

Models and forward simulations of selection, human demography,
and complex traits

by

Lawrence Hart Uricchio

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

Chair

.....
John A. Wite
.....
Jeff D. Wall
.....
.....

Committee in Charge

UMI Number: 3681226

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3681226

Published by ProQuest LLC (2015). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Copyright 2014
by
Lawrence Hart Uricchio

to Zina: my love, my partner, my inspiration

Acknowledgments

I am indebted to the many people who supported me in the endeavor to complete this research.

Prior to arriving at UCSF, I was supervised by Dan Nicolae and Carole Ober at the University of Chicago. Carole and Dan immersed me in human genomics, and reinvigorated my enthusiasm for research.

I had the great fortune to work with Cyrus Maher, Raul Torres, Nicolas Strauli, Zachary Szpiech, and Kevin Hartman during my PhD. Although a statistical test has yet to be performed, there is compelling anecdotal evidence that the Hernandez lab attracts old PhD students. Though similar in age, we are diverse in scientific interests. I have learned so much from my fellow trainees.

I am grateful to the members of my committee – John Witte, Jeff Wall, and my research mentor Ryan Hernandez. Jeff has been a source of insight on all topics population genetic, and John a fantastic collaborator on my recent projects. I thank Ryan for challenging me, introducing me to productive projects, asking stimulating questions, and encouraging me to overcome research hurdles. It's been a pleasure to study with him, and an honor.

I thank Carl Palazzotto for instilling in me a Positive Mental Attitude.

Both my father-in-law Gene and mother-in-law Zoreh made positive contributions to my general well-being over the past several years, and their own scientific expertise is a source of stimulation.

My parents Bill and Katy raised me and my sister Phoebe in a house full of books in suburban Connecticut. I am so grateful to have parents who care so deeply about the life of the mind, and a sister who is both thoughtful and compassionate. I am also thankful to my grandparents Joe (who always made me laugh), Shirley (whose sweetness cheers all around her), Susan (whose intellect is on another plane), and Orson (with whom I feel a kinship that likely exceeds our genetic relatedness). I thank them for their love.

I thank my dog Milo for his silliness, which lightened my mood on foggy San Francisco days.

My wife Zina, my partner of 13 years at the time of writing, you have been my companion for my entire adult life. PhDs, if nothing else, tend to drag on. Without your love, your affection, your patience, I could not have endured all the starts and stops. Thank you for joining me on this journey, and brightening every moment. Your kindness warms me, your partnership steadies me, your adventurousness inspires me. I love you.

Previously published work

Chapters 2 and 3 of this dissertation were previously published.

The text in Chapter 2 is a reprint of the material as it appears in URICCHIO and HERNANDEZ (2014). Ryan Hernandez supervised the research that forms the basis of this chapter.

The text in Chapter 3 is a reprint of the material as it appears in URICCHIO *et al.* (2015). Ryan Hernandez supervised the research that forms the basis of this chapter. Raul Torres and John Witte also collaborated on the design and implementation of this research. Lawrence's contributions included designing the power study, writing the Python software, running the simulations, analyzing the simulated data, and writing the majority of the manuscript.

Chapter 4 of this dissertation will be submitted to a peer-reviewed journal. The work was conceived and executed by Lawrence under the supervision of Ryan Hernandez. John Witte also collaborated on the design of the research.

References

- URICCHIO, L. H., and R. D. HERNANDEZ, 2014 Robust forward simulations of recurrent hitchhiking. *Genetics* **197**: 221–236.
- URICCHIO, L. H., R. TORRES, J. S. WITTE, and R. D. HERNANDEZ, 2015 Population genetic simulations of complex phenotypes with implications for rare variant association tests. *Genetic epidemiology* **39**: 35–44.

Abstract

Evolutionary forces such as recombination, demography, and selection can shape patterns of genetic diversity within populations and contribute to phenotypic variation. While theoretical models exist for each of these forces independently, mathematically modeling their joint impact on patterns of genetic diversity remains very challenging. Fortunately, it is possible to perform forward-in-time computer simulations of DNA sequences that incorporate all of these forces simultaneously. Here, I show that there are trade-offs between computational efficiency and accuracy for simulations of a widely investigated model of recurrent positive selection. I develop a theoretical model to explain this trade-off, and a simple algorithm that obtains the best possible computational performance for a given error tolerance. I then pivot to develop a framework for simulations of human DNA sequences and genetically complex phenotypes, incorporating recently inferred demographic models of human continental groups and selection on genes and non-coding elements. I use these simulations to investigate the power of rare variant association tests in the context of rampant selection and non-equilibrium demography. I show that the power of rare variant association tests is in some cases quite sensitive to underlying assumptions about the relationship between selection and effect sizes. This work highlights both the challenge and the promise of applying forward simulations in genetic studies that seek to infer the parameters of evolutionary models and detect statistical associations.

Contents

1	Introduction	1
1.1	Evolutionary models and simulations of DNA sequences	2
1.2	Methods of simulation for DNA sequences under evolutionary models	3
1.2.1	Coalescent simulators	3
1.2.2	Forward simulators	5
2	Robust forward simulations of recurrent hitchhiking	9
2.1	Introduction	10
2.2	Model	11
2.2.1	The expectation of π in recurrent hitchhiking	15
2.3	Materials & Methods	17
2.3.1	Simulating RHH Models	17
2.3.2	Fixing the probability of fixation	17
2.4	Results	18
2.4.1	A “naive” approach to parameter rescaling	18
2.4.2	RHH with large values of s	20
2.4.3	Robust parameter rescaling for RHH simulations	21
2.4.4	The notion of “sufficiently distant” flanking sites	23
2.4.5	The role of interference	24
2.4.6	An application to <i>Drosophila</i> parameters	25
2.5	Discussion	26
2.6	Appendix	29
2.6.1	Derivation of $p_{\tau_f}^*$ for large s	29
2.6.2	RHH simulations in SFS_CODE	31
2.6.3	Efficiency of rescaled simulations	32
3	Simulating complex phenotypes under evolutionary models	47
3.1	Introduction	48
3.2	Materials and Methods	50

3.2.1	<code>sfs_coder</code> : A Python-based interface to SFS_CODE	50
3.2.2	Simulations of human demography and selection	51
3.2.3	Simulations of genomic elements	52
3.2.4	Haplotype resampling with Hapgen2	52
3.2.5	Choosing a region under strong linked selection	53
3.2.6	Simulations of phenotypes and the power of SKAT-O	53
3.3	Results	54
3.3.1	Demography, sample size, and selection impact rare variants	54
3.3.2	Haplotype resampling under-estimates the number of rare variants in large samples	55
3.3.3	Power estimates may be impacted by local genomic context and demography	56
3.3.4	The impact of linked selection	57
3.4	Discussion	58
3.5	Acknowledgments	61
3.6	Appendix	61
4	Selection and explosive growth may hamper the power of rare variant association tests	72
4.1	Introduction	73
4.2	Materials & Methods	75
4.2.1	An evolutionary model of complex phenotypes	75
4.2.2	Calculating the impact of demographic events on genetic architecture	78
4.2.3	Three-population forward simulations of human selection and demography	79
4.2.4	Calculating the genetic variance	80
4.2.5	Power of SKAT-O	81
4.3	Results	81
4.3.1	Selection and demography impact the genetic architecture of complex traits	81
4.3.2	Architecture of complex traits in multiple human populations under a selection-based phenotype model	82

4.3.3	The power of SKAT-O is inversely proportional to variance explained by rare variants	84
4.4	The impact of increased sample size: a mock sequencing study with multiple genes .	85
4.4.1	A simple strategy to increase power may also increase the false-positive rate .	85
4.5	Discussion	86

List of Tables

1	Parameter definitions	12
---	---------------------------------	----

List of Figures

2.1	A pictorial representation of the recurrent hitchhiking model	33
2.2	Probability of fixation	34
2.3	Algorithms 1 & 2 performance	35
2.4	Adjusting the flanking sequence length	36
2.5	Intuitive explanation of Algorithm 2	37
2.6	Unlinked, strong selection	38
2.7	RHH simulations with interference	39
2.8	A practical example: Drosophila parameters	40
2.9	Comparison of theoretical forms	41
2.10	Computational burden of RHH simulations	42
2.11	Duration of rescaled RHH simulations	43
3.1	A pictorial representation of a model of human genomic sequences	63
3.2	Cumulative site frequency spectra	64
3.3	Hapgen performance when the population has recently expanded	65
3.4	Power of SKAT-O	66
3.5	Nucleotide diversity in simulations with linked selected sequences	67
4.1	Variance due to singletons	89
4.2	Genetic variance of complex traits $\tau = 1$	90
4.3	Genetic variance of complex traits $\tau = 0.75$	91
4.4	Genetic variance of complex traits $\tau = 0.5$	92
4.5	Proportion of variance due to rare variants that is due to singletons	93
4.6	Power of SKAT-O under various models of effect sizes	94
4.7	Power of SKAT-O for large samples, $\tau = 1$	95
4.8	Power of SKAT-O for large samples, $\tau = 0.75$	96
4.9	Power of SKAT-O for large samples, $\tau = 0.5$	97

4.10 Power and false positive rate of SKAT-O with adjusted weights 98

1 Introduction

1.1 Evolutionary models and simulations of DNA sequences

There is tremendous variation in human phenotypes such as height and eye color. Ancient humans observed that, to some extent, these attributes are passed on from parent to child. Genetics is the study of the physical basis of this inheritance, and population genetics concerns itself with the dynamics of genetic inheritance over long time scales.

Evolutionary forces such as mutation, recombination, genetic drift, demography, and natural selection all influence genetic variation in populations. Theoretical interest in these forces stretches back over a century, predating even the knowledge of DNA as the molecule of inheritance. Many mathematical models have been developed to explore their impact on genes, amongst which the most influential are the coalescent and diffusion theory. Thanks to these models, we have a theoretical framework for investigating patterns of genetic diversity in populations as a function of fundamental model parameters, such as mutation rate and selection strength.

In recent years, the price of both genotyping and DNA sequencing have dropped dramatically while their throughput has increased. Many large scale efforts are underway to catalog genetic variation in humans or other organisms, or have already been published (most notably INTERNATIONAL HAPMAP CONSORTIUM *et al.* 2007; 1000 GENOMES PROJECT CONSORTIUM *et al.* 2012). While these studies have largely substantiated the predictions of population genetic theory, they have also demonstrated the need for more sophisticated models to fully understand the impact of evolutionary forces on genetic diversity. For example, it is clear that both positive selection (the influence of alleles that improve an organism's fitness) and negative selection (alleles that are deleterious for the reproductive success of the organism) impact patterns of genetic diversity, but these forces are seldom modeled simultaneously. Demographic processes such as population growth, contraction, and migration also impact patterns of diversity, but these processes are challenging to model simultaneous to selection. Recombination further complicates theoretical models. General predictions about patterns of diversity that simultaneously account for all these factors are lacking in the literature.

Though it is very challenging to incorporate all these processes simultaneously in theoretical models, much recent progress has been made in simulating genetic data under complex models. Simulations can provide insight into the impact of evolutionary forces on patterns of diversity,

allow us to quantify the accuracy of theoretical predictions, assess the performance of statistical inference techniques, and generate phenotypes under quantitative models of complex genetic traits. Here, I will introduce the two most popular classes of simulation tools (namely coalescent and forward simulation), and discuss current limitations of each method and how this work contributes to the genetics simulation literature.

1.2 Methods of simulation for DNA sequences under evolutionary models

Simulations are widely used in population genetics, and many computer methods for simulating DNA sequences under population genetic models have been developed and published. The goal of these simulation programs is to generate DNA sequence data stochastically under a model of replication in a population. The two most popular categories of methods are coalescent simulators (HUDSON 2002; SPENCER and COOP 2004; HELLENTHAL and STEPHENS 2007; LIANG *et al.* 2007; EWING and HERMISSON 2010; EXCOFFIER and FOLL 2011; ZENG 2013) and forward simulators (HERNANDEZ 2008; ZANINI and NEHER 2012; MESSER 2013; THORNTON 2014). Coalescent simulators follow the genealogy of a sample of chromosomes backwards in time, while forward simulators model every chromosome in the population and follow these chromosomes forward in time.

1.2.1 Coalescent simulators

Coalescent simulators generate a random genealogical history for a sample of chromosomes by modeling the shared ancestry process backwards in time. Suppose that there are N total chromosomes in the population at all times in the past, and that the population reproduces in non-overlapping generations. Then the chance that any two chromosomes share a parent in the previous generation is $1/N$. If the chromosomes in our sample were chosen randomly and uniformly from a population of size N , then all pairs of chromosomes are equally likely to share common ancestry at any point in time. If $n \ll N$, then the total rate R_n at which common ancestry events occur is approximately proportional to the number of pairs of lineages, $R_n = \frac{1}{N} \binom{n}{2}$, and the waiting time to the next coalescent event is geometrically distributed with this rate parameter. Assuming that N is a large number, then we can approximate this geometric waiting time with an exponential process.

This process of ancestry sharing immediately suggests an algorithm for generating genealogies under the coalescent. We draw an exponential random number with rate R_n ; this is the time of

the first coalescent event. Since all pairs of lineages are equally likely to share common ancestry, we choose two lineages uniformly at random and join them together at this time. The number of lineages has now decreased to $n - 1$, so we now draw a random exponential number with rate R_{n-1} , and repeat the above lineage joining procedure. We repeat these steps until a single lineage remains.

Mutation events occur at a constant rate through time, so the the number of mutations that occur on any given branch of the tree is proportional to the branch length. Given that we have a random tree under the model, we can then generate random sequences under the model by allowing mutations to occur on the simulated tree. Note that adding mutations only after simulating the tree assumes that the mutations do not affect the genealogical process.

The coalescent has many advantages as a model for simulations. First, it is very fast to generate samples under the coalescent when there is no recombination because we only need to follow the n sampled lineages back in time. Second, the coalescent can easily accommodate many kinds of demographic events, such as changes in population size and migration events (HUDSON 2002; EXCOFFIER and FOLL 2011). These events change the overall rate at which lineages coalesce, but they do not break the fundamental assumption that the probability of common ancestry does not depend on the mutation events that are observed under demographic models. Recombination can also be accommodated, but comes at a large computational cost for chromosome scale sequences or high recombination rates. For this reason some simplifications to the coalescent have been proposed, where very unlikely coalescent events are prohibited in order to maintain computational feasibility (MCVEAN and CARDIN 2005; MARJORAM and WALL 2006).

In contrast, natural selection is very challenging to simulate under the coalescent (but see SPENCER and COOP 2004; EWING and HERMISSON 2010; ZENG 2013, for examples of selection-based coalescent simulators with some simplifying assumptions and HUDSON and KAPLAN 1988; NEUHAUSER and KRONE 1997 for early theoretical insights into the coalescent with selection). Selected mutations, by definition, alter the probability of common ancestry. Hence, it is inappropriate to simulate the genealogy backwards in time completely independent of the mutation events. To get around this problem, it is necessary to have some foreknowledge of the impact of selection on the genealogy, either by simulating some portion of the data forward in time and conditioning on these forward simulations, or modeling the impact of selected sites on the genealogy. For this

reason, it is often advantageous to adopt a different simulation methodology when the model of interest involves complex selection (e.g., simultaneous positive and negative selection, or selection at many partially linked sites).

1.2.2 Forward simulators

As an alternative to coalescent simulations, it is possible to simulate DNA sequences forward-in-time under arbitrarily complex models that include selection, recombination, and demography. Forward simulators follow the population genealogical history from some point in the distant past, until the time of sampling. Because we simulate the genealogy forward in time, mutations can be generated simultaneous to the genealogy, and any impact these mutations have on the genealogy in future generations is accommodated as the simulation progresses.

The simulation procedure proceeds directly from the model of replication introduced in the previous section, by simply reversing the direction of time. In each new generation $t + 1$, we pick parents from generation t . If there is no selection, then every individual in generation t is equally likely to be a parent of an individual in generation $t + 1$. If selection is present, then parents are picked with probability proportional to their fitness divided by the population average fitness. We repeat this process $N(t + 1)$ times, where $N(t + 1)$ is the population size in generation $t + 1$. New mutations are introduced in the offspring at a pre-specified rate. If we wish to simulate multiple populations, we allow the populations to split at pre-specified times, or merge, or accommodate migration at a pre-specified rate. At the end of the simulation, we sample n individuals from the population and report the mutations present on each of their chromosomes.

In a forward simulation, we are sampling *directly* from the model, whereas to build the coalescent simulator, we made several approximations and assumptions about the underlying genetic process. As a result, coalescent simulators only work well when the underlying assumptions are met (e.g., $n \ll N$). Forward simulators always produce patterns of diversity that are sampled exactly from the underlying stochastic model. However, this flexibility comes at a large cost. In the forward simulator, we must store every (N) chromosome in the population in RAM, whereas we only followed the sampled (n) chromosomes in the coalescent simulation. Moreover, in order to reach “steady state” it is necessary to run the simulation for a duration (in generations) proportional to N . Thus, the performance cost of the simulation scales roughly with N^2 . For many natural

populations, N may be in the millions (or larger), which is too costly to simulate for many interesting models.

Throughout this dissertation, I examine the theory and practice of using forward simulations in population genetic studies of selection and complex demography. In the second chapter, I examine approaches for reducing the computational burden of forward simulation, and show that these approaches sometimes come at the expense of accuracy of the simulated patterns of diversity. I develop a method for constraining the bias in the simulations for a simple model of recurrent selection, and discuss extensions to more sophisticated models. In the third and fourth chapters, I pivot to discuss joint forward simulation of complex traits and genotypes in models that include selection and demography. I show that the power of a widely used rare variant association test may be strongly impacted by selection, and argue that selection must be included in simulations when assessing statistical power in order to obtain sensible and interpretable power estimates. These studies demonstrate that forward simulations can provide insights into models that are difficult to simulate under the coalescent, but also highlight the potential for forward simulations to provide misleading results when applied inappropriately.

References

- 1000 GENOMES PROJECT CONSORTIUM, G. R. ABECASIS, A. AUTON, L. D. BROOKS, M. A. DEPRISTO, *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.
- EWING, G., and J. HERMISSON, 2010 Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics (Oxford, England)* **26**: 2064–2065.
- EXCOFFIER, L., and M. FOLL, 2011 fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics (Oxford, England)* **27**: 1332–1334.
- HELLENTHAL, G., and M. STEPHENS, 2007 mshot: modifying hudson’s ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics (Oxford, England)* **23**: 520–521.
- HERNANDEZ, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. *Bioinformatics (Oxford, England)* **24**: 2786–2787.
- HUDSON, R. R., 2002 Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)* **18**: 337–338.
- HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. *Genetics* **120**: 831–840.
- INTERNATIONAL HAPMAP CONSORTIUM, K. A. FRAZER, D. G. BALLINGER, D. R. COX, D. A. HINDS, *et al.*, 2007 A second generation human haplotype map of over 3.1 million snps. *Nature* **449**: 851–861.
- LIANG, L., S. ZÖLLNER, and G. R. ABECASIS, 2007 Genome: a rapid coalescent-based whole genome simulator. *Bioinformatics (Oxford, England)* **23**: 1565–1567.
- MARJORAM, P., and J. D. WALL, 2006 Fast “coalescent” simulation. *BMC genetics* **7**: 16.
- MCVEAN, G. A. T., and N. J. CARDIN, 2005 Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **360**: 1387–1393.
- MESSER, P. W., 2013 Slim: simulating evolution with selection and linkage. *Genetics* **194**: 1037–1039.
- NEUHAUSER, C., and S. M. KRONE, 1997 The genealogy of samples in models with selection. *Genetics* **145**: 519–534.
- SPENCER, C. C. A., and G. COOP, 2004 Selsim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics (Oxford, England)* **20**: 3673–3675.
- THORNTON, K. R., 2014 A c++ template library for efficient forward-time population genetic simulation of large populations. *Genetics* **198**: 157–166.
- ZANINI, F., and R. A. NEHER, 2012 Ffpopsim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics (Oxford, England)* **28**: 3332–3333.

ZENG, K., 2013 A coalescent model of background selection with recombination, demography and variation in selection coefficients. *Heredity* **110**: 363–371.

2 Robust forward simulations of recurrent hitchhiking

URICCHIO, L. H., and R. D. HERNANDEZ, 2014 Robust forward simulations of recurrent hitchhiking. *Genetics* **197**: 221–236.

2.1 Introduction

A central goal of population genetics is to determine the strength and rate of natural selection in populations. Natural selection impacts patterns of genetic diversity within populations, and is likely to influence phenotypes of biological and medical interest (BUSTAMANTE *et al.* 2005; TORGERSON *et al.* 2009; MAHER *et al.* 2012; ARBIZA *et al.* 2013). There exists a large body of literature focused on mathematical models of selection in populations and inferring the action of selection on DNA sequences under these models (recent reviews include POOL *et al.* 2010; CRISCI *et al.* 2012; CUTTER and PAYSEUR 2013). One such model is known as recurrent hitchhiking, in which patterns of diversity at a selectively neutral locus are altered due to repeated positive selection at linked loci.

Recurrent hitchhiking has been theoretically explored (SMITH and HAIGH 1974; OTA and KIMURA 1975; KAPLAN *et al.* 1989; STEPHAN *et al.* 2006; COOP and RALPH 2012) and applied to DNA sequences of various organisms (BACHTROG 2008; JENSEN *et al.* 2008; INGVARSSON 2010; SINGH *et al.* 2013). The classic work of STEPHAN *et al.* 1992 modeled the dynamics of the neutral locus in a single sweep with diffusion-based differential equations, which they solved approximately. WIEHE and STEPHAN 1993 later showed that their solution for single sweeps could be applied to a recurrent sweep model, where the expected reduction in neutral diversity is well approximated by $\frac{r}{r+\alpha\lambda}$; $\alpha = 2Ns$ where N is the population size and s is the selection coefficient, r is the recombination rate, λ is the rate of positively selected substitutions, and I is a constant that approximates the value of an integral. However, little work has been done to explore recurrent sweeps with forward simulations (but see KIM and STEPHAN 2003, and CHEVIN *et al.* 2008, where interfering substitutions were studied with forward simulations, and the discussion herein).

It is crucial to understand the dynamics of recurrent sweeps (and other population genetic models) when realistic perturbations to the model are introduced, which is often difficult in a coalescent framework. In contrast, with forward simulations it is straightforward to introduce arbitrarily complex models, including demographic processes, interference between selected sites, simultaneous negative and positive selection, and variable strength of selection or recombination rate across a chromosome. Furthermore, forward simulations can be performed exactly under a given model, and hence they can be used as a direct test of theoretical predictions. Simulations

can be used in conjunction with inference methods such as Approximate Bayesian Computation to estimate parameters when the likelihood function of the data under the model is unknown (BEAUMONT *et al.* 2002).

In population genetics, forward methods have often been overlooked in favor of reverse time coalescent simulators due to computational efficiency (HERNANDEZ 2008; for an overview of coalescent and forward simulation techniques, see KIM and WIEHE 2009). Although coalescent simulations are generally more computationally efficient, in most applications they require some *a priori* knowledge of allele trajectories. Recent improvements in computer memory and processor speeds have made forward simulations more tractable. However, simulations of recurrent hitchhiking in some parameter regimes of interest (e.g., $N > 10^5$) are still computationally prohibitive, so it is frequently necessary to rescale model parameters (e.g., N and chromosome length, L) (KIM and WIEHE 2009). Currently, the literature provides some guidelines for performing parameter scaling in forward simulations (HOGGART *et al.* 2007), but it is not clear that these methods will be generally applicable to all models or hold in all parameter regimes.

In this investigation, we examine recurrent sweeps through forward simulation and theory. We provide a detailed, practical discussion of simulations of recurrent sweeps in a forward context, focusing on scaling laws of relevant parameters such as N , λ , r , α , and L . We evaluate a “naive” parameter rescaling algorithm, and show that this technique can bias patterns of variation in the simulations because it is not conservative with respect to the underlying genealogical process, particularly in the large α , small N regime. We quantify the effect of large values of the selection coefficient s on recurrent hitchhiking through theory. Finally, we leverage these principles to make gains in computational efficiency with a simple algorithm that provides the best possible performance for a prespecified error threshold, and apply the method to simulations of parameters previously inferred in *Drosophila*.

2.2 Model

Here, we describe the recurrent hitchhiking model (shown schematically in Figure 2.1), upon which we build the results and simulations in this article. Key parameters of the model are discussed below, and summarized in Table 1.

A neutral locus is flanked on both sides by sequences experiencing repeated positively selected

substitutions at rate λ per generation per site. λ is assumed to be small enough that multiple positively selected mutations do not simultaneously sweep in the population and hence there is no interference (though interference between selected sites is not prohibited in the simulations performed herein). Population size is fixed at N . In forward simulations, there is no distinction between effective and census population size, so $N = N_e$. Recombination occurs at rate r_f (recombination fraction) per generation per chromosome. Note that the recombination fraction is the probability that the number of recombination events between two loci is an odd number, and cannot exceed 0.5. Each positively selected site has a selection coefficient $s = \alpha/2N$. Heterozygous individuals have fitness $1+s$, while individuals homozygous for the selected allele have fitness $1+2s$. The neutral locus itself is assumed to be non-recombining. Any of the above constraints and model assumptions can be relaxed in forward simulations.

Table 1: **Parameter definitions**

N	Population size
τ_f	Time of fixation of selected allele
t	Time in generations
$x(t)$	Frequency of the selected allele
$h(t)$	Relative heterozygosity at the neutral locus among selected chromosomes
q_l	Minor allele frequency at site l
l_0	Length of neutral region
L	Flanking sequence length
π	Nucleotide diversity, $\pi = \sum_{l=1}^{l_0} \frac{2q_l(1-q_l)(n-1)}{n}$
π_0	Nucleotide diversity under neutrality
π_N	Nucleotide diversity in a population of size N
$p(t)$	Probability of common ancestry at the neutral locus
n	Number of sampled sequences
μ	Mutation rate/generation/chromosome/base pair
$\theta = 4N\mu$	Population scaled mutation rate
s	Selection coefficient
$\alpha = 2Ns$	Population scaled selection strength
r	Recombination rate/generation/chromosome/base pair
$\rho = 4Nr$	Population scaled recombination rate
r_f	Recombination fraction (probability of productive recombination, per generation per chromosome)
λ	Rate of positively selected substitutions per site per generation
k_h	Rate of common ancestry induced by sweep events (see (2.4))
R_c	Total rate of common ancestry induced by sweep and coalescent events
$I_{\alpha,s}$	The integral $\int_0^{u^*} p_{\tau_f}(u)du$ (see (2.4))
$I_{\alpha,s}^*$	The integral $\int_0^{u^*} p_{\tau_f}^*(u)du$ (see (2.14))
$I = 0.075$	A constant approximating $I_{\alpha,s}$

Consider the coalescent history at the neutral locus of two sequences sampled immediately after a selective sweep. If there is no recombination between the neutral and selected loci during the sweep, then the two sequences must share a common ancestor at some point during the sweep. If selection is sufficiently strong, then the time to fixation of selected alleles is effectively instantaneous relative to the neutral fixation process (KAPLAN *et al.* 1989), and thus the expected heterozygosity at the neutral site at the completion of the sweep is nearly 0 because very few mutations are introduced during the sweep.

Recombination significantly complicates this model. Immediately after a sweep, the reduction in heterozygosity at the neutral locus is a function of the recombination distance between the selected substitution and the neutral locus, and the strength of selection. Stephan, Wiehe, and Lenz (SWL) calculated the reduction in heterozygosity at the neutral locus with a diffusion based, differential equation framework (STEPHAN *et al.* 1992). They showed that the expected reduction in heterozygosity at the neutral locus among chromosomes carrying the selected allele, relative to the baseline heterozygosity, $h(t)$, can be modeled with a simple differential equation, which they solved approximately.

KAPLAN *et al.* 1989 showed that $h(t)$ is closely related to the probability that two sequences sampled at the end of the sweep share a common ancestor at the neutral site during the sweep, $p(t)$.

$$p(t) = 1 - h(t) \tag{2.1}$$

This allows the results of SWL to be interpreted in terms of the coalescent process at the neutral locus. Note that when $t = \tau_f$ (the end of the sweep), $p(\tau_f)$ represents the probability of common ancestry at the neutral locus for a pair of sequences at some point during the sweep because all chromosomes carry the selected allele at the end of a sweep. Throughout the article, we subscript variables of interest with τ_f to denote their values at the time of fixation and emphasize their dependence on the recombination fraction r_f (e.g., $p_{\tau_f}(r_f)$). Rewriting SWL results with (2.1), we obtain

$$\frac{d}{dt}p(t) = \frac{1 - p(t)}{2Nx(t)} - 2r_f p(t) (1 - x(t)) \tag{2.2}$$

where $x(t)$ is the frequency of the selected allele at time t during the sweep. Equation (2.2) is equivalent to equation 5 of BARTON 1998 when the selected allele is at low frequency.

Equation (2.2) can be interpreted in terms of the recombination process between the neutral and selected loci. In particular, there are two mechanisms that can change the proportion of selected sequences that share common ancestry at the neutral locus. The first term on the RHS of (2.2) represents that chance of common ancestry in the previous generation among selected sequences that have already recombined off of the original background. The chance that any two such sequences share a common ancestor in the previous generation is $\frac{1}{2Nx(t)}$. The second term represents the chance that a recombination event occurs between a selected chromosome and some non-selected chromosome, thereby reducing $p(t)$. The first term is only important when the frequency of the selected site is low, because it is inversely proportional to the number of selected chromosomes, whereas the second term contributes non-negligibly to the dynamics at all allele frequencies of the selected locus.

Consider the coalescent history at the neutral locus of two lineages sampled at the current time (not necessarily immediately after a sweep event). In each preceding generation, there is some chance that they share a common ancestor at the neutral locus due to normal coalescent events, and some chance that they share common ancestry because of a sweep event. Since sweeps occur nearly instantaneously relative to the timescale of coalescence under neutrality, we can approximate the chance of common ancestry as two competing processes. Neutral events occur at rate $1/2N$ and compete with sweep events, which happen at rate $\frac{2\lambda}{r}p_{\tau_f}(r_f)dr_f$ in a window of size dr_f , assuming that sweeps occur homogeneously across the chromosome and $r_f \approx rL$. Note that r and λ appear in a quotient in this rate, which implies that multiplying both the substitution and recombination rates by a common factor has no impact on the model. The factor of 2 represents the flanking sequence on either side of the neutral locus.

Following the results of SWL, an approximate solution to (2.2) is:

$$p_{\tau_f}(r_f) = 1 - \frac{2r_f}{s} \alpha^{-\frac{2r_f}{s}} \Gamma \left[\frac{-2r_f}{s}, \frac{1}{\alpha} \right] \quad (2.3)$$

where Γ is the incomplete gamma function. Note that (2.3) is a function of r_f, α , and s , but we only denote the dependence on r_f since α and s are assumed to be fixed for the analysis herein.

Following SWL, we denote the rate at which lineages merge due to sweep events as k_h ,

$$k_h = 2N \left(\frac{2\lambda}{r} \int_0^{r_f^*} p_{\tau_f}(r_f) dr_f \right) \quad (2.4)$$

where r_f^* is taken as the value of r_f that corresponds to the end of the flanking sequence. If the flanking chromosome being modeled exceeds $\frac{s}{r}$ base pairs, previous work suggests that r_f^* can be taken to be any value sufficiently far away from the neutral locus such that the value of k_h is as close as desired to its asymptotic limit JENSEN *et al.* 2008. The factor of $2N$ is introduced to rescale in coalescent units, such that neutral coalescent events happen at rate 1 relative to sweep merger events.

2.2.1 The expectation of π in recurrent hitchhiking

In the recurrent hitchhiking (RHH) model, it is of great interest to describe the reduction in diversity as a function of the basic parameters of the model (α, r, λ , etc.). To make this dependence clearer, we perform two changes of variables in (2.4). First, we note that (2.4) was derived by SWL under the assumption that r_f^* is small, such that the recombination fraction is given by $r_f \approx rL$. Here we will frequently be concerned with values of r_f that approach its maximum value of 0.5, which invalidates this approximation. We therefore rewrite (2.4) as a function of L , substituting $r_f = \frac{1-e^{-2rL}}{2}$ for the quantity r_f (HALDANE 1919). We then substitute the quantity $u = \frac{2r}{s}L$ for L . Rewriting p_{τ_f} and k_h as functions of u , we have

$$p_{\tau_f}(u) = 1 - \frac{(1 - e^{-su})}{s} \alpha^{-\frac{(1-e^{-su})}{s}} \Gamma \left[\frac{-(1 - e^{-su})}{s}, \frac{1}{\alpha} \right] \quad (2.5)$$

and

$$k_h = \frac{2Ns\lambda}{r} \int_0^{u^*} p_{\tau_f}(u) du \quad (2.6)$$

It is useful to examine the properties of (2.5) and (2.6) as a function of s . When s is small, (2.5) can be rewritten as

$$p_{\tau_f}(u) \approx 1 - u\alpha^{-u} \Gamma\left[-u, \frac{1}{\alpha}\right] \quad (2.7)$$

which removes the dependence on s and is identical to the quantity inside the integral on the RHS of equation 4 of WIEHE and STEPHAN 1993. Thus, the integral on the RHS of (2.6) is a function only of the parameter α when s is small. This is not necessarily the case as s becomes large, but we also note that (2.3) was originally derived under the assumption that s is small, so it is possible that the large s behavior is not accurately captured by (2.5) and (2.6).

Similar to WIEHE and STEPHAN 1993, we define the integral in (2.6) as $I_{\alpha,s}$, but we include the subscript α, s to emphasize that, under some circumstances, $I_{\alpha,s}$ may be a function of both α and s and cannot be written as a function of only the population scaled strength of selection. The total rate of coalescence R_c (in coalescent units) due to both sweep and neutral coalescent events is then

$$R_c = 1 + k_h = 1 + \frac{\alpha\lambda}{r} I_{\alpha,s} \quad (2.8)$$

The expected height of the coalescent tree for two sequences is the inverse of this rate. The expected reduction in diversity at the neutral locus is proportional to the decrease in the height of the coalescent tree, relative to neutrality.

$$E_{\alpha,s}[\pi/\pi_0] = \frac{1}{R_c} = \frac{r}{r + \alpha\lambda I_{\alpha,s}} \quad (2.9)$$

WIEHE and STEPHAN 1993 found that $I_{\alpha,s}$ is approximately constant ($I=0.075$) over a range of large values of α .

$$E_{WS}[\pi/\pi_0] = \frac{r}{r + \alpha\lambda I} \quad (2.10)$$

Note that this removes the dependence on s , which is asserted by (2.9). In the following subsections we show that both (2.9) and (2.10) may not hold when s is large.

2.3 Materials & Methods

2.3.1 Simulating RHH Models

We performed forward simulations of RHH with SFS_CODE (HERNANDEZ 2008; see Appendix for details). A pictorial representation of the model is shown in Figure 2.1.

All simulations in this article were performed with $\theta = 0.002$ at the neutral locus, and reductions in diversity were calculated as the ratio of the observed diversity to 0.002, unless otherwise noted. Nucleotide diversity π and Tajima’s D were calculated with a custom script. We often report the difference proportion in π_{N_1} (diversity in a rescaled population of size N_1) as compared to π_{N_0} (diversity in the population of size N_0), which we define as $\frac{\pi_{N_1} - \pi_{N_0}}{\pi_{N_0}}$. For each simulation we sampled 10 individuals (20 chromosomes) from the population. The neutral loci in all simulations are 1 Kb in length.

2.3.2 Fixing the probability of fixation

Throughout the article, we discuss appropriate choices of r , λ , s , L , and N for simulations. However, in forward simulations, the rate of substitution is not explicitly provided to the software, but rather a rate of mutation. In order to calculate the appropriate mutation rate for a simulation, one must incorporate the probability of fixation for a positively selected site. For $s < 0.1$, the fixation probability of KIMURA 1962 is sufficient:

$$P_{Kimura}(s, \alpha) = \frac{1 - e^{-2s}}{1 - e^{-2\alpha}} \quad (2.11)$$

However, when $s > 0.1$, this approximation overestimates the probability of fixation. For $s > 0.1$, we treat the initial trajectory of the selected site as a Galton-Watson process and calculate the probability of extinction by generation i , $P_e(i)$, with procedure $P_{GW}(s)$ FISHER 1999.

In practice, this algorithm takes fewer than 200 iterations to converge for $s > 0.1$, and provides accurate results (Figure 2.2). Simulations for this figure were performed with a simple Wright-Fisher simulator that only sampled the trajectory of the selected site and followed it until either 1) loss or 2) the frequency of the selected site exceeded $100/\alpha$, which very nearly guarantees eventual fixation.

```

procedure  $P_{GW}(s)$ 
   $P_e(0) = e^{-(1+s)}$ 
   $i = 1$ 
  while  $P_e(i) - P_e(i - 1) > \delta$  do
     $P_e(i) = e^{-(1+s)(1-P_e(i-1))}$ 
     $i \leftarrow i + 1$ 
  end while

  return  $1 - P_e(i)$ 
end procedure

```

All analysis and simulation scripts used in this article are available upon request from the authors.

2.4 Results

2.4.1 A “naive” approach to parameter rescaling

In forward simulations, the most computationally costly parameters are N and L , so we seek to reduce these parameters as much as possible. A simple and widely used rescaling assumption is that patterns of diversity are conserved when population scaled parameters $\rho = 4Nr$, $\alpha = 2Ns$, and $\theta = 4N\mu$ are held fixed and N is varied (KIM and WIEHE 2009; for more discussion, see subsection 5 of the SFS_CODE manual). This is equivalent to the statement that the effective population size is not a fundamental parameter of the dynamics, and is similar to the rescaling strategy described in HOGGART *et al.* 2007, which was not designed specifically for RHH simulations.

Equation (2.9) provides an informed view of rescaling that incorporates RHH theory. Equation (2.9) predicts that the impact of the underlying genealogical process on neutral sequence depends on the compound parameters Ns and r/λ , but not directly on ρ . Hence, if s is increased and N decreased while holding their product constant, and r and λ are increased while holding their ratio constant, (2.9) predicts that patterns of variation will be maintained.

Finally, (2.3) suggests as we decrease N and increase s for fixed α we must also increase the length of the flanking sequence, because selection at more distant sites can impact the neutral locus as s is increased. Note that we can accomplish this either by fixing the recombination rate and increasing the length in base pairs of the flanking region or by increasing the recombination

rate for some fixed flanking length. Since the same number of mutations are introduced, these options are functionally identical, but the latter may require less RAM for some forward simulation implementations.

Taken together, these scaling principles suggest a simple algorithm for choosing simulated values of N_1 , α_1 , r_1 , and λ_1 that are conservative with respect to the genealogical process as predicted by (2.9). We wish to model L_0 flanking base pairs of sequence in a population of size N_0 with parameters ρ_0, α_0 , and λ_0 . We choose L_1 and N_1 to be any computationally convenient flanking length and population size. We compute the remaining simulation parameters with Algorithm 1.

procedure ALGORITHM 1($\rho_0, \alpha_0, L_0, \lambda_0, N_0, N_1, L_1$)

Let $s_0 = \frac{\alpha_0}{2N_0}$; $r_0 = \frac{\rho_0}{4N_0}$; $a = \frac{s_0}{L_0 r_0}$

$\alpha_1 = \alpha_0$

$s_1 = \frac{\alpha_1}{2N_1}$

$r_1 = \frac{s_1}{aL_1}$

$\lambda_1 = \frac{r_1 \lambda_0}{r_0}$

$\rho_1 = 4N_1 r_1$

return $N_1, \rho_1, \alpha_1, L_1, \lambda_1$

end procedure

Note that if we choose $L_1 = L_0$, we obtain $\alpha_1 = \alpha_0$, $\rho_1 = \rho_0$, and $4N_1 \lambda_1 = 4N_0 \lambda_0$, which is consistent with the rescaling strategy of HOGGART *et al.* 2007 and diffusion theory.

In Figures 2.3A-C, we show results obtained with Algorithm 1. In 2.3A, we plot the normalized difference in mean diversity between simulations performed in a population with $N_0 = 5,000$ and simulations performed with rescaled parameters and varying choices of N_1 . The dashed black line at 0 represents the expectation under perfect rescaling, because perfect rescaling will result in a normalized difference in means equal to zero between rescaled parameters and the original parameters. The colored points each represent the mean of 5,000 simulations and the solid colored curves were explicitly calculated with (2.9).

Algorithm 1 generates patterns of diversity in the rescaled populations (colored points, 2.3A) that are similar to the simulated diversity in the model population (black dashed line, 2.3A) when the strength of selection is low, but the algorithm performs poorly when the strength of selection gets arbitrarily large. Qualitatively similar results are observed for the variance in π (2.3B) and Tajima's D (2.3C). Furthermore, the mean diversity of simulations performed with Algorithm 1 does not agree well with explicit calculation of the expected diversity using (2.9) when selection is

strong, as seen by the divergence between the mean diversity in the simulations and the solid curves. In fact, (2.9) predicts that the diversity will decrease as N grows because of the dependence of $I_{\alpha,s}$ on s (2.3A, solid colored curves), but simulations show the opposite pattern. This demonstrates that the simulated value of s has some effect on the expected patterns of diversity (which is not predicted by the results of WIEHE and STEPHAN 1993, which we used to build Algorithm 1), and that (2.9) does not appropriately model this dependence.

In the next subsections we examine circumstances under which the assumptions used to derive (2.9) and (2.10) may break down, and we use insights from this analysis to design a more robust approach to parameter rescaling.

2.4.2 RHH with large values of s

We have predicated Algorithm 1 on (2.9) and (2.10), and hence it is likely that it will not perform adequately in parameter regimes in which (2.9) or (2.10) is not accurate. Equation (2.9) was derived using (2.3), which used the assumption that s is small, so it is possible that in the large s regime (2.9) will fail to accurately predict the reduction in diversity. Here, we derive a theoretical form that describes the impact of RHH in the large s regime by conditioning on the altered dynamics of the selected locus under very strong selection.

For genic selection, the dynamics of the selected locus are described by

$$\frac{d}{dt}x(t) = \frac{sx(t)(1-x(t))}{1+2sx(t)} \quad (2.12)$$

For small s , the denominator of (2.12) is very close to 1 and is typically ignored (and was ignored in the derivation of (2.3) by SWL). However, for very large s the denominator is non-negligible, which slows the rate of growth of the selected site when it is at moderate to high frequency. To investigate RHH with large s , we solved (2.2) approximately, conditioning on (2.12) for the dynamics of the selected site (see Appendix for derivation). We find

$$p_{r_f}^*(r_f) = e^{-4r_f} \left(1 - \frac{2r_f}{s} \alpha^{\frac{-2r_f}{s}} \Gamma \left[\frac{-2r_f}{s}, \frac{1}{\alpha} \right] \right) \quad (2.13)$$

This result differs by only a factor of e^{-4r_f} from (2.3), but makes very different predictions for

large s and r_f . As s increases, more distant sites can impact the diversity at the neutral locus. In fact, s can be made arbitrarily large whereas r_f is constrained to remain less than 0.5. As a result, we expect that (2.9) will underestimate the observed diversity for large s . If we use Algorithm 1 to make N arbitrarily small and s arbitrarily large, (2.13) predicts that patterns of diversity in the simulated population may be significantly different from the larger population because of this s dependence. We denote the reduction in diversity calculated with (2.13) as

$$E_{\alpha,s}^* [\pi/\pi_0] = \frac{r/\lambda}{r/\lambda + \alpha I_{\alpha,s}^*} \quad (2.14)$$

with an asterisk to differentiate it from (2.9). $I_{\alpha,s}^*$ is computed exactly as in subsection 2.1, but replacing (2.3) with (2.13).

We performed simulations of RHH with large values of s to test (2.14). We find that (2.14) accurately predicts the impact of RHH on diversity for large s , whereas (2.9) is a poor predictor in the large s regime (Figure 2.4). We have performed this analysis primarily to explain the biased patterns of diversity produced by Algorithm 1, but we note that in some cases (e.g., microbes under extreme selection pressures), it is possible that s can be much larger than 0.1. Indeed, one experimental evolution study of *Pseudomonas fluorescens* reported values of s as large as 5 and a mean value of 2.1 (BARRETT *et al.* 2006). If and when s achieves such large values in recombining organisms it will be advantageous to use equation (2.13) in place of (2.3).

2.4.3 Robust parameter rescaling for RHH simulations

Using the results in the previous subsection, we modify Algorithm 1 to guard against violating the assumptions of the RHH model as we rescale the parameters of the simulations.

Let N_1 , L_1 , etc., be defined as in the previous subsection. Our goal is to reduce N_0 as much as possible without altering the underlying dynamical process by more than a prespecified amount. Let δ_I be the maximum deviation between $I_{\alpha,s}^*$ for a population of size N_1 and the model population of size N_0 that we are willing to accept in our simulations. For example, let $\delta_I = 0.01$ if we desire simulated sequences in which $I_{\alpha,s}^*$ in a population of size N_1 differs by no more than 1% from a population of size N_0 . Let δ_p be the maximum difference between $p_{\tau_f}^*(u)$ in populations of size N_0 and N_1 that we are willing to accept in our simulations, over all u from 0 to u^* , the length of the

flanking region. Qualitatively, δ_I is a constraint on the total area under $p_{\tau_f}^*(u)$, which influences the overall level of diversity, while δ_p is a constraint on the shape of $p_{\tau_f}^*(u)$, which influences the coalescent dynamics for a substitution that occurs at a given distance from the neutral locus. See Figure 2.5 for a pictorial explanation. We formalize these constraints in Algorithm 2. Note that every parameter chosen by Algorithm 2 is exactly with consistent Algorithm 1, but in Algorithm 2 we precompute how small we can make N without altering the dynamics of the RHH model.

procedure ALGORITHM 2($\rho_0, \alpha_0, L_0, N_0, \lambda_0, L_1, \delta_I, \delta_p$)
 Let $s_0 = \frac{\alpha_0}{2N_0}$; $r_0 = \frac{\rho_0}{4N_0}$; $a = \frac{s_0}{L_0 r_0}$; $u_{max} = 2/a$
 $\alpha_1 = \alpha_0 = \alpha$
 Numerically solve $\frac{I_{\alpha, s_0} - I_{\alpha, s_1}}{I_{\alpha, s_0}} = \delta_I$ for the quantity s_1 .
 $D = \text{Max}[p_{\tau_f, s_0}^*(u) - p_{\tau_f, s_1}^*(u)]$ over all u on $[0, u_{max}]$
if $D > \delta_p$ **then** Numerically solve $\text{Max}[p_{\tau_f, s_0}^*(u) - p_{\tau_f, s_1}^*(u)] = \delta_p$ for s_1 over all u on $[0, u_{max}]$
end if
 $N_1 = \frac{\alpha_1}{2s_1}$
 $r_1 = \frac{s_1}{aL_1}$
 $\lambda_1 = \frac{r_1 \lambda_0}{r_0}$
 $\rho_1 = 4N_1 r_1$
return $N_1, \rho_1, \alpha_1, L_1, \lambda_1$
end procedure

We implemented Algorithm 2 in Python, using the numerical optimization tools in SciPy (JONES *et al.* 2001–) for the numerical optimization steps. In Figures 2.3D-F, we demonstrate the performance of Algorithm 2 for three different values of $\delta_I = \delta_p$. Smaller values of δ generate sequences that are more closely matched to the diversity in the population of size N_0 , but require a larger simulated N_1 and are hence more computationally intensive. Note that for the values of δ that we have chosen herein, only a very small change in the overall diversity is expected. While larger values of δ may be acceptable for some applications, we do not recommend large values in general because the underlying dynamics are not necessarily expected to be conserved even if the change in overall diversity is small. Indeed, small deviations in mean π and Tajima’s D are observed for the largest value of δ with strong selection (2.3F).

Computational performance of the rescaled simulations is shown in Figure 2.11 (see Appendix).

2.4.4 The notion of “sufficiently distant” flanking sites

We designed Algorithm 2 to work for any given L_0 in a population of size N_0 . In the RHH literature, flanking regions of $L_0 = \frac{s}{r}$ are of particular interest because equation (2.3) suggests that sites that are more than $\frac{s}{r}$ base pairs from the neutral locus have no impact on the neutral site JENSEN *et al.* 2008, at least for small s . However, this result does not hold in the large s regime. First, the recombination fraction is not linear in the number of base pairs of flanking sequence when $r_f > 0.1$, and r_f cannot exceed 0.5, even for arbitrarily long flanking sequences. Equations (2.3) and (2.13) are functions of $\frac{r_f}{s}$, so as s gets large ($s > 0.5$) it is not possible to make compensatory linear increases in r_f . Second, the dynamics of the selected site are altered when s is large, as we noted in the previous subsections. In particular, (2.13) suggests that sites that have $r_f = 0.5$ (unlinked sites) have a non-negligible impact on the diversity at the neutral site when s is very large. Our model predicts the impact of L_u unlinked sites to be

$$E[\pi/\pi_0] = \frac{1}{1 + 2N\lambda L_u p_{r_f}^*(r_f = 0.5)} \quad (2.15)$$

Figure 2.6 shows the reduction in diversity, relative to neutrality, for simulations of RHH that include a neutral region and L_u unlinked selected sites, and *no* linked selected sites. Equation (2.15) accurately predicts the reduction in diversity for these simulations. These results highlight another problem with Algorithm 1. In Algorithm 1, we linearly increase the flanking sequence as s increases. However, for large s , the majority of these flanking sites are essentially unlinked to the neutral locus, but can have a non-negligible impact on the neutral locus. This is fundamentally different from the dynamics in the small s regime, where unlinked sites have no impact on the neutral locus.

While this result may not be intuitive, it is a natural consequence of very large values of s . Consider the implausible but instructive case when $s \approx 2N$. In the first generation after the selected site is introduced into the population, approximately half of the offspring are expected to be descendants of the individual with the selected site. At a locus that is unlinked to the selected site, one of the two chromosomes of the individual with the selected mutant is chosen with equal probability for each of the descendants, which causes an abrupt and marked decrease in diversity. Though this effect is more subtle in our simulations in Figure 2.6 (which have $1 < s \ll 2N$), there

is a measurable decrease in π due to the accumulated effect of unlinked sites with large s .

2.4.5 The role of interference

In the previous subsections, we have restricted our analysis to parameter regimes in which interference between selected sites is very rare, which is an assumption of the RHH model. However, one of the advantages of forward simulations is that they can be performed under conditions with high levels of interference.

In Figure 2.7, we examine the performance of Algorithm 2 with very high rates of positive selection. Note that the value of λ on the x -axis is the expected value in the absence of interference in a population of size N_0 , and the observed value of λ in the simulations is slightly lower. Point sizes in Figure 2.7 indicate the amount of interference between selected sites, as measured by the fraction of selected substitutions that overlap with at least one other substitution while segregating in the population. This is a conservative metric for the total effect of interference because it does not include the fraction of selected sites that are lost due to competition with other selected sites.

As the rate of interference increases, the theoretical predictions of equation (2.13) underestimate the reduction in diversity by an increasing amount (2.7A, black points). This is expected because as interference increases, a smaller fraction of selected sites reach fixation, and furthermore the trajectories of the sites that fix are altered due to competition. Neither of these effects is modeled by equation (2.13).

More strikingly, as the rate of interference increases, the separation between the rescaled populations (green and red points) and the original population (black points) also increases. This demonstrates that Algorithm 2 does not recapitulate the expected diversity in the rescaled populations when the rate of interference is high in the population of size N_0 . This result is expected when we consider that the rate of interference is a function of both the rate of substitutions and the time that selected substitutions segregate in the population before fixation. It is well known that the time to fixation is a function of both $\alpha = 2Ns$ and N , and cannot be written naturally as a function of only one or the other. Hence, when we rescale the population with fixed α , we necessarily change the amount of interference. We analyze this effect in more detail in the next subsection when we perform rescaling for two sets of parameters inferred in *Drosophila*.

2.4.6 An application to *Drosophila* parameters

In Figure 2.8, we perform rescaling with RHH parameters that are relevant to *Drosophila*. MACPHERSON *et al.* 2007 found evidence supporting strong selection ($s = 0.01$), which occurred relatively infrequently ($\lambda = 3.6 \times 10^{-12}$) in a *Drosophila* population of size $N_0 = 1.5 \times 10^6$. JENSEN *et al.* 2008 found weaker ($s = 0.002$), more frequent selection ($\lambda = 10^{-10}$) in a population of $N_0 = 10^6$. Our goal is not to debate the “true” parameters, but rather to investigate the practicability of rescaling using previously inferred parameters. Assuming a flanking sequence length of $L_0 = s_0/r_0$ and a recombination rate of $r_0 = 2.5 \times 10^{-8}$, we apply Algorithms 1 and 2 to these parameter sets and investigate the effect on diversity.

In Figure 2.8A, we show that the trend in simulated diversity (solid red curve) as a function of N under the MACPHERSON *et al.* 2007 parameters is correctly described by (2.14), which predicts that the diversity decreases at low N . However, the model slightly underestimates the mean diversity compared to simulations. In contrast, the model predictions of the diversity are very inaccurate under the parameters estimated by JENSEN *et al.* 2008 (Figure 2.8B, solid blue curve). In both 2.8A and 2.8B, Algorithm 1 strongly alters the patterns of diversity as N is decreased. The value of N_1 calculated with Algorithm 2 and $\delta_I = \delta_p = 0.06$ are shown with the dotted vertical line.

Point sizes in 2.8A and 2.8B indicate the proportion of substitutions that are introduced while another substitution is on the way to fixation (as in Figure 2.7). While the interference is fairly mild for large values of N under the parameters of MACPHERSON *et al.* 2007, the amount of interference is extreme at all values of N under the JENSEN *et al.* 2008 parameters. In both cases the amount of interference in the simulations changes as we rescale N_1 .

We designed Algorithm 2 under the assumption that interference is negligible in the population of size N_0 . This assumption is approximately met under the parameters of MACPHERSON *et al.* 2007, where sweeps are infrequent and overlapping sweeps are rare. However, this assumption is broken by the parameters of JENSEN *et al.* 2008, where the rate of sweeps is more than an order of magnitude higher. As a result, the diversity is not well predicted by equation (2.14) at any value of N , and Algorithm 2 fails to generate sequences with accurate patterns of genetic diversity.

In general, it is useful to know *a priori* when the assumptions of the RHH model are not met as a result of high interference in a population of size N_0 . Consider the probability that a positively

selected substitution arises in the population while another substitution is heading towards fixation, p_{inter} . Under the assumption that interference is sufficiently infrequent such that the mean time to fixation and probability of fixation are not strongly altered, we can approximate p_{inter} as

$$p_{inter} \approx 1 - (1 - 2L\lambda)^{\tau_f} = 1 - \left(1 - \frac{2s\lambda}{r}\right)^{\frac{2(1+s)\log 2N}{s}} \approx \frac{4(1+s)\lambda \log 2N}{r} \quad (2.16)$$

The probability of no substitution in a single generation is $1 - 2L\lambda$, and hence the probability that no new selected substitutions are introduced while a given selected mutation is on its way to fixation is $(1 - 2L\lambda)^{\tau_f}$. Supposing that $L = \frac{s}{r}$ and plugging in the expectation of τ_f garners the rest of the terms in the equation. The final approximation is valid for very small $\frac{2s\lambda}{r}$. Note that we do not expect (2.16) to hold exactly in any parameter regime because the time to fixation is actually a random variable (and furthermore, both τ_f and λ are altered when interference is frequent), but we find that p_{inter} is a useful approximation for describing the interference in simulations during the rescaling process.

In Figures 2.8C and 2.8D, we investigate the behavior of (2.16) as we rescale the population size. As N is decreased from N_0 , the value of p_{inter} initially decreases because the product $(1+s)\log 2N$ decreases. However, as N gets very small with Algorithm 1, $(1+s)\log 2N$ eventually begins to increase because $1+s$ increases faster than $\log 2N$ decreases, increasing the amount of interference.

Although the exact calculation of the effects of interference is very challenging, it is straightforward to calculate the value of p_{inter} in a population of size N_0 . If the value of p_{inter} is large (e.g., > 0.05 , as with the parameters in Figure 2.8B), then the assumptions of Algorithm 2 are broken and there is no guarantee that the rescaled simulated sequences will be sufficiently accurate. By contrast, if there is low interference in a population of size N_0 then it is safe to perform rescaling so long as the value of p_{inter} is constrained. In practice, the value of p_{inter} (or other quantities that are related to the rate of interference) can be taken as an additional constraint in the calculation of N_1 in Algorithm 2 such that the impact of interference is limited under rescaling.

2.5 Discussion

Simulations are an integral part of population genetics because it is often difficult to obtain exact analytical expressions for many quantities of interest, such as likelihoods for sequence data

under a given model. Until recently, forward simulations were not practical because of the large computational burden that they can impose. However, several new forward simulation techniques have been proposed and published (HOGGART *et al.* 2007; HERNANDEZ 2008; ZANINI and NEHER 2012; ABERER and STAMATAKIS 2013; MESSER 2013), and their use in population genetic studies is becoming increasingly popular.

Despite these computational advances, it remains very computationally intensive to simulate large populations and long chromosomes in a forward context. It is frequently necessary to perform parameter rescaling to achieve computational feasibility for parameter regimes of interest (e.g., $N > 10^5$ with long flanking sequences), particularly for applications such as Approximate Bayesian Computation which require millions of simulations for accurate inference. The hope of such rescaling efforts is that expected patterns of diversity will be maintained after rescaling, and that the underlying genealogical process will remain unaltered.

In this investigation, we tested a “naive” approach to parameter rescaling, and showed that this approach can strongly alter the expected patterns of diversity because it does not conserve the underlying genealogical process at the neutral site. In particular, for fixed values of α , s can get arbitrarily large as N is decreased, and previous theoretical results do not accurately predict the patterns of diversity in this parameter regime. We derived a new theoretical form for the reduction in diversity when s is large, and show that it has strong predictive power in simulations. We leveraged this result to develop a simple rescaling scheme (Algorithm 2) that approximately conserves the underlying genealogical process. We note that in practice Algorithm 2 may not always be necessary, and as long as s remains small (say, < 0.1) Algorithm 1 will suffice. The advantage of Algorithm 2 is that it allows us to quantitate the effect of rescaling, and to get the best possible computational performance for a given error tolerance.

It will be of great interest to extend the rescaling results for recurrent sweeps presented herein to models that include arbitrary changes in population size and interference between selected sites. In the case of changes in population size, we note that the strategy presented in Algorithm 2 can be easily extended to perform optimization across a range of population sizes such that the constraints are simultaneously satisfied at all time points in a simulation. This strategy is consistent with previous approaches to rescaling in the context of complex demography HOGGART *et al.* 2007.

Interference poses a greater challenge, because the amount of interference is dependent both on

the rate of substitution and the time that selected sites segregate. It is well known that the time to fixation of selected alleles cannot be written as a simple function of α and depends on N as well, and hence rescaling population size with fixed α alters the effects of interference on sequence diversity COMERON and KREITMAN 2002. Improved understanding of scaling laws for interference may be necessary in order to develop appropriate rescaling strategies, to the extent that such rescaling is possible at all in a forward simulation context. However, recent progress was made in the case of very strong interference, where scaling laws were recently derived by WEISSMAN and BARTON 2012.

The rescaling results presented here pose an interesting dilemma for the use of forward simulations in population genetic studies, as previously noted by KIM and WIEHE 2009. A major appeal of forward simulations (as compared to coalescent simulators) is the ability to incorporate arbitrary models (e.g., interference between selected sites, complex demographic processes) without knowing anything about the distribution of sample paths *a priori*. However, if simulations are only feasible when the parameters are rescaled, there is no guarantee for any given theoretical model that the rescaling will maintain expected dynamics. We also note that the rescaling method proposed herein was informed by in-depth knowledge provided by the previous work of several authors, and that in general it may not always be obvious which parameters must be simultaneously adjusted to maintain expected patterns of variation in simulations for a given complex model.

Nonetheless, forward simulation in population genetics has a bright future. Forward simulation remains the only way to simulate arbitrarily complex models. For many populations of interest (e.g., ancestral human populations), population size is sufficiently small such that it can often be directly simulated without rescaling. Continued computational advances in both hardware and software in coming years will expand the boundaries of computational performance of forward simulation. Finally, active development of the theory of positive selection, interfering selected sites, background selection, demographic processes, and the joint action thereof will lend further insight into parameter rescaling and advances in the use of forward simulations in population genetic studies.

2.6 Appendix

2.6.1 Derivation of $p_{\tau_f}^*$ for large s

In this subsection, we solve for the probability of identity p_{τ_f} when the selection coefficient is large. We will be concerned with the probability of identity p at various frequencies throughout the sweep process. We will subscript p with $t(x)$ to indicate the value of p at the time when the selected site reaches frequency x (e.g., $p_{t(1/2N)}$). The trajectory of the selected site for large s is given (2.12), while the dynamics of the neutral site are given by (2.2). We transform (2.2) into allele frequency space by dividing by (2.12). We obtain:

$$\frac{d}{dx}p = \frac{(1-p)(1+2sx)}{2Nsx^2(1-x)} - \frac{2r_f p(1+2sx)}{sx} \quad (2.17)$$

This equation can be solved in *Mathematica* with the initial condition $p_{t(1/2N)} = 1$, meaning that all backgrounds carrying the selected locus are identical at the neutral locus when the selected site is introduced. At the end of the sweep, $x \approx 1$. We take the solution with $x = 1 - 1/2N$ because $x = 1$ results in a singularity.

$$\begin{aligned} p_{\tau_f} = & e^{2r_f(-2+\frac{1}{2N})+\frac{2N-2}{s-2Ns}} \left(2 - \frac{1}{N}\right)^{-\frac{1+2Nr_f+2s}{Ns}} \left(\frac{1}{N}\right)^{\frac{2+1/s}{2N}} \\ & \left(e^{\frac{2r_f}{N}} \left(\frac{1}{N}\right)^{\frac{1+4Nr_f+2s}{2Ns}} + 4 \frac{r_f}{s} e^{\frac{1}{s}} \left(2 - \frac{1}{N}\right)^{\frac{2+\frac{1}{s}}{2N}} \right. \\ & \left. \left(\int_1^{1-\frac{1}{2N}} e^{\frac{\frac{1}{C}-8Nr_f s C+(1+2s)\log[1-C]-(1+4Nr_f+2s)\log[C]}{2Ns}} \frac{(-1+2sC)}{2Ns(-1+C)C^2} dC \right. \right. \\ & \left. \left. - \int_1^{\frac{1}{2N}} e^{\frac{\frac{1}{C}-8Nr_f s C+(1+2s)\log[1-C]-(1+4Nr_f+2s)\log[C]}{2Ns}} \frac{(-1+2sC)}{2Ns(-1+C)C^2} dC \right) \right) \end{aligned} \quad (2.18)$$

Equation (2.18) can be numerically integrated in *Mathematica*. However, this solution is complicated, slow to evaluate, and provides little intuition about the dynamics. As an alternative, we employ an approximate solution strategy.

Following BARTON 1998 and others, we subdivide the trajectory of the selected allele into low

frequency and high frequency portions. For small x , the term $(1 + 2sx) \approx 1$, even for large s . As a result, there is little difference between the dynamics for small s and large s sweeps at low frequency. We rewrite (2.17) as

$$\frac{d}{dx}p = \frac{(1-p)}{2Nsx^2(1-x)} - \frac{2r_f p}{sx} \quad (2.19)$$

which is valid for low x . We define the solution to (2.19) on the interval $x = [1/2N, \epsilon]$ as $p_{t(\epsilon)}$.

For $x > \epsilon$, the second term on the RHS of (2.17) dominates the first term, because the first term is inversely proportional to the number of selected chromosomes. To obtain the high frequency dynamics of the selected allele, we take $p_{t(\epsilon)}$ as the initial condition and solve the following differential equation on the interval $x = [\epsilon, 1]$:

$$\frac{d}{dx}p = -\frac{2r_f p(1+2sx)}{sx} \quad (2.20)$$

We find the solution:

$$p_{\tau_f} = \left(e^{4r_f(\epsilon-1)} \right) \epsilon^{\frac{2r_f}{s}} p_{t(\epsilon)} \quad (2.21)$$

We can perform the exact same analysis under the assumption that the dynamics are given by $\frac{dx(t)}{dt} = sx(t)(1-x(t))$, as was done by SWL. This garners the solution:

$$p_{\tau_f}^{SWL} = \epsilon^{\frac{2r_f}{s}} p_{t(\epsilon)} \quad (2.22)$$

which differs by only a factor of $e^{4r_f(\epsilon-1)}$ from (2.21). Since (2.22) was derived under assumptions identical to those used in STEPHAN *et al.* 1992, we conclude that sweeps with large s can be modeled with the equation

$$p_{\tau_f} = e^{-4r_f} \left(1 - \frac{2r_f}{s} \alpha^{-\frac{2r_f}{s}} \Gamma \left[\frac{-2r_f}{s}, \frac{1}{\alpha} \right] \right) \quad (2.23)$$

This equation provides very similar results to (2.3) for small s , as expected, but deviates for large s (Figure 2.9).

To verify that this approximation provides accurate results to the full solution given by equation

(2.18), we compared (2.18) to (2.23) in *Mathematica*. Agreement is very good between the exact and approximate solutions for all values of s that we investigated (Figure 2.9).

2.6.2 RHH simulations in SFS_CODE

We performed forward simulations of RHH with SFS_CODE (HERNANDEZ 2008). An example command line for a RHH simulation is:

```
sfs_code 1 10 -t <θ> -Z -r <ρ> -N <N> -n <n> -L 3 <L> <l0> <L> -a N R -v L A 1 -v L
1 <Rmid> -W L 0 1 <α> 1 0 -W L 2 1 <α> 1 0
```

All of these options are described in the SFS_CODE manual, which is freely available online at sfscode.sourceforge.net, or by request from the authors. Briefly, this command line runs 10 simulations of a single population of size N and samples n individuals at the end. The recombination rate is set to ρ , and 3 loci are included in the simulation. The middle locus (locus 1) is l_0 base pairs long while the flanking loci are L bp long. The middle locus is neutral, while the flanking loci contain selected sites with selection strength $\alpha = 2Ns$. Every mutation in the flanking region is positively selected. The sequence is set to be non-coding with the option “-a N R”. The “-v” option provides the flexibility to designate different rates of mutation at different loci, and the mechanics of its usage are described in detail in the SFS_CODE manual. R_{mid} specifies the rate at which mutations are introduced into the middle segment relative to the flanking sequences. Please see the SFS_CODE manual for a detailed example of parameter choice for RHH simulations.

Forward simulations of DNA sequences can require large amounts of RAM and many computations. Recurrent hitchhiking models are particularly challenging to simulate because very long sequences must be simulated. In particular, for a given selection coefficient s , RHH theory suggests that sites as distant as $r_f \approx s$ must be included in the simulation to include all sufficiently distant sites (see (2.3)).

In many organisms, $r \approx 10^{-8}$. Assuming $r_f \approx rL$, this implies that a selection coefficient of $s = 0.1$ would require 10^7 base pairs of simulated sequence on each side of the neutral locus in order to include all possible impactful sites. This is a prohibitively large amount of sequence

for many reasonably chosen values of θ and N in forward simulations (Figure 2.10). However, in simulations of RHH, we are primarily interested in examining the diversity at a short, neutral locus. We adapted SFS_CODE such that individual loci can have different mutation rates and different proportions of selected sites. For RHH simulations, we set the proportion of selected sites to zero in the neutral locus, and 1 in the flanking sequence. This greatly increases the speed and decreases RAM requirements for SFS_CODE because much less genetic diversity is generated in the flanking sequences (Figure 2.10, blue curves). Time and RAM usage were measured with the Unix utility “time” with the command “/usr/bin/time -f ‘%e %M’ sfs_code [options]”. Note that time reports a maximum resident set size that is too large by a factor four due to an error in unit conversion on some platforms, which we have corrected herein. Simulations were performed on the QB3 cluster at UCSF, which contains nodes with a variety of architectures and differing amounts of computational load at any given time. As such, the estimates of efficiency herein should be taken only as qualitative observations.

2.6.3 Efficiency of rescaled simulations

We report the time to completion of rescaled simulations relative to non-scaled populations using Algorithm 2 (Figure 2.11). We observe reductions in time between approximately 99% and 40% for the parameters under consideration here. In general, the best performance is obtained for weaker selection, since in this case s is small in the population of size N_0 , meaning that the value of N can be changed quite dramatically without breaking the small s approximation. Better gains are also observed as the error threshold is increased, but this comes at an accuracy cost (see subsection 2.4.3).

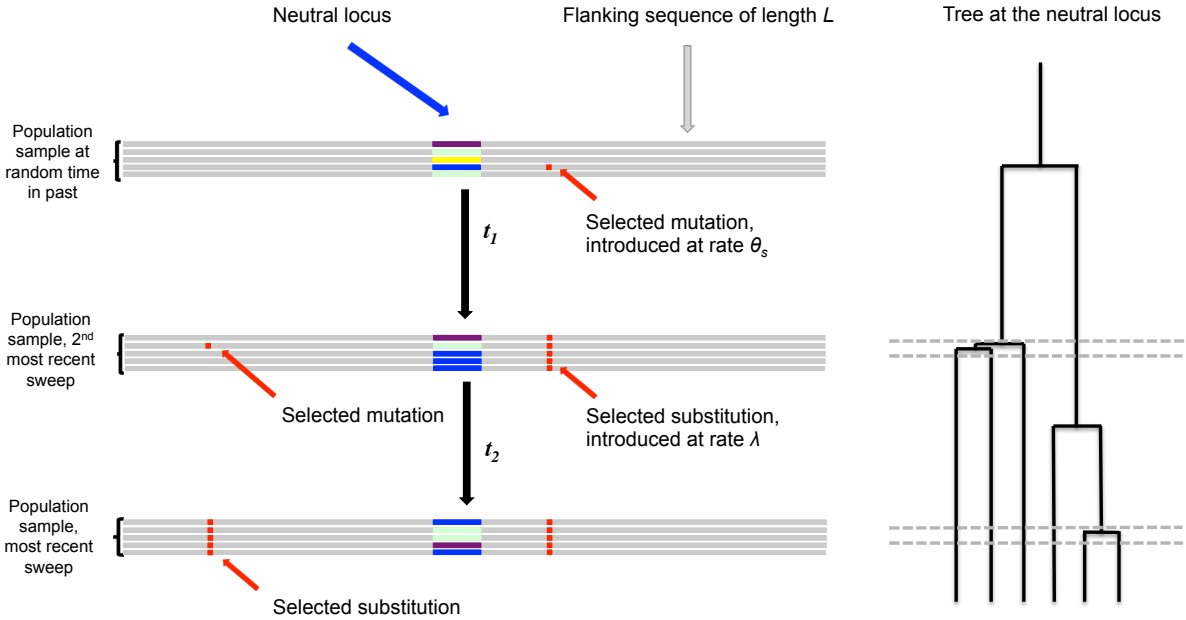


Figure 2.1: A pictorial representation of the recurrent hitchhiking model. Diverse neutral haplotypes are indicated with various colors at the neutral locus. When a selected site is introduced and eventually goes to fixation, it drags linked neutral variation to higher frequency. Selected mutations can occur at any distance from the neutral locus within the flanking sequence. Viewed from the perspective of the genealogy, sweep events generate an excess of recent common ancestry at a linked neutral site, reducing the overall height of the coalescent tree relative to neutrality. Selected sites that are more closely linked to the neutral site have a stronger impact on the overall height of the tree because they induce more common ancestry on a short time scale. The overall impact of linked selection at a neutral locus is a function of the strength of selection, the rate of recombination, and the rate at which selected sites reach fixation. The neutral site is assumed to be non-recombining, but this assumption can be relaxed in simulations.

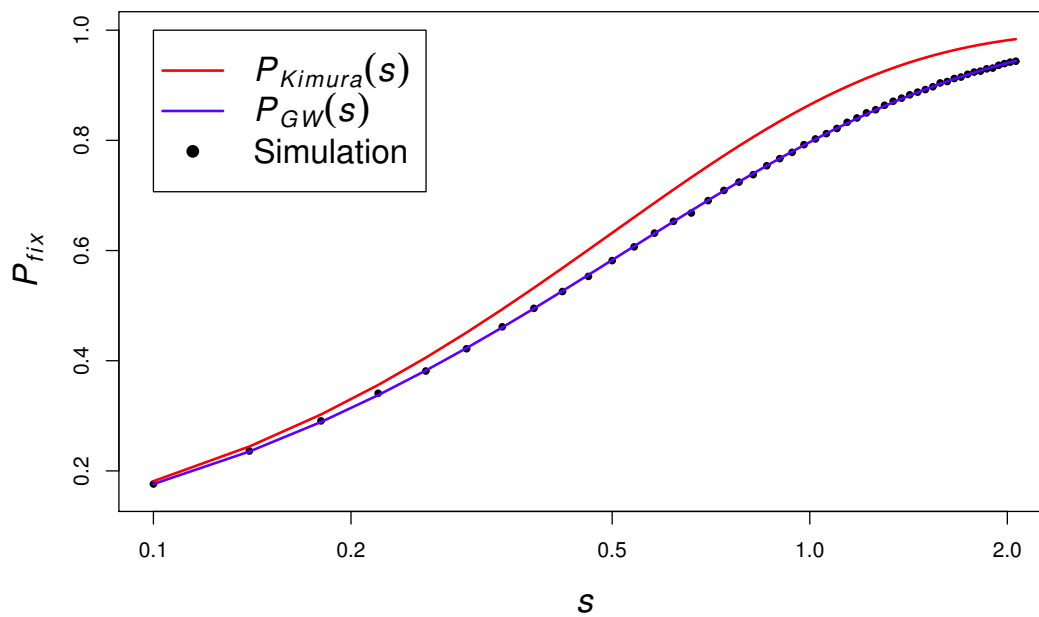


Figure 2.2: The probability of fixation (P_{fix}) as a function of s . Simulation points represent the fraction of fixations in 10^5 simulations. $N = 10^4$.

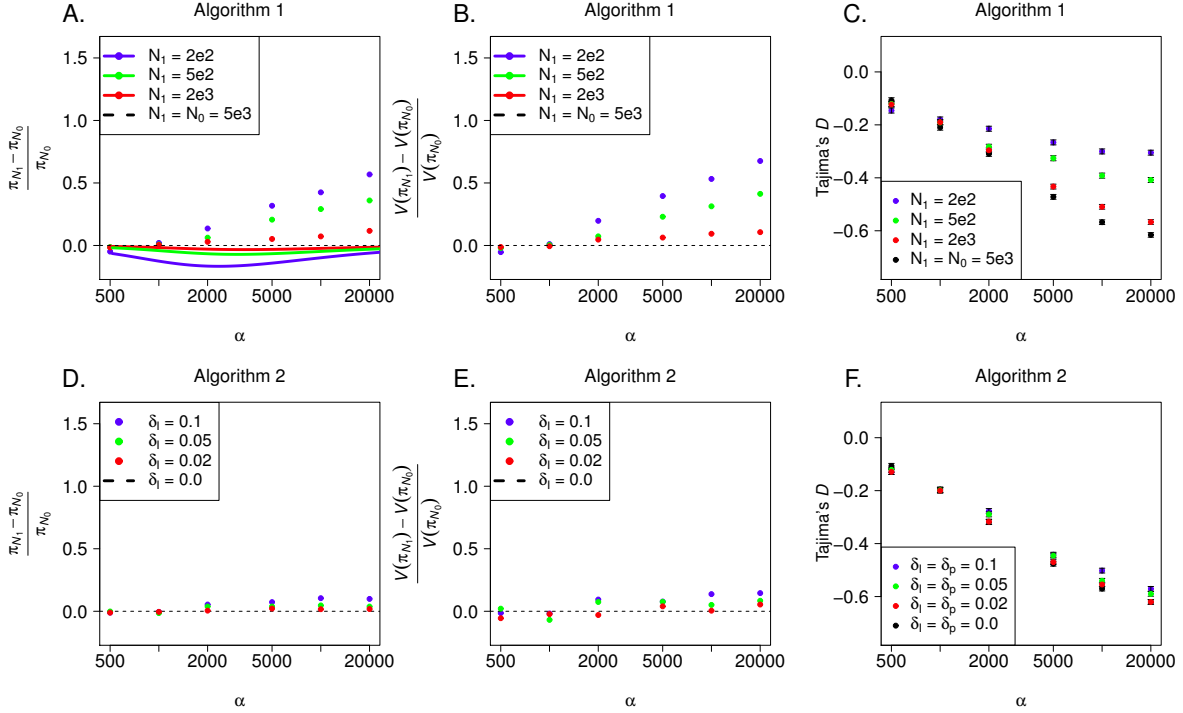


Figure 2.3: Mean and variance of observed diversity in the rescaled populations (N_1) relative to the model population ($N_0 = 5,000$). Rescaled parameters were obtained with Algorithm 1 in panels A, B, and C and with Algorithm 2 in panels D, E, and F. 10,000 simulations were performed for each parameter combination. The theoretical curves in A were calculated with (2.9) in *Mathematica* (WOLFRAM 2010). Parameters: $N_0 = 5 \times 10^3$, $\rho_0 = 10^{-3}$, $\lambda_0 = 10^{-10}$, $L_0 = 10^6$, $L_1 = 10^5$. Panels C and F show the mean Tajima's D for the same simulations. Error bars in C and F are the standard error of the mean.

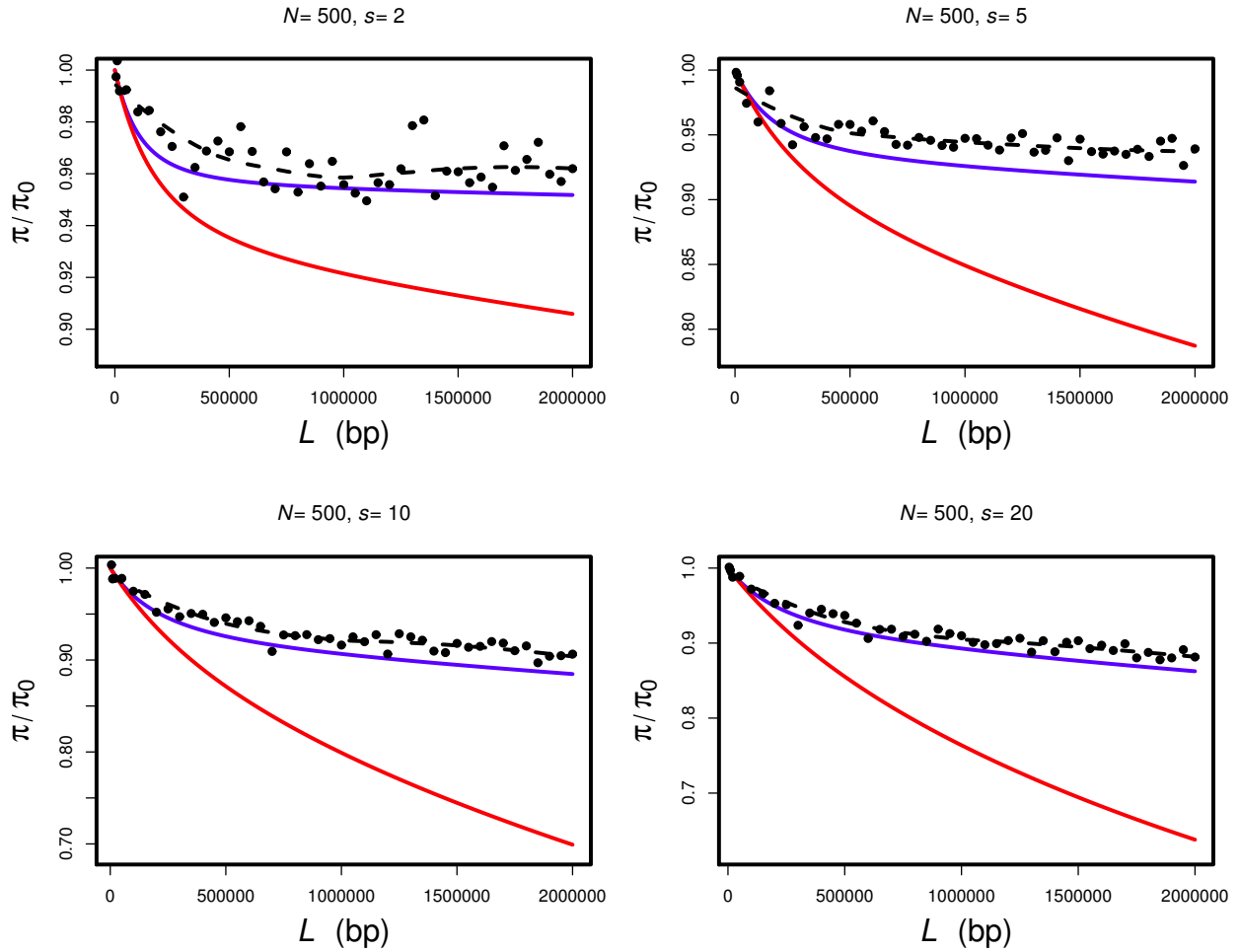


Figure 2.4: Simulations of recurrent hitchhiking with differing amounts of flanking sequence. In each panel, we vary the amount of flanking sequence and calculate the expected reduction in diversity using (2.9) (red curves) or (2.14) (blue curves), where we have used $u^* = \frac{2r}{s}L$ as the upper bound of the integration for calculating $I_{\alpha,s}$ and $I_{\alpha,s}^*$. Simulation points each represent the mean of 5,000 simulations, and the dashed black lines represent loess smoothing of the simulated data.

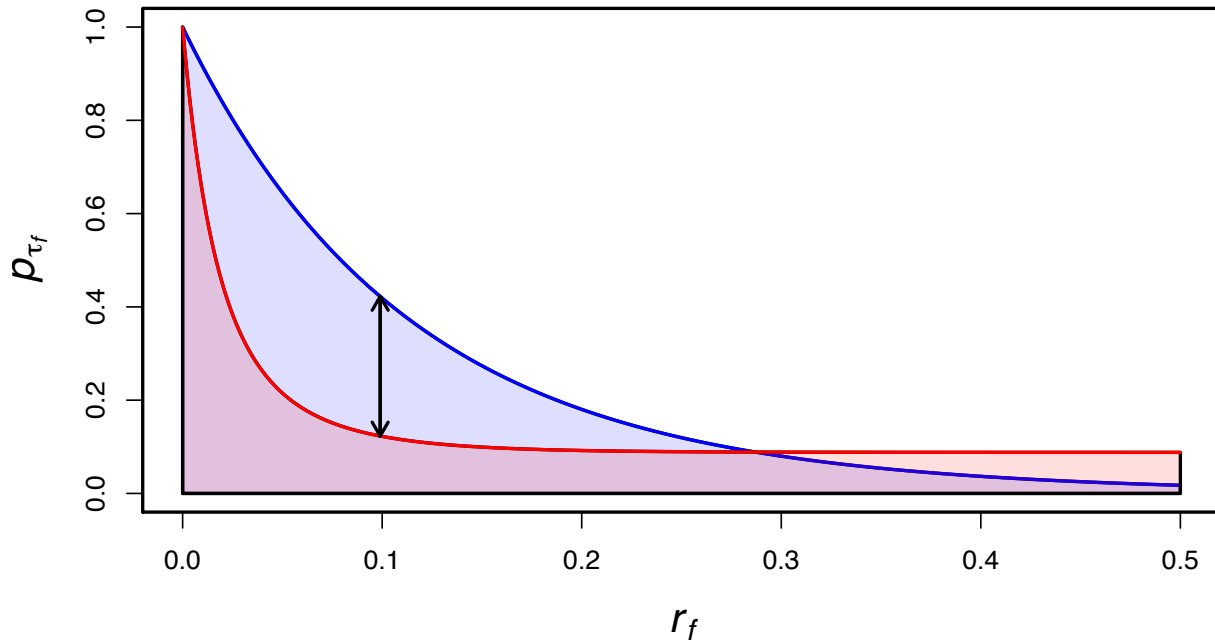


Figure 2.5: Algorithm 2 computes the difference between $p_{\tau_f}^*(r_f)$ in the population that we wish to model (blue) and a simulated population of smaller size (red) and bounds the difference in the probability of identity in the simulated and model populations. δ_I represents the maximum difference between the area under the red curve and the area under the blue curve that is acceptable in simulations. δ_p represents the maximum difference between the red and blue curves that is acceptable over all values of r_f (represented by the black arrow). Qualitatively, δ_I constrains the overall diversity in the simulated sequences, while δ_p constrains the shape of the probability of common ancestry during a sweep as a function of recombination distance.

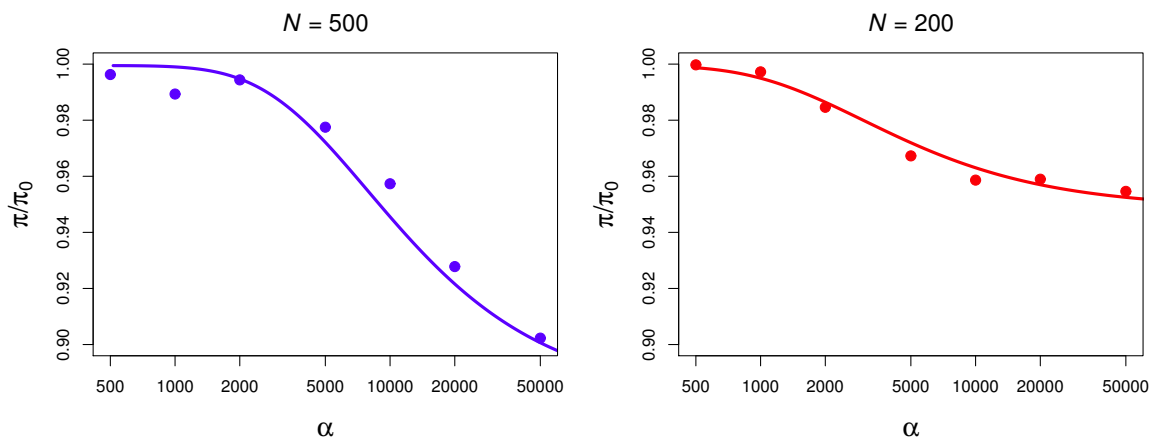


Figure 2.6: Simulations of RHH including only a neutral region and L_u explicitly unlinked selected sites. $L_u = 2000$, $\lambda = 0.5 \times 10^{-6}$. Points represent the mean of 4×10^4 simulations, solid lines were calculated with equation (2.15) in *Mathematica*.

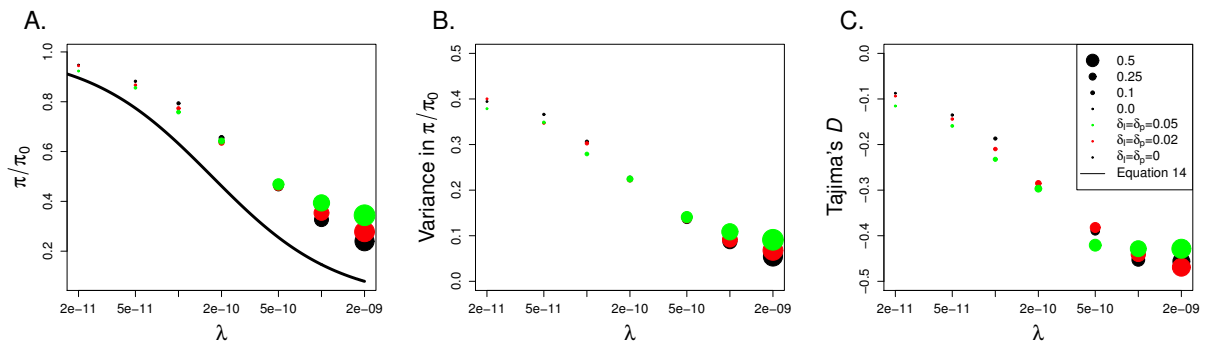


Figure 2.7: Rescaled simulations of RHH with increasing amounts of interference. 10,000 simulations were performed for each data point. Point sizes are indicative of the amount of interference in the simulations, as measured by the fraction of selected substitutions that overlap with at least one other positively selected substituted allele while both are segregating in the population. Parameters: $\alpha = 2 \times 10^3$, $L_0 = \frac{s_0}{r_0} = 4 \times 10^6$, $r_0 = 5 \times 10^{-8}$, $N_0 = 5 \times 10^3$.

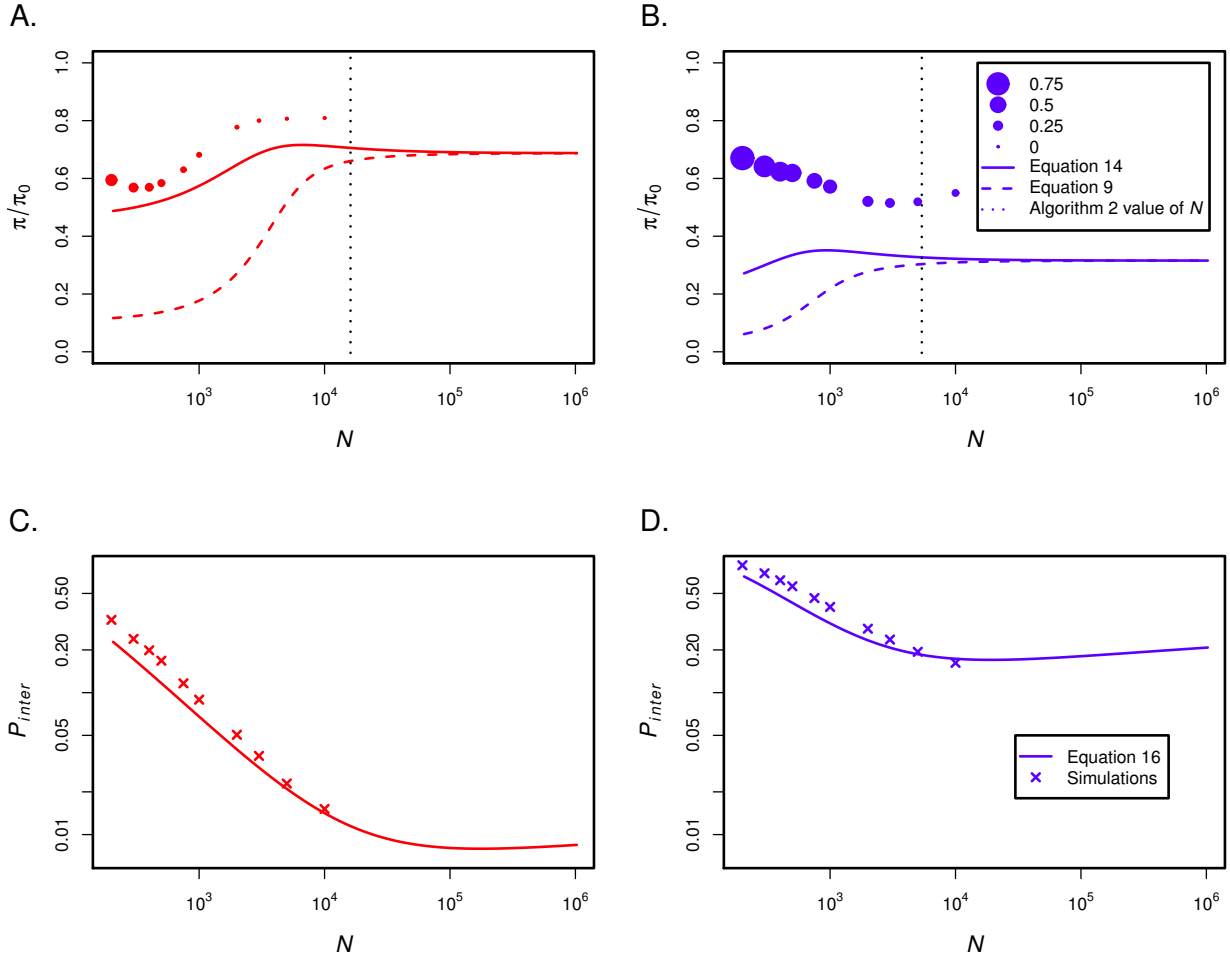


Figure 2.8: The reduction in diversity under previously inferred RHH parameters. Panels A and C use the parameters inferred by MACPHERSON *et al.* 2007 while panels B and D use the parameters inferred by JENSEN *et al.* 2008. The solid curves in A and B are given by (2.14), and the dashed curves are given by (2.9). Panels C and D show the probability of interference as measured by the proportion of substitutions that are introduced while another substitution is segregating in the population. The solid curves are given by (2.16). The size of the points in A and B is proportional to the observed value of p_{inter} , which is plotted in C and D. The dotted vertical lines show the value of N where $\delta_I = \delta_p = 0.05$.

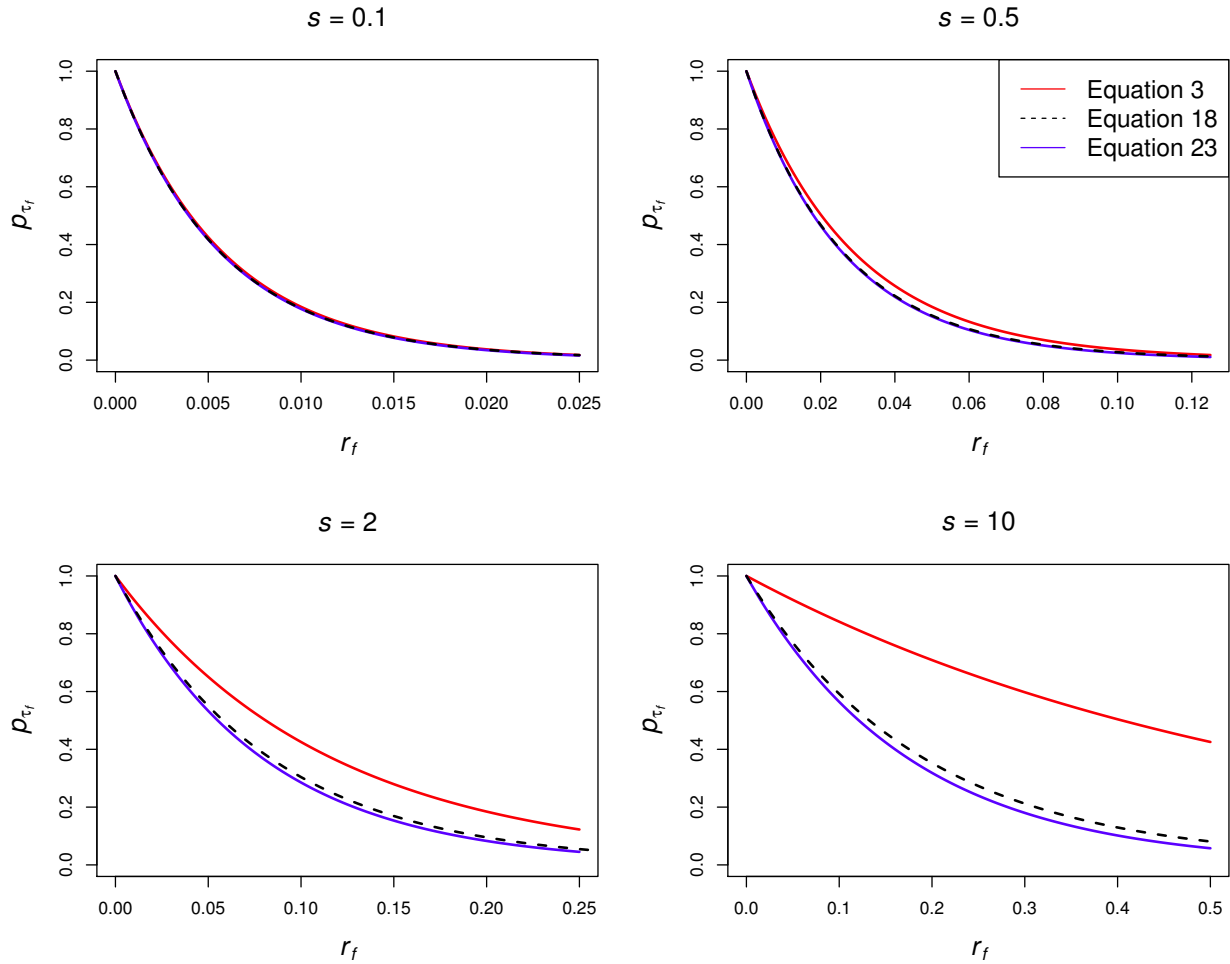


Figure 2.9: Equation (2.23) (blue) and Equation (2.18) (black) compared to (2.3) (red). As expected, (2.23) and (2.3) are in very close agreement for small s , but diverge for large s . Equation (2.23) is a good approximation to (2.18) across a wide range of values of s . $\alpha = 10^4$

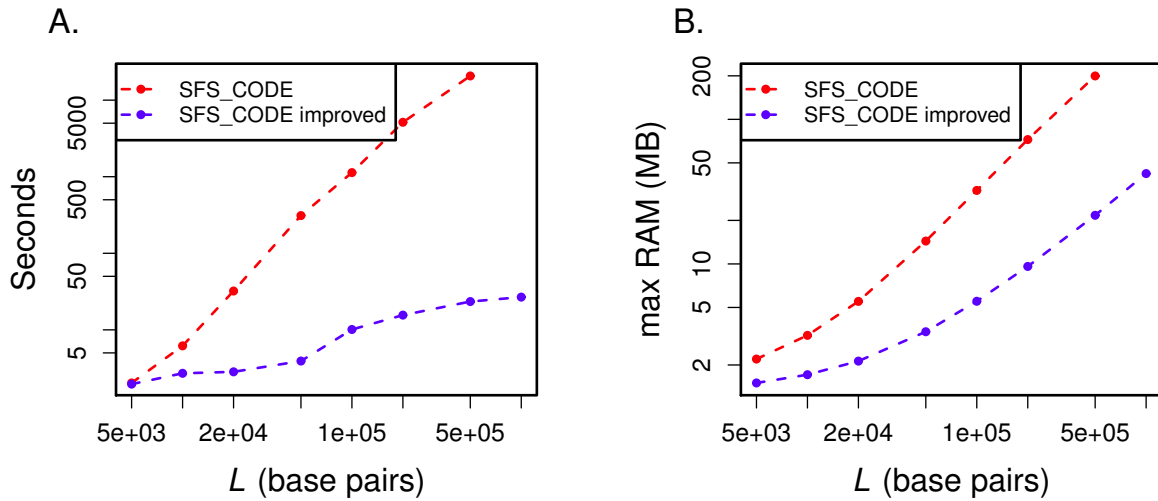


Figure 2.10: The computational burden of simulations of recurrent positive selection in SFS_CODE is much lower in the new version of SFS_CODE. Both RAM requirements and time to complete the simulations are reduced. Points represent the mean of 500 simulations. $\theta = 10^{-3}$, $\rho = 10^{-3}$, $N = 500$, $\lambda = 10^{-9}$, $\alpha = 1000$.

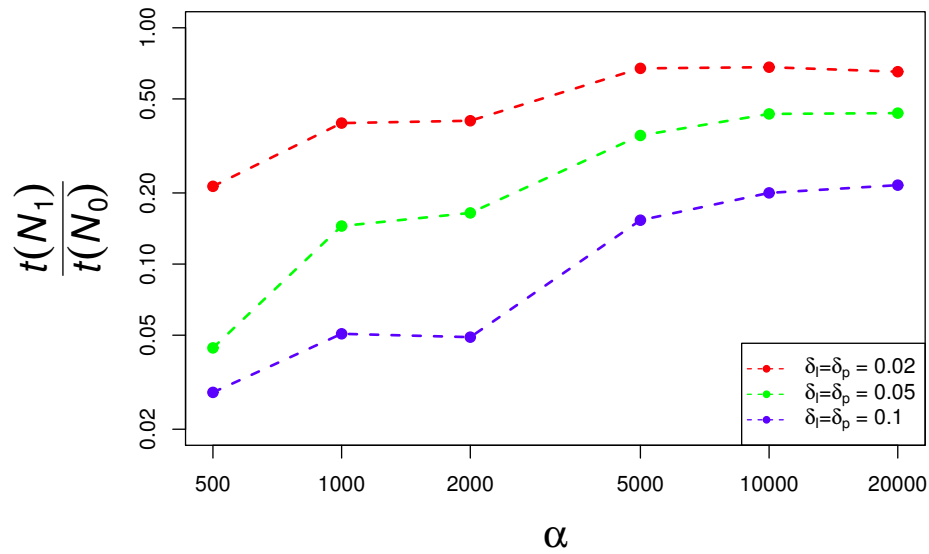


Figure 2.11: Duration of rescaled simulations ($t(N_1)$) relative to non-scaled simulations ($t(N_0)$). Parameters were chosen to match Figure 2.3. Parameters: $N_0 = 5 \times 10^3$, $\rho_0 = 10^{-3}$, $\lambda_0 = 10^{-10}$, $L_0 = 10^6$, $L_1 = 10^5$.

References

- ABERER, A. J., and A. STAMATAKIS, 2013 Rapid forward-in-time simulation at the chromosome and genome level. *BMC Bioinformatics* **14**: 216–216.
- ARBIZA, L., I. GRONAU, B. A. AKSOY, M. J. HUBISZ, B. GULKO, *et al.*, 2013 Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet* **45**: 723–729.
- BACHTROG, D., 2008 Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster*, two species with different current effective population sizes. *BMC Evol Biol* **8**: 334–334.
- BARRETT, R. D., R. C. MACLEAN, and G. BELL, 2006 Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. *Biol Lett* **2**: 236–238.
- BARTON, N. H., 1998 The effect of hitch-hiking on neutral genealogies. *Genetics Research* **72**: 123–133.
- BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate bayesian computation in population genetics. *Genetics* **162**: 2025–2035.
- BUSTAMANTE, C. D., A. FLEDEL-ALON, S. WILLIAMSON, R. NIELSEN, M. T. HUBISZ, *et al.*, 2005 Natural selection on protein-coding genes in the human genome. *Nature* **437**: 1153–1157.
- CHEVIN, L.-M., S. BILLIARD, and F. HOSPITAL, 2008 Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation. *Genetics* **180**: 301–316.
- COMERON, J. M., and M. KREITMAN, 2002 Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**: 389–410.
- COOP, G., and P. RALPH, 2012 Patterns of neutral diversity under general models of selective sweeps. *Genetics* **192**: 205–224.
- CRISCI, J. L., Y.-P. POH, A. BEAN, A. SIMKIN, and J. D. JENSEN, 2012 Recent progress in polymorphism-based population genetic inference. *Journal of Heredity* **103**: 287–296.
- CUTTER, A. D., and B. A. PAYSEUR, 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet* **14**: 262–274.
- FISHER, R. A., 1999 *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.
- HALDANE, J., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**: 299–309.
- HERNANDEZ, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**: 2786–2787.
- HOGGART, C. J., M. CHADEAU-HYAM, T. G. CLARK, R. LAMPARIELLO, J. C. WHITTAKER, *et al.*, 2007 Sequence-level population simulations over large genomic regions. *Genetics* **177**: 1725–1731.

- INGVARSSON, P. K., 2010 Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *populus tremula*. *Mol Biol Evol* **27**: 650–660.
- JENSEN, J. D., K. R. THORNTON, and P. ANDOLFATTO, 2008 An approximate bayesian estimator suggests strong, recurrent selective sweeps in *drosophila*. *PLoS Genet* **4**.
- JONES, E., T. OLIPHANT, P. PETERSON, *et al.*, 2001– SciPy: Open source scientific tools for Python.
- KAPLAN, N. L., R. R. HUDSON, and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. *Genetics* **123**: 887–899.
- KIM, Y., and W. STEPHAN, 2003 Selective sweeps in the presence of interference among partially linked loci. *Genetics* **164**: 389–398.
- KIM, Y., and T. WIEHE, 2009 Simulation of dna sequence evolution under models of recent directional selection. *Briefings in Bioinformatics* **10**: 84–96.
- KIMURA, M., 1962 On the probability of fixation of mutant genes in a population. *Genetics* **47**: 713–719.
- MACPHERSON, J. M., G. SELLA, J. C. DAVIS, and D. A. PETROV, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *drosophila*. *Genetics* **177**: 2083–2099.
- MAHER, M. C., L. H. URICCHIO, D. G. TORGERSON, and R. D. HERNANDEZ, 2012 Population genetics of rare variants and complex diseases. *Hum Hered* **74**: 118–128.
- MESSER, P. W., 2013 Slim: Simulating evolution with selection and linkage. *Genetics* .
- OTA, T., and M. KIMURA, 1975 The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). *Genet Res* **25**: 313–326.
- POOL, J. E., I. HELLMANN, J. D. JENSEN, and R. NIELSEN, 2010 Population genetic inference from genomic sequence variation. *Genome Res* **20**: 291–300.
- SINGH, N. D., J. D. JENSEN, A. G. CLARK, and C. F. AQUADRO, 2013 Inferences of demography and selection in an african population of *drosophila melanogaster*. *Genetics* **193**: 215–228.
- SMITH, J. M., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. *Genetics Research* **23**: 23–35.
- STEPHAN, W., Y. S. SONG, and C. H. LANGLEY, 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**: 2647–2663.
- STEPHAN, W., T. H. WIEHE, and M. W. LENZ, 1992 The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theoretical Population Biology* **41**: 237 – 254.
- TORGERSON, D. G., A. R. BOYKO, R. D. HERNANDEZ, A. INDAP, X. HU, *et al.*, 2009 Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet* **5**.

WEISSMAN, D. B., and N. H. BARTON, 2012 Limits to the rate of adaptive substitution in sexual populations. *PLoS Genet* **8**.

WIEHE, T. H., and W. STEPHAN, 1993 Analysis of a genetic hitchhiking model, and its application to dna polymorphism data from *drosophila melanogaster*. *Mol Biol Evol* **10**: 842–854.

WOLFRAM, 2010 *Mathematica* edition: Version 8.0.

ZANINI, F., and R. A. NEHER, 2012 Ffpopsim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics* **28**: 3332–3333.

3 Simulating complex phenotypes under evolutionary models

3.1 Introduction

Genome-wide association studies have identified many common loci that contribute to complex heritable phenotypes, but a large proportion of the heritability remains unexplained (MANOLIO *et al.*, 2009; WITTE, 2010). Proposed sources of this missing heritability include rare variants, environmental interactions, structural variants, common variants of weak effect, and upward biases in the original estimates of heritability. Sequencing studies with large numbers of samples may offer new opportunities to find the unexplained heritability of complex phenotypes, especially rare causal variants as these sites were mostly absent from and very poorly tagged by the original array-based genotyping platforms.

Unfortunately, power to detect rare causal variants using single-marker statistical tests at the genome-wide scale is generally much lower than is desirable. As a result, researchers have proposed statistical methods to pool rare variants within a putatively causal locus and jointly test for the contribution of these variants to the phenotype (HOFFMANN *et al.*, 2010; NEALE *et al.*, 2011; WU *et al.*, 2011; LEE *et al.*, 2012). While these methods all presuppose that low frequency causal sites have larger effects than high frequency causal sites, there are few mechanistic reasons for such a phenomenon other than the action of purifying selection restraining the frequencies of large effect mutations. Moreover, it has been argued that only those phenotypes with causal sites that are under selection will have a large fraction of heritability explained by rare variants (SIMONS *et al.*, 2014). Unfortunately, most rare variant association tools have not directly modeled selection on causal alleles or used simulations of selection to test their performance (but see KING *et al.* (2010) and PRICE *et al.* (2010)). Since both natural selection and demography have strong effects on the frequency spectrum of variant sites, these evolutionary forces may have considerable impact on the performance of rare variant association tests (ZUK *et al.*, 2014).

Demography and selection both impact genetic variation within populations, and population geneticists have developed a rich literature that models the effects of these forces on sampled DNA sequences (for reviews, see EMERSON *et al.* (2001); NIELSEN (2005)). In general, changes in population size alter the probability of common ancestry between two sampled sequences per generation as the genealogical history is traced backwards in time. Meanwhile, selection acts to reduce the overall amount of genetic variation by shrinking the time to common ancestry, and in

some cases also changes the shape of the genealogy. The height and shape of the genealogical tree relating sequenced chromosomes affect the total amount of variation in the samples and the frequencies of the variant sites.

However, only as the scale of sequencing experiments has increased in recent years has it become possible to apply models of selection and demography to real data sets. In particular, it is now possible to use statistical inference techniques to infer the parameters of demographic/selection models (e.g., population split times, growth rates, and the strength and rate of selection). Perhaps the most influential such model is known as the Poisson Random Field (SAWYER and HARTL, 1992), which has been used to infer both demographic events (GUTENKUNST *et al.*, 2009; GRAVEL *et al.*, 2011; TENNESSEN *et al.*, 2012) and selection (BUSTAMANTE *et al.*, 2001; WILLIAMSON *et al.*, 2005; BOYKO *et al.*, 2008; TORGERSON *et al.*, 2009).

As a result of these studies and others, we have rich information about the recent history of human continental groups and natural selection acting on human genomic elements such as conserved noncoding sequences and exons. In general, studies of human demography have found that human genetic variation is consistent with a population bottleneck as humans moved out of Africa and into Europe and Asia, and that in the recent past human populations have expanded rapidly. Studies of selection have found that most amino acid changes in proteins are weakly deleterious, and a substantial proportion of changes are strongly deleterious (BOYKO *et al.*, 2008). Moreover, conserved noncoding elements have a qualitatively similar distribution of selective constraints with a lower mean strength of selection (TORGERSON *et al.*, 2009), but there exist ultra-conserved noncoding regions in the human genome with even stronger selective constraints than coding regions (KATZMAN *et al.*, 2007).

There has been tremendous recent interest in both population genetic inference and association testing, and several studies connecting the fields have now been published (e.g., see PRITCHARD (2001); EYRE-WALKER (2010); MAHER *et al.* (2012); THORNTON *et al.* (2013); SIMONS *et al.* (2014); LOHMUELLER (2014)). Population genetics has direct implications for association studies, since the relationship between allele frequencies and effect sizes determines the power to detect causal sites. In particular, recent population growth and selection (KEINAN and CLARK, 2012; ?; TENNESSEN *et al.*, 2012) have both increased the proportion of sites at low frequency and impacted the total number of segregating sites in a sample. Accounting for the impact of selection and demography

on the frequency spectrum may be crucial to making sensible estimates of statistical power for association tests that pool putatively causal rare variants.

Perhaps the most widely used rare variant association test is the sequence kernel association test, or SKAT (WU *et al.*, 2011). SKAT provides a regression-based framework for rare variant association testing, and has several advantages over tests that count the prevalence of rare variants in cases and controls (collectively known as burden tests, e.g. COHEN *et al.* (2004); MORGENTHALER and THILLY (2007); LI and LEAL (2008)). SKAT retains statistical power when rare variants have effects with opposing directions and provides the machinery for covariate adjustment. Furthermore, many burden tests can be treated as special cases of SKAT (WU *et al.*, 2011). For these reasons, we focus on SKAT in this article (specifically SKAT-O, an optimized version of SKAT (LEE *et al.*, 2012; ?)).

Here, we introduce a simulation tool that incorporates recently inferred population genetic models of natural selection and demography, and accounts for the inferred functional elements and local recombination rate of any desired locus in the human genome. We demonstrate how local genomic features can impact patterns of variation within sampled DNA sequences, and show that accounting for these patterns may have practical implications for rare variant association test power calculations under some phenotype models. We also consider the impact of linked selection on patterns of genetic variation and discuss the simulation of phenotypes under models with selection.

3.2 Materials and Methods

3.2.1 `sfs_coder`: A Python-based interface to SFS_CODE

We built a Python-based front-end to the forward simulator SFS_CODE, which we have named `sfs_coder`. `sfs_coder` is designed to allow users to simulate human DNA sequences using inferred demographic histories and human selection models. A typical workflow in `sfs_coder` consists of 1) importing the appropriate modules, 2) performing SFS_CODE simulations of the desired population genetic model and/or locus in the human genome, and 3) analyzing the output of the command or simulating phenotype data using the simulated genetic data. The post-processing analysis tools include locus-by-locus computation of π , Tajima's D , Z_{nS} , Watterson's θ , Fay and Wu's H , and the site frequency spectrum. Each step can be accomplished with a few lines of code if the user

simulates one of the models that we have included, but any acceptable SFS_CODE command can be called from `sfs_coder`. Furthermore, advanced Python users can use `sfs_coder`'s object-oriented framework to write their own analysis tools beyond what we have provided.

In the next sections, we detail the models that are simulated in this paper, all of which are immediately accessible through our free software, except as noted. The software is available at `sfscode.sourceforge.net` and complete documentation is available at `uricchio.github.io/sfs_coder`.

3.2.2 Simulations of human demography and selection

We simulated human demography and selection for three human continental groups (African, Asian, and European) using the parameter estimates of previous studies. All of the demographic models we considered include discrete population size changes, population splits, and continuous migration (SCHAFFNER *et al.*, 2005). Two of the models incorporate the above features with recent exponential growth in the European and Asian continental groups (GUTENKUNST *et al.*, 2009; GRAVEL *et al.*, 2011), and the final model incorporates all of the above features with recent acceleration of the growth rate in the African and European continental groups (TENNESSEN *et al.*, 2012). The model of SCHAFFNER *et al.* (2005) was simulated with the coalescent simulator `cosi` (<http://www.broadinstitute.org/~sfs/cosi/>) and is not included in `sfs_coder`, while the other models were simulated using the forward simulator SFS_CODE (HERNANDEZ, 2008) and `sfs_coder`. The parameters of the population split times, migration rates, bottlenecks, and growth rates were obtained from the maximum likelihood estimates of the relevant publications (GUTENKUNST *et al.*, 2009; GRAVEL *et al.*, 2011; TENNESSEN *et al.*, 2012). Model parameters were scaled as described in the SFS_CODE manual, available at `sfscode.sourceforge.net`. Sample SFS_CODE command lines for each of the different models are provided in the Appendix.

In all simulations of selection on coding regions, we used the distribution of selection coefficients on non-synonymous sites that was inferred by BOYKO *et al.* (2008). For conserved noncoding elements, we applied the distribution inferred by TORGERSON *et al.* (2009). Both are Γ -distributed, and parameters for the distributions are given in the Appendix.

To summarize the results of our simulations, we plot the cumulative site frequency spectrum (cSFS). The value of the cSFS at frequency x is defined as the proportion of variant sites below or equal to frequency x in our simulations. We also report the nucleotide diversity, π , for some of

our simulations. π is defined as the mean number of pairwise differences per base pair between a random pair of chromosomes within the sample. The values of π that we report are the mean over a set of independent simulations.

3.2.3 Simulations of genomic elements

`sfs_coder` allows users to input the coordinates of a human genomic region and models the local genomic structure of this region. We model human genomic elements within `sfs_coder` as shown in Figure 1. The positions of exons were extracted from GENCODE v14 (HARROW *et al.*, 2012). The positions of conserved noncoding elements were inferred by SIEPEL *et al.* (2005), and recombination rates by INTERNATIONAL HAPMAP CONSORTIUM *et al.* (2007). Data sources for each of these elements are included in `sfs_coder`.

Furthermore, we allow users to specify one of several recently inferred models of human demography, namely those of GUTENKUNST *et al.* (2009), GRAVEL *et al.* (2011), and TENNESSEN *et al.* (2012), simultaneous to simulating complex genomic structure. Both the exonic regions and conserved noncoding regions are under selection in the simulations, with selection coefficients drawn from distributions that were inferred specifically for these regions by recent studies (BOYKO *et al.*, 2008; TORGERSON *et al.*, 2009).

3.2.4 Haplotype resampling with Hapgen2

Haplotype resampling methods provide an efficient mechanism for simulating large samples based on an existing reference panel, and constitute an alternative to forward simulations of DNA sequences. Such methods include Hapgen2 (SU *et al.*, 2011) and others, and are often applied based on the deep catalog of variation represented by the HapMap project (INTERNATIONAL HAPMAP CONSORTIUM *et al.*, 2007) or the 1000 Genomes Project (1000 GENOMES PROJECT CONSORTIUM *et al.*, 2012). Haplotype resampling has the demonstrated ability to recapitulate the haplotype and genetic variation of large samples when the population size remains constant, but it is not clear that they will perform well in cases of rapid population growth. Recent studies have suggested that the excess of rare variants associated with recent growth may only be detectable with very large sample sizes (KEINAN and CLARK, 2012; TENNESSEN *et al.*, 2012).

We simulated thirty unlinked 10Mb regions of the human genome under the European demo-

graphic model of TENNESSEN *et al.* (2012) and sampled 10^4 chromosomes using SFS_CODE. Each 10Mb region was based on chr15:59200000-69200000, and incorporated natural selection on all exons and conserved noncoding regions, as well as the genetic map inferred by the HapMap project (INTERNATIONAL HAPMAP CONSORTIUM *et al.*, 2007). We randomly chose 100 chromosomes from this simulation to form a reference panel. We then used Hapgen2 to resample this reference panel up to a larger sample size of 10^3 , 5×10^3 , or 10^4 chromosomes and compared the distributions of derived allele frequencies (DAF) at each sample size to a random subsample of equivalent size from SFS_CODE simulations.

3.2.5 Choosing a region under strong linked selection

We performed simulations of a genomic region under strong linked selection using `sfs_coder`. To select a candidate region for these simulations, we computed the density of phastCons elements (SIEPEL *et al.*, 2005) and total genetic distance separately in 1Mb sliding windows (10 kb sliding distance) across the human genome (hg19). We then took the intersect of those windows that were in the top 10% of the distribution of phastCons (phastConsElements46wayPlacental) density and the bottom 10% of the distribution of recombination distance (HapMapII GRCh37). From that intersection, we extracted windows that had a mean B value less than 25 (indicating very strong background selection (MCVICKER *et al.*, 2009)). We found 931 transcripts (GENCODE version 14) that fell within these regions. We then picked chr3:50320000-50350000 for our simulations of selection and power since it had among the highest densities of transcripts within this set of regions. Our background selection simulations incorporate the 2 Mb surrounding this region (chr3:49335000-51335000). This region also contains 13 Genome-Wide Association Study loci in the NHGRI GWAS Catalog.

3.2.6 Simulations of phenotypes and the power of SKAT-O

We followed WU *et al.* (2011) in simulating phenotypes and testing the power of the rare variant association method SKAT-O. Effect sizes $\beta(x)$ of causal variants were taken as $\beta(x) = -0.4 \log_{10}(x)$, where x is the minor allele frequency. Thus, lower frequency sites have larger effect sizes. Among variants under 3% frequency, 5% were taken to be causal. Phenotypes, Y , of each sampled individual were then generated as

$$Y = X_1 + X_2 + \beta G + \epsilon \quad (3.1)$$

where X_1 is a standard normal covariate, X_2 is a dichotomous covariate that takes the value 0 with probability 0.5 and the value 1 otherwise, and ϵ is a standard normal random variable (not taken as a covariate). β is the vector of effect sizes and G represents the genotypes. This phenotype model is included in `sfs_coder`, as are the models of EYRE-WALKER (2010) and SIMONS *et al.* (2014), which map selection coefficients, rather than allele frequencies, to effect sizes. Our software also allows the user to set the desired genetic variance explained by the causal sequence for each of the models.

For each of the demographic models considered in our power calculations, we generated 250 independent simulations with the relevant sample size and selective constraint for the human genomic locus at chromosome 3, hg19 positions 50320000-50350000. For each simulation, we randomly selected causal loci and generated phenotypes as described above. We resampled causal sites from each genetic simulation 4 times, for a total of 10^3 simulations of phenotypes for each demographic model. We then ran SKAT-O and computed the fraction of results with p -values under 10^{-6} .

Depending on the exact sequencing experiment performed, the number of statistical tests might range from approximately 2×10^4 (all genes) to 10^5 (all 30 kb sequences in the genome), so Bonferroni corrected significance thresholds may range from 5×10^{-7} to 2.5×10^{-6} . Here we have chosen 10^{-6} , but we emphasize that the trends in power as a function of sample size, selection, and demography are not dependent on this choice.

We obtained the SKAT R-package from <http://www.hsph.harvard.edu/skat/download/>.

3.3 Results

3.3.1 Demography, sample size, and selection impact rare variants

The results of rare variant association tests are contingent on the joint distribution of variant frequencies and effect sizes. In this section, we examine the effects of selection and demography on the simulated frequency spectrum through simulations.

We simulated human demography and selection under several previously inferred models (SCHAFFNER *et al.*, 2005; GUTENKUNST *et al.*, 2009; GRAVEL *et al.*, 2011; TENNESSEN *et al.*, 2012). In Figure 2,

we plot the simulated cumulative site frequency spectra of each of the three demographic models that we considered. Consistent with population genetic theory (KEINAN and CLARK, 2012) (and the data that was used to infer the models), exponential or two-phase exponential growth results in a large excess of rare variants relative to a constant population size model (gray dashed line). Furthermore, when sample size is large (5×10^3 chromosomes, lower panels), the two-phase model of exponential growth (TENNESSEN *et al.*, 2012) generates many more low frequency variants than the other growth models.

Our simulations also included natural selection on non-synonymous sites (BOYKO *et al.*, 2008). Negative selection tends to constrain variants to lower frequencies, so the site frequency spectra of non-synonymous sites are shifted towards the left (Figure 2, dotted lines). Note that we do not include non-synonymous sites for the model of SCHAFFNER *et al.* (2005) in Figure 2 because the coalescent simulator *cosi* does not allow for the introduction of natural selection.

3.3.2 Haplotype resampling under-estimates the number of rare variants in large samples

As an alternative to forward simulations, investigators might opt to resample haplotypes from a previously sequenced sample. In this section, we test whether a haplotype resampling method (also known as a “sideways” simulation, (CHEN *et al.*, 2014)) is able to recapitulate the extent of rare variation expected in large samples when based on a modest reference panel of 100 chromosomes in the context of rapid population growth.

We find that under the demographic model of TENNESSEN *et al.* (2012), Hapgen2 (SU *et al.*, 2011) does a poor job of recapitulating the extent of rare variation expected at large sample sizes. In Figure 3A, we show a quantile-quantile (QQ) plot of the DAF distribution inferred from Hapgen2 versus the DAF distribution expected by SFS_CODE under the TENNESSEN *et al.* (2012) European demographic model. If Hapgen2 were able to recapitulate the underlying DAF distribution Figure 3A would follow the diagonal dotted line. However, we find that as the sample size increases, the extent to which Hapgen2 underestimates the fraction of rare variants increases (indicated by curves deviating above the diagonal). In Figure 3B we look closely at the expected (based on SFS_CODE) and inferred (based on Hapgen2) frequencies of each SNP observed in the sample of 10^4 chromosomes using a scatter plot. We do not expect points to fall along the diagonal in this

case because of the resampling procedure, but we would expect the points to be symmetrically distributed about the diagonal (blue curve). Instead, we find that Hapgen2 DAF frequencies are skewed toward higher frequencies for rare variants. This is demonstrated using a loess smoothing (red curve). The loess curve shows that the Hapgen2 DAF may be strongly biased by the reference panel size.

3.3.3 Power estimates may be impacted by local genomic context and demography

Forward simulations allow investigators to model the effects of demography and selection on sampled DNA sequences (PENG *et al.*, 2014). Since recombination and natural selection jointly impact the number of segregating sites and the proportion of sites at low frequency, it may be important to accurately account for these features when performing power calculations. Moreover, in the case of a targeted resequencing study, it is desirable to model the genomic architecture of the target locus directly.

We tested the performance of the method SKAT-O with our simulations of demography and selection (LEE *et al.*, 2012). We simulated 30 kb of sequence from chromosome 3 (hg19 coordinates 50320000-50350000, which is a region under strong selection, see Methods), under two different demographic models (GUTENKUNST *et al.*, 2009; TENNESSEN *et al.*, 2012), with and without selection on coding and conserved noncoding elements. Selection coefficients were drawn from the distributions inferred by BOYKO *et al.* (2008) for coding regions and TORGERSON *et al.* (2009) for conserved noncoding regions. We also ran simulations where the entire 30 kb region was treated as a single gene (*i.e.*, ignoring the local structure of conserved elements and allowing selection on all non-synonymous sites).

Following WU *et al.* (2011), we generated phenotypes by allowing 5% of the sampled variants under 3% frequency to be causal (see Methods). We ran SKAT-O on the phenotypes and genotypes from the African and European continental groups and computed the fraction of simulations with p -values under 10^{-6} . Selection impacts power by increasing the fraction of sites at low frequency and decreasing the overall level of genetic variation. The net result is a loss of power. When the entire locus is under selection, genetic variation is pushed to even lower levels and the site frequency spectrum is further shifted to rare alleles, which substantially reduces power. Note that we have not forced the genetic variance explained by the test sequence to be the same between the neutral

and selected models, but this is precisely the point of this experiment. Because of reduced levels of overall variation, regions under direct selection must have larger effect sizes in order to explain an equal proportion of the genetic variance as neutral regions and reach the same power as neutral sequences.

In general, we observe higher power in the African continental group for this phenotype model (Figure 4, with $> 80\%$ power for a sample size of 5×10^3), which is in line with the overall higher level of genetic diversity within Africans (Figure 5). We also observe a subtle difference in power between the demographic models of TENNESSEN *et al.* (2012) (which includes rapid growth of the African population) and GUTENKUNST *et al.* (2009) (which includes only an ancient expansion in the African population) in the African population.

Figure 4 shows that the effect of selection on power for this particular phenotype model ranges from $\approx 5 - 50\%$ depending on the sample size and demographic model. However, we emphasize that these results should not be interpreted as general effects of human selection and demography on statistical power, but rather a reflection of a specific phenotype model commonly used in the literature (WU *et al.*, 2011). In general, the effects of recent selection and demography on power may be more or less severe depending on sample size and the relationship between effect sizes and allele frequency, which is still a matter of some debate. One of the main advantages of our simulation method over other methods for estimating power is that the selection strength operating on every variant in the simulation is known. In two of the phenotype models included in `sfs_coder`, we use selection coefficients rather than allele frequencies to draw effect sizes when simulating phenotypes. In the Discussion section we further deliberate on the simulation of phenotypes. Here we have examined direct selection, but linked selection may also play a role in altering patterns of variation and affecting power calculations, as we address in the next section.

3.3.4 The impact of linked selection

Patterns of genetic diversity at neutral sites can be altered via physical linkage to sites under selection (SMITH and HAIGH, 1974; CHARLESWORTH *et al.*, 1993). The effects of linked negative selection (known as “background selection”) on variation in humans have been well studied (MCVICKER *et al.*, 2009). Sites that are closely linked to deleterious variants are also prohibited from increasing in frequency, resulting in a local decrease in genetic diversity. The mean number of

variants segregating at such loci is lower than in regions that are unlinked to selected loci. Background selection also has an effect on the shape of the site frequency spectrum (i.e., the proportion of variant sites at a given frequency) (ZENG and CHARLESWORTH, 2011; NICOLAISEN and DESAI, 2013), but this effect is often subtle.

We ran 250 simulations incorporating the demographic model of GUTENKUNST *et al.* (2009) for a 2Mb region that is centered on the 30 kb region considered in Figure 4 (chr3:50320000-50350000). This locus is expected to have among the strongest effects of background selection in the human genome under the model of MCVICKER *et al.* (2009). All inferred exons and conserved noncoding elements within this region were simulated. In Figure 5, we show the mean observed nucleotide diversity π for these simulations as compared to simulations we used for Figure 4, which did not include the flanking sequences (but did incorporate selection on the 30 Kb central region). The overall level of diversity is strongly reduced in the background selection simulations. Additionally, there is a very small shift in the site frequency spectrum towards rare variants in the background selection simulations (data not shown).

We used these background selection simulations to compute power of SKAT-O with the same phenotype model that we considered in Figure 4. For a sample size of 2,000 individuals, power in Europeans drops from 44% for simulations without background selection to 38% when background selection is included, and power in Africans drops from 67% to 57%, due to the overall reduction in genetic diversity. Note that this reduction in overall diversity also implies a reduction in variance explained by the test sequence under the phenotype model of WU *et al.* (2011), as discussed in the previous section.

3.4 Discussion

Simulations play an increasingly prominent role in statistical and population genetics because they can be used to generate DNA sequence data under models that are too complex to handle analytically. In statistical genetics, simulations have been used to generate sequence data under various demographic scenarios and to assess the performance of various statistical tests of association in the presence of complex demographics (e.g., see WU *et al.* (2011)). However, natural selection also impacts patterns of genetic variation and thus might influence conclusions about statistical power.

Here, we introduced a new simulation tool (`sfs_coder`) that allows users to model human

demography, selection, recombination, and genomic elements (exons and conserved noncoding sequences) for any locus in the human genome. Both selection and demography alter the frequency spectrum of variant sites and the number of variant sites within a sample of sequenced chromosomes. Jointly modeling these evolutionary forces may prove to be an important step forward for the use of simulations in assessing the performance of tests of statistical association (KING *et al.*, 2010), and indeed we showed here that natural selection and demography can impact the power of rare variant association tests under some phenotype models.

As an alternative to forward simulations of DNA sequences, researchers have also proposed “sideways” simulations (CHEN *et al.*, 2014). These simulations use a reference panel of sequenced (or genotyped) chromosomes to generate new chromosomes under the model of LI and STEPHENS (2003). The advantage of these methods is that they rely on observed genetic data, so the impact of natural selection and demography on the genealogy and the sampled genetic diversity are present in the reference panel and do not need to be directly modeled. However, here we showed that these methods do not recapitulate expected patterns of variation for rare variants when the population has experienced recent growth and the size of the reference panel is much smaller than the desired sample size.

While it is straightforward to use forward simulation tools such as `sfs_coder` to model the effects of selection and demography on DNA sequences, an important caveat of forward simulation is model misspecification. Here, we have tapped into the deep population genetics literature to incorporate recently inferred models, but these models should not be considered absolute truth. In particular, in future studies with larger sample sizes, considerable refinement could be made in the estimates of recent human growth as well as the strength of selective constraint. For example, there remains some debate about the rate of recent human growth and its influence on DNA sequences. One study of neutral genomic regions did not find support for recent acceleration in the growth rate (GAZAVE *et al.*, 2014), in contrast to TENNESSEN *et al.* (2012). Furthermore, more diverse annotations of genomic elements than we have considered here may allow for more precise distributions of selection coefficients for each locus in the genome.

Another important consideration when performing forward simulations is the choice of simulation parameters, especially sequence length and population size. Larger sequence lengths and population sizes can dramatically increase the computation time of forward simulations, so it is

advantageous to pick these variables to be as small as possible (HOGGART *et al.*, 2007). However, we showed here that ignoring the impact of linked negative selection can alter patterns of simulated diversity and potentially affect power calculations. Furthermore, for some evolutionary models the simulated population size can also bias simulated patterns of genetic diversity if it is chosen to be too small (URICCHIO and HERNANDEZ, 2014). Some further work is needed to explore appropriate choices of sequence length and population size for forward simulations of background selection and demography, and we have left these choices up to the end-user in our software `sfs_coder`.

Applications of simulations abound in the genetics literature, but one use that is of particular interest to statistical geneticists is the estimation of statistical power. For rare variant association tests (or any test of association that pools putatively causal variants), power is a function of the joint distribution of allele frequencies and effect sizes. Here, we showed that power is higher in Africans as compared to non-Africans for a simple phenotype model that asserts a logarithmic increase in effect size as allele frequency decreases. A side-effect of this phenotype model is that populations with more genetic variation have a greater proportion of variance explained by the test sequence. For a real phenotype, this may or may not hold, since it is not necessarily true that 1) effect sizes will be the same in two different populations or 2) the environmental variance is the same in two different populations. Exactly how the joint distribution of effect sizes and allele frequencies differ between populations may depend intimately on the action of selection on causal sites in each population and recent demographic history.

Generally, the distribution of effect sizes is not known, so arbitrary distributions have been proposed in previous work (and were applied here). These distributions assign larger effect sizes to rare variants than common causal alleles, but they do not necessarily have an obvious interpretation in terms of the strength and rate of selection in the human genome. Rare variants are not likely to contribute substantially to the variance observed in complex phenotypes unless causal sites are under selection (SIMONS *et al.*, 2014), so the interpretability of power studies could be improved substantially by performing assessments with genetic models that include selection. In future studies it will be advantageous to simulate phenotypes where the effect sizes are modeled directly based on the selection coefficients of the causal sites, reminiscent of the work of EYRE-WALKER (2010). Tools such as `sfs_coder` should enable this work as human geneticists push further into the age of deep sequencing, large sample sizes, and complex genetic architectures.

3.5 Acknowledgments

This work was partially supported by the National Institutes of Health (grants 1R01HG007644 to R.D.H and 1R01CA088164 to J.S.W.) and a Sloan Foundation Research Fellowship (to R.D.H.). L.H.U. was supported by an Achievement Rewards for College Scientists fellowship and R.T. was supported by National Science Foundation Graduate Research fellowship 1144247. We thank M. Cyrus Maher for providing a useful script.

3.6 Appendix

We ran used the default parameter settings of the model “bestfit”, included in the *cosi* distribution, to simulate the model of SCHAFFNER *et al.* (2005).

The following are representative command lines for our SFS_CODE simulations.

```
GUTENKUNST et al. (2009):  sfs_code 3 10 -N 7000 -n 50 50 0 -A -L 1 100 -t 0.001
-r 0.001 -TS 0.219178 0 1 -TS 0.544658 1 2 -TE 0.60274 -Td 0 P 0 1.68493 -Td
0.219178 P 1 0.170732 -Td 0.544658 P 1 0.47619 -Tg 0.544658 P 1 58.4 -Td 0.544658
P 2 0.242857 -Tg 0.544658 P 2 80.3 -Tm 0.219178 P 0 1 6.15 -Tm 0.219178 P 1 0 0.5
-Tm 0.544658 L 0.738 0.4674 0.06 0.192 0.01938 0.09792
```

```
GRAVEL et al. (2011): sfs_code 3 10 -N 7000 -n 50 50 0 -A -L 1 100 -t 0.001 -r 0.001
-TS 0.265753 0 1 -TS 0.342466 1 2 -TE 0.405479 -Td 0 P 0 1.982738 -Td 0.265753
P 1 0.128575 -Td 0.342466 P 1 0.554541 -Tg 0.342466 P 1 55.48 -Td 0.342466 P
2 0.29554 -Tg 0.342466 P 2 70.08 -Tm 0.265753 P 0 1 4.3422 -Tm 0.265753 P 1 0
0.5583 -Tm 0.342466 L 0.7237 0.225794 0.09305 0.115754 0.00858 0.03421
```

```
TENNESSEN et al. (2012):  sfs_code 3 10 -N 7000 -n 50 50 0 -A -L 1 100 -t 0.001
-r 0.001 -TS 0.265753 0 1 -TS 0.342466 1 2 -TE 0.405479 -Td 0 P 0 1.982738
-Td 0.265753 P 1 0.128575 -Td 0.342466 P 1 0.554541 -Tg 0.342466 P 1 44.822
-Td 0.342466 P 2 0.29554 -Tg 0.342466 P 2 70.08 -Tm 0.265753 P 0 1 4.3422 -Tm
0.265753 P 1 0 0.5583 -Tm 0.342466 L 0.7237 0.225794 0.09305 0.115754 0.00858
0.03421 -Tg 0.391465 P 0 242.36 -Tg 0.391465 P 1 284.7
```

We rescaled the ancestral population size to $N = 5 \times 10^3$ (as opposed to 7×10^3 above) for computational feasibility for our simulations. We applied the distribution of selection coefficients of BOYKO

et al. (2008) to coding loci, which is given by the flag `-W 2 0 0 0 0.184 0.00040244`. For conserved noncoding elements, we applied the distribution of TORGERSON *et al.* (2009), which is given by `-W 2 0 0 0 0.0415 0.0015625`. See the SFS_CODE manual at sfscode.sourceforge.net for more information.

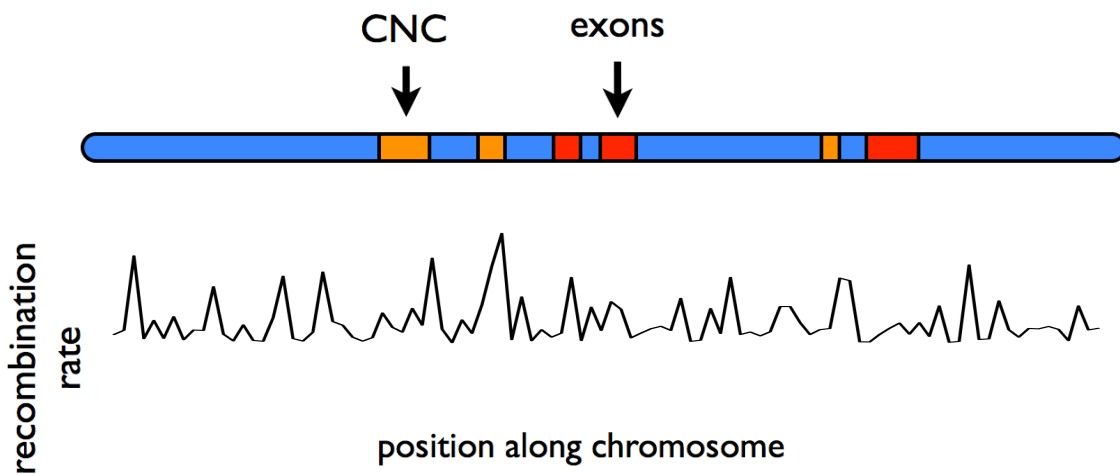


Figure 3.1: A model of human genomic sequences that incorporates selection on exons and conserved noncoding elements (with separate distributions of selection coefficients) and local recombination rates. Selection coefficients on exons and conserved noncoding elements were taken from BOYKO *et al.* (2008) and TORGERSON *et al.* (2009), respectively. The positions of the conserved noncoding elements were inferred by SIEPEL *et al.* (2005), and the recombination map was inferred by INTERNATIONAL HAPMAP CONSORTIUM *et al.* (2007). (CNC: conserved noncoding).

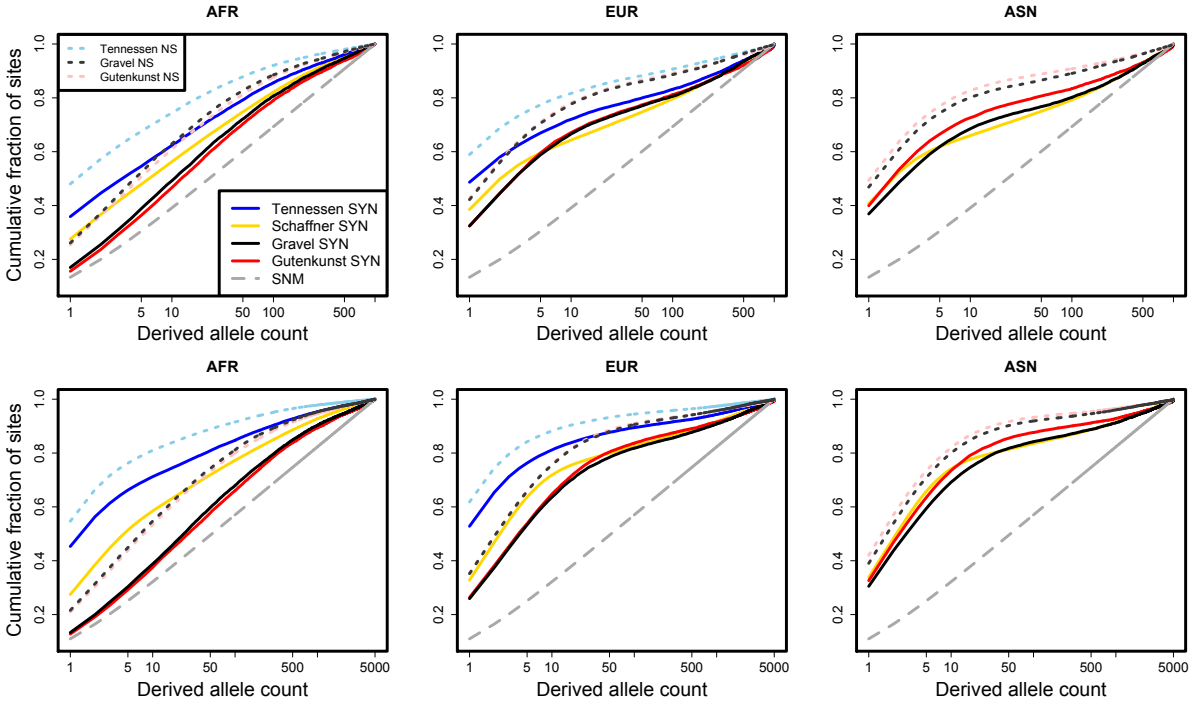


Figure 3.2: Simulated cumulative site frequency spectra in three human continental groups under several recently inferred demographic models. Sample size is 10^3 chromosomes in the top panels and 5×10^3 chromosomes in the lower panels. Note that the model of TENNESSEN *et al.* (2012) did not infer demography of the Asian continental group, so we do not plot a curve for this population. However, the Asian continental group is included in the TENNESSEN *et al.* (2012) simulations. Each curve was calculated using 10^3 independent simulations of 5×10^3 base pairs each. The gray dashed curve represents the analytical expectation based on the standard neutral model and is not the result of simulations. (SNM: standard neutral model, AFR: African population, EUR: European population, ASN: Asian population, NS: non-synonymous, SYN: synonymous).

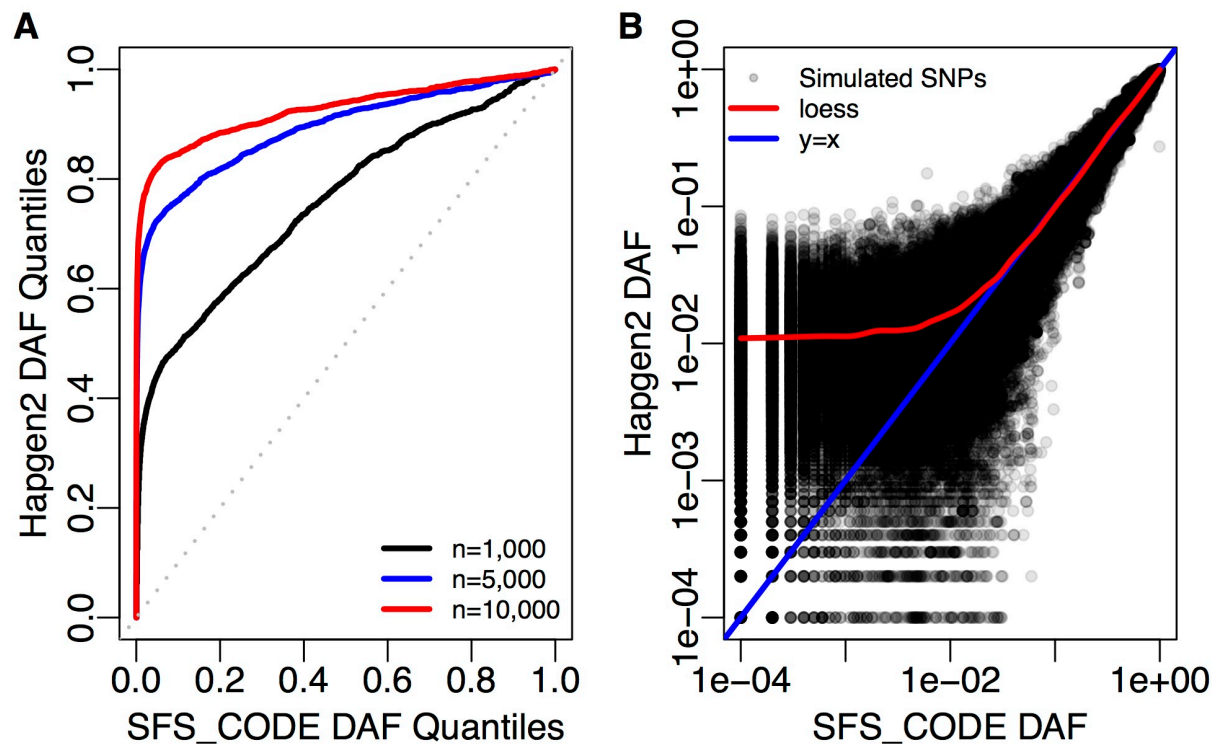


Figure 3.3: Generating large samples by haplotype resampling results in a deficiency of rare variants in the presence of population growth. For this plot we simulated 10^4 chromosomes from thirty unlinked 10Mb regions of the human genome (300Mb total) in SFS_CODE under the TENNESSEN *et al.* (2012) demographic model discussed in the text. We then used 100 chromosomes as a reference panel in Hapgen2 to generate a much larger sample (as indicated in the legend). (A) shows a quantile-quantile plot of the derived allele frequencies (DAF) for chromosomes simulated using Hapgen2 versus the expected DAF distribution from SFS_CODE (with sample sizes indicated in the legend). (B) shows a scatter plot of the frequencies of each SNP inferred by Hapgen2 compared to SFS_CODE for a sample of size 10^4 chromosomes. Both figures demonstrate that Hapgen2 fails to recapitulate the extent of rare variation expected under rapid population growth, particularly for large sample sizes. *Figure credit: Ryan Hernandez*

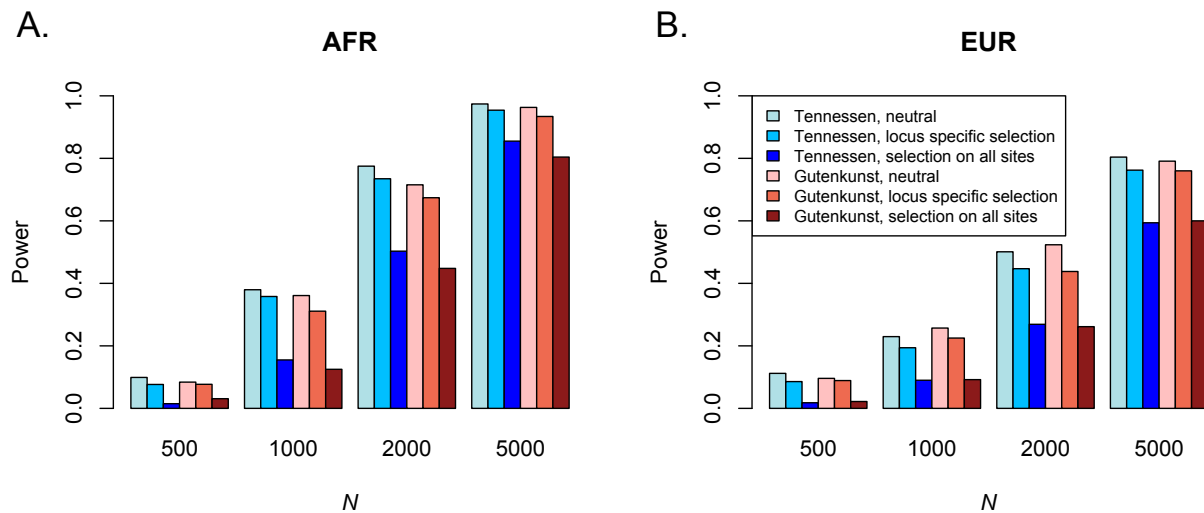


Figure 3.4: Power of SKAT-O at the $\alpha = 10^{-6}$ level for two different models of demography with and without selection, for different sample sizes N . We applied both a locus specific model of selection (and a model that treats the entire 30kb locus as a single gene. All results are for a region on chromosome 3, hg19 coordinates 50320000-50350000.

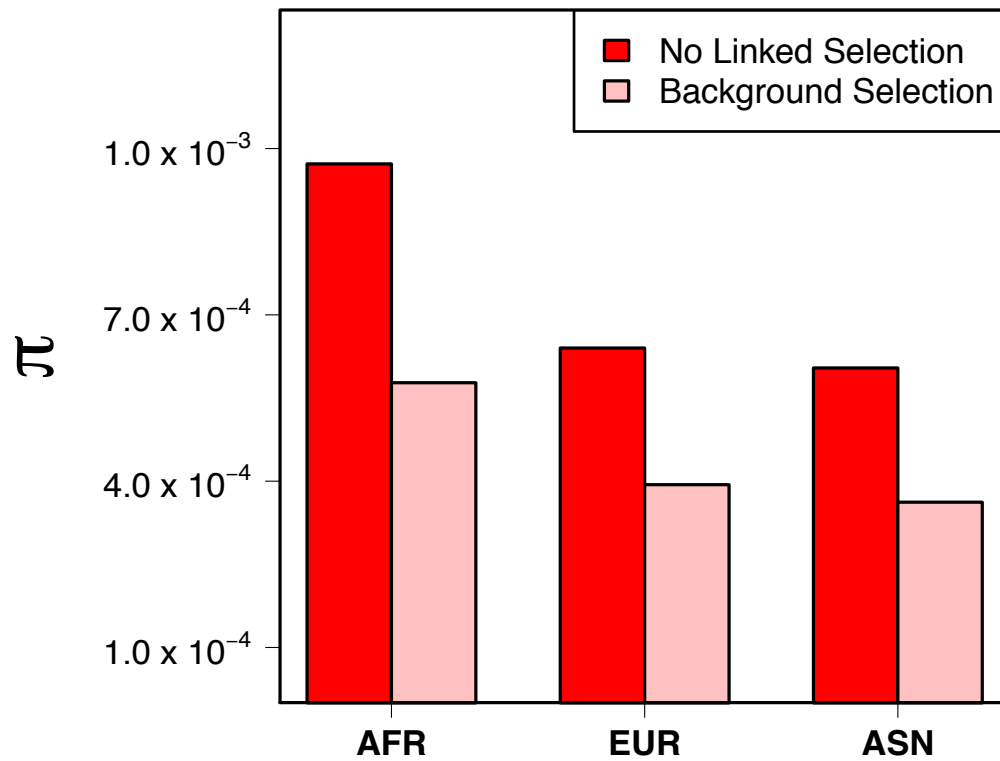


Figure 3.5: We calculated the pairwise nucleotide diversity π for for the locus at chr3:50320000-50350000, for simulations with and without background selection and the demographic model of GUTENKUNST *et al.* (2009). The simulations without background selection simulated only this locus, while the background selection simulations also included 1 Mb of flanking sequence on either side of the locus of interest. Note, sites in the flanking 1Mb are not included in the diversity calculations. *Figure credit: Raul Torres*

References

- 1000 GENOMES PROJECT CONSORTIUM, G. R. ABECASIS, A. AUTON, L. D. BROOKS, M. A. DEPRISTO, *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.
- BOYKO, A. R., S. H. WILLIAMSON, A. R. INDAP, J. D. DEGENHARDT, R. D. HERNANDEZ, *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics* **4**: e1000083.
- BUSTAMANTE, C. D., J. WAKELEY, S. SAWYER, and D. L. HARTL, 2001 Directional selection and the site-frequency spectrum. *Genetics* **159**: 1779–1788.
- CHARLESWORTH, B., M. T. MORGAN, and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.
- CHEN, H.-S., C. M. HUTTER, L. E. MECHANIC, C. I. AMOS, V. BAFNA, *et al.*, 2014 Genetic simulation tools for post-genome wide association studies of complex diseases. *Genetic Epidemiology* **Submitted**.
- COHEN, J. C., R. S. KISS, A. PERTSEMLIDIS, Y. L. MARCEL, R. MCPHERSON, *et al.*, 2004 Multiple rare alleles contribute to low plasma levels of hdl cholesterol. *Science (New York, N.Y.)* **305**: 869–872.
- EMERSON, B. C., E. PARADIS, and C. THEBAUD, 2001 Revealing the demographic histories of species using DNA sequences. *TREE* **16**: 707–716.
- EYRE-WALKER, A., 2010 Evolution in health and medicine sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences of the United States of America* **107 Suppl 1**: 1752–1756.
- GAZAVE, E., L. MA, D. CHANG, A. COVENTRY, F. GAO, *et al.*, 2014 Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences of the United States of America* **111**: 757–762.
- GRAVEL, S., B. M. HENN, R. N. GUTENKUNST, A. R. INDAP, G. T. MARTH, *et al.*, 2011 Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 11983–11988.
- GUTENKUNST, R. N., R. D. HERNANDEZ, S. H. WILLIAMSON, and C. D. BUSTAMANTE, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics* **5**: e1000695.
- HARROW, J., A. FRANKISH, J. M. GONZALEZ, E. TAPANARI, M. DIEKHANS, *et al.*, 2012 Gencode: the reference human genome annotation for the encode project. *Genome research* **22**: 1760–1774.
- HERNANDEZ, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. *Bioinformatics (Oxford, England)* **24**: 2786–2787.
- HOFFMANN, T. J., N. J. MARINI, and J. S. WITTE, 2010 Comprehensive approach to analyzing rare genetic variants. *PloS one* **5**: e13584.

- HOGGART, C. J., M. CHADEAU-HYAM, T. G. CLARK, R. LAMPARIELLO, J. C. WHITTAKER, *et al.*, 2007 Sequence-level population simulations over large genomic regions. *Genetics* **177**: 1725–1731.
- INTERNATIONAL HAPMAP CONSORTIUM, K. A. FRAZER, D. G. BALLINGER, D. R. COX, D. A. HINDS, *et al.*, 2007 A second generation human haplotype map of over 3.1 million snps. *Nature* **449**: 851–861.
- KATZMAN, S., A. D. KERN, G. BEJERANO, G. FEWELL, L. FULTON, *et al.*, 2007 Human genome ultraconserved elements are ultraselected. *Science (New York, N.Y.)* **317**: 915.
- KEINAN, A., and A. G. CLARK, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science (New York, N.Y.)* **336**: 740–743.
- KING, C. R., P. J. RATHOUZ, and D. L. NICOLAE, 2010 An evolutionary framework for association testing in resequencing studies. *PLoS genetics* **6**: e1001202.
- LEE, S., M. J. EMOND, M. J. BAMSHAD, K. C. BARNES, M. J. RIEDER, *et al.*, 2012 Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American journal of human genetics* **91**: 224–237.
- LI, B., and S. M. LEAL, 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics* **83**: 311–321.
- LI, N., and M. STEPHENS, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**: 2213–2233.
- LOHMUELLER, K. E., 2014 The impact of population demography and selection on the genetic architecture of complex traits. *PLoS genetics* **10**: e1004379.
- MAHER, M. C., L. H. URICCHIO, D. G. TORGERSON, and R. D. HERNANDEZ, 2012 Population genetics of rare variants and complex diseases. *Human heredity* **74**: 118–128.
- MANOLIO, T. A., F. S. COLLINS, N. J. COX, D. B. GOLDSTEIN, L. A. HINDORFF, *et al.*, 2009 Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.
- MCVICKER, G., D. GORDON, C. DAVIS, and P. GREEN, 2009 Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**: e1000471.
- MORGENTHALER, S., and W. G. THILLY, 2007 A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutation research* **615**: 28–56.
- NEALE, B. M., M. A. RIVAS, B. F. VOIGHT, D. ALTSHULER, B. DEVLIN, *et al.*, 2011 Testing for an unusual distribution of rare variants. *PLoS genetics* **7**: e1001322.
- NICOLAISEN, L. E., and M. M. DESAI, 2013 Distortions in genealogies due to purifying selection and recombination. *Genetics* **195**: 221–230.
- NIELSEN, R., 2005 Molecular signatures of natural selection. *Annual review of genetics* **39**: 197–218.

- PENG, B., H. CHEN, L. MECHANIC, B. RACINE, J. CLARKE, *et al.*, 2014 Genetic data simulators and their applications: an overview. *Genetic Epidemiology* .
- PRICE, A. L., G. V. KRYUKOV, P. I. W. DE BAKKER, S. M. PURCELL, J. STAPLES, *et al.*, 2010 Pooled association tests for rare variants in exon-resequencing studies. *American journal of human genetics* **86**: 832–838.
- PRITCHARD, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics* **69**: 124–137.
- SAWYER, S. A., and D. L. HARTL, 1992 Population genetics of polymorphism and divergence. *Genetics* **132**: 1161–1176.
- SCHAFFNER, S. F., C. FOO, S. GABRIEL, D. REICH, M. J. DALY, *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. *Genome research* **15**: 1576–1583.
- SIEPEL, A., G. BEJERANO, J. S. PEDERSEN, A. S. HINRICHS, M. HOU, *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**: 1034–1050.
- SIMONS, Y. B., M. C. TURCHIN, J. K. PRITCHARD, and G. SELLA, 2014 The deleterious mutation load is insensitive to recent population history. *Nature genetics* **46**: 220–224.
- SMITH, J. M., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. *Genetical research* **23**: 23–35.
- SU, Z., J. MARCHINI, and P. DONNELLY, 2011 Hapgen2: simulation of multiple disease snps. *Bioinformatics (Oxford, England)* **27**: 2304–2305.
- TENNESSEN, J. A., A. W. BIGHAM, T. D. O’CONNOR, W. FU, E. E. KENNY, *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)* **337**: 64–69.
- THORNTON, K. R., A. J. FORAN, and A. D. LONG, 2013 Properties and modeling of gwas when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. *PLoS genetics* **9**: e1003258.
- TORGERSON, D. G., A. R. BOYKO, R. D. HERNANDEZ, A. INDAP, X. HU, *et al.*, 2009 Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS genetics* **5**: e1000592.
- URICCHIO, L. H., and R. D. HERNANDEZ, 2014 Robust forward simulations of recurrent hitchhiking. *Genetics* **197**: 221–236.
- WILLIAMSON, S. H., R. HERNANDEZ, A. FLEDEL-ALON, L. ZHU, R. NIELSEN, *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 7882–7887.
- WITTE, J. S., 2010 Genome-wide association studies and beyond. *Annual review of public health* **31**: 9–20 4 p following 20.

- WU, M. C., S. LEE, T. CAI, Y. LI, M. BOEHNKE, *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* **89**: 82–93.
- ZENG, K., and B. CHARLESWORTH, 2011 The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics* **189**: 251–266.
- ZUK, O., S. F. SCHAFFNER, K. SAMOCHA, R. DO, E. HECHTER, *et al.*, 2014 Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* **111**: E455–E464.

- 4 Selection and explosive growth may hamper the power of rare variant association tests

4.1 Introduction

Much recent debate among geneticists has focused on the role of rare variants in complex traits (PRITCHARD 2001). While there is evidence that rare variants can in some cases contribute to common diseases of interest (HALLER *et al.* 2009; TORGERSON *et al.* 2012), it has not been established whether they explain a large proportion of the genetic variance for any traits of interest. Indeed, a number of recent papers have suggested that most of the variance is attributable to common variants of weak effect, at least for some phenotypes (YANG *et al.* 2010; GAUGLER *et al.* 2014). Nonetheless, several large-scale sequencing studies of cases and controls or quantitative phenotypes are underway, and these studies will uncover a large amount of novel rare variation within their samples. Most of this variation is likely to be unrelated to the phenotype of interest, so it is imperative that population and statistical geneticists continue to make progress on methods to interpret this deluge of data and provide biological insights from noisy data (MAHER *et al.* 2012).

Classically, geneticists have relied on single-marker tests of association to discover causal sites, where the contribution to the phenotype of each genotyped allele is assessed with a likelihood ratio test. The power to detect an association is a function of allele frequency, sample size, and effect size, so (ignoring complications such as linkage and cryptic population structure) it is straightforward in this framework to compute power by rastering over a grid of effect sizes and allele frequencies. Recent work has shown that the power of single-marker tests is sensitive to the relationship between selection strength and effect sizes, as well as population demography (LOHMUELLER, 2014).

Unfortunately, power to detect rare causal variants with single-marker tests of association is very low, even in large samples. For this reason, several statistical tests that pool rare variants within a putatively causal locus and jointly test for their role in the phenotype have been proposed and published (HOFFMANN *et al.* 2010; WU *et al.* 2011; LEE *et al.* 2012a,b). When pooling variants of differing effects, it is no longer a simple matter to consider all possible effect sizes and allele frequency combinations, because the state space of possible joint distributions is very large. Since this distribution is unknown, studies that assess the power of rare variant association tests have proposed arbitrary joint distributions of effects and allele frequencies. It is not obvious that these proposed distributions are biologically or evolutionarily plausible, or that power of rare variant tests is insensitive to assumptions made about this joint distribution.

It has long been appreciated by geneticists that rare variants can only contribute substantially to genetic variance when they have dramatically larger effect sizes than common variants. The most plausible explanation for an inverse relationship between allele frequency and effect size is natural selection on trait altering alleles (EYRE-WALKER 2010; SIMONS *et al.* 2014), but most rare variant association tests have been assessed using simulations that do not include the action of natural selection (URICCHIO *et al.* 2015). It is well known that both natural selection and demography impact the frequency spectrum of alleles, and hence these forces may also have some impact on the joint distribution of effects and frequencies.

Very little is known about effect sizes for very low frequency alleles, but several recent studies have examined the distribution of selection coefficients in human genomic elements such as exons and conserved non-coding elements (BOYKO *et al.* 2008; TORGERSON *et al.* 2009; MCVICKER *et al.* 2009). Selection-based models assert that every conserved base in the genome has some (potentially very small) impact on reproductive fitness, and mutations at such bases are likely to deleteriously impact the organism. While it is not possible to accurately estimate the strength of selection acting on individual rare sites, it is possible to infer the distribution of selection coefficients using the distribution of observed allele frequencies across all segregating mutations. Intuitively, the reason that some genomic sites are conserved is that they contribute to important functions, such as maintaining protein structure or activity. If there is a strong evolutionary pressure to constrain values of some phenotype to fall within a narrow range, then selection will constrain mutations with large effects on the phenotype to low frequency. Hence, there is a natural relationship between selection coefficients and effect sizes for phenotypes under direct or pleiotropic selection (by which we mean selection on another phenotype, which has some common genetic basis with the phenotype of interest). In particular, recent models relating selection strength and effect size have asserted that there is a monotonic increasing relationship between selection strength and the mean absolute value of effect size (EYRE-WALKER, 2010; SIMONS *et al.*, 2014). Combining models that relate selection strength to effect sizes with recently inferred human-specific distributions of selection coefficients, it is possible to use simulations to learn about implications of demography and selection for genetic architecture and the power of statistical tests (LOHMUELLER, 2014).

Here, we examine the power of one the most popular (and elegant) rare variant association tests (the “sequence kernel association test”, SKAT-O, which subsumes many other rare variant

association tests as special cases; WU *et al.* 2011; LEE *et al.* 2012b,a) in the context of population genetic simulations of simultaneous selection and demography. We develop a model of complex traits that explicitly captures the relationship between effect sizes and selection strength and simulate genotypes and phenotypes under the model. Following earlier studies, we show that only under very strong assumptions about the relationship between effect sizes and selection coefficients do rare variants contribute substantially to the genetic variance. We extend these results by considering the impact of various model parameters in the context of two different demographic models of human history, one of which includes recent accelerated growth, and show that most of the contribution to the genetic variance from rare variants is due to extremely rare variants for a broad range of parameters. We show that our selection-based model of complex phenotypes can result in dramatically different power calculations for SKAT-O, and that power is sensitive to assumptions about recent demographic history – particularly the rate of population growth in the past several thousand years. Moreover, we show that the power of SKAT-O (using its default settings) is inversely proportional to the genetic variance explained by rare variants under our phenotype model, even when the total genetic variance of the test sequence is fixed. This means that the test has the worst power when the rare variants make the greatest contribution to genetic variance. We show that the most obvious strategy to increase power within the SKAT-O framework – adjusting the parameters of SKAT-O’s weight distribution – increases power, but also dramatically increases false positive rates. These results suggest that the power of rare variant association tests may have been overestimated in previous studies, and that more work may need to be done to develop methods for rare variant association tests (or adapt current methods) to be well powered under realistic assumptions about demographic history and recent selection.

4.2 Materials & Methods

4.2.1 An evolutionary model of complex phenotypes

We develop a phenotype model that explicitly captures the relationship between selection strength and effect size and is a variant of models proposed by EYRE-WALKER (2010) and SIMONS *et al.* (2014). Here, we will briefly describe these two models and motivate the modifications we have made.

The model of EYRE-WALKER (2010) computes effect sizes z with $z = \delta s^\tau(1 + \epsilon)$. δ is -1 or 1 with equal probability, thereby allowing for both trait-increasing and trait-decreasing mutations. τ is an exponent that transforms the distribution of selection coefficients to allow phenotypic effects to grow faster ($\tau > 1$) or slower ($\tau < 1$) than selection coefficients. The central idea is that effect sizes may not have the same marginal distribution as selection coefficients, but sites with larger effects on fitness will also have larger effects on the phenotype. ϵ is a random normal variable with mean zero and variance σ^2 . As the variance of ϵ grows, the correlation between selection coefficient and effect size decreases.

In our study, we are concerned with how the joint distribution of effect sizes and allele frequencies impacts statistical power for discovering causal loci. It is plausible that power will depend on both the marginal distribution of effects and the relationship between effects and allele frequencies, so it is desirable to have a mechanism to hold the marginal distribution of effects constant in order to focus on the relationship between allele frequency and effects. SIMONS *et al.* (2014) proposed a model with two selection coefficients, one strong and one weak, that has this property. With probability ρ , a mutation has effect size proportional to its selection coefficient, and with probability $1 - \rho$, it has an effect size randomly sampled from the marginal distribution of selection coefficients (and then scaled by a proportionality constant). Here, we extend this model by 1) including a Γ -distribution of selection coefficients that was inferred for human coding regions (BOYKO *et al.* 2008) and 2) including both the τ and δ parameters from the model of EYRE-WALKER (2010).

Thus, our model for effect sizes z_s for a site with selection coefficient s can be summarized as:

$$z_s = \begin{cases} \delta |s|^\tau & \text{if } p \leq \rho \\ \delta |s_r|^\tau & \text{if } p > \rho \end{cases} \quad (4.1)$$

where p is a uniform random number and s_r is a random sample from the marginal distribution of selection coefficients. Since we do not have an analytical expectation for the distribution of sampled selection coefficients for the complicated demographic models that we simulate here, we use the sampled variants in any given simulation to provide a distribution on s . When $\tau = 1$, ρ is also the Pearson correlation between the selection coefficient and the effect size, but this is not the case when $\tau \neq 1$. However, the interpretation that a high value of ρ corresponds to a tight correlation between effects and a low value corresponds to a weak correlation holds across all values

of τ . Note, we only take non-synonymous sites as causal; synonymous sites always have 0 effect in our simulations.

When $\rho = 1$, we obtain exactly the model of EYRE-WALKER (2010). We do not include the σ parameter of the original model, but this parameter was shown to have no impact on genetic architecture (EYRE-WALKER, 2010). When $\tau = 1$, we obtain the model of SIMONS *et al.* (2014), but with a distribution of selection coefficients inferred from human genomic data, and the additional possibility for causal sites to be either trait increasing or decreasing.

From an evolutionary perspective, this model captures the idea that phenotypes under direct selection will have a tight correlation between selection strength and effect size (i.e., high ρ), but the marginal distribution of effects may grow faster or slower than the distribution of selection coefficients (i.e., τ can be a value greater than or less than 1). Due to pleiotropic effects, some selection coefficients may be large while their effect on the phenotype is small (i.e., decreasing ρ allows increased emphasis on pleiotropic effects).

We also simulate phenotypes under the model of WU *et al.* (2011), which was used in the original SKAT study to test the power of their method. In this model, an allele with frequency $x < 0.03$ has effect size $z(x) \propto \log_{10}(x)$ with probability 0.05, and otherwise has an effect size of zero. Here, we take all non-synonymous sites under 3% frequency to be causal, such that the total number of causal sites are roughly comparable between simulations under our model and simulations using the effect size distribution of WU *et al.* (2011). Note that our loci are ≈ 18 times shorter than the loci simulated by WU *et al.* (2011), so by taking all the variants as causal we will have close to the same expected number of causal variants within the locus as in their study, but we have far fewer non-associated variants. In this sense, our simulations represent a “best-case” scenario, since a very high proportion of the sites within each causal locus are causal for the trait, though some sites have very small effect sizes.

The statistical power of association tests is a function of the fraction of the variance in the phenotype that is explained by the test sequence. For this reason, we always fix the total contribution of the test loci at a pre-specified amount (and hence any observed differences in power cannot be explained by systematic differences in the contribution of genetics to the phenotype). In some analyses, we simulate phenotypes by allowing only a single gene to be causal for the phenotype, and fix the proportion of the variance explained by this gene at 5%. We also perform analyses

using a polygenic model where genetic variation in the trait is driven by 20 genes, and fix the total contribution of genetics to the phenotype at 50% (i.e., $h^2 = 0.5$). Thus, each simulated gene will have a different contribution to the phenotype, but we calculate power as a function of variance explained.

4.2.2 Calculating the impact of demographic events on genetic architecture

We investigated the impact of selection and demography on the site frequency spectrum, as well as the genetic architecture of complex traits, using exact numerical calculations under the Wright-Fisher model and stochastic forward simulations.

For our numerical calculations, we consider a model consisting of discrete and exponential population size changes. While our software is generalizable to other demographic models of interest, here we focused on the marginal European demographic history of the model of GUTENKUNST *et al.* (2009). This model includes population size changes of magnitudes $\nu = [1.685, 0.170732, 0.47619]$ at times $t = [0, 0.219178, 0.544658]$ (times are in coalescent units). These events correspond to an expansion in the African ancestral population, a bottleneck event as Europeans moved out of Europe, and a second later bottleneck. Immediately after the last bottleneck event, the population grows exponentially at rate 58.4 (scaled in coalescent units). For further details on the model parameters, see URICCHIO *et al.* (2015).

To calculate the site frequency spectrum as a function of time after demographic events, we propagated Wright-Fisher transition matrices forward in time. The transition probability for a site present in k copies in a population of size $2N$ to k^* in the next generation with selection coefficient s is given as

$$\text{Binomial} \left(k^*; 2N, \frac{(1+s)x}{1+sx} \right) \quad (4.2)$$

where $x = \frac{k}{2N}$, the allele frequency of the site. Discrete changes in population size change the state space on k , and hence the rate at which drift happens in each subsequent generation, as well as the equilibrium proportion of variable sites present at any given frequency. The code we developed is implemented in Python and is freely available upon request.

We performed these calculations for two selection coefficients, $s = -0.01$ and $s = -0.0002$, each with identical underlying mutation rates, exactly as in the selection-based phenotype model

proposed in SIMONS *et al.* (2014). We assumed a human ancestral population size of 7.3×10^3 , as was inferred in GUTENKUNST *et al.* (2009), such that the $\gamma = 2Ns = -146$ for the large selection coefficient and $\gamma = -2.92$ for the small selection coefficient.

We used our code to calculate the proportion of variable sites that are present in a single copy (singletons) in a sample of 100 chromosomes for each of the selection coefficients in this model, which we denote as Ψ . We also calculated the genetic variance due to singleton sites as a function of ρ in our phenotype model, assuming that $\tau = 1$ (we later relax this assumption).

We also performed simulations under this model of selection and demography, and sampled variants at time points from $t = 0$ to $t = 1$ (in coalescent units). We performed 100 simulations per time point. Scripts for the simulations, which were performed using `sfs_coder`, are available upon request of the authors.

4.2.3 Three-population forward simulations of human selection and demography

We used `sfs_coder` to perform forward simulations of human selection and demography (URICCHIO *et al.* 2015). `sfs_coder` is a python based front-end to the forward simulation software SFS_CODE (HERNANDEZ 2008) that includes several models of human demography and selection. The demographic models we simulated are those of GUTENKUNST *et al.* (2009) and TENNESSEN *et al.* (2012). Briefly, the model of GUTENKUNST *et al.* (2009) includes three populations, namely the African, European, and Asian continental groups. The European and Asian populations are formed by a series of bottlenecks as the human population moved out of Africa, and the model also includes recent exponential growth in the European and Asian continental groups. Migration between all pairs of populations is also included in the model. The model of TENNESSEN *et al.* (2012) includes all of the above features, but also adds a second (more-recent) phase of accelerated exponential growth in the European continental group and includes recent exponential growth in the African continental group. `sfs_coder` is available at http://sourceforge.net/projects/sfscode/files/sfs_coder/. Throughout the paper, we will refer to the GUTENKUNST *et al.* (2009) model as the “growth” model, and the TENNESSEN *et al.* (2012) model as the “accelerated growth” model. We chose these two models because they both represent plausible demographic histories of human continental groups inferred from human sequence data, but propose dramatically different rates of recent expansion in Europeans and Africans, and hence generate different patterns of variation in samples.

We used the selection model that was inferred by BOYKO *et al.* (2008), which is a Γ -distribution of selection coefficients on non-synonymous sites in the human genome. Throughout the simulation, every non-synonymous mutation has a selection coefficient drawn from this distribution, while synonymous sites are neutral. In each simulation, we included 20 unlinked genes. Each gene was 1.65×10^3 base pairs long, which is the mean length a gene in RefSeq. While each gene is unlinked, recombination was included within each gene. Assuming a per-base recombination rate of $\rho_r = 4Nr = 10^{-3}$ and that the typical gene is composed of exons and introns spanning an average of 5.115×10^4 bp, we set the per-base $\rho_r = \frac{5.115 \times 10^4 \times 10^{-3}}{1.65 \times 10^3} = 0.031$. While this does not maintain the intron/exon structure of a gene, average linkage disequilibrium across the entire gene should be maintained.

We performed 2×10^3 simulations under each demographic model and sampled between 10^3 and 5×10^3 individuals (2×10^3 to 10^4 chromosomes) at the end of each simulation from the African and European populations. The number of simulations was chosen in order to obtain sufficiently small standard error around our power estimates such that parameter sets investigated are easily distinguishable.

4.2.4 Calculating the genetic variance

We follow several earlier studies in calculating the genetic variance due to alleles at frequency ω , including PRITCHARD (2001) and SIMONS *et al.* (2014). Genetic variance V_ω due to variants at or below allele frequency ω is given by

$$V_\omega = \int_{x=0}^{\omega} \frac{1}{2} E(z^2|x) f(x) (1-x)(x) dx \quad (4.3)$$

where $f(x)$ is the site frequency spectrum, i.e. the proportion of sampled alleles at frequency x , $E(z^2|x)$ is the mean squared effect size of variants at frequency x , and n is the sample size. In order to obtain an accurate measure of the site frequency spectrum and the effect sizes of variants at frequency x , we pool all 250 simulations performed under each model, for a total of 5×10^3 total simulated genes. We divide by V_1 , the total variance explained by genetic factors, in order to obtain a measure that represents the proportion of the variance explained by variants under frequency ω .

4.2.5 Power of SKAT-O

We obtained the SKAT-O R package from <http://www.hsph.harvard.edu/skat/>. We computed power as the proportion of simulations with p -values below 2.5×10^{-6} . We used this threshold since our study focuses on selection in coding regions, and hence our analysis is relevant to a whole exome study with $\approx 2 \times 10^4$ genes. 2×10^4 statistical tests corresponds to a Bonferroni corrected p -value of 2.5×10^{-6} . We used the default settings for SKAT-O unless otherwise stated.

4.3 Results

4.3.1 Selection and demography impact the genetic architecture of complex traits

We investigated the impact of selection and demography on the site frequency spectrum as a function of time using numerical calculations and simulations, as well as the role of singleton variants in driving variance in genetic traits under the selection-based phenotype model discussed in the methods.

In the GUTENKUNST *et al.* (2009) model of European demographic history, our numerical calculations predict that the proportion of variable sites that are singletons is strongly impacted by demographic events (solid lines, Figure 4.1A), and that the non-equilibrium predictions made under the model are in agreement with results from stochastic forward simulations (points, Figure 4.1A). In particular, expansion events increase the proportion of sites that are singletons, while contractions decrease the proportion of sites that are singletons. The vertical gray dashed line shows the current time, while the lighter dashed line shows the model predictions for the neutral case ($\gamma = 0$, not simulated).

In Figure 4.1B, we show the proportion of the variance in a genetic trait that is explained by singleton variants under this model of demography, selection and complex traits, for various values of ρ and $\tau = 1$ (solid lines are again model predictions while the points are the results of simulations). We find that the proportion of the variance in the trait that is explained by rare variants (in this case, singletons in a sample of 100 chromosomes), is strongly impacted by demographic events, and the relationship between selection and demography. Expansions increase the proportion of variable sites that are rare, which increases the role of rare variants in the trait. Contractions have the opposite effect. When $\rho = 1$, the majority of the variance in the trait is

driven by singleton variants under this model, but this proportion rapidly drops off as ρ decreases to 0.7 and then 0.

These results show that selection, demography, and the relationship between selection and effect sizes all impact the architecture of complex traits. Since these forces (and model parameters) strongly impact the joint distribution of allele frequencies and effect sizes, there is reason to believe that they may also impact the performance of association tests that seek to uncover causal variation. In the next sections, we build on these results by performing forward simulations under more complicated models that include multiple populations, migration events, and Γ -distributed selection coefficients.

4.3.2 Architecture of complex traits in multiple human populations under a selection-based phenotype model

We examined the proportion of the variance that is due to rare variants in our model of complex phenotypes using simulations of the three human continental groups under two different demographic models (GUTENKUNST *et al.*, 2009; TENNESSEN *et al.*, 2012). In Figure 4.2, we plot the proportion of the genetic variance due to variants under allele frequency ω as a function of ω for several different values of ρ with $\tau = 1$ for a sample of 1,000 individuals. Consistent with previous studies, we find that a substantial proportion of the genetic variance is attributable to rare variants only when the selection strength is very tightly correlated with the effect size. This result echoes the findings of SIMONS *et al.* (2014), but here we have extended the previous argument by including a distribution of selection coefficients that was inferred for human coding regions.

We find that demography plays a subtle role in the proportion of variance explained by rare variants. In both Europeans and Africans, the proportion of the variance explained by rare variants increases when recent accelerated growth is included in the demographic model (i.e., in the model of TENNESSEN *et al.* (2012)). Europeans and Africans have slightly different proportions of variance explained by rare variants for all sets of model parameters, with European demographic models tending to have slightly more variance explained by extremely rare variants than the corresponding African demographic model.

Interestingly, among rare variants, the preponderance of variance explained is determined by singleton variants, and not variants at more intermediate frequencies in the sample. This result

holds across all values of ρ when $\tau = 1$, for both populations and both demographic models, but is more extreme in the accelerated growth model of TENNESSEN *et al.* (2012). In the gray dashed lines, we plot $\frac{V_\omega}{V_1}$ when effect sizes are given by $\log_{10}(x)$ for alleles with frequency x under 3%. Under this effect size function, much more of the variance is attributable to variants at intermediate rare frequencies.

In Figures 4.3 and 4.4, we further extend the results of SIMONS *et al.* (2014) by considering values of τ where effect sizes grow much slower than selection coefficients. Similar to the case when $\tau = 1$, we find that a substantial proportion of the variance is explained by rare variants only when $\rho \geq 0.9$. However, the demographic model subtly changes the shape of the $\frac{V_\omega}{V_1}$ curves, meaning that the relative proportion of the variance explained by singletons vs intermediate frequency rare variants is different under the two demographic models that we considered. Specifically, we find that nearly all of the variance due to rare variants is attributable to singleton variants when $\tau = 1$, but an increasing proportion of the variance is due to moderate frequency rare variants as τ decreases. More variance is due to intermediate frequency rare variants in the growth model of GUTENKUNST *et al.* (2009) than than the accelerated growth model of TENNESSEN *et al.* (2012).

To further emphasize the role of singleton variants as a function of the parameters τ and ρ for the two demographic models, in Figure 4.5 we plot the proportion of the genetic variance due to variants under 3% frequency that is explained singletons ($V_{\frac{1}{2^n}}/V_{0.03}$). As τ decreases, the proportion of the variance that is explained by singleton variants decreases. However, the variance due to singletons is always much larger under the model of accelerated growth (blue curves), and represents a substantial proportion of the total variance due to rare variants when ρ is large for all values of τ considered. The phenotype model of WU *et al.* (2011) also results in substantially different proportions of the variance explained by singleton variants for the two demographic models. However, variance explained by singletons is always substantially lower in the model of WU *et al.* (2011) than our model when $\rho \geq 0.9$, i.e., when a substantial proportion of the total genetic variance in the phenotype is attributable to rare variants.

These results demonstrate that the joint distribution of allele frequencies and effects are very different between our model and the $\log_{10}(x)$ model. In particular, very-rare variants have much larger effects relative to intermediate frequency rare variants than previously considered models. Since statistical power is a function of this joint distribution of frequencies and effects, there is

reason to believe that the parameters of our population-genetic model may impact the power of the test..

4.3.3 The power of SKAT-O is inversely proportional to variance explained by rare variants

We investigated the power of SKAT-O as a function of the proportion of variance explained by rare variants by altering the parameter ρ in our simulations of complex phenotypes, while fixing the variance explained by the test sequence at 5% (Figure 4.6). Note that 5% is a very large amount of variance to attribute to a single gene, but we are largely concerned with the trends in power as a function of the model parameters and not whether power is overall “high” or “low”. Later in the section we will consider power calculations across a range of values of variance explained.

We find that the power is substantially lower when effects are drawn from our model as opposed to effects given by $\log_{10}(x)$, the function that was used to test the power of SKAT (WU *et al.*, 2011; LEE *et al.*, 2012b). This result holds for all model parameters, for both Europeans and Africans, and both demographic models that we considered. We also find that under the model of WU *et al.* (2011), the power is not highly sensitive to the demographic model. This is not the case under our model of effect sizes, where power is always substantially higher under the growth model of GUTENKUNST *et al.* (2009) than the accelerated growth model of TENNESSEN *et al.* (2012). Under the accelerated growth model, a larger proportion of the genetic variance due to rare variants is driven by very-rare variants (as opposed to intermediate frequency rare variants).

Since we have fixed the total variance explained by the test sequence at 5% in each simulation, we might expect that the power of SKAT-O would be unaffected by changing ρ (if SKAT-O is entirely insensitive to the joint distribution of effect sizes and allele frequencies). Alternatively, power might increase as ρ increases since SKAT-O is tuned to test for the contributions from rare variants, and rare variants play a greater role in the genetic variance as ρ increases. Counter-intuitively, we find the opposite, and the power decreases as ρ increases (Figure 4.6). This effect is most prominent under the accelerated growth model (blue bars) and when τ is large. When τ decreases to 0.5, and intermediate rare variants play some role in the trait, the trend is much less prominent.

4.4 The impact of increased sample size: a mock sequencing study with multiple genes

We repeated our power calculations for the model of TENNESSEN *et al.* (2012) using larger sample sizes of 2.5×10^3 and 5×10^3 . In these simulations, we generated phenotypes using 20 unlinked genes. We set the total heritability $h^2 = 0.5$, so the mean variance explained by each gene is 2.5%, but there is considerable variance in variance explained among genes—the variance explained for a given gene depends on the simulated effect sizes that gene. Using the simulated phenotypes and genotypes, we ran SKAT-O independently on each gene, and then aggregated the data across all genes to investigate power as a function of variance explained and sample size.

In Figure 4.7, we investigate the power of SKAT-O for $\tau = 1$ in Africans and Europeans. As in the previous section, we find that power is anti-correlated with ρ , and is substantially lower in our model than in a $\log_{10}(x)$ effect size model. Additionally, we find that increasing the sample size to 5×10^3 only slightly increases power for large values of ρ , even for very large values of variance explained up to 5% per gene. Note, there are equal numbers of simulations in each curve, but far fewer simulations with variance explained between 2% and 5% under our model than the $\log_{10}(x)$ based model, which explains the additional noise in the blue curves.

In Figures 4.8 and 4.9, we repeat this analysis for $\tau = 0.75$ and $\tau = 0.5$. We find that power is substantially increased for $\tau = 0.5$, but only slightly increased when $\tau = 0.75$. For all values of τ , we observe large differences in power between different values of ρ , and the highest power when ρ is 0 and very little of the variance is explained by rare variants.

4.4.1 A simple strategy to increase power may also increase the false-positive rate

SKAT-O includes a flexible weight-distribution over allele frequencies, which is effectively a prior on the frequency distribution of causal variants. The default distribution is a β -distribution with shape parameters 1 and 25, which gradually puts more weight on rarer alleles.

As was noted in WU *et al.* (2011), the performance of SKAT may improve when a good choice of weights is made. We re-ran SKAT-O for $\tau = 1.0$ in the African population with the rare shifted weight distribution ($\beta[0.1, 500]$). This resulted in substantial increases in power, although power was always lower than under the the $\log_{10}(x)$ effect size model with the default β -distribution

(Figure 4.10A). Moreover, power now increases as the fraction of variance explained by rare variants increases. Unfortunately, this increased power comes at a cost. We permuted the phenotypes for the same simulations and ran the same test. Under the null, we expect only 0.00025% of these tests to have p -values under 2.5×10^{-6} , but we observe that $\approx 0.3\%$ of simulations have p -values this low, or approximately a factor of 1.2×10^3 more than expected under the null. This result holds across all values of ρ that we investigated (Figure 4.10B).

If we suppose that 10 genes harbor causal rare variants for a trait of interest and that our power is approximately 50%, then we expect to find 5 of these genes in a genome-wide scan with the above parameterization of SKAT-O. However, we also expect to find 60 ($0.003 \times 2 \times 10^4$) false positive genes, providing a false discovery rate of 92%.

4.5 Discussion

A great deal of research interest has focused on the problem of “missing heritability”, which refers to the discrepancy between variance explained by associated variants discovered by Genome Wide Association Studies (GWAS) and estimates of the narrow-sense heritability of common, genetically complex phenotypes. Although there are many possible explanations for this discrepancy, one of the most popular explanations is that rare variants of large effect may make up the difference. This hypothesis has been used as motivation for a number of large-scale sequencing studies of large cohorts.

As sequencing technology has progressed to the point where very-rare (and potentially novel) variants are routinely detected in a large samples, there has been a corresponding push to develop statistical tools to detect causal rare variants. One of the most popular tools is known as SKAT-O, or the optimized sequence kernel association test (LEE *et al.* 2012a). This test is now routinely applied to large sequencing datasets.

Although it is clear that rare variant association tests are very successful at detecting associations under some phenotype models, it does not necessarily follow that previously investigated phenotype models are biologically or evolutionarily plausible. Here, we built off previous models of the relationship between selection strength and effect size to develop a model of complex phenotypes (EYRE-WALKER 2010; SIMONS *et al.* 2014), and performed simulations under this model that include complex human demography and a previously inferred distribution of selection coeffi-

cients for human coding sequences. We showed that genetic architecture and power calculations are quite sensitive to demography and the relationship between selection strength and effect size, and that power estimates under our model are generally substantially lower than under the previously investigated model.

A principle reason for this discrepancy is the role of very-rare variants in complex phenotypes under our model. Under our evolutionary model, when rare variants explain a substantial proportion of the genetic variance, the greatest contributions are made by extremely rare variants. When accelerated growth is included in the demographic model, singleton variants have even larger effect sizes. Singleton variants are the hardest of all variants to detect since they occur only once in the data set, and non-causal singletons are ubiquitous in sequencing of large samples. We showed that some modifications to the default settings of SKAT-O can substantially improve power under our model, but these changes may also come at an increase in false positive rates. Although we cannot provide a concrete reason for this increase in false positives, we speculate that the distribution of singleton variants is not well captured by the null model of the statistical test.

If very-rare variants do play a substantial role in driving variance in phenotypes at the population level, then this may also have implications for the design of sequencing experiments, where researchers sometimes must choose between deeper sequencing and more samples. Singletons are the hardest variants to call in the data set and power and false positive rates for detecting singletons may depend strongly on variant calling methods (e.g., multi-sample vs. single-sample calling). These considerations may have important implications for power of rare variant association tests if our model accurately captures the genetic architecture for some complex phenotypes of interest.

Our results demonstrate that it may be more difficult to interpret a null signal from a rare variant association test than previously appreciated, as power is highly sensitive to several model assumptions and may be lower than was estimated in previous studies. However, it is entirely possible that while the distribution of selection coefficients that we used was inferred from human polymorphisms, the resulting phenotype model does not accurately depict complex human phenotypes. Still, it is clear that *any* phenotype model that is relevant to a rare variant association test must include selection, because rare variants do not contribute substantially to variance in phenotypes in the absence selection. Our model also has implications for the distribution of phenotypes in a population. In particular, the distribution is not expected to be exactly normal because of

the presence of very large effect mutations, which generate fatter tails. In future studies, it will be advantageous to exploit these signals to compare various models of heritability (including models with and without selection) to put firmer bounds on the proportion of the genetic variance that is determined by rare variants and provide further insight into the utility of rare variant association tests.

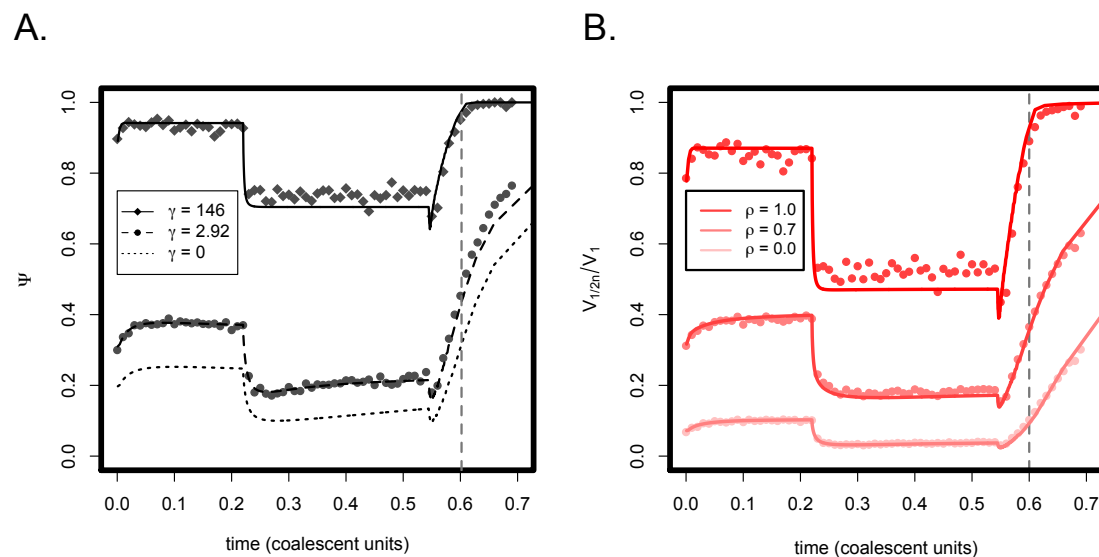


Figure 4.1: We calculated Ψ , the proportion of variable sites that are singletons (A) as well as the proportion of the genetic variance in a complex trait that is due to singleton variants (B) for a sample of 100 chromosomes for the marginal European demographic history in the model of GUTENKUNST *et al.* (2009) and the selection-based phenotype model of SIMONS *et al.* (2014). The solid, dashed, and dotted lines show the results of numerical calculations under the Wright-Fisher model, whereas the points are the results of stochastic forward simulations. Each point represents the mean across 200 simulations.

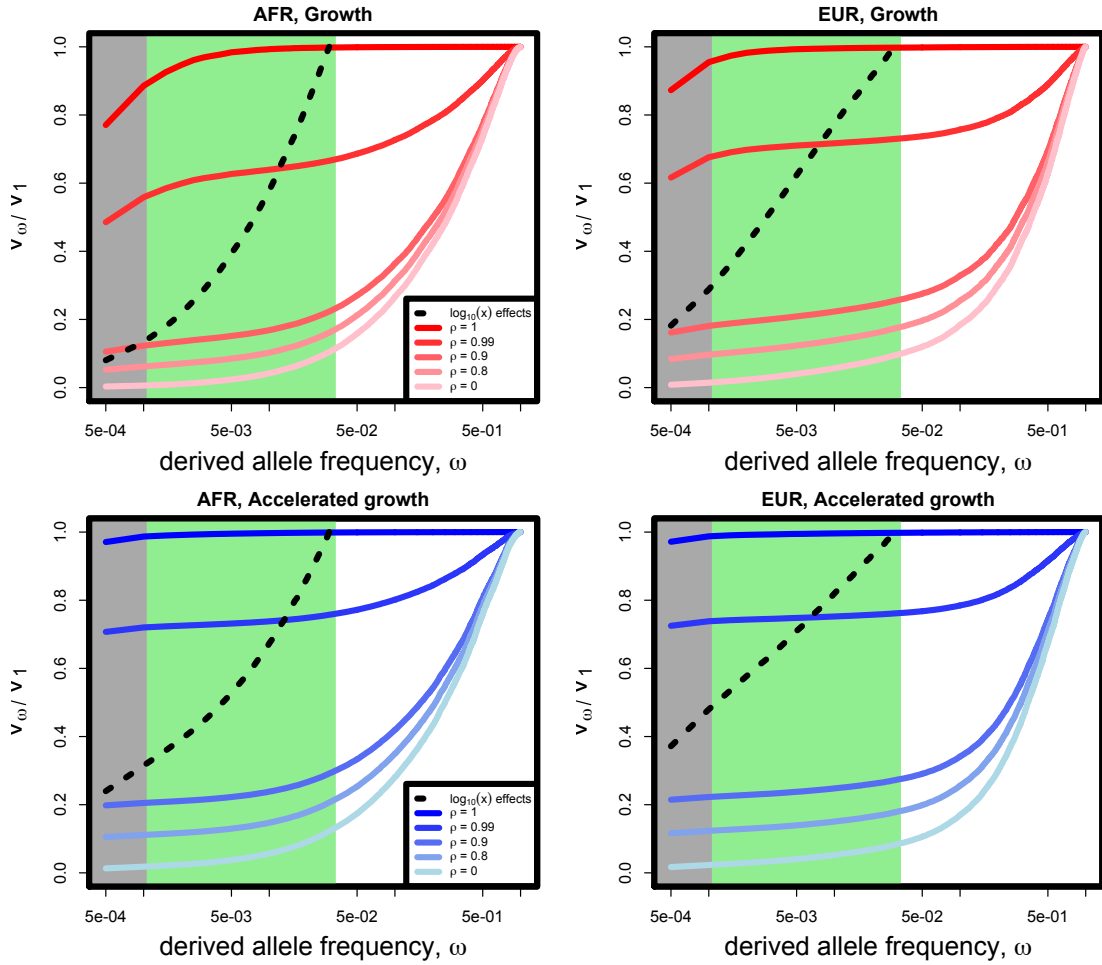


Figure 4.2: The proportion of the genetic variance $\frac{V_{\omega}}{V_1}$ explained by variants under allele frequency ω for two different demographic models of human history and two different populations with $\tau = 1.0$. The gray shaded area represents alleles at very low frequency ($< 10^{-3}$) and the green area represents moderately rare alleles (< 0.03). (Abbreviations: EUR, European continental group; AFR, African continental group)

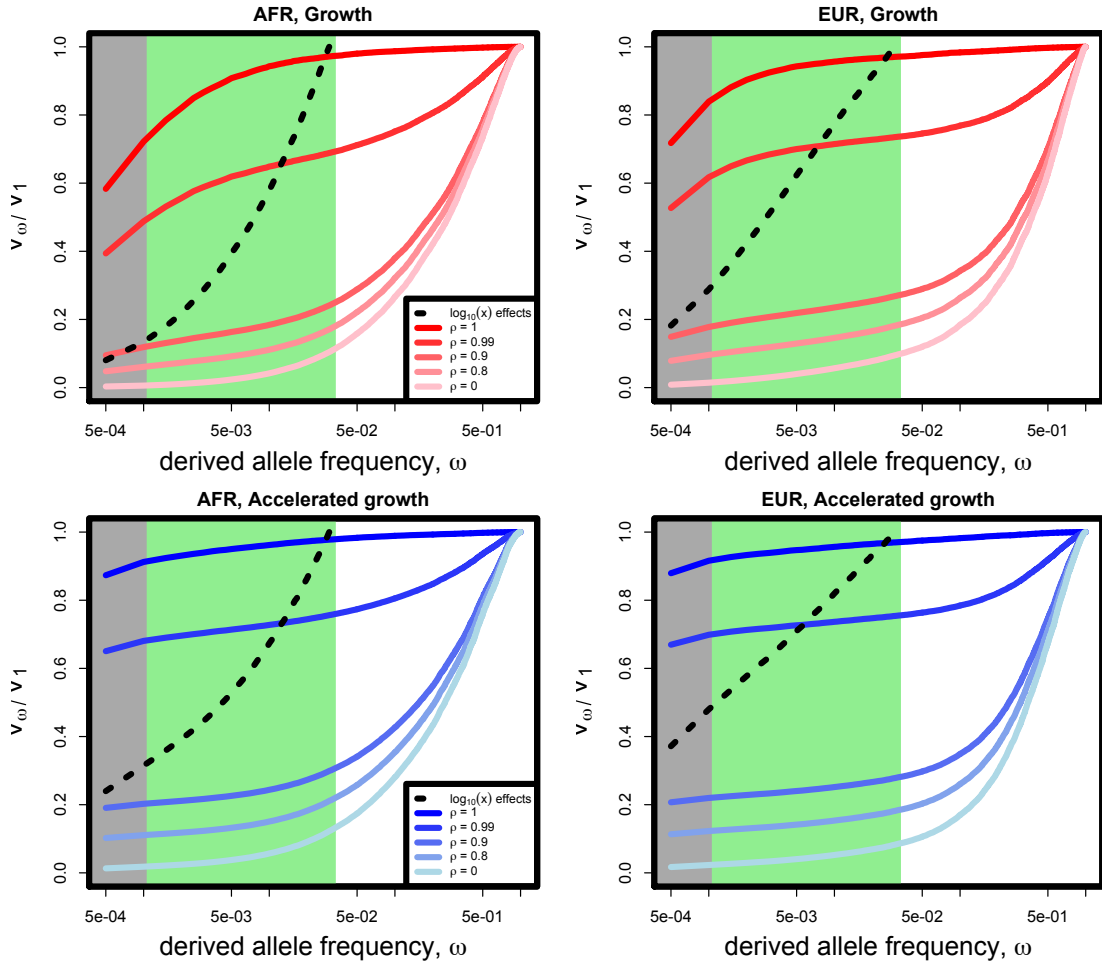


Figure 4.3: The proportion of the genetic variance $\frac{V_{\omega}}{V_1}$ explained by variants under allele frequency ω for two different demographic models of human history and two different populations with $\tau = 0.75$. The gray shaded area represents alleles at very low frequency ($< 10^{-3}$) and the green area represents moderately rare alleles (< 0.03). (Abbreviations: EUR, European continental group; AFR, African continental group)

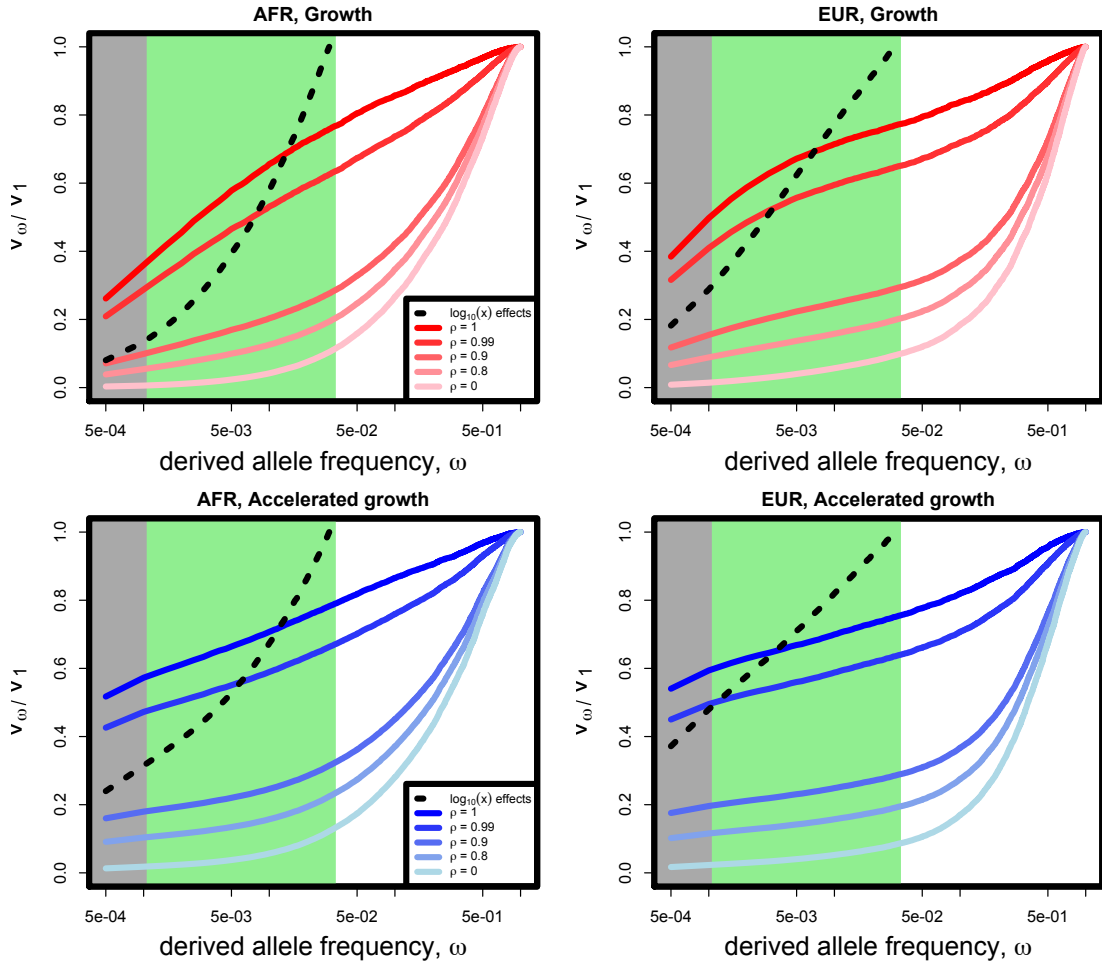


Figure 4.4: The proportion of the genetic variance $\frac{V_{\omega}}{V_1}$ explained by variants under allele frequency ω for two different demographic models of human history and two different populations with $\tau = 0.5$. The gray shaded area represents alleles at very low frequency ($< 10^{-3}$) and the green area represents moderately rare alleles (< 0.03). (Abbreviations: EUR, European continental group; AFR, African continental group)

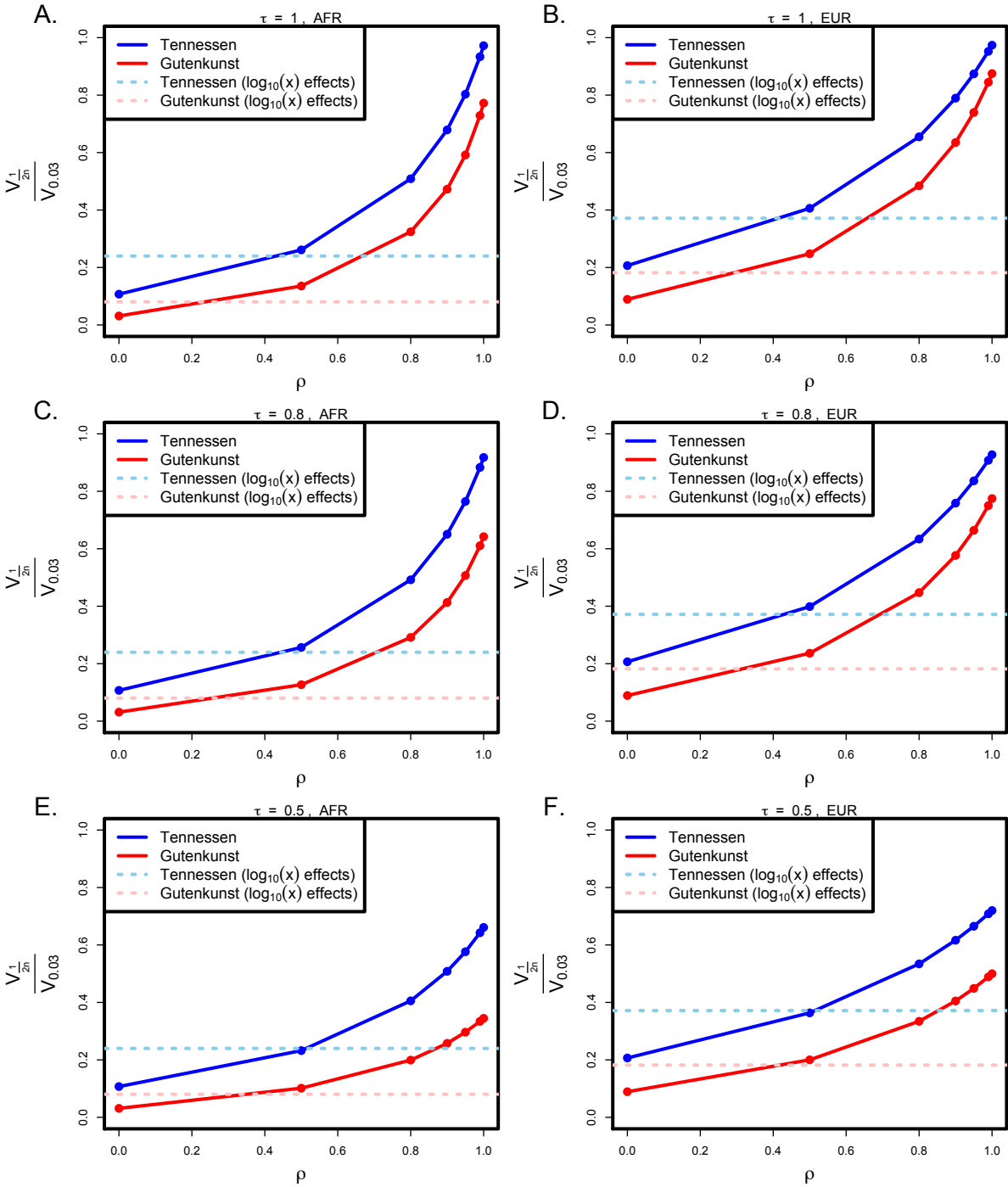


Figure 4.5: The proportion of the variance due to rare variants (allele frequency < 0.03) that is due to singleton variants under various models of effect sizes. The solid lines and points show results for our model, where effect sizes are a function of selection coefficients (see Methods). The dashed lines show results for the phenotype model where effects are taken as proportional to $\log_{10}(x)$ for alleles at frequency x . (Abbreviations: *AFR*, African continental group; *EUR*, European continental group).

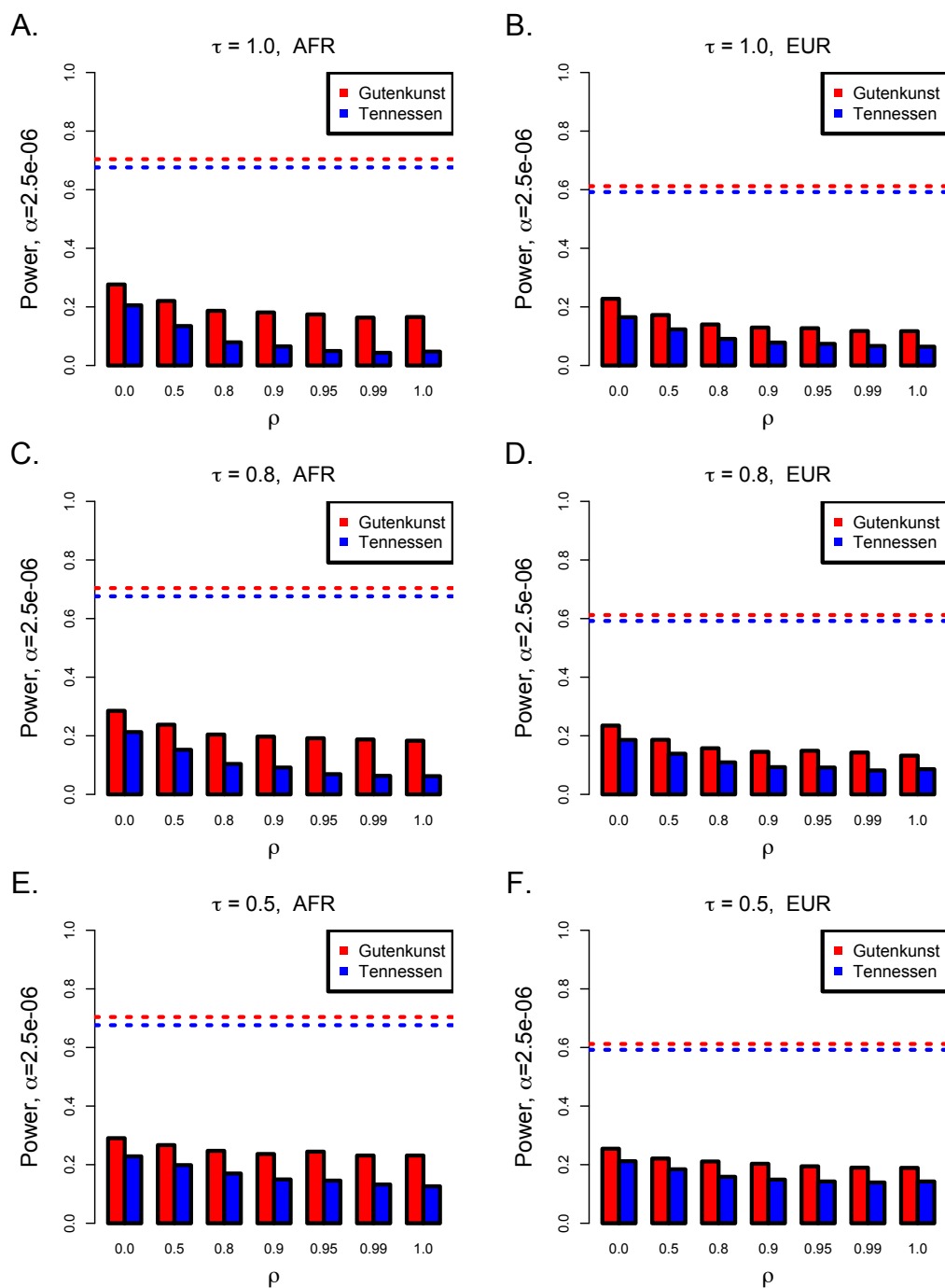


Figure 4.6: The power of SKAT-O under various effect size models. The dashed lines show the power when the effect sizes are taken to be proportional to $\log_{10}(x)$ for alleles at frequency x , while the bars show our model (see Methods). The accelerated growth model of TENNESSEN *et al.* (2012) is shown in blue, and growth model of GUTENKUNST *et al.* (2009) in red. (Abbreviations: *AFR*, African continental group; *EUR*, European continental group).

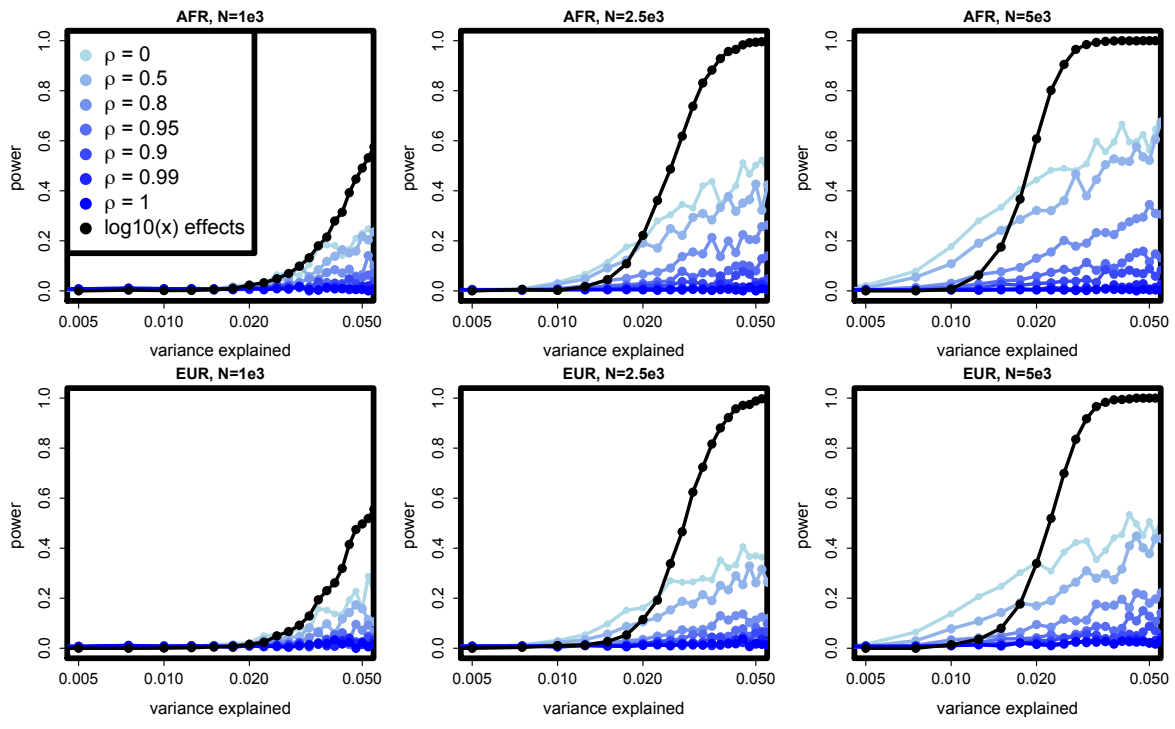


Figure 4.7: The power of SKAT-O as a function of variance explained and sample size N , for various values of ρ and $\tau = 1$. All simulations are under the model of TENNESSEN *et al.* (2012).

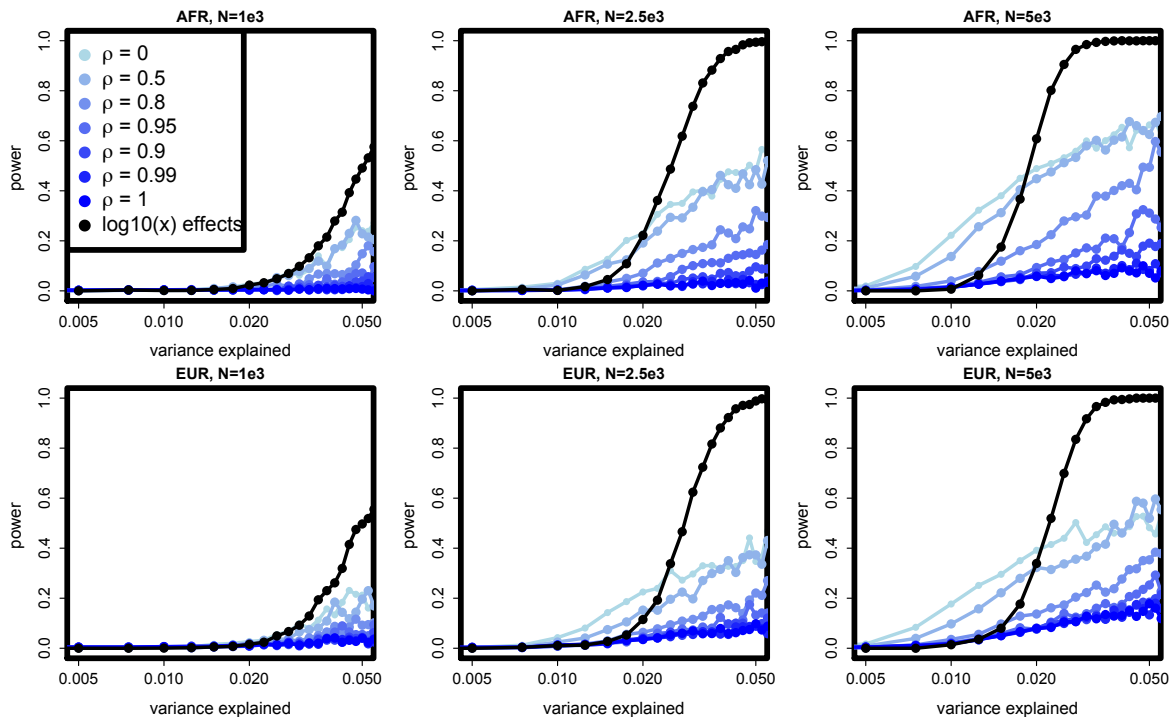


Figure 4.8: The power of SKAT-O as a function of variance explained and sample size N , for various values of ρ and $\tau = 0.75$. All simulations are under the model of TENNESSEN *et al.* (2012).

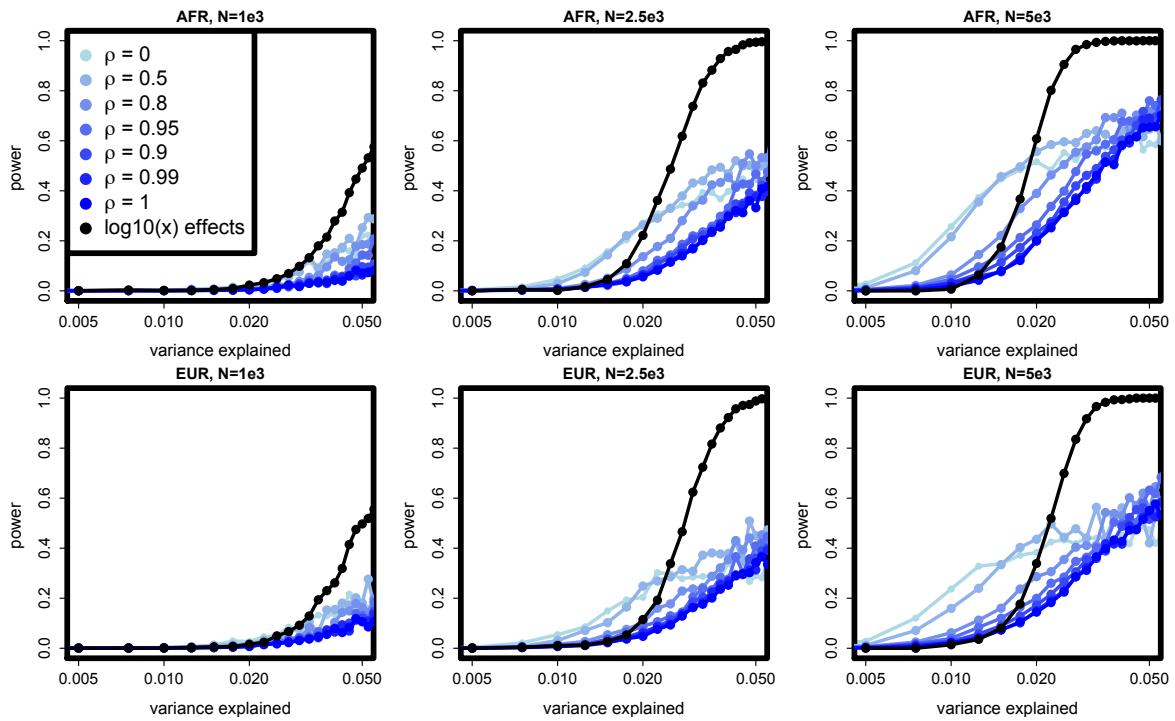


Figure 4.9: The power of SKAT-O as a function of variance explained and sample size N , for various values of ρ and $\tau = 0.5$. All simulations are under the model of TENNESSEN *et al.* (2012).

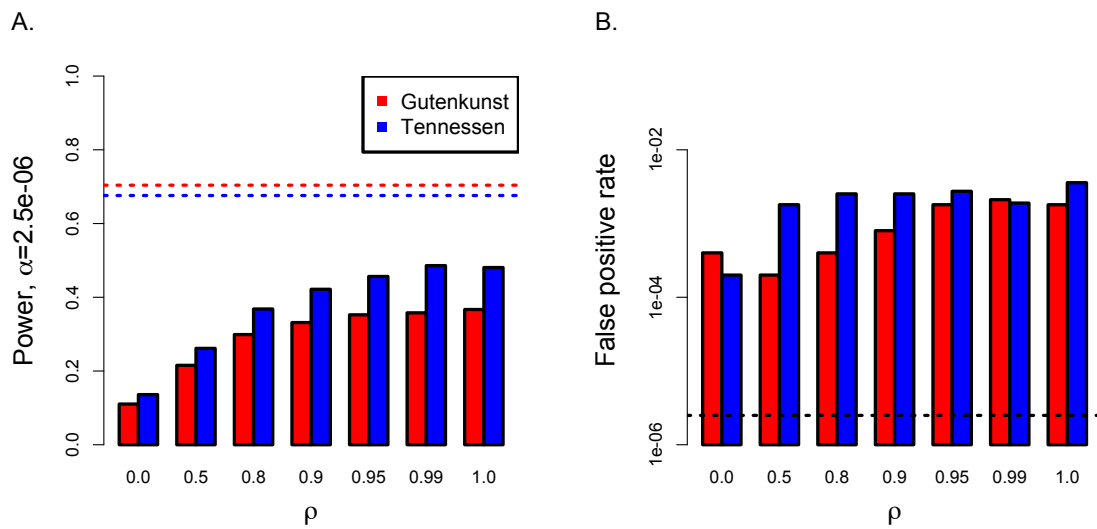


Figure 4.10: We adjusted the default weights of SKAT-O to be more rare-shifted ($\beta[0.1, 500]$). We computed the power of SKAT-O with these weights in the African continental group (A). We also computed the false-positive rate by permuting the phenotypes (B).

References

- BOYKO, A. R., S. H. WILLIAMSON, A. R. INDAP, J. D. DEGENHARDT, R. D. HERNANDEZ, *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics* **4**: e1000083.
- EYRE-WALKER, A., 2010 Evolution in health and medicine sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences of the United States of America* **107 Suppl 1**: 1752–1756.
- GAUGLER, T., L. KLEI, S. J. SANDERS, C. A. BODEA, A. P. GOLDBERG, *et al.*, 2014 Most genetic risk for autism resides with common variation. *Nature genetics* **46**: 881–885.
- GUTENKUNST, R. N., R. D. HERNANDEZ, S. H. WILLIAMSON, and C. D. BUSTAMANTE, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics* **5**: e1000695.
- HALLER, G., D. G. TORGERSON, C. OBER, and E. E. THOMPSON, 2009 Sequencing the *il4* locus in african americans implicates rare noncoding variants in asthma susceptibility. *The Journal of allergy and clinical immunology* **124**: 1204–9.e9.
- HERNANDEZ, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. *Bioinformatics (Oxford, England)* **24**: 2786–2787.
- HOFFMANN, T. J., N. J. MARINI, and J. S. WITTE, 2010 Comprehensive approach to analyzing rare genetic variants. *PloS one* **5**: e13584.
- LEE, S., M. C. WU, and X. LIN, 2012a Optimal tests for rare variant effects in sequencing association studies. *Biostatistics (Oxford, England)* **13**: 762–775.
- LEE, S., M. C. WU, and X. LIN, 2012b Optimal tests for rare variant effects in sequencing association studies. *Biostatistics (Oxford, England)* **13**: 762–775.
- LOHMUELLER, K. E., 2014 The impact of population demography and selection on the genetic architecture of complex traits. *PLoS genetics* **10**: e1004379.
- MAHER, M. C., L. H. URICCHIO, D. G. TORGERSON, and R. D. HERNANDEZ, 2012 Population genetics of rare variants and complex diseases. *Human heredity* **74**: 118–128.
- MCVICKER, G., D. GORDON, C. DAVIS, and P. GREEN, 2009 Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**: e1000471.
- PRITCHARD, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics* **69**: 124–137.
- SIMONS, Y. B., M. C. TURCHIN, J. K. PRITCHARD, and G. SELLA, 2014 The deleterious mutation load is insensitive to recent population history. *Nature genetics* **46**: 220–224.
- TENNESSEN, J. A., A. W. BIGHAM, T. D. O’CONNOR, W. FU, E. E. KENNY, *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)* **337**: 64–69.

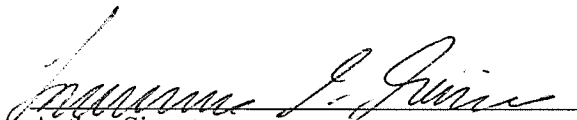
- TORGERSON, D. G., A. R. BOYKO, R. D. HERNANDEZ, A. INDAP, X. HU, *et al.*, 2009 Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS genetics* **5**: e1000592.
- TORGERSON, D. G., D. CAPURSO, R. A. MATHIAS, P. E. GRAVES, R. D. HERNANDEZ, *et al.*, 2012 Resequencing candidate genes implicates rare variants in asthma susceptibility. *American journal of human genetics* **90**: 273–281.
- URICCHIO, L. H., R. TORRES, J. S. WITTE, and R. D. HERNANDEZ, 2015 Population genetic simulations of complex phenotypes with implications for rare variant association tests. *Genetic epidemiology* **39**: 35–44.
- WU, M. C., S. LEE, T. CAI, Y. LI, M. BOEHNKE, *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* **89**: 82–93.
- YANG, J., B. BENYAMIN, B. P. MCEVOY, S. GORDON, A. K. HENDERS, *et al.*, 2010 Common snps explain a large proportion of the heritability for human height. *Nature genetics* **42**: 565–569.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.


Author Signature

1/5/2015
Date