ADAPTIVE AND NEUTRAL EVOLUTIONARY INSIGHTS FROM STATISTICAL ANALYSES OF POPULATION GENETIC DATA

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF BIOLOGY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Julie Marie Granka
March 2013
© 2013 by Julie Marie Granka. All Rights Reserved.
Re-distributed by Stanford University under license with the author.

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
http://creativecommons.org/licenses/by-nc/3.0/us/

This dissertation is online at: http://purl.stanford.edu/md970sh0103
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Marcus Feldman, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Hunter Fraser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Dmitri Petrov

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Hua Tang

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
Abstract

Organismic evolution involves both selective and neutral forces, although their relative contributions are often unknown. This thesis proposes novel statistical methods for analyzing genetic data from a variety of organisms, including yeast, *Mycobacterium tuberculosis*, and humans. The chapters of this thesis provide complimentary perspectives on the relative roles of selection and neutrality, from the molecular to the population level, and present various statistical tools for genetic data analysis. Chapter 2 proposes a maximum-likelihood based method with which to classify and identify interactions, or epistasis, between pairs of genes. Chapter 3 details a study of genetic data from *Mycobacterium tuberculosis* isolated from human Aboriginal Canadian communities; our analyses suggest that the bacterium spread to these communities via the Canadian fur trade in the 18th and 19th centuries. Chapter 4 discusses the detection of signatures of natural selection in the genomes of 12 diverse African human populations, and proposes novel considerations for identifying biological functions under selection and for comparing signals of selection between populations. Finally, Chapter 5 details the inference of the genetic basis and evolutionary history of light skin pigmentation and short stature in the genetically diverse ≠Khomani Bushmen of the Kalahari Desert of South Africa, believed to be one of the world’s most ancient human populations. These chapters emphasize that a more complete understanding of the evolutionary history of humans and other organisms requires not only the consideration of neutral and selective processes, but also both phenotypic and genetic information. The statistical methods and approaches presented in the following chapters have the potential to improve inferences of natural selection and demography from genetic data, as well as provide insight into the relative roles of both.
Acknowledgements

So many people have been tremendously supportive during my years as a graduate student at Stanford. I would like to extend my gratitude to them.

First, I would like to thank my advisor Dr. Marc Feldman. Marc has been a constant source of support, encouragement, and knowledge. I have learned so much from Marc about conducting research, formulating scientific questions, and drawing meaningful conclusions. I thank Marc for giving me the opportunity to be involved in so many exciting and unique research projects, and I truly appreciate the freedom he allowed me in pursuing my own intellectual interests.

My committee members Dr. Hua Tang, Dr. Dmitri Petrov, and Dr. Hunter Fraser have been inspirations to me during the course of my program, and I thank them for their ongoing support and feedback. A special thanks to Dr. Carlos Bustamante for introducing me to population genetics many years ago and for his continued guidance. I would also like to thank Dr. Daniel Fisher, the external chair on my thesis committee.

Dr. Brenna Henn has been an incredible mentor to me over the past several years. I have always been able to turn to Brenna for advice – receiving in return not only answers and assistance, but motivation to tackle even greater issues. Brenna has taught me a great deal, and I can’t thank her enough.

My sincere thanks to Dr. Caitlin Pepperell, whose enthusiasm and curiosity were contagious. I learned a great deal from Caitlin about the process of scientific inquiry, as well as the value of interdisciplinary collaboration. I would also like to thank Dr. Hong Gao, who worked with me on my first project as a graduate student.

Many thanks to my collaborators Dr. Jeff Kidd and Chris Gignoux for their helpful advice and suggestions, as well as Dr. Mike Palmer for his support in running jobs on our computer cluster. I would also like to thank my collaborators from the University of Stellenbosch in South Africa; in particular, Dr. Eileen Hoal, Dr. Marlo Möller and Dr. Cedric Werely for working with me in the field. Much of my work would not be possible without the support of the South African San Institute (SASI), our kind, patient and knowledgeable interpreters, and the #Khomani San community for their support and interest in our research.

The administrative staff of the Biology Department, including Dan King, Valerie Kiszka, Claudia Ortega, Matt Pinheiro, and Monica Bernal, has been extremely helpful in guiding me through the Ph.D. program. I would especially like to thank both Larry Bond and Jim Collins for their help and assistance while I have been at Stanford.

Not least, I extend a huge thank you to all of my labmates, who are among my closest friends. Graduate school would have been far less enjoyable without our hallway and doorway discussions and their inventive
ice creams and frozen yogurts. In particular, I want to thank past and current Feldman Lab members Michal Arbilly, Lily Blair, Elhanan Borenstein, Amanda Casto, Oana Carja, Nicole Creanza, Laurel Fogarty, Robert Furrow, my officemate Philip Greenspoon, Edith Katsnelson, my former officemate Laurent Lehmann, Misha Lipatov, Mike McLaren, Ricardo Pinho, Allison Rhines, Deb Rogers, Marcel Salathé, Jeremy van Cleve and Daniel Weissman. I would also like to thank Nandita Garud, Philipp Messer, Leithen M’Gonigle, Dane Klinger, and the amazing friends who have supported me prior to and during graduate school.

Finally, I extend a huge thank you to my family. My parents, Rachele and Edward Granka, have been nothing but encouraging and supportive, and I thank them for their infinite caring, advice, and understanding. A thank you isn’t enough for my sister Laura Granka and her endless sage advice, humor, motivation, and friendship. I also thank my Uncle Lu for his thoughtful support, as well as the rest of my family: uncles, aunts, cousins, and loving grandparents.
Contents

Abstract \hspace{1cm} v
Acknowledgements \hspace{1cm} vi

1 Introduction \hspace{1cm} 1

2 On the Classification of Epistatic Interactions \hspace{1cm} 4
2.1 Introduction \hspace{1cm} 5
2.2 Methods \hspace{1cm} 7
2.2.1 Epistatic Subtype Selection \hspace{1cm} 7
2.2.2 Generation of Simulated Data \hspace{1cm} 10
2.2.3 Analysis of Experimental Data \hspace{1cm} 10
2.3 Application to Simulated Data \hspace{1cm} 11
2.3.1 Estimation of Epistatic Measures \hspace{1cm} 11
2.3.2 Bias and Variance of the Epistatic Parameter (\(\epsilon\)) \hspace{1cm} 12
2.4 Application to Experimental Data \hspace{1cm} 12
2.4.1 St Onge et al. [255] Dataset \hspace{1cm} 12
2.4.2 Jasnos and Korona [131] Dataset \hspace{1cm} 13
2.5 Discussion \hspace{1cm} 14
2.6 Acknowledgements \hspace{1cm} 17
2.7 Figures and Tables \hspace{1cm} 18

3 Dispersal of *Mycobacterium tuberculosis* via the Canadian fur trade \hspace{1cm} 22
3.1 Introduction \hspace{1cm} 24
3.2 Results \hspace{1cm} 25
3.3 Discussion \hspace{1cm} 27
3.4 Materials and Methods \hspace{1cm} 29
3.4.1 Population Descriptions \hspace{1cm} 29
3.4.2 Genotyping Methods \hspace{1cm} 29
3.4.3 Mutation Rate Estimate \hspace{1cm} 30
4 Limited Evidence for Classic Selective Sweeps in African Populations

4.1 Introduction ................................................... 38
4.2 Methods ....................................................... 39
  4.2.1 Samples ....................................................... 39
  4.2.2 Single SNP Statistics ........................................ 40
  4.2.3 Haplotype Statistics ....................................... 40
  4.2.4 Coalescent Simulations ..................................... 42
  4.2.5 Gene Annotation Enrichment for Haplotype Statistics .... 43
4.3 Results ......................................................... 44
  4.3.1 Single SNP Statistics ........................................ 44
  4.3.2 Haplotype Statistics ....................................... 45
  4.3.3 Enrichment of Biological Functions ......................... 47
4.4 Discussion ..................................................... 48
  4.4.1 Population Differentiation .................................. 48
  4.4.2 Comparison of Empirical Tail Windows ......................... 49
  4.4.3 Enrichment of Biological Functions ......................... 50
  4.4.4 Complicating Factors for Studies of Selection ................ 51
4.5 Data Access .................................................... 52
4.6 Acknowledgments .............................................. 52
4.7 Figures and Tables ............................................. 54

5 Evolution of Height and Skin Pigmentation in the ≠Khomani Bushmen

5.1 Introduction ................................................... 64
5.2 Methods ....................................................... 66
  5.2.1 Samples ....................................................... 66
  5.2.2 Relatedness ................................................... 67
  5.2.3 Ancestry Estimation ......................................... 67
  5.2.4 Statistical Analyses ......................................... 68
5.3 Results ......................................................... 70
  5.3.1 Ancestry ....................................................... 70
  5.3.2 Phenotypes ..................................................... 71
5.4 Discussion ..................................................... 74
  5.4.1 Ancestry Estimation ......................................... 74
### C Supplemental Materials: Limited Evidence for Classic Selective Sweeps in African Populations

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.1</td>
<td>Supplemental Methods</td>
<td>143</td>
</tr>
<tr>
<td>C.1.1</td>
<td>Samples</td>
<td>143</td>
</tr>
<tr>
<td>C.1.2</td>
<td>Population Differentiation</td>
<td>143</td>
</tr>
<tr>
<td>C.1.3</td>
<td>Choice of Several Haplotype Statistic Parameters</td>
<td>144</td>
</tr>
<tr>
<td>C.1.4</td>
<td>Comparison of Empirical Tail Windows</td>
<td>148</td>
</tr>
<tr>
<td>C.1.5</td>
<td>Neutral Coalescent Simulations</td>
<td>148</td>
</tr>
<tr>
<td>C.1.6</td>
<td>Selective Sweep Coalescent Simulations</td>
<td>149</td>
</tr>
<tr>
<td>C.1.7</td>
<td>Gene Ontology Enrichment</td>
<td>150</td>
</tr>
<tr>
<td>C.2</td>
<td>Supplemental Results</td>
<td>152</td>
</tr>
<tr>
<td>C.2.1</td>
<td>Population Differentiation</td>
<td>152</td>
</tr>
<tr>
<td>C.2.2</td>
<td>Haplotype Statistics</td>
<td>154</td>
</tr>
<tr>
<td>C.2.3</td>
<td>Gene Ontology Enrichments</td>
<td>156</td>
</tr>
<tr>
<td>C.3</td>
<td>Supplemental Figures and Tables</td>
<td>159</td>
</tr>
</tbody>
</table>

**Bibliography** | 181
List of Tables

2.1 Fractions of simulations that recover the true model (null or epistatic) among the six models (three null models and three epistatic models) with different values of the epistasis coefficient, fitness standard deviation, and mean fitnesses of single mutants. ........................................ 19
2.2 Summary of gene pairs with the indicated epistatic subtypes, inferred using the FDR procedure with the BIC method. ................................................................. 20
2.3 Comparison of validation measures for each dataset for different variations of the FDR and BIC procedure, considering only a subset of epistatic subtypes with their corresponding null models. .................................................................................. 21
3.1 Within- and between-lineage AMOVA based on 12 minisatellite loci for *M. tuberculosis* from Quebec, Ontario, Saskatchewan, and Alberta. .................................................. 35
3.2 Genetic timing estimates for DS6Quebec *M. tuberculosis* lineage in calendar years. ....... 36
4.1 Significantly enriched Gene Ontology (GO) terms in the top empirical 0.1% of windows in each population. .............................................................. 60
4.2 Significantly enriched GO terms in the top empirical 0.1% of windows in each population, continued. .............................................................. 61
5.1 Pearson correlations of self-reported measures of ancestry with ancestries inferred for each individual. .......................................................... 92
5.2 Means (and standard deviations) of measured phenotypes for all individuals and the subset of genotyped individuals. .................................................. 93
5.3 Estimates of heritability ($h^2$) of height, skin pigmentation, and tanning (wrist minus inner arm pigmentation) using various methods. ........................................ 94
5.4 Coefficients for regressions of height and pigmentation on ancestries. .......................... 95
5.5 Top SNPs associated with height from a genome-wide association study using *EMMAX*. .... 96
A.1 Fractions of simulations that recover the true model among the six models when allowing both beneficial and deleterious mutations. ........................................ 112
A.2 Fractions of simulations that recover the true model among the six models, with different values of the epistasis coefficient and mean fitnesses of single mutants; number of replicates is equal to 10. 113
A.3 Fractions of simulations that recover the true model among the six models with different values of the epistasis coefficient and mean fitnesses of single mutants; number of replicates is equal to 2. 114
A.4 Summary of gene pairs with the indicated epistatic subtypes, inferred using the FDR procedure with the BIC method which considers all three epistatic subtypes and their corresponding null models (Study S, fitnesses measured in the absence of MMS). 115
A.5 Comparison of validation measures for each dataset for different variations of the FDR and BIC procedure, considering only a subset of epistatic subtypes with their corresponding null models (Study S, fitnesses measured in the absence of MMS). 116
A.6 Comparison of validation measures for each dataset for different variations of the FDR and BIC procedure, considering only a subset of epistatic subtypes with their corresponding null models; significance assessed by permutation. 117

B.1 Frequencies of M. tuberculosis lineages in four populations. 140
B.2 Fur trade migrants, Quebec to the Northwest, 1710 - 1870. 141
B.3 Estimated mutation rates for M. tuberculosis minisatellite loci, per locus and per transmission generation. 142

C.1 Populations analyzed. 177
C.2 Mean $F_{ST}$ values between pairs of populations. 178
C.3 Mantel correlations between matrices of the indicated variables for each haplotype statistic. 179
C.4 Significantly enriched Gene Ontology (GO) terms in the top empirical 1% and 5% of windows in each population. 180
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Distribution of the epistasis values ($\epsilon$) for significant epistatic pairs in the Jasnos and Korona [131] dataset</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>Fur trade geography, regional <em>M. tuberculosis</em> lineage frequencies and historical timeline</td>
<td>32</td>
</tr>
<tr>
<td>3.2</td>
<td>Minimum spanning tree of minisatellite haplotypes (12 loci) from <em>M. tuberculosis</em> within the DS6<em>Quebec</em> lineage, in QU, ON, SK, AB</td>
<td>33</td>
</tr>
<tr>
<td>3.3</td>
<td>Genetic diversity of <em>M. tuberculosis</em> populations</td>
<td>34</td>
</tr>
<tr>
<td>4.1</td>
<td>Permutation procedure used to assess significance of enrichment of Gene Ontology (GO) terms in the top $x%$ of 100 kb genomic windows</td>
<td>54</td>
</tr>
<tr>
<td>4.2</td>
<td>Enrichment of genic and non-genic SNPs for values of $F_{ST}$ and for average values of derived allele frequency difference ($</td>
<td>\delta</td>
</tr>
<tr>
<td>4.3</td>
<td>Extreme $</td>
<td>\delta</td>
</tr>
<tr>
<td>4.4</td>
<td>P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH MKK statistic in each population</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>Number of shared 100 kb genomic windows in the top empirical 1% for iHS vs. mean $F_{ST}$ for pairs of populations</td>
<td>58</td>
</tr>
<tr>
<td>4.6</td>
<td>Number of shared 100 kb genomic windows in the top empirical 1% for XP-EHH vs. mean $F_{ST}$ for pairs of populations</td>
<td>59</td>
</tr>
<tr>
<td>5.1</td>
<td>Inferred pedigrees for a subset of related sampled individuals</td>
<td>79</td>
</tr>
<tr>
<td>5.2</td>
<td>Individual ancestry proportions for the indicated number of ancestral populations inferred for phenotyped individuals using <em>ADMIXTURE</em></td>
<td>80</td>
</tr>
<tr>
<td>5.3</td>
<td>Cross-validation errors and log likelihoods for models estimated using <em>ADMIXTURE</em> for various numbers of ancestral populations</td>
<td>81</td>
</tr>
<tr>
<td>5.4</td>
<td>Ancestry proportions versus age of phenotyped ≠Khomani Bushmen individuals</td>
<td>82</td>
</tr>
<tr>
<td>5.5</td>
<td>Violin plots of height and skin pigmentation measurements for genotyped individuals</td>
<td>83</td>
</tr>
<tr>
<td>5.6</td>
<td>Results from a linear mixed model regression of height on sex and age</td>
<td>84</td>
</tr>
<tr>
<td>5.7</td>
<td>Results from a linear mixed model regression of tanning (wrist minus inner arm pigmentation) on sex and age</td>
<td>85</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td>Results from a linear mixed model regression of height on sex, age, and proportion European ancestry.</td>
<td></td>
</tr>
<tr>
<td>5.9</td>
<td>Results from linear mixed model regressions of inner arm skin pigmentation on proportion ancestry, for various ancestries.</td>
<td></td>
</tr>
<tr>
<td>5.10</td>
<td>Genetic relationships for pairs of individuals inferred from EMMAX using the Balding-Nichols method.</td>
<td></td>
</tr>
<tr>
<td>5.11</td>
<td>QQ plots of $-\log_{10} SNP$ p-values for genome-wide association studies of height and inner arm pigmentation using EMMAX.</td>
<td></td>
</tr>
<tr>
<td>5.12</td>
<td>Manhattan plots of $-\log_{10} SNP$ p-values for height association using EMMAX.</td>
<td></td>
</tr>
<tr>
<td>5.13</td>
<td>Effect sizes by genotype and allele frequencies in worldwide HGDP populations for top height SNPs.</td>
<td></td>
</tr>
<tr>
<td>A.1</td>
<td>Power of our method to uncover the true (simulated) epistatic model.</td>
<td></td>
</tr>
<tr>
<td>A.2</td>
<td>Power of our method to uncover the true (simulated) epistatic model assuming $\epsilon_{ij} = 0.3$, $\sigma_{ij} = 0.1$, and $\mu_i = \mu_j = 0.8$.</td>
<td></td>
</tr>
<tr>
<td>A.3</td>
<td>The distribution of the mean squared error (MSE) of the MLE of the epistatic parameter under the additive, multiplicative, and minimum epistatic models as well as the corresponding null model.</td>
<td></td>
</tr>
<tr>
<td>A.4</td>
<td>Venn diagram comparing the overlap of epistatic pairs found examining the St Onge et al. [255] dataset using the authors’ original method and our BIC method.</td>
<td></td>
</tr>
<tr>
<td>A.5</td>
<td>Distribution of the epistasis values ($\epsilon$) for significant epistatic pairs, determined using the FDR procedure with the BIC method for all gene pairs of the St Onge et al. [255] dataset in the presence of MMS.</td>
<td></td>
</tr>
<tr>
<td>A.6</td>
<td>Several epistatic networks constructed based on the St Onge et al. [255] dataset.</td>
<td></td>
</tr>
<tr>
<td>A.7</td>
<td>Several epistatic networks constructed based on the Jasnos and Korona [131] dataset.</td>
<td></td>
</tr>
<tr>
<td>A.8</td>
<td>Quantile-Quantile plot of the single-mutant fitness values for several genes analyzed in the St Onge et al. [255] dataset.</td>
<td></td>
</tr>
<tr>
<td>A.9</td>
<td>Quantile-Quantile plot of the single-mutant fitness values for several genes analyzed in the Jasnos and Korona [131] dataset.</td>
<td></td>
</tr>
<tr>
<td>B.1</td>
<td><em>M. tb.</em> minimum spanning trees based on minisatellite data and lineage typing.</td>
<td></td>
</tr>
<tr>
<td>B.2</td>
<td>Diversity calculations obtained from minisatellite data.</td>
<td></td>
</tr>
<tr>
<td>B.3</td>
<td>Distribution of minisatellite haplotype diversities obtained from the permutation procedure under a random distribution of isolates to populations.</td>
<td></td>
</tr>
<tr>
<td>B.4</td>
<td>Results from rejection sampling procedure in RS.</td>
<td></td>
</tr>
<tr>
<td>B.5</td>
<td>Validation results for rejection sampling procedure.</td>
<td></td>
</tr>
<tr>
<td>C.1</td>
<td>Enrichment of genic and non-genic SNPs for values of $</td>
<td>\delta</td>
</tr>
</tbody>
</table>
C.2 Enrichment of genic and non-genic SNPs for values of $\delta$ between selected pairs of populations. 160
C.3 Enrichment of genic and non-genic SNPs for values of $\delta$ between the HGDP Yoruba and HapMap Tuscans. 161
C.4 Extreme values of SNP differentiation over all SNPs vs. mean $F_{ST}$ for all pairs of populations. 162
C.5 P-values for the 5 most extreme 100 kb genomic windows according to the iHS statistic in each population. 163
C.6 P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH CEU statistic in each population. 164
C.7 P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH YRI statistic in each population. 165
C.8 P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH KHB statistic in each population. 166
C.9 Number of shared 100 kb genomic windows in the top empirical 0.1% vs. mean $F_{ST}$ for pairs of populations. 167
C.10 Number of shared 100 kb genomic windows in the top empirical 1% vs. correlations of window statistics over all genomic windows for pairs of populations. 168
C.11 Correlations of window statistics over all 100 kb genomic windows vs. average Bantu ancestry for pairs of populations. 169
C.12 Demographic model simulated using msms. 170
C.13 Number of overlapping 100 kb windows in the empirical tails of XP-EHH for pairs of populations simulated under neutrality. 171
C.14 Number of overlapping 100 kb windows in the empirical tails of iHS for pairs of populations simulated under neutrality. 172
C.15 Number of overlapping 100 kb windows and true positive rate in the empirical tails of XP-EHH for pairs of populations simulated under recent selection. 173
C.16 Number of overlapping 100 kb windows and true positive rate in the empirical tails of iHS for pairs of populations simulated under recent selection. 174
C.17 Number of overlapping 100 kb windows and true positive rate in the empirical tails of XP-EHH for pairs of populations simulated under less recent selection. 175
C.18 Number of overlapping 100 kb windows and true positive rate in the empirical tails of iHS for pairs of populations simulated under less recent selection. 176
Introduction

Organismic evolution involves both selective and neutral forces, although their relative contributions are often unknown. Accurate interpretation of genetic data requires a consideration of the two, their interaction, and their underlying causes. In this thesis, I use novel statistical methods to analyze genetic data from a variety of organisms, including yeast, *Mycobacterium tuberculosis*, and humans. The chapters of this thesis provide complimentary perspectives on the relative roles of selection and neutrality, from the molecular to the population level, and present various statistical tools for genetic data analysis.

Natural selection acts on phenotypes, which are shaped by complex interactions among genes and between genes and the environment. Understanding how these interactions affect phenotypes can improve our understanding of selection. Chapter 2 of this thesis, *On the Classification of Genetic Interactions*, proposes a new method with which to classify and identify interactions, or epistasis, between pairs of genes. Given a set of single and double mutant phenotypes, various mathematical definitions of epistasis have been used in population genetics. We suggest that various models of genetic interaction can be considered simultaneously in a genetic system, and develop a method to classify genetic interactions using a likelihood framework and model selection criteria. Our method has high accuracy in identifying genetic interactions of various types, and when applied to experimental data collected in yeast (*Saccharomyces cerevisiae*), identifies novel genetic interactions. Our method provides an alternative approach for studying epistasis, providing another view of how selection may act on phenotypes.

Population-level studies, incorporating genetic data from a large number of members of a species, can provide a wealth of information about that species’ evolutionary history. While understanding natural selection is often an aim of such studies, demographic analyses, especially in the case of pathogens, can be equally enlightening. In Chapter 3, *Dispersal of Mycobacterium tuberculosis via the Canadian fur trade*, we study genetic data from *Mycobacterium tuberculosis*, the obligate pathogen that causes human tuberculosis (TB). With a goal of understanding its spread, we examine *M. tuberculosis* isolated from several contemporaneous Aboriginal Canadian populations which experienced severe TB epidemics in the late 19th century. We use a variety of computational approaches to compare genetic samples of *M. tuberculosis* between populations and develop an Approximate Bayesian Computation procedure to estimate parameters of a demographic expansion of *M. tuberculosis* coincident with epidemics. In so doing, we find support for spread of *M. tuberculosis* to indigenous communities via the fur trade in the 18th and 19th centuries. In our analysis of *M. tuberculosis* migration, we are able to shed light on the dynamics of pathogen spread and the effect of a pathogen’s evolutionary history on its host.

Along with demographic forces, natural selection also leaves its signature on the genetics of populations. Identifying putative signatures of selection using computational methods applied to genome-wide data can
CHAPTER 1. INTRODUCTION

generate hypotheses about a population’s underlying selective pressures. In Chapter 4, *Limited Evidence for Classic Selective Sweeps in African Populations*, we search for signatures of selection in the genomes of 12 diverse African human populations. First, to interpret the genome-wide scans, we compare genomic regions identified to be under selection between populations. Using novel statistical approaches, we find evidence for false positive signals of selection, the presence of which may bias inter-population comparisons if demographic history is not thoroughly considered. In Chapter 4, we also develop a novel permutation approach to identify biological functions that are significantly associated with regions of the genome under selection. Improving upon previous methods, we account for a number of genomic factors that may have biased the conclusions of previous analyses. While we identify several interesting signals of selection in African populations, our analyses underscore the importance of accurately specifying null models and neutral evolutionary history when searching for signatures of selection.

Finally, together with genetic data, data on phenotypic variation can provide insight into the molecular basis and evolutionary history of complex phenotypes. In Chapter 5, *Evolution of Height and Skin Pigmentation in the ≠Khomani Bushmen*, we study a phenotypically and genetically diverse KhoeSan population of the Kalahari Desert of South Africa, believed to be one of the world’s most ancient human populations. Using phenotypic data from over 100 individuals and corresponding genetic data from almost 90 individuals, we perform statistical analyses to characterize the genetic basis of their light skin pigmentation and short stature. We find that continental ancestry strongly affects skin pigmentation variation, and in a genome-wide association study, identify several loci of large effect that contribute to variation in stature in the ≠Khomani Bushmen. While we do not arrive at definitive conclusions regarding the evolutionary history of human pigmentation and height, our study demonstrates the utility of studying phenotypic and genetic data from diverse endogamous human populations.

In conclusion, these studies emphasize that a more complete understanding of the evolutionary history of humans and other organisms requires not only the consideration of neutral and selective processes, but also both phenotypic and genetic information. The statistical methods and approaches presented in the following chapters have the potential to improve inferences of natural selection and demography from population genetic data, as well as provide insight into the relative roles of both.

Development of the likelihood framework for identifying and classifying epistatic interactions in Chapter 2 was performed by both myself and Dr. Hong Gao, my collaborator on the project. Dr. Gao coded and implemented the method, and carried out the simulations and associated analyses to test the method’s performance. I applied our method to previously-published datasets of yeast phenotypes, validated the method in these datasets, and carried out subsequent interpretations of analyzed data. Dr. Gao and I co-wrote and revised the manuscript. Each of us wrote the sections of the manuscript relevant to our work, and with input from Dr. Marc Feldman I primarily wrote the Introduction and Discussion.

All analyses and interpretations of Chapter 3 were performed by both myself and Dr. Caitlin Pepperell, my collaborator on the project. Data was collected by Dr. Pepperell and other collaborators. I performed a majority of the statistical analyses of *M. tuberculosis* genetic data: performing simulations to calibrate a
mutation rate, calculating all genetic timing estimates, calculating measures of diversity, implementing and analyzing the Approximate Bayesian Computation procedure for demographic modeling, and calculating genetic estimates of migration. Other statistical analyses of genetic networks, Analysis of Molecular Variance (AMOVA), the map of Figure 3.1, and all historical analyses were performed by Dr. Pepperell. I wrote the sections of the Methods, Supplemental Methods, Results, and Supplemental Results that were relevant to my own work. While both Dr. Pepperell and I interpreted all results, Dr. Pepperell wrote a majority of the main manuscript, with substantial input from me.

In Chapter 4, I performed almost all statistical analyses, with input from Dr. Brenna Henn and other co-authors. Preliminary calculations of XP-EHH were performed by Chris Gignoux, and iHS calculations were performed by Dr. Jeffrey Kidd. Dr. Kidd also provided several scripts and analysis tools, including those for making summary plots of genomic regions under selection colored by p-values and annotated by their genes. Dr. Henn provided assistance regarding interpretation of results. I wrote the manuscript with input from all co-authors; Dr. Henn and Dr. Feldman also provided extensive edits.

In Chapter 5, the analyzed data were collected by a number of individuals over a number of years. Samples were collected by Dr. Brenna Henn and Dr. Gignoux in 2006; additional samples and phenotypic measurements were collected by me (2010 and 2011), Dr. Henn (2010), Dr. Marlo Möller (2011), and Dr. Cedric J. Werely (2011) of the University of Stellenbosch. Interviews for age verification on a subset of individuals were conducted by Justin Myrick (2012). DNA extraction was performed by individuals in the group of Dr. Eileen Hoal. I performed all analyses, with the exception of identity-by-descent analyses for estimation of relatedness (performed by Dr. Henn). I wrote the entire chapter, with input from Dr. Henn and Dr. Feldman.
On the Classification of Epistatic Interactions

Hong Gao*, Julie M. Granka†, Marcus W. Feldman§

* Department of Genetics, Stanford University, Stanford, CA, USA
§ Department of Biology, Stanford University, Stanford, CA, USA
† Equal contributions to this work.


Abstract

Modern genome-wide association studies are characterized by the problem of “missing heritability.” Epistasis, or genetic interaction, has been suggested as a possible explanation for the relatively small contribution of single significant associations to the fraction of variance explained. Of particular concern to investigators of genetic interactions is how to best represent and define epistasis. Previous studies have found that the use of different quantitative definitions for genetic interaction can lead to different conclusions when constructing genetic interaction networks and when addressing evolutionary questions. We suggest that instead, multiple representations of epistasis, or epistatic “subtypes,” may be valid within a given system. Selecting among these epistatic subtypes may provide additional insight into the biological and functional relationships among pairs of genes. In this study, we propose maximum likelihood and model selection methods in a hypothesis-testing framework to choose epistatic subtypes which best represent functional relationships for pairs of genes based on fitness data from both single and double mutants in haploid systems. We gauge the performance of our method with extensive simulations under various interaction scenarios. Our approach performs reasonably well in detecting the most likely epistatic subtype for pairs of genes, as well as in reducing bias when estimating the epistatic parameter ($\epsilon$). We apply our approach to two available datasets from yeast (Saccharomyces cerevisiae), and demonstrate through overlap of our identified epistatic pairs with experimentally-verified interactions and functional links that our results are likely of biological significance in understanding interaction mechanisms. We anticipate that our method will improve detection of epistatic interactions and will help to unravel the mysteries of complex biological systems.
2.1 Introduction

Understanding the nature of genetic interactions is crucial to obtaining a more complete picture of complex biological systems and their evolution. The discovery of genetic interactions has been the goal of many researchers studying a number of model systems, including but not limited to *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Escherichia coli* [38, 39, 60, 65, 131, 197, 235, 245, 255, 279, 312, 316]. Recently, high throughput-experimental approaches, such as epistatic mini-array profiles (E-MAPs) and genetic interaction analysis technology for *E. coli* (GIANT-coli), have enabled the study of epistasis on a large scale [49, 50, 240, 241, 282]. However, it remains unclear whether the computational and statistical methods currently in use to identify these interactions are indeed the most appropriate.

The study of genetic interaction, or “epistasis,” has had a long and somewhat convoluted history. Bateson [16] first used the term “epistasis” to describe the ability of a gene at one locus to “mask” the mutational influence of a gene at another locus [54]. The term “epistacy” was later coined by Fisher [78] to denote the statistical deviation of multilocus genotype values from an additive linear model for the value of a phenotype [205, 206].

These origins are the basis for the two main current interpretations of epistasis. The first, as introduced by Bateson [16], is the “biological,” “physiological,” or “compositional” form of epistasis, concerned with the influence of an individual’s genetic background on an allele’s effect on phenotype [47, 54, 176, 205, 206]. The second interpretation, attributed to Fisher, is “statistical” epistasis, which in its linear regression framework places the phenomenon of epistasis in the context of a population [176, 206, 292, 293, 302]. Each of these approaches is equally valid in studying genetic interactions; however, confusion still exists about how to best reconcile the methods and results of the two [9, 54, 158, 176, 205, 206].

Aside from the distinction between the statistical and physiological definitions of epistasis, there exist inconsistencies when studying solely physiological epistasis. For categorical traits, physiological epistasis is clear as a “masking” effect. When non-categorical or numerical traits are measured, epistasis is defined as the deviation of the phenotype of the multiple mutant from that expected under independence of the underlying genes.

The “expectation” of the phenotype under independence, that is, in the absence of epistasis, is not defined consistently between studies. For clarity, consider epistasis between pairs of genes, and without loss of generality, consider fitness as the phenotype. The first commonly-used definition of independence, originating from additivity, defines the effect of two independent mutations to be equal to the sum of the individual mutational effects. A second, motivated by the use of fitness as a phenotype, defines the effect of the two mutations as the product of the individual effects [64, 69, 206]. A third definition of independence has been referred to as “Minimum,” where alleles at two loci are independent if the double mutant has the same fitness as the less-fit single mutant. Mani et al. [165] claim that this has been used when identifying pairwise epistasis by searching for synthetic lethal double mutants [59, 197, 198, 278, 279]. A fourth is the “Log” definition presented by Mani et al. [165] and Sanjuan and Elena [236]. The less-frequently used “scaled $\epsilon$” [245] measure of epistasis takes the multiplicative definition of independence with a scaling factor.
CHAPTER 2. ON THE CLASSIFICATION OF EPISTATIC INTERACTIONS

These different definitions of independence are partly due to distinct measurement “scales.” For some traits, a multiplicative definition of independence may be necessary to identify epistasis between two genes, whereas for other traits, additivity may be appropriate [75, 165, 206, 292]. An interaction found under one independence definition may not necessarily be found under another, leading to different biological conclusions [165].

Mani et al. [165] suggest that there may be an “ideal” definition of independence for all gene pairs for identifying functional relationships. However, it is plausible that different representations of independence for two genes may reflect different biological properties of the relationship [146, 229]. “Two categories of general interest [the additive and multiplicative definitions, respectively] are those in which etiologic factors act interchangeably in the same step in a multistep process, or alternatively act at different steps in the process,” [229, p. 468]. In some cases, the discovery of “epistasis” may merely be an artifact of using an incorrect null model [146]. It may be necessary to represent “independence” differently, resulting in different statistical measures of interactions, for different pairs of genes depending on their functions.

Previous studies have suggested that different pairs of loci may have different modes of interaction and have attempted to sub-classify genetic interactions into regulatory hierarchies and mutually exclusive “interaction subtypes” to elucidate underlying biological properties [8, 65, 255]. We suggest that epistatic relationships can be divided into several “subtypes,” or forms, corresponding to the aforementioned definitions of independence. As a particular gene pair may deviate from independence according to several criteria, we do not claim that these subtypes are necessarily mutually exclusive. We attempt to select the most likely epistatic subtype which is the best statistical representation of the relationship between two genes. To further sub-classify interactions, epistasis among deleterious mutations can take one of two commonly used forms: positive (equivalently; alleviating, antagonistic, or buffering) epistasis, where the phenotype of the double mutant is less severe than expected under independence; and negative (equivalently; aggravating, synergistic, or synthetic), where the phenotype is more severe than expected [49, 64, 165, 245].

Another objective of such distinctions is to reduce the bias of the estimator of the epistatic parameter ($\epsilon$), which measures the extent and direction of epistasis for a given gene pair. Mani et al. [165], assuming that the overall distribution of $\epsilon$ should be centered around 0, find that inaccurately choosing a definition of independence can result in increased bias when estimating $\epsilon$. For example, using the minimum definition results in the most severe bias when single mutants have moderate fitness effects, and the additive definition results in the largest positive bias when at least one gene has an extreme fitness defect [165]. Therefore, it is important to select an optimal estimator for $\epsilon$ for each pair of genes from among the subtypes of epistatic interactions.

Epistasis may be important to consider in genomic association studies, as a gene with a weak main effect may only be identified through its interaction with another gene or other genes [55, 56, 83, 174, 175]. Epistasis has also been studied extensively in the context of the evolution of sex and recombination. The mutational deterministic hypothesis proposes that the evolution of sex and recombination would be favored by negative epistatic interactions [76, 144]; many other studies have also studied the importance of the form
of epistasis [38, 64, 69, 141, 163, 196]. Indeed, according to Mani et al. [165, p. 3466], “the choice of
definition [of epistasis] alters conclusions relevant to the adaptive value of sex and recombination.”

Given fitness data from single and double mutants in haploid organisms, we implement a likelihood
method to determine the subtype which is the best statistical representation of the epistatic interaction for
pairs of genes. We use maximum likelihood estimation and the Bayesian Information Criteria (BIC) [244]
with a likelihood ratio test to select the most appropriate null or epistatic model for each putative interac-
tion. We conduct extensive simulations to gauge the performance of our method, and demonstrate that it
performs reasonably well under various interaction scenarios. We apply our method to two datasets with fit-
ness measurements obtained from yeast [131, 255], whose authors assume only multiplicative epistasis for all
interactions. By examining functional links and experimentally-validated interactions among epistatic pairs,
we demonstrate that our results are biologically meaningful. Studying a random selection of genes, we find
that minimum epistasis is more prevalent than both additive and multiplicative epistasis, and that the overall
distribution of $\epsilon$ is not significantly different from zero (as Jasnos and Korona [131] suggest). For genes in
a particular pathway, we advise selecting among fewer epistatic subtypes. We believe that our method of
epistatic subtype classification will aid in understanding genetic interactions and their properties.

2.2 Methods

We consider epistasis among $L$ genes (represented by $g_1, g_2, \ldots, g_L$) given the phenotypes of each of the
possible single and double mutant haplotypes, each with a given number of replicates. The phenotype we
study is fitness scaled by wildtype, obtained either through experiments or simulation [49, 50, 131, 240,
241, 245, 255]. The fitness of the wildtype genotype $g_0$ is 1, and the relative fitnesses of single mutants $g_i$
($1 \leq i \leq L$) and double mutants $g_{ij}$ ($1 \leq i < j \leq L$) are denoted by $w_i$ and $w_{ij}$, respectively. Supposing
that we have $N$ replicates of fitness values for each mutant, the superscript $n$ represents measurements in the
$n$-th replicate ($1 \leq n \leq N$). We denote by $\sigma_i^2$ and $\sigma_{ij}^2$ the variances in fitnesses of single mutants $g_i$
and double mutants $g_{ij}$, respectively. We define $\epsilon_{ij}$ as the epistatic coefficient of a genetic interaction between
genes $g_i$ and $g_j$.

2.2.1 Epistatic Subtype Selection

Pairwise physiological epistasis for measured numerical traits is defined as the deviation of the observed
phenotype of the double mutant from the phenotype expected when the alleles at each locus act independently.
As mentioned above, the additive and multiplicative models are widely used to represent independence, while
the minimum model is used by investigators searching for synthetic lethal double mutants [165]. Thus, we
focus predominantly on these three models, as do Mani et al. [165]. The “Log” model is practically equivalent
to the multiplicative model for deleterious mutations [165]; since we generally focus on deleterious mutations,
we do not consider this model. We do not include a model based on the “scaled $\epsilon$” measure of epistasis [245],
as it was found to be less meaningful than the multiplicative model in describing interactions [255].
CHAPTER 2. ON THE CLASSIFICATION OF EPISTATIC INTERACTIONS

We thus use six models: the three models of independence (additive, multiplicative, and minimum) and the three models (“subtypes”) of epistasis (additive, multiplicative, and minimum epistasis), which incorporate a deviation from the independence models. We have three major questions. First, when data follow one of the three models of independence, do we incorrectly infer the presence of epistasis? Second, can we detect epistasis between two genes if they interact? Third, can we accurately determine which “subtype” is the most appropriate for a particular pair of genes if they interact?

The three null models are as follows. If genes $g_i$ and $g_j$ act additively, i.e., there is no additive epistasis, the null model for the double mutant fitness in the $n$-th experiment ($w_{ij}^n$) is a normal distribution with mean $w_i^n + w_j^n - 1$ and variance $\sigma_{ij}^2$. If $g_i$ and $g_j$ act multiplicatively, $w_{ij}^n$ is a normal random variable with mean $w_i^n w_j^n$ and variance $\sigma_{ij}^2$. The third model specifying the absence of epistasis assumes that $w_{ij}^n$ is normal with mean $\min(w_i^n, w_j^n)$ and variance $\sigma_{ij}^2$. We emphasize that the minimum model described here is a model of independence.

In the presence of additive epistasis $\epsilon_{ij}$, the probability distribution of the double mutant fitness $w_{ij}$ in the $n$-th experiment, given $w_i^n, w_j^n$, and variance $\sigma_{ij}^2$, is:

$$\mathbb{P}(w_{ij}^n = x \mid w_i^n, w_j^n, \epsilon_{ij}, \sigma_{ij}^2) \sim N(w_i^n + w_j^n - 1 + \epsilon_{ij}, \sigma_{ij}^2)$$ (2.1)

When $g_i$ and $g_j$ exhibit multiplicative epistasis $\epsilon_{ij}$, $w_{ij}^n$ has probability distribution:

$$\mathbb{P}(w_{ij}^n = x \mid w_i^n, w_j^n, \epsilon_{ij}, \sigma_{ij}^2) \sim N(w_i^n w_j^n (1 + \epsilon_{ij}), \sigma_{ij}^2)$$ (2.2)

This form of multiplicative epistasis is slightly different from that of other authors, who suggest $w_{ij} = w_i w_j + \epsilon_{ij}$ [131, 165, 245, 255]; our formulation is simply a rescaling of this formula. Our motivation for Equation 2.2 is that it relates more closely to the large body of research representing the fitness of an individual with $k$ mutations as $w_k = e^{-sk}$ [64, 69, 196, 302].

In the case of minimum epistasis, $w_{ij}^n$ is also distributed normally:

$$\mathbb{P}(w_{ij}^n = x \mid w_i^n, w_j^n, \epsilon_{ij}, \sigma_{ij}^2) \sim N(\min(w_i^n, w_j^n) + \epsilon_{ij}, \sigma_{ij}^2)$$ (2.3)

An epistatic “subtype” refers to one of the three epistatic models shown in the equations above.

To calculate the likelihood of the observed double mutant fitnesses, we assume that each pair of genes is independent of all other pairs, i.e., there are no interactions of higher order. The log likelihood of the data based on $N$ replicates is

$$\log L = \sum_{n=1}^{N} \sum_{i<j: 1 \leq i < j \leq L} \log \mathbb{P}_E(w_{ij}^n = x \mid w_i^n, w_j^n, \epsilon_{ij}, \sigma_{ij}^2) * I_{\{g_i \text{ interacts with } g_j\}} + \log \mathbb{P}_N(w_{ij}^n = x \mid w_i^n, w_j^n, \sigma_{ij}^2) * (1 - I_{\{g_i \text{ interacts with } g_j\}}),$$ (2.4)

where $I_{\{\Phi\}}$ is an indicator function and is equal to one if $\Phi$ is true and zero otherwise, and $\mathbb{P}_E$ and $\mathbb{P}_N$ are the probabilities (given above) under the epistatic and null models, respectively. Since we assume that each
2.2. METHODS

A pair of genes is independent of all other pairs, we study each separately. We use the Bayesian Information Criterion (BIC) [244], a commonly used model selection criterion, to determine the most likely interaction subtype, if any, of each pair. To calculate the BIC, we first estimate the parameters $\epsilon_{ij}$ and $\sigma^2_{ij}$ using Maximum Likelihood Estimation (MLE). For additive epistasis, the ML estimators (MLEs) for $\epsilon_{ij}$ and $\sigma^2_{ij}$ are

$$
\hat{\epsilon}_{ij}^{(a)}_{\text{MLE}} = \bar{w}_{ij} - \bar{w}_i - \bar{w}_j + 1
$$
$$
\hat{\sigma}^2_{ij}^{(a)}_{\text{MLE}} = \frac{1}{N} \sum_{n=1}^{N} (w_{ij}^n - w_i^n - w_j^n + 1 - \hat{\epsilon}_{ij}^{(a)}_{\text{MLE}})^2,
$$

(2.5)

where $\bar{w}_*$ denotes the mean fitness of genotype $g$ over all $N$ replicates, and the right superscript $(a)$ indicates that the estimator is for additive epistasis. The MLEs for multiplicative and minimum epistasis are given in Appendix A.

We select the model that is the best statistical representation of the interaction for each pair of genes using the BIC, defined as

$$
\text{BIC} = n_p \log(N) - 2 \log L_M
$$
$$
= n_p \log(N) + N (1 + \log (2\pi \sigma^2_{ij,\text{MLE}})),
$$

(2.6)

where $L_M$ and $n_p$ are the maximum likelihood and the number of free parameters, respectively, in the given model. For the three null models, $n_p$ is one, as only $\sigma^2_{ij}$ is estimated; for the three epistatic models, $n_p = 2$, as both $\sigma^2_{ij}$ and $\epsilon_{ij}$ are estimated. The model with the smallest BIC is selected as optimal, rewarding models with a higher likelihood but penalizing for additional parameters.

For each pair, we perform a likelihood ratio test (LRT) of the epistatic model with the lowest BIC against its corresponding null model as an optional step to assess the significance of epistasis. The LRT statistic is given as

$$
\lambda = -2 [\log L_{M_{\text{null}}} - \log L_{M_{\text{epistatic}}}].
$$

(2.7)

Under the null distribution, $\lambda$ is approximately $\chi^2$ distributed with one degree of freedom (for the one additional parameter $\epsilon$ estimated in the epistatic model).

When identifying epistasis for many gene pairs, the problem of simultaneously testing multiple hypotheses arises. We consider three commonly-used multiple-testing corrections, which use the p-values obtained from the LRT above. The first is the Bonferroni correction, which sets the chosen significance level divided by the number of tests (in our case, the number of putatively epistatic gene pairs) as a p-value cutoff. This is generally considered to be the most conservative procedure to control the family-wise error rate (the probability of rejecting at least one true null hypothesis among all tests). The second multiple-testing correction, proposed by Benjamini and Hochberg [26], controls the false discovery rate (FDR, or the expected proportion of null hypotheses that are falsely rejected). Third, we control the positive false discovery rate (pFDR), which is slightly more lenient than the FDR procedure, as suggested by Storey [258]. We advocate determining significance using the FDR, which is intermediate between the other two procedures (Sections 2.3.1 and 2.4). We implement our methods in R (http://www.r-project.org, code available upon request).
2.2.2 Generation of Simulated Data

We carried out extensive simulations to assess the performance of our method. For simplicity, we assume that each mutation is deleterious, i.e., the fitnesses of single mutants cannot be greater than that of wild type ($w_0 = 1$). (In Appendix A and Table A.1, we consider relaxing this assumption to include beneficial mutations). We sample the fitnesses of single mutants $g_i$ and $g_j$ from a truncated normal distribution on $[0, 1]$ with mean $\mu$ and standard deviation $\sigma$, i.e., $w_i \sim N(\mu_i, \sigma^2_i)_{[0,1]}$ and $w_j \sim N(\mu_j, \sigma^2_j)_{[0,1]}$. Although many deletion strains found in nature have a wildtype fitness, we simulate with this support in order to encompass the entire range of fitness values possible for deleterious mutants. We choose the support for double mutant fitnesses to be $[0, 1.2]$; since both synergistic and antagonistic epistasis may be possible, the fitness of double mutants may be greater than 1. We simulate $w_{ij}$ given $w_i$ and $w_j$ following the definitions presented in Section 2.2.1. For example, if genes $g_i$ and $g_j$ act additively and independently, $w_{ij} \sim N(w_i + w_j - 1, \sigma^2_{ij})_{[0,1,2]}$; under additive epistasis, $w_{ij} \sim N(w_i + w_j - 1 + \epsilon_{ij}, \sigma^2_{ij})_{[0,1,2]}$. Distributions under multiplicative and minimum null and epistatic models are given in Appendix A.

2.2.3 Analysis of Experimental Data

We analyze two datasets of fitness measurements in yeast (*Saccharomyces cerevisiae*) [131, 255]. In addition to choosing the appropriate epistatic subtype based on the BIC, we perform a LRT to test epistatic models against their corresponding null models and correct for multiple testing with the three methods described in Sections 2.2.1. For all tests, we select an $\alpha$ significance level of 0.05.

The first dataset of St Onge et al. [255] included 26 non-essential genes known to confer resistance to the DNA-damaging agent methyl methanesulfonate (MMS), and contained double-deletion strains for all pairs that were possible to construct (323 double mutant strains). The authors measured the fitnesses of single and double mutants in media both with and without MMS. For both media types, while single mutants have approximately 10 replicate fitness measurements, double mutants have only two. Since our method assumes the same number of replicates for single and double mutants, we perform our method with two sampled single mutant fitness measurements. We repeat this sampling 1000 times, and select an epistatic subtype for the pairs that were chosen as epistatic in at least 900 out of the 1000 replicates (see Appendix A for details).

In the second study, Jasnos and Korona [131] performed 639 random crosses among 758 randomly selected gene deletions to create a series of double mutants and assayed the growth fitnesses of each single and double mutant. As an equal number of replicates are available for both single and double mutants ($n \approx 5$), we apply our method to this dataset directly.

We perform several analyses to validate the use of our method. First, we examine the number of epistatic pairs sharing “specific” functional links, a term used by both St Onge et al. [255] and Tong et al. [279]. St Onge et al. [255] define these to be genes that share Gene Ontology (GO) terms [270] that have associated with them fewer than 30 genes; Tong et al. [279] define the cutoff for specificity to be less than or equal to 200 associated genes. We examine our data using these and several intermediate cutoffs. Second, we
count epistatic pairs sharing the broader GO Slim terms (obtained from the Yeast Genome Database, http://www.yeastgenome.org, accessed 2009-09-11). To assess the significance of the number of specific functional links or shared GO Slim terms among inferred epistatic pairs, we implement both Fisher’s Exact Test and a permutation procedure (Table A.6), which produce comparable results. Third, we examine overlap of inferred epistatic pairs with experimentally-verified physical interactions in yeast (BIOGRID database, [257], http://www.thebiogrid.org, Release 2.0.56, see Appendix A).

2.3 Application to Simulated Data

2.3.1 Estimation of Epistatic Measures

We perform 10,000 simulations following Section 2.2.2 under each of the six models (three null models and three epistatic models) for each set of parameters shown in Table 2.1. For each simulation, we sample 50 fitness replicates ($N = 50$), obtain the BIC for each of the six models, and select the model with the lowest BIC. We compute accuracy as the proportion of simulations in which the model under which the data are simulated is selected (see also Tables A.1, A.2, and A.3).

Table 2.1 demonstrates that our method can select the correct null or epistatic model for various parameter combinations. When simulating under one of the three null models, our method identifies the correct model with almost 100% accuracy. Overall, the minimum null model is identified with higher accuracy than either the additive or multiplicative null models. Identification of the minimum epistatic subtype is consistently the most accurate (near 100%), while the accuracy of selecting the additive and multiplicative epistatic subtypes varies with the epistatic coefficient $\epsilon_{ij}$.

If two single mutants have the same mean fitness, our method is less able to detect multiplicative epistasis than either additive or minimum epistasis (Table 2.1). Introducing a fairly large difference between single mutant fitness values slightly increases the accuracy of selecting multiplicative epistasis when data are simulated under this definition. For all epistatic models, the probability of selecting the appropriate model generally increases with the absolute value of $\epsilon_{ij}$: the larger the absolute value of $\epsilon_{ij}$, the larger is the difference between the expected double mutant fitness under the correct model and the other models. We also have little ability to consistently distinguish between the additive and multiplicative models. See also Figures A.1 and A.2.

Testing hypotheses with the LRT and FDR procedure [26] gives nearly identical accuracy to that described above (results not shown). With data simulated under the null model, as expected, the FDR slightly decreases the false positive rate; for data simulated under an epistatic model, power decreases by at most 1%. Although the false positive rate decreases when using the Bonferroni correction, we obtain slightly lower power than with the FDR; the pFDR [258] does not greatly improve power over the FDR (results not shown). The FDR procedure thus seems preferable.
2.3.2 Bias and Variance of the Epistatic Parameter ($\epsilon$)

Mani et al. [165] show that using different definitions of independence results in different estimators of the epistatic parameter ($\epsilon$), some with larger biases and variances than others. To evaluate the bias and variance of $\hat{\epsilon}_{\text{MLE}}$ estimated with our method, we perform simulations and compute its mean squared error (MSE) under each epistatic model. Because the FDR procedure does not greatly affect the accuracy of our inference (see above), we do not implement it in these simulations. For the $n$-th simulation, $MSE_n = (\hat{\epsilon}_{\text{MLE}} - \epsilon_n)^2$, where $\epsilon_n$ is the true value used in the $n$-th simulation and $\hat{\epsilon}_{\text{MLE}}$ is estimated under the specific epistatic model. If the simulated $|\epsilon|$ is large, the MSE of $\hat{\epsilon}_{\text{MLE}}$ is the smallest and is nearly zero when the true epistatic model is selected (Figure A.3). When the simulated $|\epsilon|$ decreases, we find that while the MSE is always small when the correct model is selected, at times the MSE is even lower under an incorrect model (Figure A.3). In general, our method seems to reduce the bias and variance of $\epsilon$.

2.4 Application to Experimental Data

Epistasis has been studied extensively in yeast (Saccharomyces cerevisiae), a model organism for which generating a large number of mutants and examining growth phenotypes is experimentally tractable. Of interest are studies measuring single and double mutant fitnesses on a continuous scale. With our proposed method, we re-analyze two datasets by detecting epistasis and its subtype for each pair of mutants [131, 255].

2.4.1 St Onge et al. [255] Dataset

St Onge et al. [255] examined 26 non-essential genes known to confer resistance to MMS, constructed double-deletion strains for 323 double mutant strains (all but two of the total possible pairs), and assumed the multiplicative form of epistasis for all interactions (Section 2.2.3). Following these authors, we focus on single and double mutant fitnesses measured in the presence of MMS. (For results in the absence of MMS, see Table A.4.)

Using the resampling method described in Section 2.2.3 and Appendix A, 222 gene pairs pass the cutoff of having epistasis inferred in at least 900 out of 1000 replicates. This does not include five synthetic lethal gene pairs. Hypothesis testing and a multiple-testing procedure (for 222 simultaneous hypotheses) are necessary to determine the final epistatic pairs.

To select one among the three multiple-testing procedures, we follow St Onge et al. [255] and examine gene pairs that share “specific” functional links (see Section 2.2.3). The Bonferroni method is likely too conservative, yielding only 25 significantly epistatic pairs with only one functional link among them; alternatively, the pFDR procedure appears to be too lenient in rejecting independence for all 222 pairs. Therefore, we use the FDR procedure (although the number of functional links is not significant) and detect 193 epistatic
pairs, of which 5 (2.6%) are synthetic lethal, 19 (9.8%) have additive epistasis, 33 (17.1%) have multiplicative epistasis, and 136 (70.5%) have minimum epistasis (Table 2.2, Table A.4). We find 29 gene pairs with positive (alleviating) epistasis, and 159 gene pairs with negative (aggravating) epistasis.

To further validate the use of our method and the FDR procedure, we assess by Fisher’s Exact Test the significance of an enrichment of both Biological Process and all GO Slim term links among epistatic pairs, neither of which are significant ([270]; http://www.yeastgenome.org; Table 2.3). Our discovered epistatic interactions do overlap with experimentally-verified physical interactions (BIOGRID database ([257]); Tables 2.3, A.5, and A.6). Although some of the previously unidentified interactions that we identify could be false positives, many are likely to be new discoveries.

The epistatic subtypes we consider are not necessarily mutually exclusive. To more fully assess the assumptions of our method, we also consider several of the possible subsets of the epistatic subtypes (and their corresponding null models) in our procedure. As the minimum epistatic subtype was the most frequently selected in this data set, we first do not include the minimum null model nor the minimum epistatic model in our procedure (i.e., we select from among four rather than six models for a pair; Table 2.3, column 3). In Table 2.3, we also examine our method’s performance when it selects from only additive epistasis or independence (column 4), only multiplicative epistasis and independence (column 5), or only minimum epistasis and independence (column 6). Allowing for different subsets of subtypes does not greatly affect the number of inferred epistatic pairs with experimentally identified interactions (Table 2.3, Table A.5). However, there are a significant number of epistatic pairs with functional links only when the minimum epistatic subtype is not included (also see Table A.5). It is not immediately clear which epistatic subtypes are the most appropriate for these data, although including the minimum subtype may not be appropriate [165] (Section 2.5).

Although it may be best to consider fewer epistatic subtypes for this specific dataset, we report our results including all three epistatic subtypes and their corresponding null models (Table 3). Our method identifies epistasis for 88 of the same pairs as St Onge et al. [255], although we identify 105 epistatic pairs not identified by the original authors (Figure A.4, Table A.5). St Onge et al. [255] find that epistatic pairs with a functional link have a positively-shifted distribution of epistasis. We find no such shift in epistasis values (Table 2.2, Figure A.5). We also demonstrate (described in Section 2.3.2) that our method seems to reduce bias of the epistatic parameter ($\epsilon$) (Table 2.2). (Note that in the absence of MMS, we find that $\tau$ is negative and significantly different from 0; Table A.4.) When considering only a subset of the epistatic subtypes, however, we find $\tau$ to be positive and significantly different from zero (results not shown). See Appendix A and Figure A.6 for additional discussion of the epistatic pairs we identify.

### 2.4.2 Jasnos and Korona [131] Dataset

The Jasnos and Korona [131] dataset included 758 yeast gene deletions known to cause growth defects and reports fitnesses of only a sparse subset of all possible gene pairs ($\approx 0.2%$ of the possible pairwise genotypes, or 639 pairs out of $\binom{758}{2}$). Because the authors do not identify epistatic pairs in a hypothesis-testing framework, we cannot explicitly compare our conclusions with theirs.
To validate our method, we examine gene pairs which have “specific” functional links (Section 2.2.3). When defining a functional link using GO terms [270] with fewer than 30 genes associated with them, only 1 of 639 tested gene pairs has a functional link. Raising the threshold of associated genes to 50 and 100, the number of tested pairs with functional links rises only to 3 and 9, respectively. Because of the large number of random genes and sparse number of gene pairs in this dataset, we follow Tong et al. [279] and select GO terms that have associated with them less than or equal to 200 genes. 25 of 639 tested pairs then have a functional link.

Only the FDR multiple-testing procedure results in a significant enrichment of functional links among epistatic pairs (Table 2.3); thus, it is likely the most appropriate of the three multiple-testing methods. With the FDR procedure we find 352 significant epistatic pairs, of which 35 (9.9%) have additive epistasis, 63 (17.9%) have multiplicative epistasis, and 254 (72.2%) have minimum epistasis (Table 2.2). These proportions of inferred subtypes suggest that the authors’ original restriction to multiplicative epistasis may be inappropriate. We find 141 gene pairs with positive epistasis, and 211 gene pairs with negative epistasis.

We do not find a significant number of epistatic pairs with shared GO Slim Biological Process terms (Section 2.2.3), but do when considering all shared GO Slim terms (Table 2.3). However, the significant enrichment of functional links among epistatic pairs suggests that the epistasis found with the BIC and FDR procedures is valid and of biological significance. We also find an overlap of epistatic pairs with experimentally-validated physical interactions from the BIOGRID database).

As with the St Onge et al. [255] dataset, we also consider some of the possible subsets of the three epistatic subtypes (and their corresponding null models) in our model selection procedure (Table 2.3). In contrast to the St Onge et al. [255] dataset, using all three epistatic subtypes results in a significant number of epistatic pairs with functional links; this measure is not significant when using any of the other subsets of the subtypes. This suggests that our proposed method with three epistatic subtypes may indeed be the most appropriate for datasets with randomly selected genes.

We examined the distribution of the estimated values of the epistatic parameter ($\epsilon$) for all pairs with significant epistasis. Jasnos and Korona [131], in assuming only multiplicative epistasis, conclude that epistasis is predominantly positive. However, we find that the estimated mean of epistasis is not significantly different from zero (two-sided t-test, p-value = 0.9578; Figure 2.1 and Table 2.2). This also supports that our method reduces bias of the estimator of $\epsilon$. We report in further detail on the epistatic pairs identified in Appendix A and Figure A.7.

2.5 Discussion

We present a likelihood estimation method using the Bayesian Information Criterion (BIC) which, given fitness data from single and double mutants, determines the most appropriate epistatic subtype for pairs of genes that interact. Through extensive simulations, we demonstrate that selection of the epistatic subtype is extremely important, and that our method can indeed do so accurately. We demonstrate through the analysis
of two datasets that the epistasis we discover is likely biologically meaningful. Besides the detection of
genetic interactions and clarification of their functions, accurate assessment of epistasis is also important for
evolutionary theory. For example, the distribution of epistatic effects has important implications for the role
of epistasis in the evolution of sex and recombination. We demonstrate that our method reduces bias of the
epistatic parameter $\epsilon$.

Our approach is novel for several reasons. While in reality different subtypes and representations of
epistasis may be best for different pairs of loci, classical studies have usually assumed only one subtype
of epistasis for all pairs. Our method selects the most likely subtype for pairs of genes, providing a more
precise picture of the loci actually exhibiting epistasis and insight into their modes of genetic interaction.
Second, some past studies of epistasis have not clearly distinguished between the epistatic and independence
models. Jasnos and Korona [131], for example, calculate values of $\epsilon$ without formally assessing whether the
perceived epistasis is merely statistical noise. With the presented BIC and LRT procedures, we distinguish
between the epistatic and null models in a hypothesis-testing framework and address the multiple-hypothesis
testing problem with the FDR [26].

Applying our method to two yeast datasets [131, 255], we detect both novel and experimentally-verified
epistatic interactions. We believe that an enrichment of “specific” functional links among epistatic pairs
is likely the most direct indication of biological importance of inferred epistasis ([255, 279]), as genetic
interactions probably occur on a fine biological scale (Section 2.2.3).

Including all three epistatic subtypes and their corresponding null models in our procedure results in a
significant enrichment of functional links among identified epistatic pairs for the Jasnos and Korona [131]
dataset, but not the St Onge et al. [255] dataset (both with and without MMS). Performance for the St Onge
et al. [255] dataset is improved when only the additive or multiplicative subtypes, or both, are considered
(Table 2.3). This apparent difference between datasets may be due to the fact that the St Onge et al. [255]
dataset includes genes known to be involved in conferring resistance to MMS, whereas the Jasnos and Korona
[131] dataset measured a random selection of genes and gene pairs. In examining epistasis among genes with
similar functions, it thus may be more appropriate to consider fewer epistatic subtypes. This difference
between the datasets may also be due to the low sample size for double mutants in the St Onge et al. [255]
dataset.

We find predominantly minimum epistasis, i.e., a deviation from the minimum null model, when con-
sidering all three subtypes in both datasets. This is likely valid for the Jasnos and Korona [131] dataset, but
it remains unclear which subtypes are the most appropriate for the St Onge et al. [255] dataset. Mani et al.
[165] suggest that the minimum definition of independence may not always be useful. Examining another
dataset with randomly selected genes and gene pairs will help to assess the performance of our method.

The slight overlap of epistatic pairs inferred using different subsets of the three subtypes emphasizes that
the subtypes we select are not necessarily mutually exclusive (Table 2.3). We claim to select the statistically
best-fitting epistatic subtype for pairs of genes in order to gain additional insight into their mode of interaction.
For example, a double mutant can deviate from both additive and multiplicative independence. Selecting the
additive epistatic subtype with our method suggests that additive epistasis may be more appropriate and useful way to represent the relationship between the two genes. The different results obtained considering different subtypes again emphasizes the importance of selecting among them [165].

While Jasnos and Korona [131] find a skew towards positive epistasis, we find estimates of epistasis to be centered around zero (Table 2.2). For both datasets, however, the mean of additive epistasis (\(\epsilon\)) is positive and significantly different from 0, whereas the mean of minimum epistasis (\(\epsilon\)) is negative and significantly different from 0 (Table 2.2). Applying our method to datasets with a larger number of genes may provide a more accurate view of the types and strengths of epistasis often found in nature.

Inference about epistatic subtypes is affected by many factors, including the number of replicate experiments and the variance of mutant fitnesses. Accuracy clearly increases with the number of replicate measurements, but in practice less than 10 replications are performed [131, 255]. We assess our method’s accuracy when the number of replicates is decreased from 50 to 10 and to 2 (Tables A.2 and A.3). As expected, accuracy decreases, but varies depending on the type of epistasis examined (Appendix A). Our failure to identify epistasis for several experimentally-identified interacting pairs in the St Onge et al. [255] dataset is likely due to our low power with only two double-mutant fitness measurements.

A large variance in mutant fitnesses also decreases the accuracy of our method. Our simulations use 0.05 as the standard deviation of mutant fitnesses, which is the typical value found for single and double mutants in the Jasnos and Korona [131] dataset. To assess the effect of variance of fitness on accuracy, we ran simulations with the standard deviation increased to 0.1; this slightly reduces the ability to distinguish the true epistatic model from the others (Table 2.1).

We also address the approximation of the distribution of the LRT statistic (\(\lambda\)) to the \(\chi^2\) distribution under the null hypothesis. For both datasets, very few functional links exist among epistatic pairs inferred with the Bonferroni method. This indicates that epistatic pairs with functional links have somewhat larger p-values, and are only significant under less stringent multiple-testing protocols. The \(\chi^2\) approximation may not always be appropriate, as the measured fitnesses sometimes deviate from normality in both datasets (Appendix A, Figures A.8 and A.9). This may also explain our inability to detect some experimentally-validated interactions. We urge measuring a large number of replicates to meet the normality requirement and to increase the power of identifying epistatic interactions.

Our method is not limited to studying fitness as a phenotype and can be applied to any trait, although it is necessary to scale the phenotype of interest by the wildtype value. Our likelihood-based method is only applicable to studies of experimental data where multiple replicates are measured; it is not immediately applicable to studies of computationally-predicted fitness values, with only one measurement [245]. However, adding random noise terms to each predicted value to create multiple “replicates” could allow our method to be applied in the future. In addition, it is important to note that the epistasis identified with this method is a property of the alleles at each locus; it is conceivable that when examining other pairs of alleles at the tested loci, different results could be obtained. We also note that our method is only applicable to haploid systems. Identifying epistasis and constructing double mutants is much more difficult in diploid systems, as
more complex statistical issues arise due to dominance and other effects [79, 93, 156, 196, 293]. Since current methods to identify epistasis in haploid systems are still in their infancy, it is important to first develop and perfect these methods before addressing diploid systems.

Our method is one in a wide research field studying genetic interactions and in a long history of the study of epistasis [12, 48, 99, 226]. As mentioned above, our method leaves many open questions. Not all of the genetic interactions that we find may translate into physical interactions, and many researchers have pursued the question of how genetic and protein-protein interactions are related [152, 243, 306]. The translation of discoveries of epistasis into discoveries of functional interactions is a very intriguing, albeit complex, problem. We expect that our method will shed light on intrinsic properties of epistasis and may help to address these broader issues.

2.6 Acknowledgements

We thank Elhanan Borenstein for his helpful comments and suggestions, as well as two anonymous reviewers and the editor of Genetics. This research project is supported by National Institutes of Health (NIH) grant GM28016 to MWF, NIH grant RO1 GM073059 to Hua Tang, and a Stanford Dean’s postdoctoral fellowship to HG. JMG is supported by the Stanford Genome Training Program, which is funded by the NIH (academic year 2008-2009), and by a National Science Foundation (NSF) Graduate Research Fellowship (grant number DGE-0645962) (2009-2010).
2.7 Figures and Tables

Figure 2.1: Distribution of the epistasis values ($\epsilon$) for significant epistatic pairs in the Jasnos and Korona [131] dataset, determined using the FDR procedure and the BIC method including all three epistatic subtypes and their corresponding null models. Mean of $\epsilon$ is -0.0009, with a standard deviation of 0.3177; median value is -0.0587. A similar figure is shown in Figure 3 of Jasnos and Korona [131].
Table 2.1: Fractions of simulations that recover the true model (null or epistatic) among the six models (three null models and three epistatic models) with different values of the epistasis coefficient ($\epsilon_{ij}$), fitness standard deviation ($\sigma = \sigma_i = \sigma_j = \sigma_{ij}$), and mean fitnesses of single mutants ($\mu_i$) and ($\mu_j$).

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$\mu_i$</th>
<th>$\mu_j$</th>
<th>$P_N^{(a)}$</th>
<th>$P_N^{(p)}$</th>
<th>$P_N^{(m)}$</th>
<th>$\epsilon_{ij}$</th>
<th>$P_E^{(a)}$</th>
<th>$P_E^{(p)}$</th>
<th>$P_E^{(m)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.8</td>
<td>0.8</td>
<td>0.986</td>
<td>0.983</td>
<td>0.994</td>
<td>-0.3</td>
<td>0.996</td>
<td>0.906</td>
<td>0.929</td>
</tr>
<tr>
<td>0.05</td>
<td>0.8</td>
<td>0.8</td>
<td>-0.2</td>
<td>0.992</td>
<td>0.902</td>
<td>-0.2</td>
<td>0.992</td>
<td>0.902</td>
<td>0.958</td>
</tr>
<tr>
<td>0.05</td>
<td>0.8</td>
<td>0.8</td>
<td>-0.1</td>
<td>0.965</td>
<td>0.873</td>
<td>-0.1</td>
<td>0.965</td>
<td>0.873</td>
<td>0.981</td>
</tr>
<tr>
<td>0.05</td>
<td>0.8</td>
<td>0.8</td>
<td>0.1</td>
<td>0.739</td>
<td>0.720</td>
<td>0.1</td>
<td>0.739</td>
<td>0.720</td>
<td>0.997</td>
</tr>
<tr>
<td>0.05</td>
<td>0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>0.591</td>
<td>0.605</td>
<td>0.2</td>
<td>0.591</td>
<td>0.605</td>
<td>0.996</td>
</tr>
<tr>
<td>0.05</td>
<td>0.8</td>
<td>0.8</td>
<td>0.3</td>
<td>0.746</td>
<td>0.610</td>
<td>0.3</td>
<td>0.746</td>
<td>0.610</td>
<td>0.999</td>
</tr>
<tr>
<td>0.05</td>
<td>0.9</td>
<td>0.6</td>
<td>0.988</td>
<td>0.987</td>
<td>0.992</td>
<td>-0.3</td>
<td>0.999</td>
<td>0.958</td>
<td>0.978</td>
</tr>
<tr>
<td>0.05</td>
<td>0.9</td>
<td>0.6</td>
<td>-0.2</td>
<td>0.995</td>
<td>0.939</td>
<td>-0.2</td>
<td>0.995</td>
<td>0.939</td>
<td>0.965</td>
</tr>
<tr>
<td>0.05</td>
<td>0.9</td>
<td>0.6</td>
<td>-0.1</td>
<td>0.986</td>
<td>0.901</td>
<td>-0.1</td>
<td>0.986</td>
<td>0.901</td>
<td>0.972</td>
</tr>
<tr>
<td>0.05</td>
<td>0.9</td>
<td>0.6</td>
<td>0.1</td>
<td>0.864</td>
<td>0.832</td>
<td>0.1</td>
<td>0.864</td>
<td>0.832</td>
<td>0.994</td>
</tr>
<tr>
<td>0.05</td>
<td>0.9</td>
<td>0.6</td>
<td>0.2</td>
<td>0.830</td>
<td>0.825</td>
<td>0.2</td>
<td>0.830</td>
<td>0.825</td>
<td>0.999</td>
</tr>
<tr>
<td>0.05</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>0.888</td>
<td>0.821</td>
<td>0.3</td>
<td>0.888</td>
<td>0.821</td>
<td>0.999</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.955</td>
<td>0.952</td>
<td>0.993</td>
<td>-0.3</td>
<td>0.978</td>
<td>0.870</td>
<td>0.915</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8</td>
<td>0.8</td>
<td>-0.2</td>
<td>0.961</td>
<td>0.866</td>
<td>-0.2</td>
<td>0.961</td>
<td>0.866</td>
<td>0.884</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8</td>
<td>0.8</td>
<td>-0.1</td>
<td>0.953</td>
<td>0.758</td>
<td>-0.1</td>
<td>0.953</td>
<td>0.758</td>
<td>0.896</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.1</td>
<td>0.754</td>
<td>0.711</td>
<td>0.1</td>
<td>0.754</td>
<td>0.711</td>
<td>0.995</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>0.689</td>
<td>0.618</td>
<td>0.2</td>
<td>0.689</td>
<td>0.618</td>
<td>0.997</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.3</td>
<td>0.908</td>
<td>0.517</td>
<td>0.3</td>
<td>0.908</td>
<td>0.517</td>
<td>1.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>0.6</td>
<td>0.973</td>
<td>0.971</td>
<td>0.991</td>
<td>-0.3</td>
<td>0.780</td>
<td>0.942</td>
<td>0.946</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>0.6</td>
<td>-0.2</td>
<td>0.947</td>
<td>0.910</td>
<td>-0.2</td>
<td>0.947</td>
<td>0.910</td>
<td>0.945</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>0.6</td>
<td>-0.1</td>
<td>0.963</td>
<td>0.744</td>
<td>-0.1</td>
<td>0.963</td>
<td>0.744</td>
<td>0.915</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>0.6</td>
<td>0.1</td>
<td>0.833</td>
<td>0.731</td>
<td>0.1</td>
<td>0.833</td>
<td>0.731</td>
<td>0.987</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>0.6</td>
<td>0.2</td>
<td>0.804</td>
<td>0.758</td>
<td>0.2</td>
<td>0.804</td>
<td>0.758</td>
<td>0.994</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>0.882</td>
<td>0.781</td>
<td>0.3</td>
<td>0.882</td>
<td>0.781</td>
<td>0.997</td>
</tr>
</tbody>
</table>

$P_N^{(a)}$ is the fraction of simulations (simulated under the additive null model) with the additive null model attaining the minimum value of the BIC among the six models. $P_N^{(p)}$ and $P_N^{(m)}$ are defined in a manner similar to $P_N^{(a)}$ for the multiplicative and minimum models, respectively. $P_E^{(a)}$ represents the power of our method, i.e. the percentage of simulations (simulated under the additive epistatic model with epistasis coefficient $\epsilon_{ij}$) with the additive epistatic model attaining the minimum value of the BIC among the six models. $P_E^{(p)}$ and $P_E^{(m)}$ are defined in a manner similar to $P_E^{(a)}$ for the multiplicative and minimum models, respectively. 50 replicate fitness values are simulated.
Table 2.2: Summary of gene pairs with the indicated epistatic subtypes, inferred using the FDR procedure with the BIC method, which considers all three epistatic subtypes and their corresponding null models.

<table>
<thead>
<tr>
<th>Epistatic Subtype</th>
<th>Study S</th>
<th>Study J</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study S</td>
<td>193 (100%)</td>
<td>352 (100%)</td>
</tr>
<tr>
<td></td>
<td>(\tau = -0.060)</td>
<td>(\tau = -0.001)</td>
</tr>
<tr>
<td></td>
<td>(\epsilon_{q_{0.5}} = -0.096)</td>
<td>(\epsilon_{q_{0.5}} = -0.059)</td>
</tr>
<tr>
<td>Study J</td>
<td>352 (100%)</td>
<td>352 (100%)</td>
</tr>
<tr>
<td>Additive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study S</td>
<td>19 (9.8%)</td>
<td>35 (9.9%)</td>
</tr>
<tr>
<td></td>
<td>(\tau = 0.115^*)</td>
<td>(\tau = 0.193^{***})</td>
</tr>
<tr>
<td></td>
<td>(\epsilon_{q_{0.5}} = 0.131)</td>
<td>(\epsilon_{q_{0.5}} = 0.188)</td>
</tr>
<tr>
<td>Study J</td>
<td>35 (9.9%)</td>
<td>35 (9.9%)</td>
</tr>
<tr>
<td>Multiplicative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study S</td>
<td>33 (17.1%)</td>
<td>63 (17.9%)</td>
</tr>
<tr>
<td></td>
<td>(\tau = 0.048)</td>
<td>(\tau = 0.017)</td>
</tr>
<tr>
<td></td>
<td>(\epsilon_{q_{0.5}} = -0.166)</td>
<td>(\epsilon_{q_{0.5}} = -0.115)</td>
</tr>
<tr>
<td>Study J</td>
<td>63 (17.9%)</td>
<td>63 (17.9%)</td>
</tr>
<tr>
<td>Minimum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study S</td>
<td>136 (70.5%)</td>
<td>254 (72.2%)</td>
</tr>
<tr>
<td></td>
<td>(\tau = -0.111^{***})</td>
<td>(\tau = -0.032^{**})</td>
</tr>
<tr>
<td></td>
<td>(\epsilon_{q_{0.5}} = -0.091)</td>
<td>(\epsilon_{q_{0.5}} = -0.065)</td>
</tr>
<tr>
<td>Study J</td>
<td>254 (72.2%)</td>
<td>254 (72.2%)</td>
</tr>
</tbody>
</table>

Numbers are the counts of each type, and percentages are given of the total number of epistatic pairs. The mean (\(\bar{\tau}\)) and median (\(\epsilon_{q_{0.5}}\)) of the epistatic parameter (\(\epsilon\)) are given for each subtype, with “*” indicating that the mean of \(\epsilon\) is significantly different from 0 (p-value \(\leq 0.05\); “**” implies p-value \(\leq 0.01\), “***” implies p-value \(\leq 0.001\)). Study S refers to the St Onge et al. [255] dataset measured in the presence of MMS, and Study J refers to the Jasnos and Korona [131] dataset. (For Study S, 5 of the epistatic pairs are synthetic lethal and are not shown; as a result, percentages do not sum to 100%).
Table 2.3: Comparison of validation measures for each dataset for different variations of the FDR and BIC procedure, considering only a subset of epistatic subtypes with their corresponding null models: all epistatic subtypes (A, P, M), only the additive and multiplicative subtypes (A, P), only the additive (A), only the multiplicative (P), or only the minimum (M) subtype (see text for details).

<table>
<thead>
<tr>
<th>Subtypes Considered in BIC Procedure</th>
<th>A, P, M</th>
<th>A, P</th>
<th>A</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study J</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number Found (636)</td>
<td>352</td>
<td>273</td>
<td>263</td>
<td>231</td>
<td>329</td>
</tr>
<tr>
<td>Functional Links (25)</td>
<td>19 (0.0255)*</td>
<td>13 (0.2320)</td>
<td>11 (0.4689)</td>
<td>10 (0.4227)</td>
<td>15 (0.2619)</td>
</tr>
<tr>
<td>GO Slim Terms (Biological Process) (115)</td>
<td>69 (0.1573)</td>
<td>50 (0.4874)</td>
<td>55 (0.0736)</td>
<td>44 (0.3534)</td>
<td>68 (0.04902)*</td>
</tr>
<tr>
<td>GO Slim Terms (All) (369)</td>
<td>224 (0.0009)*</td>
<td>172 (0.01654)*</td>
<td>160 (0.1297)</td>
<td>146 (0.0273)*</td>
<td>213 (0.0003)*</td>
</tr>
<tr>
<td>Experimentally Identified (3)</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Study S</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number Found (323)</td>
<td>193</td>
<td>192</td>
<td>247</td>
<td>171</td>
<td>243</td>
</tr>
<tr>
<td>Functional Links (36)</td>
<td>21 (0.6450)</td>
<td>29 (0.0041)*</td>
<td>34 (0.0031)*</td>
<td>29 (0.0003)*</td>
<td>24 (0.9256)</td>
</tr>
<tr>
<td>GO Slim Terms (Biological Process) (283)</td>
<td>174 (0.0657)</td>
<td>174 (0.03656)*</td>
<td>223 (0.0010)*</td>
<td>153 (0.1825)</td>
<td>213 (0.5534)</td>
</tr>
<tr>
<td>GO Slim Terms (All) (307)</td>
<td>185 (0.2866)</td>
<td>182 (0.6926)</td>
<td>237 (0.1472)</td>
<td>162 (0.6997)</td>
<td>231 (0.5908)</td>
</tr>
<tr>
<td>Experimentally Identified (29)</td>
<td>17</td>
<td>22</td>
<td>24</td>
<td>23</td>
<td>21</td>
</tr>
</tbody>
</table>

Numbers in parentheses indicate p-values by Fisher’s Exact Test, “*” indicates significance. Study J refers to the Jasnos and Korona [131] dataset, and Study S refers to the St Onge et al. [255] dataset measured in the presence of MMS. Numbers in parentheses indicate the total number of tested pairs and the total number of each type of link found in each complete dataset.
Dispersal of *Mycobacterium tuberculosis* via the Canadian fur trade


*Division of Infectious Diseases, Stanford University, Stanford, CA 94305
§Department of Biology, Stanford University, Stanford, CA, USA
**Ontario Agency for Health Protection and Promotion, ON Canada M9P 3T1
§§Department of Medicine, McGill University, QC Canada H3G 1A4
a Provincial Laboratory of Public Health, AB Canada T6G 2J2
b Sioux Lookout First Nations Health Authority, ON Canada P8T 1B8
c Department of Medicine, University of Alberta, AB Canada T6G 2J3
d Division of Infectious Diseases, Queen’s University, ON Canada K7L 3N6
†Equal contributions to this work.


Abstract

Patterns of gene flow can have marked effects on the evolution of populations. To better understand the migration dynamics of *Mycobacterium tuberculosis*, we studied genetic data from European *M. tuberculosis* lineages currently circulating in Aboriginal and French Canadian communities. A single *M. tuberculosis* lineage, characterized by the DS6Quebec genomic deletion, is at highest frequency among Aboriginal populations in Ontario, Saskatchewan, and Alberta; this bacterial lineage is also dominant among tuberculosis (TB) cases in French Canadians resident in Quebec. Substantial contact between these human populations is limited to a specific historical era (1710-1870), during which individuals from these populations met to barter furs. Statistical analyses of extant *M. tuberculosis* minisatellite data are consistent with Quebec as a source population for *M. tuberculosis* gene flow into Aboriginal populations during the fur trade era. Historical and genetic analyses suggest that tiny *M. tuberculosis* populations persisted for ≈100 years among indigenous populations and subsequently expanded in the late 19th century after environmental changes favoring the pathogen. Our study suggests that spread of TB can occur by two asynchronous processes: (i) dispersal of *M. tuberculosis* by minimal numbers of human migrants, during which small pathogen populations are sustained.
by ongoing migration and slow disease dynamics, and (ii) expansion of the *M. tuberculosis* population facilitated by shifts in host ecology. If generalizable, these migration dynamics can help explain the low DNA sequence diversity observed among isolates of *M. tuberculosis* and the difficulties in global elimination of tuberculosis, as small, widely dispersed pathogen populations are difficult both to detect and to eradicate.
3.1 Introduction

Migration of human pathogens between host populations is of clear biomedical interest, as global spread of infectious diseases is predicated on pathogen dispersal. For obligate human pathogens, the paths of migration are shaped by social networks among hosts, and, at a broader scale, by human migrations and patterns of comingling. The shape and magnitude of these processes may be inferred from genetic data from both extant populations of humans [216, 219] and micro-organisms carried by humans [173]. Statistical genetic inferences about pathogen migration have been buttressed by a variety of independent data types, such as linguistic [173], epidemiological [286], and ecological [31].

Human tuberculosis (TB) is caused by Mycobacterium tuberculosis (M. tuberculosis). By current estimates, M. tuberculosis infects one third of the world’s population, of which a minority progress to disease and account for 9-10 million new, transmissible cases per year [301]. Genetic data from M. tuberculosis are characterized by low DNA sequence diversity [2] and significant population subdivision, both at continental [107, 111] and fine geographic scales [203]. TB is characterized by variable transmissibility, depending on environmental conditions, and irregular disease dynamics [7, 124]. Variation in temporal dynamics of disease is introduced by the phenomenon of clinical latency, whereby a host may be infected with M. tuberculosis for decades before reactivating the infection, becoming ill, and thus able to transmit the infection to others. These and other aspects of disease ecology would be expected to affect migration of M. tuberculosis among host populations. Broad outlines of M. tuberculosis migratory history have been described based on the geographic range of specific bacterial clades and lineages [107]. However, very little is known about M. tuberculosis migratory patterns and rates, population dynamics during migration, or effects of migratory history on M. tuberculosis population genetic diversity.

In this paper, we present evidence suggesting that M. tuberculosis dispersed into Canadian Aboriginal populations as a result of contact with European fur traders in the 18th century. Although contact between the populations involved small numbers of individuals, it was extensive enough to result in the development of Métis society, of both European and Aboriginal ancestry (see Appendix B for more historical details). Large scale TB epidemics were not evident among Western Canadian Aboriginal populations until the late 19th and 20th centuries [58, 77, 161, 203], suggesting that epidemics may be uncoupled from the process of M. tuberculosis dispersal.

Our analyses of historical, epidemiological, and M. tuberculosis genetic data suggest: 1) M. tuberculosis may be spread by small numbers of human migrants, 2) M. tuberculosis populations can persist at low levels over historical time scales, 3) these small bacterial populations may be sustained by ongoing migration and possibly by slow disease dynamics, and 4) shifts in host ecology favoring the pathogen may be accompanied by bacterial population expansions, with marked effects on genetics of M. tuberculosis populations.
3.2 Results

Results of *M. tuberculosis* population screening for lineage-defining polymorphisms [88, 185] are shown in Figure 3.1 (and Table B.1). DS6Quebec is the most frequent lineage among bacteria from all four populations: the French Canadian population of Quebec (QU), as well as Aboriginal populations in Ontario (ON), Saskatchewan (SK), and Alberta (AB). Within SK, DS6Quebec was found in all of eight intra-provincial regions (fine-scale geographic data available only for this population, see [203] for definitions of within-SK regions). Frequencies of other *M. tuberculosis* lineages were more variable among populations. A minimum spanning tree, based on 12 minisatellite loci from bacteria with the DS6Quebec polymorphism, reveals a distinctly star-like network structure (Figure 3.2), with little evidence of population differentiation. Networks based on minisatellite data within other common lineages (Rd 182, Rd 219 and H37Rv-like) were not star-like (see Figure B.1). Lineage-specific patterns were also evident in analyses of population genetic structure (Analysis of Molecular Variance, AMOVA, Table 3.1). Overall, a modest level of differentiation was observed among populations ($F_{ST} = 0.08$). However, differentiation varied substantially between lineages with the least within the DS6Quebec lineage ($F_{ST} = 0.04$) and the greatest within the H37Rv-like lineage ($F_{ST} = 0.52$). These results suggest 1) the DS6Quebec lineage was introduced to these populations in the relatively remote past, resulting in a similarly high frequency across populations, 2) *M. tuberculosis* belonging to the DS6Quebec lineage were, or are, exchanged freely among study populations, 3) the DS6Quebec lineage underwent rapid population expansion, reflected in a star-like network topology, possibly coincident with its introduction to new host populations, and 4) there is temporal and/or spatial variation in patterns of bacterial migration among human populations, reflected in variable differentiation within bacterial lineages.

Although the QU population is now geographically and socially isolated from ON, SK and AB populations, there is a history of comingling in the context of the Canadian fur trade (Figure 3.1). Table 3.2 outlines the expected history of the DS6Quebec lineage, assuming it was introduced to Canada by a French immigrant to Quebec, and dispersed to Aboriginal populations via fur trade transportation and social networks. Estimates derived from *M. tuberculosis* minisatellite data of the time to most recent common ancestor (TMRCA [23]) and divergence time between bacterial populations ($T_D$ [314]) suggest that genetic exchange between QU and Western Aboriginal *M. tuberculosis* populations occurred over approximately 100 years; timing of this exchange is consistent with historical human migrations connected with the trade in furs (additional estimates in Appendix B).

Levels of overall genetic diversity of *M. tuberculosis* restriction fragment length polymorphism (RFLP [284]) haplotypes from four populations are shown in rarefaction curves in Figure 3.3A. Published data were available from MB, and are included along with QU, SK and AB; data from ON are not included, due to its small sample size (see Appendix B and Figure B.3). The number of distinct haplotypes is highest in QU, consistent with it being a source population for pathogen migration. Levels of diversity are lowest in SK and MB, with intermediate diversity in AB. Diversity thus does not decrease in an East-to-West pattern, as we might expect with a serial founder effect [216] originating in QU.

Post-European-contact epidemics of TB among Canadian indigenous populations have been associated
with dramatic social, economic and environmental changes that characterized the industrial era (late 19th century and beyond, see Appendix B). There is evidence of regional differences in the pace of these changes: published analyses of archival fur trade documents indicate that some indigenous populations remained remote from the cash based industrial economy as late as the 1920s [217]. Regions within SK and MB fall into this historically remote/traditional category, whereas AB does not contain any such regions (Figure 3.1). Further development of transportation networks (starting with the widespread use of bush planes in the 1930s) later allowed commercial development of previously remote regions [217].

Given associations between industrialization, attendant shifts in host ecology, and TB epidemic expansion among indigenous populations (see Appendix B), we hypothesized that \textit{M. tuberculosis} diversity in SK and MB was low relative to AB as a result of more recent \textit{M. tuberculosis} population expansion in remote/traditional areas. To test this hypothesis, minisatellite and RFLP haplotypes from SK were divided into historically “remote” (RS) and “non-remote” (NRS) regions according to the classification of trading districts outlined in [217]. Consistent with this classification, diversity in NRS is similar to AB (containing no remote regions), with the lowest diversity observed in RS (Figure 3.3B and Appendix B). Permutation procedures identified RS and QU as having lower and greater diversity, respectively, than would be expected under a random distribution of minisatellite haplotypes among populations (see Appendix B).

We analyzed DS6\textsubscript{Quebec} minisatellite data with rejection sampling, an Approximate Bayesian Computation (ABC) method, to assess whether a demographic model allowing for an expansion in RS (in which the historical bacterial effective population size, \(N_e\), was smaller than the current \(N_e\)) is more likely than a null model of constant size. We estimated the parameter \(\omega\), the ratio of historical \(N_e\) to contemporary \(N_e\) (see Methods and Appendix B). Results using this method indicate that the posterior probability of the expansion model (versus the null constant size model) is 1; based on simulations under the null model, the p-value of this posterior probability is 0. We estimate \(\omega\) to be 0.077 (95% credibility interval 0.012-0.291); see Figure B.4A. Translating these results to absolute numbers (current \(N_e\approx 30\)) generates an estimate of effective number of TB cases in RS prior to 1930 (the time fixed for the expansion, based on the social history described above) equal to 2.345 individuals (95% credibility interval: 0.351-8.803). See Appendix B for parameter estimation under alternative mutation rate assumptions, all of which reject the constant size model in favor of the expansion model.

Based on the size and dynamics of the population involved in the French (Montreal-based) trade in the West prior to 1870, and prevalence of TB in European cities in the 18th century, we estimated the absolute number of \textit{M. tuberculosis} infections transmitted to Western indigenous populations (see Appendix B). The historical estimate of \(Nm\), the product of population size (\(N\)) and fraction of the indigenous \textit{M. tuberculosis} population replaced by immigrants (\(m\)) per generation is 0.16. In an island model, \(Nm < 1\) results in substantial population differentiation due to genetic drift [307]. Estimates of \(Nm\) derived from bacterial minisatellite data were higher than this historical estimate: for pooled lineages, \(Nm\) derived from \(F_{ST}\) (AMOVA) is 6.16, and Slatkin’s private alleles method [15] generated an estimate of 3.62 (empirical 95% CI 2.47-4.72).
3.3 Discussion

The recent history of human migration to Western Canada informs analyses of *M. tuberculosis* genetic data presented here. Most importantly, substantial contact between French Canadian and Western indigenous populations is limited to a specific historical period (the fur trade era). The early bound on this period of contact is provided by the dates of incursion of fur traders into the Western interior (approximately 1710 [120]), while the later bound is provided by historical and demographic analyses indicating that Westward migration of French Canadians ceased in the latter half of the 19th century for economic and other reasons [92].

Based on the observed pattern of *M. tuberculosis* lineage frequencies in this historical context, as well as bacterial population genetic diversities consistent with QU as a source population, genetic timing estimates, and patterns of genetic differentiation between populations, we infer that the DS6Quebec lineage was dispersed to indigenous populations by French Canadian fur traders, about a century before epidemic forms of TB were manifest in Western Aboriginal communities.

Several features of the human migration associated with trade in furs are noteworthy. First, the absolute number of human migrants was small: by our estimate, 5,419 individuals migrated from East to West, during 160 years of trade between Montreal and Western Canadian indigenous populations (see Appendix B and Table B.2). These early migrations are dwarfed by population movements of the late 19th and early 20th centuries. Facilitated by massive inter-provincial and international migration to the Canadian prairies - from which French Canadians were largely absent - the census of Western Canada grew from 110,000 in 1871 to 1,750,000 in 1911 [283]. Between 1900 and 1917, a total of 1,671,414 foreign-born individuals migrated to Western Canada from the UK, the Ukraine, Russia, Germany, Austria, Hungary, Norway, Denmark, China and elsewhere [294]. Despite the enormous number of migrants from regions of high TB incidence, who would be expected to introduce a broad range of Euro-American and East Asian *M. tuberculosis* lineages [111], bacterial populations in Western indigenous communities are dominated by the DS6Quebec lineage, present at a similar frequency in the French Canadian population of Quebec (Figure 3.1).

The apparent lack of *M. tuberculosis* gene flow from 19th century homesteaders to indigenous populations may be explained by social distance between populations. International migrants to the prairies, particularly those facing language and cultural barriers to assimilation, were socially and geographically isolated [294]. By the late 19th century, prairie indigenous populations were also socially segregated, at times forcibly [161, 294], from the non-Aboriginal population. This is in contrast to the earlier fur trade era, characterized by inter-marriage and trading collaborations between European immigrants and First Nations groups ([82] and Appendix B). Although contact between populations involved small numbers of individuals, we speculate that close social ties between sending and receiving host populations permitted migration of *M. tuberculosis* through the fur trade. Given that transmission of *M. tuberculosis* requires sustained, close contact (exemplified by efficient transmission in high density shared living environments such as prisons [124]) this observation is likely to be generalizable, with structure of global *M. tuberculosis* populations strongly influenced by the social architecture of host populations.
Although by no means a disease-free interval for Native peoples (there were devastating epidemics of smallpox and other infectious diseases), epidemics of TB were not a feature of the fur trade era (1710-1870). TB epidemics among Western Canadian indigenous populations occurred later, starting in the late 1800s [58, 77, 124, 161]. Expansion of rail and steamship networks into Western Canada in the late 19th century permitted agricultural development, industrial-scale extraction of natural resources, development of government institutions and mass immigration. TB epidemics were among a chain of sequelae for Aboriginal populations that included displacement, loss of traditional food sources, crowding and institutionalization ([84] and Appendix B).

Some indigenous populations (e.g. RS, parts of MB) remained remote from the evolving industrial economy into the 1920s ([217], shaded areas of Figure 3.1). We find that \textit{M. tuberculosis} genetic diversity in RS is the lowest, and diversity in QU the highest (Figure 3.3); we did not have detailed geographic data that would allow us to classify MB \textit{M. tuberculosis} strains according to the same scheme as SK (i.e. NRS vs RS). We note that molecular epidemiological studies of other historically remote/traditional regions are consistent with low \textit{M. tuberculosis} population genetic diversity [145, 186].

Differences in genetic diversity may result from two distinct phenomena. First, the process of \textit{M. tuberculosis} migration is likely to involve serial founder events, such that diversity is highest in the source population (QU), and lower in the subsequently founded populations (ON, RS, NRS, AB, MB). Another explanation is that QU, the oldest \textit{M. tuberculosis} population (founded by 17th - 18th century migrants from France, see Table 3.2), may be closer to an equilibrium distribution of haplotype frequencies; the youngest populations may still show evidence of a “founder flush” [266]. A historically remote/traditional population, RS was likely shielded from ecological antecedents of epidemic TB until the regional incursion of air travel networks in the 1930s [217], and thus had the most recent expansion in its pathogen population. Demographic modeling with Approximate Bayesian Computation demonstrated a very high probability of expansion in the RS \textit{M. tuberculosis} population since 1930; magnitude of the bacterial expansion was approximately 13-fold.

Although our historical estimates of \textit{M. tuberculosis} migration rates were low (direct $\hat{N}m = 0.16$), the lack of genetic differentiation among populations (DS6\textit{Quebec} $F_{ST} = 0.04$) and estimates of Nm from the genetic data (indirect $Nm = 3.62$) would imply a high rate of \textit{M. tuberculosis} migration. Slatkin has observed that indirect (genetic) estimates of gene flow may be higher than direct measurements of population dispersal, especially in populations characterized by colonization-extinction-re-colonization dynamics, which can limit population differentiation even when rates of migration are low [251, 252]. Taken together, our results suggest that absolute numbers of \textit{M. tuberculosis} migration events were low ($\hat{N}m < 1$, an order of magnitude estimate based on historical data), and patterns of genetic differentiation have been affected by unstable dynamics of small, pre-industrial \textit{M. tuberculosis} populations.

We have delineated two asynchronous processes involved in the spread of TB from European immigrants to Native Canadians (Figure 3.1B). The first is dispersal of the etiologic agent, \textit{M. tuberculosis}, populations of which were sustained at very low levels ($N_e \approx 2$) for approximately 100 years by small numbers of human migrants who had intimate, sustained contact with susceptible hosts. In addition to sustained migration,
variable transmission dynamics of TB may have cushioned small bacterial populations against extinction [33]. The second process is expansion of the bacterial population, following a shift in host ecology favoring the pathogen. We find evidence of bacterial population expansions in the DS6\textsuperscript{Quebec} \textit{M. tuberculosis} haplotype network (Figure 3.2), patterns of genetic diversity (Figure 3.3), and coalescent-based demographic analysis of bacterial minisatellite data.

This is a study of a specific historical phenomenon, and it is unknown whether the observed patterns of \textit{M. tuberculosis} migration are applicable to other settings. However, the fact that pathogen migration can be asynchronous with epidemiological phenomena is likely to render control of TB more difficult, and could help explain global persistence of the pathogen despite extensive efforts at eradication. As an obligate human pathogen that requires specific environmental conditions and sustained contact for transmission, barriers to geographic spread of \textit{M. tuberculosis} are largely social, and therefore somewhat fluid. Perseverance of tiny bacterial populations would permit gradual accrual of a large geographic range despite significant population subdivision. As a result, by the time that the spread of TB is obvious, \textit{M. tuberculosis} populations may already be well established.

### 3.4 Materials and Methods

#### 3.4.1 Population Descriptions

Clinical isolates of \textit{M. tuberculosis} in this study are from five Canadian populations: the French Canadian population of Quebec (QU), and Aboriginal populations in Ontario (ON), Manitoba (MB), Saskatchewan (SK), and Alberta (AB). The QU sample (\(n = 297\)) is described in [185]. The ON sample (\(n = 45\)) derives from TB cases that occurred between 1997 and 2009 in First Nations communities in a single region of the province (total population approx. 25,000). MB data (\(n = 163\)) are from a published study [32]: total number of TB cases, number of different bacterial RFLP haplotypes, number of singletons and the configuration distribution of the five most common \textit{M. tuberculosis} haplotypes from First Nations reserve communities in MB (1992-1999) are reported. We made the conservative assumption that the remaining (unreported) haplotypes were doubletons. The SK sample (\(n = 444\)) is described in Pepperell et al. [203]. AB samples (\(n = 283\)) are from TB cases among First Nations individuals in the province, excluding urban areas, from 1990-2008.

#### 3.4.2 Genotyping Methods

Isolates of \textit{M. tuberculosis} were genotyped by RFLP, based on the number and location of IS6110 elements [284]. The number of tandem repeats at 12 minisatellite loci was also determined for each isolate using a standard methodology [263]. \textit{M. tuberculosis} lineagetyping, based on genomic deletions [88, 185], was done with real-time PCR (details in Appendix B).
3.4.3 Mutation Rate Estimate

We estimated a mutation rate ($\mu$) per transmission generation [203] for minisatellite loci based on simulations using empirical estimates of intra-host *M. tuberculosis* population dynamics and in vitro estimates of mutations per cell doubling for bacterial minisatellite loci. Given a known per-year RFLP mutation rate [265], we made additional estimates by examining the number of RFLP types per minisatellite haplotype and comparing $\theta (2N\mu)$ estimates from RFLP and minisatellite data. All estimates were of a similar order of magnitude and were consistent with published estimates [223]; we used 0.001 mutations per locus (see Appendix B and Table B.3).

3.4.4 Statistical Calculations

*Genetic differentiation.* Analysis of Molecular Variance (AMOVA, implemented in Arlequin version 3.5 [74]) was used to calculate $F_{ST}$ values from *M. tuberculosis* minisatellite haplotype data.

*Network analysis.* BioNumerics 5.0 (Applied Maths, Kortrijk, Belgium) was used to generate minimum spanning trees of *M. tuberculosis* minisatellite haplotypes. This program implements the Prim-Jarnik algorithm; the BURST priority rule maximizing single and double locus variants was used during network searches.

*Genetic timing estimates.* We estimated TMRCA of the DS6Quebec lineage in each population using minisatellite genotypes and the method of *Ytime* [23]. We selected as the root the haplotype at the center of the DS6Quebec network (Figure 3.2), present at high frequency in QU, SK and AB. We obtained bootstrap confidence intervals with *Ytime* assuming a constant population size and neutral demography (see Appendix B). Divergence times for DS6Quebec isolates between pairs of populations were estimated from minisatellite genotypes by the $T_D$ estimator, which is robust to population size changes and weak gene flow [314] (see Appendix B). This procedure requires an estimate of $V_0$, the average variance in repeat number in the ancestral population at the time of divergence; we estimated $T_D$ using three different values of $V_0$ for each pair of populations (see Results and Appendix B). Alternative mutation rate and demographic assumptions have little effect on our main conclusions (Appendix B).

*Population genetic diversity.* We assessed diversity of minisatellite and RFLP haplotypes in each population, for all lineages, by calculating the number of distinct haplotypes using rarefaction [136], calculating haplotype diversity with re-sampling to the lowest sample size to correct for unequal sample sizes [247], and implementing a novel permutation procedure to assess whether observed diversities of minisatellite haplotypes could be explained by a random partitioning of isolates from all populations (details in Appendix B).

*Migration (Nm).* Under a simple island model, assuming an equilibrium population where each “island” (QU, SK, and AB) has equal size $N$, we applied Slatkin’s private alleles method with minisatellite haplotype data from all lineages to estimate $Nm$, where $m$ is the probability that an individual is a migrant each
3.5. ACKNOWLEDGMENTS

Acknowledgments

We thank T MacMillan and D Popescu for assistance with data collection; A Avelar, T Van, E Heinemeyer and C-Y Wu for technical assistance; and G Dolganov for advice and assistance with real time PCR assays. This work was supported by NIH grant 5K08AI67458-2 to CP, National Institutes of Health (NIH) grant GM28016 to MWF and a National Science Foundation (NSF) Graduate Research Fellowship (grant number DGE-0645962) to JMG.
3.6 Figures and Tables

Figure 3.1: Fur trade geography, regional M. tuberculosis lineage frequencies and historical timeline. A) Map of Canada from Natural Resources Canada [182]. The main fur trade canoe route from the St. Lawrence River to the Beaufort Sea (Montreal route) is in light blue. Canoe routes from Hudson’s Bay to the interior are shown in red. Geography of canoe routes is based on a map by A. Ray in [120]; this was checked against a map of fur trade posts [181]. Geography of railways, steamship lines and areas classified as remote/traditional on the basis of archival evidence (all ca. 1920) are from Figure 38 in [120] and [217]. Proportional lineage frequencies of M. tuberculosis isolates are shown as pie charts in the corresponding province (see also Table B.2). Nomenclature of M. tuberculosis lineages is from [88] and [185]. Pie charts are the same size for clarity, although total sample sizes differed between populations. Lineage frequencies were unavailable for MB. B) Events shown in the timeline are the founding of New France (Quebec) in 1608; incursion of fur traders to the Northwest around 1710; British conquest of New France in 1760 (and end of migration from France to Quebec); merger of North West Company and Hudson’s Bay Company (HBC) in 1820 (with subsequent abandonment of Montreal route in favor of Hudson’s Bay routes); completion of the Canadian Pacific Railway (CPR) in 1885, by which time Western buffalo herds were severely depleted; and, finally, widespread use of bush planes to reach remote areas, starting in the 1930s. Gray boxes indicate the estimated timing of processes in M. tuberculosis demographic history: dispersal of M. tuberculosis to indigenous populations (light gray), and expansion of M. tuberculosis populations as a result of shifts in host ecology (dark gray).
Figure 3.2: Minimum spanning tree of minisatellite haplotypes (12 loci) from *M. tuberculosis* within the DS6Quebec lineage, in QU, ON, SK, AB. The size of nodes is proportional to the frequency of TB cases associated with that *M. tuberculosis* haplotype. Scale is indicated by the line at the bottom of the figure, which represents a difference of one minisatellite repeat. Minisatellites not available from MB.
Figure 3.3: **Genetic diversity of *M. tuberculosis* populations.** A) Number of distinct RFLP haplotypes as a function of the number of sampled chromosomes obtained using rarefaction (see Methods). Populations include QU, MB, SK, and AB; ON is not included due to its low sample size (see Appendix B). Every fourth data point is presented for clarity. B) RFLP haplotype diversity (Shannon index), correcting for sample size by repeatedly sampling the total number of isolates in the smallest sample from each population (NRS, \( n = 123 \)). Boxplots indicate values obtained over all samples; value for the smallest sample (NRS) is indicated by a line. Populations are as in part A), except that SK is split into Remote Saskatchewan (RS) and Non-Remote Saskatchewan (NRS). Although MB contains both remote and non-remote regions (Figure 3.1), detailed geographic data were not available for this sample; diversity shown here is for the entire sample.
Table 3.1: Within- and between-lineage AMOVA based on 12 minisatellite loci for *M. tuberculosis* from Quebec, Ontario, Saskatchewan, and Alberta.

<table>
<thead>
<tr>
<th>Bacterial Lineage</th>
<th>Source of Variation</th>
<th>D.f.</th>
<th>SS</th>
<th>Variance component</th>
<th>% Variation</th>
<th>$F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS6Quebec</td>
<td>Among populations</td>
<td>3</td>
<td>16.25</td>
<td>0.04</td>
<td>3.81</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Within populations</td>
<td>583</td>
<td>521.26</td>
<td>0.89</td>
<td>96.19</td>
<td></td>
</tr>
<tr>
<td>H37Rv-like</td>
<td>Among populations</td>
<td>3</td>
<td>181.68</td>
<td>0.89</td>
<td>51.56</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Within populations</td>
<td>306</td>
<td>256.44</td>
<td>0.84</td>
<td>48.44</td>
<td></td>
</tr>
<tr>
<td>Rd 182</td>
<td>Among populations</td>
<td>3</td>
<td>8.26</td>
<td>0.33</td>
<td>22.26</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Within populations</td>
<td>59</td>
<td>67.95</td>
<td>1.15</td>
<td>77.74</td>
<td></td>
</tr>
<tr>
<td>Rd 219</td>
<td>Among populations</td>
<td>3</td>
<td>14.28</td>
<td>0.25</td>
<td>20.97</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Within populations</td>
<td>65</td>
<td>60.68</td>
<td>0.93</td>
<td>79.03</td>
<td></td>
</tr>
<tr>
<td>All lineages</td>
<td>Among populations</td>
<td>3</td>
<td>75.54</td>
<td>0.10</td>
<td>7.95</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Within populations</td>
<td>1,040</td>
<td>1227.16</td>
<td>1.18</td>
<td>92.05</td>
<td></td>
</tr>
</tbody>
</table>

D.f., degrees of freedom; $F_{ST}$, fraction of variation among populations; SS, sum of squares.
### Table 3.2: Genetic timing estimates for DS6$^{\text{Quebec}}$ M. tuberculosis lineage in calendar years.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Historical correlate</th>
<th>Point estimate(s)*</th>
<th>Confidence interval†</th>
<th>Historical prediction‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMRCA (QU)</td>
<td>French migration to Quebec</td>
<td>1740</td>
<td>1709-1771</td>
<td>1608-1760</td>
</tr>
<tr>
<td>TMRCA (SK)</td>
<td>Fur trade expansion West</td>
<td>1797</td>
<td>1777-1815</td>
<td>1730-1870</td>
</tr>
<tr>
<td>TMRCA (AB)</td>
<td></td>
<td>1779</td>
<td>1751-1804</td>
<td>1750-1870</td>
</tr>
<tr>
<td>$T_D$ (QU/SK)</td>
<td>Separation of populations</td>
<td>1789, 1884, 1910</td>
<td>1788-1979</td>
<td>1870</td>
</tr>
<tr>
<td>$T_D$ (QU/AB)</td>
<td></td>
<td>1779, 1901, 1885</td>
<td>1805-1966</td>
<td></td>
</tr>
</tbody>
</table>

*Three point estimates are shown for $T_D$. The first is the earliest bound for divergence, where a variance of 0 is assumed for minisatellite repeat number in the ancestral population. For the second estimate, we assume that variance in the ancestral population is equal to the variance of haplotypes found presently in the QU population and shared with the founded population (SK or AB). The third point estimate is based on an assumption that variance in the ancestral population was equal to the variance found presently in haplotypes in the founded population (SK or AB) that are also found in QU.

†Results presented for TMRCA assume a star-like genealogy. TMRCA confidence intervals are larger if constant population size or exponential growth are assumed (see Appendix B).

‡Estimate of timing of M. tuberculosis population events, based on historical (nongenetic) data. References for date ranges are Charbonneau [46] for timing of migration to Quebec from France and Innis [120] for other dates. The earlier bounds for fur trade expansion into Saskatchewan and Alberta are based on the time of establishment of fur trade posts in these regions. Note that this timing is later than expansion into the Northwest as a whole (ca. 1710), which includes Western Ontario and Manitoba. Timing of separation of populations is based on analyses of Innis [120] indicating that the Hudson’s Bay Company shifted to less labor-intensive methods of extracting furs (and thus had no need of French Canadian voyageurs), starting about this time. Analyses of interprovincial migration from the 1870s to the early 20th century are also consistent with minimal migration of French Canadians to the Western provinces [92].
Limited Evidence for Classic Selective Sweeps in African Populations

Julie M. Granka*, Brenna M. Henn§, Christopher R. Gignoux**, Jeffrey M. Kidd §§, Carlos D. Bustamante§, Marcus W. Feldman*

*Department of Biology, Stanford University, Stanford, CA, USA
§Department of Genetics, Stanford University, Stanford, CA, USA
**University of California, San Francisco, CA, USA
§§Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

Genetics. 2012 Nov; 192(3):1049-64.

Abstract

While hundreds of loci have been identified as reflecting strong positive selection in human populations, connections between candidate loci and specific selective pressures often remain obscure. This study investigates broader patterns of selection in African populations, which are underrepresented despite their potential to offer key insights into human adaptation. We scan for hard selective sweeps using several haplotype and allele frequency statistics with a dataset of nearly 500,000 genome-wide single nucleotide polymorphisms in 12 highly diverged African populations that span a range of environments and subsistence strategies. We find that positive selection does not appear to be a strong determinant of allele frequency differentiation among these African populations. Haplotype statistics do identify putatively-selected regions that are shared across African populations. However, as assessed by extensive simulations, patterns of haplotype sharing between African populations follow neutral expectations and suggest that tails of the empirical distributions contain false positive signals. After highlighting several genomic regions where positive selection can be inferred with higher confidence, we use a novel method to identify biological functions enriched among populations’ empirical tail genomic windows, such as immune response in agricultural groups. In general, however, it seems that current methods for selection scans are poorly suited to populations which, like the African populations in this study, are affected by ascertainment bias, have low levels of linkage disequilibrium, possibly old selective sweeps, and potentially reduced phasing accuracy. Additionally, population history can confound the interpretation of selection statistics, suggesting greater care is needed in attributing broad genetic patterns to human adaptation.
4.1 Introduction

Elucidating the selective pressures that human populations have encountered, as well as the means by which they have adapted to them, are central aims of evolutionary biology and anthropology. Recently, statistical methods in population genetics have been applied to genome-wide polymorphism data to identify genetic loci that may have experienced natural selection. Such inferences have primarily been made under the model of a hard selective sweep, where a new allele rapidly rises to fixation within a population due to positive selection [3, 207, 212, 231, 232, 233, 234, 291, 303]. While these genome-wide scans have detected hundreds of loci as focal sites of selective sweeps, most connections between the loci and their selective pressures remain unknown. Lack of phenotypic or functional information, the relative effect of background selection, demography, and statistical noise have been suggested as some of the causes of the lack of agreement between studies [3, 105, 106, 269].

In light of these issues, it is crucial to understand broader patterns of adaptation in human populations before identifying individual putatively selected loci [52, 207]. One strategy is to compare across populations the regions of the genome that appear to be undergoing positive selection while incorporating knowledge of populations’ environments and lifestyles, providing insight into the selective forces acting on the same loci [95, 96, 97, 147, 207, 291]. A second approach is to search for biological functions that could be under strong positive selection within populations, identifying functions that are more frequently associated with putatively-selected regions of the genome than expected by chance (see, for example, [291] and [170]).

Using both approaches, we sought a more complete picture of adaptation in human African populations. Previous studies [130, 147, 233, 291] have reported strong signals of positive selection in African populations. Local adaptation is expected to be common within Africa, as African populations practice a range of subsistence strategies and inhabit diverse environments to which adaptation may have occurred [40, 61, 62]. Selection may also have acted more efficiently in African populations, many of which have maintained large effective population sizes for the past 50,000 years, in contrast to populations outside of Africa [40, 103, 148, 267]. While African populations would appear ideal for understanding the extent of positive selection in humans, they have usually been underrepresented in such studies.

In a set of 12 highly diverged hunter-gatherer, pastoralist, and agriculturalist African populations, we search for genomic evidence of positive selection. We analyze genome-wide data of nearly a half a million single nucleotide polymorphisms (SNPs), and include populations from the Human Genome Diversity Project (HGDP) [157], HapMap [121, 122, 123] and four hunter-gatherer populations for which thorough selection scans have not yet been conducted [103, 242]. As potential candidates for positive selection, we identify outliers in the empirical genomic distributions of several commonly-used selection statistics, including measures of population differentiation and hard selective sweeps.
4.2 Methods

4.2.1 Samples

We performed selection scans on genomic data from individuals sampled from 12 African samples using a set of nearly 500,000 single nucleotide polymorphisms (SNPs). Our sample of African populations includes agriculturalists, pastoralists, and hunter-gatherers, as well as individuals from rainforests, deserts, and other environments.

Populations included from the Centre d’Étude du Polymorphisme Humain (CEPH)-HGDP (sample sizes in parentheses) were the South African Bantu (8), Kenyan Bantu (11), Biaka Pygmies (22), Mbuti Pygmies (13), Mandenka (22), Namibian San (5), and Yoruba (21); Mozabites were excluded [157, 227]. We also included 12 Namibian San individuals from Schuster et al. [242], who were combined with the HGDP Namibian San for a total of 17 Namibian San individuals; we also included additional individuals from the Hadza and Sandawe of Tanzania and the Khomani Bushmen of South Africa, described by Henn et al. [103]. As in Henn et al. [103], we removed closely related individuals, leaving 17 Hadza, 28 Sandawe, and 31 Khomani Bushmen. (For $F_{ST}$ calculations, we removed admixed individuals identified in Henn et al. [103], leaving 11 Hadza and 23 Khomani Bushmen). HapMap Maasai trio parents from Kinyawa, Kenya (M KK) (46 individuals) and Yoruba trio parents from Ibadan, Nigeria (YRI) (100 individuals) were also analyzed [121, 122, 123]. (As Pemberton et al. [201] found evidence of undocumented relatedness between some Maasai individuals, we removed a subset from the original dataset; see Appendix C Methods). For some analyses, we included HapMap CEPH Utah residents with ancestry from northern and western Europe (CEU) trio parents (88 individuals). HapMap Tuscans (88 individuals) were included for pairwise mean $F_{ST}$ and population differentiation calculations [103]. For a summary, see Table C.1.

After merging the SNPs genotyped in the HapMap samples, in the CEPH-HGDP samples on the Illumina 650K platform, in the samples from Henn et al. [103] on the Illumina 550K platform, and sequence data from Schuster et al. [242], a total of 461,767 SNPs remained. For haplotype-based selection statistics (see below), we pruned additional SNPs and phased genotype data from 461,154 SNPs using BEAGLE [36] (as in Henn et al. [103], see Appendix C Methods). Known parent-offspring trios from the HapMap YRI and Khomani Bushmen were used as seeds, with all default parameters, to increase phasing performance [37]. All populations were phased in this manner except for the HapMap CEU, YRI, and MKK, for which phase information was available from the HapMap website. 457,682 autosomal SNPs were used for analyses of single SNP selection statistics (see below).

Ancestral alleles for all SNPs were determined by their state in chimpanzee; chimpanzee positions were obtained from the Ensembl database (http://www.biomart.org, accessed August 2011). If the ancestral state was unavailable or ambiguous, the SNP was removed. All datasets were merged to Human Genome Build hg18 (NCBI Build 36). When needed, we used the recombination map available through the HapMap [121, 122, 123].
4.2.2 Single SNP Statistics

We calculated genome-wide (over all SNPs) mean $F_{ST}$, a measure of population differentiation, between all pairs of African populations, and for each African population against the HapMap Tuscans, using the program GENEPOP \[220, 230\]. GENEPOP was also used to calculate per-SNP $F_{ST}$ values between all pairs of African populations. We also calculated $F_{ST}$ over all populations for each SNP as in Akey et al. \[4\], following Weir and Cockerham \[299\]. For these analyses, we included only the HGDP Yorubans (not the HapMap YRI due to their close relation to the HGDP population), for a total of 11 African populations.

We classified a SNP as genic if it was within 2 kb of a gene, and non-genic otherwise. Gene locations were obtained from University of California, Santa Cruz (UCSC) refFlat mappings of RefSeq genes to hg18 (downloaded in May 2010, \url{http://www.genome.ucsc.edu}). We arranged values of $F_{ST}$ calculated over all populations for each SNP in bins with increments of 0.05. Within each bin, an enrichment value was calculated as the proportion of genic (or nongenic) SNPs in that bin divided by the genome-wide proportion of genic (or nongenic) SNPs. To assess significance, we resampled 200 kb segments of the genome and re-calculated enrichments in each $F_{ST}$ bin \[52\]. Upper and lower 2.5% values over 1,000 bootstrap samples were used to obtain 95% confidence intervals. We also calculated the absolute value of the derived allele frequency difference ($|\delta|$) between all pairs of populations (55, or $\binom{11}{2}$, pairs). We averaged the $|\delta|$ values over all pairs for each SNP (to create one measure of differentiation) and calculated genic enrichments (in bin sizes of 0.03), assessing significance as above. We also calculated enrichments in bins of (non-absolute) $\delta$ values for individual pairs of African populations, as well as between the HGDP Yorubans and HapMap Tuscans, again assessing significance by bootstrapping (see Appendix C).

For all population pairs, we evaluated the relationships between the top 99.99% tail values of $|\delta|$ (over all SNPs) and mean $F_{ST}$ (averaged over all SNPs), and fitted a curve using the \textit{lowess()} function in R (\url{http://www.r-project.org}) (see Appendix C Methods for additional analyses). We simulated the expected relationship under neutrality between the top 99.99% tail of $|\delta|$ values and mean $F_{ST}$ using the “beta-binomial” method of Balding \[11\] (also used in \[52\]), where allele frequencies in each population are drawn from a beta-binomial distribution whose variance is a function of the $F_{ST}$ between the two populations (see Appendix C Methods). For each pair of populations, we repeated this simulation independently for 457,682 loci (the number of autosomal SNPs), and determined the 99.99th percentile value of $|\delta|$ over all loci. For each population pair we plotted the mean and standard deviation of the 99.99% $|\delta|$ values over 25 simulations. A best fit curve through the simulated mean values was drawn using the \textit{lowess()} function in R.

4.2.3 Haplotype Statistics

We calculated several haplotype statistics widely used for detection of selective sweeps: the iHS (Integrated Haplotype Score) \[291\], which identifies ongoing and incomplete sweeps, and the Cross Population Extended Haplotype Homozygosity Test (XP-EHH) \[233\], which detects complete or nearly complete sweeps occurring in one population but not in another “reference” population (see Appendix C Methods for more details). Both
4.2. METHODS

Statistics were calculated using scripts from the HGDP Selection Browser website [207] (http://hgdp.uchicago.edu). Phased genotypes, ancestral SNP states, and recombination maps were determined as described previously.

Reference populations used for the XP-EHH calculations were the HapMap CEU (“XP-EHH CEU”), HapMap YRI (“XP-EHH YRI”), HapMap Maasai (“XP-EHH MKK”), and Khomani Bushmen (“XP-EHH KHB”) (see Appendix C for details). Using a more closely-related African reference population with a relatively more stable demographic history avoids artifacts caused by an out-of-Africa expansion, which could strongly affect the XP-EHH CEU statistic. iHS and XP-EHH CEU were calculated for all 12 populations, XP-EHH YRI for ten (HGDP Yorubans and HapMap YRI were excluded), and XP-EHH MKK and XP-EHH KHB for 11 (Maasai and Khomani Bushmen, respectively, were excluded) (see Table C.1). Per-SNP statistics were normalized in bins of SNPs with similar minor allele frequencies (in increments of 0.05) for iHS as in V oight et al. [291], and over all SNPs for XP-EHH as in Sabeti et al. [233].

Windows and Haplotype Statistics

We broke the genome into non-overlapping windows of 100 kb to identify regions of the genome under selection. Because of the lower linkage disequilibrium (LD) in African populations [103], we reduced the 200 kb window size used by Pickrell et al. [207] (see Appendix C Methods; see also Conrad et al. [51]).

As a summary statistic for each window, we used the maximum XP-EHH value of all SNPs within a window; for iHS, we used the proportion of SNPs in the window with $|iHS| > 2$ [207]. The variance of SNP statistics within 100 kb windows was significantly smaller than that among randomly selected SNPs (see Appendix C Methods and Results), confirming that the windows effectively grouped SNPs with more similar statistic values.

We binned genomic windows according to their numbers of SNPs in increments of ten SNPs (combining the few windows with $\geq 50$ SNPs into one bin). We then obtained an empirical p-value for each window based on its ranking (according to its window statistic) in its SNP frequency bin. (See Appendix C and [52].)

Overlap of Signals Between Populations

As a summary, we first obtained the five genomic windows with the lowest empirical p-values in each population, and their p-values in all other populations, for iHS, XP-EHH CEU, XP-EHH YRI, XP-EHH MKK, and XP-EHH KHB (see Appendix C Methods for more details).

To quantify the similarity of selection signals between pairs of populations, we counted the number of 100 kb genomic regions present in the top empirical 1% (and 0.1%) of both populations of the pair using the intersectBed program in the program suite BEDTools [215]. For each haplotype statistic, this resulted in a matrix of values over all population pairs. For each population pair we also computed a correlation between the two populations’ genome-wide window statistic values to construct another matrix for each haplotype statistic. We calculated a series of Mantel correlations of these matrices with a matrix of between-population $F_{ST}$ values; we obtained p-values by comparing observed correlations to those of 5,000 permutations, where
rows and columns of the original matrices were randomly permuted, using the `mantel.rtest()` function in R. See Appendix C for additional analyses.

We also tested the relationship of the matrix of overlaps of empirical tail windows with a matrix of average measures of Bantu ancestry for the populations of each pair (with a Mantel test as above). To calculate this ancestry measure, we first averaged the proportions of Bantu ancestry, inferred from the population structure analyses of Henn et al. [103], over all individuals in each population separately. Then, for each population pair, we averaged the two population Bantu ancestry values.

### 4.2.4 Coalescent Simulations

We performed genome-wide simulations of XP-EHH and iHS using the coalescent simulation program `msms` [72] under models that were neutral and models that included positive selection. We simulated a three-population model: one population served as the reference population for XP-EHH, and the other two were the populations for which XP-EHH and iHS were calculated.

We assumed an equilibrium model and used the formula \((1 - F_{ST}) = (1 - \frac{1}{2N_e})^t\), where \(t\) is the divergence time in generations, to translate \(F_{ST}\) values to divergence times [115]. To mimic the XP-EHH CEU statistic, we simulated an \(F_{ST}\) of 0.15 between the reference population and the two daughter (African) populations (see \(F_{ST}\) between African populations and Tuscans, Table C.2). We simulated daughter population divergence times to correspond to \(F_{ST}\) values of 0.01, 0.05, and 0.10, spanning the range of \(F_{ST}\) values between the African populations in our dataset (Table C.2). We assumed 12,000 as the effective size \((N_e)\) of each population ([238]; see Appendix C and Figure C.12 for details.)

We did not rigorously model the ascertainment strategy of the Illumina genotyping platform, as we were not primarily interested in its precise effect [68]. As a proxy, we first removed all simulated SNPs with frequency less than 0.05 in the “European” reference population. We calculated allele frequencies for each SNP across all African populations of our dataset and grouped them by minor allele frequency (MAF) in bins of size 0.05; we then randomly thinned the simulated SNPs to match this MAF distribution and observed SNP density [233]. Although this is only an approximation to the properties of the SNPs of our dataset, it should be sufficient for a window-based analysis of haplotype statistics [51].

Using these simulated and pruned SNP data, in each of the two daughter populations we calculated iHS and XP-EHH using the first population as a reference. We followed the same procedure as for observed data: we normalized statistic values across the genome, broke the genome into 100 kb windows, and calculated empirical p-values for windows based on their numbers of SNPs. We calculated the overlap in tail windows of the two daughter populations using the p-value cutoffs 0.05, 0.01, 0.005, and 0.001. We repeated this genome-wide simulation for 100 independent replicates for each daughter-population \(F_{ST}\) value (0.01, 0.05, and 0.10).

We simulated 20 haplotypes from each population. For neutral simulations, we simulated segments of 10-Mb and concatenated 300 of these segments to mimic an entire genome (3,000 Mb in total). We also simulated data under a model of a selective sweep occurring in the first daughter population; the second daughter
population and the reference population were assumed to evolve neutrally. Since we used an empirical outlier approach, we simulated only 50% of the genome under this model of selection (300 5-Mb segments); we simulated another 150 10-Mb segments under neutrality for a total of 3,000 Mb (see Appendix C Methods).

For simulations under selection, the position in the middle of each 5-Mb segment (at 2.5 Mb) was the selected site, although the site was not sampled. The positive selection strength of the selected allele was $s = 0.2$; while unrealistic, this allowed for efficient simulation of very strong selective sweeps that could be detected with a high true positive rate by both iHS and XP-EHH (see Appendix C Methods). Based on preliminary simulations, we set the time of selection as 62.4 generations before the present, again allowing both XP-EHH and iHS to have reasonable ability to detect the sweeps (see Appendix C Methods). We only included simulations where the frequency of the selected allele was non-zero at sampling time (i.e., it had not been lost), and where the selected allele had arisen only once (i.e., at a frequency $\frac{1}{N}$ for a hard sweep); we discarded all others. See Appendix C Methods for additional details.

For each simulation, we calculated the true positive rate as the percentage of genomic windows in each empirical tail that were simulated under selection (i.e., that were genomic windows within a 5-Mb selection simulation segment). Because signals of selection may not extend across an entire 5-Mb, this may be an overestimate of the true positive rate, and an underestimate of the number of false positives.

Limitations of the msms program at the time of analysis prevented reliable simulation of parallel selective sweeps at the same site in both daughter populations (Gregory Ewing, personal correspondence).

### 4.2.5 Gene Annotation Enrichment for Haplotype Statistics

To determine whether Gene Ontology (GO) biological process terms (http://www.geneontology.org) were significantly over-represented within populations in genomic regions with the most extreme selection statistics, we implemented a permutation approach similar to that of Begun et al. [22] (see Figure 4.1 and Appendix C Methods). To assess significance in the top $x\%$ of genomic windows, we randomly sampled $x\%$ of genomic windows from the genome, accounting for the number of SNPs within each window as when assigning p-values to observed windows (see Figure 4.1), and obtained the number of occurrences of each GO term. The distribution of the number of occurrences of a GO term over 20,000 of these randomizations was used as a null distribution to obtain a p-value for its number of occurrences in the observed top $x\%$. In addition to examining all GO terms, we examined a subset of terms present in the PANTHER database ($\approx 260$ terms, associated with 28,173 genes), studied by Voight et al. [291] (http://www.pantherdb.org; downloaded in November 2010). To account for multiple hypothesis testing of many GO terms, for each population we controlled the false discovery rate (FDR) at the 5% level using the procedure of Benjamini and Hochberg [26]. Python and R scripts used to perform the analysis are available upon request; see Appendix C Methods for more details.

We focused on enrichments in the 0.1% empirical tail for each haplotype statistic for each population, based on results from between-population comparisons of empirical tail windows (see Results and Table C.3). We merged adjacent empirical tail genomic windows and did not report GO terms that achieved significance
only because a gene spanned multiple contiguous windows. Since significant terms associated with only one genomic window merely indicate that they are rare, we also only reported terms associated with more than one window. To investigate the difference in results obtained when using alternative methods, we replicated the GO enrichment method used by Voight et al. [291] for iHS in the HapMap YRI (see the Appendix C Methods).

4.3 Results

4.3.1 Single SNP Statistics

We first examined population differentiation of single SNPs. We calculated per-SNP autosomal $F_{ST}$ over 11 African populations (see Methods, Tables C.1 and C.2), and plotted the enrichment of genic and nongenic SNPs, in comparison to their genome-wide frequencies, in bins of $F_{ST}$ values (Figure 4.2A). Frequent strong positive selection on genic regions would be expected to cause genic SNPs to display higher population differentiation on average than nongenic SNPs [13, 52]. However, SNPs with extreme $F_{ST}$ were not significantly more often genic than nongenic. At intermediate $F_{ST}$ values ($\approx 0.2 - 0.3$), a slight enrichment of genic SNPs was significant.

$F_{ST}$ calculated across all 11 populations could have poor power to identify loci under positive selection if populations have experienced mostly local, or independent, selection. For each SNP we also calculated the derived allele frequency difference ($\delta$) between all ($55, or \binom{11}{2}$) pairs of populations, following Coop et al. [52] (Figures C.1A and B). Less than half of the pairs of populations exhibited $|\delta|$ values greater than 0.8, though most highly differentiated SNPs were genic (Figure C.1C). The most highly differentiated SNPs ($|\delta| > 0.8$) were found in comparisons with the Hadza and Mbuti Pygmies, where one allele was nearly completely absent (see also Figure C.2D). There were also very few SNPs (often less than 5) in each population pair’s most extreme $\delta$ value bins. As a result, when assessing significance of enrichment for individual pairs, a majority of the bootstrap samples of genomic regions did not even contain SNPs with extreme $\delta$ values – invalidating the bootstrap confidence intervals (see Appendix C Results and Figure C.2). Thus, for no pair did we find a significant enrichment of genic SNPs for the most extreme $\delta$ values. Genic enrichments in the next-most-extreme $\delta$ bins, with valid bootstrap confidence intervals, were not significant.

Upon averaging $|\delta|$ over all pairs of populations to obtain a measure of differentiation similar to $F_{ST}$ for each SNP, enrichment of genic SNPs at high values was also not significant (Figure 4.2B). There was a significant enrichment of genic SNPs at low averaged $|\delta|$ values, possibly due to purifying selection; however, this enrichment was not observed for $F_{ST}$ calculated over all 11 populations (Figure 4.2A; see also Figure C.1C).

We also examined genic enrichments for $\delta$ between the HapMap Tuscans and HGDP Yorubans, as a comparison to the analysis of Coop et al. [52] for Perlegen Type A SNPs between the HapMap CEU and YRI (Hinds et al. [110]; Figure C.3). Unlike among African populations, confidence intervals were valid, since all
bootstrap samples contained SNPs in each \( \delta \) bin. However, a genic enrichment among SNPs with high \( \delta \) was not significant for either the Tuscans or the Yorubans.

For each African population pair, we examined the relationship between their mean \( F_{ST} \) (calculated over all SNPs) and the upper 99.99% tail value (over all SNPs) of \( |\delta| \) (Figure 4.3). As mean \( F_{ST} \) between African populations increased, the 99.99% tail value of \( |\delta| \) also increased (as in Coop et al. [52]; see also Figure C.4). Under a neutral equilibrium model, we simulated extreme \( |\delta| \) values for all pairs of populations given their observed pairwise \( F_{ST} \) values (following Coop et al. [52]; see Methods). The simulated data, as well as the lowess smoothed curve, corresponded closely to observed data (Figure 4.3).

4.3.2 Haplotype Statistics

Haplotype-based statistics iHS [291] and XP-EHH [233] calculated with phased genotype data may be less biased by SNP ascertainment than SNP-based measurements such as \( F_{ST} \) [51]. For a total of four statistics, we used HapMap CEU, HapMap YRI, HapMap MKK, and the \#Khomani Bushmen (KHB) as reference populations (see Methods). Candidate genomic windows for positive selection should have selection statistic values in the tails of an empirical distribution across all genomic windows, if such selection is relatively rare [106].

Comparison of Empirical Tail Windows

Genome-wide selection scans often compare candidate signals across multiple populations [103, 207]. An example of this is Figure 4.4, which shows the top five 100 kb genomic windows with the lowest empirical p-values for XP-EHH MKK in each population (see also Figures C.5, C.6, C.7, C.8).

For an analysis of patterns of sharing of selective sweep statistics beyond what can be obtained from Figure 4.4, for every pair of populations we calculated the number of 100 kb windows in the 1% tail of the empirical distributions of both populations (approximately 250 tail windows). As a population pair’s mean \( F_{ST} \) increased (i.e., as populations became more diverged), the number of windows appearing in the top 1% of both populations decreased predictably. Mantel correlations between matrices of mean \( F_{ST} \) and number of overlapping 1% tail windows between pairs of populations were significant for iHS, XP-EHH CEU, XP-EHH MKK, and XP-EHH KHB (all except XP-EHH YRI) (Figures 4.5 and 4.6 and Table C.3). We observed a similar, though less pronounced, relationship between mean \( F_{ST} \) and extent of sharing for the 0.1% tail (Figure C.9 and Table C.3). For most statistics, the extent of sharing of candidate sweep signals thus appeared to be difficult to separate from levels of population divergence (as measured by \( F_{ST} \)).

As overlap of tail genomic regions for the XP-EHH YRI statistic did not appear to be related to population divergence, we suspected that patterns might be more related to the Bantu expansion, which resulted in extensive admixture within Africa over the past several thousand years [24, 28]. We found a significant negative Mantel correlation of tail overlap (in the top 1%) with the average Bantu ancestry of the populations in the pair (p-value = 0.0006; see Table C.3 and Figure C.11). The more gene flow a pair of populations had
with the Bantu (for which the YRI are considered a proxy for a source population), the fewer empirical tail genomic windows they shared.

We then assessed whether between-population patterns in the empirical tails, expected to reflect positive selection, were qualitatively different from between-population patterns across all genomic windows (summarized by the correlation of two populations’ window statistic values across the whole genome). For all statistics, there was a significant positive Mantel correlation between 1% tail overlap and correlation of the two populations’ genome-wide window statistics (Table C.3, Figure C.10). Outlying genomic windows thus did not exhibit between-population patterns that were different from those of genome-wide windows. A more extreme tail may reveal more meaningful signatures of selection, as overlap in the top 0.1% was less related to the genome-wide correlation of two populations’ window statistics (Table C.3; see Appendix C Results for related analyses).

Neutral Coalescent Simulations

To understand the relationship between $F_{ST}$ and the extent of overlap of empirical tail genomic windows under neutrality, we performed neutral coalescent simulations of genome-wide SNP data (see Methods). XP-EHH and iHS were calculated for two daughter populations with various divergence times, and a third outgroup “reference” population was simulated to mimic the XP-EHH CEU statistic (Figure C.12).

As the two daughter populations became more highly diverged (i.e., had a larger $F_{ST}$ value), the number of windows overlapping in their 1% empirical tails for XP-EHH decreased predictably (Figure 4.6 and Figure C.13). This trend was also apparent for iHS, although there was no significant difference in overlap for different $F_{ST}$ values (Figure 4.5 and Figure C.14). This negative relationship between tail overlap and $F_{ST}$ was even stronger for less extreme empirical tails (5%), and weaker for more extreme tails (0.1%), as in our observed data; this is likely because more extreme tails contained fewer windows (Figures C.13 and C.14). Thus, depending on their levels of divergence, populations can be expected to share some genomic windows in their empirical tails, even without positive selection acting in parallel on those windows.

In both observed and simulated neutral data, there was a similar relationship between population divergence and the extent of sharing of outlying genomic windows between populations. However, there were more overlapping tail windows in the observed data than in the neutral simulations for both XP-EHH and iHS (Figures 4.5 and 4.6). While the simulations were modeled after the XP-EHH CEU statistic, simulated overlap values were comparable to those observed for XP-EHH KHB.

Selective Sweep Coalescent Simulations

To understand the relationship between $F_{ST}$ and the extent of overlap of empirical tail genomic windows given selective sweeps, we performed coalescent simulations as above with hard sweeps occurring in one of the two daughter populations, and the other populations evolving neutrally (see Methods). We simulated sweeps with a relatively high selection coefficient ($s = 0.2$) to enable them to be detected by both XP-EHH and iHS and to allow more efficient simulations (see Appendix C Methods). We also simulated a relatively
recent selection time (≈ 62 generations ago), preventing most selected alleles from reaching fixation so that they could be detected by iHS (Appendix C Methods, Figures C.15 and C.16; Voight et al. [291] and Sabeti et al. [233]).

For both XP-EHH and iHS, there was no relationship between \( F_{ST} \) and the number of 100 kb windows overlapping in the tails of the daughter populations. The extent of overlap appeared consistent with that expected by chance under independence (Figures 4.5, 4.6, C.15 and C.16). For XP-EHH, the number of overlapping windows was considerably lower than that of observed data and neutral simulations (Figure 4.6); for iHS, the number of overlaps was only slightly lower (Figure 4.5).

In simulations of less recent positive selection (144 generations ago), the true positive rate of detecting sweeps for iHS was low – genomic windows simulated under neutrality, as well as selection, appeared in the empirical tails (Figure C.18, see also Figure C.17). Because of these false positive windows, the relationship of \( F_{ST} \) with the extent of sharing of tail windows between populations was more apparent (Figure C.18, see Appendix C Results). With the same simulation parameters for XP-EHH, the true positive rate remained high, and there was no relationship between \( F_{ST} \) and overlap of tail windows.

### 4.3.3 Enrichment of Biological Functions

Another approach used to integrate results of genome-wide selection scans asks whether any biological functions (Gene Ontology biological process terms, or GO terms) appear to be more frequent targets of positive selection than expected by chance (http://www.geneontology.org). Within each population, we used a novel permutation procedure (see Methods and Figure 4.1) to search for biological functions significantly associated with genes in outlying genomic windows (according to values of iHS, XP-EHH CEU, YRI, MKK and KHB). We focused on the 0.1% tail, motivated by the lower correlation of between-population patterns of this more extreme tail with those of the whole genome (see previous section and Table C.3).

Many GO terms appeared to be under strong selection within populations (Tables 4.1 and 4.2 and Table C.4). For example, the Mbuti Pygmies showed an enrichment of genes related to phosphate transport, and the Namibian San showed an enrichment for terms related to apoptosis (Table 4.1). In the Mandenka, several GO terms related to antigen processing were enriched (Table 4.2). Several terms appeared to be under selection in multiple populations. In the Biaka Pygmies, Mbuti Pygmies, ≠Khomani Bushmen, Mandenka, and HapMap YRI, terms related to transcription were enriched primarily in the tails for XP-EHH CEU (Table C.4). In the Mandenka and Yoruba, immune response was significantly enriched in top windows for XP-EHH YRI and XP-EHH MKK, respectively. In the Namibian San and the Yoruba, the term brain development was enriched, while the ≠Khomani Bushmen and HapMap YRI both showed enrichment for ubiquitin cycle. The term sodium ion transport was enriched in both the Kenyan Bantu and HapMap YRI, which could motivate future studies in light of the salt retention hypothesis (Table C.4; Thompson et al. [271]). No GO terms were significant in the Hadza for any statistic, and terms were only significant for the Kenyan Bantu at less extreme p-value cutoffs (see Appendix C Results and Table C.4).
CHAPTER 4. SELECTIVE SWEEPS AND AFRICAN POPULATIONS

We found no significantly enriched terms in the empirical tails of iHS for the HapMap YRI at any significance level (Table 4.2). In contrast, in their iHS analysis of the same population, Voight et al. [291] found enrichment of several terms including olfaction and MHC-1 mediated immunity. There could be several reasons for this lack of agreement. While we searched for enrichment among genomic windows, Voight et al. [291] searched for enrichments among genes using statistics from 50 contiguous SNPs (see Appendix C Methods); they also analyzed a larger set of SNPs. Since we were able to replicate the results of Voight et al. [291] using our SNP dataset and their enrichment procedure (see Appendix C Methods and Results), the primary cause of the difference in results appeared to be the statistical method for assessing enrichment.

4.4 Discussion

We computed statistics for hard selective sweeps and patterns of SNP variation in a diverse set of human African populations following the approaches of many recent genome-wide scans for positive selection [3, 207, 231, 232, 233, 234, 291]. Patterns of allele frequency differentiation and sharing of haplotype selection signals among African populations were consistent with expectations under neutrality, suggesting the presence of false positive signals in empirical tails. We confirmed that the effects of positive selection do contribute to some empirical tail loci by identifying biological functions significantly associated with them using a novel permutation procedure. However, population history can confound the interpretation of selection statistics, suggesting that greater care is needed in attributing broad genetic patterns to human adaptation.

4.4.1 Population Differentiation

Positive selection does not appear to have substantially shaped present-day allele frequency differences among the African populations in our dataset (Figures 4.2 and 4.3 and Figures C.1 and C.2). SNP density may partially explain this finding. While Coop et al. [52] found a significant enrichment of genic SNPs among those highly differentiated between HapMap CEU and YRI for Perlegen Type A SNPs (which have a uniform ascertainment based on a multi-ethnic discovery set; Hinds et al. [110]), our analysis of just the Illumina 550K genotype data for HGDP Yorubans and HapMap Tuscans showed no significant enrichment (Figure C.3). We were also unable to identify a significant genic enrichment for SNPs with extreme $\delta$ values for any pair of African populations (Figures C.1A, C.1B, C.3 and Appendix C Results). This suggests that in addition to SNP density, ascertainment bias could contribute to our inability to identify significant genic enrichments.

In another analysis we discovered phenomena that may be unique to African populations. Extreme values of allele frequency differences between pairs of African populations were closely related to a pair’s overall divergence (as measured by mean $F_{ST}$), in near perfect concordance with data simulated under a neutral equilibrium model (Figure 4.3 and Figure C.4). (Since selective sweeps are believed to have been relatively rare in recent human evolution, mean $F_{ST}$ is likely a valid measure of divergence [106].) In contrast, a study of worldwide populations using a comparable set of SNPs [52] suggested a stronger role of positive
4.4. DISCUSSION

Selection in increasing levels of allele frequency differentiation beyond that predicted by population history. In addition, most strong signals of differentiation in our dataset resulted from extreme allele frequencies in the Hadza and Mbuti Pygmies, which have each undergone recent extreme bottlenecks (see Henn et al. [103], Appendix C Results, and Figure C.2D).

Our findings from SNP array data do not support positive selection as a more important determinant of allele frequency differentiation among African populations than population history. This cautions that population differentiation measures used with an empirical outlier approach could lead to false positive selection signals in this dataset.

4.4.2 Comparison of Empirical Tail Windows

With the haplotype statistics iHS and XP-EHH, we examined empirically-outlying genomic windows of 100 kb as candidates for selection. A common strategy used to interpret the generated lists of putatively selected loci is to compare them between populations (see, for example, Figure 4.4). Such comparisons have been used to generate hypotheses, such as shared selective pressures, to explain why particular regions of the genome might be under selection in multiple populations [52, 135, 207]. Analyzing coalescent simulations (performed with msms [72]) under models of both neutrality and local adaptation, we find that conclusions from comparative methods can be misleading if there are a large number of false positive signals.

Under neutrality, empirical tail genomic windows are expected to be shared between populations, with the amount of sharing determined by a population pair’s level of divergence (Figures 4.5 and 4.6). In simulations under a model of local adaptation, with selective sweeps occurring in only one population of a pair, there was no relationship between divergence and the extent of sharing of outlier haplotype statistics if the statistics had a high ability to detect sweeps (i.e., a high true positive rate and a low false negative rate). Under this scenario, the amount of sharing of haplotype signals between populations was that expected by chance under independence regardless of \( F_{ST} \) value (Figures 4.5, 4.6, C.15, and C.16). In contrast, if haplotype statistics had insufficient ability to detect selection (due in our simulations to the timing of selection), a slight negative relationship was expected (Figure C.18). This occurred because false positive neutral genomic windows, rather than only those under selection, appeared in empirical tails. Lower power and a higher false positive rate may indeed be expected for sweeps with more realistic parameter values than those we simulated (chosen for efficiency and to maximize the ability of iHS and XP-EHH to detect selection, see Methods) [231, 233, 291].

Due to limitations of msms at the time of analysis, we were unable to simulate selection under a model of parallel, convergent selective sweeps at the same locus in multiple populations (see Methods). Given the above results, if statistics have a high true positive rate, empirical tail signals shared between populations should correctly reflect populations’ true amount of convergent adaptation and appear to have no relationship with population history.

In African populations, we observed patterns of between-population divergence and overlap of outlier loci that were very similar to those simulated under neutrality (Figures 4.5 and 4.6; while simulated data
did not precisely match observed data, differences were likely due to the simulations’ simplicity and non-uniform SNP density of observed data). The observed strong negative relationship between $F_{ST}$ and extent of overlap of tail windows suggests that many signals in the 1% empirical tails may not be driven primarily by local (or convergent) adaptation – under which only a weak relationship would be expected (Figures 4.5 and 4.6). (In the more extreme 0.1% tail, this negative relationship was less pronounced.) Furthermore, between-population empirical tail overlap for the top 1% seemed to follow genome-wide correlations (Figure C.10 and Table C.3). In light of our simulations, false positive outlier loci seem to be a plausible explanation for the close relationship of population history with selective sweep patterns in African populations.

Our results agree with Coop et al. [52] and Pickrell et al. [207], who found that selective sweep signals tend to cluster by broad geographic and continental regions, and who emphasize that demography can influence haplotype selection statistics (see also [51, 129]). When methods do not fully or accurately identify signals of selection, sharing of selection signals between populations can have an important contribution from population history.

Between-population patterns of observed data and data simulated under both selection and neutrality can thus illuminate the prevalence of false positives in observed data. Previous studies may have spuriously attributed a common sweep signal among populations to selection because they did not examine these broader patterns. In light of this, summaries such as Figure 4.4 can be misleading (see Pickrell et al. [207]). Determining significance of selection with empirical p-values has also elicited criticism; since the true proportion of the genome that is under natural selection is unknown, so are the most appropriate empirical cutoffs (see Results; [3, 106, 159, 269, 272, 273]).

Finally, it is not usual to calculate XP-EHH using multiple, diverse reference populations, as we do with XP-EHH versus Europeans, Yorubans, Maasai and ≠Khomani San. Previous studies of XP-EHH in Africans have primarily used an out-of-Africa reference population (i.e., XP-EHH CEU), which identifies different signals of selection from when using a more closely-related reference population (see Appendix C Results). Divergence time of the reference population also affects the amount of sharing between populations of outlying genomic windows: fewer tail windows are shared for XP-EHH using as references the Maasai and Yorubans, which are more recently diverged compared to Europeans and ≠Khomani Bushmen (see Figure 4.6 and Figures C.5, C.6, C.7, C.8). As seen with the Yorubans as a reference, gene flow may also play a role (Figure C.11). Thus, divergence time and demography of the reference population should be considered when interpreting comparisons of XP-EHH between populations.

### 4.4.3 Enrichment of Biological Functions

As an alternative approach to identifying important signals of selection, we searched for biological functions over-represented in the tail of genomic window distributions for the iHS and XP-EHH. Motivated by the haplotype statistic analyses and simulations, which suggested a high rate of false positives in the 1% empirical tail, we selected the top 0.1% as a more appropriate cutoff (see Results). Using a statistically rigorous permutation procedure (see Methods), we detected enrichment of many Gene Ontology (GO) biological process
terms in outlying genomic windows of nearly all studied populations (except for the Hadza, likely because of their recent severe bottleneck; Henn et al. [103]); Tables 4.1 and 4.2. Many of these enrichments suggest previously-unreported selective pressures within individual populations (see Results): examples include the terms blood coagulation in the Sandawe of Tanzania and response to virus in the Maasai.

Such findings increase our confidence that some signals inferred using selective sweep statistics do indeed identify areas of the genome that are under selection in African populations. Immune response, for instance, was significantly enriched in empirical tail genomic windows of both the Mandenka and Yoruba, for which other terms such as B-cell differentiation and antigen processing and presentation were also enriched. Since the Mandenka and Yoruba both inhabit western Africa, this may indicate strong pressure from infectious or parasite-related diseases in this region, or may reflect recent selective pressure associated with adoption of agriculture only 5,000 years ago [89]. These enrichments agree with recent studies, which have highlighted pathogens as a primary driver of recent human evolution (Fumagalli et al. [87], Novembre and Han [190]), and which have presented evidence that HLA and other disease-related genes are under selection in West Africans [29].

By preserving the genomic features of our dataset, our permutation method accounts for important factors such as non-uniform SNP density, different gene lengths, the existence of regions of the genome where genes with similar functions are clustered, and the method used to assign empirical p-values to genomic windows (Figure 4.1). While our method follows that of Begun et al. [22], permutation methods are not often implemented for GO enrichment analyses in human genetics [170]. Those commonly used are designed for microarray studies, which present different statistical challenges from analyses of genomic windows, within which several genes can exist [117, 118, 170]. Fumagalli et al. [87], for example, used the method GONOME [256]; although this method accounts for gene size, it does not account for gene clustering within genomic windows (see also [80, 96, 97, 114, 153, 262, 296]). While McLean et al. [169] developed a method for enrichments among genomic segments, it is not appropriate for larger genomic windows containing multiple genes.

Studies have acknowledged that these biases could result in false positive signals of enrichment. Voight et al. [291], for example, suggested that gene clusters like the HLA region may have contributed to their significant results (see also Metspalu et al. [170]). In fact, our results for outlying genomic regions for iHS in the HapMap YRI differed from those of Voight et al. [291], likely due to our use of different statistical methodology.

### 4.4.4 Complicating Factors for Studies of Selection

It has been suggested that selection may be more common outside of Africa than within Africa [52, 113, 134, 140, 179, 259, 303]. While we have identified plausible signals of selection in our African dataset, demographic and technical factors complicate the application of selection statistics.

The absence of observed genic enrichment for highly differentiated polymorphisms may be due to ascertainment bias, as the SNPs analyzed here were ascertained primarily in European populations and are...
 unlikely to be in LD with unique alleles that may be under selection in Africans [68, 103, 188, 291]. A similar explanation has been invoked to explain the failure of genome-wide association studies (GWAS) in some African populations, e.g., for studies of malaria [184]. The lower LD of African populations may also make it difficult to identify SNPs undergoing allele frequency differentiation due to selection, to detect signatures of increased LD due to a sweep, and to infer chromosomal phasing accurately [103]. To address some of these concerns related to LD, we analyzed genomic windows of 100 kb, in contrast to previous studies that have used windows of twice this size [103, 170, 207]. Our analyses suggest that SNP density can also affect statistical significance, even with a relatively large dataset of nearly 500,000 SNPs and in populations where ascertainment bias is minimal (Figure C.3).

Additionally, small sample sizes in some populations may have reduced our ability to accurately estimate allele frequency differentiation or detect hard selective sweeps. The power of the iHS and XP-EHH statistics to detect selective sweeps in African populations may also be limited if the selective events have occurred too far in the past (greater than ≈30,000 years ago; Sabeti et al. [231]). One such example could be the transition to a desert environment in the Khomani Bushmen, which may have exerted selective pressure too long ago to be detected.

All of these factors could contribute to the difficulty of unequivocally inferring positive selection. While SNP density and ascertainment bias on existing genotyping arrays are influential, whole-genome sequencing may not be a panacea for the issues we raise (preliminary results; see also Lachance et al. [147]). The sensitivity of widely-used selective sweep statistics to population history, selection timing, phasing error, lower LD, and sample size will likely remain problematic.

Although we focus exclusively on hard selective sweeps, other models of selection, such as polygenic adaptation or adaptation based on standing genetic variation, also have the potential to uncover signals of positive selection [53, 87, 95, 96, 97, 106, 211, 212, 268]. Awareness of the issues raised here should also improve analyses of these alternative models of selection.

### 4.5 Data Access

All genotype data has been published in previous work and is publicly available. See Henn et al. [103], International HapMap 3 Consortium et al. [121], International HapMap Consortium [122], International HapMap Consortium et al. [123], Li et al. [157], Schuster et al. [242].

### 4.6 Acknowledgments

We would like to thank Nandita Garud, Laura Rodríguez-Botigué, Philip Greenspoon, Philipp Messer, David Enard, Martin Sikora, and Jonathan Pritchard for helpful comments and discussions. J.M.G is supported by the National Science Foundation (NSF) Graduate Research Fellowship, grant number DGE-0645962. B.M.H. and C.D.B. are supported by National Institutes of Health (NIH) grant R01HG003229. C.R.G. is supported
in part by NIH Training Grant T32 GM007175 and the University of California, San Francisco Chancellor’s Research Fellowship. J.M.K. is supported by NIH Grant T32HG000044. This research is supported in part by NIH grant GM28016 to M.W.F.
4.7 Figures and Tables

Figure 4.1: Permutation procedure used to assess significance of enrichment of Gene Ontology (GO) terms in the top $x\%$ of 100 kb genomic windows. Procedure is performed for a particular statistic (iHS, XP-EHH CEU, XP-EHH YRI, XP-EHH MKK, or XP-EHH KHB) for a particular population. Note that the third and fourth boxes demonstrate the method used to assign p-values to genomic windows based on the number of SNPs per window; see Methods for details.
4.7. FIGURES AND TABLES

Figure 4.2: Enrichment of genic and non-genic SNPs for values of $F_{ST}$ and for average values of derived allele frequency difference ($|\delta|$) between pairs of populations. (A) Enrichments of genic and non-genic SNPs for values of $F_{ST}$ (calculated across all 11 populations for each SNP), divided into bins of width 0.05. Dashed lines indicate upper and lower 95% confidence intervals for each of the genic and non-genic categories, assessed by bootstrap re-sampling (see Methods). (B) Enrichments for average $|\delta|$ values ($|\delta|$ is calculated between all pairs of populations for each SNP; average is taken over the pairs), divided into bins of width 0.03. Dashed lines are as in (A).
Figure 4.3: **Extreme $|\delta|$ values versus mean $F_{ST}$ for pairs of populations.** Each point is a different population pair (55 pairs in total). Y-axis is the upper 99.99% tail value (over all autosomal SNPs) of derived allele frequency difference ($|\delta|$) between the pair of populations; X-axis is mean $F_{ST}$ (over all SNPs) between the two populations. Black points are observed data; red points are the mean of data points simulated using the beta-binomial method (described in Methods) over 25 independent simulations (error bars indicate +/- 2 standard deviations over the simulations). Best-fit lowess curves are drawn through the points of the same color.
### 4.7. FIGURES AND TABLES

#### Figure 4.4: P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH MKK statistic in each population. Red indicates \( p < 0.01 \), orange indicates \( p < 0.05 \), yellow indicates \( p < 0.10 \), and white indicates \( p > 0.10 \). “Genes” column lists the genes located within the indicated windows.
Figure 4.5: Number of shared 100 kb genomic windows in the top empirical 1% for iHS vs. mean $F_{ST}$ for pairs of populations. X-axis is mean $F_{ST}$. Y-axis indicates the number of shared windows in the top empirical 1%. (A) Each point is a different population pair (some points are labeled, see key). A best-fit lowess curve is drawn through the points. Significance of the p-value for the Mantel correlation between the $x$ and $y$ variables is indicated in the upper right corner: “NS”: not significant, “*”: $< 0.05$, “**”: $< 0.01$, “***”: $< 0.001$. (B) Boxplot of overlap in the top empirical 1% of iHS between populations with the indicated simulated $F_{ST}$ values, for 100 independent neutral simulations. (C) Means +/- standard errors of overlap, for the neutral simulations of part (B). (D) Boxplot of overlap in the top empirical 1% of iHS between populations with the indicated simulated $F_{ST}$ values, for 100 independent simulations of a selective sweep in only one population with selection strength $s = 20\%$ and time of selection 62 generations ago. (E) Means +/- standard errors of overlap, for the simulations in part (D).
4.7. FIGURES AND TABLES

Figure 4.6: **Number of shared 100 kb genomic windows in the top empirical 1% for XP-EHH vs. mean $F_{ST}$ for pairs of populations.** X-axis is mean $F_{ST}$, Y-axis indicates the number of shared windows in the top empirical 1%. (A-D) Each point is a different population pair (some points are labeled, see key and legend of Figure 4.5.) (E) Boxplot of overlap in the top empirical 1% of XP-EHH between populations with the indicated simulated $F_{ST}$ values, for 100 independent neutral simulations. (F) Boxplot of overlap in the top empirical 1% of XP-EHH between populations with the indicated simulated $F_{ST}$ values, for 100 independent simulations of a selective sweep only in one population with selection strength $s = 20\%$ and time of selection 62 generations ago.
Table 4.1: Significantly enriched Gene Ontology (GO) terms in the top empirical 0.1% of windows in each population.

<table>
<thead>
<tr>
<th>Population</th>
<th>Statistic</th>
<th>GO term*</th>
<th>Name of GO term</th>
<th>p value</th>
<th>Windows</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biaka Pygmy</td>
<td>XP-EHH CEU</td>
<td>GO:0045893</td>
<td>positive regulation of transcription, DNA-dependent</td>
<td>0.0002</td>
<td>8</td>
<td>TGFBIII, TCF3, PCRD2, NSD1, TFE3, FOXK1, CREB5, SMARCD3</td>
</tr>
<tr>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:0006412</td>
<td>translation</td>
<td>0.01035</td>
<td>3</td>
<td>AC107956.8, EIF4A1, AL121591.10</td>
</tr>
<tr>
<td>Mbati Pygmy</td>
<td>XP-EHH CEU</td>
<td>GO:0006366 (P)</td>
<td>transcription from RNA polymerase II promoter</td>
<td>0.0085</td>
<td>3</td>
<td>ILF2, NFIA, NFKB1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0007186</td>
<td>G-protein coupled receptor protein signaling pathway</td>
<td>0.02015</td>
<td>4</td>
<td>CENP12, NFIA, GPR35, NFKB1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006817</td>
<td>phosphate transport</td>
<td>0.0067</td>
<td>2</td>
<td>COL5A3, COL6A1</td>
</tr>
<tr>
<td>Sandawe</td>
<td>XP-EHH YRI</td>
<td>GO:0007596</td>
<td>blood coagulation</td>
<td>0.01185</td>
<td>2</td>
<td>ADORA2A, IL22RA2</td>
</tr>
<tr>
<td>Namibian San</td>
<td>XP-EHH YRI</td>
<td>GO:00048038</td>
<td>cAMP metabolic process</td>
<td>0.0081</td>
<td>2</td>
<td>PTH, PTHrH</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:00042981</td>
<td>regulation of apoptosis</td>
<td>0.00775</td>
<td>2</td>
<td>TNFSF3P1, MALT1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006916</td>
<td>anti-apoptosis</td>
<td>0.0135</td>
<td>2</td>
<td>MALT1, YWHAZ</td>
</tr>
<tr>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006784</td>
<td>ATP biosynthetic process</td>
<td>0.00002</td>
<td>2</td>
<td>ATP5B6, SLCA2A5, AIP6VIR2</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0001764</td>
<td>neuron migration</td>
<td>0.0025</td>
<td>2</td>
<td>HIG3, NT1, MYH10, PAFAH1B1, YWHAE, CDK5B1, RELN, TWiST1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0003036</td>
<td>actin cytoskeleton organization</td>
<td>0.01555</td>
<td>2</td>
<td>CAPT, CRK</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0007420</td>
<td>brain development</td>
<td>0.00905</td>
<td>2</td>
<td>PPT1, RELN</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:00043687</td>
<td>post-translational protein modification</td>
<td>0.00135</td>
<td>2</td>
<td>UBE2L3, UBE2E2</td>
</tr>
<tr>
<td></td>
<td>Khoi Bushmen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:00051246</td>
<td>regulation of protein metabolic process</td>
<td>0.00135</td>
<td>2</td>
<td>UBE2L3, UBE2E2</td>
</tr>
<tr>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006457</td>
<td>protein folding</td>
<td>0.0143</td>
<td>2</td>
<td>DNAJC1, PPI1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006512</td>
<td>ubiquitin cycle</td>
<td>0.0193</td>
<td>3</td>
<td>HEC6V, UBE2L3, UBE2E2</td>
</tr>
<tr>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:00068114</td>
<td>sodium ion transport</td>
<td>0.01905</td>
<td>2</td>
<td>SLCA45, ACCN5</td>
</tr>
<tr>
<td>Maasai</td>
<td>XP-EHH MKK</td>
<td>GO:00039615</td>
<td>response to virus</td>
<td>0.00223</td>
<td>2</td>
<td>CREB2P, IFNAS</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:00034293</td>
<td>response to drug</td>
<td>0.00375</td>
<td>2</td>
<td>UGT1A5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006487</td>
<td>protein N-linked glycosylation</td>
<td>0.00845</td>
<td>2</td>
<td>MGA15, KEL</td>
</tr>
</tbody>
</table>

Significance determined as described in Methods and Figure 4.1, using an FDR cutoff of 0.05, within each population. A population is not shown if no terms were significantly enriched (i.e., the Hadza and Kenyan Bantu). The “Statistic” column indicates the statistic resulting in significance. *Next to a GO term, “(P)” indicates that the term was only significant when examining the PANTHER subset of terms; if nothing is next to the term, it indicates significance only when all GO terms were examined. We count the number of windows in the 0.1% empirical tail to which each term was associated (“Windows” column); rather than reporting those windows, we report the genes within those windows that were associated with the given GO term (in the “Genes” column).
Table 4.2: Significantly enriched GO terms in the top empirical 0.1% of windows in each population, continued.

<table>
<thead>
<tr>
<th>Population</th>
<th>Statistic</th>
<th>GO term*</th>
<th>Name of GO term</th>
<th>p value</th>
<th>Windows</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandenka</td>
<td>XP-EHH CEU</td>
<td>GO:0007588</td>
<td>excretion</td>
<td>0.00125</td>
<td>2</td>
<td>HMOX1, KCNK5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0006813</td>
<td>potassium ion transport</td>
<td>0.00850</td>
<td>3</td>
<td>ABC1, KCNK11, KCNK12, KCNK5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0045893</td>
<td>positive regulation of transcription, DNA-dependent</td>
<td>0.00805</td>
<td>2</td>
<td>MAP2K3, TCF3</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0016311</td>
<td>dephosphorylation</td>
<td>0.01355</td>
<td>2</td>
<td>PPPFRS, PPPFRN2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>antigen processing and presentation of peptide or polysaccharide antigen via MHC Class II</td>
<td>0.0</td>
<td>3</td>
<td>IL2, HLA-DRB5, HLA-DRB9, HLA-DQB1, HLA-DQA1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>positive regulation of calcium-mediated signaling</td>
<td>0.0</td>
<td>2</td>
<td>FCER1, P2RX5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>response to extracellular stimulus</td>
<td>0.0</td>
<td>2</td>
<td>RASGRF4, RPS19</td>
</tr>
<tr>
<td>Yoruba</td>
<td>XP-EHH MKK</td>
<td>GO:0051056</td>
<td>regulation of small GTPase mediated signal transduction</td>
<td>0.0535</td>
<td>2</td>
<td>NF1, RAPGEF4</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0042552</td>
<td>myelination</td>
<td>0.0085</td>
<td>2</td>
<td>SBF2, GAL3S11</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0008900</td>
<td>cognition</td>
<td>5.5E-14</td>
<td>2</td>
<td>CTNS, NF1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:00074720</td>
<td>brain development</td>
<td>0.00825</td>
<td>2</td>
<td>CTNS, NF1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:00050850</td>
<td>positive regulation of calcium-mediated signaling</td>
<td>0.0002</td>
<td>2</td>
<td>FCER1A, P2RX5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>positive regulation of calcium-mediated signaling</td>
<td>0.0066</td>
<td>2</td>
<td>GCM1, ROD1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>intracellular protein kinase cascade</td>
<td>0.0091</td>
<td>2</td>
<td>GSG2, ICK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>regulation of transcription, DNA-dependent</td>
<td>0.00045</td>
<td>7</td>
<td>CAMTA1, SCMH1, TCF3, JMD2B, MCM5, AC093376.5, DLD5, DLD6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>camera-type eye morphogenesis</td>
<td>0.00015</td>
<td>2</td>
<td>NF1, RING1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>collagen fibril organization</td>
<td>0.0003</td>
<td>2</td>
<td>NF1, COL11A2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B cell differentiation</td>
<td>0.00095</td>
<td>2</td>
<td>CD79A, POUL1F1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>neurotransmitter transport</td>
<td>0.0185</td>
<td>2</td>
<td>SLC32A1, SLC5A20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ubiquitin cycle</td>
<td>0.0021</td>
<td>4</td>
<td>USP31, RING1, TULP4, HECW1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>lipid transport</td>
<td>0.0185</td>
<td>2</td>
<td>OSBP2, APOE5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>immune response</td>
<td>0.00525</td>
<td>3</td>
<td>CD79A, CC9, AL645940.4, AL645931.7, HLA-DRB1, HLA-DQA1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>positive regulation of calcium-mediated signaling</td>
<td>0.0004</td>
<td>2</td>
<td>FCER1A, P2RX3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>regulation of small GTPase-mediated signal transduction</td>
<td>0.0005</td>
<td>2</td>
<td>MAP2K2, RASGRF4</td>
</tr>
</tbody>
</table>
Evolution of Height and Skin Pigmentation in the ≠Khomani Bushmen

Julie M. Granka*, Christopher R. Gignoux§, Cedric J. Werely†, Marlo Möller†, Justin Myrick**, Jeffrey M. Kidd §§, Eileen G. Hoal†, Carlos D. Bustamante ††, Marcus W. Feldman*, Brenna M. Henn‡

* Department of Biology, Stanford University, Stanford, CA, USA
§ University of California, San Francisco, CA, USA
† Division of Molecular Biology and Human Genetics, University of Stellenbosch, South Africa
** Department of Anthropology, University of California, Los Angeles, CA, USA
 §§ Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA
 †† Department of Genetics, Stanford University, Stanford, CA, USA
‡ Department of Ecology and Evolution, State University of New York, Stony Brook, NY, USA

Abstract

No two traits are as immediately recognizable for their variability among and within human populations as height and skin pigmentation. Many evolutionary hypotheses have been proposed to explain their variability, but relatively little is known about the genetic basis of these traits across populations. The KhoeSan hunter-gatherers of southern Africa, believed to be one of the world’s most ancient human populations, have been the focus of extensive anthropological research for their subsistence strategy, “click” language, and great genetic diversity. While anthropologists have also documented the short stature and light skin pigmentation of the KhoeSan in comparison to most other African populations, the genetic basis of these phenotypes in this population remain completely unknown. For the first time, we analyze the heritability of height and skin pigmentation in 103 ≠Khomani Bushmen individuals of the Northern Cape of South Africa, and perform a genome-wide association analysis for these traits using almost 300,000 single nucleotide polymorphisms (SNPs). Using both genetic and kinship methods, we obtain estimates of heritability for height that agree with previous studies (≈ 0.80), find no evidence for heritability of darkening of skin pigmentation (tanning), and find that most variation in innate skin pigmentation (≈ 99%) has the potential to be explained by the SNPs in our dataset. Variation in amounts of European, Bantu, and KhoeSan ancestry strongly contribute to variation in innate skin pigmentation; however, continental ancestry does not contribute significantly to variation in height or tanning. Finally, in a genome-wide association study, we identify several loci associated
with height that have large effect sizes of \( \approx 5\text{-}10 \) cm and are located near genes previously identified to be associated with height. We find no loci with a significant effect on skin pigmentation variation; because of skin pigmentation’s high heritability and its strong relationship with ancestry, we believe admixture mapping to be more appropriate for future research. Our study emphasizes that analyses of phenotypically and genetically diverse endogamous populations have the potential to reveal novel insights into the genetic basis and evolutionary history of complex traits.
5.1 Introduction

A major aim of research in human genetics has been to characterize phenotypic variation by estimating the relative roles of genes and the environment, as well as natural selection, in affecting complex traits. No two traits are as immediately recognizable for their variability among and within human populations as height and skin pigmentation. Both easily measurable, the traits have been studied historically by physical anthropologists, and more recently by human geneticists using various methods including genome-wide association studies (GWAS), candidate gene approaches, animal models, and inferences of natural selection [14, 142, 149]. Despite their relatively frequent study, much remains to be understood about the genetic loci underlying their variation, as well as the role of natural selection and adaptation in maintaining this variation across and within human populations.

Human height shows great diversity among human populations. While the average height of European males is approximately 178 cm [287], average heights of “Pygmy” males of Africa, Southeast Asia, and South America range from 140-160 cm [204]. The latter variation has led to extensive research and a number of evolutionary explanations for the apparent convergence of short stature in Pygmies. One hypothesis is that a limitation in food resources may have exerted selective pressure for reduced body size to facilitate a reduction in caloric intake [204, 248]. Smaller stature could also have been an adaptation for increased mobility for tasks such as foraging, tree climbing, or maneuvering through dense forest cover [204]. Though it fails to explain the evolution of Pygmies in non-tropical environments, another hypothesis suggests that since thermoregulation by evaporative cooling is less effective in tropical environments, shorter stature may have evolved to generate less body heat [204]. A final hypothesis suggests that the Pygmy phenotype is a life-history trade-off: because of high mortality, growth cessation occurs earlier to allow for earlier reproduction [19, 171]. Interestingly, taller stature has itself been proposed to have been under positive selection in European populations [281].

Uncertainty regarding the environments of ancestral human populations complicates the assessment of the selective forces that have acted on human height. Knowing whether and why certain populations may have evolved towards shorter or taller stature can be aided by understanding height’s genetic basis. Candidate gene studies and genome-wide association studies have identified many loci associated with stature in human populations [20, 130, 149, 183, 192, 281, 311]. However, although height is known to be highly heritable (narrow sense heritability is believed to be ≈ 80%), the proportion of phenotypic variance explained by associated genetic variants is only ≈ 5-10% [149, 310, 313]. Characterizing the genetic basis of height in diverse populations is crucial to reconstructing its evolutionary history.

Unlike human genetic variation in general, a majority of variation in human skin pigmentation can be explained by between-population structure [168]. Skin pigmentation is highly correlated with latitude, with equatorial populations having generally darker pigmentation [14, 222]. Due to this apparent relationship with ultraviolet radiation (UVR), a number of evolutionary hypotheses for variation in pigmentation have been proposed. Darker skin can provide a protection against sunburn and cancer, as well as against photolysis of the nutrient folate [127, 168, 199, 260]. Thus, in geographic regions with high levels of UVR, dark pigmentation
5.1. INTRODUCTION

could have been a response to any number of these selective pressures. Under low UVR conditions where sufficient Vitamin D is not acquired through diet, lighter skin may have evolved as an adaptation to ensure adequate Vitamin D synthesis [199]. Additional selective pressures on pigmentation are possible, including bacterial defense and sexual selection [199, 260].

As with height, human skin pigmentation is believed to be highly heritable [14]. A number genes associated with human pigmentation, many with strong signals of selection, have been identified primarily in European and Asian populations [42, 52, 125, 125, 168, 180, 207, 233, 260, 260]. Again, debate regarding the selective pressures underlying human skin pigmentation could be resolved with an even more thorough characterization of its genetic basis, but the genetic determinants of pigmentation identified from a larger global sample will be needed [14, 125, 126, 168, 180, 199, 221, 222, 260].

We focus our study of height and pigmentation on “Bushmen” hunter-gatherers of the Kalahari desert, believed to be one of the world’s most ancient human populations based on their high genetic diversity and low levels of LD [101, 103, 143, 178, 242, 274, 275]. While the KhoeSan have long been studied by linguists for their Khoisan language (characterized by the presence of click consonants), anthropologists have also stressed the short “pygmyoid” stature of the KhoeSan in comparison to Europeans and central African Bantu speakers [133, 193, 194, 239, 277, 300, 308]. Anthropologists have also emphasized the unique skin pigmentation of the KhoeSan, yellow in tone and relatively light, intermediate between that of European and central African populations [133, 193, 194, 204, 239, 277, 285, 300, 308]. Despite their potential to contribute to knowledge of possibly ancestral variation in human height and pigmentation, the KhoeSan have not been studied jointly for their phenotypic and genetic variation.

We examine height and skin pigmentation with genome-wide single nucleotide polymorphism (SNP) data in a sample of ≠Khomani Bushmen, a KhoeSan population in the Northern Cape of South Africa. The ≠Khomani Bushmen include two Khoisan-speaking groups: the N|u, who traditionally practiced a hunter-gatherer lifestyle, and the Nama, a widespread KhoeSan group associated with the spread of pastoralism in southern Africa. That the ≠Khomani Bushmen have experienced extensive admixture with neighboring Bantu groups during the past 800 years, as well as with European Dutch settlers during the past 500 years, can be used to further explore genetic contributions to phenotypic variation via admixture patterns [202]. For example, given knowledge of worldwide height and pigmentation variation, we hypothesize that taller stature may be correlated with increased European and Bantu ancestry, shorter stature with increased KhoeSan ancestry, dark pigmentation with increased Bantu ancestry, and light pigmentation with increased European and KhoeSan ancestry. The phenotypic diversity and unique population and admixture history of the ≠Khomani Bushmen make them an ideal population for studying the evolutionary history of, as well as the contribution of different ancestries to, height and pigmentation.

In a dataset of 103 individuals, we estimate heritability, explore relationships of skin pigmentation and height with genetic ancestry, and perform genome-wide association studies to identify loci associated with each phenotype (height, innate skin pigmentation, and darkening of skin pigmentation (tanning)). Our study
underscores that analyses of phenotypic and genetic data in diverse human populations can provide novel insights into these two well-studied phenotypes.

5.2 Methods

5.2.1 Samples

Sampling of the ≠Khomani Bushmen took place in the in the Northern Cape of South Africa in the southern Kalahari Desert region (near Upington and neighboring villages) in 2006, 2010, and 2011. Institutional review board (IRB) approval was obtained from both Stanford University and the University of Stellenbosch, South Africa. ≠Khomani Bushmen N|u-speaking individuals, local community leaders, traditional leaders, nonprofit organizations, and a legal counselor were all consulted regarding the aims of the research before collection of DNA [103]. All individuals gave signed written consent, with a witness present, before participating. Individuals collected in 2006 were re-consented under an updated protocol. Ethnographic interviews of all individuals were conducted, including questions about age, language, place of birth, and ethnic group of the individual and of his/her mother, maternal grandmother, father, and paternal grandfather. We recorded the relationships between any sampled individuals if revealed during the interview. Ages of older individuals were verified with separate interviews regarding reproductive history. DNA was obtained via saliva, collected using Oragene saliva collection kits (DNAGenotek, Ontario, Canada).

A portable reflectance spectrophotometer (DermaSpectrometer DSMII ColorMeter, Cortex Technology, Hadsund, Denmark) was used to measure skin pigmentation (melanin content, \( M = \log_{10}(1/\% \text{ red reflectance}) \)) in 103 individuals. The device was standardized to 0 as suggested by the manufacturer twice a day while sampling. Five measurements of \( M \) were taken on each of the left and right upper inner arms (areas less affected by sunlight and representative of constitutive skin pigmentation), as well as on the dorsal side of the left or right wrist (an area more exposed to sunlight) for a measure of tanning [250]. Using an anthropometer, two to five replicate measurements of height (in centimeters) were taken of 102 individuals (the height of one individual could not be measured); measurements standing against a wall were not possible. Measurements of all individuals were taken by J.M.G., and recorded by a different individual. For the remainder of the analyses, we examined the trimmed phenotype means (highest and lowest values removed); we also averaged the inner arm skin pigmentation measurements over the two arms.

18 individuals were previously genotyped on the Illumina 550K platform [103] and 68 individuals were genotyped on the Illumina OmniExpress array, for a total of 86 genotyped and phenotyped individuals. (Some individuals were genotyped on the 550K platform but had no corresponding phenotype data, for a total of 103 genotyped individuals.) Overlap between the two genotype datasets for all genotyped individuals was determined to be 298,346 SNPs using the program PLINK (version 1.07; Purcell et al. [214]; http://pngu.mgh.harvard.edu/purcell/plink). Both datasets were standardized to Human Genome Build hg19 (NCBI GRCh37).
5.2. METHODS

5.2.2 Relatedness

Pedigrees of all sampled individuals were constructed based on relationships recorded during ethnographic interviews. Estimates of identity by descent (IBD) from SNPs were calculated using the program BEAGLE [35]. These estimates were used to verify all relationships inferred from ethnographic data (as in Henn et al. [104]) and to infer and confirm additional ones. Pedigree plots and kinship matrices were constructed using the kinship package in R (http://www.r-project.org).

5.2.3 Ancestry Estimation

For estimation of individual ancestry proportions, data were merged with additional samples that could be informative about the admixture history of the ǂKhomani Bushmen. Populations included from the Centre d’Étude du Polymorphisme Humain Human Genome Diversity Project (CEPH-HGDP) (sample sizes in parentheses) were the South African Bantu (8), Kenyan Bantu (11), Namibian San (5), Mozabites (29), and French (28). We also included data of 12 Namibian San individuals from Schuster et al. [242], as well as individuals from the Hadza (17) and Sandawe (28) of Tanzania, described by Henn et al. [103]. Populations included from the HapMap were: Yoruba trio parents from Ibadan, Nigeria (YRI) (100), CEPH Utah residents with ancestry from northern and western Europe (CEU) trio parents (88), and Maasai trio parents from Kinyawa, Kenya (M KK) (46 individuals; related individuals NA21384, NA21475, NA21399, NA21365, NA21362, NA21382, NA21423, NA21453, NA21615, and NA21634 were removed; see Pemberton et al. [201]) [121, 122, 123].

After merging the SNPs genotyped in the HapMap, CEPH-HGDP, and South African samples, and removing SNPs with a missing genotype rate > 5% and SNPs with minor allele frequency < 1% (i.e. singletons), a total of 284,279 SNPs remained. All datasets were merged to Human Genome Build hg19 as above.

We ran the program ADMIXTURE [5] on this merged dataset to estimate ancestral clusters and admixture proportions in all individuals. Only SNPs not in close linkage disequilibrium (LD) ($r^2 < 0.9$) were analyzed (all other SNPs were removed using the program PLINK), leaving a total of 259,506 SNPs. Since our dataset of ǂKhomani Bushmen included many related individuals (see Figure 5.1), we ran ADMIXTURE in six separate runs of unrelated individuals, using all default parameters in unsupervised mode. This ensured that sets of related individuals were not identified as ancestral clusters. We examined a range of numbers of ancestral populations ($K$), from 4 to 13, and found that a large number of ancestral clusters ($K > 7$) produced cluster inferences that were unstable.

We chose $K = 6$ as the best representation of the number of ancestral populations for the ǂKhomani Bushmen and other included populations (see Results). We repeated the ADMIXTURE inference 25 times for each of the six separate runs, and averaged the results over the 25 runs using the program CLUMPP (with 100 random input sequences) [128]. These averaged values were then aggregated from the six runs into one plot.

To verify the accuracy of the above ancestry estimation, we compared the estimated ancestry proportions
against each individual’s self-reported ancestry based on ethnographic interviews. For each individual, we created a measure of ancestry based on the reported ancestries of their maternal and paternal relatives. For an ancestry \( A \), the self-reported measure \( R \) was calculated as:

\[
R(A) = 0.5(I_{m=A} + I_{f=A}) + 0.25(I_{gma=A} + I_{gpa=A})
\]

(5.1)

where \( I_{x=A} \) indicates whether relative \( x \) (\( m = \) mother; \( f = \) father; \( gma = \) maternal grandmother; \( gpa = \) paternal grandfather) was reported to be of ancestry \( A \). Ancestries (represented by \( A \)) were: European (including the reported ethnicities “Dutch” and “Coloured”), European/Griqua (including the previous ethnicities as well as “Griqua,” another coloured South African group); San (including the reported ethnicities “Bushmen,” “San,” “#Khomani Bushmen,” and other synonyms); San/Nama (the previous ethnicities as well as “Nama”); and Bantu (including the ethnicities “Tswana,” “Black,” “Zulu,” “!Xhosa,” and “Ovambo”). The statistical package R was used to assess significance of the Pearson correlation of the self-report measures (\( R(A) \)) with ancestry as inferred by ADMIXTURE. These correlations were examined for all 103 genotyped individuals, including those for whom phenotype information was absent.

### 5.2.4 Statistical Analyses

Violin plots for height and skin pigmentation measurements of genotyped #Khomani Bushmen individuals were drawn using the `vioplot()` function in the `vioplot` package in R. As a measure of the amount of tanning of an individual during their lifetime, we subtracted each individual’s inner arm pigmentation measurement from their wrist pigmentation measurement.

Regressions of height, inner arm skin pigmentation, and tanning (wrist minus inner arm pigmentation) on sex, age, and ancestry proportions were performed using a linear mixed model, allowing all individuals to be included despite close relatedness (see Figure 5.1). The model for the phenotype, \( Y \), is:

\[
Y = X\beta + Zb + \epsilon
\]

(5.2)

where \( X \) and \( Z \) are vectors of the fixed and random effects, respectively. Sex, age, and ancestry comprise the fixed effects, and all individuals comprise \( Z \), the random effects. The \( \beta \) coefficients estimated for this model include \( \beta_{sex} \), \( \beta_{age} \), and \( \beta_{ancestry} \) for a given ancestry. The random coefficient \( b \) is distributed as \( N(0, \sigma^2\Theta) \), where the matrix \( \Theta \) corresponds to the kinship matrix (between pairs of individuals) constructed from the inferred pedigrees (see Figure 5.1). The error term, \( \epsilon \), is distributed as \( N(0, \sigma^2_\epsilon) \). Parameters \( \beta \), \( \sigma^2 \) and \( \sigma^2_\epsilon \) were estimated using maximum likelihood with the `lmekin()` function in the package `kinship` in R, for all 86 phenotyped and genotyped individuals (85 individuals for height).

Estimates of narrow-sense heritability \( (h^2) \), the proportion of phenotypic explained by additive genetic variance, were calculated using several methods (for a discussion see Visscher et al. [288] and Zaitlen and Kraft [313]). The first method used the linear mixed model above (Equation 5.2). \( h^2 \) can be estimated as \( \sigma^2/(\sigma^2 + \sigma^2_\epsilon) \), or the fraction of phenotypic variance explained by the kinship matrix. For height, fixed
5.2. METHODS

Effect covariates of sex and age were included for estimation of $h^2$; for pigmentation, no covariates were included (see Results). This analysis was performed on all 103 individuals with phenotypic information (102 individuals for height).

We also estimated heritability from the regression of offspring phenotype on parent phenotype for 18 parent/offspring pairs for which height and skin pigmentation information was available; here, $h^2$ is equal to twice the regression coefficient [162]. For height, $h^2$ was calculated using the residuals from the regression of height on sex and age (residuals from the mixed model of Equation 5.2; see Results); $h^2$ for pigmentation was obtained from the averaged arm values without accounting for covariates (see Results). Since not all parent/offspring pairs were independent (for instance, multiple children from the same parent were included in our dataset), we obtained heritability estimates using all parent/offspring pairs, as well as several independent sets of pairs.

We also estimated heritability using the GCTA method of Yang et al. [310], which calculates a lower bound for $h^2$ as the proportion of trait variability explained by genotyped SNPs. This estimate can also be interpreted as an upper bound for the fraction of phenotypic variance that can be explained by SNPs identified by a GWAS (see Zaitlen and Kraft [313]). The model of this method takes a similar form to Equation 5.2; however, the random coefficient $b$ is distributed as $N(0, a^2\Theta)$ where $\Theta$ is represented by a “genetic relationship matrix” calculated from genotyped SNPs instead of by the kinship matrix. We first estimated the genetic relationship matrix (between pairs of individuals) using all default parameters of GCTA, and with this matrix, we then used GCTA to estimate the variance parameters and calculate heritability by restricted maximum likelihood. For height, we used sex and age as covariates; for pigmentation, we used no covariates. Because this analysis requires genetic information, it was performed on only 86 phenotyped and genotyped individuals (85 individuals for height).

For each phenotype, we performed a genome-wide association study on 297,963 SNPs (with minor allele frequency $> 0.01$) in all phenotyped #Khomani Bushmen individuals, estimating a linear mixed model for each SNP using the program EMMAX [138] to account for substructure and relatedness of individuals [209]. The linear mixed model has a similar form as Equation 5.2, but includes SNP genotype as a fixed effect (with corresponding coefficient $\beta_{SNP}$); as above, a pairwise genetic relatedness matrix between individuals, approximated from genotyped SNPs, is used to represent the matrix $\Theta$ (see Kang et al. [138]). We estimated this pairwise genetic relatedness matrix between individuals using the Balding-Nichols (BN) model with EMMAX, as recommended by the authors [138], and plotted and clustered the matrix using the heatmap.2() function from the R package gplots. Parameters of the regression model were then estimated for each SNP by restricted maximum likelihood using EMMAX. For height, sex and age were included as covariates; for pigmentation, ancestry proportions in the San, European, and/or Bantu clusters (see Results) were included as covariates. Quantile-Quantile (QQ) plots of estimated p-values were plotted using the qqunif() function in the R package gap. Plots of candidate regions were created using LocusZoom (http://csg.sph.umich.edu/locuszoom/; Pruim et al. [213]), and recombination rates were plotted using 1000 Genomes Pilot 1 SNP calls for 59 Yoruba (YRI) individuals (November 2010; 1000 Genomes Project Consortium et al. [1]).
To identify the quantitative effect of identified loci on the phenotypes (see Results), we estimated effect sizes under a model where $\text{mean}(Y_{BB}) = -a$, $\text{mean}(Y_{Bb}) = d$, and $\text{mean}(Y_{bb}) = a$, where $Y_{XX}$ is the phenotype given genotype $XX$ at the locus; $B$ and $b$ are the alternative alleles at a locus [162]. For estimation of $a$ and $d$ for height loci, we set $Y$ as the residuals from a regression of height on sex and age from a linear mixed model using the kinship matrix estimated from the inferred pedigrees (Figure 5.1) as the variance/covariance matrix $\Theta$ (see above).

5.3 Results

5.3.1 Ancestry

Pedigrees of sampled individuals are shown in Figure 5.1, and includes relationships between 50 genotyped and phenotyped individuals based on IBD and ethnographic information; the remaining 36 genotyped and phenotyped individuals were not related to others in the dataset. Relationships between individuals include parent/offspring, avuncular, grandparent/grandchild, cousins, and others.

For each individual, ancestral clusters and admixture proportions in these clusters were estimated using the program $\text{ADMIXTURE}$ [5] with various values of $K$, the number of ancestral populations (Figure 5.2). Because our sample included related individuals, we ran $\text{ADMIXTURE}$ in six separate runs of unrelated individuals as described in the Methods. The $K$ with the lowest cross-validation error was 4, 5, or 6, depending on the run (Figure 5.3); cross-validation error increased greatly for $K > 6$ for all runs. We believe results at $K = 6$ to be most representative of the admixture of the $\neq$Khomani Bushmen (Figure 5.2). At $K = 6$, major admixture clusters corresponded to European (orange cluster), Bantu (light blue), Eastern African (blue), Mozabite (green), Hadza (yellow), and San (purple). In the $\neq$Khomani Bushmen, the major ancestries represented are Bantu, San, and European, with a minor amount of Eastern African ancestry. For phenotyped individuals, the mean amount of San ancestry per individual was 0.717 (standard deviation (sd) = 0.19); mean European ancestry was 0.129 (sd = 0.15), mean Bantu ancestry was 0.094 (sd = 0.114), and mean East African ancestry was 0.049 (sd = 0.043).

At $K = 7$, all but one of the six runs of unrelated $\neq$Khomani Bushmen individuals resulted in a split between the Namibian San (purple cluster), and South African San (pink). The $F_{ST}$ between these two clusters as calculated by $\text{ADMIXTURE}$ had an average of 0.054 across the five runs; this $F_{ST}$ is low compared to $F_{ST}$ between other inferred clusters (see also Henn et al. [103]).

While collecting samples, we noticed that older individuals tended to be less admixed than younger individuals. In order to statistically assess this bias in our sampling, we performed a simple linear regression of ancestry on age for phenotyped individuals using R. We found a significant positive relationship of age on the proportion of San ancestry ($p = 0.007$) and significant negative relationship of age on proportion European ancestry ($p = 0.045$) (see Figure 5.4). Similarly, Pearson correlations between ancestry and age were significant for the proportion of San ancestry ($r = 0.295; p = 0.006$) and European ancestry ($r = -0.262; p = 0.015$), but not for Bantu ancestry ($r = -0.06; p = 0.560$) or East African ancestry ($r = -0.211; p = 0.051$).
5.3. RESULTS

We also tested the concordance of our genetic admixture estimates with self-reported ancestry obtained from ethnographic interviews for all genotyped individuals (Table 5.1). Despite the crudeness of our self-reported ancestry measure which included information about only one grandparent on each of the maternal and paternal sides (see Equation 5.1), we found significant correlations between self-reported European, Bantu, and San ancestry and genetic estimates. Thus, the ancestries estimated by ADMIXTURE appear to be appropriate.

5.3.2 Phenotypes

Measurements of height and skin pigmentation (M) are shown for genotyped individuals in Figure 5.5 and for all phenotyped individuals in Table 5.2. For genotyped individuals, male and female heights were significantly different; average female and male heights were 151.11 cm and 163.64 cm, respectively (Wilcoxon test p-value = 7.43x10^{-10}). Skin pigmentation (M) did not differ by sex (Wilcoxon test p-value = 0.832); however, wrist minus inner arm pigmentation, as a measurement of tanning, was significantly different between males and females (Wilcoxon test p-value = 0.003). All three phenotypes were roughly normally distributed (Figure 5.5); for each phenotype, Kolmogorov-Smirnov tests performed in R showed no significant difference from a normal distribution (all p-values > 0.05 across all individuals as well as separately for males and females).

We used a linear mixed model to estimate the effects of sex and age on each phenotype for genotyped individuals (see Methods). For height, sex and age were significant ($\beta_{\text{male}} = 12.632, p = 0; \beta_{\text{age}} = -0.098, p = 0.003$), and were included as covariates in future analyses (see Figure 5.6). An interaction term between sex and age was not significant. Age and sex were not significant covariates for arm skin pigmentation. For measurements of wrist minus inner arm pigmentation (tanning), both sex and age were significant predictors, with males and older individuals having a greater positive difference between their wrist and arm pigments ($\beta_{\text{male}} = 6.375, p = 0.001; \beta_{\text{age}} = 0.176, p = 0$) (see Figure 5.7).

We made several estimates of narrow-sense heritability, or $h^2$, for both height and skin pigmentation (Table 5.3). Heritabilities for height estimated from parent/offspring pairs (after accounting for sex and age) ranged from 0.608 to 0.895, with very large standard errors, likely due to the small number of pairs (see Table 5.3). Heritabilities for height using a mixed model given our inferred pedigree produced a higher estimate ($\hat{h}^2 = 0.999$). The lower-bound $\hat{h}^2$ estimate calculated using GCTA is 0.588, though with a large standard error, similar to the value of 0.45 reported by Yang et al. [310].

Arm pigmentation heritability estimates using parent/offspring regressions ranged from 0.409 to 0.808, again with large standard errors (Table 5.3). Pigmentation heritability was estimated to be 0.999 using the inferred pedigree with a linear mixed model; in agreement with this estimate, the lower-bound $\hat{h}^2$ estimate using GCTA was 0.999 and highly significant. A large proportion of variation in skin pigmentation appears to be explained by genotyped SNPs. In contrast, for measurements of tanning (wrist minus inner arm pigmentation), heritabilities using parent/offspring pairs are all of small magnitude, sometimes negative, and highly variable (Table 5.3). The proportion of variance explained by genotyped SNPs for the tanning measurement
using the GCTA method (0.385) is the smallest of all three phenotypes, and is not significant. Thus, genetic architecture of tanning seems to be very different than that of height and constitutive skin pigmentation.

Using the mixed model regression of Equation 5.2, we tested whether ancestry proportions estimated from ADMIXTURE were significant predictors of height and skin pigmentation. A separate regression of height accounting for sex and age was performed on each ancestry; Figure 5.8 shows the results of the regression of height on sex, age, and proportion European ancestry. No ancestry (neither European, Bantu, San, or East African) was a significant predictor of height (Table 5.4), and the estimated $\beta_{\text{ancestry}}$ coefficients were near 0 for all ancestries. Given the relationship between ancestry and age in our dataset (Figure 5.4), we also tested for an interaction between each ancestry and age; however, neither the interaction terms nor ancestry terms were significant (results not shown). When more than one ancestry was included in the model, no ancestries were significant predictors of height (results not shown).

In contrast, ancestry does appear to have a strong effect on skin pigmentation variation. In separate regressions of pigmentation on each ancestry (not including sex and age as covariates), Bantu and European ancestry proportion were both significant; San and East African ancestry were not significant (Table 5.4). As expected, Bantu ancestry significantly increased skin pigmentation, whereas European ancestry significantly decreased skin pigmentation (Figure 5.9). Combinations of Bantu and San, European and San, and Bantu and European ancestries also yielded significant regression coefficients. However, a model including all three ancestries gave no significant coefficients, likely due to the covariance between the proportions (Table 5.4). No regression models including East African ancestry gave significant regressions (results not shown). Agreeing with our low estimates of heritability for the tanning measurement (wrist minus inner arm pigmentation), after accounting for sex and age we found no significant effect of any ancestry on the amount of tanning (Table 5.4).

We then asked whether there are large effect SNPs correlated with skin pigmentation and height in the †Khomani Bushmen. In order to correct for population and family structure in a GWAS framework, we used the program EMMAX [138]. First, a genetic relationship matrix between all pairs of individuals was calculated with EMMAX using the Balding-Nichols model in order to account for relatedness of individuals and substructure (see Figure 5.10). Pairs of individuals with high values of genetic relatedness correspond to relatives in the pedigree of Figure 5.1, indicating that the method can appropriately account for the inclusion of relatives.

We then used EMMAX to calculate a p-value and $\beta$ coefficient for each SNP in a mixed model regression (as in Equation 5.2), accounting for the genetic relationship matrix in Figure 5.10. For height, we included sex and age as covariates, and the QQ plot of Figure 5.11A shows that p-values were uniformly distributed. Several SNPs had p-values slightly more extreme than expected by chance (see Table 5.5). However, no p-values fell below the genome-wide significance cutoff of 1.68$\times 10^{-7}$ ($-\log_{10}(p) = 6.77$; based on a Bonferroni correction at the $\alpha = 0.05$ level) (Figure 5.12A).

The top SNP, rs9315851 on chromosome 13, is in an intron in the gene VWA8, which is associated with ATP binding and ATPase activity according to the Gene Ontology (http://amigo.geneontology).
5.3. RESULTS

The effect allele (C), associated with taller stature, is the ancestral allele, and is at high frequency in the ≠Khomani Bushmen (79%) (Figure 5.13A and D). Interestingly, the gene VWA8 has been associated with body weight and body mass index (BMI) in the Framingham Heart Study [81]. Located 428,145 bases away (at position 42,631,719) in the gene DGKH is SNP rs6561030, found to be associated with body height in Japanese individuals [192]. Means of residuals of height accounting for sex and age by rs9315851 SNP genotype suggest a strong effect of the locus: CC (1.631 cm); CT (-2.457 cm); TT (-11.538 cm) (Figure 5.13A). Under a quantitative genetic model for the phenotype (see Methods), \( a = 6.584 \) cm, suggesting a large effect, and \( d = 2.497 \), suggesting dominance of the C allele. However, only three individuals had the TT genotype, compared to 30 and 52 with the CT and CC genotypes, respectively.

The second top SNP, rs2967055 on chromosome 5, is located in an intronic region surrounded by many genes (Table 5.5 and Figure 5.12C). Again, the allele associated with taller stature (G) is the ancestral allele, and at high frequency (Figure 5.13B and E). Interestingly, this SNP is 247 bases away from rs2937550 (at position 36,473,237), which was found to be associated with height in a meta-analysis of African Americans [183]. One of the genes located near rs2967055 is SLC1A3, which, in addition to being associated with episodic ataxia and glutamate transport, was also found to be associated with height in the NHLBI Family Heart Study (dbGaP Study Accessions: phs000221 and phs000501). SNPs in SLC1A3 have also been found to be associated with BMI and waist-hip ratio in the same study, as well as with LDL cholesterol levels [139]. Another nearby gene is NIPBL, which is associated with abnormal skeletal growth [183]. Means of the residuals of height accounting for sex and age again suggest a strong effect of the locus: GG (1.205 cm); AG (-4.894 cm); AA (-8.835 cm) (Figure 5.13B; under the described quantitative genetic model (see Methods), \( a = 5.020 \) cm and \( d = -1.077 \), suggesting some dominance of the A allele). As with the previous example, a majority of individuals (65) had the GG genotype, while only 19 had the AG, and only 1 had the AA, genotypes.

Finally, the third top SNP, rs2531691 on chromosome 10, is in an intron of the gene ENO4 (Table 5.5). The only variation in this SNP exists in sub-Saharan Africa and the derived allele is absent in the HGDP Namibian San; HGDP populations outside of Africa are nearly fixed for the T (ancestral) allele (Figure 5.13F). As with the previous associated SNPs, the ancestral allele is associated with taller stature and is at high frequency (97%) in our sample of ≠Khomani Bushmen. According to the Gene Ontology, associations with ENO4 include glycolysis and magnesium ion binding. SNPs in ENO4 have been found to be associated with heart disease and heart failure [150]. While the mean of height residuals (after accounting for age and sex) for TT individuals (0.432 cm) is greater than for CT individuals (-9.610 cm), there are only 6 individuals with the CT genotype, and no CC homozygotes (Figure 5.13C). Nonetheless, this suggests a large effect of the locus.

In our genome-wide association study for pigmentation using EMMAX, we do not account for sex and age as covariates; however, we do account for ancestry (see Table 5.4). Because of the non-independence of European, Bantu, and San ancestry proportions, only two ancestries were included at a time (European and Bantu, Bantu and San, or European and San); we also attempted to simultaneously include all three
ancestries as covariates. In Figure 5.11B, we show a QQ plot of p-values when including only European and San ancestries as covariates; QQ plots for all other ancestry combinations are similar. There are no outlying SNPs with extreme p-values based on the QQ plots of Figure 5.11B, and no SNP p-values fell below the genome-wide significance cutoff.

5.4 Discussion

Height and skin pigmentation vary greatly within and between human populations. While they are both easily measurable traits, a great deal remains to be understood about their evolutionary history and genetic basis. Given their high genetic diversity, short stature, and light skin pigmentation, Southern African KhoeSan hunter-gatherers can provide novel insight into the ancestral features and genetic architecture of these phenotypes. In our study of the ≠Khomani Bushmen of the Northern Cape of South Africa, we leverage the population’s recent admixture with Bantu and European populations to better characterize the genetic contribution to these two traits.

5.4.1 Ancestry Estimation

Genetic estimates of ancestry reveal that the ≠Khomani Bushmen are an admixed population, with many individuals having appreciable amounts of European and Bantu admixture (as previously reported by Henn et al. [103]). Genetic admixture estimates are consistent with self-reported ethnicity information, supporting that genetic ancestry measurements accurately reflect the admixture history of this population [202, 239, 277] (Table 5.1).

5.4.2 Height

With average female and male adults heights of 151.78 cm and 163.90 cm, respectively, our measurements of height in the ≠Khomani Bushmen were slightly greater than previous measurements in various KhoeSan populations of the Kalahari (Table 5.2). For instance, measurements of southern KhoeSan populations reported by Schapera [239], Tobias [277] and Dart [57] were 5-10 cm shorter for both males and females. Our estimates agreed more closely with measurements taken in more northerly KhoeSan and Nama (or “Hottentot”) populations, although our sample had slightly greater mean measurements [193, 194, 239, 277, 300, 308]. These measurement differences can be explained by the inclusion of Nama individuals in our sample, our inclusion of admixed individuals, difference in ages of sampled individuals, or sampling bias for shorter individuals in previous studies [277]. Previous studies did, however, identify a large difference between male and female heights of ≈10 cm, as in our study (Table 5.2). As expected, height of the ≠Khomani Bushmen individuals was greater than most Pygmy populations, for whom adult male height is generally 140-160 cm [18, 20, 130, 204].
5.4. DISCUSSION

We used several methods to obtain estimates of narrow-sense heritability \( (h^2) \) for height (Table 5.3). While standard errors of our estimates were large because of small sample size, our point estimates agreed with previous estimates (\( \approx 80\% \)). Our results suggest that height in the ǂKhomani Bushmen, as in other populations, has a large contribution from additive genetic variance. However, the SNPs genotyped in our study appear to explain only a moderate proportion of this variance (as found by Yang et al. [310]; see GCTA estimate of Table 5.3). In addition to genetic factors affecting height, environmental and nutritional factors may play a role in its variance [193, 202].

Interestingly, we found no effect of ancestry on height in the ǂKhomani Bushmen (Table 5.4 and Figure 5.8). This is in contrast to Pygmy populations for whom other studies have claimed a correlation between non-Pygmy ancestry and increased height [18, 130]. However, although Becker et al. [18] found a significant positive relationship between non-Pygmy ancestry and height when pooling across several Pygmy populations from Cameroon, Gabon, and Central African Republic (including Aka, Baka, and Bongo Pygmies), they did not find a significant relationship within Pygmy populations. In addition, though Jarvis et al. [130] found a significant association of height with non-Pygmy ancestry, their results may be biased by including multiple Pygmy populations (Baka, Bakola, and Bedzan), as well as Bantu-speaking populations from Cameroon. That we were able to detect a significant relationship between ancestry and skin pigmentation (see Results) suggests that our lack of relationship for height is not necessarily due to lack of power due to small sample size. Since we obtained relatively high point-estimates of height heritability, it seems that the loci affecting height are not strongly differentiated by ancestry. Differences in height among ancestral populations (European, Bantu, and KhoeSan) may not be great enough to produce a significant signal.

In a genome-wide association study, we identified several loci of large effect potentially related to height in the ǂKhomani Bushmen (Table 5.5 and Figure 5.13A, B and C). While none of these loci reached genome-wide significance, several SNPs appeared to be outliers from a uniform distribution of p-values (based on the QQ plot of Figure 5.11A). Promisingly, the associations we identified are located near or within genes that have previously been found to have modest associations with height or other body size characteristics (see Results and Figure 5.12).

Differences in height between genotypes for the identified loci suggest large effects for the quantitative contribution of each allele, with possible dominance effects. Effect sizes under an additive quantitative genetic model (see Methods) were very large (\( \approx 5\text{-}10 \) cm), much higher than effect sizes estimated from previous studies (\( \approx 0.5 \) cm) (Figure 5.13A, B and C) [154, 183, 237, 297]. We note that for all three identified SNPs, the effect allele associated with taller height is at very high frequency in the ǂKhomani Bushmen; the few heterozygotes and homozygotes with the “short” allele could bias effect size estimates (Figure 5.13D, E, and F). Additionally, it is possible that unaccounted-for substructure (either due to admixture or relatedness) could cause spurious associations. Nonetheless, for each association (Table 5.5), the effect allele associated with increased height is the ancestral allele and major allele in the ǂKhomani Bushmen. The high frequency of the “taller” allele in the ǂKhomani Bushmen seems surprising given their shorter stature, and indicates the likely existence of additional loci underlying the overall short stature of the ǂKhomani Bushmen.
The prevailing view in human height research is that there are many loci of small effect that contribute to height variation, requiring large sample sizes to detect genetic associations [287]. Our study of a modest number of individuals suggests that loci with large effect sizes may also contribute to human height variation. Loci of smaller effect likely still contribute to height variation in the ≠Khomani Bushmen, supported by our estimates that common tag SNPs on the Illumina arrays explain 58% of the variance of height in our sample (see also Visscher [287]).

Because the alleles associated with taller stature are the ancestral allele for all three identified height SNPs, this could imply that shorter stature is a derived trait in the ≠Khomani Bushmen (Figure 5.13). A number of hypotheses regarding the evolution of human Pygmy short stature may be relevant to KhoeSan populations, such as the evolution of short stature for increased foraging mobility [204], though other hypotheses such as thermoregulation in a rainforest environment and life-history tradeoffs are not applicable. While our analyses are a promising first step, further research is necessary in order to more fully understand the evolution of KhoeSan shorter stature. Additional samples would allow us to verify the loci we have identified and to identify other loci of large or small effect that may be associated with height variation.

5.4.3 Skin Pigmentation

Skin pigmentation measurements of the ≠Khomani Bushmen as measured by a spectrophotometer (mean $M \approx 54$) are much greater than those of Europeans ($\approx 20 < M < 30$), just above those observed in Asians ($\approx 40 < M < 50$), in the low range of those of African-Americans ($\approx 50 < M < 90$), and lower than those observed in Melanesians ($\approx 60 < M < 90$) (measurements taken with the same measurement tool; Candille et al. [42], Norton et al. [189], Shriver and Parra [250]) (Table 5.2). Comparison of our measurements of skin pigmentation with previous studies of the Kalahari KhoeSan, which primarily reported qualitative measurements, were not possible [239, 300]. Age was not significantly related to skin pigmentation, confirming that measurements taken on the upper inner arm may reflect innate pigmentation instead of UV exposure. However, while previous studies have adjusted for effects of sex in pigmentation measurements, we found no significant effect of sex on inner arm pigmentation [42, 189].

Interestingly, in our study of tanning (wrist minus inner arm pigmentation) (Figure 5.5), males and older individuals tended to have a greater positive difference between the two pigmentation measurements (Figure 5.7). Age can be seen as a proxy for the amount of time that an individual has been exposed to UV radiation, causing older individuals having greater tanning. The effect of sex suggests that males in this population may experience more sun exposure than females, perhaps due to spending more time outdoors. Though less likely, it could also imply a greater intrinsic tanning ability of males compared to females.

We obtained slightly higher estimates of heritability for innate pigmentation than those of previous studies, which have ranged between 55% and 80% [85, 225] (Table 5.3). Though heritability estimates calculated from parent/offspring regressions in our study had large standard errors (as for height), the high point estimates emphasize a strong genetic contribution to pigmentation. The GCTA method of Yang et al. [310]
yielded a significant heritability estimate of 0.99, suggesting that a large proportion of variability in pigmentation may be explained by the genotyped SNPs.

In contrast, tanning (wrist minus inner arm pigmentation) did not appear to be heritable, especially when compared to heritability estimates obtained for height and innate arm skin pigmentation (Table 5.3) (the large difference in estimates between independent sets of parent/offspring pairs is due to several outlying pairs). Variability in tanning among ≠Khomani Bushmen individuals is likely to have a larger contribution from the environment and sun exposure than from genetic factors. Comparing the heritabilities among height, arm pigmentation, and tanning verifies that despite large standard errors, the point estimates of heritability that we obtain can be informative about the traits’ genetic bases.

Much of the heritability of innate pigmentation appears to be explained by individuals’ ancestries, as expected based on worldwide distributions of pigmentation phenotypes [199, 222]. As hypothesized, Bantu ancestry is significantly associated with darker pigmentation, and European ancestry with lighter pigmentation (Table 5.4 and Figure 5.9). While San ancestry alone is not a significant predictor of pigmentation, it is significant when included in a regression along with either European or Bantu ancestry, suggesting that there may exist KhoeSan-specific loci that contribute to variation in pigmentation. Given the strong associations of innate pigmentation with ancestry, genetic loci affecting pigmentation appear to be strongly differentiated by ethnicity [168]. In contrast to innate pigmentation, the amount of tanning does not appear to be affected by ancestry, agreeing with its low estimated heritability.

Because of the identified associations with ancestry, we included ancestry proportions as covariates in a genome-wide association study of innate arm pigmentation. This analysis was complicated, however, by the non-independence of the Bantu, European, and San ancestry proportions. Because of this collinearity, only two out of the three ancestries were included as covariates at a time; however, this may not account for all genetic structure (see Figure 5.11B). We are unable to find any loci significantly associated with pigmentation.

Our inability to identify loci associated with pigmentation in a GWAS is surprising, especially given the strong associations of pigmentation with ethnicity and the high heritability estimates of pigmentation estimated using the GCTA method. While it is possible that our small sample size may reduce our ability to identify novel associations, future studies using alternative methods may be more effective. First, instead of using ancestry proportions as covariates, accounting for ancestry using principal components calculated from genetic data may be more appropriate [10, 208]. Since principal components are by nature independent, they would allow for all underlying effects of ancestry to be incorporated. The strong effect of ancestry on pigmentation also suggests that admixture mapping will be an attractive tool for identifying loci associated with pigmentation in this population in the future [246, 249, 304]. With such an approach, KhoeSan-specific, European-specific, or Bantu-specific loci contributing to variation in pigmentation could be discovered. Given that genetic and phenotypic analyses of the KhoeSan have never before been performed, KhoeSan-specific pigmentation loci may reveal novel genes associated with pigmentation.

The ancestral nature of the KhoeSan make them a particularly attractive population for further understanding human pigmentation. Not only would the discovery of novel genes provide insight into pigmentation’s
genetic basis, but identifying the ages of alleles associated with lighter or darker pigmentation could reveal the evolutionary history of human pigmentation across worldwide populations [25, 168].

5.4.4 Conclusions

Despite our small sample size, our analysis of 103 #Khomani Bushmen reveals high heritability of both skin pigmentation and height, strong associations between ancestry and innate pigmentation, and several promising loci of large effect associated with height. We recognize that our power would be improved with a larger set of unrelated individuals as well as a denser set of genotypes, obtained potentially through imputation [166]. Nonetheless, this study emphasizes that analyses of phenotypically and genetically diverse individuals from endogamous populations have the potential to provide novel insights into the genetic basis and evolutionary history of complex traits – even those as often-studied as height and skin pigmentation.

5.5 Acknowledgments

We would like to thank the #Khomani San community, the South African San Institute (SASI), and our translators from the #Khomani San community for their assistance and support in carrying out this research. We would also like to thank Sandra Beleza, Oana Carja, and Karla Mendoza for helpful comments and discussions. We would like to acknowledge the Morrison Institute for Population Studies at Stanford University and a Stanford University Department of Biology SCORE grant for partially funding this research. J.M.G is supported by the National Science Foundation (NSF) Graduate Research Fellowship, grant number DGE-0645962. B.M.H. is supported by NIH grant R01HG003229. This research was supported in part by NSF grant GM28016 to M.W.F.
5.6 Figures and Tables

Figure 5.1: **Inferred pedigrees for a subset of related sampled individuals.** Relationships are inferred from ethnographic interviews and IBD estimation (see Methods). Circles are females, squares are males. Numbers below the circles or squares indicate the individual’s anonymized code; codes with an “X” denote individuals who were not sampled. Shading indicates the genotyping platform, and a slash indicates that the individual was not phenotyped (see legend). Note: some individuals were genotyped but not phenotyped; other individuals were phenotyped but not genotyped.
Figure 5.2: Individual ancestry proportions for the indicated number of ancestral populations inferred for phenotyped individuals using \textit{ADMIXTURE}. Each vertical bar is an individual, whose ancestry is indicated by its colors. Individuals are sorted by population, listed below the plot. $K$, the number of ancestral populations, ranges from 4 to 7. Inferences were made for six runs of unrelated individuals, and then aggregated into one plot (see Methods). At $K = 7$, one of the six runs did not infer the same ancestral clusters as the others; as a result, there are fewer individuals represented in the bottom panel (see Results).
Figure 5.3: Cross-validation errors and log likelihoods for models estimated using *ADMIXTURE* for various numbers of ancestral populations. Each colored line indicates an independent run of unrelated ≠Khomani Bushmen individuals, as described in the Methods (Section 5.2.3). X-axes: $K$, the number of ancestral populations. Top panel y-axis: cross-validation error as inferred by *ADMIXTURE* [5]. Bottom panel y-axis: log likelihood of model.
CHAPTER 5. HEIGHT AND PIGMENTATION IN THE KHOMANI BUSHMEN

Figure 5.4: Ancestry proportions versus age of phenotyped Khomani Bushmen individuals. Admixture proportions are inferred by ADMIXTURE, as in Figure 5.2. Regressions are simple linear regressions. Estimated coefficients and associated p-values are indicated in the corners of each plot.
Figure 5.5: Violin plots of height and skin pigmentation measurements for genotyped individuals. (A) Height (cm), (B) average arm pigmentation (M), and (C) tanning defined as wrist pigmentation minus inner arm pigmentation (Wrist M - Arm M) for all individuals, males, and females. Boxplots are drawn within the blue shaded regions, and the estimated kernel densities are drawn along each side.
Figure 5.6: Results from a linear mixed model regression of height on sex and age. Lines indicate the partial regression coefficients of height on sex (left panel), and height on age (right panel). \( \hat{\beta}_{\text{male}} \) and \( \hat{\beta}_{\text{age}} \) coefficients, with their p-values, are located in the upper left and lower left corners of the respective plots. Points in red are males, black are females. Mixed model equation is in Equation 5.2.
Figure 5.7: Results from a linear mixed model regression of tanning (wrist minus inner arm pigmentation) on sex and age. Lines indicate the partial regression coefficients of tanning on sex (left panel), and tanning on age (right panel). \( \hat{\beta}_{male} \) and \( \hat{\beta}_{age} \) coefficients, with their p-values, are located in the upper left and lower left corners of the respective plots. Points in red are males, black are females. Mixed model equation is in Equation 5.2.
Figure 5.8: Results from a linear mixed model regression of height on sex, age, and proportion European ancestry. Lines indicate the partial regression coefficients of height on European ancestry (top left panel), sex (top right panel), and age (bottom left panel), using the mixed model equation of Equation 5.2. \( \hat{\beta}_{\text{European}}, \hat{\beta}_{\text{male}}, \) and \( \hat{\beta}_{\text{age}} \) coefficients, with their p-values, are located in the corners of the respective plots. Points in red are males, black are females.
5.6. FIGURES AND TABLES

Figure 5.9: Results from linear mixed model regressions of inner arm skin pigmentation on proportion ancestry, for various ancestries. Each plot is a separate regression model for each ancestry, using the mixed model equation of Equation 5.2. Lines indicate the regression coefficients of the model, $\hat{\beta}_{Eur}$, $\hat{\beta}_{Bantu}$, $\hat{\beta}_{San}$, and $\hat{\beta}_{EastAfr}$, with their p-values, are located in the corners of the respective plots.
Figure 5.10: Genetic relationships for pairs of individuals inferred from EMMAX using the Balding-Nichols method. Higher values in the matrix correspond to greater relatedness between individuals. Individuals in the matrix were clustered to facilitate interpretation. Anonymized codes for each individual are indicated on the bottom and right sides of the matrix.
Figure 5.11: **QQ plots of $-\log_{10}$ SNP p-values for genome-wide association studies of height and inner arm pigmentation using EMMAX.** Y-axes indicate observed $-\log_{10}(p$-value), and x-axes indicate expected $-\log_{10}(p$-value) based on the uniform distribution. (A) Height analysis, where sex and age are included as covariates. (B) Inner arm pigmentation analysis, where the proportion European and the proportion San ancestry are included as covariates.
Figure 5.12: Manhattan plots of $-\log_{10}$ SNP p-values for height association using EMMAX. Y-axis indicates observed $-\log_{10}$(p-value). Covariates included are sex and age. (A) Whole genome. SNPs along the x-axis are ordered by chromosome and position, and shaded by chromosome. (B) 1000kb region surrounding rs9315851 (in yellow). Y-axis also indicates recombination rate (cM/Mb) estimated from African 1000 Genomes data. LD ($r^2$) with rs9315851 is indicated for all other genotyped SNPs (see legend). Locations of nearby genes are listed below the plot. (C) 1000kb region surrounding rs2967055; see legend for part (B).
Figure 5.13: Effect sizes by genotype and allele frequencies in worldwide HGDP populations for top height SNPs. A), B), C): Boxplots of effect sizes for the indicated SNPs: rs9315851, rs2967055, and rs2531691, respectively. Y-axes indicate the residuals of height after accounting for sex and age using a linear mixed model (see Methods); x-axis indicates the genotype for the plotted individuals. D), E), F): Map of allele frequencies indicated by pie charts in worldwide HGDP populations for rs9315851, rs2967055, and rs2531691, respectively (from http://hgdp.uchicago.edu).
Table 5.1: Pearson correlations of self-reported measures of ancestry with ancestries inferred by AD-MIXTURE for each individual.

<table>
<thead>
<tr>
<th>Self-Report Measure(^a)</th>
<th>Correlation(^b)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>European*</td>
<td>0.545</td>
<td>2.6x10(^{-9})</td>
</tr>
<tr>
<td>European/Griqua*</td>
<td>0.514</td>
<td>2.7x10(^{-8})</td>
</tr>
<tr>
<td>San†</td>
<td>0.336</td>
<td>0.0005</td>
</tr>
<tr>
<td>San/Nama†</td>
<td>0.387</td>
<td>5.4x10(^{-5})</td>
</tr>
<tr>
<td>Bantu**</td>
<td>0.231</td>
<td>0.019</td>
</tr>
</tbody>
</table>

\(^a\)Self-reported measure is defined in the Methods, Equation 5.1.

\(^b\)Correlations are performed against genetic ancestries inferred by ADMIXTURE (Figure 5.2); the genetic ancestry against which the measure is correlated is given by the symbol: * – European, † – San, and ** – Bantu.
Table 5.2: Means (and standard deviations) of measured phenotypes for all individuals and the subset of genotyped individuals.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Overall</th>
<th>Females</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>All Individuals (n = 103)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height (cm) (n = 102)</td>
<td>155.23 (7.95)</td>
<td>151.78 (5.18)</td>
<td>163.90 (7.08)</td>
</tr>
<tr>
<td>Inner Arm Pigmentation (M)</td>
<td>55.01 (10.59)</td>
<td>55.41 (11.24)</td>
<td>54.00 (8.80)</td>
</tr>
<tr>
<td>Wrist Pigmentation (M)</td>
<td>70.96 (12.37)</td>
<td>69.12 (12.04)</td>
<td>75.64 (12.14)</td>
</tr>
<tr>
<td>Wrist - Inner Arm Pigmentation (M)</td>
<td>15.94 (9.60)</td>
<td>13.71 (9.06)</td>
<td>21.64 (8.67)</td>
</tr>
<tr>
<td><em>Genotyped Individuals (n = 86)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height (cm) (n = 85)</td>
<td>154.94 (8.17)</td>
<td>151.11 (5.02)</td>
<td>163.64 (7.24)</td>
</tr>
<tr>
<td>Inner Arm Pigmentation (M)</td>
<td>54.35 (10.54)</td>
<td>54.66 (10.54)</td>
<td>53.65 (8.72)</td>
</tr>
<tr>
<td>Wrist Pigmentation (M)</td>
<td>70.71 (12.40)</td>
<td>68.99 (12.46)</td>
<td>74.67 (11.54)</td>
</tr>
<tr>
<td>Wrist - Inner Arm Pigmentation (M)</td>
<td>16.35 (9.29)</td>
<td>14.33 (8.90)</td>
<td>21.02 (8.62)</td>
</tr>
</tbody>
</table>

* Sample size (n) for each set of individuals applies to each phenotype unless otherwise indicated.
Table 5.3: Estimates of heritability ($h^2$) of height, inner arm skin pigmentation, and tanning (wrist minus inner arm pigmentation) using various methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>$h^2_{\text{height}}$ (S.E.)</th>
<th>$h^2_{\text{pigm}}$ (S.E.)</th>
<th>$h^2_{\text{tanning}}$ (S.E.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent/Offspring</td>
<td>0.608 (0.454)</td>
<td>0.808 (0.5426)</td>
<td>-0.025 (0.6778)</td>
</tr>
<tr>
<td>(18 non-independent pairs)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/Offspring</td>
<td>0.852 (0.843)</td>
<td>0.528 (0.685)</td>
<td>0.110 (0.7712)</td>
</tr>
<tr>
<td>(12 independent pairs – set 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/Offspring</td>
<td>0.895 (0.489)</td>
<td>0.4085 (0.5704)</td>
<td>-0.415 (0.546)</td>
</tr>
<tr>
<td>(12 independent pairs – set 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed Model</td>
<td>0.999</td>
<td>0.999</td>
<td>0.455</td>
</tr>
<tr>
<td>(kinship matrix from pedigree)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCTA [310]</td>
<td>0.588 (0.379)</td>
<td>0.999 (0.173)</td>
<td>0.385 (0.272)</td>
</tr>
</tbody>
</table>

- See Methods (Section 5.2.4) for details of the indicated methods.
- S.E. indicates the standard error of the estimate; no standard errors could be calculated for the mixed model method.
- For height and tanning (wrist minus inner arm pigmentation), parent/offspring heritability estimates were calculated using the residuals from a regression of height on sex and age (see Methods). Non-adjusted values were used for inner arm pigmentation.
- For height and tanning, both the mixed model and GCTA methods account for sex and age as covariates. No covariates were included for analyses of inner arm pigmentation.
- Only genotyped individuals were included for the GCTA analysis; all other estimates include all individuals, even those who were not genotyped.
### Table 5.4: Coefficients for regressions of height, pigmentation, and tanning on ancestries.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Model</th>
<th>European</th>
<th>Bantu</th>
<th>San</th>
<th>East African</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height(^a)</td>
<td>European Only</td>
<td>-0.365</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bantu Only</td>
<td>-1.053</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>San Only</td>
<td>-0.211</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>East African Only</td>
<td>8.369</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inner Arm Pigmentation(^c)</td>
<td>European Only</td>
<td>-24.98***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bantu Only</td>
<td>33.10***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>San Only</td>
<td>2.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>East African Only</td>
<td>0.320</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>European and Bantu</td>
<td>-21.01**</td>
<td>28.54**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>European and San</td>
<td>-51.93***</td>
<td>-27.70***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bantu and San</td>
<td>45.11***</td>
<td>15.93*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>European, Bantu, and San</td>
<td>-43.70</td>
<td>8.33</td>
<td>-20.43</td>
<td></td>
</tr>
<tr>
<td>Tanning (Wrist - Arm)(^b)</td>
<td>European Only</td>
<td>3.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bantu Only</td>
<td>2.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>San Only</td>
<td>-3.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>East African Only</td>
<td>8.57</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) The “Model” column indicates the ancestries that are included in the model. P-values from the regression are indicated after each $\hat{\beta}_{\text{ancestry}}$ coefficient as follows: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$.

\(^b\) For height and tanning (wrist minus inner arm pigmentation), coefficients are calculated from a model also including sex and age (i.e., $\hat{\beta}_{\text{ancestry}}$ values are partial regression coefficients); $\hat{\beta}_{\text{sex}}$ and $\hat{\beta}_{\text{age}}$ are not shown.

\(^c\) For pigmentation, for models with more than one ancestry, $\hat{\beta}_{\text{ancestry}}$ coefficients are the partial regression coefficients with their associated p-values.
### Table 5.5: Top SNPs associated with height from a genome-wide association study using EMMAX.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Location</th>
<th>Position</th>
<th>Effect (Other)*</th>
<th>Frequency*</th>
<th>$\beta_{SNP}$ (S.E.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs9315851</td>
<td>chr 13q14.11</td>
<td>42203574</td>
<td>C* (T)</td>
<td>0.788</td>
<td>5.3686 (1.0752)</td>
<td>$3.066 \times 10^{-7}$</td>
</tr>
<tr>
<td>rs2967055</td>
<td>chr 5p13</td>
<td>36472990</td>
<td>G* (A)</td>
<td>0.878</td>
<td>5.9926 (1.26013)</td>
<td>$1.462 \times 10^{-6}$</td>
</tr>
<tr>
<td>rs2531691</td>
<td>chr 10q25.3</td>
<td>118639142</td>
<td>T* (C)</td>
<td>0.965</td>
<td>10.8652 (2.5348)</td>
<td>$6.093 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

*a* Effect allele indicates the allele with the associated $\beta_{SNP}$ effect. * denotes the ancestral allele as indicated by dbSNP.

*b* Frequency of the effect allele in the ≠Khomani Bushmen.
A

Supplemental Materials: On the Classification of Genetic Interactions


A.1 Supplemental Methods

A.1.1 Derivation of MLEs for Multiplicative and Minimum Epistasis

For multiplicative epistasis, the Maximum Likelihood (ML) estimators for the epistasis coefficient and fitness variance of double mutants are

$$\hat{\epsilon}_{ij}^{(p)} = \frac{w_i w_j w_{ij}}{w_i^2 w_j^2} - 1,$$

$$\hat{\sigma}^2_{ij}^{(p)} = \frac{1}{N} \sum_{n=1}^{N} (w_{ij}^n - w_i^n w_j^n(1 + \hat{\epsilon}_{ij}^{(p)}))^2,$$  \hspace{1cm} (A.1)

where the right superscript $(p)$ indicates that the estimator is for multiplicative epistasis.

For minimum epistasis, the ML estimators for the epistasis coefficient and fitness variance of double mutants are

$$\hat{\epsilon}_{ij}^{(m)} = \frac{1}{N} \sum_{n=1}^{N} (w_{ij}^n - \min(w_i^n, w_j^n)),$$

$$\hat{\sigma}^2_{ij}^{(m)} = \frac{1}{N} \sum_{n=1}^{N} (w_{ij}^n - \min(w_i^n, w_j^n) - \hat{\epsilon}_{ij}^{(m)})^2,$$  \hspace{1cm} (A.2)

where the right superscript $(m)$ indicates that the estimator is for minimum epistasis. Formulas for the additive MLEs are in Chapter 2.

A.1.2 Simulation of Double Mutant Fitness Values

When simulating double mutant fitness values, the support is $[0, 1.2]$, allowing for both synergistic and antagonistic epistasis. If $g_i$ and $g_j$ act multiplicatively, that is, there is no multiplicative epistasis, $w_{ij} \sim N(w_i w_j, \sigma_{ij}^2)[0,1.2]$, and if $g_i$ and $g_j$ act according to the minimum definition of independence, $w_{ij} \sim N(\min(w_i, w_j), \sigma_{ij}^2)[0,1.2]$. (Distribution for under additive independence is shown in Chapter 2). For multiplicative epistasis, $w_{ij} \sim N(w_i w_j(1 + \epsilon_{ij}), \sigma_{ij}^2)[0,1.2]$, and for minimum epistasis, $w_{ij} \sim N(\min(w_i, w_j) + \epsilon_{ij}, \sigma_{ij}^2)[0,1.2]$. (Distribution for additive epistasis is given in Chapter 2).

A.1.3 Details of Analysis of Experimental Data

The first dataset (St Onge et al. [255]) included 26 non-essential genes known to confer resistance to the DNA-damaging agent methyl methanesulfonate (MMS), and contained double-deletion strains for all pairs
that were possible to construct (a total of 323 double mutant strains). The authors measured the fitnesses of single and double mutants in media both with and without MMS. For both media types, while single mutants have approximately 10 replicate fitness measurements, double mutants have only two replicates.

We modify our procedure to analyze this dataset. For each gene in each pair, we randomly partition its single-mutant fitness values into two parts and use the mean of each part as the two single-mutant fitness replicates. Because of this procedure’s variability, we repeat this procedure 1000 times, each time determining the subtype of epistasis for each pair. We then select gene pairs for which one of the three epistatic subtypes was chosen in at least 900 of the replicates, and select as the final epistatic subtype for each pair that which was chosen in the majority of the replicates. Changing the cutoff value to either 950 or 800 replicates to create this “high-confidence” epistatic set does not greatly affect our results. We average values of the LRT statistic ($\lambda$) and $\epsilon$ over replicates where the given subtype was selected to obtain the final values.

When examining pairs of genes which share “specific” functional links (defined in Chapter 2), we obtain the appropriate Gene Ontology (GO) terms in all three GO categories [270] for the genes of interest with AmiGO (for St Onge et al. [255] data, GO database release 2009-09-08; for Jasnos and Korona [131] data, GO database release 2009-09-22). In order to assess significance of both specific functional links and shared GO Slim terms among the inferred epistatic pairs, we implement both Fisher’s Exact Test and a permutation procedure with 5000 iterations. For each permutation, we randomly partition the $n$ gene pairs originally tested for epistasis into two groups, $x$ pairs and $n - x$ pairs, where $x$ is the number of inferred epistatic pairs. We then count the number out of $x$ pairs that share either a functional link or GO Slim term link. We set the $p$-value as the proportion of iterations where the number of links exceeds our observation.

To examine only experimentally-verified interactions in the BIOGRID database [257], we ignored interactions inferred quantitatively from phenotype data and whose “evidence codes” were one the following: “Dosage Growth Defect,” “Dosage Lethality,” “Dosage Rescue,” “Phenotypic Enhancement,” “Phenotypic Suppression,” “Synthetic Defect,” “Synthetic Haploinsufficiency,” “Synthetic Lethality,” and “Synthetic Rescue.” We focused only on experimentally-validated (physical) interactions, identified by co-localization, co-purification, etc.

In addition, we note that in our formulation (see Chapter 2), the multiplicative epistatic model is undefined when double mutants have a fitness of 0 ($\hat{\epsilon}_{ij} = -1$ and $\hat{\sigma}_{ij} = 0$). In our current implementation of the method, we do not attempt to further sub-classify these interactions as either additive or minimum epistasis. We feel that because synthetic lethals have a double mutant fitness of 0, any further sub-classification is unnecessary.

### A.2 Supplemental Results

#### A.2.1 Additional Analyses of Simulated Data

We assess the accuracy of inferring the correct epistatic subtype when decreasing the sample size to 10 (Table A.2; see also Table 2.1, Chapter 2). The ability to distinguish null models from epistatic models is high, with accuracy of approximately $0.85 \sim 0.9$ for various values of $\epsilon$. While the accuracy of inferring
minimum epistasis decreases slightly (to about 80%), inference of multiplicative epistasis is heavily affected by a reduction of sample size (accuracy of $0.4 \sim 0.6$). The inference of additive epistasis is robust to a decrease in replicates when $\epsilon$ is negative; however, when $\epsilon$ is positive, accuracy decreases substantially ($0.5 \sim 0.6$). See also Table A.3.

We also relaxed our restriction to deleterious mutations to assess the effect of a more complex fitness landscape on the estimation of epistasis. As illustrated in Table A.1, our method is robust to beneficial mutations, which do not decrease the accuracy of inferring the correct epistatic subtype. If beneficial mutations are of specific interest, in the future it may be worthwhile to consider the Log model as an epistatic subtype [165, 236], as its exclusion may only be warranted for deleterious mutations. We also note that when at least one of the single mutant fitness values in a pair is 1, the predictions under each null model are equivalent and we have no power to detect the correct epistatic subtype (not shown).

Another practical issue is that the accuracy of constructing an epistatic network (indicating all pairwise interactions) decreases substantially as the number of genes increases. Our method is effective in detecting the existence and type of epistasis for one gene pair (accuracy is $\sim 0.95$ for minimum epistasis; Table 2.1, Chapter 2). However, for more than 10 genes we must select the epistatic subtype for more than 45 pairs, decreasing accuracy to at most $0.95^{45} \approx 0.1$. This issue arises in all large-scale studies of epistasis, and is not unique to our method. To improve performance, we advocate controlling the FDR.

### A.2.2 Additional Analyses of Experimental Data

To select one among the three possible multiple-hypothesis testing procedures for the St Onge et al. [255] dataset, we examine the number of inferred epistatic pairs which share “specific” functional links (described in Chapter 2). 36 of the 323 pairs examined in the dataset share a specific functional link (including five synthetic lethals in the dataset). When using the Bonferroni correction, out of the 25 gene pairs for which we identify epistasis, only 1 gene pair has a functional link. (See Discussion in Chapter 2 for implications of this result.) Implementing the FDR procedure [26], we find 188 gene pairs with epistasis (not including synthetic lethal double mutants), of which 20 have a functional link. St Onge et al. [255] discover 133 gene pairs with significant epistasis; as they find this to be a reasonable number, 188 pairs also appears reasonable. Finally, with the pFDR procedure [258], we find 222 gene pairs with epistasis; this procedure rejects neutrality for all pairs for which an epistatic model was originally selected (for the pFDR, the value above which p-values are estimated to be uniformly distributed is 0.1). In order to assess significance of these values, we perform Fisher’s exact test and a permutation procedure (described in Chapter 2); however, the number of functional links observed among epistatic pairs is not significant for any of the multiple-testing measures.

We show several examples of epistatic networks found with our method which considers all three epistatic subtypes and their corresponding null models (with the FDR procedure) when examining the St Onge et al. [255] dataset. In Figure A.6(b), we show the epistatic interactions inferred between all pairs of genes with functional links, as well as the type of epistasis for each connection. We find a mixture of positive and negative epistasis, and slightly different results than the original authors. In Figure A.6(c), we show, following
the original authors, the homologous recombination pathway including the genes RAD51, RAD52, RAD54, RAD55 and RAD57. While St Onge et al. [255] find 10 epistatic pairs among these genes, we find only 7 epistatic pairs, all of which are classified as minimum epistasis. In Figure A.6(d), we present epistasis found for genes of the SHU complex (SHU1, SHU2, CSM2 and PSY3), as do St Onge et al. [255]. We do not find PSY3 to interact with any of the genes in this pathway, and find a mixture of multiplicative and minimum epistasis. In constructing the “high-confidence” set of epistatic pairs through sample replications (Chapter 2), we are likely more conservative than St Onge et al. [255] in identifying epistasis. In addition, these authors do not implement any multiple-hypothesis testing procedures. This probably slightly increases the power of their method over ours, yet unfortunately at the expense of detecting more false positives of gene pairs with epistasis.

We also searched for over-representation of GO Slim terms in genes involved in each additive, multiplicative, and minimum subtypes, as well as positive and negative epistasis; however, we found no significant results (The Gene Ontology Consortium [270], http://www.yeastgenome.org).

We next examine the Jasnos and Korona [131] dataset and look for “specific” functional links that exist among epistatic pairs inferred with each of the three multiple hypothesis testing procedures. We examine Gene Ontology (GO) [270] terms with fewer than 200 genes associated with them (described in Chapter 2) and find 25 links among all 636 tested pairs. Using the Bonferroni procedure, 56 epistatic pairs are significant, of which 5 have a functional link. The FDR procedure results in 352 significant epistatic pairs, of which 19 have a functional link. For the pFDR, we detect 471 epistatic pairs, of which 21 have a functional link. When assessing significance of the number of functional links among epistatic pairs by permutation (see Chapter 2), we find that only the FDR results in a significant number of functional links (p-value = 0.024).

The epistatic networks discovered with our method and the FDR procedure in the Jasnos and Korona [131] dataset are small and isolated, likely due to the examination of only a subset of the possible pairwise genotypes in the study. We annotate functions of the genes in the networks according to the yeast Gene Ontology (GO) Slim Biological Process terms (following Jasnos and Korona [131], see Chapter 2). Four examples of constructed epistatic networks denoting pairwise epistatic interactions are shown in Figure A.7. Figure A.7(b) shows minimum epistatic interactions between three genes involved in the biological processes of mitochondrion organization and translation, all of which are functionally linked (as defined above) with the GO term mitochondrial protein formation and synthesis (GO:0032543); MRPL37 and MSK1 are also linked through the mitochondria (GO:005739). These functional links suggest that the epistatic interactions we detect may in fact be meaningful.

In Figure A.7(c), we show that genes IFM1 and SNT309 have alleviating positive epistasis; we also find a positive minimum epistatic interaction between SNT309 and BUD28. We find a larger linked group of six genes, among which there are two pairs of genes that have epistasis and share a functional link (Figure A.7(d)). In this network, we find a mixture of both minimum and additive epistasis, and both positive and negative epistasis. The functional links and shared GO Slim annotations in this network suggest that this novel epistasis is likely biologically significant.
A.2. SUPPLEMENTAL RESULTS

Finally, Figure A.7(e) illustrates an example of a larger group of twelve genes connected by three types of epistasis. Of the twelve genes, four (RDH54, HCM1, CIN8 and SGS1) have roles in the cell cycle. It is possible that other genes in this network, such as SWA2, may also be involved in this process, although they have not yet been studied. A specific functional link also exists between genes SGS1 and CIN8, which are both involved in mitotic chromosome segregation. We believe that these, along with the other interactions identified, are true epistatic interactions.

Although we search for significant enrichment of particular GO terms amongst genes involved in additive, multiplicative, and minimum epistasis, we fail to find any, indicating that the subtypes of epistasis are likely distributed over many functional classes of genes. This lack of significance is also likely due to the small number of double mutants examined.

As discussed in Chapter 2, we assess the normality of the replicate fitness data for both studied datasets. A deviation from normality may suggest that the $\chi^2$ approximation for the distribution of the likelihood ratio test statistic under the null hypothesis is inappropriate in some cases. Double-mutant fitness values of the St Onge et al. [255] dataset certainly deviate from normality, as only two replicates are measured (Figure A.8). In Figure A.8 we also show QQ plots for several of the genes of the SHU complex, CSM2, CSM3, and PSY3 (presented in St Onge et al. [255]). Although functional links exist, we do not find PSY3 to interact with either gene; this could be due to the deviation of the CSM3 single-mutant fitness values from normality. For the Jasnos and Korona [131] dataset, we examine the QQ plots of the genes NCS2 and SGO1 (Figure A.9); although a previous interaction has been identified for this pair, we do not detect epistasis. The deviation from normality of the single-mutant fitness values of SGO1 and the small number of double-mutant fitness replicates could explain our inability to detect this interaction. The distributions of fitness values for several other genes show similar deviations from normality (not shown).

A.2.3 Consideration of Alternative Model Selection Procedures

To explore whether the accuracy of inferring the epistatic subtype for pairs of genes could be improved, we also tried using two alternatives to the BIC procedure. The first method is similar to cross-validation. The sample was split randomly into a modeling set and a testing set. We fit each of the six models to the modeling data set and computed the mean squared error per model for the testing set. After repeating this procedure 1000 times, we selected the model with the smallest accumulated mean squared error. Overall, this method failed to outperform the BIC (data not shown).

The difficulty with testing for epistasis is that the null hypothesis can be any one of the three null models of independence, and is thus not well-specified. We explored using the Expectation-Maximization (EM) algorithm [63] for this problem in order perform a different likelihood ratio test. Under the null model of no epistasis, double mutants can be distributed as a mixture of the three null models; under the epistatic model, the distribution is a mixture model of the three epistatic subtypes. Using the EM algorithm, we maximized the likelihood under each model, and performed a likelihood ratio test. If the null hypothesis of independence was rejected, we selected the epistatic measure for the gene pair as the mixture component that had the highest
estimated proportion. However, we found that this method resulted in poorer performance (data not shown). Thus, it appears that the more simplistic BIC method currently implemented is the most appropriate method for this estimation problem.
A.3 Supplemental Figures and Tables

Figure A.1: Power of our method to uncover the true (simulated) epistatic model assuming $\epsilon_{ij} = -0.3$, $\sigma_{ij} = 0.1$, and $\mu_i = \mu_j = 0.8$. (A) The area under each curve and less than zero represents the percentage of simulations with the additive epistatic model favored over the other models shown in the legend. The red, blue, yellow, green, and purple lines represent the comparison of the BIC between the additive epistatic model and the multiplicative epistatic model, minimum epistatic model, additive null model, multiplicative null model, and minimum null model, respectively. (B) and (C) are shown in a manner similar to (A) for the multiplicative and minimum epistatic models, respectively.
Figure A.2: Power of our method to uncover the true (simulated) epistatic model assuming $\epsilon_{ij} = 0.3$, $\sigma_{ij} = 0.1$, and $\mu_i = \mu_j = 0.8$. Legends are as defined in Figure A.1.
Figure A.3: The distribution of the mean squared error (MSE) of the MLE of the epistatic parameter under the additive, multiplicative, and minimum epistatic models as well as the corresponding null model (with mean fitnesses of single mutants $\mu_i = 0.9$ and $\mu_j = 0.6$). a)1-3 show the distribution of MSE assuming epistatic coefficient $\epsilon_{ij} = -0.3$; b)1-3 assume $\epsilon_{ij} = -0.1$; c)1-3 assume $\epsilon_{ij} = 0.1$; d)1-3 assume $\epsilon_{ij} = 0.3$. The red, blue, and yellow lines represents the distribution of the MSE of $\hat{\epsilon}_{MLE}$ under the additive, multiplicative, and minimum epistatic models, respectively; the green line represents the MSE of $\hat{\epsilon}_{MLE}$ under the null model (of the simulated definition).
Figure A.4: Venn diagram comparing the overlap of epistatic pairs found examining the St Onge et al. [255] dataset using the authors’ original method and our BIC method (including all three epistatic subtypes and their corresponding null models with the FDR procedure). For the pairs identified by only one method, we indicate the number for which an interaction has already been experimentally identified (obtained using the BIOGRID database).
Figure A.5: Distribution of the epistasis values (ε) for significant epistatic pairs, determined using the FDR procedure with the BIC method (including all three epistatic subtypes and their corresponding null models) for all gene pairs of the St Onge et al. [255] dataset in the presence of MMS (bin size = 0.05). ε values for gene pairs with specific functional links as determined by GO terms (explained in text) are shown in red (mean of -0.060, not significantly different than 0), and for those without specific functional links are shown in black. A similar figure is shown in St Onge et al. [255].
Figure A.6: Several epistatic networks constructed based on the St Onge et al. [255] dataset (using the FDR procedure with the BIC method including all three epistatic subtypes and their corresponding null models). Legend in (a) denotes the possible subtypes for each epistatic interaction. (b) depicts all epistatic interactions that were found between genes that also share “specific” functional links (defined in text); note that epistatic interactions were inferred between other genes in this network that do not have a specific functional link (not shown). (c) represents all inferred interactions in the RAD homologous recombination pathway. (d) represents all inferred interactions in the SHU pathway. “+” sign indicates positive epistasis, “−” sign indicates negative epistasis, and “I” indicates that the interaction has been experimentally identified.
A.3. **SUPPLEMENTAL FIGURES AND TABLES**

<table>
<thead>
<tr>
<th>Epistatic Subtypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>—— Additive</td>
</tr>
<tr>
<td>— — Multiplicative</td>
</tr>
<tr>
<td>— — Minimum</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GO Slim Biological Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome Organization</td>
</tr>
<tr>
<td>Cytoskeleton Organization</td>
</tr>
<tr>
<td>Mitochondrion Organization</td>
</tr>
<tr>
<td>Transcription</td>
</tr>
<tr>
<td>Translation</td>
</tr>
<tr>
<td>DNA Metabolic Processes</td>
</tr>
<tr>
<td>RNA Metabolic Processes</td>
</tr>
<tr>
<td>Cell Cycle</td>
</tr>
<tr>
<td>Cellular Component Morphogenesis</td>
</tr>
<tr>
<td>Cellular Homeostasis</td>
</tr>
<tr>
<td>Cytokinesis</td>
</tr>
<tr>
<td>Response to Stress</td>
</tr>
<tr>
<td>Unknown/Other</td>
</tr>
</tbody>
</table>

Figure A.7: **Several epistatic networks constructed based on the Jasnos and Korona [131] dataset (using the FDR procedure with the BIC method including all three epistatic subtypes and their corresponding null models).** Legend in (a) shows each type of epistatic interaction and each color which corresponds to one biological process as listed by the Gene Ontology (GO Slim terms). (b), (c), (d), and (e) are examples of inferred three-gene, three-gene, six-gene, and twelve-gene networks, respectively. “+” sign indicates positive epistasis, “−” sign indicates negative epistasis, “*” indicates that the gene pair has a specific functional link (described in text).
Figure A.8: Quantile-Quantile plot of the single-mutant fitness values for the genes CSM2, CSM3, and PSY3 (top row), and the double-mutant fitness values for CSM2/CSM3, CSM2/PSY3, and CSM3/PSY3 (bottom row) (St Onge et al. [255] dataset). We identify an interaction between CSM2 and CSM3, but not for the other two pairs, when analyzing this dataset.
A.3. SUPPLEMENTAL FIGURES AND TABLES

Figure A.9: Quantile-Quantile plot of the single-mutant fitness values for the genes \( NCS2 \) and \( SGO1 \) (top row), and the double-mutant fitness values of \( NCS2/SGO1 \) (bottom row) (Jasnos and Korona [131] dataset). We do not identify an interaction between these two genes.
Table A.1: Fractions of simulations that recover the true model among the six models when allowing both beneficial and deleterious mutations, i.e., the range for fitness of single mutants is \([0, 1.2]\) and the range for fitness of double mutants is \([0, 1.5]\).

<table>
<thead>
<tr>
<th>(\mu_i)</th>
<th>(\mu_j)</th>
<th>(P_N^{(a)})</th>
<th>(P_N^{(p)})</th>
<th>(P_N^{(m)})</th>
<th>(\epsilon_{ij})</th>
<th>(P_E^{(a)})</th>
<th>(P_E^{(p)})</th>
<th>(P_E^{(m)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.987</td>
<td>0.986</td>
<td>0.993</td>
<td>-0.3</td>
<td>0.996</td>
<td>0.903</td>
<td>0.929</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.987</td>
<td>0.986</td>
<td>0.993</td>
<td>-0.2</td>
<td>0.991</td>
<td>0.899</td>
<td>0.956</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.987</td>
<td>0.986</td>
<td>0.993</td>
<td>-0.1</td>
<td>0.964</td>
<td>0.869</td>
<td>0.982</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.987</td>
<td>0.986</td>
<td>0.993</td>
<td>0.1</td>
<td>0.742</td>
<td>0.719</td>
<td>0.997</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.987</td>
<td>0.986</td>
<td>0.993</td>
<td>0.2</td>
<td>0.581</td>
<td>0.602</td>
<td>0.998</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.987</td>
<td>0.986</td>
<td>0.993</td>
<td>0.3</td>
<td>0.740</td>
<td>0.613</td>
<td>0.998</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.989</td>
<td>0.987</td>
<td>0.994</td>
<td>-0.3</td>
<td>0.999</td>
<td>0.964</td>
<td>0.982</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.989</td>
<td>0.987</td>
<td>0.994</td>
<td>-0.2</td>
<td>0.997</td>
<td>0.948</td>
<td>0.971</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.989</td>
<td>0.987</td>
<td>0.994</td>
<td>-0.1</td>
<td>0.986</td>
<td>0.901</td>
<td>0.979</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.989</td>
<td>0.987</td>
<td>0.994</td>
<td>0.1</td>
<td>0.878</td>
<td>0.855</td>
<td>0.997</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.989</td>
<td>0.987</td>
<td>0.994</td>
<td>0.2</td>
<td>0.836</td>
<td>0.843</td>
<td>0.999</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.989</td>
<td>0.987</td>
<td>0.994</td>
<td>0.3</td>
<td>0.891</td>
<td>0.837</td>
<td>0.999</td>
</tr>
</tbody>
</table>

For all the simulations, the fitness standard deviation \(\sigma_i = \sigma_j = \sigma_{ij} = 0.05\). The rest of the parameters are defined in Table 2.1, Chapter 2. 50 replicate fitness values are simulated.
Table A.2: Fractions of simulations that recover the true model among the six models, with different values of the epistasis coefficient ($\epsilon_{ij}$) and mean fitnesses of single mutants ($\mu_i$ and $\mu_j$); number of fitness replicates is equal to 10.

<table>
<thead>
<tr>
<th>$\mu_i$</th>
<th>$\mu_j$</th>
<th>$P_N^{(a)}$</th>
<th>$P_N^{(p)}$</th>
<th>$P_N^{(m)}$</th>
<th>$\epsilon_{ij}$</th>
<th>$P_E^{(a)}$</th>
<th>$P_E^{(p)}$</th>
<th>$P_E^{(m)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.842</td>
<td>0.824</td>
<td>0.926</td>
<td>-0.3</td>
<td>0.836</td>
<td>0.574</td>
<td>0.705</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>-0.2</td>
<td>0.807</td>
<td>0.533</td>
<td>0.732</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>-0.1</td>
<td>0.758</td>
<td>0.396</td>
<td>0.762</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.1</td>
<td>0.59</td>
<td>0.423</td>
<td>0.871</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>0.472</td>
<td>0.335</td>
<td>0.875</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.3</td>
<td>0.491</td>
<td>0.506</td>
<td>0.898</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.856</td>
<td>0.821</td>
<td>0.903</td>
<td>-0.3</td>
<td>0.85</td>
<td>0.629</td>
<td>0.742</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>-0.2</td>
<td>0.833</td>
<td>0.589</td>
<td>0.753</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>-0.1</td>
<td>0.796</td>
<td>0.405</td>
<td>0.732</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.1</td>
<td>0.677</td>
<td>0.374</td>
<td>0.848</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.2</td>
<td>0.64</td>
<td>0.487</td>
<td>0.873</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>0.645</td>
<td>0.527</td>
<td>0.884</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For all simulations, the fitness standard deviation $\sigma_i = \sigma_j = \sigma_{ij} = 0.05$. The rest of the parameters are defined in Table 2.1, Chapter 2.
Table A.3: Fractions of simulations that recover the true model among the six models with different values of the epistasis coefficient ($\epsilon_{ij}$) and mean fitnesses of single mutants ($\mu_i$) and ($\mu_j$); number of replicates is equal to 2.

<table>
<thead>
<tr>
<th>$\mu_i$</th>
<th>$\mu_j$</th>
<th>$P_{N}^{(a)}$</th>
<th>$P_{N}^{(p)}$</th>
<th>$P_{N}^{(m)}$</th>
<th>$\epsilon_{ij}$</th>
<th>$P_{E}^{(a)}$</th>
<th>$P_{E}^{(p)}$</th>
<th>$P_{E}^{(m)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.261</td>
<td>0.195</td>
<td>0.338</td>
<td>-0.3</td>
<td>0.527</td>
<td>0.315</td>
<td>0.469</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.261</td>
<td>0.195</td>
<td>0.338</td>
<td>-0.2</td>
<td>0.519</td>
<td>0.285</td>
<td>0.511</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.261</td>
<td>0.195</td>
<td>0.338</td>
<td>-0.1</td>
<td>0.498</td>
<td>0.232</td>
<td>0.533</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.261</td>
<td>0.195</td>
<td>0.338</td>
<td>0.1</td>
<td>0.444</td>
<td>0.175</td>
<td>0.591</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.261</td>
<td>0.195</td>
<td>0.338</td>
<td>0.2</td>
<td>0.36</td>
<td>0.199</td>
<td>0.599</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.261</td>
<td>0.195</td>
<td>0.338</td>
<td>0.3</td>
<td>0.255</td>
<td>0.344</td>
<td>0.603</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.273</td>
<td>0.161</td>
<td>0.301</td>
<td>-0.3</td>
<td>0.514</td>
<td>0.341</td>
<td>0.455</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.273</td>
<td>0.161</td>
<td>0.301</td>
<td>-0.2</td>
<td>0.509</td>
<td>0.302</td>
<td>0.494</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.273</td>
<td>0.161</td>
<td>0.301</td>
<td>-0.1</td>
<td>0.512</td>
<td>0.231</td>
<td>0.505</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.273</td>
<td>0.161</td>
<td>0.301</td>
<td>0.1</td>
<td>0.492</td>
<td>0.173</td>
<td>0.567</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.273</td>
<td>0.161</td>
<td>0.301</td>
<td>0.2</td>
<td>0.446</td>
<td>0.202</td>
<td>0.557</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.273</td>
<td>0.161</td>
<td>0.301</td>
<td>0.3</td>
<td>0.394</td>
<td>0.274</td>
<td>0.555</td>
</tr>
</tbody>
</table>

For all the simulations, the fitness standard deviation $\sigma_i = \sigma_j = \sigma_{ij} = 0.05$. The rest of the parameters are defined in Table 2.1, Chapter 2.
Table A.4: **Summary of gene pairs with the indicated epistatic subtypes, inferred using the FDR procedure with the BIC method which considers all three epistatic subtypes and their corresponding null models (St Onge et al. [255] dataset, with fitnesses measured in the absence of MMS).**

<table>
<thead>
<tr>
<th>Epistatic Subtype</th>
<th>Study S (No MMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>171 (100%)</td>
</tr>
<tr>
<td></td>
<td>τ = -0.063***</td>
</tr>
<tr>
<td></td>
<td>ϵ_{q_{0.5}} = -0.057</td>
</tr>
<tr>
<td>Additive</td>
<td>4 (2.3%)</td>
</tr>
<tr>
<td></td>
<td>τ = 0.049*</td>
</tr>
<tr>
<td></td>
<td>ϵ_{q_{0.5}} = 0.074</td>
</tr>
<tr>
<td>Multiplicative</td>
<td>7 (4.1%)</td>
</tr>
<tr>
<td></td>
<td>τ = -0.109**</td>
</tr>
<tr>
<td></td>
<td>ϵ_{q_{0.5}} = -0.084</td>
</tr>
<tr>
<td>Minimum</td>
<td>155 (90.6%)</td>
</tr>
<tr>
<td></td>
<td>τ = -0.063***</td>
</tr>
<tr>
<td></td>
<td>ϵ_{q_{0.5}} = -0.057</td>
</tr>
</tbody>
</table>

Numbers are the counts of each type, and percentages are given of the total number of epistatic pairs. The mean (τ) and median (ϵ_{q_{0.5}}) of the epistatic parameter (ϵ) are given for each subtype, with “*” indicating that the mean of ϵ is significantly different from 0 (p-value ≤ 0.05; “**” implies p-value ≤ 0.01, “***” implies p-value ≤ 0.001). Note that 5 of the epistatic pairs are synthetic lethal (not shown in table).
<table>
<thead>
<tr>
<th>Subtypes Considered in BIC Procedure</th>
<th>14</th>
<th>12</th>
<th>18</th>
<th>14</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO Slim Terms (All) (307)</td>
<td>165 (0.1558)</td>
<td>114 (0.0010)*</td>
<td>149 (0.4151)</td>
<td>145 (0.0037)*</td>
<td>114 (0.0010)*</td>
</tr>
<tr>
<td>GO Slim Terms (Biological Process) (283)</td>
<td>158 (0.0058)*</td>
<td>114 (0.0010)*</td>
<td>145 (0.0037)*</td>
<td>20 (0.9682)</td>
<td></td>
</tr>
<tr>
<td>Functional Links (36)</td>
<td>171</td>
<td>130</td>
<td>165</td>
<td>130</td>
<td>171</td>
</tr>
<tr>
<td>Number Found (323)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study S (No MMS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Numbers in parentheses indicate p-values by Fisher's Exact Test, "*" indicates significance. Study S refers to the St Onge et al. [255] dataset.
Table A.6: Comparison of validation measures for each dataset for different variations of the FDR and BIC procedure, considering only a subset of epistatic subtypes with their corresponding null models: all epistatic subtypes (A, P, M), only the additive and multiplicative subtypes (A, P), only the additive (A), only the multiplicative (P), or only the minimum (M) subtype (see text for details).

<table>
<thead>
<tr>
<th>Subtypes Considered in BIC Procedure</th>
<th>A, P, M</th>
<th>A, P</th>
<th>A</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study J</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number Found (636)</td>
<td>352</td>
<td>273</td>
<td>263</td>
<td>231</td>
<td>329</td>
</tr>
<tr>
<td>Functional Links (25)</td>
<td>19 (0.0240)*</td>
<td>13 (0.2274)</td>
<td>11 (0.4700)</td>
<td>10 (0.4216)</td>
<td>15 (0.2466)</td>
</tr>
<tr>
<td>GO Slim Terms (Biological Process) (115)</td>
<td>69 (0.2720)</td>
<td>50 (0.6226)</td>
<td>55 (0.1328)</td>
<td>44 (0.4726)</td>
<td>68 (0.0)*</td>
</tr>
<tr>
<td>GO Slim Terms (All) (369)</td>
<td>224 (0.0220)*</td>
<td>172 (0.1180)</td>
<td>160 (0.4112)</td>
<td>146 (0.1172)</td>
<td>213 (0.007)*</td>
</tr>
<tr>
<td>Experimentally Identified (3)</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Study S (MMS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number Found (323)</td>
<td>193</td>
<td>192</td>
<td>247</td>
<td>171</td>
<td>243</td>
</tr>
<tr>
<td>Functional Links (36)</td>
<td>21 (0.473)</td>
<td>29 (0.0004)*</td>
<td>34 (0.0)*</td>
<td>29 (0.0)*</td>
<td>24 (0.8102)</td>
</tr>
<tr>
<td>GO Slim Terms (Biological Process) (283)</td>
<td>174 (0.0652)</td>
<td>174 (0.0368)*</td>
<td>223 (0.0108)*</td>
<td>153 (0.1822)</td>
<td>213 (0.5560)</td>
</tr>
<tr>
<td>GO Slim Terms (All) (307)</td>
<td>185 (0.2806)</td>
<td>182 (0.6940)</td>
<td>237 (0.1458)</td>
<td>162 (0.7006)</td>
<td>231 (0.5794)</td>
</tr>
<tr>
<td>Experimentally Identified (29)</td>
<td>14</td>
<td>22</td>
<td>24</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td><strong>Study S (No MMS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number Found (323)</td>
<td>171</td>
<td>120</td>
<td>156</td>
<td>99</td>
<td>203</td>
</tr>
<tr>
<td>Functional Links (36)</td>
<td>20 (0.2986)</td>
<td>20 (0.0052)*</td>
<td>25 (0.0018)*</td>
<td>16 (0.0258)*</td>
<td>18 (0.9198)</td>
</tr>
<tr>
<td>GO Slim Terms (Biological Process) (283)</td>
<td>158 (0.0052)*</td>
<td>114 (0.0008)*</td>
<td>145 (0.0036)*</td>
<td>95 (0.001)*</td>
<td>186 (0.004)*</td>
</tr>
<tr>
<td>GO Slim Terms (All) (307)</td>
<td>165 (0.1520)</td>
<td>115 (0.4182)</td>
<td>149 (0.4510)</td>
<td>95 (0.4264)</td>
<td>196 (0.088)</td>
</tr>
<tr>
<td>Experimentally Identified (29)</td>
<td>14</td>
<td>14</td>
<td>18</td>
<td>12</td>
<td>14</td>
</tr>
</tbody>
</table>

Numbers in parentheses indicate p-values by permutation, “*” indicates significance. Study J refers to the Jasnos and Korona [131] dataset, and Study S refers to the St Onge et al. [255] dataset. Numbers in parentheses indicate the total number of tested pairs and the total number of each type of link found in each complete dataset.
Supplemental Materials: Dispersal of *Mycobacterium tuberculosis* via the Canadian fur trade


**B.1 Supplemental Background Material**

**B.1.1 A Brief History of the Canadian Fur Trade, 1640-1870**

The Canadian fur trade involved the barter of European goods for animal furs. European goods were imported (and furs exported) from Montreal, in present day Quebec, and Hudson’s Bay. In this report, we focus on the Montreal route, which would have been the means of dispersal of *M. tuberculosis* (*M. tb*) from French Canadians to indigenous peoples of the Canadian interior (it is likely that British immigrants using the Hudson’s Bay route also transmitted *M. tb* to Aboriginal populations). We do not know whether Aboriginal populations involved in fur trading had been exposed previously to *M. tb*. There is evidence of TB among some indigenous Canadian populations prior to contact with European immigrants, but the extent and severity of the disease are not known [295].

Trade in furs necessitated a network of transportation routes over a vast geographic area that eventually encompassed what are now referred to as Ontario (ON), Manitoba (MB), Saskatchewan (SK), Alberta (AB), British Columbia (BC), and the Northwest Territories (NWT). The earliest means of transportation was the canoe, and movement of goods through lakes, rivers and portages (gaps in the navigable system of rivers and lakes) required canoemen (voyageurs), guides, translators, navigators, administrators and others. Employees of fur trading companies were male; voyageurs, with “few exceptions... were French” [120]. Fur traders stayed in the Canadian interior for variable lengths of time (discussed further in the Supplemental Results (Section B.3) and contact with indigenous populations commonly involved sexual relationships with Native women, some of which were formalized as marriages, and others that involved cohabitation and procreation outside of marriage [82]. Native women in these relationships played an important role in fur trade commerce (e.g. by processing furs), although this role was not acknowledged formally [82]. Children born of these alliances became the Métis. In a recent census, almost 400,000 Canadians identified themselves as Métis [41].

**B.1.2 Western Canadian Indigenous Populations in the Industrial Era, 1870-present**

While historians debate the degree to which traditional fur trade activities affected indigenous Canadian populations, it is clear that Euro-Canadian society had a profound impact on Native Canadians in the late
B.2  SUPPLEMENTAL METHODS

19th century and beyond [84]. Industrial scale extraction of resources such as buffalo, furs and fish, as well as agricultural development, resulted in destruction of traditional food sources for Native peoples and their displacement from the land [84]. Legal agreements between the Canadian government and Native leaders resulted in confinement of Native populations to specific tracts of land ("reserves"), where subsistence was difficult if not impossible in some cases [84, 218]. Attempts to acculturate Native Canadian populations included creation of residential educational institutions that housed large numbers of children in conditions highly conducive to the spread of tuberculosis [172]. All of these influences promote high rates of tuberculosis by the following mechanisms: malnutrition, which increases penetrance of disease following infection [45, 160], increased population density, which supports large scale epidemics of infectious disease, and crowded, inadequate housing, which is known to promote transmission of tuberculosis [30, 44, 100, 164]. Epidemics of tuberculosis followed the environmental and social changes described above; they were severe enough that some worried Native Canadian populations would succumb completely to tuberculosis [58, 161]. Native Canadians are still disproportionately affected by TB [67], but the disease is no longer associated with the very high mortality rate observed in the 19th and early 20th centuries.

B.2  Supplemental Methods

B.2.1  Aboriginal Population Descriptions

The *M. tb.* isolates from ON, MB, SK, and AB analyzed in this study were originally collected from reserves (lands formally designated for the use of specific First Nations), Métis communities (definition from Canada census, based on > 25% of total population identified as Métis), and rural communities not formally associated with specific Aboriginal populations. We did not include data from Aboriginal populations living in urban areas. ON samples are from a single region in Ontario, whereas the population base for each of the other samples was the entire province. *M. tb.* samples analyzed in this study were collected from Métis populations and a range of First Nations encompassing a diversity of Aboriginal languages, ethnicities and cultures found in the four provinces.

B.2.2  Real-time PCR Determination of Bacterial Lineage

Primer sequences for real-time PCR determination of *M. tb.* lineage (defined by polymorphisms described in [88, 185]) can be found at http://www.pnas.org. Two sets of primers and probes were designed for each polymorphism. Genomic DNA from *M. tb.* strains was tested in parallel for the wild type and the mutant sequence at each locus. If the strain was wild type at a particular locus, there would be a product from the former reaction and none from the latter (and vice versa for strains with the mutation). Wild type primer probe sets are designated with "(37)" in the name, after the laboratory reference strain H37Rv. Nomenclature of lineage-defining genomic deletions is adopted from [88, 280]. The mutant primer probe sets for Rd 219 and Rd 193 are not specific (due to technical constraints imposed by nature of flanking sequence) so mutants
were identified by absence of product with the wild type primer probe set, plus a product with the mutant primer probe set. Strains that were wild type at these loci would generate a product with both primer probe sets. For DS6\textsuperscript{Quebec}, we designed a primer probe set for one flank of the deletion. We also used a primer probe set - courtesy of Gregory Dolganov - that was within the deletion (Rv 1760). Isolates with DS6\textsuperscript{Quebec} were identified by presence of a product with the flanking primer probe set, and absence of a product with the primer probe set within the deletion (i.e. Rv 1760). Spoligotyping data (presence or absence of 43 oligonucleotide spacers at a direct repeat locus [137]) were available for some populations (QU, ON, SK): DS6\textsuperscript{Quebec} lineage was further confirmed for \textit{M. tb.} isolates from these populations by absence of spacers 9 and 10 [88]. The results of lineage typing were extrapolated to isolates from the same population, with an identical RFLP (± 1 band) and/or minisatellite genotype.

\subsection*{B.2.3 Mutation Rate Estimation}

As we have sampled only one \textit{M. tuberculosis} isolate per infected individual in this study, the appropriate timescale for mutation is per transmission generation, rather than per \textit{M. tuberculosis} cell doubling. A transmission generation (estimated to be 1.2 years in length [203], also consistent with [264]) is defined as the time from initial infection by \textit{M. tb.} to the time that the infection spreads to another individual. To estimate a per transmission generation mutation rate for minisatellite loci, we performed simulations in SIMCOAL 2.0 [73, 151]. Since no per generation mutation rate estimate in \textit{M. tuberculosis} has been made previously for minisatellite loci, we assumed a mutation rate of $2.3 \times 10^{-5}$ per locus per generation, based on estimates for minisatellite loci from parallel serial passaging experiments in \textit{Yersinia pestis} and \textit{Escherichia coli} O157:H7 [289, 290]. In the simulations, we also assumed a simple stepwise mutation model (SMM) for the minisatellite loci, where a mutation at a locus only increases or decreases the repeat number by one [177, 195]. Proceeding forward in time in the simulations, we began with one bacterium, as animal experiments suggest that most infections are initiated by a single droplet [124, 224], which, as a result of its small size, cannot contain more than a few bacilli [34]. Based on radiographic studies of pulmonary tuberculosis, we assume that the disease involves an entire lung lobe by the time it is extensive enough to result in transmission [119]. We estimated the weight of adult lung lobes from autopsy studies [132]. Finally, for an estimate of the number of \textit{M. tb.} bacilli/gram of infected lung tissue, we relied on Cannetti’s classic microscopy studies of autopsy specimens (we used the average result across different clinical-pathological classifications of specimens) [43]. Based on these data sources, our final estimate of the \textit{M. tb.} population size, at the time an infected human host transmits that infection, is $10^7$ bacilli. In the simulation experiments, we simulated exponential doubling until $10^7$ bacteria were present. At this point, we sampled 3 bacteria (an estimate of the number of bacilli contained in a single infectious droplet [34]) and counted the number of mutations in the sample compared to the original genotype. Repeating this procedure 1000 times, we obtained an estimate of 0.00125 mutations per locus per transmission generation. This agrees with the concept of a “transmission bottleneck,” where most mutations in the host’s bacterial population are lost during transmission [98].

To validate the above rate, we made use of the known RFLP mutation rate in \textit{M. tb.} of approximately
B.2. SUPPLEMENTAL METHODS

0.287 events per year for a bacterial strain with 10 copies of IS6110 [6, 228, 265]; note that we multiplied this mutation rate by 1.2 to convert the per-year RFLP mutation rate to per transmission generation. We first counted the number of minisatellite haplotypes per RFLP haplotype (for isolates from QU, SK, and the MIRU-VNTRplus database [116]) to calibrate a minisatellite mutation rate. We also estimated $\theta = 2N\mu$ using homozygosity ($H = 1/(1 + \theta)$) for all RFLP and all minisatellite haplotypes in QU using Arlequin [74]; the ratio of the $\theta$ estimates should equal the ratio of the mutation rates for the two loci. Mutation rate estimates are shown in Table B.3, and are all of a similar order of magnitude. As mentioned in Chapter 3, we settled on a mutation rate of 0.001 per locus per transmission generation. This lower rate is motivated by the fact that the RFLP mutation rates are pedigree-derived, which may be too fast for historical time scales [102, 112]. Our final estimate is also within the range from a recent report analyzing mutation rates at these loci [223] ([0.00084-0.018] per locus per generation assuming a generation time of 1.2 years).

B.2.4 Statistical Calculations

Given the limited documentation of horizontal gene transfer in M. tb. [107, 111, 254, 263, 305] we used haplotype statistics, rather than single locus statistics, in a majority of our analyses. To correct for sample size differences, we used rarefaction to obtain the number of distinct minisatellite and RFLP haplotypes (and the number of private minisatellite haplotypes) in each population (as described in [136]). Note that the number of private haplotypes is not calculated for RFLP haplotypes, which could not be compared across populations. We also calculated diversity of minisatellite and RFLP haplotypes in each population as $-\sum \pi \log(\pi)$ (where $\pi$ is the haplotype frequency and the sum is over all haplotypes; [247]). In order to correct for unequal sample sizes, we sampled $x$ individuals without replacement from each population (where $x$ is the number of individuals of the smallest sample), calculated diversity, and repeated this 5000 times. In addition, we used a permutation procedure to assess whether the observed population diversities (calculated using the same formula) could be explained by randomly partitioning isolates to all populations. We pooled all haplotypes (from QU, ON, RS, NRS, AB), and for each population with sample size $n$, calculated diversity for $n$ minisatellite haplotypes sampled with replacement from the pool. For each population, an empirical distribution was generated for 1000 samples and compared to the observed diversity. We also examined the effect of excluding QU haplotypes from the pool and repeating the same procedure. In all of the above procedures, we removed any haplotype with at least one missing genotype position.

In order to reduce noise in the following calculations, we focused only on isolates from the DS6Quebec lineage, as they share a more recent common ancestor than do isolates from all lineages and are thus less likely to be affected by homoplasy. We used the method of Ytime to calculate the time to the most recent common ancestor (TMRCA) of the DS6Quebec lineage in each population using minisatellite haplotypes [23]. This calculation, which assumes a simple SMM, involves dividing the average squared difference (ASD) in repeat length between each sample minisatellite haplotype and the putative root haplotype by the mutation rate (0.001 per locus per generation). We eliminated missing data at each locus and allowed loci to have different sample sizes in our calculation; this is in contrast to the Ytime implementation, which does not allow missing
data. Changing the root haplotype slightly alters the obtained estimates. However, the chosen root haplotype presented in Chapter 3 (the center of the DS6\textit{Quebec} haplotype network, see Figure 3.2 of Chapter 3) does seem to be the most appropriate given the star-like structure of the network (see Results).

While assumptions of demography do not affect point estimates of the TMRCA, they do affect the confidence intervals obtained for the estimates. In addition to calculating confidence intervals under the assumption of a star-like genealogy (presented in Chapter 3), we obtain confidence intervals under the assumption of constant population size and exponential population growth (shown in the Supplemental Results (Section B.3)). We also assess sensitivity of the results to the mutation rate assumed by examining estimates using mutation rates of 0.005 and 0.01, both within the range of values in Table B.3.

To calculate the divergence time between two populations “A” and “B,” we used Zhivotovsky’s [90, 91, 314, 315] $T_D$ estimator using minisatellite genotypes of the DS6\textit{Quebec} lineage in each population. The estimate is calculated as $T_D = t_D 2w$, where the average is taken over each locus and $w$ is the average mutation rate per locus (0.001 per locus per generation) (see Equation 1 of [314]). For each locus, $t_D = D_1 2V_0$, where $D_1 = (r_A - r_B)^2 + V_A + V_B$. $V_0$ is the estimated variance in repeat number in the ancestral population at the time of divergence, and in population X, $r_X$ is the mean and $V_X$ is the variance in repeat number. $V_0$ is typically unknown; results reported indicate the value of $V_0$ used. A value of $V_0 = 0$ assumes no ancestral allele size variance and provides an absolute earliest bound for $T_D$. To obtain intermediate estimates, we examined the haplotypes shared between populations “A” and “B” and calculated the variance of these haplotypes in either population; in the results, these estimates for $V_0$ are presented as Var(A) and Var(B), respectively. Standard errors are obtained using the standard deviations of the locus $t_D$ estimates [314].

As for the TMRCA estimates, we assess sensitivity of the estimates of $T_D$ to mutation rate by performing estimation under the mutation rates of 0.005 and 0.01.

Under the assumption of a simple island model, estimates of $Nm$, where $m$ is the probability that each gamete is an immigrant each generation and $Nm$ is the absolute number of migrants each generation, were obtained using Slatkin’s private alleles method for minisatellite haplotypes [251, 253]. Estimates were obtained using the equation $\log_{10}(p(1)) = a log_{10}(Nm) + b$, where $p(1)$ is the proportion of private alleles, and $a$ and $b$ are pre-determined constants for a reference sample size of 50 ([253], Equation 14). To account for unequal sample sizes of the “islands” (QU, SK, and AB), we calculated $Nm$ for a sample of 50 individuals (without replacement) from each island and obtained a distribution of $Nm$ estimates over 1000 samples. The upper and lower 2.5% tails of the $Nm$ estimates were used as an empirical 95% confidence interval. For these calculations, we removed any haplotype missing a genotype in at least one position. We exclude ON from this calculation due to its low sample size; MB is not included since minisatellite genotypes were unavailable.

As a note, in our study we refer to \textit{M. tb}. $N_e$ in the standard population genetic context. We define effective population size at the “epidemiological level,” as do Reyes and Tanaka [223]. As with molecular epidemiological studies, we analyze a single \textit{M. tb} isolate from each individual with TB, and do not attempt to model bacterial genetic diversity within individual hosts. Where each infected host is associated with a single pathogen, $N_e$ of the pathogen should be proportional to the number of infected hosts - i.e. prevalence
and a number of studies have demonstrated congruence between the demographic history inferred from pathogen genetic data and temporal trends in prevalence (e.g. [191] for a bacterial infection and [27] for a viral illness). A recent simulation study [86] found pathogen Ne to be better approximated by disease incidence. In this study (see Supplemental Results (Section B.3 below), we used the harmonic mean of annual incident TB cases as an estimate of M. tb. Ne.

B.2.5 Rejection Sampling Details

Background

Under the Bayesian framework for parameter inference, we assume data D is generated by a probabilistic model governed by the parameters \( \theta \) with prior density \( f(\theta) \). Of interest is the posterior distribution, \( f(\theta|D) \propto f(D|\theta)f(\theta) \), where \( f(D|\theta) \) is the data likelihood. For complicated models, this likelihood is difficult to compute [155, 167, 187]. For example, the likelihood for data generated under the coalescent framework, as in this study, can only be obtained after conditioning on a particular genealogy, \( G \), and integrating over the entire space of possible genealogies \( G \); i.e., \( \int_G f(D|\theta) f(G) dG \) [187]. This is an integration over an extremely large space, and is impractical to carry out.

A number of Monte-Carlo methods have been proposed for sampling from the parameter posterior distribution, \( f(\theta|D) \). Available packages employing Markov-Chain Monte Carlo methods have been used with high accuracy and success [21, 108, 109, 187]; however, these methods do not enable inference of more complicated demographic scenarios or inference with multiple types of genetic loci [298]. In our study, we attempted to use IM [108, 109, 187] for parameter inference, and observed poor performance and convergence. Alternatively, an Approximate Bayesian Computation (ABC) method, rejection sampling, has been shown to produce accurate and interpretable results for inference of coalescent models [71, 200, 210, 219, 298]. In the rejection sampling framework, rather than conditioning on the full data \( D \), several statistics \( S \) are used to summarize the full data.

The general approach is to first calculate from the data observed values for the chosen summary statistics \( (S_{obs}) \). Then:

1. Draw \( \theta' \) from its prior distribution, \( f(\theta) \).
2. Simulate data \( D \) using the coalescent and calculate summary statistics \( S \).
3. Accept \( \theta' \) if the Euclidean distance \( ||S - S_{obs}|| < \epsilon \). (Rather than an absolute cutoff \( \epsilon \), in our implementation, the threshold was chosen as a particular quantile \( (P_\delta) \) of all performed simulations (1,000,000 in total)).
4. Repeat steps 1-3 until a desired number of samples of \( \theta \) are accepted. (In our implementation, we performed 1,000,000 simulations, of which a percentage \( (P_\delta) \) were accepted).
5. Use the accepted simulations to form a posterior distribution for \( \theta \).
The outlined approach (hereafter referred to as the “Rejection” approach) assumes that the likelihood of the statistics \( f(S|\theta) \) is constant in the range where \( S \approx S_{\text{obs}} \) (for more details, see [155]). Recently, a number of modifications to the above procedure have been made that do not make this assumption [17, 70, 71, 155]. One method [155] proposes a general linear model for the summary statistics given the parameters \((\theta)\), i.e., \( S|\theta = C\theta + c_0 + \epsilon \) with \( \epsilon \sim N(0, \sigma_s) \). The coefficients \( C \) and \( c_0 \) can be estimated by ordinary least squares, and are used when calculating the posterior distribution of \( \theta \) in step 5 [155]. The authors claim that this method (hereafter referred to as the “GLM” method) takes into account the local linearity and strong colinearity of the summary statistics.

The rejection sampling framework described also allows for selection between demographic models [70, 71, 210, 219]. A model’s posterior probability can be estimated by the proportion of simulations under the model that are accepted when also considering the distances \( ||S - S_{\text{obs}}|| \) from other models; see details below.

**Implementation details**

We present in Chapter 3 the rejection sampling procedure for demographic modeling of an expansion in the Remote Saskatchewan (RS) population. Based on the diversity results (see Chapter 3 and Supplemental Results (Section B.3), it appears likely that the RS population was historically smaller than its current size, and thus likely underwent a recent expansion. Based on timing of historical events (i.e. expanded transportation networks) that would have opened these areas for development and thus rendered host populations susceptible to epidemic TB (see Supplemental Historical Background for details (Section B.1), we estimate the expansion time in RS to be around 1930. This is equal to 54 transmission generations (1.2 years in length, as described earlier) prior to the midpoint of the interval in which study samples were collected (1995). We estimated the expansion parameter \( \omega \), which is equal to the ratio of the historical effective population size (prior to the expansion) to the current effective population size. The expansion described by the parameter \( \omega \) in this model is within a reasonable time frame for estimation, as found from previous work using the rejection sampling procedure (see below, in the section Exploration of Alternative Demographic Models (Section B.2.5)).

In order to reduce noise in parameter estimation, we focus only on isolates from the DS6\textsubscript{Quebec} lineage. For the proposed demographic expansion model, we fix the time of expansion at 54 generations, and draw \( \omega \) from a prior Uniform distribution on [0.01, 1] (step 1 in the Background section (B.2.5) above). Assuming a constant size in RS after the expansion, the current effective population size was estimated to be 544. As described in Chapter 3, this estimate was based on the harmonic mean of RS TB case counts from 1986-2004, which was then multiplied by the frequency of the DS6\textsubscript{Quebec} lineage and 18, the number of years over which exhaustive samples were taken (see “Statistical Calculation” section above for an interpretation of \( N_e \)). As in step 2 above and using these demographic parameters, we performed 1,000,000 coalescent simulations, with a sample size of 195 individuals, of minisatellite haplotypes under a SMM (with a mutation rate of 0.001 per locus per transmission generation and using SIMCOAL 2.0 [73, 151]). Also as part of step 2, we calculated summary statistics from the observed and simulated minisatellite data using Python scripts. Haplotype statistics (such as the number of distinct haplotypes, the number of singleton haplotypes,
and haplotype diversity), rather than only single locus statistics, were again employed given the limited documentation of horizontal gene transfer in *M. tb*. We also calculated the mean variance in repeat size (over loci and isolates) as a summary statistic, which has desirable theoretical properties [90, 315]. These statistics have also been used in previous studies [17, 70, 71, 219]. When calculating summary statistics for individuals with one missing allele in RS (38 individuals), we imputed the missing allele as the most frequent allele at that position. Analogously, in the coalescent simulations, we randomly masked and imputed one position in 38 individuals. This imputation has little quantitative effect on the results.

For steps 3 and 4 above, to increase speed, the program *ABCEstimator* was used to calculate Euclidean distances between the observed and simulated summary statistics ($||S - S_{obs}||$) and to output a list of simulated parameter sets accepted under a particular distance threshold $P_\delta$ [298]. Results were not sensitive to the $P_\delta$ threshold; we used a cutoff of 5,000 retained simulations out of 1,000,000 simulations ($P_\delta = 0.005$). For step 5, the *ABCEstimator* program was used to obtain posterior distributions under the “GLM” post-rejection approach; posterior distributions under the regular “Rejection” approach were obtained using the density() function in R (see Background section (B.2.5)).

We also performed model selection under the rejection sampling framework by estimating the posterior probability of a population expansion model versus a null constant size model. We first performed an additional 1,000,000 simulations under a null constant size model. Then, we ranked the 2,000,000 simulations of both models (1,000,000 from each model) according to their Euclidean distances $||S - S_{obs}||$ from the observed summary statistics. (Note that for the constant size model, no parameter is estimated, and only the Euclidean distances $||S - S_{obs}||$ are computed). We calculated the posterior probability of each model as the proportion of simulations of each model in the smallest 500 (of the 2,000,000) distance values; results were not sensitive to these cutoffs. The method naturally penalizes for models with more parameters, as the parameter space is less explored for these models (see [298]). We performed this analysis using scripts in R.

**Assessment of Sensitivity to Mutation Rate**

We also assessed the sensitivity of the ABC procedure to the assumed mutation rate. In addition to performing parameter estimation assuming a mutation rate of 0.001 per locus per generation, we followed the identical estimation procedure using mutation rates of 0.0005, 0.005, and 0.01. These mutation rate estimates are in the range of values presented in Table B.3 and are also in the range of values estimated by Reyes and Tanaka (2010) ([0.00084-0.018] per locus per generation assuming a generation time of 1.2 years) [223].

In addition to using alternative point estimates of the mutation rate, we allowed the mutation rate to have a Uniform prior on [0.005, 0.02] (again motivated by the range of estimates described above). We follow the same procedure in estimating $\omega$ by examining $\omega$ values in the top $P_\delta$ of simulations with the lowest distance values; analogously, we examined the mutation rate values in the top $P_\delta$ of simulations to co-estimate the mutation rate parameter.
Method Validation Procedures

We validated the proposed method for inference of the expansion parameter ($\omega$) in the one-population model (where again, $\omega$ is the ratio of the historical population size before the expansion to the current population size). First, for 100,000 coalescent simulations with $\omega \sim U[0.01, 1]$ (as described above), we examined the correlations between summary statistics, and with $\omega$. Then, to assess the accuracy of parameter estimation, we simulated datasets with known values of $\omega$ (300 simulated datasets each for $\omega = 0.1, 0.5$ and 0.8) under an identical expansion model, again using SIMCOAL 2.0. To estimate the $\omega$ parameter for these simulated datasets, we employed the rejection sampling procedure with the same 1,000,000 simulations (under the expansion model) used for inference of the observed RS data. Using a tolerance of $P_\delta = 0.005$ as described above, means and modes of the posterior distribution for $\omega$ were obtained for each dataset using both the Rejection and GLM methods. The mean squared error was calculated for each dataset as $\text{MSE} = \sum (\omega - \hat{\omega})^2$, where $\omega$ is the true assumed value ($\omega = 0.1, 0.5$ or 0.8) and $\hat{\omega}$ is the estimated value (taken as the posterior mean). Standard errors of the MSE were calculated by bootstrapping over all datasets.

To assess the performance of model selection, we followed the procedure used for the RS data and used the 1,000,000 simulations under the expansion and constant size models (2,000,000 simulations in total). The ability to select the expansion model over the constant size model was assessed for each of the 300 simulations under each assumed $\omega$ (0.1, 0.5, and 0.8, as above). We also performed model selection for 300 datasets simulated under a constant size model ($\omega = 1$). For each of the 300 datasets (for each $\omega$ value), we calculated the posterior probability of both models as described (by examining the proportion of each model in the smallest 500 (out of 2,000,000) Euclidean distance values). A posterior probability above the liberal cutoff value of 0.5 was used to select one model over another.

Exploration of Alternative Demographic Models

Originally, our goal was to estimate the parameters of a bottleneck in the Non-Remote Saskatchewan (NRS) population when founded from Quebec. We intended to compare several demographic models: the first model was a founder model with a bottleneck of size $N_{BN}$ in the NRS population when founded from QU; the second model allowed for bi-directional migration at rate $m$ during the bottleneck. We fixed the time of the bottleneck ($T_1 = 164$) and the time of population expansion ($T_2 = 100$) based on historical timings described in Chapter 3 (see Table 3.2). The parameters to be estimated were $N_{BN}$ (ratio of the bottleneck size to the current size) and $m$ (the rate of migration between the two populations, in the migration model only). Uniform priors were assumed for $N_{BN}$ ([0.01, 1]) and $m$ ([0.001, 1]). As described for the RS inference, current effective population sizes ($N_e$) in each population were estimated from historical TB incidence data: because we assumed constant population sizes (before and after the expansion in NRS and at all times in QU), we took the harmonic mean of case counts and multiplied this by the DS6Quebec lineage frequency and the number of years over which samples were collected to obtain the $N_e$ of each population (as for the RS $N_e$ estimation).

We performed simulations of 12 minisatellite loci (with the fixed parameters $T_1$, $T_2$, $N_eQU$ and $N_eSK$).
and variable parameters $N_{BN} \sim U[0.01, 1]$ and $m \sim U[0.001, 1]$) for use in the rejection sampling procedure. The summary statistics $S$ of the minisatellite data used in the procedure provided information not only about the relationships between the two populations but also about each individual population and have been used in previous studies [17, 70, 71, 219]. As in the RS inference, we examined haplotype-based statistics such as haplotype diversity, as well as the proportion of private and shared minisatellite haplotypes out of the total number of distinct haplotypes. We also calculated $\delta \mu^2$, the average squared distance in allele size between populations, and the variance in allele size within populations [90, 315]. As in the RS inference, for individuals with one missing allele in NRS, we imputed it as the most frequent allele at that position for these calculations.

After implementing the rejection sampling approach and method validation procedures as described previously (modified as above), we were unable to estimate even $N_{BN}$ alone with reasonable accuracy. We were also not able to accurately estimate $m$, nor perform model selection between the bottleneck and migration models. We explored possible reasons for the poor performance of the method (for example, by allowing for only uni-directional migration and by decreasing the ranges of the prior distributions). We discovered differences in performance when the fixed timings of the model - the founding time ($T_1$) and the expansion time ($T_2$) - were changed. Performance was consistently poorer for demographic models where either event occurred further in the past; this is likely due to the high mutation rate of the minisatellite loci. Accuracy considerably improved when the demographic events occurred within 25-50 generations in the past. Given the fixed historical times of the bottleneck and expansion in NRS, we concluded that we were unable to estimate the parameters of the proposed model with our method and data.

B.3 Supplemental Results

B.3.1 Genetic Timing Estimates

We calculated estimates of TMRCA for additional populations (other than QU and SK, presented in Chapter 3) using the method of Ytime. Point estimates and 95% confidence intervals in generations from the present for each population are as follows; Non-Remote Saskatchewan (NRS): 150.4 (24.4, 376.7); Remote Saskatchewan (RS): 192.7 (38.4, 385.1); and Alberta (AB): 149.3 (45.3, 417.2). These results also follow historical expectations under the Canadian fur trade (as described in Chapter 3), with calendar year estimates for the preceding TMRCA point estimates approximately equal to 1814, 1791, and 1779, respectively. Since the DS6Quebec network is not perfectly star-like (not all haplotypes are derived immediately from the root haplotype), the accuracy of TMRCA estimates may be affected by this violation of Ytime model assumptions. We also explored sensitivity of TMRCA estimates to the assumed root haplotype by obtaining estimates with alternative root haplotypes. Alternative root haplotypes generally increased the TMRCA estimates, supporting that the haplotype at the center of the DS6Quebec network (haplotype “233325153324,” at high frequencies in all populations, results presented in Chapter 3) is the most appropriate. For example, using instead the root haplotype “233325153325” (at low frequency in QU but the next most frequent haplotype in
both SK and AB), increases the TMRCA in QU from 212.5 to 298.2 and in SK from 164.8 to 209.9 years from the present.

For the populations for which TMRCA is presented in Chapter 3 (QU, AB, SK), we obtained point estimates under alternative mutation rate estimates. As expected, the estimates scale directly with the mutation rate: for mutation rates of 0.005 and 0.01, respectively, the TMRCAs for QU are 1944 and 1969, for SK are 1955 and 1975, and for AB are 1952 and 1973. Although TMRCA estimates certainly are sensitive to mutation rate, our conclusions regarding spread of the DS6\textsuperscript{Quebec} lineage via the fur trade do not rely on these estimates. We cite other strong evidence (lineage frequencies, haplotype networks and genetic differentiation, and genetic diversity) in support of the fur trade hypothesis, none of which depend on these timing estimates. In fact, given their large confidence intervals, these timing estimates do not contradict our conclusions.

We calculated additional confidence intervals for the TMRCA estimates under different demographic assumptions (using the original mutation rate of 0.001). First, we obtained estimates under an exponential growth model with $r = 0.0473$. (Given the point estimate obtained from the ABC procedure of the ratio of the current to ancestral population size in RS ($1/\omega = 12.886$), we estimate $r = 0.0473$ given exponential growth over 54 generations). Though this rate may not hold for all populations, it is well-motivated in that it assumes a similar rate of growth over a similar time period in all populations. Confidence intervals under this demography were, in calendar years: QU (1435-1932); SK (1552-1960); AB (1516-1950). These results are very similar to those obtained under the assumption of a constant population size: QU (1414-1934); SK (1537-1952); AB (1495-1941).

We then calculated TMRCA confidence intervals under the assumption of a “star-like” genealogy. For all populations, the confidence intervals of the timing estimates are much narrower, likely because actually specifying a haplotype genealogy reduces uncertainty in the TMRCA estimation. Confidence intervals for each population under a star-like demography were, in calendar years: QU (1709-1771); SK (1777-1815); AB (1751-1804). Thus, under the assumption of a star-like genealogy and a mutation rate of 0.001, these TMRCA estimates are quite consistent with introduction of the DS6\textsuperscript{Quebec} lineage to indigenous populations via the fur trade.

We also estimated the divergence time ($T_D$) between NRS and QU, using the method of Zhivotovsky [314] as described in Chapter 3 and Supplemental Methods (Section B.2). Estimates using several values of $V_0$ were obtained: 1) $V_0 = 0$, 2) variance in NRS of haplotypes shared between NRS and QU, and 3) variance in QU of haplotypes shared between NRS and QU. For these $V_0$ values, the respective divergence time estimates between QU and NRS in generations from the present (with corresponding dates in calendar year), are 167.5 (1794), 83.2 (1895), and 89.6 (1887). Standard deviations of each of these estimates are 71.3, 30.5, and 31.5, respectively. Although confidence intervals and ranges of possible $T_D$ values are large, historical timing estimates of the fur trade are within these ranges, as described in Chapter 3.

For the population pairs for which $T_D$ is presented in Chapter 3 (QU/SK and QU/AB), we obtained point estimates under alternative mutation rate estimates. Using a higher mutation rate of 0.005, the three point estimates (in the same order as in Chapter 3 Table 3.2 and above), are more recent: QU/SK: 1954, 1973,
1978; QU/AB: 1952, 1976, 1973. Using an even higher mutation rate of 0.01, the three point estimates are: QU/SK: 1974, 1984, 1986; QU/AB: 1973, 1986, 1984. As mentioned for the TMRCA estimates, although the mutation rate does greatly affect the TD estimates obtained, our conclusions do not rely on these timing estimates, nor do these more recent timing estimates contradict our conclusions.

B.3.2 Patterns of Genetic Differentiation

We observed lineage-specific patterns of *M. tb.* genetic differentiation, with as much as 52% of total variation (H37Rv-like lineage) partitioned among populations. Differentiation of lineages other than DS6^Quebec^ may be due to the fact that contact occurred between small segments of sending and receiving populations, and/or the temporary nature of contact (i.e. other lineages were introduced later than DS6^Quebec^). Along with recurrent population bottlenecks, such patchy and stuttering *M. tb.* migration dynamics may help explain low levels of DNA sequence diversity observed in *M. tb.* populations.

B.3.3 Estimates of Population Diversity

Results of diversity calculations from different combinations of populations and other genetic loci gave results consistent with the RFLP results described in Chapter 3. Using rarefaction to examine the number of distinct RFLP types in each population (as in Figure 3.3A of Chapter 3) on splitting SK into RS and NRS, we find that QU has the highest, and RS and MB the lowest, numbers of distinct haplotypes; NRS, AB, and ON all have intermediate numbers of distinct haplotypes. Calculating diversity of RFLP types without splitting the SK population (with re-sampling to the size of the smallest sample to correct for sample size differences as in Figure 3.3B in Chapter 3) we find results consistent with Figure 3.3A of Chapter 3: SK and MB have the lowest, QU has the highest, and AB has intermediate diversity. Note that for the calculation of diversities using this re-sampling procedure, we did not include ON due to its small sample size (*n* = 45). Rarefaction using minisatellite haplotypes for both the number of distinct and the number of private haplotypes produces a pattern entirely consistent with RFLP results when splitting SK into NRS and RS (Figure B.2A and B.2B) and when not splitting the SK population; the ordering of the populations is similar for both the number of private and the number of distinct haplotypes. We also find that minisatellite haplotype diversity data follows the patterns described above and presented in Chapter 3 (not splitting SK, Figure B.2C; splitting SK, Figure B.2D). Results from the permutation procedure assessing whether the observed minisatellite haplotype diversity in each population can be explained by a random partitioning of haplotypes into each population are shown in Figure B.3. In Figure B.3A, we include QU isolates in the pool of common haplotypes (along with NRS, RS, AB, and ON isolates), and find that when isolates are randomly distributed among populations, the observed diversity in QU is higher than expected; in RS, it is lower than expected. We then exclude QU isolates from the pool of haplotypes (Figure B.3B), and find that when isolates are randomly distributed, only the diversity in RS is lower than expected. (Again, minisatellite haplotypes are not available from MB strains).
Based on these diversity and permutation calculations using multiple genetic loci, we confirm the conclusion stated in Chapter 3 that the observed genetic diversities are consistent with the pattern of social networks created as a byproduct of the French Canadian fur trade. Consistently, QU has the highest diversity (and highest number of private haplotypes), as expected from its role as a source population for the isolates in all other populations. Diversity and the number of private haplotypes are generally intermediate in the founded populations of AB, NRS, and ON. Finally, diversity and the number of private haplotypes are lowest in historically remote regions (parts of MB and RS, see Figure 3.1A of Chapter 3), consistent with \textit{M. tb.} from these regions undergoing a more recent expansion in population size.

**B.3.4 Rejection Sampling Validation Results**

First, we examined the correlations between the summary statistics $S$ (number of distinct haplotypes, number of singleton haplotypes, haplotype diversity, and mean variance in repeat sizes) used in the rejection sampling procedure for the expansion model in RS. Diversity and the number of distinct haplotypes are highly correlated (correlation of 0.8888), as are the number of singletons and the number of distinct haplotypes (0.8765). Variance in allele size is only moderately correlated with the number of distinct haplotypes and the number of singletons (correlations of 0.6032 and 0.5818, respectively); haplotype diversity and the number of singletons have a modest correlation of 0.6625. Below, we discuss performance of the rejection sampling procedure in estimating $\omega$ in order to assess whether these correlations warrant the removal of any of the summary statistics.

Motivated by the above correlations, we examined the accuracy of the $\omega$ estimates obtained when using only two summary statistics (the variance in allele size and haplotype diversity), rather than all four correlated statistics (including the number of singletons and the number of distinct haplotypes). For the Rejection procedure, we find that there is little effect of removal of the two summary statistics (the number of singletons and distinct haplotypes) on the $\omega$ estimates (assessed by mean squared error, or MSE) (Figure B.4A; conclusions are similar for the GLM results). Because of this, we decided to use all four summary statistics in our subsequent analyses. As expected, accuracy of the procedure is consistently higher for more pronounced expansions (low $\omega$), and accuracy decreases as the historical population size before the expansion becomes more similar to the current size ($\omega$ close to 1).

In Figure B.4C, we show estimates of $\omega$ from the 300 test simulations for each of the assumed $\omega$ values. For pronounced expansions ($\omega = 0.1$), both posterior means and posterior modes are accurately centered on low $\omega$ values for the Rejection approach. For intermediate $\omega$ values, posterior modes are almost uniformly distributed across the prior interval, indicating more noise in the estimated posterior densities. True $\omega$ values near 1 result in generally higher posterior mean and mode estimates, although the spread of $\omega$ estimates is quite large. For all $\omega$ values, we find that the post-rejection GLM procedure proposed by Wegmann et al. [298] does not significantly improve performance. For the remainder of our results, we use only estimates obtained from the Rejection, and not the GLM, procedure.

Next, we examine the ability of the method to distinguish between the expansion model and a null constant
population size model based on the posterior probabilities of each model (selecting a particular model if its posterior probability is above 0.5). Interestingly, the combinations of summary statistics used slightly alter the model selection results (Figure B.5B); another motivation for choosing to use all four summary statistics, rather than only diversity and variance in allele size, is that it generally improves model selection performance for the expansion models. Model selection is 100% accurate for pronounced expansions ($\omega = 0.1$), and as expected, becomes less accurate as the expansions become milder (with $\omega = 0.8$, the expansion model is selected over the constant size model only 60% of the time). Model posterior probabilities are close to 0.5 for mild expansions (histograms of the actual posterior probabilities from each simulation are shown in Figure B.5D), indicating that as the population expansions become less pronounced, the two models become harder to distinguish. The false positive rate of our method (selecting the expansion model when data are simulated under the null constant size model) is $\approx 40\%$ when using all four statistics (Figure B.5B). Despite this high rate, we are able to confidently assign p-values to the posterior probabilities of the expansion model. This is done by using a null distribution of posterior probabilities estimated from data simulated under the null hypothesis of constant population size (Figure B.5D shows the converse, bottom right panel); see results below.

### B.3.5 Results of Rejection Sampling Procedures

Summary statistics and observed values calculated for the imputed RS genotypes are as follows: number of distinct haplotypes (14); number of singletons (6); minisatellite haplotype diversity (1.6925); and average variance in allele size over loci (0.1077). These four summary statistics used with the rejection procedure outlined previously gave the posterior distribution for $\omega$ shown in Figure B.4A. We also find that the observed summary statistics for RS are similar to those for the retained simulations, with slightly more variation for variance in allele size and the number of distinct haplotypes (Figure B.4B).

With the methods described above, we calculated a posterior probability of 1 for the expansion model versus the null constant size model. As a null distribution for this posterior probability, we used the posterior probabilities of the expansion model estimated for each of the 300 constant size model simulations (described above). From this null distribution, we determined that the p-value of our observed posterior probability (1) is 0; thus, the expansion model with the additional parameter $\omega$ is indeed significant in the RS population. As stated in the Methods (Chapter 3), we assessed the sensitivity of the procedure with the observed RS data to the minisatellite mutation rate assumed. First, in the coalescent simulations used for inference, we used different point estimates of the mutation rate (0.0005, 0.005, and 0.01; i.e., both above and below the current assumed value of 0.001). From these simulations, we inferred the expansion parameter $\omega$ for the observed data, as well as the posterior probability of the expansion model and its corresponding p-value.

For mutation rates of 0.005 and 0.01, nearly all observed summary statistics are outside of the range of simulated summary statistics for both the expansion and contraction models (for example, the observed value of 14 distinct haplotypes is far lower than the number distinct haplotypes in any simulation with these mutation rates - in contrast to Figure B.4B). This indicates that these mutation rates are likely too high to
produce the patterns of diversity that we observe. For all mutation rates, the 95% credibility intervals of $\omega$ for the observed data are generally larger than those obtained using a mutation rate of 0.001. Point estimates and 95% credibility intervals for $\omega$ are as follows for the indicated mutation rates: 0.0005: $\omega = 0.384$ [0.08, 0.89]; 0.005: $\omega = 0.144$ [0, 0.76]; 0.01: $\omega = 0.321$ [0.02, 0.94]. The wider credibility intervals for $\omega$ under these mutation rates are expected. In our analyses examining alternative demographic scenarios (see previous section), we inferred that the mutation rate of 0.001 allowed us to accurately infer events between 25-50 generations in the past. The mutation rate of 0.0005 is thus likely too slow to allow accurate inference for events occurring relatively recently. Aside from producing summary statistics largely inconsistent with the observed data, the high mutation rates likely result in too much homoplasy to accurately infer events at this time scale, resulting in larger credibility intervals.

For all mutation rates, we examine the posterior probability of the contraction model over the constant size model for our observed data. In all cases, the expansion model is significant. Posterior probabilities and p-values (in parentheses) for each mutation rate are as follows: 0.0005: 0.852 (0.02); 0.005: 1 (0); 0.01: 0.976 (0). Thus, for the observed RS data, we can reject the constant size model regardless of mutation rate. We also assessed the effects of drawing the mutation rate for the ABC procedure on a uniform distribution from [0.0005, 0.02] (mirroring the range of Reyes and Tanaka [223]). In performing estimation for $\omega$ in a similar manner, the estimate ($\omega = 0.234$) had a large credibility interval ([0.016, 0.824]), likely reflecting the increased uncertainty in the mutation rate parameter. When estimated in conjunction with $\omega$, the mutation rate parameter had a posterior credibility interval on [0.0005, 0.0012]. Notably, this credibility interval for the mutation rate includes our assumed rate of 0.001.

This result should be qualified, as given that our method examines only four summary statistics, a confident estimation of two parameters from these statistics may not be possible. However, we do conclude that our assumed mutation rate of 0.001 is in fact reasonable based on this joint estimation using the ABC procedure. For the observed data, the posterior probability of the expansion model over the constant size model (0.96) was again significant (p-value = 0.0033). In all cases, regardless of mutation rate, the constant size model is rejected in favor of the contraction model in RS. Due to the timing of the expansion event of interest (54 generations in the past), estimates of the $\omega$ parameter have larger credibility intervals at alternative mutation rates for the reasons described above. Nevertheless, our conclusions regarding the expansion in RS are robust to alternative assumptions of minisatellite mutation rates.

### B.3.6 Historical Estimates of Migration between Quebec and the Northwest, 1710-1870

In order to estimate the number of French Canadian migrants to Western provinces, we have relied on fur trade documents (employee records), as there are no reliable censuses for this region, during this time period. Fur trade documents refer to the “Northwest” (NW), which encompasses the region described in the Supplemental background material (Section B.1) (present day Ontario west of “Grand Portage” at the Western edge of Lake Superior, Manitoba, Saskatchewan, Alberta, British Columbia, the Northwest Territories).
Expansion of the fur trade into this region (from the Montreal route) occurred after 1710 [120]. From 1710-1800, trade in the region increased. Between 1720 and 1730, the annual employee count in the Northwest was \( \approx 49 \) men. This increased to \( \approx 140 \) men employed annually in the years from 1739-1752. Employee accounts from 1773 and 1798 were 205 and 917 employees, respectively. For the period 1710-1800, we use the average of these censuses, 328, as an estimate of the number of employees per year. From 1800 to 1820, trade remained relatively stable [120]. For this period we use an account from 1805 of 1,610 men. After the merger of the two major fur trading companies (the North West Company and Hudson’s Bay Company) in 1820, trade was consolidated and the Montreal route was abandoned, although a number of French Canadian voyageurs were kept on by the Hudson’s Bay Company. For this period we use as an estimate of the annual count of employees the number of voyageurs employed by the Hudson’s Bay Company in 1857, equal to 500 men (all preceding figures are from Innis [120] and references within). After 1870, traditional forms of fur trading concluded [120], as did Westward migration of French Canadians [92].

In order to estimate the absolute number of migrants to the Northwest using these employee counts, it is necessary to have an estimate of employee turnover: that is, how long did fur trade employees remain in the interior before dying, or returning to Quebec? A contract extended over at least two years, as the length and difficulty of the journey precluded outbound and return travel in a single season. Contracts for low level employees (including voyageurs) were generally a few years in length, but could extend for as long as 6 years [120]. Employees commonly extended their residence in the interior over multiple contracts. In order to estimate the average length of time spent in the interior, we used the published analysis of a 1785-6 account book from a Saskatchewan fur trade post [66]. The authors of this study cross-referenced names and demographic information from European traders mentioned in the account book with other archival material and were thus able to describe the career trajectories of the 105 individuals named therein. For each individual, we calculated the number of years spent in the interior, assuming that it was two years for individuals who appear in this journal but for whom no other records exist. Under these assumptions, European traders spent an average of 16 years in the interior.

Calculations performed to obtain our historical estimate of \( Nm \) (absolute number of \( M. \) tuberculosis infections introduced per transmission generation) are shown in Table B.2, with each column indicating the three historical periods over which estimates were obtained. We assume that the population turned over every 16 years, so that the total number of migrants per population is the product of the annual census and the number of turnovers expected for the period. With these assumptions, the total number of migrants from 1710-1870 (over all three periods) is equal to 964 individuals per population (i.e. to each of ON, MB, SK, AB, BC and NWT, assuming that migrants were divided equally among receiving populations). Based on 1% annual tuberculosis mortality in European cities in the mid 18th century [94], 50% mortality among cases and disease duration of 2 years [261], we estimate a prevalence of 4% of infectious tuberculosis among the French Canadian colonists. This may be high, since the healthy may be over-represented among immigrants [46]. With 4% prevalence of tuberculosis disease among (human) migrants to the Northwest, 39 infections would have been introduced to each population.
For the purpose of comparison with genetic analyses of migration based on an island model, contact under this model extends from 1710 to 1995 (midpoint of the study period). Dividing this by the generation time of 1.2 years results in an estimate of 237.5 transmission generations of contact. Finally, our historical estimate of $N_m$ is equal to the absolute number of migrant infections (39) divided by the number of generations (237.5) - this yields $N_m = 0.16$. If we assume that the number of migrants to each receiving population was proportional to the number of regional trading posts (instead of assuming that migration was equal to each population), historical estimates of $N_m$ range between 0.04 (BC) and 0.25 (SK, AB). These calculations are made under a number of simplifying assumptions; as a result, we intend this result to be an order of magnitude estimate (i.e., historical $N_m < 1$).
B.4 Supplemental Figures and Tables

Figure B.1: \textit{M. tb.} minimum spanning trees based on minisatellite data and lineage typing. \textit{M. tb.} isolates from QU are in red, ON in yellow, SK in green and AB in purple (note minisatellite genotypes unavailable for MB). The Rd 219 lineage is shown in panel A, Rd 182 is shown in panel B, and the “H37Rv-like” lineage is shown in panel C. H37Rv-like isolates have the katG CTG to CGG mutation, indicating that they are members of the Euro-American clade, but lack any of the genomic deletions that would identify them with a specific lineage within the clade.
Figure B.2: Diversity calculations obtained from minisatellite data. A) Number of distinct minisatellite haplotypes as a function of the number of sampled chromosomes obtained using rarefaction (see Methods, Chapter 3). Populations are Quebec (QU), Ontario (ON), Non-Remote Saskatchewan (NRS), Remote Saskatchewan (RS), and Alberta (AB) (MB not included since minisatellite genotypes not available). B) Number of private minisatellite haplotypes as a function of the number of sampled chromosomes; as in part A. C) Minisatellite haplotype diversity, corrected for sample size differences by re-sampling to the size of the smallest sample (AB, $n = 266$), as described in text. Populations are Quebec (QU), Saskatchewan (SK), and Alberta (AB). D) Minisatellite haplotype diversity, corrected for sample size differences by re-sampling to the size of the smallest sample (NRS, $n = 107$). Populations are Quebec (QU), Non-Remote Saskatchewan (NRS), Remote Saskatchewan (RS), and Alberta (AB).
Figure B.3: Distribution of minisatellite haplotype diversities obtained from the permutation procedure under a random distribution of isolates to populations (described in Supplemental Methods (Section B.2)), with single vertical lines indicating the observed diversities in each population. A) Histograms are obtained by pooling isolates from all populations (QU, ON, AB, NRS, RS), and for each population, randomly selecting the indicated number of isolates ($n$) and calculating haplotype diversity; this is repeated 1000 times. B) Histograms are obtained from pooling isolates from only the populations Ontario, Alberta, Non-Remote Saskatchewan, and Remote Saskatchewan (not Quebec).
Figure B.4: **Results from rejection sampling procedure in RS.** A) Posterior density (main figure) and cumulative posterior density (inset) of $\omega$, the size of RS prior to its expansion in 1930 (relative to the current population size). After translating the $\omega$ estimates to absolute numbers, the current size of RS is 30.222 and the historical size was 2.345; 95% Credibility Interval: [0.351, 8.803] (obtained from the cumulative density). B) Summary statistics from the retained simulations used to generate the posterior distribution for $\omega$ in the Rejection procedure. Red vertical lines indicate the observed summary statistics (given in Supplemental Results (Section B.3)).
Figure B.5: Validation results for rejection sampling Procedure. A) Mean squared errors (MSEs) for (estimated using rejection sampling) when data are simulated under different true values assumed for $\omega$ (x-axis); $\omega = 1$ is the constant size model. Different colors indicate the different statistics used for parameter inference (“All Statistics” includes the variance in allele size, haplotype diversity, the number of singletons, and the number of distinct haplotypes). Solid lines indicate use of the rejection procedure, and dashed lines indicate the GLM procedure (described in Supplemental Methods (Section B.2.5)). B) Model selection performance (y-axis), given as a proportion of the 300 simulations for which the correct demographic model was chosen (i.e., had a posterior probability greater than 0.5); as in part A), this is plotted against the true assumed $\omega$ value of the simulation (x-axis). For true $\omega$ values of 0.1, 0.5, and 0.8 (expansion model), the y-axis indicates power; for $\omega = 1$ (constant size model), the y-axis indicates 1 minus the false positive rate. Different color lines indicate the different statistics used, as in part A. C) Rejection estimates of $\omega$ obtained from 300 simulations with the indicated assumed true $\omega$ values (left to right: $\omega = 0.1$, 0.5, and 0.8). Histograms of the posterior means and the posterior modes are shown. D) Histograms of posterior probabilities of the expansion demographic model obtained for the 300 simulations under various true assumed $\omega$ values (expansion model: $\omega = 0.1$, 0.5, and 0.8; constant size model: $\omega = 1$).
Table B.1: **Frequencies of *M. tuberculosis* lineages in four populations.**

<table>
<thead>
<tr>
<th>Lineage</th>
<th>QU n (%)</th>
<th>ON n (%)</th>
<th>SK n (%)</th>
<th>AB n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS6\textsuperscript{Quebec}</td>
<td>144 (48)</td>
<td>17 (38)</td>
<td>276 (62)</td>
<td>157 (55)</td>
</tr>
<tr>
<td>H37Rv-like</td>
<td>53 (18)</td>
<td>14 (31)</td>
<td>152 (34)</td>
<td>96 (34)</td>
</tr>
<tr>
<td>Rd 115</td>
<td>5 (2)</td>
<td>0</td>
<td>1 (0.2)</td>
<td>2 (0.7)</td>
</tr>
<tr>
<td>Rd 174</td>
<td>1 (0.3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rd 182</td>
<td>55 (19)</td>
<td>1 (2)</td>
<td>3 (0.7)</td>
<td>4 (1)</td>
</tr>
<tr>
<td>Rd 183</td>
<td>1 (0.3)</td>
<td>5 (11)</td>
<td>0</td>
<td>2 (0.7)</td>
</tr>
<tr>
<td>Rd 219</td>
<td>32 (11)</td>
<td>7 (16)</td>
<td>12 (3)</td>
<td>19 (7)</td>
</tr>
<tr>
<td>Non-Euro-American clade</td>
<td>6 (2)</td>
<td>1 (2)</td>
<td>0</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Total sample</td>
<td>297</td>
<td>45</td>
<td>444</td>
<td>283</td>
</tr>
</tbody>
</table>

*These isolates have the katG CTG to CGG mutation, indicating that they are in the Euro-American clade, but they do not have any of the lineage defining genomic deletions (DS6\textsuperscript{Quebec}, Rd 115, and so forth).*
### Table B.2: Fur trade migrants, Quebec to the Northwest, 1710 - 1870.

<table>
<thead>
<tr>
<th>Population Data</th>
<th>Phase of Fur Trade</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Growth (1710-1800)</td>
<td>Stabilization (1800-1820)</td>
<td>Contraction (1820-1870)</td>
<td></td>
</tr>
<tr>
<td>Northwest (NW) census*</td>
<td>328</td>
<td>1,610</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Populations within NW†</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Census per population</td>
<td>66</td>
<td>268</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Time (years)</td>
<td>90</td>
<td>20</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Employee turnovers</td>
<td>5.625</td>
<td>1.25</td>
<td>3.125</td>
<td></td>
</tr>
<tr>
<td>Total migrants per population</td>
<td>369</td>
<td>335</td>
<td>260</td>
<td></td>
</tr>
</tbody>
</table>

*Annual census of French Canadian fur trade employees resident in the Northwest.
†The number of distinct populations within the NW. For 1710-1800, this includes Western Ontario, Manitoba, Saskatchewan, Alberta and the Northwest Territories. After 1800, there was significant trading activity in British Columbia as well.
Table B.3: Estimated mutation rates for *M. tuberculosis* minisatellite loci, per locus and per transmission generation; mutation rates on which the estimates are based are shown also, with appropriate units.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mutation rate used</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmission simulations</td>
<td><em>Y. pestis</em> and <em>E. coli</em> (per generation): 2.3 x 10^5</td>
<td>0.00125</td>
</tr>
<tr>
<td>Comparison with RFLP (SK and QU)</td>
<td>RFLP (per year): 0.287</td>
<td>0.0074</td>
</tr>
<tr>
<td>Comparison with RFLP (MIRU-VNTR_plu database)</td>
<td>As above</td>
<td>0.0115</td>
</tr>
<tr>
<td>(\theta_{RFLP}:\theta_{MIRU}) (QU)</td>
<td>As above</td>
<td>0.0069</td>
</tr>
</tbody>
</table>
Supplemental Methods

C.1.1 Samples

Table C.1 lists details of the samples used for each analysis. HapMap Maasai trio parents from Kinyawa, Kenya (MKK) (46 individuals) were included for population differentiation and haplotype analyses [121, 122, 123]. As Pemberton et al. [201] found evidence of undocumented relatedness between individuals in the Maasai, we removed a subset of the related individuals (NA21384, NA21475, NA21399, NA21365, NA21362, NA21382, NA21423, NA21453, NA21615, and NA21634), leaving a total of 46 Maasai individuals.

For haplotype statistic calculations, SNPs with a missing genotype rate > 10%, SNPs with minor allele frequency < 0.5% (i.e. singletons), and SNPs out of Hardy-Weinberg equilibrium (assessed in each population independently) with a p > 0.001 were removed.

C.1.2 Population Differentiation

We first analyzed patterns of population differentiation of all autosomal SNPs for evidence of local adaptation driven by positive selection. Differences in allele frequencies between populations measured by statistics such as $F_{ST}$ may have the power to detect selective events that have occurred less than 50,000 to 75,000 years in the past [4, 231].

We calculated a per-SNP $F_{ST}$ over all 11 populations (all populations except the HapMap YRI; see Methods, Chapter 4), a per-SNP derived allele frequency difference ($\delta$) between all pairs of populations, and mean $F_{ST}$ averaged over all genotyped SNPs between all pairs of populations. For these calculations, several admixed individuals were removed based on their inferred proportions of ancestry in ancestral clusters (see Figure 1 of Henn et al. [103]). Excluded individuals in the Hadza were BAR09, END13, END15, END21, END22, and END07; in the $\neq$Khomani Bushmen, excluded individuals were SA25, SA59, SA39, SA58, SA45, SA49, SA40, and SA03 (leaving 11 Hadza and 23 $\neq$Khomani Bushmen) (Table C.1).

Enrichments of genic and non-genic SNPs (see Chapter 4) were calculated for the absolute value of derived allele frequency difference ($|\delta|$) between pairs of populations for each pair separately (55, or $\binom{11}{2}$, pairs), as in Coop et al. [52] (using a bin size of 0.1). An enrichment in a given bin was calculated as the
proportion of genic (or nongenic) SNPs in that bin divided by the proportion of genic (or nongenic) SNPs among all genotyped SNPs. For a clearer picture, we counted the number of population pairs for which each given \( |\delta| \) value was defined (some closely-related populations, for instance, may not have SNPs with a \( |\delta| \) as high as 0.9.) We then counted the number of pairs for which, at each \( |\delta| \) value, there appeared to be an enrichment of genic SNPs (an enrichment value > 1). We also averaged the enrichment values of genic and non-genic SNPs in each \( |\delta| \) bin over all 55 population pairs as another summary of enrichment. Here, we examined the absolute value of \( |\delta| \), as when examining multiple population pairs, the polarization of a SNP is arbitrary.

We then examined more closely population differentiation between several individual population pairs (out of the 55 total pairs). For each pair, we calculated \( \delta \) (since we examined each pair individually, we did not take the absolute value) and assessed the significance of genic and non-genic enrichments using bootstrapping. Briefly, bootstrapping involved resampling 200kb regions of the genome with replacement, and re-calculating enrichment values in each bootstrap sample to generate confidence intervals (see Chapter 4).

Over all 55 pairs of populations, we examined the relationships of the maximum (over all SNPs) per-SNP \( F_{ST} \) and \( |\delta| \) values between population pairs (as well as the top 99.99% tail values of \( F_{ST} \) and \( |\delta| \)), versus mean pairwise \( F_{ST} \). A best fit curve between the extreme \( F_{ST} \) (or \( |\delta| \)) values and mean \( F_{ST} \) was drawn using the lowess function in R (http://www.r-project.org, as in Coop et al. [52]).

We simulated the expected relationship under neutrality between the top 99.99% tail of \( |\delta| \) values and mean \( F_{ST} \) using the “beta-binomial” method of Balding [11] (also used in Coop et al. [52]). The model assumes that a diverged pair of populations were once part of a common ancestral population, with the ancestral frequency of some allele equal to \( p \). In the model, the allele frequency in each population follows a beta distribution with mean \( p \) and variance \( p(1-p)F_{ST} \), where \( F_{ST} \) is that between the two populations. Within each population, the sampled allele frequencies follow a binomial distribution, with the success probability parameter determined by the value drawn from the beta distribution. Since the ancestral allele frequency of each SNP is unknown, we first sampled an ancestral allele frequency \( p \) for each SNP from a uniform distribution on \([0, 1]\) (as in Coop et al. [52]). We then simulated the sample allele frequencies in each population using the beta-binomial distribution and this value of \( p \) to calculate \( |\delta| \). We repeated this simulation of \( |\delta| \) for the number of SNPs in our dataset to obtain a simulated value of the 99.99% tail (see Chapter 4 Methods).

C.1.3 Choice of Several Haplotype Statistic Parameters

We used several haplotype statistics to search for patterns of selective sweeps: the Integrated Haplotype Score (iHS) [291], and the Cross Population Extended Haplotype Homozygosity Test (XP-EHH) [233]. For the XP-EHH, we used the HapMap CEU, HapMap YRI, HapMap MKK, and Khomani Bushmen as reference populations.

For the results presented in Chapter 4, we binned the genome into 100 kb non-overlapping windows, and assigned to each a statistic based on the per-SNP haplotype statistics within that window. For XP-EHH,
the window statistic used was the maximum XP-EHH value within the window (as in Pickrell et al. [207]), shown to be a powerful summary statistic to identify selection. For iHS, we used the proportion of SNPs in that window with $|iHS| > 2$ (as in Voight et al. [291]).

We assigned empirical p-values to windows based on their numbers of SNPs by binning windows in increments of 10 SNPs per window. A p-value for a given window $w$ was obtained as the proportion of windows within that bin with a statistic greater than or equal to the statistic of window $w$. We assessed the sensitivity of results to the size of genomic windows used, to the bin sizes (in SNPs per window) for assigning p-values to windows, and to the XP-EHH reference population. Our selection of these parameters is described in detail below.

### 100 kb vs. 200 kb windows size

We studied two sizes for genomic windows: 200 kb, following Pickrell et al. [207], and 100 kb, following Voight et al. [291] and due to the lower linkage disequilibrium (LD) in African populations.

To explore how strongly the choice of window size affected which windows appeared in the most extreme tails of the empirical distribution, we compared the top empirical windows for those obtained using each of the two window sizes. We examined iHS and XP-EHH using HGDP Europeans as a reference population (as in Pickrell et al. [207]). (Note that although this particular reference population was not used for our main results, these analyses of window size should be insensitive to the reference population used.) We performed calculations only for the #Khomani Bushmen, Sandawe, and Hadza, since genome-wide scans for selection in these populations have not been conducted previously.

We obtained an empirical p-value for each window based on the statistic value of each window and accounting for the number of SNPs within each window. For both 100 kb and 200 kb windows, p-values were obtained by binning windows in increments of 20 SNPs (as in Pickrell et al. [207]). Windows with $\geq$ 100 SNPs were combined into 1 bin, as few genomic windows contained $\geq$ 100 SNPs.

The top 1% of genomic windows were compared for each population between the method using 200 kb windows and the method using 100 kb windows. As expected, the top 1% of 100 kb windows contained approximately twice as many windows as the top 1% of 200 kb genomic windows. The proportion of 200 kb windows that did not overlap with at least one 100 kb genomic window was near 0% for XP-EHH, and was $\approx$40% for the iHS. The proportion of 100 kb windows without an overlap in at least one 200 kb window (in other words, windows not identified by the 200 kb approach) was $\approx$30-50% for XP-EHH, and $\approx$40-70% for the iHS. Results did not change greatly for XP-EHH when we examined the top 0.1% or top 5% of genomic windows. When only the top 0.1% of windows were examined for iHS, the percentage of 100 kb windows that did not overlap a 200 kb window increased to $\approx$80%.

The results from XP-EHH indicate that selection signals can be localized to 100 kb within nearly all 200 kb windows (indicated by the $\approx$0% of 200 kb windows without an overlapping 100 kb window). The results from iHS suggest that this statistic is more influenced by the choice of window size; this is likely because the window statistic is the fraction of SNPs in a window with $|iHS| > 2$ (rather than the maximum value of...
the statistic, as it is for XP-EHH). Because of the lower levels of LD typically found in African populations [103], we chose the smaller size of 100 kb.

**Binning by 20 vs. 10 SNPs per window**

Another parameter of interest is the most appropriate bin size (in numbers of SNPs) within which to calculate empirical p-values for genomic windows. As in the previous analysis, we first binned windows in increments of 20 SNPs per window, combining all windows with $\geq 100$ SNPs into one window. As an alternative, we also binned windows in increments of 10 SNPs per window; for these bins, all windows with $\geq 50$ SNPs were combined into one window, because few 100 kb genomic windows contain $\geq 50$ SNPs. Again, the only populations for which we calculated these statistics were the $\neq$Khomani Bushmen, Sandawe, and Hadza, and we used HGDP Europeans as a reference for XP-EHH (as previously).

For each population, the top 1% of genomic windows using increments of 20 SNPs per bin were compared with those using increments of 10 SNPs. The binning strategy does seem to have an effect on the empirical tail windows. For both XP-EHH and iHS calculations, $\approx 10\%$ of the top 1% genomic windows from the 20 SNP binning method did not overlap with any of the top 1% of windows obtained using the 10 SNP method, and vice versa.

With increments of 20 SNPs per window, a majority of windows were in the lowest SNP frequency bin (0-20 SNPs), while with increments of 10 SNPs the number of 100 kb windows per bin was more consistent across bins. Thus, we decreased the SNP increment of bins to 10 SNPs per bin.

**Selection of XP-EHH Reference Populations**

Finally, we selected the most appropriate populations to use as reference populations for the XP-EHH statistic. Originally, we selected from the following reference populations: (i) all HGDP European populations, pooled together (152 individuals), as in Pickrell et al. [207]; (ii) HapMap CEU Trio Parents (88 individuals); (iii) HGDP Yorubans (21 individuals); (iv) HapMap YRI Trio Parents (100 individuals). The motivation for including both a European population and one of Yoruban origin was to search for selective sweeps of different ages. With trio information, HapMap individuals have more reliable phase inferences than those of the HGDP.

As with our other analyses of haplotype statistic parameters (window size and bin size), we compared the top 1% of 100 kb genomic windows for the XP-EHH between those obtained using different reference populations. Again, we calculated the statistics only for the $\neq$Khomani Bushmen, Sandawe, and Hadza.

**HGDP Europeans vs. HapMap CEU.** We first compared the top 1% of windows for XP-EHH using HGDP Europeans as a reference population with XP-EHH using HapMap CEU as a reference population. For all populations, $\approx 20$-30% of the top 1% of windows were unique when using HapMap CEU as a reference, while $\approx 25$-35% of the top 1% were unique when using HGDP Europeans. Thus, while a majority of the windows did show consistent signals using either reference population, $\approx 25\%$ of the top windows depended on the reference population used.
HGDP Yorubans vs. HapMap YRI. When we compared the calculations using either HGDP Yorubans or HapMap YRI as reference populations, between 40-50% of windows in the top 1% were unique to the calculations using each reference population. The difference in empirical tails between calculations using each of the two Yoruban reference populations was much greater than the difference between calculations using each of the two European reference populations (described above); this could be due to differences in the quality of phase information.

HGDP Europeans vs. Yorubans. We also assessed the overlap between the top 1% of windows obtained using HGDP Europeans as a reference population versus those using HGDP Yorubans as a reference population. The overlap between the two methods was not high, ranging from 7-20% for all populations. As expected, using Europeans as a reference population resulted in different signals from those when using an African reference population.

Due to greater confidence in phasing, we selected the HapMap CEU and HapMap YRI as the European and Yoruban reference populations, respectively. We also used the HapMap Maasai (M KK) and Khomani Bushmen (KHB) as reference populations. For these two reference populations, we had no alternative population against which to compare results (as we did for the HapMap CEU vs. HGDP Europeans and HapMap YRI vs. HGDP Yorubans).

Variation of Statistics within Windows

To ensure that breaking the genome into windows (as done by Voight et al. [291] and Pickrell et al. [207]) was a valid approach, we determined whether the variance of statistics within 100 kb windows was less than the variance among randomly selected SNPs. We examined the XP-EHH statistic in the HapMap YRI using HapMap CEU as a reference (as conducted in our main study); similar results should hold for all other sampled populations and statistics.

We used two methods to assess the validity of the windowing approach. First, we performed a Kruskal-Wallis test independently for each chromosome. We assigned to each SNP a categorical variable to indicate the identity of the 100 kb window in which it was located. We then determined whether there was a significant effect of window on the XP-EHH value of each SNP. We found the window effect to be significant, with a p-value of 0 for all chromosomes.

We also used a permutation approach to assess whether the variance of statistic values among SNPs within the same window was significantly less than the variance among groups of SNPs that were not within the same genomic window. To do so, we obtained the variance of the SNP XP-EHH values within each window, and then averaged the variance over all windows (average variance = 0.0554). For a null distribution, we sampled groups of non-adjacent SNPs on the same chromosome, with replacement, to mimic windows chromosome-wide. Since each genomic window contained a different number of SNPs, we used the same distribution of numbers of SNPs per window when drawing random groups of SNPs. We then computed a variance within each random group of SNPs, and averaged the variances across all random groups to obtain one draw of the average variance under the null hypothesis. We repeated this procedure 1,000 times. Our observed average
variance within 100 kb windows was significantly less than all of the 1,000 draws of average variances under
the null hypothesis (p-value = 0).

Thus, from both analyses, it appears that breaking the genome into 100 kb windows effectively groups
together SNPs with more similar statistic values, as desired.

C.1.4 Comparison of Empirical Tail Windows

As a summary, we obtained the five genomic windows with the lowest empirical p-values in each population,
and their p-values in all other populations, for iHS, XP-EHH CEU, XP-EHH YRI, XP-EHH MKK, and XP-
EHH KHB. To do so, we first merged adjacent windows in the top 10% of each population using the program
mergeBed in the program suite BEDTools [215]. If a top window in one population overlapped with a window
in at least the top 10% of another population, we noted its p-value.

For each of iHS, XP-EHH CEU, YRI, MKK, and KHB, we performed a Mantel correlation between a
matrix of mean $F_{ST}$ over all autosomal SNPs between all pairs of populations and matrices of (i) correlations
of window statistics across all genomic windows between pairs of populations, (ii) overlap in the top 1%
empirical tail of windows between pairs of populations, and (iii) overlap in the top 0.1%. We also performed
Mantel correlations between a matrix of the correlations of window statistics across all genomic windows
between pairs of populations ((i) above), and matrices of both (iv) the overlap in the top 1% between pairs
of populations and (v) overlap in the top 0.1%. We made scatter-plots of these five comparisons for each
statistic for a visual representation of the Mantel correlations (see Supplemental Results (Section C.1)).

We also performed Mantel correlations between a matrix of the average Bantu ancestry of each population
pair (see Methods, Chapter 4), with matrices of (vi) overlap in the top 1% tail between pairs of populations,
and (vii) the overall correlation of window statistics between pairs of populations. Note that the Namibian
San are not included in this analysis, as Bantu ancestral proportions for all individuals are not available in
Henn et al. [103].

C.1.5 Neutral Coalescent Simulations

We performed neutral coalescent simulations of a three-population model and compared empirical tail win-
dows for iHS and XP-EHH between populations (see Figure C.12). We assumed an $N_e$ in each population
of 12,000, and obtained divergence times under an equilibrium model given values of $F_{ST}$ using the formula
$(1 - F_{ST}) = \left(1 - \frac{1}{2N_e}\right)^t$, where $t$ is the divergence time in generations [115]. Divergence time between
the reference population and populations 1 and 2 was 3900 generations ($F_{ST} = 0.15$); divergence times between
populations 1 and 2 were: 2529 ($F_{ST} = 0.10$), 1231 ($F_{ST} = 0.05$), and 241 ($F_{ST} = 0.01$) generations. We
assumed a uniform mutation rate of $1.5 \times 10^{-8}$ per bp per generation (see Schaffner et al. [238]), and a uni-
form recombination rate of $10^{-8}$ per bp per generation. We also assumed uniform recombination to translate
the msms results to physical and genetic positions, required as input for XP-EHH and iHS calculations.

An example command line for msms corresponding to $F_{ST} = 0.10$ between the two daughter populations
C.1. SUPPLEMENTAL METHODS

(Populations 1 and 2 in Figure C.12) is below. The simulation is for one 10-Mb segment, where population “1” is the reference population, and populations “2” and “3” are the two daughter populations. Scaling is in units of $4N_e$ generations.

```
msms -ms 60 1 -t 7200 -I 3 20 20 20 0 -ej 0.053 3 2 -ej 0.081 1 2
-r 4800 10000000
```

C.1.6 Selective Sweep Coalescent Simulations

For simulations under positive natural selection, we simulated a selective sweep in the first daughter population, and allowed the parent population and second daughter population to evolve neutrally (all parameters remained as in the neutral simulations, see above and Chapter 4). Segments of 5 Mb were simulated, since generally the signals of long haplotypes were contained within this region (the selected allele was in the center of each segment). In order to ensure that each simulated sweep was a hard sweep, we retained only the simulations for which the “origin count” of the selected allele was 1, and in which the selected allele was not lost.

We ran preliminary analyses and simulations in `msms` to determine which parameters of selection were the most appropriate. First, we chose a selection coefficient of $s = 0.2$. Smaller selection coefficients resulted in a large number of unusable simulations (with < 5% of all simulations fitting the above criteria); ≈ 5-10% of all simulations fit the above criteria for $s = 0.2$.

We examined several different timings for the start of selection of the allele. The first was 144 generations ago. For XP-EHH, this timing resulted in a very high true positive rate (see description of calculation below) for detecting selective sweeps, but for iHS, the true positive rate was much lower (see Supplemental Results (Section C.2)). This is because in nearly all simulations, the selected allele had reached fixation, resulting in lower power for iHS (see Results and Sabeti et al. [233], Voight et al. [291]). For these simulations, we simulated 80 5-Mb segments under selection, and 260 10-Mb segments under neutrality (for a total genomic length of 3000 Mb).

Through additional simulations, we found that a more recent selection time of 62.4 generations ago led to a more intermediate final frequency (10-90%) of the selected allele, allowing iHS to have greater power [233]. As a result, the true positive rate for both iHS, as well as XP-EHH, was high with this timing (see Supplemental Results (Section C.2)). For these simulations, we simulated 300 5-Mb segments (50% of the simulated “genome”) to ensure that we simulated enough genomic regions with selection; the remaining 1500 Mb were simulated under neutrality.

The selection options in `msms` that were used to simulate selection in the first daughter population (62.4 generations ago) are below (where timing is in units of $4N_e$ generations; arguments are used in conjunction with the demographic parameters, specified above, modified to simulate 5 Mb):

```
-SI 0.0013 3 0 0.00001 0 -Sp 0.5 -Sc 0 2 9600 4800 0 -N 12000 -oOC
-oTrace -Smu 0
```

As for the neutral coalescent simulations, for each of the 100 genome-wide simulations we calculated iHS
and XP-EHH, binned the genome into windows of 100 kb, and ranked windows to obtain empirical tails. To examine the performance, or true positive rate, of the iHS and XP-EHH statistics, for a given empirical tail we calculated the proportion of its 100 kb windows that were within a selective sweep (non-neutral) segment. This rough measure indicates the ability of the statistics to detect signals of selection in 100 kb windows that are within 2.5 Mb of a selected site (since the selected site was located in the center of each simulated 5-Mb region). Depending on the exact lengths of haplotype signals produced from the sweeps, this may be an overestimate of the true positive rate, and an underestimate of the false positive rate.

C.1.7 Gene Ontology Enrichment

Replication of Voight et al. 2006 Results

We followed the procedure of Voight et al. [291], who searched for Gene Ontology (GO) term enrichment in the HapMap YRI for the iHS statistic. First, we obtained the 50 closest SNPs centered around each gene (obtained from the UCSC refFlat mappings of RefSeq genes to hg18; downloaded in May 2010, http://www.genome.ucsc.edu). If there were more than 50 SNPs located within a gene, we kept all of the SNPs. Then, we assigned to each gene a summary statistic, calculated as the proportion of its 50 SNPs with normalized $|iHS| > 2$. We ranked the genes in order of their summary statistics, and examined enrichments in the top 10% of genes. We assigned GO terms to genes using the biomaRt package in R (http://www.r-project.org; described in more detail below). We also pruned terms to include only those present in the PANTHER database (http://www.pantherdb.org), as Voight et al. [291] use this dataset. Note that in our analysis, we examined $\approx 500,000$ SNPs; this is only a subset of the SNPs analyzed by Voight et al. [291].

Then, we assessed enrichments for each PANTHER biological process GO term using the hypergeometric distribution. The hypergeometric distribution for a random variable $X$ can be written as:

$$P(X = k) = \binom{r}{k} \binom{n-r}{m-k} \binom{n}{m}^{-1}.$$  

(C.1)

Here, for a particular GO term, $k$ is the number of genes in the top 10% of genes with that GO term, $r$ is the number of genes annotated with the GO term in the genome overall, $n$ is the total number of genes in the genome, and $m$ is the number of genes in the top empirical 10%. To calculate a p-value for a particular GO term, given its count of occurrences in the top empirical 10% of genes ($k$), we calculate $P(X \geq k)$ by summing $P(X = i)$ for $i \in [k, r]$ if $r \leq m$ and $i \in [k, m]$ if $m \leq r$.

Haplotype Statistics

We explain our novel permutation approach for assessing significance of GO terms within populations in more detail, as in Figure 4.1 of Chapter 4. The method is particularly designed for analyses of genomic windows, within which several genes can exist. Significance is assessed for the top $x\%$ of 100 kb genomic
windows for a particular haplotype statistic (iHS or XP-EHH); in our analyses, we examine the top 0.1%, 1%, and 5%. (We focus on the top 0.1% in the Chapter 4.)

1. **Observed empirical tails.** For the top \( x \% \) of genomic windows, we determined the GO terms associated with the genes in each genomic window. (The `biomaRt` package in R was used to perform these lookups, using “ensembl_mart_51” to ensure that the genomic coordinates (Build 36) were consistent). Adjacent genomic windows were merged, so that a signal > 100 kb would have each of its GO terms counted only once. (This helps to reduce possible dependence between 100 kb windows.) For each unique GO term, we counted its number of occurrences.

2. **Genome-wide windows.** We determined the associated GO terms for each genomic window containing at least one SNP with a statistic value (using `biomaRt`, as above). (Some SNPs were removed in the XP-EHH and iHS calculations due to certain criteria. For instance, for iHS, the minor allele frequency of each SNP must be \( \geq 5\% \); Voight et al. [291] also stated, “if the region spanned by EHH > 0.05 reached a chromosome end or the start of a gap > 200 kb, then no iHS value was reported for the core SNP.” Similarly for XP-EHH, Sabeti et al. [233] stated, “if there is no SNP with such an EHH between 0.03 and 0.05, the XP-EHH test is skipped.” Thus, not all genotyped SNPs resulted in a calculated XP-EHH or iHS value.)

   We then binned the above windows by the number of SNPs per window, as done when calculating p-values for the observed data (described in section “Choice of Several Haplotype Statistic Parameters”).

3. **Randomization step.** In each SNP count bin, we randomly selected \( x \% \) of windows, and combined the top windows from all bins to create a “random top \( x \% \)” set of windows. Then, for each unique GO term, we recorded its number of occurrences across all of the windows in the “random top \( x \% \).”

   We repeated this randomization step 20,000 times, recording for each GO term its number of occurrences in each of the randomizations to form the null distribution.

4. **P value calculation.** For each GO term \( G \), we calculated a p-value for over-representation of \( G \) in the observed \( x \% \) tail as the proportion of the 20,000 permutations for which the number of occurrences of \( G \) is greater than or equal to the observed number of occurrences.

5. **Multiple hypothesis testing correction.** To account for multiple hypothesis testing, we controlled the false discovery rate (FDR) at a 5% level using the procedure of Benjamini and Hochberg [26]. This is less conservative than the Bonferroni correction, but more conservative than using no multiple hypothesis testing correction.

   In some cases, we analyzed only a subset of all GO terms that are found in the PANTHER database, as in Voight et al. [291]. This only affects our assignment of GO terms to windows in steps 1 and 2 above.
**Top $F_{ST}$ Values**

To assess whether any GO terms were enriched among SNPs with the most extreme values of $F_{ST}$ (calculated across all 11 populations), we first followed the approach described previously for genomic windows. For each SNP, we found the gene, if any, in which it was located, and obtained the GO terms associated with that gene. We then randomly selected $x\%$ of all SNPs, and counted the occurrence of each GO term; we repeated this 20,000 times to obtain a null distribution for the counts of each GO term. We obtained p-values for observed counts of GO terms from this null distribution with an FDR correction. To be more conservative when identifying over-represented terms among top genomic windows, we restricted our results to terms associated with extreme $F_{ST}$ SNPs occurring in different, not the same, genes.

In addition, we searched for significant GO terms with the program GREAT [169], which uses a binomial test to assess significance among a foreground set of genomic regions (in our case, SNPs with large $F_{ST}$ values). We used the SNPs genotyped in our dataset as a background set of genomic regions to control for non-uniform SNP density and ascertainment. Using the parameters described below, GREAT first infers gene regulatory domains, and then the GO terms associated with a given genomic region based on the domains it overlaps. It calculates the percentage of the genome annotated with each GO term; then, it counts the SNPs in the tails that are associated. A binomial test over the genomic regions is performed to obtain a p-value for each GO term, using an FDR correction (see McLean et al. [169] for more details). (This program was not used for 100 kb genomic windows, as the genomic regions of GREAT are required to be smaller).

GREAT allows for a number of “association rule” parameters, which set the boundaries of gene regulatory regions. We first used the default settings of the “basal plus extension” rule, allowing each gene to be associated with a basal regulatory domain a minimum distance of 5 kb upstream and 1 kb downstream of the transcription start site (TSS). This domain is extended both upstream and downstream until the nearest gene’s basal domain, and no more than 1000 kb away from the gene; any SNPs within this domain are associated with the gene. We then allowed the domain to be extended no more than 50 kb or 0 kb away (instead of 1000 kb).

**C.2 Supplemental Results**

**C.2.1 Population Differentiation**

Pairwise $F_{ST}$ estimates averaged over all autosomal SNPs agreed with those of Henn et al. [103], who calculated $F_{ST}$ between inferred ancestral populations (Table C.2). On average, our estimates were slightly lower, as we included admixed individuals with ancestry in several ancestral populations.

We calculated for each SNP the absolute values of derived allele frequency difference ($|\delta|$) between all (55, or $\binom{11}{2}$) pairs of populations, following Coop et al. [52]. Not all population pairs exhibited high $|\delta|$ values (Figures C.1A and C). For some pairs of populations, there was an enrichment of genic SNPs at high $|\delta|$ values, but for others, there was no enrichment (Figure C.1A). Among the population pairs for
which which genic SNPs were enriched among the highest $|\delta|$ values for the pair were the Kenya Bantu and Maasai; the South African Bantu and the Khomani Bushmen and Sandawe; the Hadza and the Maasai; the Mbuti and the Hadza, Maasai, Mandenka, Namibian San, Khomani Bushmen, Sandawe, and Yoruba; and the Khomani Bushmen and Yoruba. Interestingly, all population pairs had SNPs with allele frequency differences in the smallest bin ($0 < |\delta| < 0.1$), with an apparent raw genic enrichment of SNPs for all (Figure C.1C). Though this enrichment of genic SNPs may not be significant in all populations, it may indicate the action of purifying selection. When enrichment values in each $|\delta|$ value bin (in Figure C.1A) were averaged over all population pairs, there was only a slight increase in enrichment of genic SNPs at extreme values (Figure C.1B).

We examined several of the population pairs individually for evidence of genic enrichment using a bootstrapping procedure to assign significance (mentioned in the Supplemental Methods (Section C.1) and Chapter 4) and present results in Figure C.2. For individual population pairs, we examined the non-absolute value of $\delta$. In most cases, the SNPs in the most extreme $\delta$ value bins were indeed genic – but there were very few ($<5$) SNPs. In the bootstrapping procedure, we found that this small number of SNPs with extreme $\delta$ values was problematic. In general, less than 95% of the 1,000 bootstrap simulations actually contained a SNP with a value of $\delta$ in the highest (or lowest) bin (Figure C.2). Thus, the actual existence of SNPs in this bin was not significant, and so we could not assign significance to genic enrichment in these bins. (In Figure C.2C and D, while $>95\%$ bootstrap samples were valid, enrichment was not significant). Interestingly, the next-most-extreme $\delta$ bin was never significant according to the 95% bootstrap confidence intervals. While it may be interesting that a majority of the most strongly differentiated SNPs were genic (and that the few bootstrap samples that actually contained these SNPs showed an enrichment), we cannot state with statistical certainty that there is evidence of genic enrichment of strongly differentiated SNPs among pairs of African populations.

We also examined the SNPs with $|\delta| > 0.8$ in at least one population pair. In most cases, we found that these SNPs were primarily due to one allele being completely fixed or absent from the Mbuti Pygmies or Hadza. However, these high-$|\delta|$ SNPs between the two populations did not show a genic enrichment (Figure C.2D). The lack of genic enrichment among the many highly differentiated SNPs is likely due to the fact that they have both undergone severe bottlenecks in recent history [103].

We also analyzed population differentiation between the HGDP Tuscans and Yorubans (Figure C.3). Here, while we found that the bootstrap confidence intervals were valid, genic enrichment was not significant for either population. See Chapter 4 for discussion of these results.

Although the analyses here (see also Figure C.4) and in Chapter 4 (Figure 4.2 and 4.3) indicated that allele frequency differentiation is potentially strongly affected by population history, there were several interesting genes within which several SNPs lied in the empirical top 1% of all $F_{ST}$ values. Of these, several SNPs were within the lactase gene ($LCT$), involved in lactase persistence, in several genes known to be involved in skin pigmentation ($HERC2, HPS2, KIT, MITF, NF1, OCA2, PTGER3$), and in other genes known to be
involved in disease resistance (HLA, LARGE, TLR5, TLR6). However, as differentiation of these SNPs may not necessarily be due to positive selection (see Chapter 4 and above), we did not explore these signals further.

C.2.2 Haplotype Statistics

Overlap of Windows Between Populations

There was considerable overlap between populations of top windows for the XP-EHH CEU statistic (Figure C.6), but less overlap for iHS (Figure C.5), and even less for XP-EHH MKK, YRI, and KHB (Chapter 4 Figure 4.4 and Figures C.7 and C.8). Note that our results for the iHS and XP-EHH CEU were slightly different from those of Pickrell et al. [207] and Henn et al. [103], as they examined genomic windows of 200 kb.

We examined the relationship between mean $F_{ST}$ and extent of sharing of the top 0.1% of windows between pairs of populations (see Supplemental Methods (Section C.1)). We saw results similar to the top 1% (see Chapter 4 Results) for the XP-EHH CEU, XP-EHH MKK, and XP-EHH KHB, although the relationship between these two variables appeared weaker than for the top 1% of windows (Figure C.9). Mantel correlations remained significant for these statistics, but not for the iHS or XP-EHH YRI (Table C.3). (Since the maximum number of windows shared in the top 0.1% for any two populations was 1 for the iHS and XP-EHH YRI, there may be little power to detect any relationship.)

For each of iHS, XP-EHH CEU, YRI, MKK, and KHB, we also computed a correlation of the selection statistics per window across all genomic windows between all pairs of populations (denoted “correlation of window statistics”). For all five statistics, including the XP-EHH YRI, the number of windows in the top 1% of both populations of a pair increased as the correlation of window statistics for that pair increased, with a significant Mantel correlation between these two variables (Table C.3, Figure C.10). For the top 0.1% of genomic windows, XP-EHH CEU, XP-EHH MKK, and XP-EHH KHB still exhibited significant relationships between the correlation of window statistics and the number of shared windows in the top 0.1%, though to a lesser degree than for the top 1% (Table C.3). For the iHS and XP-EHH YRI, no more than 1 window was shared between any pair of populations, again resulting in non-significant Mantel correlations (Table C.3).

We also found that more closely related population pairs (those with lower mean $F_{ST}$) had more highly correlated window statistic values for iHS, XP-EHH CEU, XP-EHH MKK, and XP-EHH KHB (all except for XP-EHH YRI). Mantel tests for these four statistics indicated a significant correlation between the two variables (Table C.3). This negative trend between $F_{ST}$ and correlations of window statistics suggests that values of these selection statistics in the genome as a whole correspond closely with population divergence.

However, the XP-EHH YRI statistic appeared to behave in a fundamentally different manner from the other statistics. We observed no relationship between $F_{ST}$ and the correlations of window statistics between pairs of populations, nor between $F_{ST}$ and the extent of overlap in empirical tails. Since the empirical tails did seem to be related to overall correlations of window statistics, however, the different behavior of
XP-EHH YRI may not necessarily indicate that between-population patterns are driven by shared selective pressures. Interestingly, we found that the correlations of window statistics across the whole genome appeared to be related to the average proportion of Bantu ancestry. The correlation of window statistics between pairs of populations for XP-EHH YRI decreased as the average proportion of inferred Bantu ancestry for each population in the pair increased (p-value = 0.009; see Figure C.11 and Table C.3). We found a similar result for sharing in the 1% empirical tail (Table C.3). The greater amount of sharing of windows between populations with low amounts of Bantu ancestry appears to be an artifact of using a population with a large amount of Bantu ancestry as a reference.

Further Examination of Shared Windows

Population history appeared to be difficult to separate from the extent of sharing of candidate signals (see previous section). However, positive selection on the same genomic regions might be able to explain an increased number of shared windows beyond a “baseline” expected from shared ancestry. We examined the pairs of populations that appeared to have a greater than expected number of windows shared in their top 0.1% when compared to other population pairs with similar values of $F_{ST}$ (Figure C.9).

The Mandenka and Yoruba were an outlying population pair for the XP-EHH MKK, with five common windows in their top 0.1%. Of these, two adjacent regions on chromosome 11 (9,800,000-10,100,000) were outliers only in the Mandenka and Yoruba and the South African Bantu. This region contains the gene $SBF2$, mutations in which are known to be associated with Charcot-Marie-Tooth disease (characterized by abnormal myelin sheaths) [309]. For the XP-EHH YRI, the Sandawe, Hadza, and Maasai all shared a region on chromosome 2 (136,300,000-136,400,000) containing the genes $MCM6$, $LCT$ and $DARS$. This region was also found in at least the top empirical 5% of the Mbuti and Biaka Pygmies, as well as the Khomani Bushmen. Because this region appeared to be significant in hunter-gatherer, pastoralist, and agricultural populations, sharing may not be due to convergent adaptation of lactase persistence but by another unknown selective pressure or by a sweep at another gene, as suggested by Tishkoff et al. [276]. A region shared in the top 0.1% of the Kenyan Bantu and Maasai for XP-EHH YRI was on chromosome 2 (38,800,000-38,900,000), and was otherwise found only in the top 5% tail of the Mandenka. Among other genes, this region contains $GALM$, involved in galactose metabolism. Thus, a potentially diet-related trait could be under selection in all three of these agricultural/pastoralist populations (Kenyan Bantu, Maasai, Mandenka).

There were several outlying population pairs for the XP-EHH MKK (Figure C.9). For the Mandenka and Yoruba, among the five shared windows in their top 0.1% was a window that was not in the top empirical 5% of any other population except the Kenyan Bantu; however, it contains no known genes (chromosome 13: 56,900,000-57,000,000). Two adjacent windows on chromosome 3 (56,500,000-56,700,000) were not in the top empirical 5% of any other tested population. This region contains the genes $CCDC66$ and $C3orf13$, whose functions are unknown.

For iHS, one pair with a shared window was the Sandawe and Maasai; this window (chromosome 11: 84,600,000-84,700,000) was not an outlier in any other population and contains $DLG2$, potentially involved
in synapse function. The Homani Bushmen and Maasai shared an empirical outlying region for XP-EHH YRI, found only also in the empirical tail windows of the Hadza. This region on chromosome 2 (196,900,000-197,000,000) contains the gene HECW2, possibly related to ubiquitination.

**Coalescent Simulations**

For data simulated under the neutral coalescent, in nearly all empirical tails for XP-EHH, there appeared to be a strong relationship between the $F_{ST}$ of the two daughter populations and the extent of sharing of windows in their empirical tails (Figure C.13). For iHS, this relationship was less strong, though still apparent (Figure C.14).

For data simulated under selection for XP-EHH, there was generally no relationship between the number of overlapping windows and $F_{ST}$, regardless of timing of selection (Figures C.15 and C.17). In Figure C.17, where selection is less recent, there was a more apparent relationship between the two quantities (especially for the 0.001 tail). While our calculated true positive rate was still 100%, this observation may be due to the fact that our estimate of the true positive rate is an overestimate (see Supplemental Methods (Section C.1)). This observation may also be due to the inadequate number of regions simulated for this selection timing (80 5-Mb regions, versus 300 5-Mb regions for the more recent selective sweeps in Figure C.15).

For data simulated under selection for iHS, there was a more apparent relationship of the number of overlapping tail windows with $F_{ST}$ for older selective sweeps (Figure C.18) than for recent sweeps (Figure C.16). The true positive rate for iHS was nearly 100% for more recent sweeps (62.4 generations ago), where the selected alleles had not yet swept to fixation. However, for older sweeps (144 generations ago), the selected allele had generally risen to fixation, resulting in a drop of power [233, 291], and decrease of the true positive for iHS to as low as 60%. This caused the empirical tails to generally contain more neutral false positive genomic windows. This existence of false positive signals in empirical tails likely explains the apparent relationship between extent of sharing of empirical tail windows with a population pairs' $F_{ST}$ (Figure C.18).

**C.2.3 Gene Ontology Enrichments**

**Replication of Voight et al. 2006 Results**

We replicated the procedure of Voight et al. [291] to assess enrichment of Gene Ontology (GO) terms in the top 10% of genes according to iHS calculations in the HapMap YRI. Voight et al. [291] reported enrichment for steroid metabolism (GO:0008202), MHC-I mediated immunity (GO:0032393 and GO:0032394), chemosensory perception (GO:0007606), olfaction (GO:0007608), and peroxisome transport (GO:0043574). Repeating their analysis with their methodology, we also found enrichments for MHC-I mediated immunity (p-value = 0.002) and olfaction (p-value = 0.02625). Although we did not find an enrichment for chemosensory perception, this is likely due to a change in the GO database (http://amigo.geneontology.org, accessed June 2011): surprisingly, we found very few genes in the GO database associated with
chemosensory perception. We also found a higher p-value for steroid metabolism than do Voight et al. [291] (p-value = 0.6684). This is again likely due to the GO database: while we found many genes associated with the term when we look up the term in the database, the term was not associated with any of them when the genes were looked up in the database (as done for our analysis). Finally, the term peroxisome transport was not in the PANTHER database, though Voight et al. [291] reported it to be.

Thus, most differences in results when replicating the method of Voight et al. [291] were likely due to changes in the GO database since their study. In addition, we examined only a subset of the SNPs analyzed by Voight et al. [291], resulting in slightly different p-values. While we found several other significant terms when using the method of Voight et al. [291], none of them were significant when using our more rigorous permutation method (described in the Supplemental Methods (Section C.1) and in Chapter 4 Methods and Results).

**Haplotype Statistics**

Using our permutation method, we found that all haplotype statistics (iHS, and XP-EHH CEU, YRI, MKK and KHB) exhibited at least some signatures of GO term enrichment. In addition, GO terms appeared to be enriched in all empirical tails for the statistics (the top 0.1%, 1%, and 5%) (Table C.4 and Chapter 4 Tables 4.1 and 4.2). However, terms that appeared enriched in a less extreme tail (i.e. top 1%) were not necessarily enriched in a more extreme tail (i.e. top 0.1%), potentially complicating the choice of the most appropriate cutoff. Our neutral simulations also highlight that if an improper empirical tail is analyzed, attributing sharing of signals between populations to positive selection could be largely inaccurate (see also Figures C.13 and C.14). In Chapter 4 (Tables 4.1 and 4.2), we present results from the top 0.1%, motivated by the results of Table C.3.

**F\textsubscript{ST} Values**

We performed a similar Gene Ontology (GO) term enrichment analysis for SNPs with the top empirical values of F\textsubscript{ST} (calculated across all 11 populations). We first used a permutation method, analogous to the one performed for genomic windows and haplotype statistics, but found no significant results (see Supplemental Methods (Section C.1)). We also used the program GREAT [169] to search for enrichment of terms associated with SNPs with the largest values of F\textsubscript{ST}. Our results varied depending on the “extension” rules determining the association of a SNP with a particular gene (see Supplemental Methods (Section C.1)). When we allowed the domain associated with a gene to extend up to 1000 kb away beyond the basal domain, we found five significant GO terms with multiple gene associations: regulation of saliva secretion (GO:0046877), regulation of ATPase activity (GO:0043462), positive regulation of ATPase activity (GO:0032781), tricarboxylic acid cycle (GO:0006099), and coenzyme catabolic process (GO:0009109). However, when we allowed an extension of only 50 kb, none of these terms were significant, and instead we found associations of a different nature: spermatogenesis (GO:0007283), regulation of estrogen receptor signaling pathway (GO:0033146), chromatin modification (GO:0016568), regulation of cytokine biosynthetic process (GO:0042089), gamete
generation (GO:0007276), muscle filament sliding (GO:0030049), atrial cardiac muscle tissue morphogenesis (GO:0055009), and male sex determination (GO:0030238). When we required that a SNP be located within a gene in order to be associated with it (0 kb extension) we found no enriched GO terms, consistent with our permutation analysis.

Since the GREAT analysis depended heavily on certain parameters, and our permutation procedure found no evidence of over-representation for any GO terms, there did not appear to be any particular functions associated with SNPs that were strongly differentiated between African populations (as measured by $F_{ST}$). This agrees with Figures 4.2 and 4.3 of Chapter 4 and Figures 4.2, C.2, and C.4, which suggest that selection is not a primary driver of extreme allele frequency differentiation between African populations for the SNPs in our dataset.
C.3 Supplemental Figures and Tables

Figure C.1: **Enrichment of genic and non-genic SNPs for values of $|\delta|$ between pairs of populations.** Enrichments of genic and nongenic SNPs are calculated within each $|\delta|$ value bin (width 0.1). (A) Enrichments calculated separately for each pair of populations among the 11 populations (55 pairs in total). Each line indicates a unique population pair. (B) Enrichments of genic and non-genic SNPs averaged over population pairs – i.e., an average of the lines in panel (A), for genic and non-genic enrichments separately. (C) Summary of panel (A). Black line indicates the number of population pairs with a defined value of $|\delta|$ for each bin (i.e., the number of population pairs where there exists at least one SNP in the given $|\delta|$ bin). Red line indicates the number of population pairs where there appears to be an enrichment of genic SNPs (enrichment value > 1) in the given $|\delta|$ bin.
Figure C.2: Enrichment of genic and non-genic SNPs for values of $\delta$ between selected pairs of populations. Enrichments of genic and nongenic SNPs are calculated within each $\delta$ value bin, as in Figure 4.2 and Chapter 4 (width 0.1). Bootstrap confidence intervals are indicated by dotted lines (see Chapter 4). SNPs where the derived allele is fixed in the first population and absent in the second are depicted on the right side of the plot (near 1); SNPs where the derived allele is fixed in the second population and absent in the first are depicted on the left side of the plot (near -1). The number above the most extreme $\delta$ bins indicates the proportion of bootstrap simulations where an enrichment value was defined for the given bin (see Results (Chapter 4)); if no value is listed, the proportion is 1. Values less than 0.95 can be considered insignificant. (A) Kenyan Bantu and Maasai. (B) $\neq$Khomani Bushmen and Yoruba. (C) Mbuti and Namibian San. (D) Hadza and Mbuti Pygmies.
C.3. SUPPLEMENTAL FIGURES AND TABLES

Figure C.3: Enrichment of genic and non-genic SNPs for values of $\delta$ between the HGDP Yoruba and HapMap Tuscans. Enrichments of genic and nongenic SNPs are calculated within each $\delta$ value bin, with bootstrap confidence intervals, as in Figures 4.2 and C.2 and Chapter 4. SNPs where the derived allele is fixed in the Yorubans and absent in the Tuscans are depicted on the right hand side of the plot (near 1); SNPs where the derived allele is fixed in the Tuscans and absent in the Yorubans are depicted on the left hand side of the plot (near -1).
Figure C.4: Extreme values of SNP differentiation over all SNPs vs. mean $F_{ST}$ for all pairs of populations. Each point is a population pair; there are 55 pairs in total. Line is the best-fit lowess curve through the points. Top panel Y-axis: maximum value over all SNPs of derived allele frequency difference ($|\delta|$) between pairs of populations; Middle panel Y-axis: maximum value over all SNPs of $F_{ST}$ between pairs of populations; Bottom panel Y-axis: upper 99.99% tail cutoff value over all SNPs of $F_{ST}$ between pairs of populations.
### C.3. Supplemental Figures and Tables

#### Figure C.5: P-values for the 5 most extreme 100 kb genomic windows according to the iHS statistic in each population. Red indicates $p < 0.01$, orange indicates $p < 0.05$, yellow indicates $p < 0.10$, and white indicates $p > 0.10$. “Genes” column lists the genes located within the indicated windows.
Figure C.6: P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH CEU statistic in each population. Red indicates $p < 0.01$, orange indicates $p < 0.05$, yellow indicates $p < 0.10$, and white indicates $p > 0.10$. “Genes” column lists the genes located within the indicated windows.
Figure C.7: P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH YRI statistic in each population. Red indicates $p < 0.01$, orange indicates $p < 0.05$, yellow indicates $p < 0.10$, and white indicates $p > 0.10$. “Genes” column lists the genes located within the indicated windows.
Figure C.8: P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH KHB statistic in each population. Red indicates $p < 0.01$, orange indicates $p < 0.05$, yellow indicates $p < 0.10$, and white indicates $p > 0.10$. “Genes” column lists the genes located within the indicated windows.
Figure C.9: Number of shared 100 kb genomic windows in the top empirical 0.1% vs. mean $F_{ST}$ for pairs of populations. Each point is a different population pair (some points are labeled, see key). Titles indicate the haplotype statistics for which the shared windows occur. Line is a best-fit lowess curve through the points. Significance of the p-value for the Mantel correlation between the $x$ and $y$ variables is indicated in the upper right corner: “NS”: not significant, “*”: $< 0.05$, “**”: $< 0.01$, “***”: $< 0.001$. 

iHS

XP–EHH CEU

XP–EHH YRI

XP–EHH MKK

XP–EHH KHB

BPY – Biaka Pygmy
MPY – Mbuti Pygmy
HDZ – Hadza
SWE – Sandawe
SNB – Namibian San
KHB – Khomani Bushmen
BSA – Bantu South Africa
BKN – Bantu Kenya
MKK – Maasai
MDK – Mandinka
YOR – Yoruba
Figure C.10: Number of shared 100 kb genomic windows in the top empirical 1% vs. correlations of window statistics over all genomic windows for pairs of populations. Each point is a different population pair. Titles indicate the haplotype statistics for which the shared windows occur. Line is a best-fit lowess curve through the points. Significance of the p-value for the Mantel correlation between the x and y variables is indicated in the upper left corner: “NS”: not significant, “*”: < 0.05, “**”: < 0.01, “***”: < 0.001.
C.3. SUPPLEMENTAL FIGURES AND TABLES

Figure C.11: Correlations of window statistics over all 100 kb genomic windows vs. average Bantu ancestry for pairs of populations. Average proportion of Bantu ancestry is inferred from Henn et al. [103], see Methods (Chapter 4). Each point is a different population pair. Titles indicate the haplotype statistics for which the correlations are calculated. Line is a best-fit lowess curve through the points. Significance of the p-value for the Mantel correlation between the x and y variables is indicated in the upper right or left corner: “NS”: not significant, “*”: < 0.05, “**”: < 0.01, “***”: < 0.001. Note that the Namibian San are not included, as Bantu ancestral proportions for all individuals are not available from Henn et al. [103].
Figure C.12: **Demographic model simulated using msms.** $F_{ST}$ values determine the divergence times in the model; see Chapter 4.
Figure C.13: **Number of overlapping 100 kb windows in the empirical tails of XP-EHH for pairs of populations simulated under neutrality.** Means +/- standard errors over 100 simulations are shown for each value of $F_{ST}$ simulated between Populations 1 and 2 (see Figure C.12). The empirical tail examined in Populations 1 and 2 is shown above each plot.
Figure C.14: Number of overlapping 100 kb windows in the empirical tails of iHS for pairs of populations simulated under neutrality. Means +/- standard errors over 100 simulations are shown for each value of $F_{ST}$ simulated between Populations 1 and 2 (see Figure C.12). The empirical tail examined in Populations 1 and 2 is shown above each plot.
Figure C.15: **Number of overlapping 100 kb windows and true positive rate in the empirical tails of XP-EHH for pairs of populations simulated under recent selection.** Strength of selection $s = 20\%$, time of selection equal to 62 generations ago. The empirical tail examined is shown above each column, and the x-axes indicate the value of $F_{ST}$ simulated between Populations 1 and 2 (see Figure C.12). Top row y-axis: number of windows overlapping between the empirical tails of populations 1 and 2 (means +/- standard errors over 100 simulations). Bottom row y-axis: as an estimate of true positive rate, the proportion of 100 kb windows in the empirical tails of population 1 that fell within one of the 5 Mb regions in which a selective sweep was simulated (+/- standard deviation over 100 simulations).
Figure C.16: **Number of overlapping 100 kb windows and true positive rate in the empirical tails of iHS for pairs of populations simulated under recent selection.** Strength of selection $s = 20\%$, time of selection equal to 62 generations ago. The empirical tail examined is shown above each column, and the x-axes indicate the value of $F_{ST}$ simulated between Populations 1 and 2 (see Figure C.12). Top row y-axis: number of overlapping 100 kb windows; bottom row y-axis: true positive rate. See caption for Figure C.15.
Figure C.17: **Number of overlapping 100 kb windows and true positive rate in the empirical tails of XP-EHH for pairs of populations simulated under less recent selection.** Strength of selection $s = 20\%$, time of selection equal to 144 generations ago. The empirical tail examined is shown above each column, and the x-axes indicate the value of $F_{ST}$ simulated between Populations 1 and 2 (see Figure C.12). Top row y-axis: number of overlapping 100 kb windows; bottom row y-axis: true positive rate. See caption for Figure C.15.
Figure C.18: Number of overlapping 100 kb windows and true positive rate in the empirical tails of iHS for pairs of populations simulated under less recent selection. Strength of selection $s = 20\%$, time of selection equal to 144 generations ago. The empirical tail examined is shown above each column, and the x-axes indicate the value of $F_{ST}$ simulated between Populations 1 and 2 (see Figure C.12). Top row y-axis: number of overlapping 100 kb windows; bottom row y-axis: true positive rate. See caption for Figure C.15.
Table C.1: Populations analyzed.

<table>
<thead>
<tr>
<th>Population</th>
<th>Source</th>
<th>Phased?</th>
<th>(F_{ST})</th>
<th>iHS</th>
<th>XP-EHH CEU</th>
<th>XP-EHH YRI</th>
<th>XP-EHH MKK</th>
<th>XP-EHH KHB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kenyan Bantu</td>
<td>HGDP</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>S. African Bantu</td>
<td>HGDP</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Biaka</td>
<td>HGDP</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hadza</td>
<td>Henn et al. [103]</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Maasai</td>
<td>HapMap</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>REF</td>
<td>X</td>
</tr>
<tr>
<td>Mandenka</td>
<td>HGDP</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Mbuti</td>
<td>HGDP</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Namibian Bushmen</td>
<td>HGDP &amp; Schuster et al. [242]</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>≠ Khomani Bushmen</td>
<td>Henn et al. [103]</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>REF</td>
</tr>
<tr>
<td>Sandawe</td>
<td>Henn et al. [103]</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Yoruba</td>
<td>HGDP</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>REF</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>YRI</td>
<td>HapMap</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>REF</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>CEU</td>
<td>HapMap</td>
<td>Yes</td>
<td>REF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuscans</td>
<td>HapMap</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Europeans</td>
<td>HGDP</td>
<td>Yes</td>
<td>REF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

"Phased?" column indicates whether phased data were used (for haplotype analyses). An “X” indicates whether the population was used in the analyses of \(F_{ST}\) (and allele frequency differentiation), iHS, XP-EHH CEU, XP-EHH YRI, XP-EHH MKK, or XP-EHH KHB. “REF” indicates that population was used as a reference population. HGDP Europeans and HGDP Yorubans were only used as test reference populations for XP-EHH (see Supplemental Methods (Section C.1)).
Table C.2: Mean $F_{ST}$ values between pairs of populations.

<table>
<thead>
<tr>
<th>Population 1</th>
<th>0.128</th>
<th>0.108</th>
<th>0.072</th>
<th>0.066</th>
<th>0.059</th>
<th>0.060</th>
<th>0.118</th>
<th>0.127</th>
<th>0.135</th>
<th>0.133</th>
<th>0.138</th>
<th>0.143</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuscan</td>
<td>0.063</td>
<td>0.043</td>
<td>0.037</td>
<td>0.033</td>
<td>0.030</td>
<td>0.031</td>
<td>0.057</td>
<td>0.068</td>
<td>0.078</td>
<td>0.080</td>
<td>0.081</td>
<td>0.081</td>
</tr>
<tr>
<td>Yoruba</td>
<td>0.047</td>
<td>0.031</td>
<td>0.037</td>
<td>0.043</td>
<td>0.043</td>
<td>0.044</td>
<td>0.056</td>
<td>0.066</td>
<td>0.071</td>
<td>0.073</td>
<td>0.071</td>
<td>0.074</td>
</tr>
<tr>
<td>Namibian</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Bushmen</td>
<td>0.072</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Sandawe</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Maasai</td>
<td>0.072</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Hadza</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Tuscan</td>
<td>0.063</td>
<td>0.043</td>
<td>0.037</td>
<td>0.033</td>
<td>0.030</td>
<td>0.031</td>
<td>0.057</td>
<td>0.068</td>
<td>0.078</td>
<td>0.080</td>
<td>0.081</td>
<td>0.081</td>
</tr>
<tr>
<td>Yoruba</td>
<td>0.047</td>
<td>0.031</td>
<td>0.037</td>
<td>0.043</td>
<td>0.043</td>
<td>0.044</td>
<td>0.056</td>
<td>0.066</td>
<td>0.071</td>
<td>0.073</td>
<td>0.071</td>
<td>0.074</td>
</tr>
<tr>
<td>Namibian</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Bushmen</td>
<td>0.072</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Sandawe</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Maasai</td>
<td>0.072</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Hadza</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Tuscan</td>
<td>0.063</td>
<td>0.043</td>
<td>0.037</td>
<td>0.033</td>
<td>0.030</td>
<td>0.031</td>
<td>0.057</td>
<td>0.068</td>
<td>0.078</td>
<td>0.080</td>
<td>0.081</td>
<td>0.081</td>
</tr>
<tr>
<td>Yoruba</td>
<td>0.047</td>
<td>0.031</td>
<td>0.037</td>
<td>0.043</td>
<td>0.043</td>
<td>0.044</td>
<td>0.056</td>
<td>0.066</td>
<td>0.071</td>
<td>0.073</td>
<td>0.071</td>
<td>0.074</td>
</tr>
<tr>
<td>Namibian</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Bushmen</td>
<td>0.072</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Sandawe</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Maasai</td>
<td>0.072</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
<tr>
<td>Hadza</td>
<td>0.071</td>
<td>0.067</td>
<td>0.068</td>
<td>0.069</td>
<td>0.071</td>
<td>0.073</td>
<td>0.076</td>
<td>0.078</td>
<td>0.079</td>
<td>0.080</td>
<td>0.081</td>
<td>0.083</td>
</tr>
</tbody>
</table>

Calculated over all autosomal SNPs. All populations are included except for the HapMap YRI, since the HGDP Yorubans are examined instead.
### Table C.3: Mantel correlations between matrices of the indicated variables for each haplotype statistic.

<table>
<thead>
<tr>
<th>Haplotype Statistic</th>
<th>$F_{ST}$ and Correlation of Window Statistics and Average Bantu Ancestry and Overlap Top 1%</th>
<th>Overlap Top 0.1%</th>
<th>Overlap Top 1%</th>
<th>Overlap Top 0.1%</th>
<th>Overlap Top 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>iHS</td>
<td>-0.626 (0.0002)</td>
<td>-0.116 (0.212)</td>
<td>0.833 (0.0002)</td>
<td>0.223 (0.057)</td>
<td>-0.669 (0.003)</td>
</tr>
<tr>
<td>XP-EHH CEU</td>
<td>-0.885 (0)</td>
<td>-0.664 (0)</td>
<td>0.894 (0.0002)</td>
<td>0.637 (0.0004)</td>
<td>-0.913 (0)</td>
</tr>
<tr>
<td></td>
<td>-0.200 (0.137)</td>
<td>-0.096 (0.346)</td>
<td>0.834 (0.0002)</td>
<td>0.310 (0.066)</td>
<td>-0.243 (0.164)</td>
</tr>
<tr>
<td></td>
<td>-0.500 (0.001)</td>
<td>-0.341 (0)</td>
<td>0.866 (0.0002)</td>
<td>0.514 (0.007)</td>
<td>-0.747 (0)</td>
</tr>
<tr>
<td>XP-EHH YRI</td>
<td>-0.500 (0)</td>
<td>-0.341 (0)</td>
<td>0.866 (0.0002)</td>
<td>0.514 (0.007)</td>
<td>-0.747 (0)</td>
</tr>
<tr>
<td></td>
<td>-0.500 (0)</td>
<td>-0.647 (0)</td>
<td>0.869 (0.0002)</td>
<td>0.655 (0.0002)</td>
<td>-0.784 (0)</td>
</tr>
<tr>
<td>XP-EHH MKK</td>
<td>-0.500 (0)</td>
<td>-0.647 (0)</td>
<td>0.869 (0.0002)</td>
<td>0.655 (0.0002)</td>
<td>-0.784 (0)</td>
</tr>
<tr>
<td>XP-EHH KHB</td>
<td>-0.500 (0)</td>
<td>-0.647 (0)</td>
<td>0.869 (0.0002)</td>
<td>0.655 (0.0002)</td>
<td>-0.784 (0)</td>
</tr>
</tbody>
</table>

The indicated variables are calculated between all pairs of populations (creating matrices for which Mantel correlations can be calculated). P-values for the Mantel correlations, assessed by 5,000 randomizations, are indicated in parentheses and in bold if significant (p-value < 0.05). $F_{ST}$: mean $F_{ST}$ over all SNPs between pairs of populations; Correlations of Window Statistics: correlations of window statistics across all genomic windows between pairs of populations; Overlap Top 1%: number of shared windows in the empirical top 1% between pairs of populations; Overlap Top 0.1%: number of shared windows in the empirical top 0.1% between pairs of populations.
Table C.4: Significantly enriched Gene Ontology (GO) terms in the top empirical 1% and 5% of windows in each population.

<table>
<thead>
<tr>
<th>Population</th>
<th>Statistic</th>
<th>GO term*</th>
<th>Name of GO term</th>
<th>p value</th>
<th>Windows</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biaka Pygmy</td>
<td>XP-EHH CEU (0.05)</td>
<td>GO:0016571</td>
<td>histone methylation</td>
<td>0.00015</td>
<td>6</td>
<td>MEN1, EED, PRMT5, SUV420H2, EHMT1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CDK2, MDM2, CDC6, CDC25C, MTBP</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU (0.05)</td>
<td>GO:0007089</td>
<td>traversing start control point of mitotic cell cycle</td>
<td>5.2 x 10^-5</td>
<td>5</td>
<td>PICALM, ASGR1, LRPIB, DAB2, IGF2B</td>
</tr>
<tr>
<td>Namibian San</td>
<td>XP-EHH CEU (0.01)</td>
<td>GO:0006898 (P)</td>
<td>receptor-mediated endocytosis</td>
<td>0.00045</td>
<td>5</td>
<td>CDK2, MDM2, CDC6, CDC25C, MTBP</td>
</tr>
<tr>
<td>#Khomani Bushmen</td>
<td>XP-EHH CEU (0.05)</td>
<td>GO:0006355</td>
<td>regulation of transcription, DNA-dependent</td>
<td>0.00015</td>
<td>180</td>
<td>PML, GTSE1, GML</td>
</tr>
<tr>
<td>Bantu SA</td>
<td>XP-EHH YRI (0.01)</td>
<td>GO:0006977</td>
<td>DNA damage response, signal transduction by p53 class mediator, cell cycle arrest</td>
<td>5.2 x 10^-5</td>
<td>3</td>
<td>CDK2, MDM2, CDC6, CDC25C, MTBP</td>
</tr>
<tr>
<td>Bantu Kenya</td>
<td>XP-EHH CEU (0.05)</td>
<td>GO:0030178</td>
<td>negative regulation of Wnt signaling pathway</td>
<td>5.2 x 10^-5</td>
<td>7</td>
<td>PICALM, ASGR1, LRPIB, DAB2, IGF2B</td>
</tr>
<tr>
<td>Mandenka</td>
<td>iHS (0.01)</td>
<td>GO:002504 (also P)</td>
<td>antigen processing and presentation of peptide or polysaccharide via MBc Class II</td>
<td>0.00015</td>
<td>3</td>
<td>HLA-DRB5, HLA-DQB2, HLA-DPA1</td>
</tr>
<tr>
<td></td>
<td>iHS (0.01)</td>
<td>GO:0019882 (also P)</td>
<td>antigen processing</td>
<td>5.2 x 10^-5</td>
<td>5</td>
<td>CDK2, MDM2, CDC6, CDC25C, MTBP</td>
</tr>
<tr>
<td></td>
<td>iHS (0.01)</td>
<td>GO:0015031 (P)</td>
<td>protein transport</td>
<td>0.00015</td>
<td>17</td>
<td>HLA-DRB5, HLA-DQB2, HLA-DPA1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0006307 (P)</td>
<td>nitrogen compound metabolic process</td>
<td>0.00015</td>
<td>4</td>
<td>PICALM, ASGR1, LRPIB, DAB2, IGF2B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0006366 (P)</td>
<td>transcription from RNA polymerase II promoter</td>
<td>0.0009</td>
<td>8</td>
<td>HLA-DRB5, HLA-DQB2, HLA-DPA1</td>
</tr>
<tr>
<td>Yoruba</td>
<td>XP-EHH CEU (0.01)</td>
<td>GO:0007417</td>
<td>central nervous system development</td>
<td>5.2 x 10^-5</td>
<td>9</td>
<td>ACCN1, NPAS1, ADAM23, DNER, CHRD, PARK2, POUSF2, G1L3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0045586</td>
<td>negative regulation of cell cycle</td>
<td>0.0186</td>
<td>2</td>
<td>HrasL3, MOV10L1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0006366 (P)</td>
<td>transcription from RNA polymerase II promoter</td>
<td>0.00015</td>
<td>9</td>
<td>PRDM4, MAF, RNF, YBX2, HLF, SNACP2, BAPX1, CREB5, POUSF2, GATA4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0005088</td>
<td>lipid homeostasis</td>
<td>5.2 x 10^-7</td>
<td>3</td>
<td>USF1, USF2, PPARG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0003523</td>
<td>positive regulation of transcription from RNA polymerase II promoter</td>
<td>0.00015</td>
<td>2</td>
<td>USF1, USF2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0009991</td>
<td>response to extracellular stimulus</td>
<td>0.0186</td>
<td>2</td>
<td>RASGRF4, RPS19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0009991</td>
<td>response to extracellular stimulus</td>
<td>0.0186</td>
<td>2</td>
<td>RASGRF4, RPS19</td>
</tr>
</tbody>
</table>

Significance determined as described in Methods, using an FDR cutoff of 0.05, within each population. The “Statistic” column indicates the statistic resulting in significance; the numbers in parentheses after the statistic indicate the empirical tail for which an enrichment was tested (top 0.001 are shown in Chapter 4). Significant GO terms are indicated in the “GO Term” column. *(P)* indicates that the term is only significant when examining the PANTHER subset of terms; *(also P)* indicates that the term is significant both when examining PANTHER terms and when examining all GO terms; if nothing is next to the term, it indicates significance only when all GO terms are examined. We count the number of windows in the indicated empirical tail to which each term was associated (“Windows” column); rather than reporting those windows, we report the genes within those windows that were associated with the given GO term (in the “Genes” column).
Bibliography


