reported in four studies, each of which considered one or more environmental or genetic challenges, including high salinity [28], low glucose [28], rich media [29], and gene knockout [30]. The list contains 713 adaptive events across 534 mutational paths (Dataset S1).

For *E. coli*, the adaptive events were reported in a single study of 115 replicate populations adapting to temperature stress [8]. The list contains 602 adaptive events across 492 mutational paths (Dataset S2).

For *M. tuberculosis*, the adaptive events were reported in a single study of the influence of mutation bias on adaptation to antibiotic stress [5]. The underlying mutational paths were derived from two separate meta-analysis of the literature on antibiotic resistance (one performed for the study and another previously published [4]), with each mutational path required to pass stringent tests for conferring antibiotic resistance. A total of 11 antibiotics or antibiotic classes were considered: Rifampicin, ethambutol, isoniazid, ethionamide, ofloxacin, pyrazinamide, streptomycin, kanamycin, pyrazinamide, fluoroquinolones, and aminoglycosides. The adaptive events were inferred from a phylogenetic reconstruction of public *M. tuberculosis* genomes. We merged the adaptive events from the two meta-analyses. The resulting list contains 4413 adaptive events across 283 mutational paths (Dataset S3). Analyzing the adaptive events from the two meta-analyses separately (*SI Appendix*, Table S3.1) produced qualitatively similar results to those reported in Table 3.1.

For each species, we constructed the spectrum of adaptive substitutions $\mathbf{n}$ from the list of adaptive events described above, assigning each adaptive event to its respective codon-to-amino-acid change. Each element $\mathbf{n}(c, a)$ of the spectrum of adaptive substitutions therefore tallies the number of adaptive events that changed codon c to amino acid a. Note the adaptive events tallied for any codon-to-amino-acid change often reflect more than one genomic coordinate and/or nucleotide change (i.e., different mutation paths). These spectra are reported in Dataset S4.

Codon frequencies.

We used the tables of codon frequencies reported in the Codon Usage Database [78], found via query to an exact match to *Saccharomyces cerevisiae*, *Escherichia coli*, and *Mycobacterium tuberculosis*. These frequencies are reported in Dataset S5 and shown in *SI Appendix*, Fig. S3.1e-g.

Empirical mutation spectra.

For *S. cerevisiae* and *E. coli*, we used mutation rates derived from mutation accumulation experiments, as reported in Figure 3 of reference [15] and Table 3 of reference [14], respectively. For *E. coli*, we corrected the mutation rates for GC content, following [12]. For *S. cerevisiae*, the rates were already corrected [15]. For *M. tuberculosis*, we used mutation rates derived from single-nucleotide polymorphism data [5] (Dataset S6). We restricted our analysis to synonymous mutations in the 3rd codon position, and corrected the rates for GC content in that position. We also corrected for the probability that each type of mutation causes a synonymous change. For instance, of all the possible synonymous mutations in the 3rd position allowed by the standard genetic code, 23% are G/C→A/T transitions, whereas only 12% are G/C→C/G transversions.

These spectra are reported in Dataset S7 and shown in *SI Appendix*, Fig. S3.1a. We used these estimated mutation rates to define a total codon-to-amino acid mutation rate $\mu(c, a)$ for each of the 354 codon-to-amino acid changes allowed by the standard genetic code, summing the rates of all point mutations in codon c that lead to amino acid a. For example, the probability of the mutation from codon CAC to Glutamine (Q) is the sum of the probabilities of point mutations C→A and C→G, since both mutations in the third position of CAC lead to codons for Glutamine (Q).

Transition-transversion ratio vs. the full mutation spectrum

The influence of the mutation spectrum can be partitioned into an overall transition-transversion bias, and biases among different types of transitions and transversions. The model that only considers the contribution of the species-specific transition-transversion bias is given by:

$$\log \mathbb{E}[\mathbf{n}(c, a)] = \beta_0 + \log f(c) + \beta_{\text{ti/tv}} \log \mu_{\text{ti/tv}}(c, a). \tag{3.3}$$

As in Eq.3.2, β_0 is the logarithm of the constant of proportionality and $f(c)$ is the genomic frequency of codon c. The mutation term $\mu_{\text{ti/tv}}(c, a)$ is defined only by the species-specific transition-transversion ratio, and thus assigns one rate to all transitions and one (different) rate to all transversions. The corresponding regression coefficient is $\beta_{\text{ti/tv}}$.

The complete model contains all of the terms of the model above (Eq. 3.3), with an additional mutation term $\mu_{\text{rest}}(c, a)$ that accounts for the rest of the mutation spectrum (such that $\mu_{\text{ti/tv}}(c, a)\mu_{\text{rest}}(c, a) = \mu(c, a)$), along with its respective regression coefficient β_{rest}. This complete model is given by:

$$\begin{aligned} \log \mathbb{E}[\mathbf{n}(c, a)] = \beta_0 + \log f(c) + \beta_{\text{ti/tv}} \log \mu_{\text{ti/tv}}(c, a) \\ + \beta_{\text{rest}} \log \mu_{\text{rest}}(c, a). \end{aligned} \tag{3.4}$$

As in our main analyses, we used negative binomial regression to estimate the regression coefficients. Because the two models are nested, we compared their performance using a likelihood ratio test (*SI Appendix*, Table S3.2).

Entropy of the spectrum of adaptive substitutions.

The spectrum of adaptive substitutions $\mathbf{n}$ describes the number of adaptive events per codon-to-amino acid change. We calculate the entropy H of this spectrum as

$$H = \frac{-\sum_{i=1}^{m} p(n_i) \log p(n_i)}{\log(m)} \tag{3.5}$$

where $p(n_i)$ is the proportion of adaptive events that correspond to the ith codon-amino acid change, and $m = 354$ is the number of codon-to-amino acid changes allowed by the standard genetic code.

Evolutionary simulations.

We used SLiM v3.4 for the evolutionary simulations [43]. We ran each simulation until the first fixation event, repeating this process 1000 times and recording each beneficial mutation that went to fixation. We performed 50 replicates per combination of the parameters N, μ, and B. Each of the 1000 simulations per replicate used the same initial population, which comprised N copies of a nucleotide sequence of length $L = 1500$ (i.e., 500 codons), randomly generated using the codon frequencies for *S. cerevisiae*.

All sequences in the initial population were assigned a fitness of one. The fitness effects assigned to each of the possible codon-to-amino acid changes from each of the 500 codons were drawn at random from a distribution of fitness effects, and were held constant across the 1000 simulations per replicate.

A unique distribution of fitness effects was constructed for each replicate, such that synonymous mutations were neutral, a fraction B of missense codon-to-amino acid changes were beneficial, and a fraction $1 - B$ of missense codon-to-amino acid changes were deleterious. The fitness effects of beneficial codon-to-amino acid changes were drawn from an exponential distribution with density

$$f_b(x) = \lambda e^{-\lambda x} \tag{3.6}$$

where $\lambda = 33.33$, so that the expected advantageous selection coefficient was 0.03. The fitness effects of deleterious codon-to-amino acid changes were drawn from a gamma distribution with density

$$f_d(x) = \frac{x^{(a-1)} e^{-(x/s)}}{s^a \, \Gamma(a)} \tag{3.7}$$

where $a = 0.4$ and $s = 0.15$, so that the magnitude of the expected deleterious selection coefficient was twice the advantageous one [79]. For sequences with more than one mutation, we summed the effects of the individual mutations. *SI Appendix*, Fig. S3.5 shows representative distributions of fitness effects for different proportions of beneficial mutations B.

Each simulation proceeded until a single sequence went to fixation and any beneficial mutations were recorded. Our simulations thus correspond to single-step adaptive walks, extending prior theoretical work considering just a few possible adaptive mutations [19], [33] into a codon-based model of a whole gene with thousands of possible mutations. Single-step adaptive walks are particularly germane to the *M. tuberculosis* data, in which antibiotic resistance is often strongly associated with single mutations. Multi-step walks are also relevant for long-term evolution, but they would require further assumptions about the structure of the fitness landscape. In each generation t, N sequences were chosen from the population at generation $t - 1$ with replacement and with a probability proportional to their fitness. Mutations were introduced according to the product of the genome-wide mutation rate μ and the per-nucleotide mutation rate defined by the mutation spectrum for *S. cerevisiae*, with each mutation affecting fitness as defined at the onset of the simulation.

Contamination estimates.

For each type of mutation, we calculated the number of synonymous and non-synonymous sites for each possible codon, and we estimated the total number of synonymous and non-synonymous sites in the genome by taking into account the codon usage patterns of *S. cerevisiae* and *E. coli* (*SI Appendix*, Fig. S3.1e-f). We then calculated dN/dS ratios among all substitutions in the adapted lines correcting for the mutation rates of each type of mutation (*SI Appendix*, Fig. S3.1a). Following [8], we estimated the proportion of adaptive non-synonymous mutations from such ratios as $y = (x - 1.0)/x$, where x is the estimated dN/dS ratio (4.24 and 7.76 for *S. cerevisiae* and *E. coli*, respectively). Finally, we estimated the fraction of hitch-hikers in our data sets as $1 - y$.

Data Availability.

All study data are included in the article and/or supporting information. The scripts used to analyze these data and to run the evolutionary simulations can be found at https://github.com/alejvcano/Mutbias2022.

3.6 ACKNOWLEDGMENTS

The identification of any specific commercial products is for the purpose of specifying a protocol, and does not imply a recommendation or endorsement by the National Institute of Standards and Technology. This project / publication was made possible through the support of a grant from the John Templeton Foundation (grant #61782, D.M.M.) and from the Swiss National Science Foundation (grants #PP00P3_170604 and #310030_192541, J.L.P.). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. D.M.M. also acknowledges additional support from an an Alfred P. Sloan Research Fellowship and from the Simons Center for Quantitative Biology. We thank Fabrizio Menardo for assistance with the *M. tuberculosis* polymorphism data and the reviewers for their helpful comments.

REFERENCES

[1] S. Yokoyama and F. B. Radlwimmer, "The molecular genetics and evolution of red and green color vision in vertebrates", *Genetics*, vol. 158, no. 4, pp. 1697–710, 2001.

[2] B. Ujvari, N. R. Casewell, K. Sunagar, *et al.*, "Widespread convergence in toxin resistance by predictable molecular evolution", *Proceedings of the National Academy of Sciences*, vol. 112, no. 38, pp. 11 911–6, 2015.

[3] C. Natarajan, J. Projecto-Garcia, H. Moriyama, *et al.*, "Convergent evolution of hemoglobin function in high-altitude Andean waterfowl involves limited parallelism at the molecular sequence level", *PLoS Genetics*, vol. 11, no. 12, e1005681, 2015.

[4] A. Manson, K. Cohen, T. Abeel, *et al.*, "Genomic analysis of globally diverse *Mycobacterium tuberculosis* strains provides insights into the emergence and spread of multidrug resistance", *Nature Genetics*, vol. 49, pp. 395–402, 2017.

[5] J. L. Payne, F. Menardo, A. Trauner, *et al.*, "Transition bias influences the evolution of antibiotic resistance in Mycobacterium tuberculosis", *PLoS Biology*, vol. 17, no. 5, 2019.

[6] W. Liu, D. K. Harrison, D. Chalupska, *et al.*, "Single-site mutations in the carboxyltransferase domain of plastid acetyl-coa carboxylase confer resistance to grass-specific herbicides", *Proceedings of the National Academy of Sciences*, vol. 104, no. 9, pp. 3627–32, 2007.

[7] J. R. Meyer, D. T. Dobias, J. S. Weitz, J. E. Barrick, R. T. Quick, and R. E. Lenski, "Repeatability and contingency in the evolution of a key innovation in phage lambda", *Science*, vol. 335, no. 6067, pp. 428–32, 2012.

[8] O. Tenaillon, A. Rodríguez-Verdugo, R. L. Gaut, *et al.*, "The molecular diversity of adaptive convergence", *Science*, 2012.

[9] R. M. Schaaper and R. L. Dunn, "Spectra of spontaneous mutations in *Escherichia coli* strains defective in mismatch correction: The nature of *in vivo* DNA replication errors", *Proceedings of the National Academy of Sciences*, vol. 84, pp. 6220–6224, 1987.

[10] Z. Zhang and M. Gerstein, "Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes", *Nucleic Acids Research*, vol. 31, pp. 5338–48, 2003.

[11] P. Keightley, U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. Blaxter, "Analysis of the genome sequences of 3 *Drosophila melanogaster* spontaneous mutation accumulation lines", *Genome research*, vol. 19, pp. 1195–201, 2009.

[12] R. Hershberg and D. A. Petrov, "Evidence that mutation is universally biased towards AT in bacteria", *PLoS Genetics*, 2010.

[13] S. Ossowski, K. Schneeberger, J. Lucas-Lledó, *et al.*, "The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*", *Science*, vol. 327, pp. 92–4, 2010.

[14] H. Lee, E. Popodi, H. Tang, and P. L. Foster, "Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing", *Proceedings of the National Academy of Sciences*, 2012.

[15] Y. O. Zhu, M. L. Siegal, D. W. Hall, and D. A. Petrov, "Precise estimates of mutation rate and spectrum in yeast", *Proceedings of the National Academy of Sciences*, 2014.

[16] S. Kucukyildirim, H. Long, W. Sung, S. Miller, T. Doak, and M. Lynch, "The rate and spectrum of spontaneous mutations in *Mycobacterium smegmatis*, a bacterium naturally devoid of the post-replicative mismatch repair pathway", *G3*, vol. 6, pp. 2157–2163, 2016.

[17] M. D. Pauly, M. C. Procario, and A. S. Lauring, "A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses", *eLife*, vol. 6, e26437, 2017.

[18] V. Katju and U. Bergthorsson, "Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches", *Genome Biology and Evolution*, vol. 11, no. 1, pp. 136–165, 2019.

[19] D. Rokyta, P. Joyce, S. Caudle, and H. Wichman, "An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus", *Nature Genetics*, vol. 37, pp. 441–444, 2005.

[20] C. Maclean, G. Perron, and A. Gardner, "Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for Rifampicin resistance in *Pseudomonas aeruginosa*", *Genetics*, vol. 186, pp. 1345–54, 2010.

[21] A. Couce, A. Rodríguez-Rojas, and J. Blazquez, "Bypass of genetic constraints during mutator evolution to antibiotic resistance", *Proceedings of the Royal Society London B*, vol. 282, p. 20 142 698, 2015.

[22] A. M. Sackman, L. W. McGee, A. J. Morrison, *et al.*, "Mutation-driven parallel evolution during viral adaptation", *Molecular Biology and Evolution*, vol. 34, no. 12, pp. 3243–3253, 2017.

[23] A. Stoltzfus and D. M. McCandlish, "Mutational biases influence parallel adaptation", *Molecular Biology and Evolution*, vol. 34, no. 9, pp. 2163–2172, 2017.

[24] J. F. Storz, C. Natarajan, A. V. Signore, C. C. Witt, D. M. McCandlish, and A. Stoltzfus, "The role of mutation bias in adaptive molecular evolution: Insights from convergent changes in protein function", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, no. 1777, p. 20 180 238, 2019.

[25] F. Bertels, C. Leemann, K. J. Metzner, and R. R. Regoes, "Parallel evolution of HIV-1 in a long-term experiment", *Molecular Biology and Evolution*, vol. 36, no. 11, pp. 2400–2414, 2019.

[26] S. Leighow, C. Liu, H. Inam, B. Zhao, and J. Pritchard, "Multi-scale predictions of drug resistance epidemiology identify design principles for rational drug design", *Cell Reports*, vol. 30, pp. 3951–3963, 2020.

[27] S. Katz, S. Avrani, M. Yavneh, H. S., J. Gross, and H. R., "Dynamics of adaptation during three years of evolution under long-term stationary phase", *Molecular Biology and Evolution*, 2021, In press.

[28] L. M. Kohn and J. B. Anderson, "The underlying structure of adaptation under strong selection in 12 experimental yeast populations", *Eukaryotic Cell*, vol. 13, no. 9, pp. 1200–1206, 2014.

[29] G. I. Lang, D. P. Rice, M. J. Hickman, *et al.*, "Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations", *Nature*, vol. 500, no. 7464, pp. 571–574, 2013.

[30] B. Szamecz, G. Boross, D. Kalapis, *et al.*, "The Genomic Landscape of Compensatory Evolution", *PLoS Biology*, vol. 12, no. 8, 2014.

[31] S. Yu, S. Girotto, C. Lee, and R. Magliozzo, "Reduced affinity for Isoniazid in the S315T mutant of *Mycobacterium tuberculosis* KatG is a key factor in antibiotic resistance", *The Journal of Biological Chemistry*, vol. 278, pp. 14 769–14 775, 2003.

[32] P. McCullagh and J. Nelder, *Generalized Linear Models, Second Edition*, ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[33] L. Y. Yampolsky and A. Stoltzfus, "Bias in the introduction of variation as an orienting factor in evolution", *Evolution & Development*, vol. 3, no. 2, pp. 73–83, 2001.

[34] A. Stoltzfus, "Mutation-biased adaptation in a protein NK model", *Molecular Biology & Evolution*, vol. 23, pp. 1852–1862, 2006.

[35] K. Gomez, J. Bertram, and J. Masel, "Mutation bias can shape adaptation in large asexual populations experiencing clonal interference", *Proceedings of the Royal Society B: Biological Sciences*, vol. 287, no. 1937, p. 20 201 503, 2020.

[36] A. V. Cano and J. L. Payne, "Mutation bias interacts with composition bias to influence adaptive evolution", *PLOS Computational Biology*, vol. 16, pp. 1–26, Sep. 2020.

[37] D. McCandlish and A. Stoltzfus, "Modeling evolution using the probability of fixation: History and implications", *Quarterly Review of Biology*, vol. 89, no. 3, pp. 225–252, 2014.

[38] J. H. Gillespie, "A simple stochastic gene substitution model", *Theor Popul Biol*, vol. 23, no. 2, pp. 202–15, 1983.

[39] P. J. Gerrish and R. E. Lenski, "The fate of competing beneficial mutations in an asexual population", *Genetica*, vol. 102, pp. 127–144, 1998.

[40] A. Stoltzfus, "Mutationism and the dual causation of evolutionary change", *Evolution & Development*, vol. 8, pp. 304–317, 2006.

[41] R. A. Neher, "Genetic draft, selective interference, and population genetics of rapid adaptation", *Annual review of Ecology, evolution, and Systematics*, vol. 44, pp. 195–215, 2013.

[42] B. Charlesworth, M. Morgan, and D. Charlesworth, "The effect of deleterious mutations on neutral molecular variation.", *Genetics*, vol. 134, no. 4, pp. 1289–1303, 1993.

[43] P. W. Messer, *SLiM: Simulating evolution with selection and linkage*, 2013.

[44] S. F. Bailey, F. Blanquart, T. Bataillon, and R. Kassen, "What drives parallel evolution?: How population size and mutational variation contribute to repeated evolution", *Bioessays*, vol. 39, no. 1, pp. 1–9, 2017.

[45] R. D. Blake, S. T. Hess, and J. Nicholson-Tuell, "The influence of nearest neighbors on the rate and pattern of spontaneous point mutations", *J Mol Evol*, vol. 34, no. 3, pp. 189–200, 1992.

[46] M. Krawczak, E. V. Ball, and D. N. Cooper, "Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes", *American journal of human genetics*, vol. 63, no. 2, pp. 474–88, 1998.

[47] A. Hodgkinson and A. Eyre-Walker, "Variation in the mutation rate across mammalian genomes", *Nature Reviews Genetics*, vol. 12, no. 11, pp. 756–766, 2011.

[48] W. Sung, M. S. Ackerman, J.-F. Gout, *et al.*, "Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation–Accumulation Experiments", *Molecular Biology and Evolution*, vol. 32, no. 7, pp. 1672–1683, 2015.

[49] J. W. Schroeder, W. G. Hirst, G. A. Szewczyk, and L. A. Simmons, "The effect of local sequence context on mutational bias of genes encoded on the leading and lagging strands", *Curr Biol*, vol. 26, no. 5, pp. 692–7, 2016.

[50] V. Katju, A. Konrad, T. C. Deiss, and U. Bergthorsson, "Mutation rate and spectrum in obligately outcrossing Caenorhabditis elegans mutation accumulation lines subjected to RNAi-induced knockdown of the mismatch repair gene msh-2", *G3*, 2021.

[51] V. Eldholm and F. Balloux, "Antimicrobial resistance in *Mycobacterium tuberculosis*: The odd one out", *Trends in Microbiology*, vol. 24, pp. 637–648, Apr. 2016.

[52] S. Gagneux, "Ecology and evolution of *Mycobacterium tuberculosis*", *Nature Reviews Microbiology*, vol. 16, no. 4, pp. 202–213, 2018.

[53] C. Ford, R. Shah, M. Maeda, *et al.*, "*Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis", *Nature Genetics*, vol. 45, pp. 784–790, Jun. 2013.

[54] B. Good, M. Mcdonald, J. Barrick, R. Lenski, and M. Desai, "The dynamics of molecular evolution over 60,000 generations", *Nature*, vol. 551, pp. 45–50, 2017.

[55] D. Deatherage, J. Kepner, A. Bennett, R. Lenski, and J. Barrick, "Specificity of genome evolution in experimental populations of *Escherichia coli* evolved at different temperatures", *Proceedings of the National Academy of Sciences*, vol. 114, p. 201 616 132, 2017.

[56] I. Comas, M. Coscolla, T. Luo, *et al.*, "Out-of-africa migration and neolithic coexpansion of mycobacterium tuberculosis with modern humans", *Nature Genetics*, vol. 45, no. 10, pp. 1176–1182, 2013.

[57] M. Godfroid, T. Dagan, and A. Kupczok, "Recombination signal in mycobacterium tuberculosis stems from reference-guided assemblies and alignment artefacts", *Genome Biology and Evolution*, vol. 10, no. 8, pp. 1920–1926, 2018.

[58] A. Y. Morales-Arce, S. J. Sabin, A. C. Stone, and J. D. Jensen, "The population genomics of within-host mycobacterium tuberculosis", *Heredity*, vol. 126, no. 1, pp. 1–9, 2021.

[59] J. Franke, A. Klozer, J. A. de Visser, and J. Krug, "Evolutionary accessibility of mutational pathways", *PLoS computational biology*, vol. 7, no. 8, e1002134, 2011.

[60] D. L. Stern and V. Orgogozo, "Is genetic evolution predictable?", *Science*, vol. 323, no. 5915, pp. 746–51, 2009.

[61] M. Lässig, V. Mustonen, and A. M. Walczak, "Predicting evolution", *Nature Ecology & Evolution*, vol. 1, no. 3, p. 77, 2017.

[62] D. Temko, I. Tomlinson, S. Severini, B. Schuster-Bockler, and T. Graham, "The effects of mutational processes and selection on driver mutations across cancer types", *Nature Communications*, vol. 9, p. 1857, 2018.

[63] R. Poulos, Y. Wong, R.Ryan, H. Pang, and J. Wong, "Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations", *PLoS Genetics*, vol. 14, e1007779, 2018.

[64] J. M. Cannataro V.L. and J. Townsend, "Attribution of cancer origins to endogenous, exogenous, and actionable mutational processes", *bioRxiv*, pp. 10.1101/2020.10.24.352989, 2020.

[65] S. Borrell, Y. Teo, F. Giardina, *et al.*, "Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis", *Evolution, Medicine, and Public Health*, vol. 14, 65–74, 2013.

[66] J. Haldane, "A mathematical theory of natural and artificial selection. v. selection and mutation.", *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26, pp. 220–230, 1927.

[67] J. Haldane, "The part played by recurrent mutation in evolution", *The American Naturalist*, vol. 67, no. 708, pp. 5–19, 1933.

[68] R. Fisher, *The Genetical Theory of Natural Selection*. London: Oxford University Press, 1930.

[69] A. Stoltzfus and L. Y. Yampolsky, "Climbing mount probable: Mutation as a cause of nonrandomness in evolution", *J Hered*, vol. 100, no. 5, pp. 637–47, 2009.

[70] A. Stoltzfus, *Mutation, Randomness and Evolution*. Oxford, 2021, ISBN: 9780198844457.

[71] T. D. Lieberman, K. B. Flett, I. Yelin, *et al.*, "Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures", *Nature Genetics*, vol. 46, no. 1, pp. 82–87, 2014.

[72] E. I. Svensson and D. Berger, "The role of mutation bias in adaptive evolution", *Trends in Ecology and Evolution*, vol. 34, no. 5, pp. 422–434, 2019.

[73] J. Maynard Smith, R. Burian, S. Kauffman, *et al.*, "Developmental constraints and evolution", *Quart. Rev. Biol.*, vol. 60, no. 3, pp. 265–287, 1985.

[74] S. Green and N. Jones, "Constraint-based reasoning for search and explanation: Strategies for understanding variation and patterns in biology", *Dialectica*, vol. 70, no. 3, pp. 343–374, 2016.

[75] S. Kauffman, *The origins of order: Self-organization and evolution*. New York: Oxford University Press, 1993.

[76] K. Dingle, F. Ghaddar, P. Šulc, and A. A. Louis, "Phenotype bias determines how RNA structures occupy the morphospace of all possible shapes", *bioRxiv*, p. 2020.12.03.410605, 2020.

[77] K. Xiong, M. Gerstein, and J. Masel, "Non-adaptive factors determine which equally effective regulatory motif evolves to generate pulses", *bioRxiv*, p. 2020.12.02.409151, 2020.

[78] Y. Nakamura, T. Gojobori, and T. Ikemura, "Codon usage tabulated from international DNA sequence databases: status for the year 2000", *Nucleic acids research*, vol. 28, no. 1, pp. 292–292, 2000.

[79] P. Tataru, M. Mollion, S. Glémin, and T. Bataillon, "Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data", *Genetics*, vol. 207, no. 3, pp. 1103–1119, Sep. 2017.

[80] R. Karki, D. Pandya, R. C. Elston, and C. Ferlini, "Defining "mutation" and "polymorphism" in the era of personal genomics", *BMC medical genomics*, vol. 8, no. 1, pp. 1–7, 2015.

3.7 SUPPLEMENTARY MATERIAL

Model description

Our motivation is to develop and apply a phenomenological model that allows us to define a statistic that quantifies the influence of the mutation spectrum on the spectrum of adaptive substitutions. We focus on the most general and widely available data on adaptive genetic change: Single-nucleotide changes that alter amino acids, i.e., single-nucleotide missense changes. The standard genetic code specifies a set of 354 different types of single-nucleotide missense changes defined by a starting codon and an ending amino acid. For genomes that use the standard genetic code, any given episode of adaptation involving missense changes induces a distribution of adaptive substitution events over these 354 types, which we refer to as the spectrum of adaptive substitutions.

Because our goal is to model the observed number of counts for each type of adaptive substitution, we use negative binomial regression [32], which is a type of generalized linear model that is often employed for modeling count data. It is appropriate because the 354 mutational types are discrete and the substitution events that correspond to each type occur independently of one another. The general form of the model is

$$\log \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \log(\text{exposure}) + \beta\mathbf{x},$$

where Y is a vector of response variables, $\mathbf{x}$ are the explanatory variables, β_0 is the logarithm of the constant of proportionality, and exposure quantifies differences in the potential to observe each type of response. In our case, Y is the number of substitution events of each type, $\mathbf{x}$ is the logarithm of the mutation rate, and the exposure is given by codon frequency, which controls for the number of times each codon appears in protein-coding regions of the genome. Our model then takes the form

$$\log \mathbb{E}[\mathbf{n}(c,a)|\log(\mu(c,a))] = \beta_0 + \log(f(c)) + \beta \log \mu(c,a),$$

where $\mathbf{n}$ is the spectrum of adaptive substitutions (i.e., $\mathbf{n}(c,a)$ is the number of substitution events from codon c to amino acid a), $\mu(c,a)$ is the mutation rate from codon c to amino acid a, and $f(c)$ is the frequency of codon c in the genome. The coefficient β is a single statistic that captures the influence of the mutation spectrum on the spectrum of adaptive substitutions. The expected range of β is from 0 to 1: If $\beta = 0$, the mutation spectrum has no influence on the spectrum of adaptive substitutions. If $\beta = 1$, the mutation spectrum has a proportional influence on the spectrum of adaptive substitutions. Values of β between 0 and 1 represent an intermediate influence.

The above model only describes the expected number of counts for each type of substitution, however to fit the parameters of the model by maximum likelihood we must specify a full distribution for $\mathbf{n}(c,a)$. One common choice would be to assume that these counts are Poisson distributed (i.e. Poisson regression). However, Poisson regression assumes that the variance in the counts data is equal to the mean. In our data, we instead observe overdispersion, i.e. that the variance is larger than the mean. Such overdispersion is a common problem in Poisson regression. The standard solution is to instead use negative binomial regression, a more general model that

allows the variance to be different from the mean [32]. In the main text, we therefore use negative binomial regression to model the influence of the mutation spectrum on the spectrum of adaptive substitutions.

Meaning of key terms

Several important terms used in our study, such as "mutation", have meanings that are interpreted differently in different parts of the scientific community [80]. Moreover, our study design requires additional precision in being able to describe genetic and evolutionary changes, for example distinguishing a possible beneficial change to the genome of an organism from a realized instance where a heritable change of that type arises in a particular individual. In order to avoid any terminological ambiguity, we therefore provide formal definitions for these key terms below.

ADAPTIVE SUBSTITUTION An adaptive substitution is an evolutionary change in a population or sub-population, where each adaptive substitution is understood (in the present context) to result from an event of mutational introduction and an episode of selective enrichment that raises the mutant allele to a frequency close to 1.

EVENT An instance of change, having a particular time and place of occurrence, is an event. Compare to path or type. Here we assume that events occur independently from each other, and distinguish e.g. the number of times a particular mutational variant is observed from the number of distinct mutational events that introduced that variant into the population.

MISSENSE (NONSYNONYMOUS) In the literature of molecular evolution, codon changes that alter the amino acid are missense changes, and this class of change is often called "non-synonymous" (e.g., in the dN / dS literature) although technically non-synonymous changes include both missense and nonsense changes.

MUTATION A mutation is a heritable change to the genetic material in an individual lineage. The process of such change is also called mutation. The product of a mutational change is also called "a mutation" or a "mutant allele," and in population genetics this kind of usage is often extended to refer generally to derived alleles, e.g., the "concurrent mutations" regime refers to mutant alleles segregating concurrently in a population.

MUTATIONAL TYPE An event of evolutionary or mutational change can be assigned to a variety of mutational types or categories defined by a class of starting states (e.g., ATG codons) and a class of ending states (e.g., TTG codons). Here we are mainly focused on the 6 (reversible) nucleotide-to-nucleotide types and the 354 codon-to-amino-acid types. We use "path" for a specific kind of mutational type (see path).

PATH For the purposes of describing observed data sets for specific organisms, a path is a mutational type defined by a specific genomic site and a codon-to-amino-acid change. Parallel or recurrent events within a dataset are events that take place along the same path.

SPECTRUM A set of intensities or frequencies over some space of possibilities (e.g. a collection of different types of mutations) is a spectrum.

	Data		Neg. binomial regression		Prediction model		Spectrum elements	
Study	Paths	Events	β	p_β	Correlation	p_{corr}	Non-zero elements	Entropy
Basel [5]	126	2319	0.83 ± 0.27	0.002	0.15	0.005	78	0.53
Manson [4]	168	2094	0.84 ± 0.27	0.001	0.17	0.002	80	0.52

TABLE S3.1: **Separately analyzing the adaptive events from the two meta-analyses of antibiotic resistance substitutions in *M. tuberculosis* yields qualitatively similar results to analyzing them together.** Shown are the observed numbers of paths and events, the mutation coefficient β (with standard error) and its p-value, the Pearson's correlation between observed and predicted spectra of adaptive substitutions and its p-value, as well as the number of non-zero elements of the spectrum of adaptive substitutions and the entropy of the spectrum of adaptive substitutions.

	Influence of ti/tv ratio			Influence of rest of mutation spectrum			Model comparison	
Species	$\beta_{\text{ti/tv}}$	95% CI	$p_{\beta_{\text{ti/tv}}}$	β_{rest}	95% CI	$p_{\beta_{\text{rest}}}$	log likelihood	p_{LRT}
S. cerevisiae	0.79 ± 0.13	$[0.53, 1.05]$	$< 10^{-8}$				-1266.15	$< 10^{-16}$
	0.85 ± 0.11	$[0.63, 1.07]$	$< 10^{-16}$	1.25 ± 0.11	$[1.03, 1.47]$	$< 10^{-16}$	-1156.48	
E. coli	0.80 ± 0.17	$[0.46, 1.14]$	$< 10^{-5}$				-1109.90	$< 10^{-6}$
	0.85 ± 0.17	$[0.51, 1.19]$	$< 10^{-6}$	1.28 ± 0.26	$[0.77, 1.80]$	$< 10^{-6}$	-1083.80	
M. tuberculosis	0.84 ± 0.33	$[0.19, 1.50]$	0.01				-1233.32	0.03
	0.89 ± 0.32	$[0.26, 1.52]$	0.01	0.80 ± 0.36	$[0.09, 1.51]$	0.02	-1228.67	

TABLE S3.2: **The entire mutation spectrum provides better model fits than just the transition-transversion ratio.** Shown are the regression coefficient of the transition-transversion ratio $\beta_{\text{ti/tv}}$ (with standard error), its 95% confidence interval and its p-value, the regression coefficient associated to the rest of the mutation spectrum β_{rest} (with standard error), its 95% confidence interval and its p-value, as well as the p-value of the likelihood ratio test comparing both models p_{LRT}, which indicates that the more complex model including the full mutation spectrum provides a significantly better fit than the simpler model that only includes the transition-transversion ratio.

	Only codon frequencies		Complete model	
Species	Correlation [CI]	p_{corr}	Correlation [CI]	p_{corr}
S. cerevisiae	0.36 $[0.25, 0.44]$	$< 10^{-11}$	0.68 $[0.62, 0.73]$	$< 10^{-16}$
E. coli	0.31 $[0.22, 0.40]$	$< 10^{-9}$	0.41 $[0.31, 0.49]$	$< 10^{-14}$
M. tuberculosis	0.10 $[-0.0004, 0.2059]$	0.05	0.16 $[0.05, 0.26]$	0.003

TABLE S3.3: **A model using codon frequencies and the mutation spectrum provides better predictions than a model using only codon frequencies ($\beta = 0$).** Shown are the correlation coefficients for the two models, with 95 % confidence intervals and p-values.

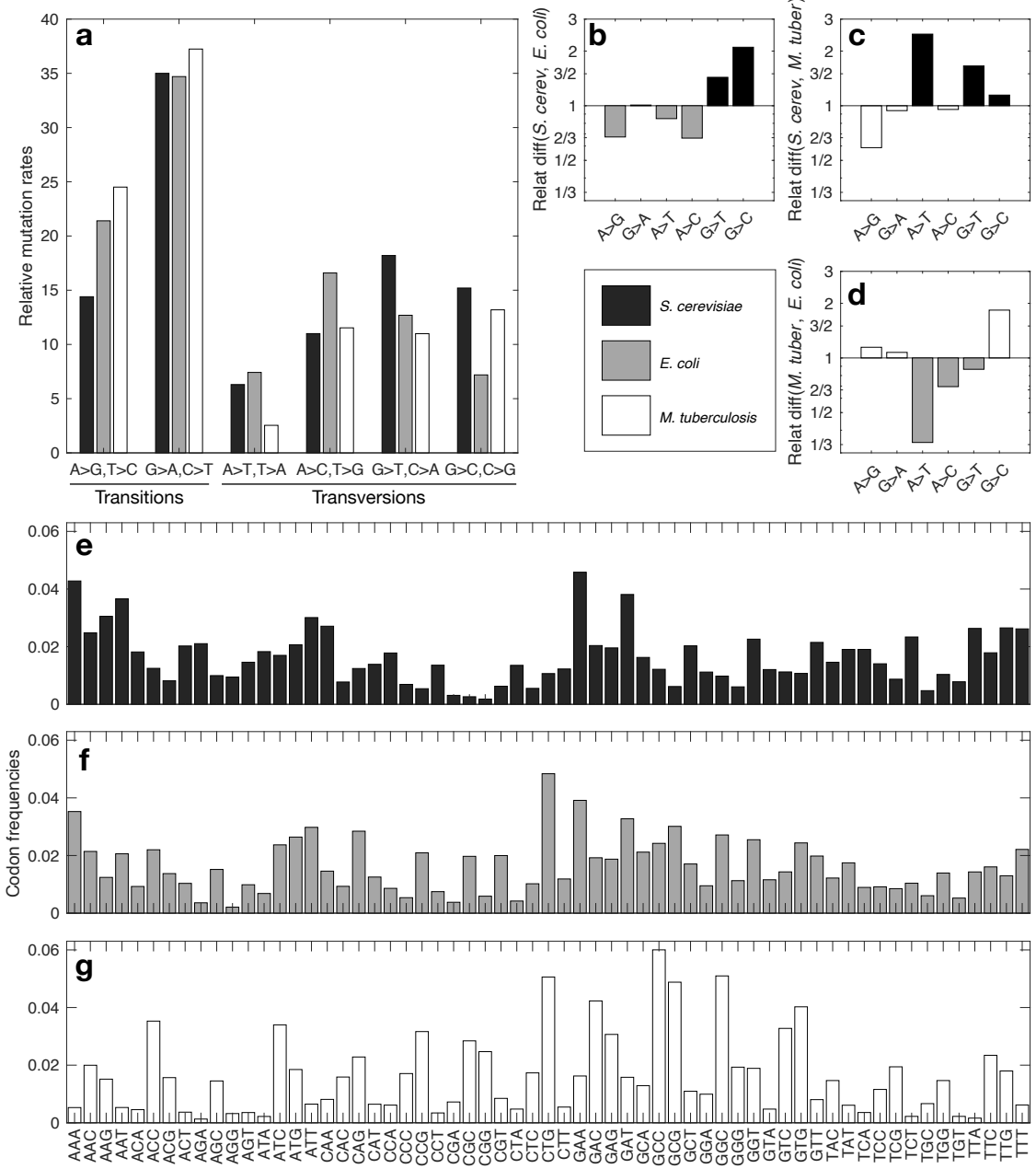

FIGURE S3.1: **Empirical mutation spectra and codon frequencies.** (a) Bar plots of the empirical mutation spectra for *S. cerevisiae*, *E. coli*, and *M. tuberculosis*. Bar color indicates the species; see legend. (b-d) Relative difference in mutation rates per mutation type, Relat diff$(b, a) = b/a$. Bar color indicates the species with the higher mutation rate for each mutation type. The vertical axis is logarithmically scaled for visual clarity. (e-g) Bar plots of the empirical codon frequencies for (e) *S. cerevisiae*, (f) *E. coli*, and (g) *M. tuberculosis*.

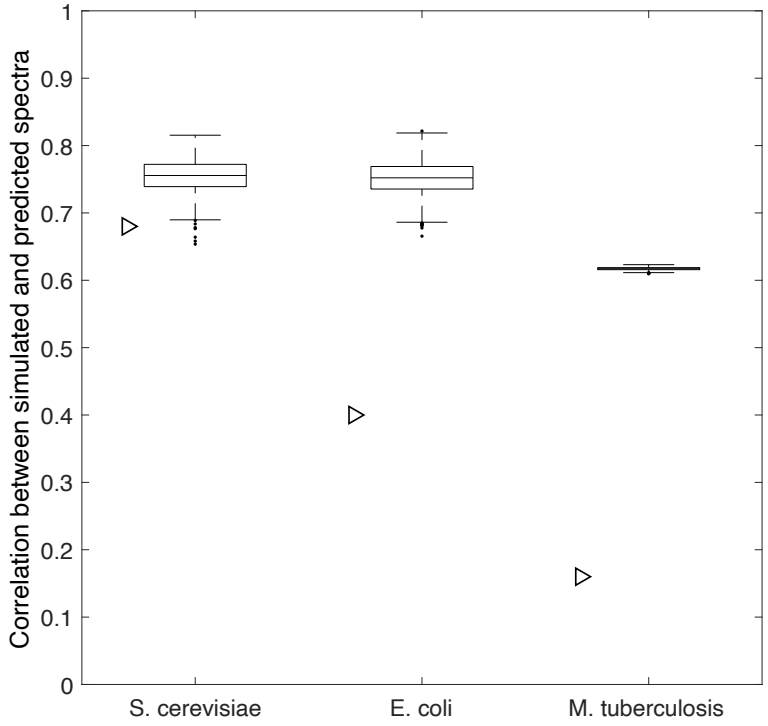

FIGURE S3.2: **The correlation between predicted and simulated spectra of adaptive substitutions depends on mutational target size, even under origin-fixation dynamics.** The distribution of correlations between predicted and simulated spectra of adaptive substitutions using the codon frequencies, mutation spectra, and number of non-zero elements in the spectrum of adaptive substitutions are shown for *S. cerevisiae*, *E. coli*, and *M. tuberculosis*. Data pertain to 10^3 simulations. Triangles show the correlations reported in Table 1, for reference.

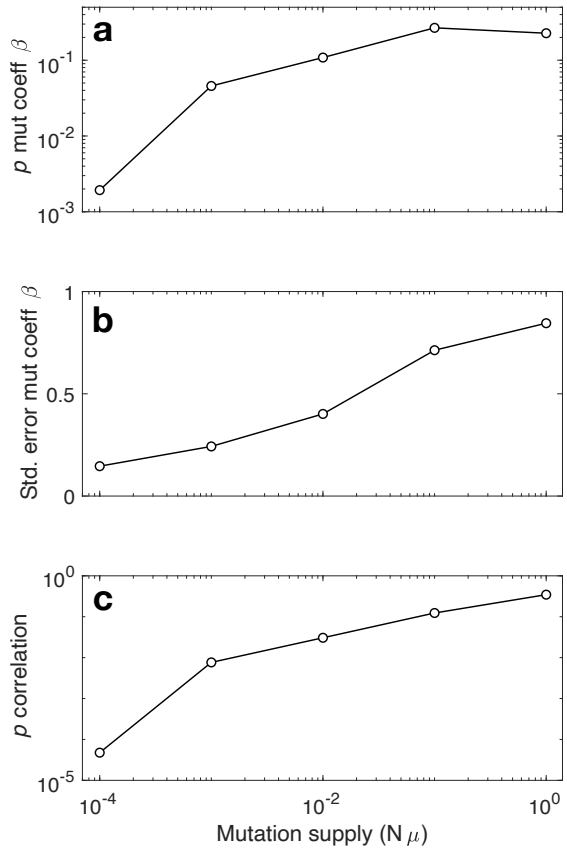

FIGURE S3.3: **High mutation supply diminishes the influence of mutation bias on adaptive evolution.** The (a) average p-value and (b) standard error of the mutation coefficient β, and (c) the average p-value of the correlation between predicted and simulated spectra of adaptive substitutions are shown in relation to mutation supply $N\mu$. Data pertain to those shown in Figs. 4a-c.

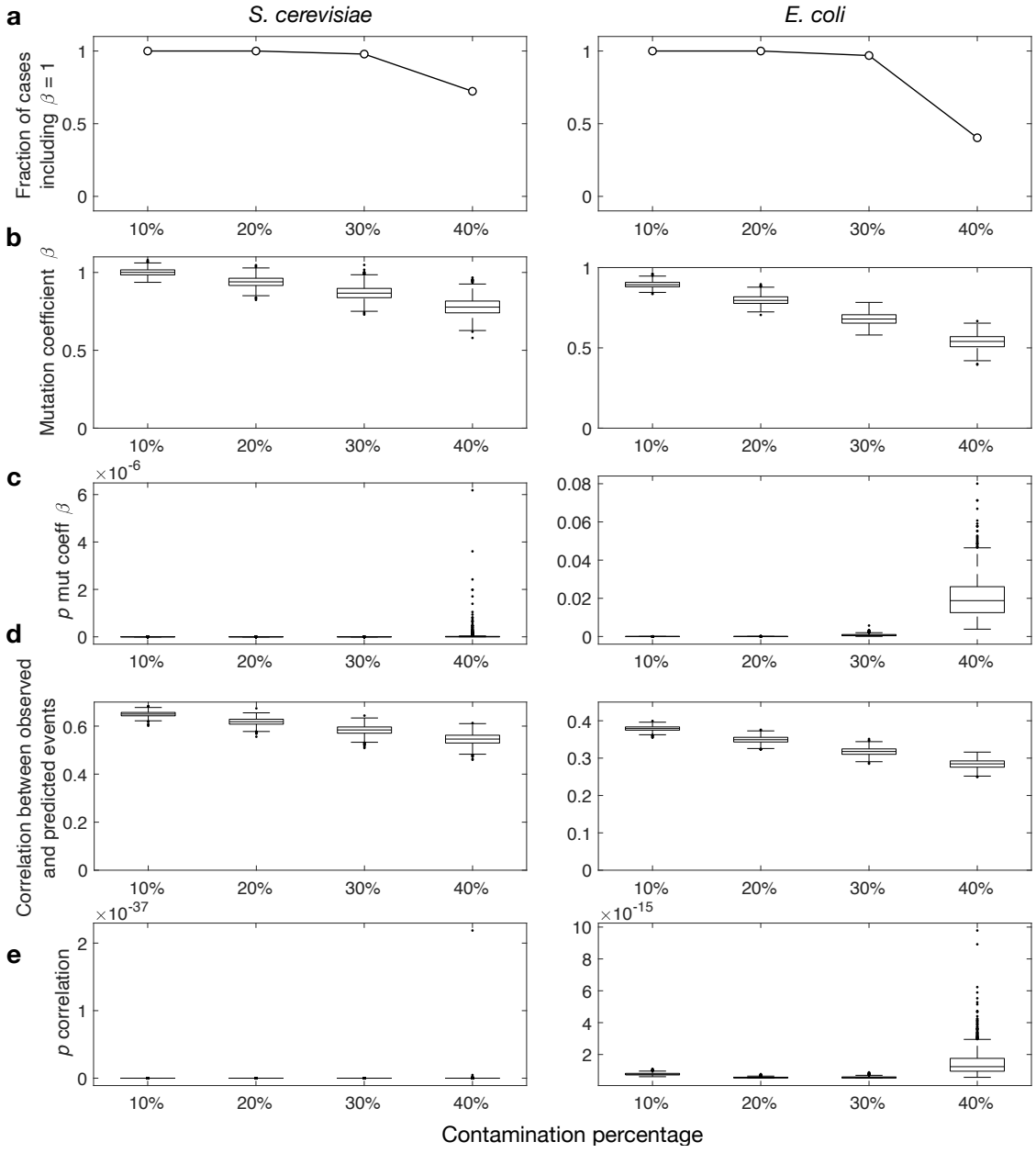

FIGURE S3.4: **Contamination analysis supports the influence of mutation bias on adaptation.** (a) Fraction of simulated data sets in which the confidence interval includes $\beta = 1$. (b) Inferred mutation coefficients β, (c) p-values of the regression coefficients β, (d) Pearson's correlation coefficients between observed and predicted spectra of adaptive substitutions, and (e) the p-values of the correlation coefficients, are all shown in relation to the percentage of substitutions randomly removed from the data sets of adaptive substitutions.

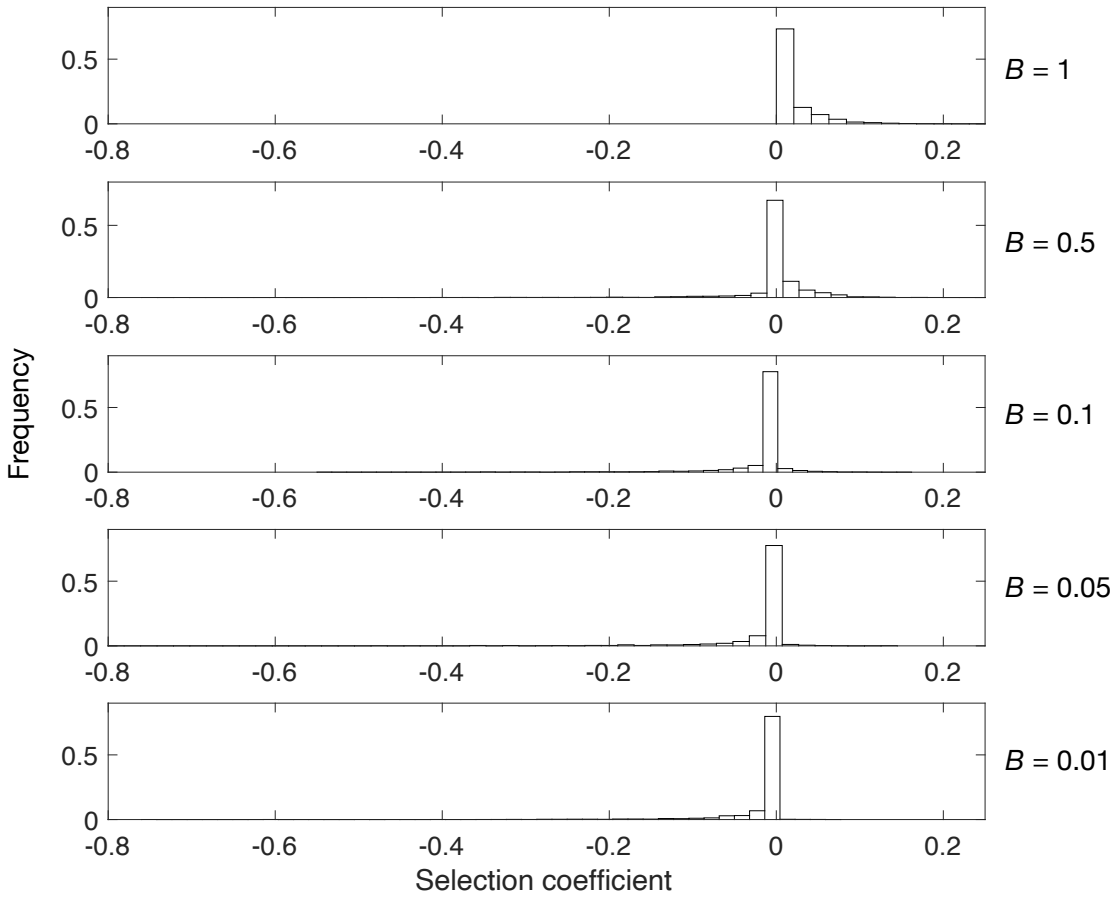

FIGURE S3.5: **Distributions of fitness effects.** Representative distributions of fitness effects used in the evolutionary simulations for five different proportions of beneficial mutations B.

4

ON THE CORRELATION OF MUTATION RATES AND SELECTION COEFFICIENTS

Currently in preparation as: Alejandro V. Cano*, Bryan Gitschlag*, Joshua L. Payne, David M. McCandlish, & Arlin Stoltzfus (2022). On the correlation of mutation rates and selection coefficients. (* co-first authors).

Author's contributions: A.V.C., B.G., J.L.P., D.M.M., and A.S. designed research; A.V.C. analyzed data and performed simulations; B.G. and D.M.M derived analytical results; and A.V.C., B.G., J.L.P., D.M.M., and A.S. wrote the manuscript.

4.1 ABSTRACT

Evolutionary change can be understood statistically as the transformation of one distribution into another. This notion is familiar in classical quantitative genetics, where evolutionary change is represented as the selective transformation of a distribution of continuous trait values in standing variation. In molecular evolution, the space is discrete and, rather than taking a measurable quantity (standing variation) as a starting point, one often begins with the concept of a universe of prior possibilities (e.g., all possible 1-residue changes) that is sampled by a process of mutagenesis, followed by selective filtering. This duality leads to a more complex set of questions, including those relating to the joint distribution of mutation rates and selection coefficients, which (in the simplest model) is first transformed by mutagenesis from a prior nominal distribution to a de novo mutation distribution, and then transformed by selection into a distribution of fixed changes. Here we consider what happens to this nominal distribution under mutagenesis and selection, using a combination of mathematical theory, computer simulations, and analysis of available data from deep mutational scanning, cancer informatics, and evolution experiments. We show that, in principle, this kind of joint conditioning can induce a great variety of association patterns. However, we argue that natural systems tend to have the kinds of joint distributions that induce negative associations, leading in some cases to Berkson's paradox. Finally, we show that the magnitude of the induced associations between mutation rates and selection coefficients is modulated by the shape of the nominal distribution and population genetics conditions.

4.2 INTRODUCTION

Recent work in evolutionary genetics has revealed that, in some evolutionary and population genetic conditions, adaptation reflects variation in both mutation rates and selection coefficients [1]–[3]. An unexplored implication is that the dual causation of adaptation by mutation and selection may induce paradoxes of joint conditioning, because in general, joint conditioning can induce associations between the causal variables (e.g., mutation rates and selection coefficients) [4]. For instance, if we sample values for X and Y in a manner that depends on the sum $Z = X + Y$, this joint conditioning makes the covariance of X and Y in the sample more negative than in the population, so that it can lead to an anti-correlation even if the prior distributions of X and Y are uncorrelated [4]. Such induced associations (often negative) can be seen as generalizations of Berkson's Paradox [5]. A common example to portray such effects considers applicants that are admitted to a university based on satisfying a minimum score that combines academic and athletic prowess. We may see a negative correlation between these two properties among admitted students, even if there was no correlation in the applicant pool. In fields where causal relationships of variables are represented with graphs, this type of problem is called "conditioning on a collider," where the collider is a node with multiple input nodes, i.e., a variable dependent on multiple other variables [4].

To the extent that substitutions implicated in an episode of evolutionary change reflect both the likelihood that an option is mutationally introduced, and its subsequent likelihood of being established by selection or drift, the process of adaptation is a collider that may induce non-causal associations between mutation rate and selection coefficient. That is, one may define a vector

of mutation rates μ, and a corresponding vector s of selection coefficients. In the simplest case of adaptation via origin-fixation dynamics, changes occur at the rate $N\mu * \pi$, i.e., rate of origin times probability of fixation, where $\pi = 2s$ in the simplest case of beneficial changes. Because this dual process is ordered, i.e., mutation comes first, three different joint distributions of μ and s are of interest: the prior nominal distribution representing the universe of possible changes, the distribution of *de novo* mutations (i.e., the mutation spectrum), and the distribution of fixed changes.

What do we know about the nominal distributions of possible changes, before mutation or selection has taken place? Empirical work is difficult to find. Of course, many deep mutational scanning (DMS) studies report fitness values for all (or nearly all) possible amino acid changes to a target protein, but such approaches typically do not involve the measurement of mutation rates. However, [6] recently used deep-sequencing methods to estimate both selection coefficients for tens of thousands of single-nucleotide changes in laboratory cultures of Dengue virus, and average mutation rates for 12 nucleotide substitution classes (see also [7]).

In the absence of a solid empirical foundation, evolutionary thinking on this issue has been dominated by assumptions and theories. The textbook doctrine that mutation is random, a central tenet of the neo-Darwinian theory of evolution [8], suggests that the frequency of occurrence of mutations will be uncorrelated with fitness effects, and this is sometimes an explicit assumption in formal modeling [9]. However, this doctrine seems to function more as a heuristic or guiding assumption than a rigorously grounded conclusion [2]. Meanwhile, diverse ideas about the evolution of mutational strategies (implying mutation-fitness correlations) have been proposed, from relatively narrow and modest claims of adaptive amelioration lowering the mutation rate in functionally important regions [10], [11], to the emergence of specialized mutation systems (e.g., cassette-shuffling systems) in the context of immune evasion or host-phage arms races ([12]; Ch. 5 of [2]), to ideas about "directed" or "smart" mutation systems [13].

What about the distribution of mutation rates and selection coefficients fixed in adaptation? The largest joint distribution for adaptive evolutionary changes, to our knowledge, is from results of [14], who reported mutation rates and selection coefficients for 11 Rifampicin-resistant isolates identified in laboratory adaptation of *Pseudomonas aeruginosa*. However, several studies of tumor prevalence have reported paired estimates of growth rates and selection coefficients. [15] report that the most prevalent tumors are often not the ones with the highest growth rates, but the ones with high rates of mutational origination. Likewise, several studies have reported on selection coefficients and mutation rates for clonal expansion in somatic haematopoesis [16].

Though the theory for the joint distribution of mutation rates and selection coefficients before and after adaptation has not received much explicit attention, classical and recent work provides guidance on expected effects. New alleles are introduced at rates that follow from the mutation spectrum, and the kinetic bias imposed on evolution by such mutation rate biases ranges from nothing (no influence) to a proportional influence, as shown recently by [3]. Likewise, the effect of selection is to establish new alleles according to their fitness effects, from the limiting case of an independent probability of fixation of an allele i dependent solely on s_i (and N), to the case of clonal interference in which differential effects of fitness are amplified further by competition among concurrent mutations, to the limiting case in which i is fixed deterministically if it is the fittest allele.

Our purpose, in this context, is to consider generally the associations of mutation rate and fitness on a discrete space such as a genome space, and how those associations change when conditioned on the effects of mutagenesis and selection. Our initial exploration of this topic draws on mathematical theory, computer simulations, and analysis of data currently available. We begin by defining the distributions of interest, and explaining some general properties of size-biased distributions. We then describe a simple stochastic model of mutagenesis and adaptation in which mutation rates and fitnesses each take on just 2 or 3 discrete values. Using this model, which is easily visualized, we show that a variety of negative and positive non-causal associations are possible, including the strong (sign-changing) and weak forms of Berkson's paradox. We then derive a more general theory of joint evolutionary conditioning where the distributions of μ and s are known. We use results from the simple model to explain and illustrate how effects of conditioning emerge from higher moments of underlying joint distributions.

Next we consider the expected effects of joint conditioning in Dengue whole genome evolution using its naturalistic underlying nominal distribution for μ and s, and representing the process of adaptation using either population simulations or explicit origin-fixation dynamics. We show that this particular empirical nominal distribution tends to be transformed in the direction of Berkson's paradox after adaptation, namely, a negative association between mutation rates and selection coefficients is induced by the joint conditioning. Moreover, we show that the magnitude of such induced association depends on the population genetic conditions, as they modulate the influence of the mutation rates in the joint conditioning [3]. Finally, we consider empirical data on the joint distribution of μ and s from a DMS study of tumor protein TP53, which allows for the construction of the nominal distribution for all possible amino acid changes. The further integration of a dataset on the observed frequency of different somatic mutations to this framework allows for the empirical characterisation of the fixed distribution of changes, which reflects the influence of both mutation and selection. We show that associations with different signs can be induced in the fixed distribution depending on the intensity of mutation bias, which is smaller for amino acid changes resulting from single point mutations than for the ones associated with multinucleotide mutations.

We conclude with a prescription for future work. For mutation-limited evolution on a discrete space, the theory of mutation-fitness associations describes how a potential distribution of possibilities is transformed into an actual one. This theory was not of much use until recent technological advances made it possible to measure mutation rates and selection coefficients with precision. Therefore, the integration of empirically characterised mutation rates and selection coefficients to evolutionary models is key to improve our understanding about the correspondence between what is selectively beneficial and what is mutationally likely.

4.3 RESULTS

Analytical Results

Here we consider the joint distribution of mutation rates and selection coefficients for beneficial changes from three different perspectives. First, we can ask about the joint distribution of selection coefficients and mutation rates when we pick randomly from an abstract set of possibilities. For

example, if we have a DNA sequence of length ℓ, then there are 3ℓ alternative sequences that differ by a single nucleotide, out of which some number n are beneficial. Any one of these n beneficial changes occurs by mutation at some definite rate and has some definite selective advantage. Accordingly, we could ask about the expected selection coefficient, for instance, or the correlation between mutation rate and selection coefficient, when we draw one of these n possibilities uniformly at random. We call this the nominal distribution and denote the expected selection coefficient or mutation rate of a draw from this distribution as $E_{\text{nom}}(s)$ or $E_{\text{nom}}(\mu)$, respectively. The nominal distribution is therefore equivalent to what we observe in a deep mutational scanning study that comprehensively determines selection coefficients for all single-nucleotide variants (or in other designs, all amino acid variants), or when we scan a model of context-dependent mutation rates across the genome and calculate the mutation rate for each possible mutation in turn. This is also the distribution relevant for determining the overall mutation rate, since the total beneficial mutation rate is the sum of the mutation rates to each possible beneficial mutation.

Second, in addition to the joint distribution of selection coefficients and mutation rates among mutational possibilities, we can consider the distribution that arises as new mutations are introduced. Here, instead of picking a random position and a random nucleotide, as in the nominal distribution, we ask about the distribution of the next beneficial mutation likely to occur in a given genome. This distribution differs from the nominal distribution in that mutations with higher rates are more likely to be the next mutation to occur. Accordingly, this is also the distribution that we observe in a mutation accumulation experiment and which determines the average selection coefficient for new mutations. We call this the distribution of de novo mutations and write the expected selection coefficient of a new mutation as $E_{\text{de novo}}(s)$.

Third, we can consider mutations that become fixed in evolution; that is, those that actually contribute to adaptation. This is equivalent to asking about the selection coefficient and mutation rate of the next mutation that is going to fix in the population. We call this distribution the fixed distribution and write the expected selection coefficient of the next mutation to fix as $E_{\text{fixed}}(s)$

To understand the necessary relationships between these distributions, it is helpful to introduce the concept of a size-biased distribution [17]. In particular, given a non-negative random variable X, we can define the size-biased distribution as the random variable X^* where $P(X^* = x)$ is proportional to $xP(X = x)$ for any non-negative value x. Thus, the size-biased distribution is a re-weighted version of the original probability distribution. More specifically, it is the probability distribution obtained when the new weights on each outcome are given by the value of that outcome. This re-weighting favors larger outcomes and results in a systematic change to the moments of the distribution. More precisely, the k^{th} raw moment of the size-biased distribution is determined by the $(k+1)^{\text{th}}$ moment of the original distribution by $E((X^*)^k) = E(X^{k+1})/E(X)$. Size-biasing is well known to result in certain paradoxes. For example, if the number of children per family is given by the random variable X then the number of siblings in a random individual's family is given by X^*, which results in the "sibling paradox" that the expected number of siblings in a given person's family is larger than the average number of children per family. Another classical example is the "waiting time paradox", where (for instance) if the time interval between buses is distributed as X, the expected time spent waiting for a bus, for a person who arrives at the bus stop at a random time, is $E(X^*)/2$ instead of the shorter time $E(X)/2$, because a person is more likely to arrive in a longer interval between buses than a smaller one.

The concept of size-biasing is helpful in thinking about the relationships between the nominal, de novo, and fixed mutational distributions because the distributions are related to each other by size-biasing according to either the mutation rate or the selection coefficient. In particular, the de novo distribution is obtained by size-biasing the nominal distribution with the mutation rate. In symbols, we write this as:

$$P_{\text{de novo}}(s = s_i \text{ and } \mu = \mu_i) = \frac{\mu_i P_{\text{nom}}(s = s_i \text{ and } \mu = \mu_i)}{E_{\text{nom}}(\mu)} \tag{4.1}$$

where s_i and μ_i are the selection coefficient and mutation rate for the i-th class of mutations, respectively.

This size-biasing means that the average mutation rate of mutations observed in the de novo distribution will be at least as large as the average mutation rate with respect to the nominal distribution. In fact, the increase in the expected mutation rate is given by

$$E_{\text{de novo}}(\mu) - E_{\text{nom}}(\mu) = \frac{\text{Var}_{\text{nom}}(\mu)}{E_{\text{nom}}(\mu)} \geq 0. \tag{4.2}$$

Importantly, size-biasing the nominal distribution with respect to mutation rate will also typically affect the mean selection coefficient, provided that a non-zero correlation exists between mutation and selection in the nominal distribution. In particular,

$$E_{\text{de novo}}(s) - E_{\text{nom}}(s) = \frac{\text{Cov}_{\text{nom}}(\mu, s)}{E_{\text{nom}}(\mu)}. \tag{4.3}$$

In fact, the form of these equations will likely be familiar to many readers as they are directly analogous to Fisher's fundamental theorem and Robertson's secondary theorem, respectively (see [18]). This is because the fitness distribution after selection is the size-biased transformation of the fitness distribution before selection, and the response to selection corresponds to size-biasing the distribution of another random variable according to its fitness. Thus, these classical results for fitness provide an easy mnemonic for the results of size-biasing more generally.

Whereas the de novo distribution is obtained from the nominal distribution by size-biasing with respect to mutation rate, the fixed distribution is obtained from the de novo distribution by size-biasing with respect to selection coefficient. This size-biasing is based on a strong-selection, weak-mutation (SSWM) approximation [19], [20]. Under this approximation, the probability of fixation of a new mutation is given by $2s$ [21], which is proportional to the selection coefficient and thus size-biases the joint distribution with respect to the selection coefficient. Importantly, when selection is relatively strong ($s \gg 1/N$), the influence of neutral drift is more negligible and probabilities of fixation more accurately reflect selection coefficients [22]. Moreover, biased mutation rates are most likely to shape the outcome of adaptive evolution when overall mutation rates are low (that is, $\mu N << 1$), since this forces selection to act on new variation as it arises instead of standing variation [23], [24]. Accordingly, the joint distributions of mutation and

selection described here approximate the origin-fixation regime, where evolution is proportional to both μ and s [23]. In particular, we have

$$P_{\text{fixed}}(s = s_i \text{ and } \mu = \mu_i) = \frac{s_i P_{\text{de novo}}(s = s_i \text{ and } \mu = \mu_i)}{E_{\text{de novo}}(s)} \tag{4.4}$$

$$= \frac{\mu_i s_i P_{\text{nom}}(s = s_i \text{ and } \mu = \mu_i)}{E_{\text{nom}}(\mu s)} \tag{4.5}$$

where the second line shows that we can also derive the fixed distribution from the de novo distribution by reweighting each class of mutations i according to the product of its mutation rates and selection coefficients $\mu_i s_i$. Similarly to the difference between the nominal and de novo distributions, moving from the de novo distribution to the fixed distribution will typically change the expected selection coefficient and expected mutation rate, with

$$E_{\text{fixed}}(s) - E_{\text{de novo}}(s) = \frac{\text{Var}_{\text{de novo}}(s)}{E_{\text{de novo}}(s)} \geq 0. \tag{4.6}$$

and

$$E_{\text{fixed}}(\mu) - E_{\text{de novo}}(\mu) = \frac{\text{Cov}_{\text{de novo}}(\mu, s)}{E_{\text{de novo}}(\mu)}. \tag{4.7}$$

So far we have discussed the fact that the mean mutation rate and selection coefficient of mutations may differ depending on whether we consider random genetic perturbations (nominal distribution), random mutations introduced into a population (the de novo distribution) or random fixed mutations. However, our main interest here is in asking about whether there is a systematic relationship between mutation and selection, and how this relationship appears from each of these perspectives. We begin by asking how such a systematic relationship is transformed across these three distributions by considering the simplest possible cases, with the caveat that the results in these simple cases are, as we will show below, deeply misleading concerning our expectations in more realistic situations.

Clearly the simplest possible case is when mutation and selection are completely independent. In particular, if $P_{\text{nom}}(s = s_i \text{ and } \mu = \mu_i) = P_{\text{nom}}(s = s_i) P_{\text{nom}}(\mu = \mu_i)$ for all i, then $P_{\text{de novo}}(s = s_i \text{ and } \mu = \mu_i) = (\mu_i P_{\text{nom}}(\mu = \mu_i)/E_{\text{nom}}(\mu))\, P_{\text{nom}}(s = s_i)$ so that mutation and selection are independent with respect to the de novo distribution where the de novo distribution of selection coefficients remains unchanged and the de novo distribition of mutation rates is simply the size-biased version of the nominal distribution of mutation rates. Similarly, $P_{\text{fixed}}(s = s_i \text{ and } \mu = \mu_i) = (\mu_i P_{\text{nom}}(\mu = \mu_i)/E_{\text{nom}}(\mu))\, (s_i P_{\text{nom}}(s = s_i)/E_{\text{nom}}(s))$, so that the fixed distribution corresponds to independent draws from the sized-biased distribution of mutation rates and the size-biased distribution of selection coefficients. Thus, in the case of exact independence between mutation and selection, this independence is maintained in all three distributions. Importantly, as we shall see, exact independence is rarely realized when considering a specific finite set of mutations, which typically leads to substantial departures from these simple expectations.

Another simple case that suggests that the qualitative relationship between mutation and selection should be similar across all three of these distributions is the case when we have only two mutation rates and two selection coefficients. For convenience we will work with relative mutation rates and selective coefficients, so that we will write the lower mutation rate as 1 and the larger mutation rate as $k > 1$, and write the lower selection coefficient as

1 and the larger selection coefficient as $b > 1$. For this case it is also helpful to be able to work with the individual probabilities for the four possible combinations of mutation rate and selection coefficient. These probabilities are written as $p^{\text{nom}}_{1,b}$, for example, referring to the probability of observing the lower mutation rate and the larger selection coefficient when drawing from the nominal distribution; likewise $p^{\text{fix}}_{k,b}$ refers to the probability of observing the greater selection coefficient and greater mutation rate when drawing a random fixed mutation. We will also write e.g. $p^{\text{de novo}}_{k} = p^{\text{de novo}}_{k,1} + p^{\text{de novo}}_{k,b}$ for the marginal frequency of the higher mutation rate in the de novo distribution and $\text{Var}_{\text{nom}}(p(s)) = p^{\text{nom}}_{b}(1 - p^{\text{nom}}_{b})$ as the variance in the frequency of the high versus low selection coefficient under the nominal distribution. Finally, let $D_{\text{nom}} = p^{\text{nom}}_{k,b} p^{\text{nom}}_{1,1} - p^{\text{nom}}_{k,1} p^{\text{nom}}_{1,b}$, which is a measure of the association between the high and low levels for the mutation rate and selection coefficient.

With this basic setup in hand, we can now consider the correlation between mutation rate and selection coefficient for our three distributions. In particular, we have

$$\rho_{\text{nom}} = \frac{D_{\text{nom}}}{\sqrt{\text{Var}_{\text{nom}}(p(\mu)^2)\,\text{Var}_{\text{nom}}(p(s))}} \tag{4.8}$$

$$\rho_{\text{de novo}} = \frac{k}{E_{\text{nom}}(\mu)^2} \frac{D_{\text{nom}}}{\sqrt{\text{Var}_{\text{de novo}}(p(\mu))\,\text{Var}_{\text{de novo}}(p(s))}} \tag{4.9}$$

$$\rho_{\text{fix}} = \frac{kb}{E_{\text{nom}}(\mu s)^2} \frac{D_{\text{nom}}}{\sqrt{\text{Var}_{\text{fix}}(p(\mu))\,\text{Var}_{\text{fix}}(p(s))}}. \tag{4.10}$$

Note that k, b, μ and s are all positive, and variances are non-negative (and must be positive for the correlation to be defined), so that the sign of each expression depends only on D_{nom}. Thus, we see that in the case where there are only 2 selective and 2 mutational classes, the sign of the correlation between mutation rates and selection coefficients is the same between all three distributions. This is illustrated in the first two rows of Fig. 4.1. In the first row, while the mean selection coefficient and mean mutation rate change across the three distributions, all three distributions remain uncorrelated. In the second row, the three distributions are all negatively correlated with the strength of the negative correlation increasing from nominal to de novo to fixed. Note that the negative correlation also results in a slightly non-monotonic pattern of change in the mean selection coefficient and mean mutation rate.

While a consistent sign arises in the correlation of mutation and selection in this simplest case of 2 possible values (high or low) for each, if we extend the possibilities to 3 classes each of mutation rate and selection coefficient, the possibilities are far less constrained. The third row of Fig. 4.1 shows that mutation and selection can be uncorrelated in the nominal and de novo distributions but correlated among fixed mutations. The reason for this is best understood in relation to Simpson's paradox, where the sign of a correlation within groups may be different than the sign of the correlation when groups are aggregated. Here we can see the original "plus" pattern as two separate groups with perfect negative correlations, one with $\mu s = 3$ and another with $\mu s = 15$. The latter group is enriched among fixed mutations, resulting in the observed negative correlation.

Another useful example is considering high, medium and low possibilities for each of mutation rate and selection coefficient, with 3 possible mutations that each take on a unique pairwise value

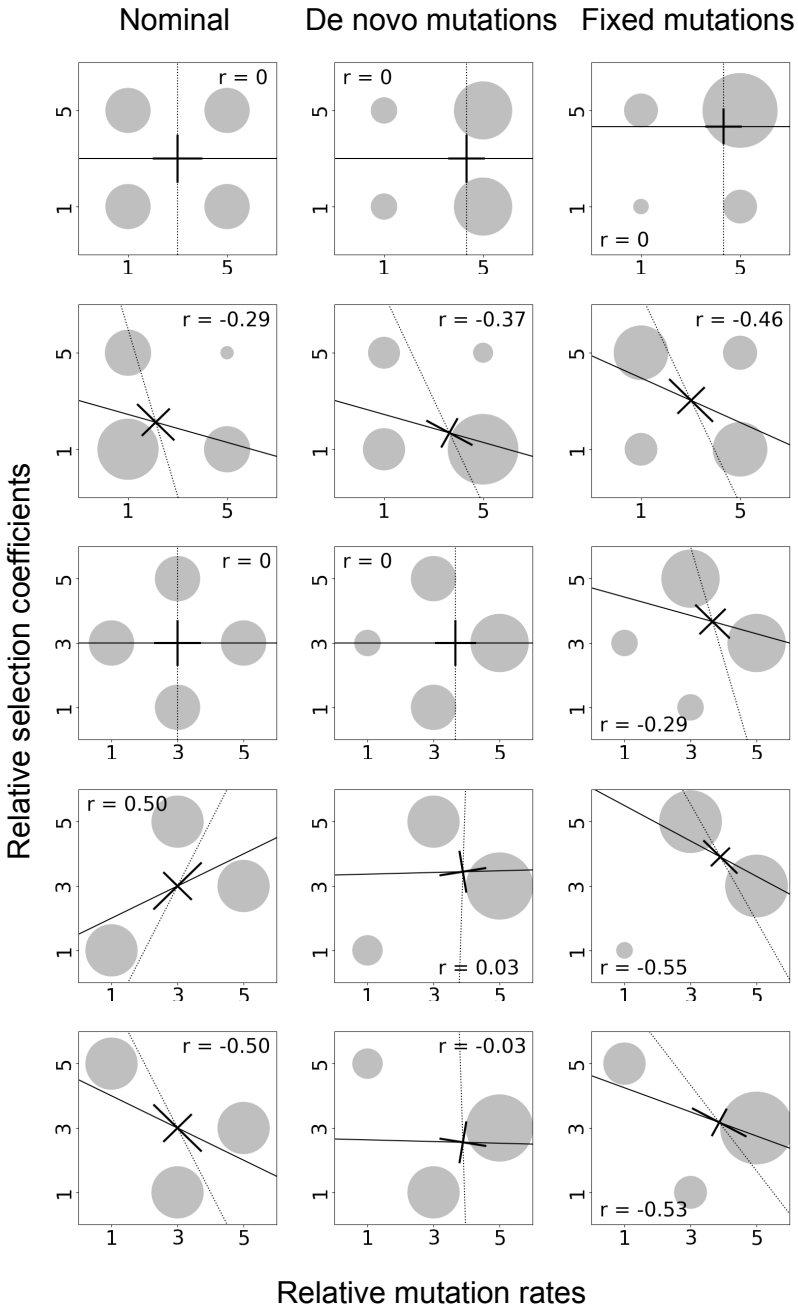

FIGURE 4.1: **Correlations of mutation and selection under some simple models.** In the simplified models shown here, mutation rates and selection coefficients each take on just 2 or 3 values. The columns from left to right show the nominal distribution (representing the landscape of possible mutations across varying mutation rates and selection coefficients), the distribution of de novo mutations (the possible changes weighted by mutation rates), and the distribution of changes resulting from a mutation-fixation process. Transecting lines show the regression of selection on mutation (solid) and mutation on selection (dotted), respectively; central crossed lines show the first and second principal components, with the lengths proportional to the variance explained, where by definition $V_1 \geq V_2$.

of mutation rate and selection coefficient. In this case, 17% of such configurations result in at least one sign-flip between the 3 distributions. More generally, we provide examples in Fig. S4.1 showing that any pattern of signs of correlation among the 3 distributions is possible once we admit 3 distinct values for mutation rate and selection coefficient. In addition, allowing mutation rates and selection coefficients to have different ranges of relative rates means that the signs of the correlations for the de novo and fixed distribution may depend on these specific ranges (Fig. S4.2). Taken together, these results suggest that when assessing the relationship between mutation and selection, it is essential to specify which of these distributions is being described.

What explains this counterintuitive behavior? The key observation here is that the lower moments of a size-biased distribution depend on higher moments of the original distribution. Thus, for example, the sign of the correlation between between mutation rate and selection coefficient for the de novo distribution depends on one of the third mixed moments of μ and s in the nominal distribution, and specifically has the same sign $E_{\text{nom}}(u^2 s)E_{\text{nom}}(\mu) - E_{\text{nom}}(\mu^2)E_{\text{nom}}(\mu s)$ (see Supplemental Information for general formulas for the variances and covariances of mutation rates and selection coefficients for all three distributions). Similarly, the fixed distribution depends on the fourth mixed moments of u and s and its correlation coefficient has the same sign as $E_{\text{nom}}(\mu^2 s^2)E_{\text{nom}}(\mu) - E_{\text{nom}}(\mu^2 s)E_{\text{nom}}(\mu s^2)$. Thus, many of the paradoxical results above reflect our poor intuition for higher mixed moments. As we shall see below, another important consequence of this dependence on higher mixed moments is that even if the selection coefficients and mutation rates of beneficial mutations are in principle independent, the finite number of pairs (of mutation rate and selection coefficient) that are drawn from such a distribution as realized beneficial mutations may poorly approximate an independent distribution in these higher moments and thus lead to non-negligible correlations for the de novo and fixed distributions.

Simulations with a biologically realistic nominal distribution

As noted above, distributions of μ and s following mutation and selection may show a variety of associations, of variable strength. Therefore, it is of interest to consider whether natural joint distributions will show any particular kind of correlation. For instance, Dolan *et. al* carried out serial passages of Dengue virus on mosquito or human host cells, and then used a high-accuracy deep-sequencing method to estimate the fitnesses of all possible 32166 single-nucleotide variants across the entire Dengue genome [6]. The same study provides estimates for the mutation rates of the 12 possible single nucleotide changes, uncovering the presence of a strong mutation bias towards C>T transitions [6]. This nominal distribution includes 325 single point mutations for which the lower bound of the selection coefficient exceeds 1. Among these beneficial changes, mutation rates show a slight and insignificant negative association with selection coefficients (Fig.4.2A, Pearson's correlation coefficient $r = -0.06$, $P = 0.28$).

To test the joint effect of mutation and selection on this relationship, we perform population-genetic simulations of evolution in a haploid genome, for five different values of mutation supply $N\mu$ (population size times mutation rate) using the nominal distribution of beneficial mutations from [6]. We implemented the simulations in SLiM v3.4 [25]. For each run of the simulation, we recorded the identity of all adaptive mutations on the first sequence to reach fixation, repeating this process until we obtain 500 beneficial substitutions. For each value of mutation supply $N\mu$,

we simulated three data sets and calculated the correlation between selection coefficients and mutation rates for the distribution of fixed mutations (Methods). Note that this is not an attempt to replicate the actual population biology of Dengue virus, which is complex: we are merely taking advantage of an empirical nominal distribution that is known for Dengue virus, and using it in a simple population model.

The results of these simulations are shown in Fig. 4.2. The effect of mutation and selection on the joint distribution depends on mutation supply ($N\mu$), as shown by the examples in Fig. 4.2C to 4.2E. At the lowest mutation supply of 10^{-4} (Fig. 4.2C), the correlation is substantially negative and takes its lowest value. As mutation supply increases to 10^{-2} or 10^{0} (Fig. 4.2D or 4.2E), the correlation becomes less negative, reflecting a diminished influence of differential mutation rates. This effect of mutation supply is summarized in 4.2F, showing that at high mutation supply, the correlation converges on roughly -0.1.

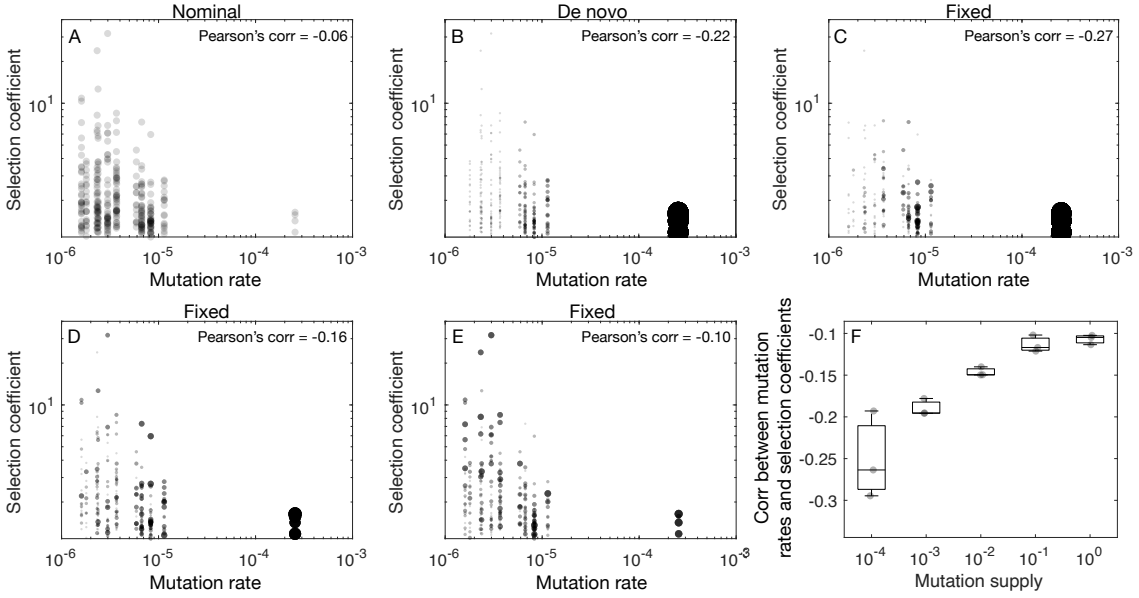

FIGURE 4.2: **Evolutionary simulations show that adaptation shapes the association between mutation rate and selection coefficient. A**. The nominal distribution shows the mutation rates and selection coefficients for all 325 beneficial point mutations. Pearson's correlation coefficient $r = -0.06$ ($P = 0.28$). The dashed line corresponds to the regression line. **B**. The de novo distribution shows the mutation rates and selection coefficients weighted by the mutation rates. Pearson's correlation coefficient $r = -0.22$ ($P < 10^{-6}$). **C-E**. The fixed distribution shows the mutation rates and selection coefficients weighted by the simulated frequency of mutations for three simulation runs with different mutation supply values. Pearson's correlation coefficient are $r = -0.27$ ($P < 10^{-6}$), $r = -0.16$ ($P < 10^{-4}$) and $r = -0.10$ ($P = 0.005$), for mutation supply values 10^{-4}, 10^{-2} and 10^{0}, respectively. In panels **B-E** the dot size is proportional to its frequency. Log-scale is used for visualisation purposes. Pearson's correlation coefficients are calculated on the unlogged values. **F.** The correlation between mutation rates and selection coefficients for the distribution of fixed mutations, as a function of mutation supply ($N\mu$). Each box shows the distribution of correlations for three replicate simulations, and the horizontal line shows the median.

Evidence from joint distributions after selection

In the previous section we used empirical data to infer the nominal distribution of possible mutations. However, one important limitation of that dataset is the lack of information about to the actual set of mutation that reached fixation, which is why we turned to evolutionary simulations to model the distributions of fixed mutations in different population genetic conditions.

In the case of somatic mutations in cancer, useful data are available relating both to the nominal distribution, and to the distribution of variants after mutation and somatic expansion, i.e., cancerous growth. Tumor protein p53 (TP53) is the most frequently mutated gene in many forms of human cancer, sometimes called "the guardian of the genome" for its role in conferring genetic stability and preventing both genome mutation and cancer formation [26]. The DMS (deep mutational scanning) study by [27] may be used to depict the nominal distribution of fitness effects of p53 variants. [27] generated all 8258 possible non-synonymous amino acid changes and assayed them in human lung carcinoma cells in the presence and absence of endogenous p53. As with the previous dataset, we focus only on those amino acid changes that confer a selective advantage (combined $E_{\text{score}} > 0$, see Methods). In the context of studying cancer, an advantage refers to somatic overgrowth, thus a large fraction of p53 changes are advantageous due to loss of the protective function of p53.

To empirically characterise the mutation rates of each possible amino acid change, we used data from the Pancancer Analysis of Whole Genomes database [28]. The mutation rate for each amino acid change is the sum of the rates for each of its associated single nucleotide changes, and the rates of each nucleotide change are specified given the identity of the bases that immediately flank the mutated base (i.e., trinucleotide mutation context) (Methods). We combine these empirical estimations to construct a nominal distribution that contains mutation rates and selection coefficients for all 2443 possible beneficial amino acid changes Fig. 4.3A. The shape of the nominal distribution for TP53 is different than the one for Dengue (Fig. 4.2A), however, the correlation between mutation rates and selection coefficients is similar, being close to zero and slightly negative (Pearson's correlation coefficient $r = -0.04$ and $P = 0.03$).

The analog of the fixed distribution for TP53 variants after mutation and selection is the clinical distribution. Here we use frequency data from the GENIE database of the American Association for Cancer Research [29], allowing for the assignment of observed frequencies to each possible beneficial amino acid change identified in the nominal distribution (Methods). The distribution of fixed mutations shown in Fig. 4.3C reveals an induced positive correlation between mutation rates and selection coefficients (Pearson's correlation coefficient $r = 0.23$ and $P < 10^{-6}$). This positive correlation reverses the weak (but marginally significant) negative correlation between mutation rates and selection coefficients in the nominal distribution.

This result prompts the question of whether clonal interference might induce a positive correlation between mutation rates and selection coefficients. To address this issue, we turned again to evolutionary simulations with SLiM v3.4 [25], using the nominal distribution for TP53 described above, and with a variable mutation supply $N\mu$ (Methods). For each value of mutation supply $N\mu$, we simulated 20 data sets and calculated the correlation between the observed and simulated frequencies of fixed mutations, as well as the correlation between mutation rates and selection coefficients for the distribution of fixed mutations (Methods). The results in Fig. 4.4

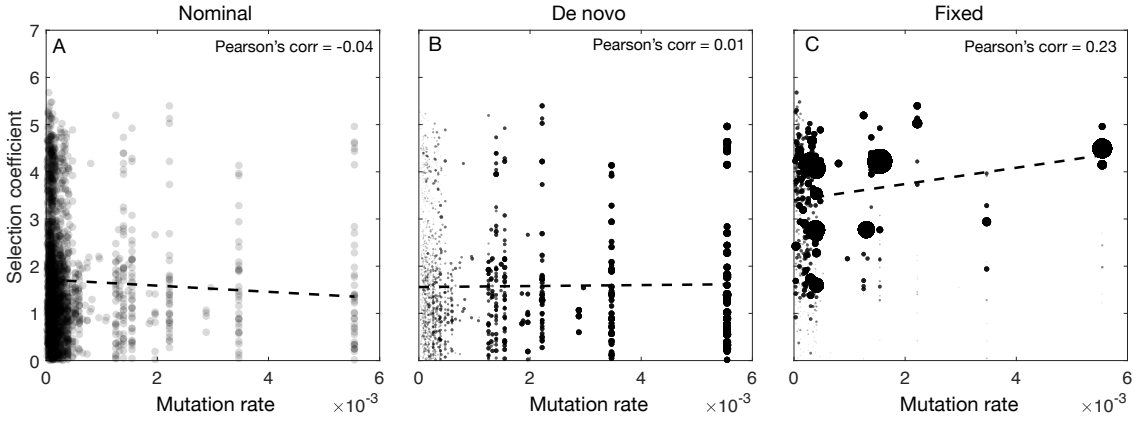

FIGURE 4.3: **Adaptation shapes the association between mutation rate and selection coefficient in TP53.** **A**. The nominal distribution shows the mutation rates and selection coefficients for all possible 2443 beneficial amino acid changes. Pearson's correlation coefficient $r = -0.04$ ($P = 0.03$). **B**. The de novo distribution shows the mutation rates and selection coefficients weighted by the mutation rates. Pearson's correlation coefficient $r = 0.01$ ($P = 0.8$). **C**. The fixed distribution shows the mutation rates and selection coefficients weighted by the observed frequency of mutations in the GENIE database. Pearson's correlation coefficient $r = 0.23$ ($P < 10^{-6}$). In panels **B** and **C** the dot size is proportional to its frequency. The dashed lines correspond to the regression lines.

indicate that the simulated results match observed results most closely when mutation supply is 10^{-2} (Fig. 4.4A), and this is also the value that maximizes the positive correlation between mutation rates and selection coefficients in the simulated adaptive changes, reaching a value of 0.1 (Fig. 4.4B). This maximum correlation, however, is considerably lower than the value of 0.25 seen in the observed TP53 data.

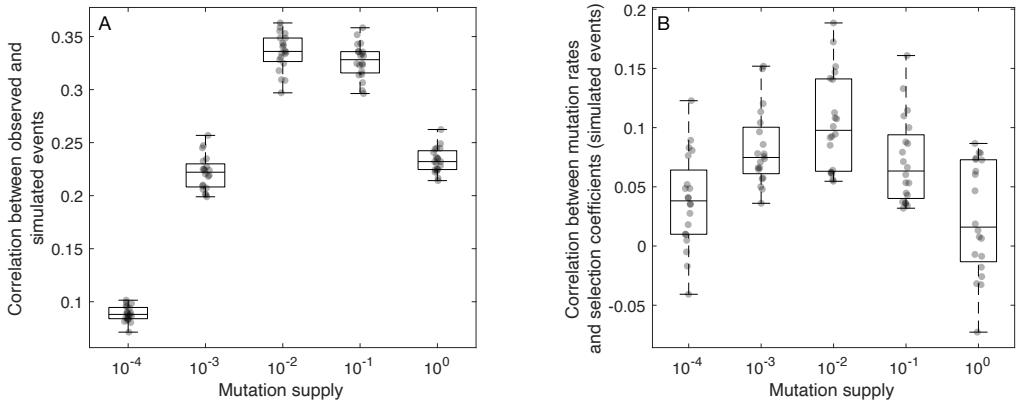

FIGURE 4.4: **Population genetic conditions shape the distribution of fixed adaptive mutations.** **A**. The correlation between observed and simulated adaptive events and **B**. The correlation between mutation rates and selection coefficients for the distribution of fixed mutations in TP53, both as a function of mutation supply ($N\mu$). Each box shows the distribution of correlations for 20 replicate simulations, and the horizontal line shows the median.

Finally, we wished to probe the effect of mutation rate further by considering multi-nucleotide mutations. Typically these are ignored, as in our simulations above, but multi-nucleotide mutations are known to occur widely in nature, at a combined rate roughly 2 orders of magnitude lower than that of single-nucleotide mutations [30], [31]. This large difference allows us to probe a

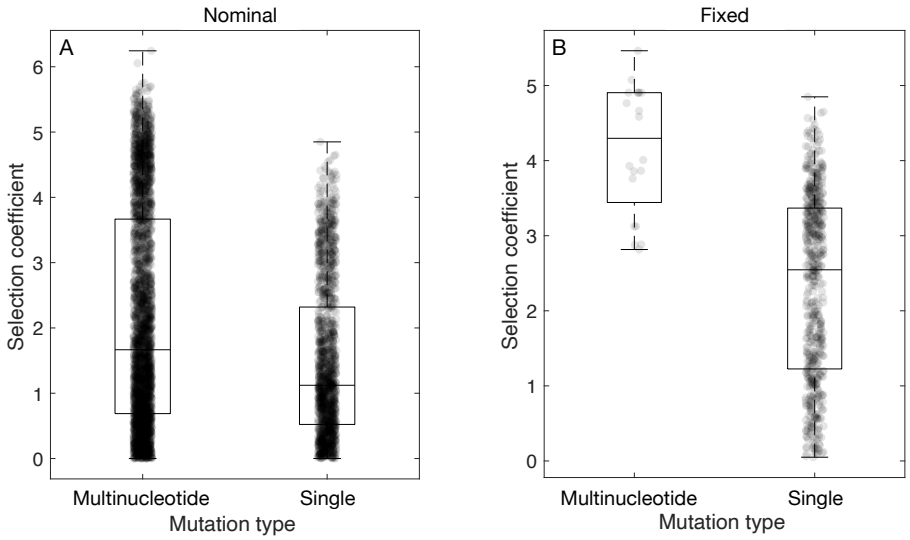

FIGURE 4.5: **Adaptation magnifies the negative association between mutation rates and selection coefficients in TP53 when including multinucleotide mutations. A.** Nominal distribution of beneficial amino acid changes. **B.** Observed distribution of fixed beneficial amino acid changes. In both panels we aggregate amino acid changes into two categories, depending on whether they are associated to single point mutations or to multinucleotide mutations.

different part of the range of mutation supply, and seem to suggest a strong expectation of Berkson's paradox: the multi-nucleotide variants that rise to prominence must have had strong selective effects to compensate for their low mutation rates, particularly in a clonal interference regime with more frequent single-nucleotide mutations.

The study of TP53 by [27] mentioned above, like most DMS studies, covers all amino acid changes, not just the 150 types of replacements (out of 380) possible via single-nucleotide mutations. Thus, we have the nominal distribution of selection coefficients for the single- and multi-nucleotide variants. In addition, the data on cancer prevalence provides the fixed distribution for multi-nucleotide variants. However, we do not have a nominal distribution of rates for multi-nucleotide mutations. In the absence of a detailed mutation rate model for multi-nucleotide mutations, we simply compare two mutational categories and apply a rank test, the Mann-Whitney U test, to the selection coefficients, reporting the chance that a random multi-nucleotide variant is fitter than a random single-nucleotide variant, which has a null expectation of 50%. The results in Fig 4.5 show a slight difference in the nominal distribution between single- vs. multi-nucleotide variants, in that a multi-nucleotide variant has a $\sim 61\%$ chance of being fitter than a single-nucleotide variant (Mann-Whitney U test, $U/(n_1 * n_2) = 0.61$, $P < 10^{-6}$, 95% CI, 0.59 to 0.62). This difference is magnified considerably in the fixed distribution (i.e., clinical prevalence). In the fixed distribution, a random multi-nucleotide variant is fitter than a random single-nucleotide variant $\sim 88\%$ of the time (Mann-Whitney U test, $U/(n_1 * n_2) = 0.88$, $P < 10^{-6}$, 95% CI, 0.82 to 0.94), which is a significantly different proportion of the time than was observed for the nominal distributions (95% bootstrap confidence intervals for the normalized Mann-Whitney test statistic $U/(n_1 * n_2)$ based on 10^3 bootstrap samples do not overlap).

4.4 DISCUSSION

The statistical consequence of evolutionary adaptation is that an underlying joint distribution of mutation rates and selection coefficients is sampled by a dual process of mutational introduction and selective filtering. Here we have developed some theory for the effects of this sampling, focusing on the case where the differential effects of mutation and selection are both strong (as in the SSWM regime or the related origin-fixation regime). We are particularly interested in whether conditioning on adaptation induces positive associations, or negative associations as with Berkson's paradox.

In theory, a variety of effects are possible, depending on higher moments of the underlying joint distributions. This dependence makes the issue much more empirical, in that expectations depend on what kinds of underlying joint distributions are actually found in nature. Using available data on mutation rates and selection coefficients from Dengue virus, we find that a simple population-genetic model of adaptation leads to a negative correlation between mutation rate and selection coefficient, one that becomes stronger when mutation supply is low. The data available on TP53 allows us to simulate the fixed distribution from a nominal distribution, and also examine the actual fixed distribution (clinical prevalence). For single-nucleotide changes, we find a positive association of mutation rate and selection intensity in clinical prevalence data. Simulations from the nominal distribution also yield a mostly positive association. However, when we compare single- to multi-nucleotide variants, we observe a weak negative association in the nominal distribution and a much stronger negative association in the fixed distribution. That is, the multi-nucleotide drivers in clinical data, which occur by mutation at a lower rate, tend to be more strongly selected than single-nucleotide drivers.

These two empirical nominal distributions (Dengue and TP53) exhibit qualitative and quantitative differences that explain the opposition in sign of the induced correlation between mutation rates and selection coefficients in the distribution of fixed mutations. While in both distributions C>T transitions occur at the highest rate, mutation bias in the Dengue dataset is several orders of magnitude higher than the one for TP53. In addition, in the Dengue nominal distribution, such mutationally favored C>T transitions generally provide a lower selective advantage, generating an "L" shape due to the lack of adaptive paths in the region corresponding to high mutation rate and high selection coefficient. Based on our analytical results, we expect this particular shape of the nominal distribution to readily induce a negative correlation in the distribution of fixed mutations. In contrast, in the case of TP53, due to both the lower mutation bias in the nominal distribution and the clonal interference in the adaptive process, selection coefficients have a considerably stronger influence in the probability of fixation than the mutation rates. Thus, the theoretical framework developed here, along with simulations, provides some guidance for understanding how associations between mutation rate and selection coefficient are influenced by the nominal distribution of beneficial mutations, its associated mutation biases, and population-genetic conditions.

The arguments that we offer based on empirical distributions must be interpreted cautiously. Clearly, researchers conducting deep mutational scanning studies have worked hard to improve measurements of fitness (and other functional effects) so that they are not confounded by effects of mutability (e.g., [7]). Likewise, mutation-accumulation studies are designed to remove effects

of selection, so as to accurately measure mutational spectra [32]. However, our analysis here, focusing specifically on correlations, subjects these data to a much higher level of scrutiny of the joint distribution than was imagined for the original uses of the data. With clinical data on cancer, for instance, there is no clear quantitative standard of ascertainment that specifies what qualifies as rapidly growing cancerous tissue. Ultimately, our understanding of what types of associations are induced in nature may depend on new methods designed to characterize a joint distribution without bias.

With these caveats in mind, the significance of the results reported here are that (1) in theory, the dual causation of adaptation can induce strong associations between mutation rates and selection coefficients, and (2) in theory, the shape of the underlying nominal joint distribution matters for the size and direction of effects, and (3) in practice, according to limited evidence currently available, both of these theoretical points are relevant to natural cases. Though these results are modest, they have some immediate implications. For instance, Stoltzfus and Norris [33], evaluating the hypothesis that transitions are more conservative (transversions more damaging) in their effects on proteins, pointed to the lack of an advantage of transitions among adaptive changes as evidence against the hypothesis. However, the observed small advantage of transversions among adaptive changes instead may represent a case of Berkson's paradox, i.e., the slight fitness advantage of transversions in the fixed distribution may represent the kind of compensation for a lower mutational rate of arrival induced by conditioning on a dual arrival-fixation process.

Some other types of data may also reveal evidence of negative associations in the fixed distribution, though detailed quantitative data to resolve this issue often are unavailable. For instance, Cannataro *et. al* note with interest that the most clinically common cancer driver mutations in two types of cancer are not the fastest growing, and their Fig. 2 appears to show a negative correlation between selection intensity and mutation rates [15] (as does the comparable Fig. 3 of [34]). Likewise, Fig. 2 of [16] shows a slight negative association between mutation rate and selection intensity for changes associated with clonal haematopoesis. Whether such results are surprising or not depends strongly on the underlying joint distributions, which typically are unknown.

As noted above, the data used in this study were not designed to be used in this way and are not ideal. In this context, it may be helpful to imagine the characteristics of an ideal experimental system. Foremost, the ideal system would allow us to interrogate the nominal distribution systematically and with precision. The ideal system would have a target for adaptation that includes many (dozens or hundreds) of possible beneficial changes with a range of selection coefficients and mutation rates. Perhaps the target of adaptation would be tunably narrow or broad. An ideal system also would be amenable to evolution experiments that cover a broad range of population-genetic conditions, including modulation of the mutation supply and of the strength of selection. For instance, imagine a bacterial system in which one may study resistance to a broad or narrow range of anti-microbials. The strength of selection could be modulated by the concentration of antimicrobials. The mutation supply could be modulated by the population size. Droplet technology as in [35] makes it possible to design high-throughput evolution experiments with many replicates of a very small population size.

Finally, let us consider two general directions for further theoretical work. The first concerns the differentiating power of selection. Although our focus here has been on cases of strong

positive selection in adaptation and cancer, various modes of evolutionary change in reproducing organisms, including neutral evolution, also depend dually on the kinetics of the introduction of variants and on their differential reproductive sorting. The size biases imposed by mutation and selection will differ depending on conditions. When mutation supply is low, i.e., origin-fixation conditions, the effect of mutation is proportional to u and the effect of selection is given by the probability of fixation, which generally is a function of s and N, per [36], regardless of whether mutations are deleterious, neutral, or beneficial (whereas our mathematical theory only covers the special case where $p_{fix} \approx 2s$). As mutation supply increases, clonal interference comes into play and, where studied, this amplifies the effect of selection and diminishes the influence of mutation bias (e.g., [1], [3]). Presumably, strong effects of conditioning will emerge under a variety of conditions, wherever effects of mutation and selection are strong.

A more challenging theoretical issue is what will happen to the joint distribution of mutation rates and selection coefficients in long-term evolution. Oddly, this more complex issue has received considerably more attention in the literature than the simpler foundational issue addressed here, with a variety of provocative results. [37] explore the issue of what happens to the overall mutation pattern under context-dependence, e.g., mutation hotspots will tend to disappear and this may be expected to lower the total mutation rate over time. Some results refer to patterns of reduction in deleterious mutation. Although classic theory focuses on the total rate of mutation U [38], [10] suggested how a pattern of mutation-reduction in genes vs. non-genes could emerge when repair mechanisms leverage structural features associated with functionality. Other results refer to adaptation or innovation. In adaptive walks, simply reversing mutation biases partway through the walk appears to improve adaptation by shifting the de novo distribution to probabilize previously unlikely benefits [39]. For polygenic quantitative traits, continued evolution under correlated selection can lead to a shift in how the traits are encoded so that mutational variability aligns more closely with trait correlations favored by selection [40]. Constructional selection per [41] enhances the mutational contribution of evolvable modules, making further innovation more likely. Thus, a challenge for future theoretical work is to consider such results together in a common framework for mutation-selection associations.

4.5 METHODS

Nominal distribution of adaptive mutations in Dengue

We use data containing estimates of selection coefficients and mutation rates for all possible single point mutations in the Dengue genome, provided in Extended Table Data 1 ([6]). The mutation rates are described for each of the 12 possible single nucleotide changes, regardless of their sequence context. We curated the dataset to only include beneficial mutations (fitnessLowerCI > 1 in Extended Table Data 1) for passage 9, replicate A with human host cells. The final list contains 325 beneficial mutations.

Nominal and fixed distributions of adaptive mutations in TP53

The study from Giacomelli *et. al* [27] provide enrichment scores for the wild-type as well as for all 8258 possible non-synonymous variants of TP53 in selection assays under the presence and absence of endogenous p53: WT TP53 (p53WT) and null TP53 (p53NULL), respectively, in isogenic human lung carcinoma cell populations. They treated those populations with two p53-activating agents, nutlin3 or etoposide for 12 days in three selection assays designed to enrich for dominant-negative (WT+nutlin3), lost of function (NULL+nutlin3), or WT-like (NULL+etoposide) alleles. We combined the estimated enrichments for the three assays ($E_{\text{score}}^{(\text{WT+nutlin3})} + E_{\text{score}}^{(\text{NULL+nutlin3})} - E_{\text{score}}^{(\text{NULL+etoposide})}$), and calculated selection coefficients for all possible non-synonymous variants relative to the enrichment of the wild-type TP53.

To construct the nominal distribution, we assigned mutation rates to all possible non-synonymous variants using data from the Pancancer Analysis of Whole Genomes database [28]. We queried a total of 28717344 whole genome single point somatic mutations to construct a trinucleotide mutational signature, that is, the mutation rates specified by the identity of the bases that immediately flank the mutated base (Fig. S4.3). We assigned mutation rates to all possible beneficial amino acid changes summing up the rates for each of its associated nucleotide changes.

We turned to a third independent dataset to extract the observed frequency of 639 different somatic mutations in TP53 in human tumors from the GENIE database of the American Association for Cancer Research Consortium [29]. We integrated the selection coefficients, mutation rates and observed frequencies of beneficial variants to construct the distribution of observed fixed mutations.

Evolutionary simulations

We used SLiM v3.4 for the evolutionary simulations [25]. We ran each simulation until the first fixation event, repeating this process 500 times for the Dengue dataset and 1000 times for the TP53 dataset, recording only the beneficial mutations that went to fixation. We performed several replicates per value of mutation supply $N\mu$, 3 for the Dengue dataset and 20 for the TP53 dataset. Each of the simulations per replicate used the same initial population, which comprised N copies of the wild-type sequence of Dengue's genome and TP53 coding region. All sequences in the initial population were assigned a fitness of one. In each generation t, N sequences were chosen from the population at generation $t-1$ with replacement and with a probability proportional to their fitness. The fitness effects assigned to each of the possible adaptive changes were taken from their respective datasets (single point mutation for Dengue and amino acid change for TP53).

4.6 ACKNOWLEDGMENTS

The identification of any specific commercial products is for the purpose of specifying a protocol, and does not imply a recommendation or endorsement by the National Institute of Standards and Technology. This work was made possible through the support of a grant from the John Templeton Foundation (grant #61782, D.M.M.) and from the Swiss National Science Foundation (grants #PP00P3_170604 and #310030_192541, J.L.P.). The opinions expressed in this publication

are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. D.M.M. also acknowledges additional support from an Alfred P. Sloan Research Fellowship and from the Simons Center for Quantitative Biology.

REFERENCES

[1] K. Gomez, J. Bertram, and J. Masel, "Mutation bias can shape adaptation in large asexual populations experiencing clonal interference", *Proceedings of the Royal Society B: Biological Sciences*, vol. 287, no. 1937, p. 20 201 503, 2020.

[2] A. Stoltzfus, *Mutation, Randomness and Evolution*. Oxford, 2021.

[3] A. V. Cano, H. Rozhoňová, A. Stoltzfus, D. M. McCandlish, and J. L. Payne, "Mutation bias shapes the spectrum of adaptive substitutions", *Proceedings of the National Academy of Sciences*, vol. 119, no. 7, e2119720119, 2022.

[4] S. Greenland, "Quantifying biases in causal models: Classical confounding vs collider-stratification bias", *Epidemiology*, vol. 14, no. 3, pp. 300–6, 2003.

[5] J. Berkson, "Limitations of the application of fourfold table analysis to hospital data", *Biometrics Bulletin*, vol. 2, no. 3, pp. 47–53, 1946.

[6] P. T. Dolan, S. Taguwa, M. A. Rangel, *et al.*, "Principles of dengue virus evolvability derived from genotype-fitness maps in human and mosquito cells", *eLife*, vol. 10, e61921, 2021.

[7] A. Acevedo, L. Brodsky, and R. Andino, "Mutational and fitness landscapes of an rna virus revealed through population sequencing", *Nature*, vol. 505, no. 7485, pp. 686–690, 2014.

[8] R. E. Lenski and J. E. Mittler, "The directed mutation controversy and neo-darwinism", *Science*, vol. 259, no. 5092, pp. 188–94, 1993.

[9] L. M. Chevin, G. Martin, and T. Lenormand, "Fisher's model and the genomics of adaptation: Restricted pleiotropy, heterogenous mutation, and parallel evolution", *Evolution*, vol. 64, no. 11, pp. 3213–31, 2010.

[10] I. Martincoreña and N. M. Luscombe, "Non-random mutation: The evolution of targeted hypermutation and hypomutation", *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 35, no. 2, pp. 123–30, 2013.

[11] J Monroe, T. Srikant, P. Carbonell-Bejerano, *et al.*, "Mutation bias reflects natural selection in *Arabidopsis thaliana*", *Nature*, pp. 1–5, 2022.

[12] J. Foley, "Mini-review: Strategies for variation and evolution of bacterial antigens", *Computational and Structural Biotechnology Journal*, vol. 13, pp. 407–16, 2015.

[13] J. R. Roth, E. Kugelberg, A. B. Reams, E. Kofoid, and D. I. Andersson, "Origin of mutations under selection: The adaptive mutation controversy", *Annual Review of Microbiology*, vol. 60, pp. 477–501, 2006.

[14] C. Maclean, G. Perron, and A. Gardner, "Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for Rifampicin resistance in *Pseudomonas aeruginosa*", *Genetics*, vol. 186, pp. 1345–54, 2010.

[15] V. L. Cannataro, S. G. Gaffney, and J. P. Townsend, "Effect sizes of somatic mutations in cancer", *Journal of the National Cancer Institute*, vol. 110, no. 11, pp. 1171–1177, 2018.

[16] C. J. Watson and J. R. Blundell, "Mutation rates and fitness consequences of mosaic chromosomal alterations in blood", *bioRxiv*, p. 2022.05.07.491016, 2022.

[17] R. Arratia, L. Goldstein, and F. Kochman, "Size bias for one and all", *Probability Surveys*, vol. 16, pp. 1–61, 2019.

[18] D. C. Queller, "Fundamental theorems of evolution", *The American Naturalist*, vol. 189, no. 4, pp. 345–353, 2017.

[19] J. H. Gillespie, *The causes of molecular evolution*. Oxford University Press On Demand, 1994, vol. 2.

[20] D. McCandlish and A. Stoltzfus, "Modeling evolution using the probability of fixation: History and implications", *Quarterly Review of Biology*, vol. 89, no. 3, pp. 225–252, 2014.

[21] J. Haldane, "A mathematical theory of natural and artificial selection. v. selection and mutation.", *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26, pp. 220–230, 1927.

[22] W.-H. Li, "Maintenance of genetic variability under the joint effect of mutation, selection and random drift", *Genetics*, vol. 90, pp. 349–382, 2 1978.

[23] L. Y. Yampolsky and A. Stoltzfus, "Bias in the introduction of variation as an orienting factor in evolution", *Evolution & Development*, vol. 3, no. 2, pp. 73–83, 2001.

[24] G. Sella and A. E. Hirsh, "The application of statistical physics to evolutionary biology", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9541–9546, 2005.

[25] P. W. Messer, *SLiM: Simulating evolution with selection and linkage*, 2013.

[26] P. Hainaut and G. P. Pfeifer, "Somatic TP53 mutations in the era of genome sequencing", *Cold Spring Harbor Perspectives in Medicine*, vol. 6, no. 11, a026179, 2016.

[27] A. O. Giacomelli, X. Yang, R. E. Lintner, *et al.*, "Mutational processes shape the landscape of tp53 mutations in human cancer", *Nature Genetics*, vol. 50, no. 10, pp. 1381–1387, 2018.

[28] J. N. Weinstein, E. A. Collisson, G. B. Mills, *et al.*, "The cancer genome atlas pan-cancer analysis project", *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[29] AACR Project Genie Consortium *et al.*, "Aacr project genie: Powering precision medicine through an international consortium", *Cancer discovery*, vol. 7, no. 8, pp. 818–831, 2017.

[30] N. G. Smith, M. T. Webster, and H. Ellegren, "A low rate of simultaneous double-nucleotide mutations in primates", *Molecular Biology and Evolution*, vol. 20, no. 1, pp. 47–53, 2003.

[31] D. R. Schrider, J. N. Hourmozdi, and M. W. Hahn, "Pervasive multinucleotide mutational events in eukaryotes", *Curr Biol*, vol. 21, no. 12, pp. 1051–4, 2011.

[32] V. Katju and U. Bergthorsson, "Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches", *Genome Biology and Evolution*, vol. 11, no. 1, pp. 136–165, 2019.

[33] A. Stoltzfus and R. W. Norris, "On the causes of evolutionary transition:transversion bias", *Molecular Biology and Evolution*, vol. 33, no. 3, pp. 595–602, 2016.

[34] V. L. Cannataro, S. G. Gaffney, T. Sasaki, *et al.*, "APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma", *Oncogene*, vol. 38, no. 18, pp. 3475–3487, 2019.

[35] P. Ruelens and J. de Visser, "Clonal interference and mutation bias in small bacterial populations in droplets", *Genes (Basel)*, vol. 12, no. 2, 2021.

[36] M Kimura, "On the probability of fixation of mutant genes in a population", *Genetics*, vol. 47, no. 6, pp. 713–719, 1962.

[37] M. Oman, A. Alam, and R. W. Ness, "How sequence context-dependent mutability drives mutation rate variation in the genome", *Genome Biology and Evolution*, vol. 14, no. 3, 2022.

[38] M. Lynch, "Evolution of the mutation rate", *Trends in Genetics*, vol. 26, no. 8, pp. 345–52, 2010.

[39] M. Sane, G. D. Diwan, B. A. Bhat, L. M. Wahl, and D. Agashe, "Shifts in mutation spectra enhance access to beneficial mutations", *bioRxiv*, p. 2020.09.05.284158, 2022.

[40] A. G. Jones, R. Burger, and S. J. Arnold, "Epistasis and natural selection shape the mutational architecture of complex traits", *Nature Communications*, vol. 5, p. 3709, 2014.

[41] L. Altenberg, "Evolution of the genotype-phenotype map", in *Evolution and Biocomputation: Computational Models of Evolution*, ser. Springer-Verlag Lecture Notes in Computer Sciences. Springer-Verlag, 1995, vol. 899, pp. 205–259.

4.7 SUPPLEMENTARY MATERIAL

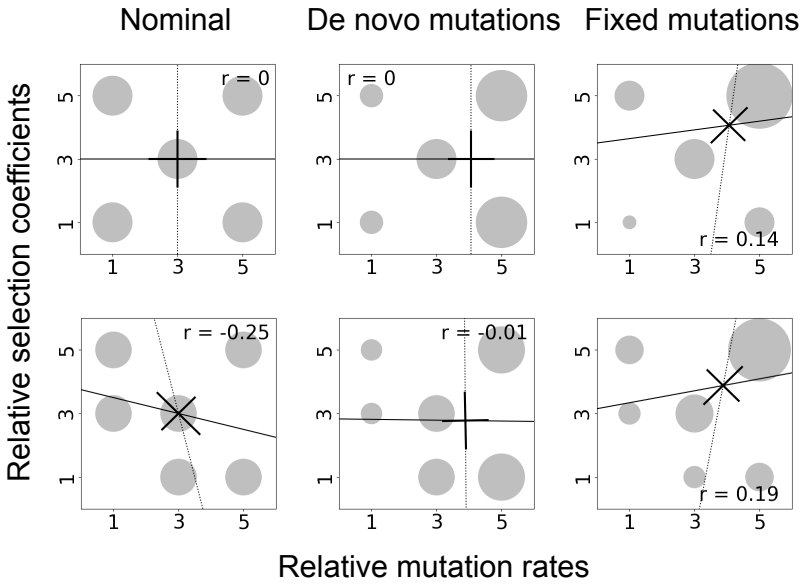

FIGURE S4.1: **Correlations of mutation and selection under some alternative models to Fig. 1.** Mutation rates and selection coefficients each take on 3 possible values, similar to the bottom 3 rows in Fig. 1. As in the case of Fig. 1, the columns from left to right show the nominal distribution (representing the landscape of possible mutations across varying mutation rates and selection coefficients), the distribution of de novo mutations (the possible changes weighted by mutation rates), and the distribution of changes resulting from a mutation-fixation process. Transecting lines show the regression of selection on mutation (solid) and mutation on selection (dotted); central crossed lines show the first and second principal components, with the lengths proportional to the variance explained, where by definition $V_1 \geq V_2$.

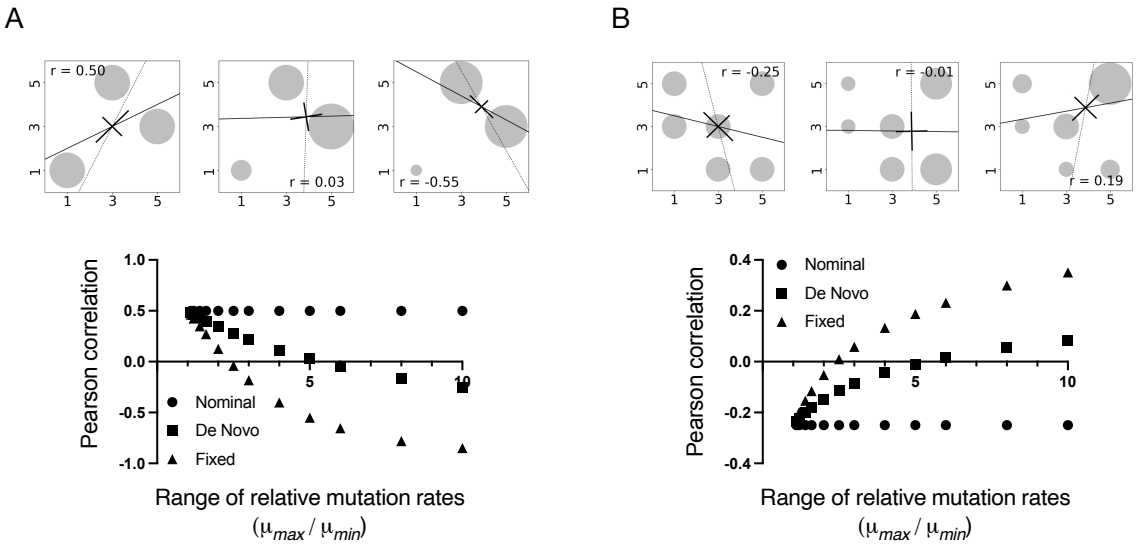

FIGURE S4.2: **Correlations of mutation and selection as a function of the magnitude of size-biasing. A.** Correlations between mutation and selection among nominal, de novo, and fixed mutational distributions in a case where conditioning on mutation fixation induces a sign-change in the correlation from positive in the nominal distribution to negative in the fixed distribution. Top: the nominal, de novo, and fixed distributions when mutation rate and selection coefficient each vary by 5-fold between their highest and lowest respective values (same as Fig. 1, fourth row). Bottom: the correlations between mutation and selection for the same 3-mutation-class distribution shown above, plotted as a function of the range, or fold-change, between highest and lowest value of mutation rate. **B.** Correlations between mutation and selection among nominal, de novo, and fixed mutational distributions in a case where conditioning on mutation fixation induces a sign-change in the correlation from negative in the nominal distribution to positive in the fixed distribution. Top: similar to panel **A**, the nominal, de novo, and fixed distributions when mutation rate and selection coefficient each vary by 5-fold between their highest and lowest respective values (same as Fig. S4.1, second row). Bottom: the correlations between mutation and selection for the same 3-mutation-class distribution shown above, similar to panel **A**, plotted as a function of the range, or fold-change, between highest and lowest value of mutation rate. To derive each correlation in the bottom plots, the range or, fold-change, of selection coefficients is updated to match that of mutation rates.

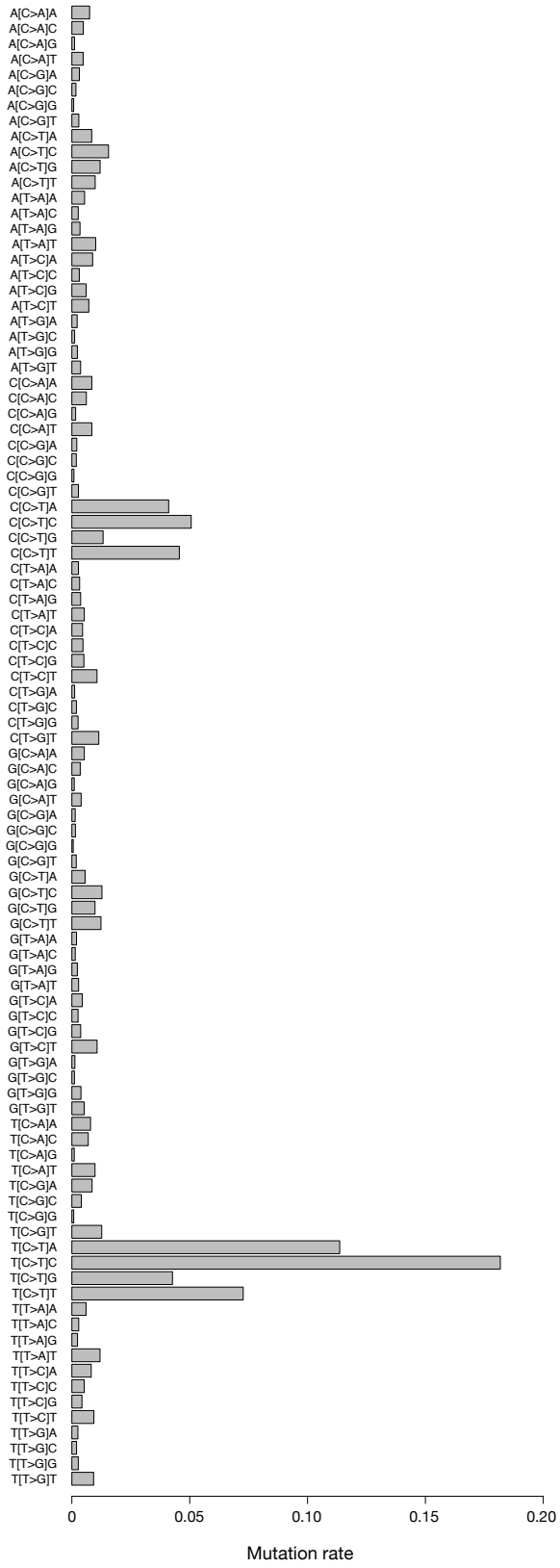

FIGURE S4.3: **Mutational signature.** We used 28717344 whole genome single point somatic mutations from the Pancancer Analysis of Whole Genomes database [28] to estimate a trinucleotide mutational signature, namely, the mutation rates specified by the identity of the bases that immediately flank the mutated base.

Supplemental information

The relationship between mutation and selection in the simplest case: Two possible values (high and low) for each

The covariance and correlation coefficient between mutation rate u and selection coefficient s is derived for a simple scenario. Consider mutations that could have either of two values of mutation rate, m_i, as well as either of two values of selection coefficient, s_j. By normalizing to the lower mutation rate and selection coefficient, the relative values of mutation rate m_i can be 1 or k, where $k > 1$, and the relative value of selection coefficient s_j can be 1 or b, where $b > 1$.

The relationship between mutation and selection: possible mutations (nominal distribution)

The proportion of possible mutations, p_{ij}, therefore falls into one of four classes (p_{11}, p_{k1}, p_{1b}, p_{kb}), corresponding to the four possible combinations of mutation rate and selection coefficient. Because p_{ij} represents the proportion of mutations with mutation rate i and selection coefficient j, the proportion of possible mutations across all combinations of mutation rate and selection coefficient add up to 1:

$$\sum_{ij} p_{ij} = p_{11} + p_{k1} + p_{1b} + p_{kb} = 1$$

Expected mutation rate u:

$$E_{poss}[u] = \sum_{i} u_i p_{ij}$$

Expected selection coefficient s:

$$E_{poss}[s] = \sum_{j} s_j p_{ij}$$

Variance of u:

$$var_{poss}(u) = E_{poss}[u^2] - E_{poss}[u]^2$$

$$= \sum_{ij} u_i^2 p_{ij} - \left(\sum_{i} u_i p_{ij} \right)^2$$

Variance of s:

$$var_{poss}(s) = E_{poss}[s^2] - E_{poss}[s]^2$$

$$= \sum_{j} s_j^2 p_{ij} - \left(\sum_{j} s_j p_{ij} \right)^2$$

Covariance between u and s:

$$cov_{poss}(u, s) = E_{poss}[ms] - E_{poss}[u] * E_{poss}[s]$$

$$= \sum_{ij} u_i s_j p_{ij} - \left(\sum_i u_i p_{ij}\right)\left(\sum_j s_j p_{ij}\right)$$

Substituting the values 1 and k for mutation rate and the values 1 and b for selection coefficient yields the following covariance expression:

$$\mathrm{cov}_{poss}(u,s) = (p_{11} + p_{k1}k + p_{1b}b + p_{kb}kb) - (p_{11} + p_{k1}k + p_{1b} + p_{kb}k) * (p_{11} + p_{k1} + p_{1b}b + p_{kb}b)$$

By substituting the following value for p_{11},

$$p_{11} = 1 - p_{k1} - p_{1b} - p_{kb}$$

the covariance simplifies as follows:

$$\begin{aligned}\mathrm{cov}_{poss}(u,s) &= (1 - p_{kb} - p_{k1} - p_{1b} + p_{k1}k + p_{1b}b + p_{kb}kb) \\ &\quad - (1 - p_{kb} - p_{k1} + p_{k1}k + p_{kb}k)(1 - p_{kb} - p_{1b} + p_{1b}b + p_{kb}b)\end{aligned}$$

$$= p_{kb}(kb-1) + p_{k1}(k-1) + p_{1b}(b-1) + 1 - [1 + (p_{kb} + p_{1b})(b-1)][1 + (p_{kb} + p_{k1})(k-1)]$$

$$= p_{kb}(kb-1) - p_{kb}(k-1) - p_{kb}(b-1) - (p_{kb} + p_{1b})(p_{kb} + p_{k1})(b-1)(k-1)$$

$$= p_{kb}(kb - k - b + 1) - (p_{kb} + p_{1b})(p_{kb} + p_{k1})(b-1)(k-1)$$

$$= p_{kb}(b-1)(k-1) - (p_{kb} + p_{1b})(p_{kb} + p_{k1})(b-1)(k-1)$$

$$= [p_{kb} - (p_{kb}^2 + p_{kb}p_{1b} + p_{kb}p_{k1} + p_{k1}p_{1b})](b-1)(k-1)$$

$$= [p_{kb}(1 - p_{kb} - p_{1b} - p_{k1}) - p_{k1}p_{1b}](b-1)(k-1)$$

$$= (p_{kb}p_{11} - p_{k1}p_{1b})(b-1)(k-1)$$

Correlation coefficient between u and s:

$$\rho_{poss}(u,s) = \frac{\mathrm{cov}_{poss}(u,s)}{\sqrt{\mathrm{var}_{poss}(u)\mathrm{var}_{poss}(s)}}$$

$$= \frac{(p_{11}p_{kb} - p_{1b}p_{k1})(b-1)(k-1)}{\sqrt{(b-1)^2(k-1)^2[p_{k1} + p_{kb} - (p_{k1} + p_{kb})^2] * [p_{1b} + p_{kb} - (p_{1b} + p_{kb})^2]}}$$

$$= \frac{p_{11}p_{kb} - p_{1b}p_{k1}}{\sqrt{p_k(1-p_k) * p_b(1-p_b)}}$$

The relationship between mutation and selection: fixed mutations

Mutations of each class (p_{11}, p_{k1}, p_{1b}, p_{kb}) fix at the following relative rate:

$$f_{ij} = \frac{p_{ij}u_i s_j}{\sum \mathrm{pms}}$$

In this formula, the proportion of possible mutations p_{ij} for each mutation rate and selection coefficient is multiplied by the respective mutation rate and selection coefficient and divided by the summation over all products of possible mutations, mutation rates, and selection coefficients.

Expected mutation rate u:

$$E_{fix}[u] = \sum_i u_i f_{ij} = \sum_{ij} u_i \left(\frac{p_{ij} u_i s_j}{\sum pms}\right) = \frac{\sum_{ij} p_{ij} u_i^2 s_j}{\sum pms} = \frac{E_{poss}[u^2 s]}{E_{poss}[ms]}$$

Expected selection coefficient s:

$$E_{fix}[s] = \sum_j s_j f_{ij} = \sum_{ij} s_j \left(\frac{p_{ij} u_i s_j}{\sum pms}\right) = \frac{\sum_{ij} p_{ij} u_i s_j^2}{\sum pms} = \frac{E_{poss}[us^2]}{E_{poss}[ms]}$$

Variance of u:

$$var_{fix}(u) = E_{fix}[u^2] - E_{fix}[u]^2 = \sum_{ij} u_i^2 \left(\frac{p_{ij} u_i s_j}{\sum pms}\right) - \left(\frac{\sum_{ij} p_{ij} u_i^2 s_j}{\sum pms}\right)^2$$

$$= \frac{\sum_{ij} p_{ij} u_i^3 s_j}{\sum pms} - \left(\frac{\sum_{ij} p_{ij} u_i^2 s_j}{\sum pms}\right)^2$$

Variance of s:

$$var_{poss}(s) = E_{poss}[s^2] - E_{poss}[s]^2 = \sum_{ij} s_j^2 \left(\frac{p_{ij} u_i s_j}{\sum pms}\right) - \left(\frac{\sum_{ij} p_{ij} u_i s_j^2}{\sum pms}\right)^2$$

$$= \frac{\sum_{ij} p_{ij} u_i s_j^3}{\sum pms} - \left(\frac{\sum_{ij} p_{ij} u_i s_j^2}{\sum pms}\right)^2$$

Covariance between u and s:

$$cov_{fix}(u, s) = E_{fix}[ms] - E_{fix}[u] * E_{fix}[s] = \frac{E_{poss}[u^2 s^2]}{E_{poss}[ms]} - \frac{E_{poss}[u^2 s]}{E_{poss}[ms]} * \frac{E_{poss}[us^2]}{E_{poss}[ms]}$$

$$= \frac{E_{poss}[u^2 s^2] * E_{poss}[ms] - E_{poss}[u^2 s] * E_{poss}[us^2]}{E_{poss}[ms]^2}$$

Substituting the values 1 and k for mutation rate and the values 1 and b for selection coefficient yields the following covariance expression:

$$cov_{fix}(u, s) = E_{fix}[ms] - E_{fix}[u] * E_{fix}[s]$$

$$= \frac{E_{poss}[u^2 s^2]}{E_{poss}[ms]} - \frac{E_{poss}[u^2 s]}{E_{poss}[ms]} * \frac{E_{poss}[us^2]}{E_{poss}[ms]}$$

$$= \frac{E_{poss}[u^2 s^2] * E_{poss}[ms] - E_{poss}[u^2 s] * E_{poss}[us^2]}{E_{poss}[ms]^2}$$

$$= \frac{(p_{11}+p_{k1}k^2+p_{1b}b^2+p_{kb}k^2b^2)*(p_{11}+p_{k1}k+p_{1b}b+p_{kb}kb)-(p_{11}+p_{k1}k^2+p_{1b}b+p_{kb}k^2b)(p_{11}+p_{k1}k+p_{1b}b^2+p_{kb}kb^2)}{(p_{11}+p_{k1}k+p_{1b}b+p_{kb}kb)^2}$$

$$= \frac{kbp_{11}p_{kb}(kb+1-k-b)+kbp_{1b}p_{k1}(k+b-1-kb)}{(p_{11}+p_{k1}k+p_{1b}b+p_{kb}kb)^2}$$

$$= \frac{kb(p_{11}p_{kb}-p_{1b}p_{k1})(kb+1-k-b)}{(p_{11}+p_{k1}k+p_{1b}b+p_{kb}kb)^2}$$

$$= \frac{kb(p_{11}p_{kb}-p_{1b}p_{k1})(k-1)(b-1)}{(p_{11}+p_{k1}k+p_{1b}b+p_{kb}kb)^2}$$

Correlation coefficient between u and s:

$$\rho_{fix}(u,s) = \frac{cov_{fix}(u,s)}{\sqrt{var_{fix}(u)var_{fix}(s)}}$$

$$= \frac{kb(p_{11}p_{kb}-p_{1b}p_{k1})(k-1)(b-1)}{(p_{11}+p_{k1}k+p_{1b}b+p_{kb}kb)^2} / \sqrt{\frac{(k-1)^2k(bp_{1b}+p_{11})(bp_{kb}+p_{k1})}{(bkp_{kb}+bp_{1b}+kp_{k1}+p_{11})^2} * \frac{(b-1)^2b(kp_{k1}+p_{11})(kp_{kb}+p_{1b})}{(bkp_{kb}+bp_{1b}+kp_{k1}+p_{11})^2}}$$

$$= \frac{kb(p_{11}p_{kb}-p_{1b}p_{k1})}{\sqrt{(bp_{1b}+p_{11})(kbp_{kb}+kp_{k1})(kp_{k1}+p_{11})(bkp_{kb}+bp_{1b})}}$$

$$= \frac{kb(p_{11}p_{kb}-p_{1b}p_{k1})}{\sqrt{E_{poss}[ms]^4 * p_b^{fix}(1-p_b^{fix})p_k^{fix}(1-p_k^{fix})}}$$

$$= \frac{kb}{E_{poss}[ms]^2} * \frac{p_{11}p_{kb}-p_{1b}p_{k1}}{\sqrt{p_b^{fix}(1-p_b^{fix})p_k^{fix}(1-p_k^{fix})}}$$

Argument 1: a correlation between mutation and selection among fixed mutations requires that possible mutations be non-uniformly distributed across values of mutation rate and selection coefficient

Under the simplest assumption, whereby the nominal mutations are equally represented across mutation rates and selection coefficients (that is, the list of possible mutations that could have a higher mutation rate or selection coefficient is not longer than the list of possible mutations that could have a lower mutation rate or selection coefficient: $p_{11} = p_{k1} = p_{1b} = p_{kb}$), then the covariance between u and s among fixed mutations is zero:

$$cov_{fix}(u,s) = \frac{kb(p_{11}p_{kb}-p_{1b}p_{k1})(k-1)(b-1)}{(p_{11}+p_{k1}k+p_{1b}b+p_{kb}kb)^2} = \frac{kb(0)(k-1)(b-1)}{(p_{11}+p_{k1}k+p_{1b}b+p_{kb}kb)^2} = 0$$

Since the correlation coefficient depends on the covariance,

$$\rho_{fix}(u,s) = \frac{cov_{fix}(u,s)}{\sqrt{var_{fix}(u)var_{fix}(s)}}$$

a covariance of zero implies a correlation coefficient of zero. In other words, unless the number of possible (nominal) mutations already differs across different values of mutation rate, selection coefficient, or both, then mutation rate and selection coefficient will not be correlated even among fixed mutations.

Argument 2: mutation and selection will be uncorrelated among fixed mutations if they are uncorrelated among possible mutations

How does the sign (positive, negative, or zero) of the correlation between u and s among possible mutations compare to the correlation among fixed mutations? Because the sign of the correlation coefficient depends on the covariance (the numerator of the correlation coefficient), the relationship between the correlation coefficients among possible versus fixed mutations with depend on the relationship between their covariances.

If mutation and selection are uncorrelated among possible mutations, their covariance will equal zero:

$$\mathrm{cov}_{\mathrm{poss}}(u,s) = (p_{kb}p_{11} - p_{k1}p_{1b})(b-1)(k-1) = 0$$

Since k and b represent the higher relative values of mutation rate and selection coefficient, respectively, the terms (k – 1) and (b – 1) are both positive. Accordingly, for the covariance between mutation and selection to be zero among possible mutations, the difference of the diagonal products of p_{ii} ($p_{kb}p_{11} - p_{k1}p_{1b}$) must be zero:

$$\mathrm{cov}_{\mathrm{poss}}(ms) = (p_{kb}p_{11} - p_{k1}p_{1b})(b-1)(k-1) = (0)(b-1)(k-1) = 0$$

$$p_{kb}p_{11} - p_{k1}p_{1b} = 0$$

$$p_{kb}p_{11} = p_{k1}p_{1b}$$

Because the difference of the diagonal products of p_{ii} ($p_{kb}p_{11} - p_{k1}p_{1b}$) also occurs in the covariance among fixed mutations, the covariance among fixed mutations will also be zero:

$$\mathrm{cov}_{\mathrm{fix}}(u,s) = \frac{kb(p_{11}p_{kb} - p_{1b}p_{k1})(k-1)(b-1)}{(p_{11} + p_{k1}k + p_{1b}b + p_{kb}kb)^2} = \frac{kb(0)(k-1)(b-1)}{(p_{11} + p_{k1}k + p_{1b}b + p_{kb}kb)^2} = 0$$

Argument 3: mutation and selection are positively correlated among fixed mutations if they are positively correlated among possible mutations

If mutation and selection are positively correlated among possible mutations, their covariance will be positive:

$$\mathrm{cov}_{\mathrm{poss}}(ms) = (p_{kb}p_{11} - p_{k1}p_{1b})(b-1)(k-1) > 0$$

If the covariance among possible mutations is positive, the terms (b – 1) and (k – 1) must be multiplied by a positive value:

$$\mathrm{cov}_{\mathrm{poss}}(ms) = (p_{kb}p_{11} - p_{k1}p_{1b})(b-1)(k-1) > 0$$

$$p_{kb}p_{11} - p_{k1}p_{1b} > 0$$

$$p_{kb}p_{11} > p_{k1}p_{1b}$$

The covariance among fixed mutations will also be positive:

$$\mathrm{cov}_{\mathrm{fix}}(u,s) = \frac{kb(p_{11}p_{kb} - p_{1b}p_{k1})(k-1)(b-1)}{(p_{11} + p_{k1}k + p_{1b}b + p_{kb}kb)^2} > 0$$

Argument 4: mutation and selection are negatively correlated among fixed mutations if they are negatively correlated among possible mutations

If mutation and selection are negatively correlated among possible mutations, their covariance will be negative:

$$\mathrm{cov}_{\mathrm{poss}}(ms) = (p_{kb}p_{11} - p_{k1}p_{1b})(b-1)(k-1) < 0$$

If the covariance among possible mutations is negative, the terms (b – 1) and (k – 1) must be multiplied by a negative value:

$$\mathrm{cov}_{\mathrm{poss}}(ms) = (p_{kb}p_{11} - p_{k1}p_{1b})(b-1)(k-1) < 0$$

$$p_{kb}p_{11} - p_{k1}p_{1b} < 0$$

$$p_{kb}p_{11} < p_{k1}p_{1b}$$

The covariance among fixed mutations will also be negative:

$$\mathrm{cov}_{\mathrm{fix}}(u,s) = \frac{kb(p_{11}p_{kb} - p_{1b}p_{k1})(k-1)(b-1)}{(p_{11} + p_{k1}k + p_{1b}b + p_{kb}kb)^2} < 0$$

Comparison of correlations between nominal and fixed distributions of mutation and selection:

$$\frac{\rho_{\mathrm{fix}}(u,s)}{\rho_{\mathrm{poss}}(u,s)} = \frac{\left(\frac{kb}{E_{\mathrm{poss}}[ms]^2}\right)\left(\frac{p_{11}p_{kb} - p_{1b}p_{k1}}{\sqrt{p_b^{\mathrm{fix}}(1-p_b^{\mathrm{fix}})p_k^{\mathrm{fix}}(1-p_k^{\mathrm{fix}})}}\right)}{\left(\frac{p_{11}p_{kb} - p_{1b}p_{k1}}{\sqrt{p_k(1-p_k) * p_b(1-p_b)}}\right)}$$

$$= \left(\frac{kb}{E_{\mathrm{poss}}[ms]^2}\right)\left(\frac{\sqrt{p_k(1-p_k) * p_b(1-p_b)}}{\sqrt{p_b^{\mathrm{fix}}(1-p_b^{\mathrm{fix}})p_k^{\mathrm{fix}}(1-p_k^{\mathrm{fix}})}}\right)$$

$$= \frac{kb}{E_{\mathrm{poss}}[ms]^2}\sqrt{\frac{\mathrm{var}_{\mathrm{poss}}(u)\mathrm{var}_{\mathrm{poss}}(s)}{\mathrm{var}_{\mathrm{fix}}(u)\mathrm{var}_{\mathrm{fix}}(s)}}$$

$$= \frac{kb}{E_{\mathrm{poss}}[ms]^2} * \frac{\sigma_{\mathrm{poss}}(u)\sigma_{\mathrm{poss}}(s)}{\sigma_{\mathrm{fix}}(u)\sigma_{\mathrm{fix}}(s)}$$

A “strong form” of Berkson’s paradox, whereby mutation and selection are either positively correlated or uncorrelated among possible mutations (the nominal distribution) but negatively correlated among fixed mutations, does not occur in the simplest case, when mutation rate and

selection coefficient are each distributed across two possible values of "high" and "low" (see Arguments 1-4). In a "weak form" of Berkson's paradox, the correlation among mutation and selection is more negative in the fixed distribution than among the nominal distribution. For a negative correlation, the absolute value of the correlation coefficient of fixed mutations is therefore greater than that of the nominal distribution:

$$\frac{\rho_{fix}(u,s)}{\rho_{poss}(u,s)} = \frac{kb}{E_{poss}[ms]^2} * \frac{\sigma_{poss}(u)\sigma_{poss}(s)}{\sigma_{fix}(u)\sigma_{fix}(s)} > 1$$

$$kb * \sigma_{poss}(u)\sigma_{poss}(s) > E_{poss}[ms]^2 * \sigma_{fix}(u)\sigma_{fix}(s)$$

For a positive correlation, the correlation coefficient of fixed mutations is less than that of the nominal distribution:

$$\frac{\rho_{fix}(u,s)}{\rho_{poss}(u,s)} = \frac{kb}{E_{poss}[ms]^2} * \frac{\sigma_{poss}(u)\sigma_{poss}(s)}{\sigma_{fix}(u)\sigma_{fix}(s)} < 1$$

$$kb * \sigma_{poss}(u)\sigma_{poss}(s) < E_{poss}[ms]^2 * \sigma_{fix}(u)\sigma_{fix}(s)$$

5

CONCLUDING REMARKS

This thesis used empirical data, statistical modeling, theoretical analyses, and evolutionary simulations to uncover novel population genetic and evolutionary conditions that facilitate mutation-biased adaptation, and investigated the evolutionary consequences of different forms of mutation bias on adaptation in both molecules and entire organisms, in the lab and in nature.

More specifically, we showed that empirical genotype-phenotype landscapes can exhibit composition bias, namely, the enrichment of a particular type of mutation in adaptive trajectories, and how composition bias interacts with mutation bias to influence different aspects of adaptive evolution, such as its predictability, as well as the evolution of mutational robustness and genetic diversity. Moreover, we developed a statistical framework to quantify the influence of mutation bias on the spectrum of adaptive substitutions, a distribution for types of genetic changes fixed during adaptation. We applied this framework to three large datasets of adaptive mutations for different microbial species, and found a strong and statistically significant influence of mutation bias for all three species. We showed how the influence of mutation bias on adaptation can be modulated by population genetic conditions and the breadth and heterogeneity of the mutational target. Finally, we showed that the relationship between mutation rates and selection coefficients among the set of mutations that reach fixation can be distorted with respect to the relationship inferred from the nominal distribution of possible mutations, and that this distortion can be influenced both by the shape of the nominal distribution of mutations and population genetic conditions.

Additional studies of empirical genotype-phenotype landscapes could reveal composition bias in further protein or macromolecular contexts beyond transcription factor-DNA interactions [1], and such composition bias could interact with more complex forms of mutation bias than transition bias to influence adaptation [2]. Furthermore, a recent study has suggested that changes in mutation bias can facilitate access to beneficial mutations, thus influencing the outcomes of adaptation [3]. How composition bias could influence shifts in mutation bias is an open question for future research.

The discovery that mutation bias strongly shapes the spectrum of adaptive substitutions, along with the framework we have developed, could encourage further experimental characterisations of both mutation spectra and spectra of adaptive substitutions, and improve our understanding of mutation-biased adaptation in additional species and evolutionary conditions. Such a framework, however, could be improved in different ways. One is by integrating more accurate forms of mutation bias, such as context dependent mutation spectra [4]. Another way is by incorporating empirical measurements of fitness for each possible adaptive path to be able to provide a clearer picture of the influence of both selection and mutation bias on different adaptive processes.

To date, research on mutation-biased adaptation have ignored ecological interactions, only focusing on the study of long-term evolutionary dynamics [5]–[10]. While insightful, it remains unclear how different ecological interactions could influence the conditions for mutation-biased adaptation. This is because these previous studies consider mutations to have effects only on

the reproduction rates and not traits that affect direct interactions (e.g., competition, facilitation) between individuals. Such assumptions preclude the stable coexistence of several variants, where frequency-dependent effects, coexistence and nonlinear dynamics are usually observed [11]–[14]. For example, a population whose individuals engage in facilitation interactions exhibits higher clonal interference than a population whose individuals compete for one resource, under the same mutation supply conditions. Because clonal interference favors the fixation of selectively-favored variants over mutationally-favored variants, one would expect a reduced influence of mutation bias on adaptation. In general, whether ecological conditions can facilitate or hinder mutation-biased adaptation remains an open question for future research.

A further aspect of evolutionary processes that has been so far overlooked by mutation-biased adaptation research is spatial heterogeneity. Natural populations often evolve in complex and heterogeneous spatial structures, with homogeneous interactions holding only at a local scale. Such spatial structure can strongly impact evolutionary outcomes by affecting the fixation probability of mutants [15], [16]. Moreover, clonal interference could be drastically affected by spatial heterogeneities [17]–[19]. For instance, low density compartments (with low clonal interference) enhance the fixation of beneficial mutationally-favoured variants. This would imply that the spectrum of adaptive substitutions can be more predictable in such regions of the population, when prior knowledge of the biased mutation spectrum is available. To what extent the spatial distribution of the population in such compartments can influence mutation-biased adaptation is a further open question.

Such questions could be addressed by using a general eco-evolutionary framework following Lotka-Volterra type models, where interaction matrices can model the type of ecological interaction present in the population (e.g. facilitation and/or competition), as well as the spatial distribution of the population in time, affecting both the sign and the magnitude of selection and the clonal interference. As the population evolves towards quasi-stable states, one could quantify the enrichment for the different types of mutations to estimate their contributions to the trajectories of the eco-evolutionary process. If the molecular changes that are more likely to occur are also more likely to contribute to adaptation, one would expect a high similarity between the mutation spectrum and the spectrum of adaptive substitutions. Reshaping the interaction matrix would allow one to assess to what extent the type of ecological interaction and the initial spatial distribution of the population influence the types of mutations that contribute to adaptation.

In sum, the future research on mutation-biased adaptation has a variety of potential directions to follow, and could provide further insights about the generality of the effects of mutation bias on adaptive evolution, as well as increase the predictive power of evolutionary models.

REFERENCES

[1] D. Ray, H. Kazan, K. B. Cook, *et al.*, "A compendium of RNA-binding motifs for decoding gene regulation", *Nature*, vol. 499, no. 7457, pp. 172–177, 2013.

[2] L. Y. Yampolsky and A. Stoltzfus, "Mutational biases", *eLS*, 2008.

[3] M. Sane, G. D. Diwan, B. A. Bhat, L. M. Wahl, and D. Agashe, "Shifts in mutation spectra enhance access to beneficial mutations", *bioRxiv*, 2020.

[4] H. Lee, E. Popodi, H. Tang, and P. L. Foster, "Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing", *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[5] L. Y. Yampolsky and A. Stoltzfus, "Bias in the introduction of variation as an orienting factor in evolution", *Evolution & Development*, vol. 3, no. 2, pp. 73–83, 2001.

[6] K. Gomez, J. Bertram, and J. Masel, "Mutation bias can shape adaptation in large asexual populations experiencing clonal interference", *Proceedings of the Royal Society B: Biological Sciences*, vol. 287, no. 1937, p. 20 201 503, 2020.

[7] A. Stoltzfus, "Mutation-biased adaptation in a protein NK model", *Molecular Biology & Evolution*, vol. 23, pp. 1852–1862, 2006.

[8] A. d. A. Soares, L. Wardil, L. B. Klaczko, and R. Dickman, "Hidden role of mutations in the evolutionary process", *Physical Review E*, vol. 104, p. 044 413, 4 2021.

[9] A. V. Cano and J. L. Payne, "Mutation bias interacts with composition bias to influence adaptive evolution", *PLOS Computational Biology*, vol. 16, pp. 1–26, 2020.

[10] A. V. Cano, H. Rozhoňová, A. Stoltzfus, D. M. McCandlish, and J. L. Payne, "Mutation bias shapes the spectrum of adaptive substitutions", *Proceedings of the National Academy of Sciences*, vol. 119, no. 7, e2119720119, 2022.

[11] G. I. Lang, D. Botstein, and M. M. Desai, "Genetic variation and the fate of beneficial mutations in asexual populations", *Genetics*, vol. 188, no. 3, pp. 647–661, 2011.

[12] R. Maddamsetti, R. E. Lenski, and J. E. Barrick, "Adaptation, clonal interference, and frequency-dependent interactions in a Long-Term Evolution Experiment with *Escherichia coli*", *Genetics*, vol. 200, no. 2, pp. 619–631, 2015.

[13] B. Good, M. Mcdonald, J. Barrick, R. Lenski, and M. Desai, "The dynamics of molecular evolution over 60,000 generations", *Nature*, vol. 551, pp. 45–50, 2017.

[14] M. G. Behringer, B. I. Choi, S. F. Miller, *et al.*, "*Escherichia coli* cultures maintain stable subpopulation structure during long-term evolution", *Proceedings of the National Academy of Sciences*, vol. 115, no. 20, E4642–E4650, 2018.

[15] M Kimura, "On the probability of fixation of mutant genes in a population", *Genetics*, vol. 47, no. 6, pp. 713–719, 1962.

[16] M. Slatkin, "Fixation probabilities and fixation times in a subdivided population", *Evolution*, vol. 35, no. 3, pp. 477–488, 1981.

[17] E. A. Martens, R. Kostadinov, C. C. Maley, and O. Hallatschek, "Spatial structure increases the waiting time for cancer", *New Journal of Physics*, vol. 13, no. 11, p. 115 014, 2011.

[18] S. F. Bailey, A. Trudeau, K. Tulowiecki, *et al.*, "Spatial structure affects evolutionary dynamics and drives genomic diversity in experimental populations of *Pseudomonas fluorescens*", *bioRxiv*, 2021.

[19] D. Ally, V. R. Wiss, G. E. Deckert, *et al.*, "The impact of spatial structure on viral genomic diversity generated during adaptation to thermal stress", *PLoS ONE*, vol. 9, no. 2, e88702, 2014.

ACKNOWLEDGEMENTS

I would like to begin by thanking myself for engaging in this majestic journey loaded with heart-warming emotions and both individual and collective struggles. I would like to thank my supervisor not only for being a valuable source of scientific knowledge, but also for his friendly spirit. I will always be thankful for being chosen to be a part of this group, it was a real life-changing experience. I would also like to thank the co-examiners of this work for accepting to go through all this in such an altruistic manner. Thanks to all the collaborators for the synergistic interactions and fascinating scientific discussions. Thanks as well to all the people around in the lab, from all the different groups, both academic and administrative staff, you are an outstanding crew.

Finally, I would like to thank my families for all the support and the love they provide, making my life more beautiful. One of these families I did not choose, but from a distance their good wishes and warm blessings fill me with determination, gracias. The second one I did choose, and although members come and go, I will always be thankful for the amazing moments together. In Zurich this family expanded more than I would have never expected; I found a beautiful love, incredible friends and an incalculable amount of charming experiences, merci vielmal Zuri.

CURRICULUM VITAE

PERSONAL DATA

Name	Alejandro Viloria Cano
Date of Birth	December 20, 1991
Nationality	Venezuelan
Email address	alejvcano@gmail.com

EDUCATION

2017 – 2022	**PhD in computational biology** PhD thesis: "Mutation-biased adaptation" under the supervision of Prof. Dr. Joshua L. Payne. Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland.
2015 – 2017	**MSc in fundamental physics)** Master thesis: "Breaking of dynamical symmetry in globally coupled systems" under the supervision of Prof. Dr. Mario Cosenza. Universidad de Los Andes, Mérida, Venezuela.
2013 – 2015	**BSc in Physics** Master thesis: "Syncronization transitions in chaotic systems" under the supervision of Prof. Dr. Mario Cosenza. Universidad de Los Andes, Mérida, Venezuela.

TEACHING

2018 - 2020	**Supervisor Master students** Supervised one master thesis and two master projects of students from the ETH department of environmental sciences. ETH Zurich, Zurich, Switzerland.
2016 - 2017	**Professor (Instructor)** Teaching physics at the Engineering department. Universidad de Los Andes, Mérida, Venezuela.
2012 - 2015	**Teaching assistant** Teaching physics at bachelor level Universidad de Los Andes, Mérida, Venezuela.

MAIN CONTRIBUTIONS TO CONFERENCES

2020 **Oral presentation** SIB days 2020, online in Switzerland.

2019 **Poster presentation** SMBE 2019, in Manchester, United Kingdom.

2019 **Poster presentation** ESEB 2019, in Turku, Finland.

2019 **Oral presentation** Modelling Ecology & Evolution Zurich in Zurich, Switzerland.

2019 **Oral presentation** Workshop "From sequences to functions: challenges in the computation of realistic genotype-phenotype maps", in Zaragoza, Spain.

PUBLICATIONS

Articles in peer-reviewed journals:

[1] A. V. Cano, H. Rozhoňová, A. Stoltzfus, D. M. McCandlish, and J. L. Payne, "Mutation bias shapes the spectrum of adaptive substitutions", *Proceedings of the National Academy of Sciences*, vol. 119, no. 7, e2119720119, 2022. DOI: 10.1073/pnas.2119720119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2119720119. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2119720119.

[2] M. G. Cosenza, O Alvarez-Llamoza, and A. V. Cano, "Chimeras and Clusters Emerging from Robust-Chaos Dynamics", *Complexity*, vol. 2021, L. V. Gambuzza, Ed., p. 8 878 301, 2021, ISSN: 1076-2787. DOI: 10.1155/2021/8878301. [Online]. Available: https://doi.org/10.1155/2021/8878301.

[3] S. Manrubia, J. A. Cuesta, J. Aguirre, *et al.*, "From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics", *Physics of Life Reviews*, vol. 38, pp. 55–106, 2021, ISSN: 1571-0645. DOI: https://doi.org/10.1016/j.plrev.2021.03.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1571064521000300.

[4] A. V. Cano and J. L. Payne, "Mutation bias interacts with composition bias to influence adaptive evolution", *PLOS Computational Biology*, vol. 16, pp. 1–26, Sep. 2020. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1008296.

[5] A. V. Cano and M. G. Cosenza, "Asymmetric cluster and chimera dynamics in globally coupled systems", *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 11, p. 113 119, 2018. DOI: 10.1063/1.5043398. eprint: https://doi.org/10.1063/1.5043398. [Online]. Available: https://doi.org/10.1063/1.5043398.

[6] A. V. Cano and M. G. Cosenza, "Chimeras and clusters in networks of hyperbolic chaotic oscillators", *Phys. Rev. E*, vol. 95, p. 030 202, 3 2017. DOI: 10.1103/PhysRevE.95.030202. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.95.030202.