Statistical methods for analyzing human genetic variation in diverse populations

by

Chaolong Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2012

Doctoral Committee:

Associate Professor Noah A. Rosenberg, Co-chair
Professor Michael L. Boehnke, Co-chair
Professor Margit Burmeister
Associate Professor Ji Zhu
Associate Professor Sebastian K. Zöllner

To my parents and my wife

ACKNOWLEDGEMENTS

First of all, I would like to give my sincere thanks to Dr. Noah Rosenberg, who has been an excellent and supportive advisor to me through the past few years. He has provided me with strong instrumental guidance in research, as well as great freedom to pursue independent work. From Noah, I have not only learned about research skills, but also how to conduct research with rigor, integrity, and independence. By day to day interactions, I have been impressed by Noah's enthusiasm and innovation in science; discussing with him in front of a set of equations and figures is one of the most enjoyable moments for me.

I would also like to acknowledge my committee, Drs. Mike Boehnke, Margit Burmeister, Ji Zhu, and Sebastian Zöllner, for their insightful and constructive comments on my research, and for writing recommendation letters for me. I am especially grateful to Mike and Sebastian for giving me helpful suggestions and great scientific support after the Rosenberg lab moved to Stanford University. While not a committee member, I would also like to thank Dr. Gonçalo Abecasis, for productive discussions on research and pointing me to an interesting question that extends my work in this dissertation.

Next, I would like to thank all members from the Rosenberg lab. The time we spent together is what I phrase as "those good old days." In particular, I would like to thank Dr. Trevor Pemberton, for being such a great cubical mate, for bringing delicious food to the lab, and for officiating my wedding in Ann Arbor. I would also like to thank Lucy Huang for introducing me to the lab, Mike DeGiorgio for

constantly sharing and discussing research ideas, and Zach Szpiech for helping me with implementation of the *MicroDrop* program, as well as Ethan Jewett and Cuong Than for their help during my visits to Stanford.

I would also like to thank all members at the Center of Statistical Genetics for their friendship and nice conversations that always cheered me up. Specifically, I would like to thank Xiaowei Zhan for his great support in preparing data sets for my research and helping me with computational questions. I would also like to show my gratitude to Drs. Sebastian Zöllner, Tom Blackwell, and William Wen for their time and patience in helping me improve my presentations.

I would like to extend my gratitude to individuals that have contributed to my career path. I would like to thank Drs. Dan Burns and Nina Lin, for recruiting me to the Bioinformatics program and helping me with transition to graduate school during my first year here. I clearly remember that it was February 14, 2008 when I received their emails, giving me my first graduate school offer. Thank you Dan and Nina, for sending me such an awesome gift on Valentine's day! I would also like to thank my teacher and lifelong friend, Dr. Kwok-Yip Szeto at the Hong Kong University of Science and Technology, for his mentorship in my early research in Physics and for his faith in my ability that encourages me to pursue a career in science. In addition, I would like to thank Drs. Minping Qian and Minghua Deng at Peking University, from whom I have learned a lot on mathematical and statistical modeling.

I am truly indebted and thankful to my parents, Conghai Wang and Liping Huang, for their endless love and support throughout my life. They always encourage me to pursue my dreams and provide me with all help they could have made. Whenever in difficulty, I won't be afraid, because I know there is a home waiting for me.

A special thanks goes to my beloved wife, Shanshan Cheng, who has always been with me through good times and bad times during my Ph.D. Thank you for your emotional support and all the delicious dishes you have made for me. It is my fortune

to meet you here and have you in the rest of my life.

Finally, this dissertation would not have been possible without contributions from my collaborators. I would like to thank all coauthors of Chapters II-IV in this dissertation, including Drs. James Degnan (III), John Hardy (III), Mattias Jakobsson (III), Trevor Pemberton (III), Noah Rosenberg (II, III, IV), Kari Schroeder (II), Andrew Singleton (III), Zachary Szpiech (III), and Sebastian Zöllner (IV).

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Statistical methods for analyzing human genetic variation in diverse populations

by

Chaolong Wang

Co-chairs: Noah A. Rosenberg and Michael L. Boehnke

The recent expansion of genetic datasets in diverse populations has allowed researchers to investigate human genetic structure and evolutionary history with unprecedented resolution. The huge amount of data also poses new statistical challenges, in both quality control and data analysis. In this dissertation, I develop statistical methods to address some challenges arising from recent population-genetic studies, and apply the methods to study the geographic structure of human genetic variation.

First, I develop a method to correct for allelic dropout, a common source of genotyping error in microsatellite data. Traditional solutions for allelic dropout often require replicate genotyping, which is costly and often impossible in population-genetic studies. To address this problem, I propose a maximum likelihood approach to estimate dropout rates from nonreplicated microsatellite genotypes. Based on simulations and empirical data, I show that this method is both accurate and fairly robust to some violations of model assumptions.

Next, I introduce a Procrustes analysis approach to compare spatial maps of genetic variation. Multivariate techniques, such as principal components analysis (PCA), have been widely used to summarize population structure, typically in two-

dimensional maps, which often resemble the geographic maps of sampling locations. Using the Procrustes approach, I quantitatively demonstrate that genetic coordinates based on SNPs and CNVs are similar to each other, and are highly concordant with the geographic coordinates.

Finally, applying PCA and Procrustes analysis on SNP data from worldwide populations, I perform a systematic study to compare genes and geography across the globe. By considering examples in different regions, I find that significant similarity between genes and geography exists in general. Further, the similarity is highest in Asia and once isolated populations have been removed, Sub-Saharan Africa. The results provide a quantitative assessment of the geographic structure of human genetic variation worldwide.

In summary, this dissertation contributes both statistical tools for analyzing large-scale genetic data and biological insights on the spatial patterns of human genetic variation. Results from this dissertation provide a basis for evaluating the role of geography in giving rise to human population structure, and can facilitate statistical methods for inferring individual geographic origin from genetic variation.

CHAPTER I

Introduction

The development of molecular-genetic technology has significantly contributed to the expansion of genetic data sets for various types of DNA variation. Especially for humans, a large number of genetic markers have been assayed in over thousands of individuals from diverse populations worldwide (e.g. *Rosenberg et al.*, 2002; *Li et al.*, 2008; *The International HapMap 3 Consortium*, 2010). The large amounts of data allow researchers to investigate questions of human genetic structure and evolutionary history with previously unavailable resolution. However, they also pose new statistical challenges for genetic research. These challenges include how to assess data quality and correct for systematic errors inherent to large data sets, and how to efficiently analyze data and provide biological interpretation. In this dissertation, I focus on developing statistical methods to address some of the challenges arising from recent population-genetic studies of two of the most widely used types of molecular markers: microsatellites and single-nucleotide polymorphisms (SNPs). Using the methods, I investigate the geographic structure of human genetic variation in worldwide populations.

Microsatellites are tandemly repeated sequences in DNA, with the length of a repeat unit typically ranging from 1 to 6 base pairs (bp). At a microsatellite locus, distinct alleles represent DNA fragments of different numbers of repeat units. The

number of repeats for a microsatellite allele can range from 3 to over 100, such that microsatellites are highly polymorphic and informative (*Rosenberg et al.*, 2003). In addition, microsatellites are fairly abundant, compromising about 3% of the human genome (*The International Human Genome Sequencing Consortium*, 2001; *Payseur et al.*, 2011), and they distribute evenly across the genome (*Subramanian et al.*, 2003). These properties make microsatellites popular as genetic markers in many research areas, including forensics (e.g. *Urquhart et al.*, 1994), linkage studies (e.g. *Weissenbach et al.*, 1992), paternity testing (e.g. *Pena and Chakraborty*, 1994), and population genetics (e.g. *Rosenberg et al.*, 2002). Further, the high mutation rate of microsatellites (10^{-4} to 10^{-3} per base per generation, *Ellegren*, 2000) allows microsatellites to be used as molecular clocks to infer human evolutionary history (*Goldstein et al.*, 1995a,b; *Slatkin*, 1995). The availability of large microsatellite data sets has led to great advances in our understanding of worldwide human population structure and the evolutionary history of modern humans (*Bowcock et al.*, 1994; *Jorde et al.*, 1995, 1997; *Kimmel et al.*, 1998; *Rosenberg et al.*, 2002; *Cavalli-Sforza and Feldman*, 2003; *Ramachandran et al.*, 2005; *Prugnolle et al.*, 2005; *Wang et al.*, 2007b; *Friedlaender et al.*, 2008; *Tishkoff et al.*, 2009).

As the most popular genetic markers in human research from the 1990s to the early 2000s, microsatellites still play an important role in many areas. Especially in molecular ecology, microsatellites remain the primary genetic markers to study non-model organisms, for which SNP arrays are not well developed (*Schlötterer*, 2004; *Seeb et al.*, 2011). Even in humans, microsatellites are still used in many population-genetic studies because of the existence of large amounts of microsatellite data from diverse populations (e.g. *Ramachandran and Rosenberg*, 2011), and because of special properties of microsatellites, such as high mutation rates and high levels of polymorphism (e.g. *Sun et al.*, 2009).

An important data quality issue with microsatellites is "allelic dropout." Genetic

variation at microsatellite loci is typically detected using polymerase chain reaction (PCR) amplification followed by electrophoresis (*Lai and Sun*, 2004). This genotyping procedure is generally reliable when using high-quality DNA samples. However, when DNA samples have poor quality, one or both allelic copies at a locus often fail to be amplified with PCR, resulting in either spurious homozygotes or missing data. This problem is known as "allelic dropout," the most significant type of genotyping error for microsatellites (*Pompanon et al.*, 2005; *Hoffman and Amos*, 2005). Allelic dropout can substantially decrease data accuracy and lead to mistaken results in many microsatellite-based studies, especially for studies in molecular ecology, where DNA samples are often collected from noninvasive sources and thus have relatively low quality (e.g. *Fernando et al.*, 2003; *Broquet and Petit*, 2004). Existing approaches, both experimental and computational, for avoiding allelic dropout often require repeated genotyping to minimize the effects of experimental error (*Miller et al.*, 2002; *Wang*, 2004; *Hoffman and Amos*, 2005; *Hadfield et al.*, 2006; *Johnson and Haydon*, 2007; *Wright et al.*, 2009). These approaches, however, are costly and are only suitable when enough DNA is available for repeated genotyping. Chapter II (*Wang et al.*, 2012a) investigates the allelic dropout problem when replicated genotypes are not available. I develop a maximum likelihood method with an expectation-maximization (EM) algorithm to jointly estimate allele frequencies, the inbreeding coefficient, and both sample-specific and locus-specific dropout rates from a single set of genotypes. Further, I propose a multiple imputation strategy to correct for allelic dropout in downstream analyses. I implement this method in a software program *MicroDrop*, which can be useful for preparing data sets to circumvent allelic dropout in diverse applications in population genetics and molecular ecology.

Upon the arrival of high-throughput genotyping technologies, SNPs have gradually replaced microsatellites and become the most popular genetic markers in human population-genetic research. The high density of SNPs across the genome also makes

SNPs suitable genetic markers for genome-wide association studies, which search across the genome for disease-susceptibility genes by comparing large numbers of diseased individuals (cases) and healthy individuals (controls) (*The International HapMap Consortium*, 2003; *Hirschhorn and Daly*, 2005; *Hindorff et al.*, 2009). Unlike multi-allelic microsatellites, most SNP markers only have two allelic states. Nevertheless, the number of SNP loci in the human genome is much larger than the number of microsatellite loci (*The International Human Genome Sequencing Consortium*, 2001; *The 1000 Genome Project Consortium*, 2010). Further, cost-effective array technologies allow SNPs to be genotyped at a very large scale, both in number of samples and number of markers. Recent population-genetic studies based on hundreds of thousands of SNP markers can reach an unprecedented resolution in identifying fine-scale population structure (for a review, see *Novembre and Ramachandran*, 2011).

Understanding population structure is crucial for both evolutionary biology and disease association studies. In evolutionary biology, population structure provides the basis to infer evolutionary processes, such as migration, admixture, and natural selection, that shape the spatial distribution of human genetic variation (e.g. *DeGiorgio et al.*, 2009; *Pickrell et al.*, 2009; *Bryc et al.*, 2010a,b). In disease association studies, when cases and controls are sampled from groups with different genetic backgrounds, population structure can lead to false association signals between non-causal genetic variants and common genetic diseases (*Pritchard and Rosenberg*, 1999; *Marchini et al.*, 2004). To study population structure, multivariate analyses, as statistical methods for summarizing high-dimensional data into a few synthetic variables, are often used to extract information from hundreds to millions of genetic markers (*Jombart et al.*, 2009). Principal components analysis (PCA) and mutidimensional scaling (MDS) are two of the most popular multivariate techniques in population genetics, and in disease association studies to control for population structure (*Price et al.*, 2006; *Engelhardt and Stephens*, 2010). Especially for large-scale SNP data sets, PCA and MDS are fa-

vored because of their computational efficiency and high level of power to decompose the complex structure of human genetic variation (*Patterson et al.*, 2006; *Paschou et al.*, 2007).

Both PCA and MDS summarize the structure of human genetic variation by projecting individuals in a low-dimensional space such that Euclidean distances between individuals approximately reflect their genetic distances. Many studies have reported that statistical maps of genetic variation generated by the first two components of PCA or MDS closely match geographic sampling locations, from the continental level such as in Europe (*Novembre et al.*, 2008; *Lao et al.*, 2008; *Heath et al.*, 2008) and West Africa (*Bryc et al.*, 2010b), to more local levels, such as in Finland (*Jakkula et al.*, 2008; *Hoggart et al.*, 2012), Iceland (*Price et al.*, 2009), and Sweden (*Salmela et al.*, 2011). These studies, however, mostly focusing on specific regions, often used different data processing procedures and different statistical techniques. Further, their comparisons with geography have been based largely on qualitative observations. These problems have caused great difficulty for integrating results from different studies to obtain a comprehensive picture of the spatial pattern of human genetic variation across the world. Chapters III (*Wang et al.*, 2010) and IV (*Wang et al.*, 2012b) address this challenge by developing novel methods to quantify similarity between spatial maps of genetic variation, and by providing a systematic study on comparing genes and geography in worldwide human populations.

Despite the popularity of PCA and MDS in population-genetic studies, relatively few quantitative approaches are available to interpret results in PCA or MDS maps. Chapter III (*Wang et al.*, 2010) introduces a quantitative approach to formally evaluate the similarity between different spatial maps of human genetic variation. This approach is based on Procrustes analysis, a classic multivariate statistical technique in shape analysis (*Dryden and Mardia*, 1998; *Cox and Cox*, 2001). Given two sets of coordinates — two maps — Procrustes analysis identifies the optimal transformations

of one set of coordinates to maximize its similarity to the other set, and it provides a similarity score between the transformed maps. The statistical significance of the similarity score is then evaluated by a permutation test (*Jackson*, 1995; *Peres-Neto and Jackson*, 2001). Several examples are analyzed in Chapter III to illustrate the versatility of the Procrustes approach in population-genetic applications, including comparisons between (1) statistical maps of population-genetic variation and geographic maps of sampling locations in European and worldwide samples, (2) statistical maps of partially overlapped samples generated by different statistical techniques (PCA vs. MDS), and (3) statistical maps of the same samples based on different types of genetic markers (SNPs vs. copy-number variants, CNVs). With these examples, I show that statistical maps of human genetic variation based on SNPs and CNVs have a high level of agreement with each other and match closely with geographic sampling locations.

Chapter IV presents a subsequent study that employs PCA and the Procrustes approach to systematically assess the geographic structure of human genetic variation across different regions of the world. I integrate genome-wide SNP data and geographic coordinates for 149 worldwide populations, including data in the Human Genome Diversity Cell Line Panel (HGDP, *Li et al.*, 2008) and HapMap Phase 3 (*The International HapMap 3 Consortium*, 2010), as well as data previously reported by several other studies (*Novembre et al.*, 2008; *Xing et al.*, 2009, 2010; *Bryc et al.*, 2010b; *Simonson et al.*, 2010). I evaluate the similarity between genes and geography in different geographic regions using a common analysis framework and a common set of markers, such that results for different locations can be directly compared to each other. By considering examples sampled from Europe, Sub-Saharan Africa, Asia, Central/South Asia, and East Asia, as well as a worldwide sample, I find that significant similarity between genes and geography exists in general at different geographic levels. Further, the highest similarity scores appear in Asia, and once isolated pop-

ulations have been removed, Sub-Saharan Africa. The results provide a quantitative assessment of the geographic structure of human genetic variation worldwide, supporting a view that geography plays a strong role in giving rise to human population structure.

Together, I develop two statistical methods including a software program in this dissertation, addressing different statistical challenges arising from recent population-genetic studies. In addition, with this dissertation, I contribute novel biological insights on the geographic structure of human genetic variation, which can benefit studies that demand knowledge of human population structure, such as evolutionary and disease association studies. Three chapters in this dissertation have been published or in press, by *Genetics* (II, *Wang et al.*, 2012a), *Statistical Applications in Genetics and Molecular Biology* (III, *Wang et al.*, 2010), and *PLoS Genetics* (IV, *Wang et al.*, 2012b).

CHAPTER II

A maximum likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes

2.1 Introduction

Microsatellite markers are widely used in population genetics and molecular ecology. In microsatellite data, distinct alleles at a locus represent DNA fragments of different sizes, typically detected by amplification using the polymerase chain reaction (PCR). Frequently, during microsatellite genotyping in diploid organisms, one or both of an individual's two copies of a locus fail to amplify with PCR, yielding a spurious homozygote or a spurious occurrence of missing data. This problem is known as "allelic dropout" (e.g. *Gagneux et al.*, 1997; *Pompanon et al.*, 2005). For example, if an individual has genotype AB at a locus, but only allele A successfully amplifies, then only allele A will be detected, and the genotype will be erroneously recorded as AA. If neither allelic copy amplifies, then the genotype will be recorded as missing. Here we follow *Miller et al.* (2002) by using "copies" to refer to the paternal and maternal variants in an individual and "alleles" to specify the distinct allelic types possible at a locus.

Allelic dropout is common in microsatellite studies and can lead to statistical

errors in subsequent analyses (e.g. *Bonin et al.*, 2004; *Broquet and Petit*, 2004; *Hoffman and Amos*, 2005). For example, in estimating population-genetic statistics, because allelic dropout can cause mistaken assignment of heterozygous genotypes as homozygotes, it can lead to underestimation of the observed heterozygosity and overestimation of the inbreeding coefficient (*Taberlet et al.*, 1999). Circumventing allelic dropout is therefore important for microsatellite studies. One general strategy for correcting for allelic dropout involves repeated genotyping, particularly for the apparent homozygotes (e.g. *Taberlet et al.*, 1996; *Morin et al.*, 2001; *Wasser et al.*, 2007). Additionally, computational approaches have been proposed to assess allelic dropout, primarily when replicate genotypes are available (*Miller et al.*, 2002; *Wang*, 2004; *Hadfield et al.*, 2006; *Johnson and Haydon*, 2007; *Wright et al.*, 2009). In practice, however, replicate genotyping is costly and often uninformative or impossible owing to insufficient DNA or logistical constraints, especially for natural populations with limited DNA samples from noninvasive sources (e.g. *Taberlet and Luikart*, 1999; *Taberlet et al.*, 1999). Therefore, in this study, we develop a maximum likelihood approach that can correct for allelic dropout without using replicate genotypes.

It is believed that the cause of allelic dropout is stochastic sampling of the molecular product, which can occur at two stages of the genotyping process (Fig. 2.1). If DNA concentration is low, then one or both of the allelic copies might not be present in sufficient quantity for successful amplification (e.g. *Navidi et al.*, 1992; *Taberlet et al.*, 1996; *Sefc et al.*, 2003). Poor quality of the template DNA (e.g. high degradation) can also prevent binding by the PCR primers and polymerase, resulting in dropout. An additional problem in the binding step is that some loci might be less likely than others to be bound. Previous studies have found that although different alleles at the same locus have similar probabilities of dropping out, loci with longer alleles tend to have higher dropout rates than those with shorter alleles (e.g. *Sefc et al.*, 2003; *Buchan et al.*, 2005; *Broquet et al.*, 2007); differences in primer anneal-

ing efficiency and in template DNA secondary structures might also contribute to different dropout rates across loci (*Buchan et al.*, 2005).

In this study, we explicitly model the two sources of allelic dropout using sample-specific dropout rates $\gamma_{i\cdot}$ and locus-specific dropout rates $\gamma_{\cdot\ell}$, such that the probability of allelic dropout at locus ℓ of individual i is determined by a function of both $\gamma_{i\cdot}$ and $\gamma_{\cdot\ell}$. With a single nonreplicated set of genotypes, we jointly estimate the parameters of the model, including allele frequencies, sample-specific dropout rates, locus-specific dropout rates, and an inbreeding coefficient, thereby correcting for the underestimation of observed heterozygosity and overestimation of inbreeding caused by allelic dropout. We use an expectation-maximization (EM) algorithm to obtain maximum likelihood estimates (MLEs). With the estimated parameter values, we perform multiple imputation to correct the bias caused by allelic dropout in estimating the observed heterozygosity. We have implemented this method in *MicroDrop*, which is freely available at `http://rosenberglab.stanford.edu`.

We first employ the method for analyzing a set of human microsatellite genotypes from Native American populations. Using the estimated parameter values, we generate a simulated data set that mimics the Native American data, and we employ this simulated data set to evaluate the performance of our model. First, we compare the patterns of missing data and heterozygosity between the simulated and real data to check if our model correctly reproduces the observed patterns. Next, we compare estimated and true values of the allelic dropout rates for the simulated data. Finally, we compare the corrected heterozygosity with the "true" heterozygosity calculated from the true genotype data prior to allelic dropout. We further evaluate the robustness of our model using simulations with different levels of inbreeding, population structure, and genotyping errors from sources other than allelic dropout. We conclude our study by using simulations to argue that our MLEs of dropout rates and the inbreeding coefficient are consistent. That is, we show that as the number of individuals and the

number of genotyped loci increase, our estimated values appear to converge to the true values of the parameters.

2.2 Data and preliminary analysis

The data set on which we focus consists of genotypes for 343 microsatellite markers in 152 Native North Americans collected from 14 populations over many years by the laboratory of D. G. Smith at the University of California, Davis. We identify the populations according to their sampling locations: three populations from the Arctic/Subarctic region, two from the Midwest of the United States (US), two from the Southeast US, two from the Southwest US, three from the Great Basin/California region, and two from Central Mexico. In this data set, the number of distinct alleles per locus has mean 8.0 across loci, with a minimum of 4 and a maximum of 24.

Allelic dropout can generate both spurious homozygotes, when one allelic copy drops out at a heterozygous locus, and missing data, when both copies drop out at either homozygous or heterozygous loci. Thus, under the hypothesis that missing data are caused by allelic dropout, we expect a higher proportion of missing data to be accompanied by a higher proportion of homozygous genotypes. If allelic dropout is caused by low DNA concentration or low quality in certain samples, then a positive correlation will be observed across individuals between missing data and individual homozygosity. Alternatively, if allelic dropout is caused by locus-specific factors such as differences across loci in the binding properties of the primers or polymerase, we instead expect a positive correlation across loci between missing data and locus homozygosity. This type of correlation is also expected if missing data are due to "true missingness"—for example, null alleles segregating in the population at certain loci, as a result of polymorphic deletions in primer regions (e.g. *Pemberton et al.*, 1995; *Dakin and Avise*, 2004). Here, we disregard true missingness and assume that all missing genotypes are attributable to allelic dropout.

For each individual, we evaluated the proportion of loci at which missing data occurred and the proportion of homozygotes among those loci for which data were not missing. As shown in Fig. 2.2A, missing data and homozygosity have a strong positive correlation: the Pearson correlation is $r = 0.729$ ($P < 0.0001$, by 10,000 permutations of the proportions of homozygous loci across individuals). This observation matches the prediction of the hypothesis that missing data result from sample-specific dropout rather than locus-specific dropout or "true missingness." By contrast, an analogous computation for each locus rather than for each individual (Fig. 2.2B) finds that the correlation between homozygosity and missing data is much smaller ($r = 0.099$ and $P = 0.0341$, by 10,000 permutations of the proportions of homozygous individuals across loci). We therefore suspect that missing genotypes in this data set arise primarily from the allelic dropout caused by low DNA concentration or quality in some samples, and that locus-specific factors such as poor binding affinity of primers and polymerase have a smaller effect. In any case, for our subsequent analyses, we continue to consider both sample-specific and locus-specific factors.

2.3 Model

Consider N individuals and L loci. Denote alleles at locus ℓ by $A_{\ell k}$ with $k = 1, 2, \ldots, K_\ell$, where K_ℓ is the number of distinct alleles at locus ℓ. Denote the observed genotype data by $W = \{w_{i\ell} : i = 1, 2, \ldots, N; \ell = 1, 2, \ldots, L\}$, where genotyping has been attempted for all individuals at all loci. Here, $w_{i\ell}$ is the observed genotype of the ith individual at the ℓth locus. Each entry of W consists of the two observed copies at a locus in a specific individual. If the observed genotype is missing at locus ℓ of individual i, then we specify $w_{i\ell} = XX$. Otherwise, $w_{i\ell} = A_{\ell k}A_{\ell h}$ for some $k, h \in \{1, 2, \ldots, K_\ell\}$, where k and h are not necessarily distinct. The true genotypes are denoted by $G = \{g_{i\ell} : i = 1, 2, \ldots, N; \ell = 1, 2, \ldots, L\}$. A description of the notation appears in Table 2.1.

To model the dropout mechanism, we specify a set of dropout states $Z = \{z_{i\ell} : i = 1, 2, \ldots, N; \ell = 1, 2, \ldots, L\}$ that connects G and W and that indicates which alleles "drop out." For a heterozygous true genotype $g_{i\ell} = A_{\ell k}A_{\ell h}$ $(h \neq k)$, supposing allele $A_{\ell k}$ drops out, the dropout state is $z_{i\ell} = A_{\ell h}X$ and the observed genotype is $w_{i\ell} = A_{\ell h}A_{\ell h}$. For a homozygous true genotype $g_{i\ell} = A_{\ell k}A_{\ell k}$, the dropout state $z_{i\ell} = A_{\ell k}X$ means that exactly one of the two allelic copies drops out.

We make five assumptions in our model:

1. All distinct alleles are observed at least once in our data set;

2. All missing and incorrect genotypes are attributable to allelic dropout;

3. Both copies at a locus ℓ of an individual i have equal probability $\gamma_{i\ell}$ of dropping out. This probability is a function of a sample-specific dropout rate $\gamma_{i\cdot}$ and a locus-specific dropout rate $\gamma_{\cdot\ell}$:

$$\gamma_{i\ell} = \gamma_{i\cdot} + \gamma_{\cdot\ell} - \gamma_{i\cdot}\gamma_{\cdot\ell}; \tag{2.1}$$

4. All individuals are unrelated and have the same inbreeding coefficient ρ, such that for any locus of any individual, the two allelic copies are identical by descent (IBD) with probability ρ;

5. Each pair of loci is independent (i.e. each pair of loci is at linkage equilibrium).

Denote $\Gamma = \{\gamma_{i\cdot}, \gamma_{\cdot\ell} : i = 1, 2, \ldots, N; \ell = 1, 2, \ldots, L\}$ and $\Phi = \{\phi_{\ell k} : \ell = 1, 2, \ldots, L; k = 1, 2, \ldots, K_\ell\}$, in which $\phi_{\ell k}$ is the true frequency of allele $A_{\ell k}$ at locus ℓ, $\gamma_{i\cdot}$ is the probability of dropout caused by sample-specific factors for any allelic copy at any locus of individual i, and $\gamma_{\cdot\ell}$ is the probability of dropout caused by locus-specific factors for any allelic copy at locus ℓ in any individual. Eq. 2.1 arises by noting that the dropout probability for an allelic copy at locus ℓ of individual i, considering the two possible causes as independent, is $\gamma_{i\ell} = 1 - (1 - \gamma_{i\cdot})(1 - \gamma_{\cdot\ell})$.

Using assumption 3, the conditional probability $\mathbb{P}(z_{i\ell}|g_{i\ell}, \Gamma)$ can be expressed as shown in Table 2.2. The conditional probability of observing genotype $w_{i\ell}$ given true

genotype $g_{i\ell}$ and dropout rates $\gamma_{i\cdot}$ and $\gamma_{\cdot\ell}$ can be calculated as

$$\mathbb{P}(w_{i\ell}|g_{i\ell},\Gamma) = \sum_{z_{i\ell}} \mathbb{P}(w_{i\ell}|z_{i\ell},g_{i\ell})\mathbb{P}(z_{i\ell}|g_{i\ell},\Gamma). \tag{2.2}$$

Here, $\mathbb{P}(w_{i\ell}|z_{i\ell},g_{i\ell})$ is either 0 or 1 because W is fully determined by Z and G, and the summation proceeds over all dropout states $z_{i\ell}$ possible given the observed genotype $w_{i\ell}$ (Table 2.2).

We use a set of binary random variables $S = \{s_{i\ell}\}$ to indicate the IBD states of the true genotypes G, such that $s_{i\ell} = 1$ if the two allelic copies in genotype $g_{i\ell}$ are IBD, and $s_{i\ell} = 0$ otherwise. Under assumption 4, we have (e.g. *Holsinger and Weir*, 2009)

$$\mathbb{P}(s_{i\ell}|\rho) = \begin{cases} \rho & \text{if } s_{i\ell} = 1 \\ 1-\rho & \text{if } s_{i\ell} = 0 \end{cases} \tag{2.3}$$

$$\mathbb{P}(g_{i\ell}|s_{i\ell},\Phi) = \begin{cases} \phi_{\ell k}^2 & \text{if } g_{i\ell} = A_{\ell k}A_{\ell k} \text{ and } s_{i\ell} = 0 \\ 2\phi_{\ell k}\phi_{\ell h} & \text{if } g_{i\ell} = A_{\ell k}A_{\ell h}\ (h \neq k) \text{ and } s_{i\ell} = 0 \\ \phi_{\ell k} & \text{if } g_{i\ell} = A_{\ell k}A_{\ell k} \text{ and } s_{i\ell} = 1 \\ 0 & \text{if } g_{i\ell} = A_{\ell k}A_{\ell h}\ (h \neq k) \text{ and } s_{i\ell} = 1 \end{cases} \tag{2.4}$$

$$\mathbb{P}(g_{i\ell}|\Phi,\rho) = \begin{cases} (1-\rho)\phi_{\ell k}^2 + \rho\phi_{\ell k} & \text{if } g_{i\ell} = A_{\ell k}A_{\ell k} \\ 2(1-\rho)\phi_{\ell k}\phi_{\ell h} & \text{if } g_{i\ell} = A_{\ell k}A_{\ell h}\ (h \neq k). \end{cases} \tag{2.5}$$

When $\rho = 0$, the genotype frequencies in eq. 2.5 follow Hardy-Weinberg equilibrium (HWE).

With the quantities in eqs. 2.2-2.5, the probability of observing $w_{i\ell}$ given parameters Ψ is

$$\mathbb{P}(w_{i\ell}|\Psi) = \sum_{g_{i\ell}} \mathbb{P}(w_{i\ell}|g_{i\ell},\Gamma)\mathbb{P}(g_{i\ell}|\Phi,\rho). \tag{2.6}$$

The summation proceeds over the set of all possible true genotypes $g_{i\ell}$, that is, over

all two-allele combinations at locus ℓ. The likelihood function of the parameters $\Psi = \{\Phi, \Gamma, \rho\}$ is then given by

$$\mathbb{P}(W|\Psi) = \prod_{i=1}^{N} \prod_{\ell=1}^{L} \mathbb{P}(w_{i\ell}|\Psi). \tag{2.7}$$

This likelihood assumes that dropout at a locus is independent across individuals, so that each observed diploid genotype of an individual at the locus is a separate trial independent of all others. Further, assumption 5 enables us to take a product across loci, as genotypes at separate loci are independent. A graphical representation of the relationships among the parameters Φ, Γ, and ρ, the latent variables G, S, and Z, and the observation W appears in Fig. 2.3.

2.4 Estimation procedure

Given the observed genotypes W, we can use an EM algorithm (e.g. *Lange*, 2002) to obtain the MLEs of the allele frequencies Φ, the sample-specific and locus-specific dropout rates Γ, and the inbreeding coefficient ρ. Under the inbreeding assumption (assumption 4), two allelic copies at the same locus need not be independent. If two allelic copies are IBD, then the allelic state of one copy is determined given the allelic state of the other copy, so that the number of independent allelic copies is 1. If two copies at the same locus are not IBD, then the number of independent allelic copies is 2. We introduce a random variable $n_{\ell k}$ to represent the number of "independent" copies of allele $A_{\ell k}$ in the whole data set, considering all individuals. We also define a random variable $d_{i\ell}$ as the number of copies that drop out at locus ℓ of individual i ($d_{i\ell} = 0, 1, \text{or } 2$).

In the E-step of our EM algorithm, we calculate (1) the expectation of the number of independent copies for all alleles, $\mathbb{E}[n_{\ell k}|W, \Psi]$, summing across individuals; (2) for each individual, the total number of dropouts caused by sample-specific fac-

tors, $\mathbb{E}[d_{i\cdot}|W,\Psi] = \sum_{\ell=1}^{L}\mathbb{E}[d_{i\ell}|W,\Psi](\gamma_{i\cdot}/\gamma_{i\ell})$; (3) for each locus, the total number of dropouts caused by locus-specific factors, $\mathbb{E}[d_{\cdot\ell}|W,\Psi] = \sum_{i=1}^{N}\mathbb{E}[d_{i\ell}|W,\Psi](\gamma_{\cdot\ell}/\gamma_{i\ell})$; and (4) the expectation of the total number of genotypes that are IBD, summing across the whole data set, $\mathbb{E}[s|W,\Psi] = \sum_{i=1}^{N}\sum_{\ell=1}^{L}\mathbb{E}[s_{i\ell}|W,\Psi]$. The factors $\gamma_{i\cdot}/\gamma_{i\ell}$ and $\gamma_{\cdot\ell}/\gamma_{i\ell}$ specify the respective probabilities that sample-specific factors and locus-specific factors contribute to the allelic dropouts at locus ℓ of individual i.

To obtain the expectations required for the E-step, we need the posterior probabilities of $g_{i\ell}$, $d_{i\ell}$, and $s_{i\ell}$ given the observed genotype $w_{i\ell}$ and the parameters Ψ, for each (i,ℓ) with $i = 1,2,\ldots,N$ and $\ell = 1,2,\ldots,L$. The posterior joint probabilities of $g_{i\ell}$ and $s_{i\ell}$ given $w_{i\ell}$ and Ψ are listed in Table 2.3, and they are calculated from Bayes' formula:

$$\begin{aligned}\mathbb{P}(g_{i\ell}, s_{i\ell}|w_{i\ell},\Psi) &= \frac{\mathbb{P}(g_{i\ell}, s_{i\ell}|\Psi)\mathbb{P}(w_{i\ell}|g_{i\ell}, s_{i\ell},\Psi)}{\sum_{g_{i\ell}}\sum_{s_{i\ell}=0}^{1}\mathbb{P}(g_{i\ell}, s_{i\ell}|\Psi)\mathbb{P}(w_{i\ell}|g_{i\ell}, s_{i\ell},\Psi)} \\ &= \frac{\mathbb{P}(s_{i\ell}|\rho)\mathbb{P}(g_{i\ell}|s_{i\ell},\Phi)\mathbb{P}(w_{i\ell}|g_{i\ell},\gamma_{i\ell})}{\sum_{g_{i\ell}}\sum_{s_{i\ell}=0}^{1}\mathbb{P}(s_{i\ell}|\rho)\mathbb{P}(g_{i\ell}|s_{i\ell},\Phi)\mathbb{P}(w_{i\ell}|g_{i\ell},\gamma_{i\ell})}. \end{aligned} \tag{2.8}$$

The second equality holds because the probability of being IBD ($s_{i\ell} = 1$) depends only on the inbreeding coefficient ρ, the true genotype $g_{i\ell}$ is independent of ρ and the dropout rate $\gamma_{i\ell}$ given $s_{i\ell}$ and the allele frequencies Φ, and the observed genotype $w_{i\ell}$ is independent of Φ and ρ given $g_{i\ell}$ and $\gamma_{i\ell}$.

For example, suppose the observed genotype is $w_{i\ell} = A_{\ell k}A_{\ell k}$, and we wish to evaluate $\mathbb{P}(g_{i\ell} = A_{\ell k}A_{\ell k}, s_{i\ell} = 1|w_{i\ell} = A_{\ell k}A_{\ell k},\Psi)$, the posterior joint probability that the true genotype is $g_{i\ell} = A_{\ell k}A_{\ell k}$ and the two allelic copies are IBD. If $w_{i\ell} = A_{\ell k}A_{\ell k}$ is observed, then the true genotype $g_{i\ell}$ can be a homozygote $A_{\ell k}A_{\ell k}$ or a heterozygote $A_{\ell k}A_{\ell h}$, with $h \in \{1,2,\ldots,K_\ell\}$ and $h \neq k$. Each term in the summation in eq. 2.8 is a joint probability $\mathbb{P}(g_{i\ell}, s_{i\ell}, w_{i\ell}|\Psi)$. To calculate this quantity, $\mathbb{P}(s_{i\ell}|\rho)$ and $\mathbb{P}(g_{i\ell}|s_{i\ell},\Phi)$ are obtained using eqs. 2.3 and 2.4 respectively. The values of $\mathbb{P}(w_{i\ell} = A_{\ell k}A_{\ell k}|g_{i\ell},\gamma_{i\ell})$ are given by Table 2.2 and can be obtained using eq. 2.2. The resulting probabilities

$\mathbb{P}(g_{i\ell}, s_{i\ell}, w_{i\ell}|\Psi)$ appear in Table 2.3. Therefore, for example,

$$\begin{aligned}\mathbb{P}(g_{i\ell} = A_{\ell k}A_{\ell k}, s_{i\ell} = 1|w_{i\ell} = A_{\ell k}A_{\ell k}, \Psi) &= \frac{\mathbb{P}(g_{i\ell} = A_{\ell k}A_{\ell k}, s_{i\ell} = 1, w_{i\ell} = A_{\ell k}A_{\ell k}|\Psi)}{\sum_{g_{i\ell}}\sum_{s_{i\ell}=0}^{1}\mathbb{P}(g_{i\ell}, s_{i\ell}, w_{i\ell} = A_{\ell k}A_{\ell k}|\Psi)} \\ &= \frac{\rho\phi_{\ell k}(1-\gamma_{i\ell}^2)}{\rho\phi_{\ell k}(1-\gamma_{i\ell}^2) + (1-\rho)\phi_{\ell k}^2(1-\gamma_{i\ell}^2) + \sum\limits_{\substack{h\in\{1,2,\ldots,K_\ell\}\\ h\neq k}} 2(1-\rho)\phi_{\ell k}\phi_{\ell h}\gamma_{i\ell}(1-\gamma_{i\ell})} \\ &= \frac{\rho\phi_{\ell k}(1-\gamma_{i\ell}^2)}{\rho\phi_{\ell k}(1-\gamma_{i\ell}^2) + (1-\rho)\phi_{\ell k}^2(1-\gamma_{i\ell}^2) + 2(1-\rho)\phi_{\ell k}(1-\phi_{\ell k})\gamma_{i\ell}(1-\gamma_{i\ell})} \\ &= \frac{\rho(1+\gamma_{i\ell})}{\rho(1+\gamma_{i\ell}) + (1-\rho)(2\gamma_{i\ell} - \phi_{\ell k}\gamma_{i\ell} + \phi_{\ell k})}. \end{aligned} \tag{2.9}$$

With the values of $\mathbb{P}(g_{i\ell}, s_{i\ell}|w_{i\ell} = A_{\ell k}A_{\ell k}, \Psi)$, the posterior probabilities of $g_{i\ell}$ and $s_{i\ell}$ can be easily calculated with eqs. 2.10 and 2.11, respectively. Results appear in Tables 2.4 and 2.5.

$$\mathbb{P}(g_{i\ell}|w_{i\ell}, \Psi) = \sum_{s_{i\ell}=0}^{1}\mathbb{P}(g_{i\ell}, s_{i\ell}|w_{i\ell}, \Psi), \tag{2.10}$$

$$\mathbb{P}(s_{i\ell}|w_{i\ell}, \Psi) = \sum_{g_{i\ell}}\mathbb{P}(g_{i\ell}, s_{i\ell}|w_{i\ell}, \Psi). \tag{2.11}$$

The posterior probabilities of $d_{i\ell}$ given $w_{i\ell}$ and Ψ appear in Table 2.6, and they are obtained by

$$\mathbb{P}(d_{i\ell}|w_{i\ell}, \Psi) = \frac{\mathbb{P}(d_{i\ell}, w_{i\ell}|\Psi)}{\mathbb{P}(w_{i\ell}|\Psi)} = \frac{\mathbb{P}(d_{i\ell}, w_{i\ell}|\Psi)}{\sum_{d_{i\ell}=0}^{2}\mathbb{P}(d_{i\ell}, w_{i\ell}|\Psi)}. \tag{2.12}$$

Here,

$$\begin{aligned}\mathbb{P}(d_{i\ell}, w_{i\ell}|\Psi) &= \sum_{g_{i\ell}}\mathbb{P}(d_{i\ell}, w_{i\ell}, g_{i\ell}|\Psi) \\ &= \sum_{g_{i\ell}}\mathbb{P}(d_{i\ell}, w_{i\ell}|g_{i\ell}, \gamma_{i\ell})\mathbb{P}(g_{i\ell}|\Phi, \rho) \\ &= \sum_{g_{i\ell}}\mathbb{P}(w_{i\ell}|g_{i\ell}, d_{i\ell})\mathbb{P}(d_{i\ell}|\gamma_{i\ell})\mathbb{P}(g_{i\ell}|\Phi, \rho). \end{aligned} \tag{2.13}$$

Therefore, $\mathbb{E}[n_{\ell k}|W,\Psi]$, $\mathbb{E}[d_{i\cdot}|W,\Psi]$, $\mathbb{E}[d_{\cdot\ell}|W,\Psi]$, and $\mathbb{E}[s|W,\Psi]$ are calculated as

$$\mathbb{E}[n_{\ell k}|W,\Psi] = \sum_{i=1}^{N}\sum_{g_{i\ell}}\sum_{s_{i\ell}=0}^{1} f(A_{\ell k}|g_{i\ell},s_{i\ell})\mathbb{P}(g_{i\ell},s_{i\ell}|w_{i\ell},\Psi), \tag{2.14}$$

$$\mathbb{E}[d_{i\cdot}|W,\Psi] = \sum_{\ell=1}^{L}\sum_{d_{i\ell}=0}^{2} d_{i\ell}\mathbb{P}(d_{i\ell}|w_{i\ell},\Psi)(\gamma_{i\cdot}/\gamma_{i\ell}), \tag{2.15}$$

$$\mathbb{E}[d_{\cdot\ell}|W,\Psi] = \sum_{i=1}^{N}\sum_{d_{i\ell}=0}^{2} d_{i\ell}\mathbb{P}(d_{i\ell}|w_{i\ell},\Psi)(\gamma_{\cdot\ell}/\gamma_{i\ell}), \tag{2.16}$$

$$\mathbb{E}[s|W,\Psi] = \sum_{i=1}^{N}\sum_{\ell=1}^{L}\sum_{s_{i\ell}=0}^{1} s_{i\ell}\mathbb{P}(s_{i\ell}|w_{i\ell},\Psi), \tag{2.17}$$

in which $f(A_{\ell k}|g_{i\ell},s_{i\ell})$ indicates the number of independent copies of allele $A_{\ell k}$ in genotype $g_{i\ell}$ given the IBD state $s_{i\ell}$, as defined below:

$$f(A_{\ell k}|g_{i\ell},s_{i\ell}) = \begin{cases} 2 & \text{if } g_{i\ell} = A_{\ell k}A_{\ell k} \text{ and } s_{i\ell} = 0 \\ 1 & \text{if } g_{i\ell} = A_{\ell k}A_{\ell k} \text{ and } s_{i\ell} = 1 \\ 1 & \text{if } g_{i\ell} = A_{\ell k}A_{\ell h}\ (h \neq k) \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

In the M-step of the EM algorithm, we update the estimation of parameters Ψ by

$$\phi_{\ell k} = \mathbb{E}[n_{\ell k}|W,\Psi] / \sum_{h=1}^{K_\ell}\mathbb{E}[n_{\ell h}|W,\Psi] \quad \text{for } k = 1,2,\ldots,K_\ell \text{ and } \ell = 1,2,\ldots,L, \tag{2.19}$$

$$\gamma_{i\cdot} = \mathbb{E}[d_{i\cdot}|W,\Psi]/(2L) \quad \text{for } i = 1,2,\ldots,N, \tag{2.20}$$

$$\gamma_{\cdot\ell} = \mathbb{E}[d_{\cdot\ell}|W,\Psi]/(2N) \quad \text{for } \ell = 1,2,\ldots,L, \tag{2.21}$$

$$\rho = \mathbb{E}[s|W,\Psi]/(NL). \tag{2.22}$$

Justification of these expressions appears in *Appendix A*. With the updated parameter values, we calculate the likelihood $\mathbb{P}(W|\Psi)$ using eq. 2.7 and then repeat the E-step and M-step. The likelihood is guaranteed to increase after each itera-

tion in this EM process and will converge to a maximum (e.g. *Lange*, 2002); the estimated parameter values are MLEs if this maximum is the global maximum. To lower the chance of convergence only to a local maximum, we repeat our EM algorithm with 100 sets of initial values of Ψ. For each set, the allele frequencies, $\Phi = \{\phi_{\ell k} : k = 1, 2, \ldots, K_\ell; \ell = 1, 2, \ldots, L\}$, are sampled independently at different loci from Dirichlet distributions, $Dir(1_{(1)}, 1_{(2)}, \ldots, 1_{(K_\ell)})$ for locus ℓ; the sample-specific dropout rates $\gamma_{i\cdot}$ $(i = 1, 2, \ldots, N)$, the locus-specific dropout rates $\gamma_{\cdot\ell}$ $(\ell = 1, 2, \ldots, L)$, and the inbreeding coefficient ρ are independently sampled from the uniform distribution $U(0, 1)$. An EM replicate is considered to be "converged" if the increase of the log-likelihood $\log_{10} \mathbb{P}(W|\Psi)$ in one iteration is less than 10^{-4}; when this condition is met, we terminate the iteration process. The parameter values that generate the highest likelihood among the 100 EM replicates are chosen as our estimates.

2.5 Imputation procedure

To correct the bias caused by allelic dropout in estimating the observed heterozygosity and other quantities, we create 100 imputed data sets by drawing genotypes from the posterior probability $\mathbb{P}(G|W, \hat{\Psi}) = \mathbb{P}(G|W, \hat{\Phi}, \hat{\Gamma}, \hat{\rho})$, in which $\hat{\Phi}$, $\hat{\Gamma}$, and $\hat{\rho}$ are the MLEs of Φ, Γ, and ρ, and $\mathbb{P}(G|W, \Psi)$ is specified in eq. 2.10 and Table 2.4. In using this strategy, we not only impute the missing genotypes but also replace some of the observed homozygous genotypes with heterozygotes, as it is possible that observed homozygous genotypes represent false homozygotes resulting from allelic dropout. This imputation strategy accounts for the genotype uncertainty that allelic dropout introduces.

2.6 Application to Native American data

We found that in sequential observations of the likelihood of the estimated parameter values, our EM algorithm converged quickly for all 100 sets of initial values for Φ, Γ and ρ (results not shown). For each of the 100 sets, the EM algorithm reached the convergence criterion within 300 iterations. The difference in the estimated parameter values among the 100 replicates was minimal after convergence, indicating that the method was not sensitive to the initial values (results not shown).

Histograms of the estimated sample-specific dropout rates $\hat{\gamma}_{i\cdot}$ and the estimated locus-specific dropout rates $\hat{\gamma}_{\cdot\ell}$ appear in Fig. 2.4. The mean of the $\hat{\gamma}_{i\cdot}$ is 0.094, and for most individuals, $\hat{\gamma}_{i\cdot} < 0.1$ (Fig. 2.4A). The maximum $\hat{\gamma}_{i\cdot}$ is 0.405; this high rate indicates that some samples have low quantity or quality, and is compatible with the fact that some of the samples are relatively old. Samples from some populations, such as Arctic/Subarctic 1 and Central Mexico 2, have higher overall quality, as reflected in low estimated sample-specific dropout rates.

Compared to the sample-specific dropout rates, the estimated locus-specific dropout rates are much smaller, with mean 0.036 and maximum 0.160 (Fig. 2.4B). The large spread of the $\hat{\gamma}_{i\cdot}$ compared to the small values of the $\hat{\gamma}_{\cdot\ell}$ is consistent with the observation that the positive correlation between missing data and homozygotes is much greater across individuals than across loci (Fig. 2.2).

The estimated inbreeding coefficient is $\hat{\rho} = 0$, the minimum possible value, smaller than the positive values typical of human populations. Several explanations could potentially explain the estimate of 0. First, our samples might be close to HWE. Second, our method might systematically underestimate the inbreeding coefficient, a hypothesis that we test below using simulations. Third, genotyping errors other than allelic dropout, such as genotype miscalling, can potentially also contribute to the underestimation. We use simulations to examine this hypothesis as well.

In a given individual, the L loci can be divided into three classes according to the

observed genotypes: n_{hom} homozygous loci, n_{het} heterozygous loci, and $L-n_{hom}-n_{het}$ loci that have both allelic copies missing. For each individual, we calculated the observed heterozygosity as $H_o = n_{het}/(n_{hom}+n_{het})$, as shown by gray points in Fig. 2.4C. High variation exists in H_o for different individuals, and the mean H_o across individuals is 0.590 (standard deviation 0.137). The observed heterozygosities are negatively correlated with the MLEs of the sample-specific dropout rates (Fig. S2.1), as is expected from the underestimation of heterozygosity caused by allelic dropout. Averaging the estimated observed heterozygosity over 100 imputed data sets, we see that variation across individuals in estimated heterozygosities is reduced compared to the values estimated directly from the observed genotypes, and the mean heterozygosity increases to 0.730 (standard deviation 0.035, Fig. 2.4C). The estimated individual heterozygosity does not vary greatly across different imputed data sets (standard deviation 0.014, averaging across all individuals).

2.7 Simulations

We perform three sets of simulations to examine the performance of our method. First, we consider simulations that assume that the model assumptions hold, using as true values the estimated parameter values from the Native American data set (Experiment 1). Next, we consider simulations that do not satisfy the model assumptions, by inclusion of population structure (Experiment 2) and genotyping errors not resulting from allelic dropout (Experiment 3). These latter simulations examine the robustness of the estimation procedure to model violations.

2.7.1 Simulation methods

To generate simulated allelic dropout rates for use in Experiments 2 and 3, we first fit the distributions of the estimated sample-specific and locus-specific dropout rates from the Native American data using beta distributions $Beta(\alpha, \beta)$. Denote the sam-

ple mean and sample variance of the MLEs of the sample-specific (or locus-specific) dropout rates as m and v, respectively. We estimated α and β using the method of moments, with $\hat{\alpha} = m[m(1-m)/v - 1]$ and $\hat{\beta} = (1-m)[m(1-m)/v - 1]$ (*Casella and Berger*, 2001). The estimated sample-specific and locus-specific dropout rates approximately follow $Beta(0.55, 5.30)$ and $Beta(1.00, 27.00)$, respectively (Figs. 2.4A and 2.4B).

2.7.1.1 Experiment 1. Native American data

We simulate data under model assumptions 2-5 with parameter values estimated from the actual Native American data (results from *Application to Native American data*). The simulation procedure appears in Fig. 2.5A. Suppose $\hat{\Phi}$, $\hat{\Gamma}$ and $\hat{\rho}$ are the MLEs of Φ, Γ, and ρ estimated from the data. First, we draw the true genotypes $\tilde{G}$ using probabilities specified by eq. 2.5, assuming that the allele frequencies are given by $\hat{\Phi}$ and the inbreeding coefficient by $\hat{\rho}$. Next, we simulate the dropout state $\tilde{Z}$ by randomly dropping out copies with probability specified by eq. 2.1, independently across alleles, loci, and individuals. Using $\tilde{G}$ and $\tilde{Z}$, we then obtain our simulated observed genotypes $\tilde{W}$. This simulation approach does not guarantee that model assumption 1 will hold, because some alleles might not be observed owing either to allelic dropout or to a stochastic failure to be drawn in the simulation. We simulate one set of genotypes at $L = 343$ loci for $N = 152$ individuals.

2.7.1.2 Experiment 2. Data with population structure

To test our method in a setting in which genotypes are taken from a structured population, we simulate data for two subpopulations with equal sample size ($N_1 = N_2 = 76$), genotyped at the same set of loci ($L = 343$). We then apply our method on the combined data set, disregarding the population structure. The procedure appears in Fig. 2.5B.

First, we use the F-model (*Falush et al.*, 2003) to generate allele frequencies for two populations that have undergone a specified level of divergence from a common ancestral population. We use the MLEs of the allele frequencies of the 343 loci in the Native American data (results from *Application to Native American data*) as the allele frequencies of the ancestral population, $\Phi^{(A)} = \hat{\Phi}$. Denote the estimated allele frequencies at locus ℓ by a vector $\hat{\phi}_{\ell\cdot}$. Under the F-model, allele frequencies of locus ℓ for population 1, $\phi^{(1)}_{\ell\cdot}$, and for population 2, $\phi^{(2)}_{\ell\cdot}$, are independently sampled from the Dirichlet distribution $Dir(\frac{1-F}{F}\hat{\phi}_{\ell\cdot})$, in which F is a parameter constant across loci that describes the divergence of the descendant populations from the ancestral population. F can differ for populations 1 and 2, but for simplicity, we set F to the same value for both populations. Using eqs. B1-B2 in *Appendix B* and the independence of $\phi^{(1)}_{\ell k}$ and $\phi^{(2)}_{\ell k}$, the squared difference of allele frequencies between the two populations satisfies $\mathbb{E}[(\phi^{(1)}_{\ell k} - \phi^{(2)}_{\ell k})^2] = 2F\hat{\phi}_{\ell k}(1 - \hat{\phi}_{\ell k})$, which is linearly proportional to F. In the limit as $F \to 0$, we get $\phi^{(1)}_{\ell\cdot} = \phi^{(2)}_{\ell\cdot} = \hat{\phi}_{\ell\cdot}$ for each ℓ, so that no divergence exists between either descendant population and the ancestral population.

We choose six values of F (0, 0.04, 0.08, 0.12, 0.16, and 0.20) in different simulations. For each value, we first generate allele frequencies, $\Phi^{(1)}$ and $\Phi^{(2)}$, at all 343 loci for populations 1 and 2. Next, we draw genotypes separately for each population according to the genotype frequencies in eq. 2.5, with the same value of the inbreeding coefficient ρ. We consider 16 values for ρ, ranging from 0 to 0.15 in increments of 0.01. In total, we generate $6 \times 16 = 96$ sets of simulated genotypes with different combinations of settings for F and ρ (although for ease of presentation, some plots show only 36 of the 96 cases). Last, we simulate allelic dropout on each of the simulated genotype data sets using $\gamma_{i\cdot}$ and $\gamma_{\cdot\ell}$ sampled independently from a $Beta(\alpha, \beta)$ distribution, in which $\alpha = 0.55$ and $\beta = 5.30$ are estimated from the MLEs of sample-specific dropout rates of the Native American data (Fig. 2.4A). We do not use the estimated α and β from the MLEs of locus-specific dropout rates because these MLEs lie in a

relatively small range (Fig. 2.4B) that would not permit simulation of high dropout rates for testing our method. Instead, use of the same beta distribution estimated from the sample-specific dropout rates produces a greater spread in the values of the simulated "true" locus-specific dropout rates, providing a more complete evaluation.

2.7.1.3 Experiment 3. Data with other genotyping errors

In our third experiment, we simulate data with stochastic genotyping errors other than allelic dropout. The simulation procedure appears in Fig. 2.5C. Each simulated data set contains a single population of $N = 152$ individuals genotyped for $L = 343$ loci. True genotypes are drawn with probabilities calculated from eq. 2.5, with allele frequencies Φ chosen as the maximum likelihood estimated frequencies from the Native American data, and the inbreeding coefficient ρ ranging from 0 to 0.15 incremented in units of 0.01 for different simulated data sets. Next, we simulate genotyping errors using a simple error model, in which at a K-allele locus in the simulated true genotypes, any allele can be mistakenly assigned as any one of the other $K - 1$ alleles, each with the same probability of $e/(K - 1)$. The parameter e specifies the overall error rate from sources other than allelic dropout, such as genotype miscalling and data entry errors (e.g. *Wang*, 2004; *Johnson and Haydon*, 2007). We consider six values for e (0, 0.02, 0.04, 0.06, 0.08, and 0.10), such that we simulate 96 ($= 6 \times 16$) data sets with different combinations of e and ρ. In the last step, as in Experiment 2, we simulate allelic dropout in each data set with both sample-specific and locus-specific dropout rates independently sampled from a $Beta(0.55, 5.30)$ distribution.

2.7.2 Simulation results

2.7.2.1 Experiment 1. Native American data

Because we simulate under assumptions 2-5 with parameter values estimated from the real data, we expect that if our model is correctly specified, the simulated data can capture patterns observed in the real data. By comparing plots of the fraction of missing data versus the fraction of homozygotes in the real and simulated data (Figs. 2.2 and 2.6), we can see that our simulated data effectively capture the observed positive correlation across individuals and the lack of correlation across loci observed from the real data. For the simulated data set, the Pearson correlation coefficient between the fraction of missing genotypes and the fraction of homozygotes is $r = 0.900$ ($P < 0.0001$) across individuals and $r = 0.143$ ($P = 0.0045$) across loci. We can also compare the observed heterozygosity for the simulated data (purple points in Fig. 2.7C) and the real data (gray points in Fig. 2.4C). The simulated data again reproduce the pattern of variation among individual heterozygosities observed in the real data. These two empirical comparisons display the similarity between the real data and the data simulated on the basis of estimates obtained from the real data, and thus support the validity of the allelic dropout mechanism specified in our model.

We can formally compare the estimated dropout rates for the simulation with the true dropout rates $\tilde{\Gamma}$ specified by the MLEs of the dropout rates for the Native American data. Fig. 2.7A shows that our method accurately estimates the sample-specific dropout rates for all 152 individuals (mean squared error 2.6×10^{-4}). The estimated locus-specific dropout rates are also close to their true values, but with a slightly higher mean squared error of 5.2×10^{-4} (Fig. 2.7B). This difference between the estimation of sample-specific and locus-specific dropout rates can be explained by the fact that the number of loci ($L = 343$) is more than twice the number of individuals ($N = 152$). Consequently, more information is available for estimating a sample-

specific rather than a locus-specific dropout rate. For the inbreeding coefficient ρ, our estimated value is 1.7×10^{-5}, close to the true value of 0 that we used to generate the simulated genotypes.

Finally, in Fig. 2.7C, we can see that our method successfully corrects the bias in estimating heterozygosity from the simulated data. The true observed heterozygosity is calculated using the true genotypes $\tilde{G}$, and has mean 0.716, averaging across all individuals. The mean estimated observed heterozygosity, obtained from the observed uncorrected genotypes $\tilde{W}$, is 0.565, lower than the true value. With imputed data sets, we obtain corrected heterozygosities that are close to the true values. The mean and standard deviation of the corrected heterozygosities, evaluated from 100 imputed data sets and averaged across individuals, are 0.715 and 0.014 respectively. The low standard deviation across different imputed data sets indicates that our imputation strategy is relatively robust in correcting the underestimation of observed heterozygosity.

2.7.2.2 Experiment 2. Data with population structure

To further test the robustness of our method, we applied our method to 96 simulated data sets with different levels of population structure (parameterized by F) and inbreeding (parameterized by ρ). In Figs. 2.8A and 2.8C, we compare the estimated dropout rates to their true values. Considering the 36 simulated data sets that are displayed, our method accurately estimates both the sample-specific and the locus-specific dropout rates. The accuracy of our estimates is then quantified by mean squared errors for each simulated data set separately, as displayed in Figs. 2.8B and 2.8D. The performance in estimating the sample-specific dropout rate is not greatly affected by either the degree of population structure or the level of inbreeding (Fig. 2.8B). By contrast, while the mean squared error of the estimated locus-specific dropout rates is roughly constant for different levels of inbreeding, it increases with

the degree of population structure (Fig. 2.8D).

One possible explanation for this observation is that the accuracy of allelic dropout estimates is closely related to the accuracy of the estimated allele frequencies. This accuracy may decrease as the level of population structure increases, because we do not incorporate population structure in our model for estimation. The estimation of locus-specific dropout rates is more sensitive to inaccurate estimates of allele frequencies because the estimated accuracy of a locus-specific rate relies on the estimation of allele frequencies at that particular locus. By contrast, a sample-specific dropout rate is obtained by averaging the expected number of sample-specific dropouts across all loci in an individual, and is less dependent on the accuracy of estimated allele frequencies at any particular locus. Therefore, sample-specific dropout rate estimates are less sensitive to population structure than are locus-specific estimates. When $F = 0$, with no population structure, the difference between the mean squared error for the sample-specific and locus-specific rates arises simply from differences in the numbers of loci and individuals, as discussed for Experiment 1.

Fig. 2.8E shows the estimated inbreeding coefficient for all 96 simulated data sets, compared to the simulated true inbreeding coefficient in the subpopulations. With $F = 0$, a scenario for which no population structure exists and the data are generated under model assumptions 2-5, our method tends to slightly underestimate the inbreeding coefficient. As F increases, the estimate becomes greater than the simulated inbreeding coefficient (Fig. 2.8F). This result is consistent with our expectation, because according to the Wahlund effect (e.g. *Hartl and Clark*, 1997), a pooled population consisting of two subpopulations is expected to have more homozygous genotypes than an unstructured population, resulting in a pattern similar to that caused by a higher level of inbreeding within the unstructured population. Indeed, with no allelic dropout, a structured population under the F model has identical expected allele frequencies and genotype frequencies to an unstructured population

with a higher inbreeding coefficient $\rho^* = \rho + (1-\rho)[F/(2-F)]$ (*Appendix B*). By comparing our estimated inbreeding coefficient $\hat{\rho}$ with the "effective inbreeding coefficient" ρ^* (dashed lines in Fig. 2.8E), we find that most of our estimated inbreeding coefficients are slightly smaller than the corresponding ρ^*, indicating that the MLE of ρ is biased downward. It is worth noting that with a single parameter ρ, we capture the deviation of genotype frequencies from HWE introduced by population structure, thereby obtaining accurate estimated allelic dropout rates without explicitly incorporating population structure in our model.

We applied the imputation procedure to correct the bias in estimating heterozygosity for each of the 96 simulated data sets. Similarly to our application in Experiment 1, we calculated the uncorrected and true heterozygosities for each individual from the simulated observed genotypes $\tilde{W}$ and the simulated true genotypes $\tilde{G}$, respectively. The corrected heterozygosity was averaged across 100 imputed data sets for each simulated data set. Results for 36 simulated data sets appear in Fig. S2.2, in which heterozygosities were averaged across all individuals in each data set. Our results show a significant improvement of the corrected heterozygosity over the uncorrected heterozygosity in all simulations, in that the corrected heterozygosity is considerably closer to the true heterozygosity. This improvement is fairly robust to the presence of population structure.

2.7.2.3 Experiment 3. Data with other genotyping errors

This set of simulations tested our method at different levels of genotyping error from sources other than allelic dropout. In all simulated data sets, with genotyping error ranging from 0 to 10% and ρ ranging from 0 to 0.15, our method is successful in estimating both sample-specific and locus-specific dropout rates (Figs. 2.9A and 2.9C). The estimation accuracy of dropout rates is not strongly affected by the genotyping error rate (Figs. 2.9B and 2.9D). We can again see that a smaller number of

individuals than loci has led to higher mean squared error for estimated locus-specific rates (Fig. 2.9D) than for sample-specific rates (Fig. 2.9B).

Similar to the $F = 0$ case in our simulations with population structure, the simulated data sets with no genotyping error ($e = 0$) are generated under model assumptions 2-5. Consistently with the results for $F = 0$, our method slightly underestimates the inbreeding coefficient ρ for most simulated data sets with $e = 0$. As genotyping error increases, the underestimation also increases (Figs. 2.9E and 2.9F). This result can be explained by noting that the simulated genotyping error, which changes the allele frequencies only slightly, tends to create false heterozygotes more frequently than false homozygotes. Therefore, the observed heterozygosity is increased while the expected heterozygosity changes little, leading to a decrease in the estimated inbreeding coefficient. Although our estimation of the inbreeding coefficient ρ becomes less accurate when the genotyping error rate is higher, the underestimation of ρ does not prevent the method from accurately estimating allelic dropout rates.

For the heterozygosity, the corrected values obtained using our imputation strategy are closer to the true values than are the uncorrected values directly obtained from the observed genotypes (Fig. S2.3). However, as the genotyping error rate e increases, our method starts to overcorrect the downward bias in estimating the observed heterozygosity, and the corrected values exceed the true values. Similarly to our explanation for the underestimation of the inbreeding coefficient, this overcorrection is introduced by the simulated genotyping error, which creates an excess of false heterozygotes. This excess is in turn incorporated into the corrected estimates of heterozygosity, because we do not model genotyping errors other than those due to allelic dropout.

2.8 Discussion

In this study, we have developed a maximum likelihood approach to jointly estimate sample-specific dropout rates, locus-specific dropout rates, allele frequencies, and the inbreeding coefficient from only one nonreplicated set of microsatellite genotypes. Our algorithm can accurately recover the allelic dropout parameters, and an imputation strategy using the method provides an alternative to ignoring high empirical missing data rates or excluding samples and loci with large amounts of missing data. Investigators can then use the imputed data in subsequent analyses, such as in studies of genetic diversity or population structure, or in software that disallows missing values in the input data. We have demonstrated our approach using extensive analyses of an empirical data set and data sets simulated using parameter values chosen on the basis of the empirical example.

We have found that our method works well on simulated data. In particular, it performs well in estimating the sample-specific dropout rates $\gamma_{i\cdot}$ and locus-specific dropout rates $\gamma_{\cdot\ell}$. Further, in the examples we have considered, it is reasonably robust to violations of the model assumptions owing to the existence of population structure or genotyping error other than allelic dropout. This robustness arises partly from the inclusion of the inbreeding coefficient ρ in our model, which enables us to capture the deviation from HWE caused by multiple factors, such as true inbreeding, population structure, and genotyping errors. Because the various sources of deviation from HWE are incorporated into the single parameter ρ, the estimation of ρ itself is more sensitive to violation of model assumptions; therefore, it is important to be careful when interpreting the estimated value of ρ, as it may reflect phenomena other than inbreeding. When data are simulated under our model, such as in the cases of $F = 0$ and $e = 0$, our method tends to slightly underestimate ρ (Figs. 2.8E and 2.9E), indicating that our MLEs are biased, at least for the inbreeding coefficient.

We can use simulation approaches to further explore the statistical properties

of our estimates. To examine the consistency of the estimators, we performed two additional sets of simulations, in which we generated genotype data under our model with either different numbers of individuals N or different numbers of loci L (*Appendix C*). When L is fixed, although estimates of the sample-specific dropout rates $\gamma_{i\cdot}$ are not affected by the value of N, our estimates of the locus-specific dropout rates $\gamma_{\cdot\ell}$ and the inbreeding coefficient ρ become closer to the true values as N increases (Fig. S2.4). When N is sufficiently large (e.g. $N = 1600$), the estimates of $\gamma_{\cdot\ell}$ and ρ are almost identical to the true values. If we instead fix N and increase L, then the estimates of $\gamma_{i\cdot}$ and ρ eventually approach the true values, while the estimates of $\gamma_{\cdot\ell}$ remain unaffected (Fig. S2.5). These results suggest, without a strict analytical proof, that our MLEs of the dropout rates and inbreeding coefficient are consistent.

For the Native American data, we can compare the estimated heterozygosities under our model with other data on similar populations. *Wang et al.* (2007b) studied microsatellites in 29 Native American populations, including eight populations from regions that overlap those considered in our data. We reanalyzed these populations, three from Canada and five from Mexico, by calculating observed heterozygosity H_o from the same 343 loci as were genotyped in our data. We obtained a mean H_o of 0.670 with standard deviation 0.051 across 176 individuals in the pooled set of eight populations. In comparison, mean H_o across our 152 Native American samples is 0.590 (standard deviation 0.137) before correcting for allelic dropout, substantially lower than in *Wang et al.* (2007b), and it is 0.730 (standard deviation 0.035) after correcting for allelic dropout, higher than in *Wang et al.* (2007b). Several possible reasons can explain the imperfect agreement between our corrected heterozygosity and the estimate on the basis of the *Wang et al.* (2007b) data. First, the sets of populations might differ in such factors as the extent of European admixture, so that they might truly differ in underlying heterozygosity. Second, the *Wang et al.* (2007b) data might have some allelic dropout as well, so that our H_o estimates from those

data underestimate the true values. Third, our method might have overcorrected the underestimation of H_o; our simulations show that because we do not model genotyping errors from sources other than allelic dropout, the existence of such errors can lead to overestimation of H_o (Fig. S2.3). It is also possible that missing genotypes caused by factors other than allelic dropout could have been erroneously attributed to allelic dropout, leading to overestimation of dropout rates, and hence, to overcorrection of H_o.

Our model assumes that all individuals are sampled from the same population with one set of allele frequencies, and that inbreeding is constant across individuals and loci. We applied this assumption to the whole Native American data set as an approximation. However, evidence of population structure can be found by applying multidimensional scaling analysis to the Native American samples. As shown in Fig. S2.6, individuals from different populations tend to form different clusters, indicating that underlying allele frequencies and levels of inbreeding differ among populations. Although our simulations have found that estimation of allelic dropout rates is robust to the existence of population structure, estimation of allele frequencies and the inbreeding coefficient can become less accurate in structured populations. It would therefore have been preferable in our analysis to apply our method on each population instead of on the pooled data set; however, such an approach was impractical owing to the small sample sizes in individual population. To address this problem, it might be possible to directly incorporate population structure into our model (e.g. *Falush et al.*, 2003), thereby enabling allele frequencies and inbreeding coefficients to differ across the subpopulations in a structured data set. Further, because samples from the same population are typically collected and genotyped as a group, full modeling of the population structure might allow for a correlation in dropout rates across individuals within a population.

An additional limitation of our approach is that during data analysis, we do not

take into account the uncertainty inherent in estimating parameters. We first obtain the MLEs of allele frequencies $\hat{\Phi}$, allelic dropout rates $\hat{\Gamma}$, and the inbreeding coefficient $\hat{\rho}$, and then create imputed data sets by drawing genotypes using $\hat{\Phi}$, $\hat{\Gamma}$, and $\hat{\rho}$. This procedure is "improper" because it does not propagate the uncertainty inherent in parameter estimation (*Little and Rubin*, 2002). To obtain "proper" estimates, instead of using an EM algorithm to find the MLEs of the parameters, we could potentially use a Gibbs sampler or other Bayesian sampling methods to sample parameter values, and then create imputed data sets using these sampled parameter sets. In such approaches, parameters sampled from their underlying distributions would be used for different imputations, instead of using the same MLEs for all imputations.

Finally, we have not compared our approach with methods that rely on replicate genotypes. While we expect that replicate genotypes will usually lead to more accurate estimates of model parameters, our method provides a general approach that is relatively flexible and accurate in the case that replicates cannot be obtained. Compared with existing models that assume HWE (e.g. *Miller et al.*, 2002; *Johnson and Haydon*, 2007), our model uses a more general assumption of inbreeding, and we also incorporate both sample-specific and locus-specific dropout rates. The general model increases the applicability of our method for analyzing diverse genotype data sets, such as those that have significant dropout caused by locus-specific factors (e.g. *Buchan et al.*, 2005). It is worth noting that HWE is the special case of $\rho = 0$ in our inbreeding model; when it is sensible to assume HWE, we can simply initiate the EM algorithm with a value of $\rho = 0$. This choice restricts the search for MLEs to the $\rho = 0$ parameter subspace, because eq. 2.22 stays fixed at 0 in each EM iteration. Similarly, if we prefer to only consider sample-specific dropout rates (or only locus-specific dropout rates), then we can simply set the initial values of $\gamma_{\cdot\ell}$ to 0 for all loci (or initial values of $\gamma_{i\cdot}$ to 0 for all individuals). These choices also restrict the search to subspaces of the full parameter space. We have implemented these options

in our software program *MicroDrop*, which provides flexibility for users to analyze their data with a variety of different assumptions.

2.9 Acknowledgements

We are grateful to Dr. Roderick Little for helpful advice, Michael DeGiorgio for comments on a draft of the manuscript, and Zachary Szpiech for help in evaluating and testing the *MicroDrop* software. We also thank two anonymous reviewers for constructive comments that have led to substantial improvement of this article. This work was supported by National Institutes of Health grants R01 GM081441 and R01 HG005855, by the Burroughs Wellcome Fund, and by a Howard Hughes Medical Institute International Student Research Fellowship.

2.10 Appendix A. The EM algorithm

The main text describes an EM algorithm for estimating parameters in our model. Here, we provide the derivation of eqs. 2.19-2.22 for parameter updates in each EM iteration. We start from a general description of the EM algorithm (e.g. *Casella and Berger*, 2001; *Lange*, 2002).

To obtain the maximum likelihood estimates (MLEs), our goal is to maximize the likelihood $\mathcal{L} = \mathbb{P}(W|\Psi)$. Because $\mathcal{L}$ is difficult to maximize directly, we use an EM algorithm to replace the maximization of $\mathcal{L}$ with a series of simpler maximizations. We introduce three sets of latent variables: the true genotypes G, IBD states S, and dropout states Z, each representing an $N \times L$ matrix. Instead of directly working on likelihood $\mathcal{L}$, the EM algorithm starts with a set of initial values arbitrarily chosen for Ψ, and in each of a series of iterations, maximizes the Q function defined by eq. A1. This iterative maximization is easier and sequentially increases the value of $\mathcal{L}$ (e.g. *Lange*, 2002), so that the parameters eventually converge to values at a

maximum of $\mathcal{L}$.

In the E-step of iteration $t+1$, we want to calculate the following expectation:

$$Q(\Psi|\Psi^{(t)}) = \mathbb{E}_{G,S,Z|W,\Psi^{(t)}}[\ln \mathbb{P}(W,G,S,Z|\Psi)]. \tag{A1}$$

This computation is equivalent to calculating $\mathbb{E}[G|W,\Psi^{(t)}]$, $\mathbb{E}[S|W,\Psi^{(t)}]$ and $\mathbb{E}[Z|W,\Psi^{(t)}]$, and then inserting these quantities into the expression for $\ln \mathbb{P}(W,G,S,Z|\Psi)$, such that eq. A1 is a function of parameters $\Psi = \{\Phi, \Gamma, \rho\}$. In the M-step, the parameters are updated with values $\Psi^{(t+1)}$ that maximize eq. A1. The explicit expression for eq. A1 is cumbersome, but given the dependency described in Fig. 2.3, we can greatly simplify our EM algorithm by a decomposition of $\mathbb{P}(W,G,S,Z|\Psi)$:

$$\begin{aligned}
\mathbb{P}(W,G,S,Z|\Psi) &= \mathbb{P}(G,S|\Psi)\mathbb{P}(Z|G,S,\Psi)\mathbb{P}(W|Z,G,S,\Psi) \\
&= \mathbb{P}(G,S|\Phi,\rho)\mathbb{P}(Z|\Gamma)\mathbb{P}(W|Z,G) \\
&\propto \mathbb{P}(G,S|\Phi,\rho)\mathbb{P}(Z|\Gamma).
\end{aligned} \tag{A2}$$

Eq. A2 implies that we can maximize $\mathbb{E}_{G,S|W,\Psi^{(t)}}[\ln \mathbb{P}(G,S|\Phi,\rho)]$ and $\mathbb{E}_{Z|W,\Psi^{(t)}}[\ln \mathbb{P}(Z|\Gamma)]$ separately in order to maximize $Q(\Psi|\Psi^{(t)})$ (eq. A1). Further, it can be shown that $n_{\ell k}$, $d_{i\cdot}$, $d_{\cdot\ell}$, and s are sufficient statistics for $\phi_{\ell k}$, $\gamma_{i\cdot}$, $\gamma_{\cdot\ell}$ and ρ, respectively. Therefore, in the E-step, we can simply calculate the expectations of these four sets of statistics (eqs. 2.14-2.17) rather than evaluating the full matrices $\mathbb{E}[G|W,\Psi]$, $\mathbb{E}[S|W,\Psi]$, $\mathbb{E}[Z|W,\Psi]$.

In the M-step, the dropout rates Γ are updated by maximizing $\mathbb{E}_{Z|W,\Psi^{(t)}}[\ln \mathbb{P}(Z|\Gamma)]$, resulting in eqs. 2.20 and 2.21, quantities that can be obtained intuitively by considering each dropout as an independent Bernoulli trial. The allele frequencies Φ and the inbreeding coefficient ρ are updated by maximizing $\mathbb{E}_{G,S|W,\Psi^{(t)}}[\ln \mathbb{P}(G,S|\Phi,\rho)]$, resulting in eqs. 2.19 and 2.22 after some algebra. As an example, we show the derivation of eqs. 2.19 and 2.22 for a single biallelic locus ($L=1$, $K_\ell = 2$).

Denote the alleles by A_1 and A_2, and the corresponding allele frequencies by ϕ_1 and ϕ_2, with $\phi_1 + \phi_2 = 1$. Suppose that in the whole data set, $x_{hk,u}$ individuals have true genotype A_hA_k ($1 \le h \le k \le 2$) and IBD state u ($u = 0$ or 1). Then $\mathbb{P}(G, S|\Phi, \rho)$ can be written as

$$\begin{aligned}
\mathbb{P}(G, S|\Phi, \rho) &= \prod_{h=1}^{2}\prod_{k=h}^{2}\prod_{u=0}^{1}[\mathbb{P}(A_hA_k, u|\Phi, \rho)]^{x_{hk,u}} \\
&= [(1-\rho)\phi_1^2]^{x_{11,0}}(\rho\phi_1)^{x_{11,1}}[(1-\rho)\phi_2^2]^{x_{22,0}}(\rho\phi_2)^{x_{22,1}}[(1-\rho)2\phi_1\phi_2]^{x_{12,0}} \\
&= 2^{x_{12,0}}\rho^{x_{11,1}+x_{22,1}}(1-\rho)^{x_{11,0}+x_{22,0}+x_{12,0}}\phi_1^{2x_{11,0}+x_{11,1}+x_{12,0}}\phi_2^{2x_{22,0}+x_{22,1}+x_{12,0}} \\
&\propto \rho^s(1-\rho)^{N-s}\phi_1^{n_1}(1-\phi_1)^{n_2}, \qquad \text{(A3)}
\end{aligned}$$

in which s is the total number of genotypes that are IBD ($u = 1$), and n_1 and n_2 are the numbers of independent copies for alleles A_1 and A_2, respectively. We can see from eq. A3 that s is a sufficient statistic for ρ, and n_1 and n_2 are sufficient statistics for Φ. Following eq. A3, $\mathbb{E}_{G,S|W,\Psi^{(t)}}[\ln \mathbb{P}(G, S|\Phi, \rho)]$ can be expressed as

$$\begin{aligned}
\mathbb{E}_{G,S|W,\Psi^{(t)}}[\ln \mathbb{P}(G, S|\Phi, \rho)] =& c + \mathbb{E}[s|W, \Psi^{(t)}] \ln \rho + (N - \mathbb{E}[s|W, \Psi^{(t)}]) \ln(1-\rho) \\
&+ \mathbb{E}[n_1|W, \Psi^{(t)}] \ln \phi_1 + \mathbb{E}[n_2|W, \Psi^{(t)}] \ln(1-\phi_1), \qquad \text{(A4)}
\end{aligned}$$

in which $c = \mathbb{E}[x_{12,0}|W, \Psi^{(t)}] \ln 2$ is a constant with respect to parameters ρ and Φ. To maximize $\mathbb{E}_{G,S|W,\Psi^{(t)}} \ln \mathbb{P}(G, S|\Phi, \rho)$, we can solve the following equations:

$$\frac{\partial}{\partial \rho}\mathbb{E}_{G,S|W,\Psi^{(t)}}[\ln \mathbb{P}(G, S|\Phi, \rho)] = \frac{\mathbb{E}[s|W, \Psi^{(t)}] - N\rho}{\rho(1-\rho)} = 0 \qquad \text{(A5)}$$

$$\frac{\partial}{\partial \phi_1}\mathbb{E}_{G,S|W,\Psi^{(t)}}[\ln \mathbb{P}(G, S|\Phi, \rho)] = \frac{\mathbb{E}[n_1|W, \Psi^{(t)}]}{\phi_1} - \frac{\mathbb{E}[n_2|W, \Psi^{(t)}]}{1-\phi_1} = 0. \qquad \text{(A6)}$$

The solutions for the case of $L = 1$ and $K_\ell = 2$ agree with eqs. 2.19 and 2.22:

$$\phi_1 = \mathbb{E}[n_1|W, \Psi^{(t)}]/(\mathbb{E}[n_1|W, \Psi^{(t)}] + \mathbb{E}[n_2|W, \Psi^{(t)}]) \tag{A7}$$

$$\rho = \mathbb{E}[s|W, \Psi^{(t)}]/N. \tag{A8}$$

2.11 Appendix B. Inbreeding and the F model

In the presence of population structure, the proportion of homozygotes in the pooled population exceeds that of an unstructured population, leading to a deviation from Hardy-Weinberg equilibrium similar to inbreeding. Therefore, we expect our algorithm to overestimate the inbreeding coefficient when population structure in the genotype data is not taken into account for the estimation. In this section, we derive an expression for this overestimation in a structured population under the F model (*Falush et al.*, 2003). We show that a structured population with two subpopulations, whose inbreeding coefficients are ρ_1 and ρ_2, has expected allele and genotype frequencies identical to an unstructured population with a certain inbreeding coefficient ρ^* higher than ρ_1 and ρ_2.

Consider a structured population with $N_1 = c_1 N$ and $N_2 = c_2 N = (1 - c_1)N$ individuals sampled from subpopulations 1 and 2, respectively (Fig. S2.7). Without loss of generality, we only examine a single locus with K alleles. Under the F model, the allele frequencies of subpopulation j ($j = 1, 2$), $\Phi_j = \{\phi_{j1}, \ldots, \phi_{jK}\}$, follow a Dirichlet distribution $\Phi_j \sim Dir\left(\frac{1-F_j}{F_j}\Phi_A\right)$, in which $\Phi_A = \{\phi_{A1}, \ldots, \phi_{AK}\}$ denotes the allele frequencies of a common ancestral population of the two subpopulations and F_j measures the divergence of subpopulation j from the ancestral population. We need the first and second moments of the allele frequencies Φ_j, quantities that can be obtained from the mean, variance, and covariance of a Dirichlet distribution.

For $h \neq k$,

$$\mathbb{E}[\phi_{jk}] = \phi_{Ak}, \tag{B1}$$

$$\mathbb{E}[\phi_{jk}^2] = \mathbb{E}[\phi_{jk}]^2 + \mathrm{Var}(\phi_{jk}) = \phi_{Ak}^2 + F_j\phi_{Ak}(1-\phi_{Ak}), \tag{B2}$$

$$\mathbb{E}[\phi_{jk}\phi_{jh}] = \mathbb{E}[\phi_{jk}]\mathbb{E}[\phi_{jh}] + \mathrm{Cov}(\phi_{jk},\phi_{jh}) = \phi_{Ak}\phi_{Ah}(1-F_j). \tag{B3}$$

Suppose the two subpopulations have inbreeding coefficients ρ_1 and ρ_2, respectively. Under the inbreeding model (e.g. *Holsinger and Weir*, 2009), the frequency of genotype A_kA_h in subpopulation j can be written as

$$P_{j,kh} = \begin{cases} (1-\rho_j)\phi_{jk}^2 + \rho_j\phi_{jk} & \text{if } h = k \\ 2(1-\rho_j)\phi_{jk}\phi_{jh} & \text{if } h \neq k. \end{cases} \tag{B4}$$

Using eqs. B1-B4, in the structured population, homozygote A_kA_k has expected genotype frequency

$$\begin{aligned} \mathbb{E}[P_{kk}] &= \mathbb{E}\Big[\sum_{j=1}^{2} c_j P_{j,kk}\Big] = \sum_{j=1}^{2} c_j \mathbb{E}[(1-\rho_j)\phi_{jk}^2 + \rho_j\phi_{jk}] \\ &= \phi_{Ak}\Big(1 - \sum_{j=1}^{2} c_j(1-\rho_j)(1-F_j)\Big) + \phi_{Ak}^2 \sum_{j=1}^{2} c_j(1-\rho_j)(1-F_j). \end{aligned} \tag{B5}$$

Similarly, the expected genotype frequency of heterozygote A_kA_h $(h \neq k)$ is

$$\begin{aligned} \mathbb{E}[P_{kh}] &= \mathbb{E}\Big[\sum_{j=1}^{2} c_j P_{j,kh}\Big] = \sum_{j=1}^{2} c_j \mathbb{E}[2(1-\rho_j)\phi_{jk}\phi_{jh}] \\ &= 2\phi_{Ak}\phi_{Ah} \sum_{j=1}^{2} c_j(1-\rho_j)(1-F_j). \end{aligned} \tag{B6}$$

We now search for the value of ρ^* at which genotype frequencies in an unstructured population satisfy eqs. B5 and B6. If we are unaware of the population structure,

then the allele frequencies in the pooled population are

$$\Phi^* = \sum_{j=1}^{2} c_j \Phi_j. \tag{B7}$$

Our goal is to derive an inbreeding coefficient ρ^* for an unstructured population with allele frequencies Φ^*, such that expected genotype frequencies of an unstructured population with inbreeding are identical to those of the structured population (eqs. B5-B6).

The expected genotype frequency of a homozygote $A_k A_k$ in an unstructured population with an inbreeding coefficient ρ^* can be written as

$$\begin{aligned}
\mathbb{E}[P^*_{kk}] &= \mathbb{E}[(1-\rho^*)(\phi^*_k)^2 + \rho^* \phi^*_k] \\
&= (1-\rho^*)\mathbb{E}\Big[\sum_{j=1}^{2} c_j \phi_{jk}\Big]^2 + \rho^* \mathbb{E}\Big[\sum_{j=1}^{2} c_j \phi_{jk}\Big] \\
&= \phi_{Ak}[c_1^2 F_1 + c_2^2 F_2 + \rho^*(1 - c_1^2 F_1 - c_2^2 F_2)] + \phi_{Ak}^2 (1-\rho^*)(1 - c_1^2 F_1 - c_2^2 F_2).
\end{aligned} \tag{B8}$$

For a heterozygote $A_k A_h$ $(h \neq k)$, the expected genotype frequency is

$$\begin{aligned}
\mathbb{E}[P^*_{kh}] &= \mathbb{E}[2(1-\rho^*)\phi^*_k \phi^*_h] \\
&= 2(1-\rho^*)\mathbb{E}\Big[\Big(\sum_{j=1}^{2} c_j \phi_{jk}\Big)\Big(\sum_{j=1}^{2} c_j \phi_{jh}\Big)\Big] \\
&= 2\phi_{Ak}\phi_{Ah}(1-\rho^*)(1 - c_1^2 F_1 - c_2^2 F_2).
\end{aligned} \tag{B9}$$

Comparing eqs. B5-B6 and eqs. B8-B9, the genotype frequencies in the two scenarios agree if

$$\rho^* = 1 - \frac{c_1(1-\rho_1)(1-F_1) + c_2(1-\rho_2)(1-F_2)}{1 - c_1^2 F_1 - c_2^2 F_2}. \tag{B10}$$

In summary, under the F model, for both homozygotes and heterozygotes, the expected genotype frequencies in a structured population are identical to those in an unstructured population with allele frequencies Φ^* (eq. B7) and inbreeding coefficient ρ^* (eq. B10). For testing the robustness of our method for allelic dropout, we simulated genotype data with population structure using $c_1 = c_2 = 0.5$, $F_1 = F_2 = F$, and $\rho_1 = \rho_2 = \rho$ (Experiment 2). In this setting, eq. B10 reduces to

$$\rho^* = \rho + (1 - \rho)\frac{F}{2 - F}. \tag{B11}$$

The values of eq. B11 for our simulated data sets are indicated by dashed lines in Fig. 2.8.

2.12 Appendix C. Additional simulation procedures

To assess the performance of our method as a function of the size of the data set, we performed two additional sets of simulations. In one, we fixed the number of loci and modified the number of individuals, and in the other, we fixed the number of individuals and modified the number of loci.

Experiment C1. Simulating data with different numbers of individuals

We used a similar procedure to that shown in Fig. 2.5A, following assumptions 2-5 of our model. We fixed the number of loci at $L = 250$. This value is chosen to be between 152 (the number of individuals in the Native American data) and 343 (the number of loci in the data). The numbers of individuals were chosen to be $N = 50$, 100, 200, 400, 800, and 1600. For each pair consisting of a choice of N and L, we simulated data sets with the inbreeding coefficient ρ ranging from 0 to 0.15 in increments of 0.01. Therefore, we generated $6 \times 16 = 96$ simulated data sets.

For each simulated data set, the allele frequencies Φ at L loci were independently

sampled (with replacement) from the estimated allele frequencies of the 343 loci in the Native American data (results from *Application to Native American data*). Given the allele frequencies Φ and the inbreeding coefficient ρ, true genotypes $\tilde{G}$ were drawn according to the inbreeding assumption. Next, the observed genotype data $\tilde{W}$ were created by adding allelic dropout. The sample-specific dropout rates $\gamma_{i\cdot}$ and the locus-specific dropout rates $\gamma_{\cdot\ell}$ were both independently sampled from $Beta(0.55, 5.30)$, as in Experiments 2 and 3 in the main text.

Experiment C2. Simulating data with different numbers of loci

The procedure we used to simulate data with different numbers of loci was similar to Experiment C1, except that we fixed the number of individuals at $N = 250$, and varied the number of loci ($L = 50$, 100, 200, 400, 800, and 1600). Therefore, we generated 96 simulated data sets, each of which has the same amount of data as a corresponding data set generated by Experiment C1.

Table 2.1: Notation used in the article. In this table, $i \in \{1, 2, ..., N\}$, $\ell \in \{1, 2, ..., L\}$, and $k, h \in \{1, 2, ..., K_\ell\}$.

Notation	Meaning	Type
i	Index of an individual	Basic notation
ℓ	Index of a locus	Basic notation
k, h	Index of an allele	Basic notation
N	Number of individuals	Basic notation
L	Number of loci	Basic notation
K_ℓ	Number of distinct alleles at locus ℓ	Basic notation
$A_{\ell k}$, $A_{\ell h}$	Allele k (h) at locus ℓ	Basic notation
X	Missing data (dropout)	Basic notation
$\gamma_{i\ell}$	Dropout probability at locus ℓ of individual i	Basic notation
$w_{i\ell}$	Observed genotype at locus ℓ of individual i	Observed data point
W	Observed genotypes, $W = \{w_{i\ell}\}$	Observed data set
$g_{i\ell}$	True genotype at locus ℓ of individual i	Latent variable
$s_{i\ell}$	IBD state at locus ℓ of individual i	Latent variable
$z_{i\ell}$	Dropout state at locus ℓ of individual i	Latent variable
G	True genotypes, $G = \{g_{i\ell}\}$	Latent variable set
S	IBD states, $S = \{s_{i\ell}\}$	Latent variable set
Z	Dropout states, $Z = \{z_{i\ell}\}$	Latent variable set
ρ	Inreeding coefficient	Parameter
$\phi_{\ell k}$	Frequency of allele $A_{\ell k}$	Parameter
$\gamma_{i\cdot}$	Sample-specific dropout rate for individual i	Parameter
$\gamma_{\cdot\ell}$	Locus-specific dropout rate for locus ℓ	Parameter
Φ	Allele frequencies, $\Phi = \{\phi_{\ell k}\}$	Parameter set
Γ	Dropout rates, $\Gamma = \{\gamma_{i\cdot}, \gamma_{\cdot\ell}\}$	Parameter set
Ψ	Model parameters, $\Psi = \{\rho, \Phi, \Gamma\}$	Parameter set
$n_{\ell k}$	Number of independent copies of allele $A_{\ell k}$	Summary statistic
$d_{i\ell}$	Number of dropouts at locus ℓ for individual i	Summary statistic
$d_{i\cdot}$	Number of sample-specific dropouts for individual i	Summary statistic
$d_{\cdot\ell}$	Number of locus-specific dropouts at locus ℓ	Summary statistic
s	Number of genotypes having two alleles IBD	Summary statistic

Table 2.2: Illustration of the outcomes of allelic dropout using two distinct alleles at locus ℓ, $A_{\ell k}$ and $A_{\ell h}$. Genotype frequencies are calculated from allele frequencies using eq. 2.5, where ρ is the inbreeding coefficient, a parameter used to model the total deviation from Hardy-Weinberg equilibrium. Dropout is assumed to happen independently to each copy at locus ℓ of individual i, with probability $\gamma_{i\ell}$ specified by eq. 2.1.

True genotype $g_{i\ell}$	Genotype frequency $\mathbb{P}(g_{i\ell}\|\Phi,\rho)$	Dropout state $z_{i\ell}$	Conditional probability $\mathbb{P}(z_{i\ell}\|g_{i\ell},\Gamma)$	Observed genotype $w_{i\ell}$	Conditional probability $\mathbb{P}(w_{i\ell}\|g_{i\ell},\Gamma)$
$A_{\ell k}A_{\ell k}$	$(1-\rho)\phi_{\ell k}^2+\rho\phi_{\ell k}$	$A_{\ell k}A_{\ell k}$ $A_{\ell k}X$ XX	$(1-\gamma_{i\ell})^2$ $2\gamma_{i\ell}(1-\gamma_{i\ell})$ $\gamma_{i\ell}^2$	$A_{\ell k}A_{\ell k}$ XX	$1-\gamma_{i\ell}^2$ $\gamma_{i\ell}^2$
$A_{\ell k}A_{\ell h}$ $(h\neq k)$	$2(1-\rho)\phi_{\ell k}\phi_{\ell h}$	$A_{\ell k}A_{\ell h}$ $A_{\ell h}X$ $A_{\ell k}X$ XX	$(1-\gamma_{i\ell})^2$ $\gamma_{i\ell}(1-\gamma_{i\ell})$ $\gamma_{i\ell}(1-\gamma_{i\ell})$ $\gamma_{i\ell}^2$	$A_{\ell k}A_{\ell h}$ $A_{\ell h}A_{\ell h}$ $A_{\ell k}A_{\ell k}$ XX	$(1-\gamma_{i\ell})^2$ $\gamma_{i\ell}(1-\gamma_{i\ell})$ $\gamma_{i\ell}(1-\gamma_{i\ell})$ $\gamma_{i\ell}^2$

Table 2.3: Posterior joint probabilities of true genotypes $g_{i\ell}$ and IBD states $s_{i\ell}$ at a single locus ℓ of an individual i. The calculation of $\mathbb{P}(g_{i\ell}, s_{i\ell}|w_{i\ell}, \Psi)$ is based on eq. 2.8. In this table, $h \neq k$.

Observed genotype $w_{i\ell}$	True genotype $g_{i\ell}$	IBD state $s_{i\ell}$	Joint probability $\mathbb{P}(g_{i\ell}, s_{i\ell}, w_{i\ell}\|\Psi)$	Posterior probability $\mathbb{P}(g_{i\ell}, s_{i\ell}\|w_{i\ell}, \Psi)$
$A_{\ell k}A_{\ell h}$	$A_{\ell k}A_{\ell h}$	1	0	0
		0	$2(1-\rho)\phi_{\ell k}\phi_{\ell h}(1-\gamma_{i\ell})^2$	1
	others	1	0	0
		0	0	0
$A_{\ell k}A_{\ell k}$	$A_{\ell k}A_{\ell h}$	1	0	0
		0	$2(1-\rho)\phi_{\ell k}\phi_{\ell h}\gamma_{i\ell}(1-\gamma_{i\ell})$	$\frac{2(1-\rho)\phi_{\ell h}\gamma_{i\ell}}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
	$A_{\ell k}A_{\ell k}$	1	$\rho\phi_{\ell k}(1-\gamma_{i\ell}^2)$	$\frac{\rho(1+\gamma_{i\ell})}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
		0	$(1-\rho)\phi_{\ell k}^2(1-\gamma_{i\ell}^2)$	$\frac{(1-\rho)\phi_{\ell k}(1+\gamma_{i\ell})}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
XX	$A_{\ell k}A_{\ell h}$	1	0	0
		0	$2(1-\rho)\phi_{\ell k}\phi_{\ell h}\gamma_{i\ell}^2$	$2(1-\rho)\phi_{\ell k}\phi_{\ell h}$
	$A_{\ell k}A_{\ell k}$	1	$\rho\phi_{\ell k}\gamma_{i\ell}^2$	$\rho\phi_{\ell k}$
		0	$(1-\rho)\phi_{\ell k}^2\gamma_{i\ell}^2$	$(1-\rho)\phi_{\ell k}^2$

Table 2.4: Posterior probabilities of true genotypes $g_{i\ell}$ at a single locus ℓ of an individual i. The calculation of $\mathbb{P}(g_{i\ell}|w_{i\ell}, \Psi)$ is based on eq. 2.10. In this table, $h \neq k$.

Observed genotype $w_{i\ell}$	True genotype $g_{i\ell}$	Posterior probability $\mathbb{P}(g_{i\ell}\|w_{i\ell}, \Psi)$
$A_{\ell k}A_{\ell h}$	$A_{\ell k}A_{\ell h}$	1
	others	0
$A_{\ell k}A_{\ell k}$	$A_{\ell k}A_{\ell h}$	$\frac{2(1-\rho)\phi_{\ell h}\gamma_{i\ell}}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
	$A_{\ell k}A_{\ell k}$	$\frac{[\rho+(1-\rho)\phi_{\ell k}](1+\gamma_{i\ell})}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
XX	$A_{\ell k}A_{\ell h}$	$2(1-\rho)\phi_{\ell k}\phi_{\ell h}$
	$A_{\ell k}A_{\ell k}$	$\rho\phi_{\ell k} + (1-\rho)\phi_{\ell k}^2$

Table 2.5: Posterior probabilities of the IBD state $s_{i\ell}$ at a single locus ℓ of an individual i. The calculation of $\mathbb{P}(s_{i\ell}|w_{i\ell}, \Psi)$ is based on eq. 2.11. In this table, $h \neq k$.

Observed genotype $w_{i\ell}$	IBD state $s_{i\ell}$	Posterior probability $\mathbb{P}(s_{i\ell}\|w_{i\ell}, \Psi)$
$A_{\ell k}A_{\ell h}$	1	0
	0	1
$A_{\ell k}A_{\ell k}$	1	$\frac{\rho(1+\gamma_{i\ell})}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
	0	$\frac{(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
XX	1	ρ
	0	$1-\rho$

Table 2.6: Posterior probabilities of the number of dropouts $d_{i\ell}$ at a single locus ℓ of an individual i. The calculations are based on eqs. 2.12 and 2.13. In this table, $h \neq k$.

Observed genotype $w_{i\ell}$	Number of dropouts $d_{i\ell}$	Joint probability $\mathbb{P}(d_{i\ell}, w_{i\ell}\|\Psi)$	Posterior probability $\mathbb{P}(d_{i\ell}\|w_{i\ell}, \Psi)$
$A_{\ell k}A_{\ell h}$	0	$2(1-\rho)\phi_{\ell k}\phi_{\ell h}(1-\gamma_{i\ell})^2$	1
	1	0	0
	2	0	0
$A_{\ell k}A_{\ell k}$	0	$[\rho+(1-\rho)\phi_{\ell k}](1-\gamma_{i\ell})^2$	$\frac{[\rho+(1-\rho)\phi_{\ell k}](1-\gamma_{i\ell})}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
	1	$2\phi_{\ell k}\gamma_{i\ell}(1-\gamma_{i\ell})$	$\frac{2\gamma_{i\ell}}{\rho(1+\gamma_{i\ell})+(1-\rho)(2\gamma_{i\ell}-\phi_{\ell k}\gamma_{i\ell}+\phi_{\ell k})}$
	2	0	0
XX	0	0	0
	1	0	0
	2	$\gamma_{i\ell}^2$	1

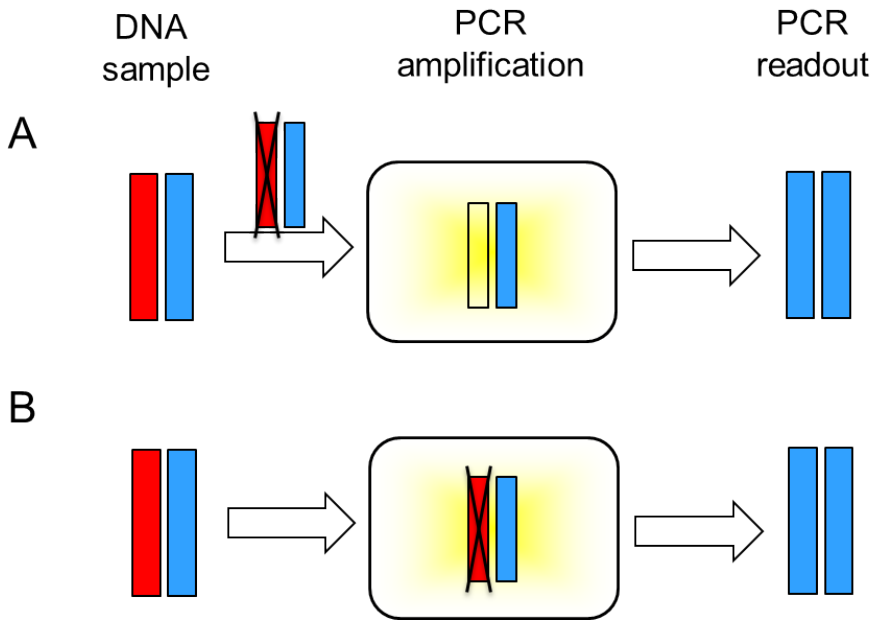

Figure 2.1: Two stages of allelic dropout. The red and blue bars are two allelic copies of a locus in a DNA sample. The black cross indicates the location at which allelic dropout occurs. (A) Owing to sample-specific factors such as low DNA concentration or poor DNA quality, one of the two alleles drops out when preparing DNA for PCR amplification. (B) Owing to either locus-specific factors such as low binding affinity between primers or polymerase and the target DNA sequences or sample-specific factors such as poor DNA quality, one of the two alleles fails to amplify with PCR. In both examples shown, allelic dropout results in an erroneous PCR readout of a homozygous genotype.

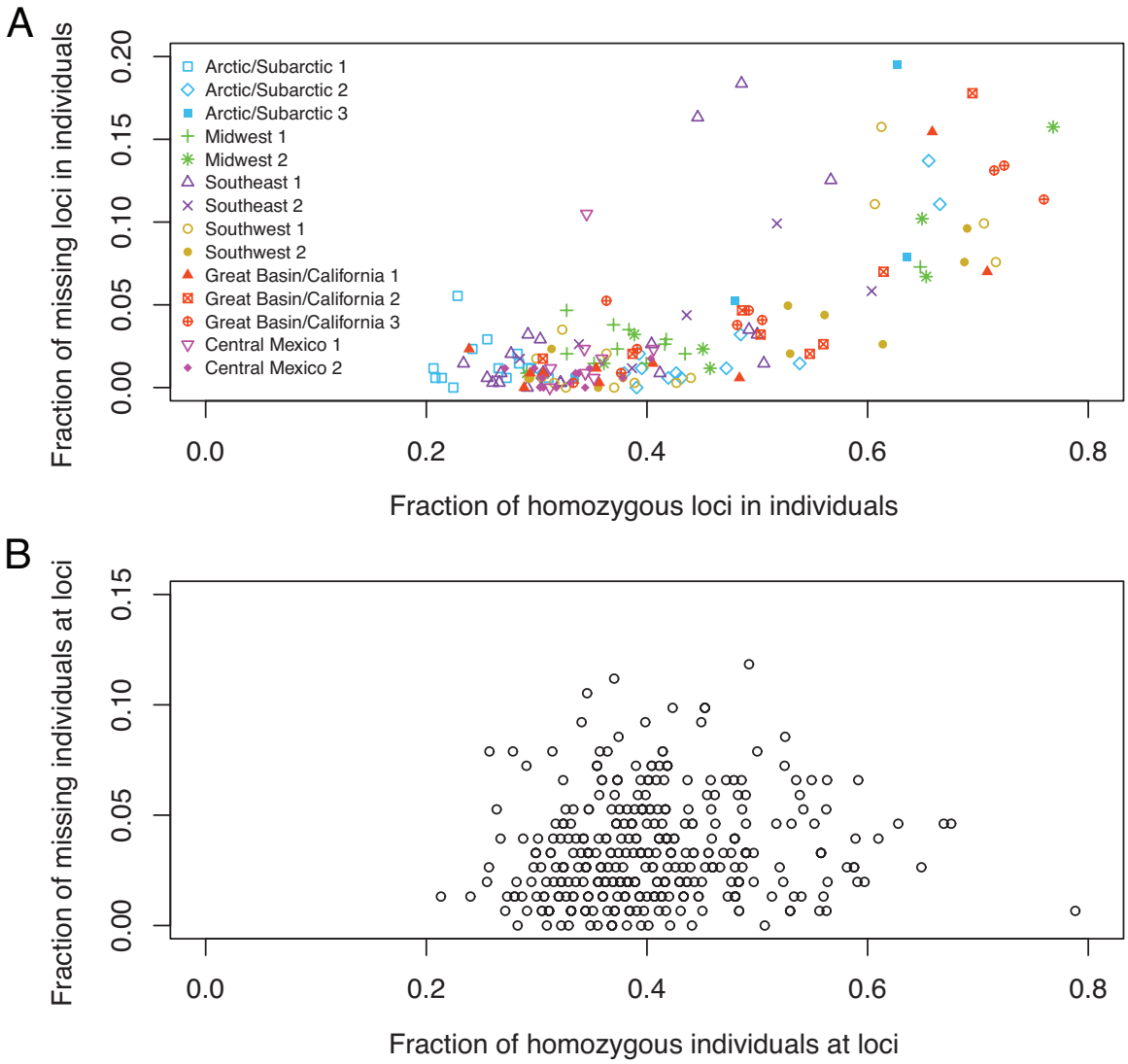

Figure 2.2: Fraction of observed missing data versus fraction of observed homozygotes. (A) Each point represents an individual with fraction x of its nonmissing loci observed as homozygous and fraction y of its total loci observed to have both copies missing. The Pearson correlation between X and Y is $r = 0.729$ ($P < 0.0001$, by 10,000 permutations of X while fixing Y). (B) Each point represents a locus at which fraction x of individuals with nonmissing genotypes are observed to be homozygotes and fraction y of all individuals are observed to have both copies missing. $r = 0.099$ ($P = 0.0326$).

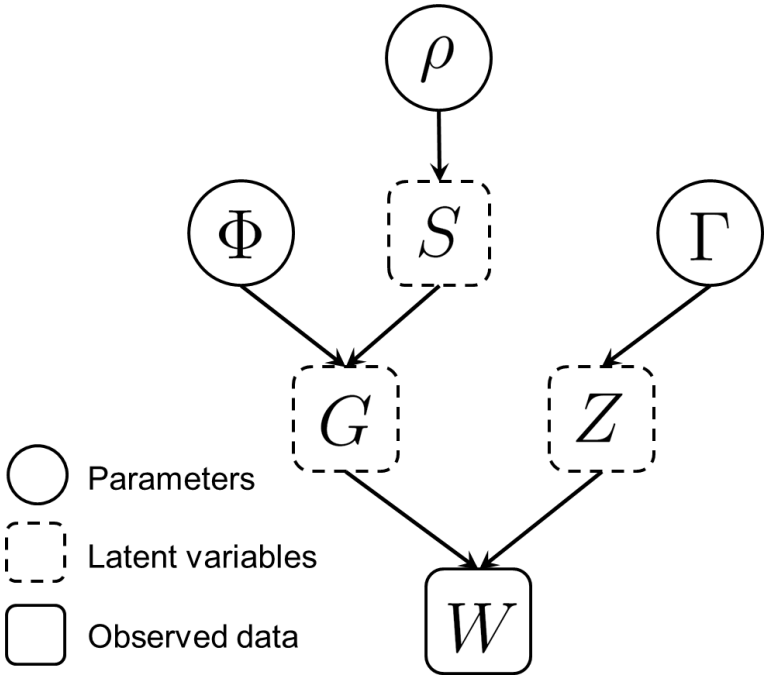

Figure 2.3: Graphical representation of the model. Each arrow denotes a dependency between two sets of quantities: Φ, allele frequencies; ρ, inbreeding coefficient; Γ, sample-specific and locus-specific dropout rates; G, true genotypes; S, IBD states; Z, dropout states; W, observed genotypes. W is the only observed data, consisting of $N \times L$ independent observations and providing information to infer parameters Φ, ρ, and Γ.

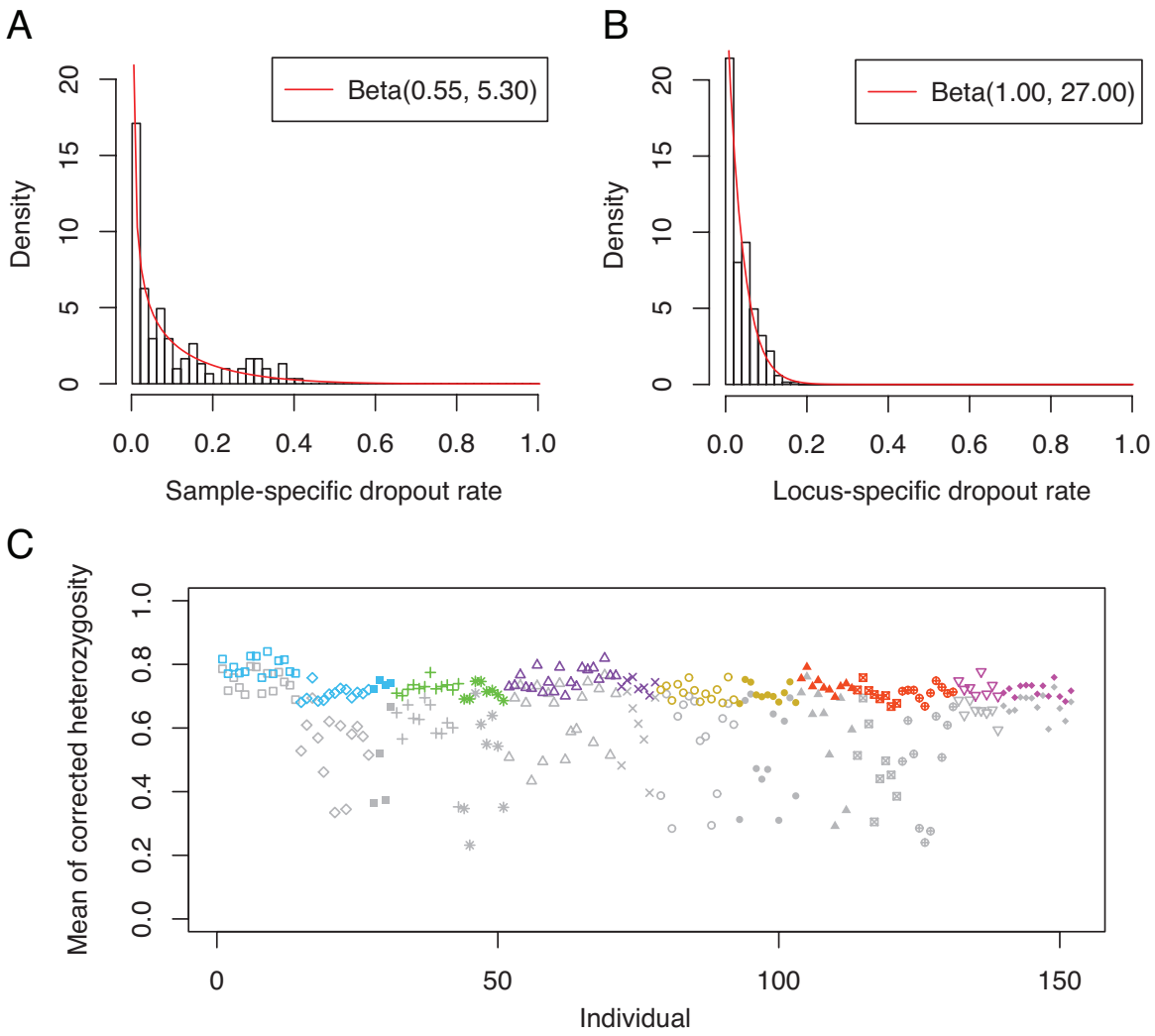

Figure 2.4: Estimated dropout rates and corrected heterozygosity for the Native American data. (A) Histogram of the estimated sample-specific dropout rates. The histogram is fit by a beta distribution with parameters estimated using the method of moments. (B) Histogram of the estimated locus-specific dropout rates. The histogram is again fit by a beta distribution using the method of moments. (C) Corrected individual heterozygosity calculated from data imputed using the estimated parameter values, averaged over 100 imputed data sets. Colors and symbols follow Fig. 2.2. The corresponding uncorrected observed heterozygosity for each individual is indicated in gray.

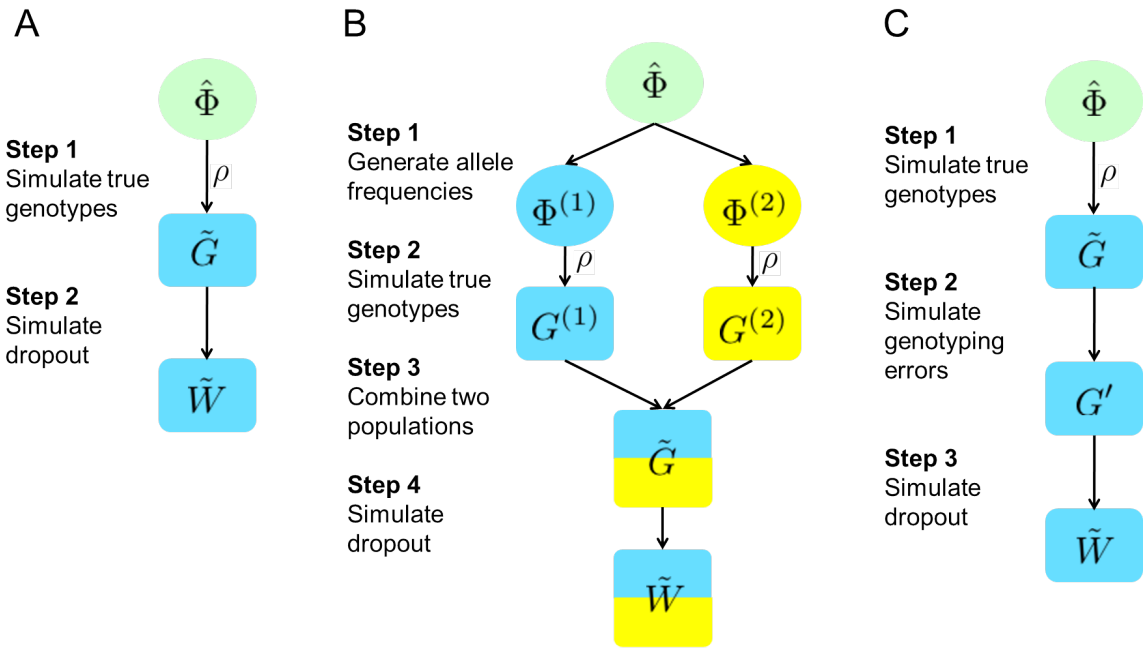

Figure 2.5: Simulation procedures. In all procedures, $\hat{\Phi}$ represents the allele frequencies estimated from the Native American data; $\tilde{G}$ represents the true genotypes generated under the inbreeding assumption; $\tilde{W}$ is the observed genotypes with allelic dropout. (A) Procedure to generate the simulated Native American data (Experiment 1). (B) Procedure to generate simulated data with population structure (Experiment 2). In step 1, the allele frequencies of two subpopulations are generated using the F model. (C) Procedure to generate simulated data with genotyping errors other than allelic dropout (Experiment 3).

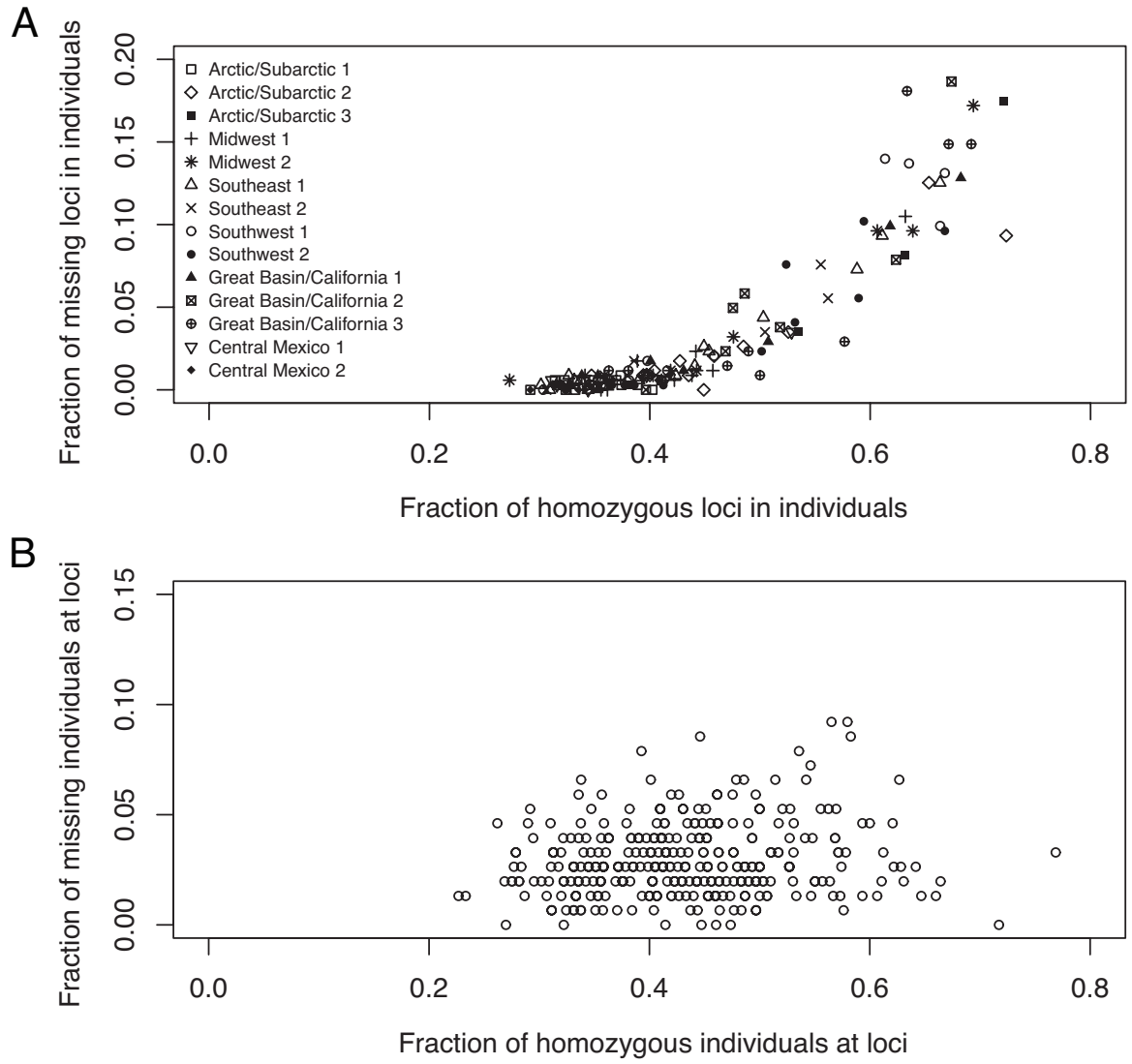

Figure 2.6: Fraction of observed missing data versus fraction of observed homozygotes for one simulated data set. (A) Each point represents an individual with fraction x of its nonmissing loci observed as homozygous and fraction y of its total loci observed to have both copies missing. The Pearson correlation between X and Y is $r = 0.900$ ($P < 0.0001$, by 10,000 permutations of X while fixing Y). (B) Each point represents a locus at which fraction x of individuals with nonmissing genotypes are observed to be homozygotes and fraction y of all individuals are observed to have both copies missing. $r = 0.143$ ($P = 0.0045$).

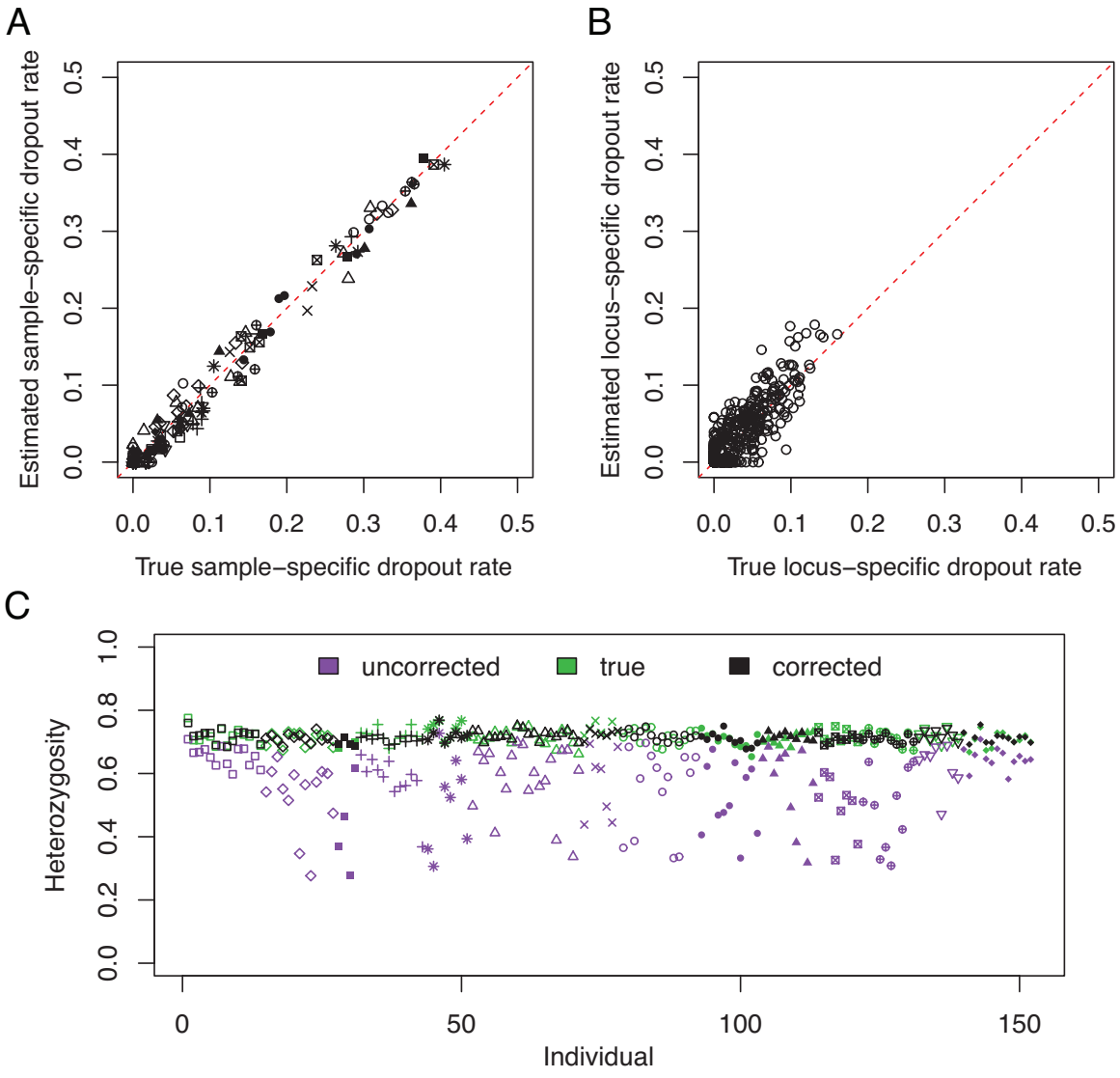

Figure 2.7: Estimated dropout rates and corrected heterozygosity for the data simulated on the basis of the Native American data set. (A) Comparison of the estimated sample-specific dropout rates and the assumed true sample-specific dropout rates. (B) Comparison of the estimated locus-specific dropout rates and the assumed true locus-specific dropout rates. (C) Individual heterozygosities in the simulated data. True values of heterozygosity are indicated by green points. With allelic dropout applied to true genotypes to generate "observed" data, the uncorrected values of heterozygosity are colored purple. Means of corrected heterozygosities across 100 imputed data sets are colored black. Symbols follow Fig. 2.6.

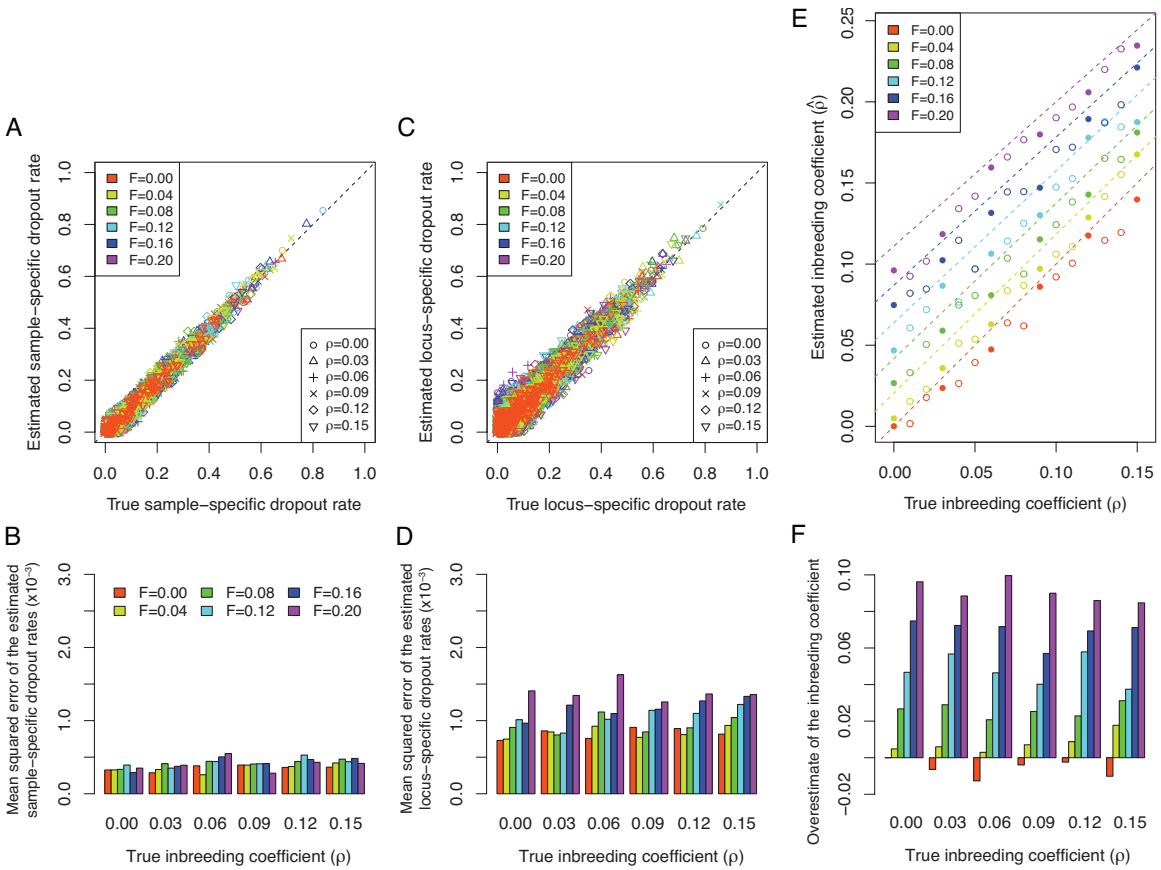

Figure 2.8: Estimated dropout rates and inbreeding coefficients for simulated data with population structure. (A) Comparison of the estimated sample-specific dropout rates and the assumed true sample-specific dropout rates. (B) Mean squared errors across all the estimated sample-specific dropout rates for each of the 36 data sets shown in panel A. (C) Comparison of the estimated locus-specific dropout rates and the assumed true locus-specific dropout rates. (D) Mean squared errors across all the estimated locus-specific dropout rates for each of the 36 data sets shown in panel C. (E) Comparison of the estimated inbreeding coefficient and the assumed true inbreeding coefficient, in which each point corresponds to one of 96 simulated data sets. The 36 solid points correspond to the simulated data sets shown in the other panels (A, B, C, D, and F). Dashed lines indicate the effective inbreeding coefficients of structured populations under the F model (eq. B11). (F) Overestimation of the inbreeding coefficient, calculated by subtracting the assumed true inbreeding coefficient from the estimated inbreeding coefficient, or $\hat{\rho} - \rho$.

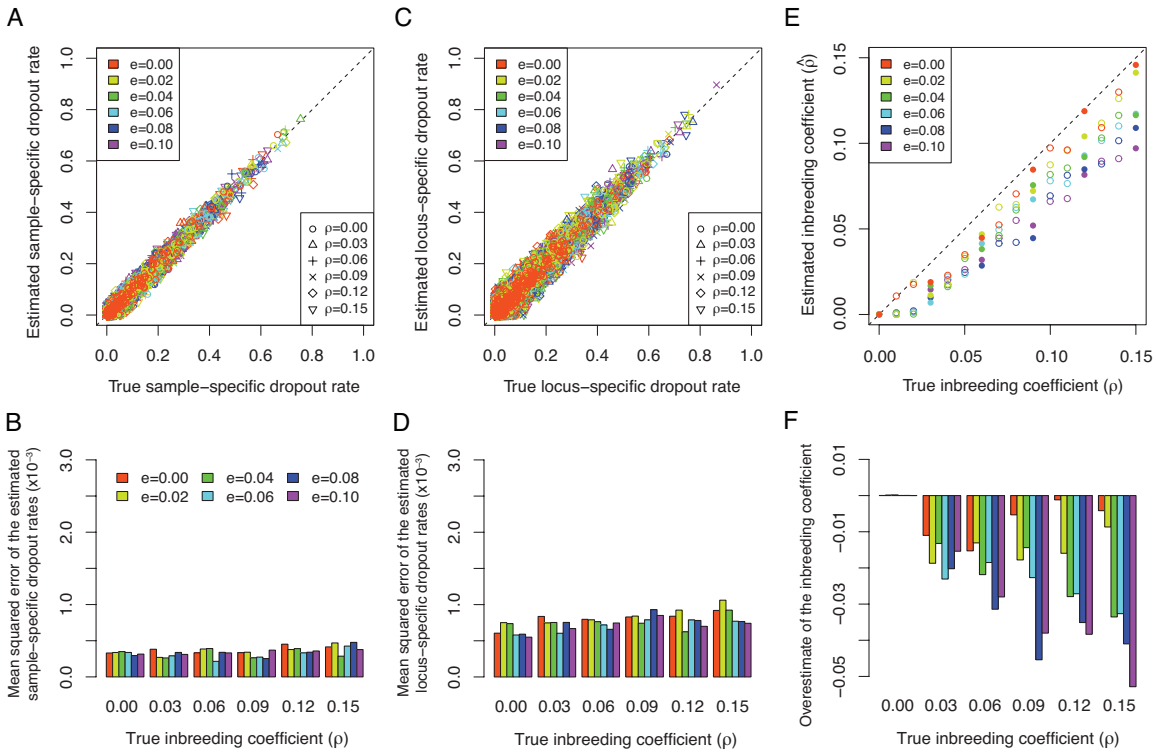

Figure 2.9: Estimated dropout rates and inbreeding coefficients for simulated data with other genotyping errors. (A) Comparison of the estimated sample-specific dropout rates and the assumed true sample-specific dropout rates. (B) Mean squared errors across all the estimated sample-specific dropout rates for each of the 36 data sets shown in panel A. (C) Comparison of the estimated locus-specific dropout rates and the assumed true locus-specific dropout rates. (D) Mean squared errors across all the estimated locus-specific dropout rates for each of the 36 data sets shown in panel C. (E) Comparison of the estimated inbreeding coefficient and the assumed true inbreeding coefficient, in which each point corresponds to one of 96 simulated data sets. The 36 solid points correspond to the simulated data sets shown in the other panels (A, B, C, D, and F). (F) Overestimation of the inbreeding coefficient, calculated by subtracting the assumed true inbreeding coefficient from the estimated inbreeding coefficient, or $\hat{\rho} - \rho$.

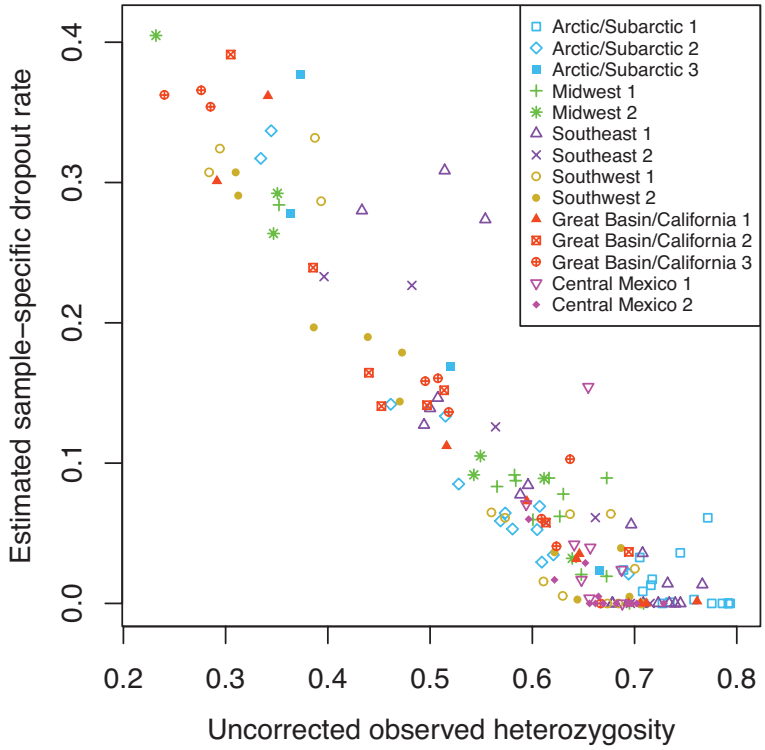

Figure S2.1: The estimated sample-specific dropout rate versus the observed heterozygosity before correcting for allelic dropout in the Native American data. For each individual, loci with both copies missing are excluded from the calculation of observed heterozygosity.

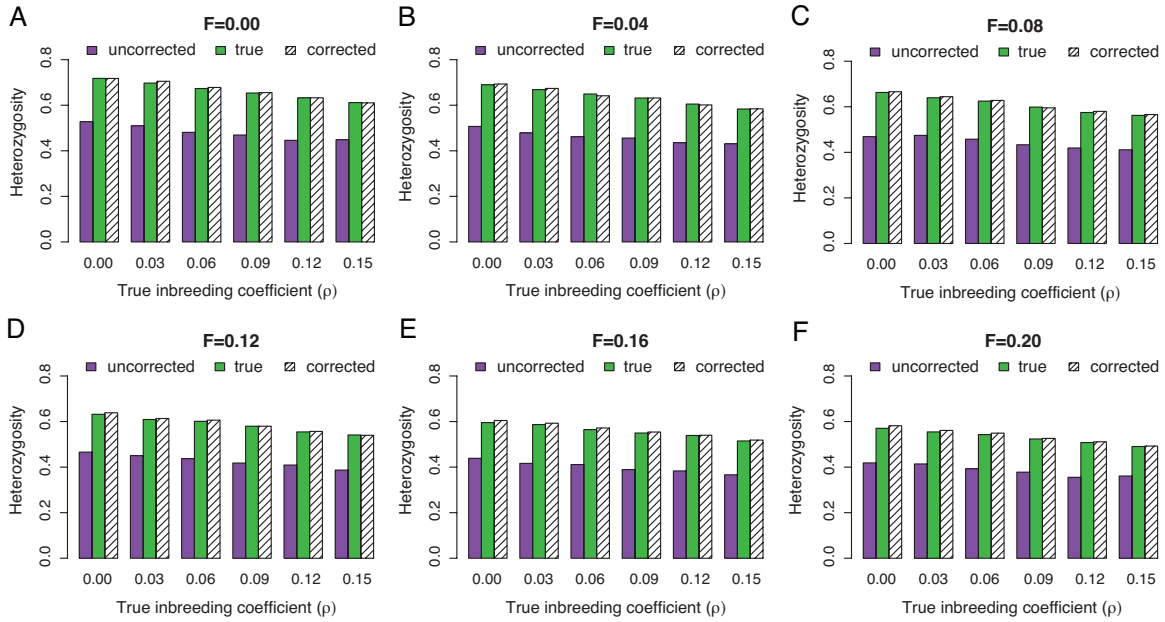

Figure S2.2: Correcting the underestimation of observed heterozygosity for simulated data with population structure. In each panel, a purple bar indicates the uncorrected observed heterozygosity averaged across all individuals in a simulated data set after applying allelic dropout; a green bar indicates the "true" observed heterozygosity averaged across all individuals in the same simulated data set before applying allelic dropout; and a striped black bar indicates the corrected observed heterozygosity averaged across all individuals and across 100 imputed data sets. The x-axis indicates values of the inbreeding coefficient that were set for different simulations. Different panels correspond to different values of the F parameter in the F-model for simulating structured populations. (A) $F = 0$; (B) $F = 0.04$; (C) $F = 0.08$; (D) $F = 0.12$; (E) $F = 0.16$; (F) $F = 0.20$.

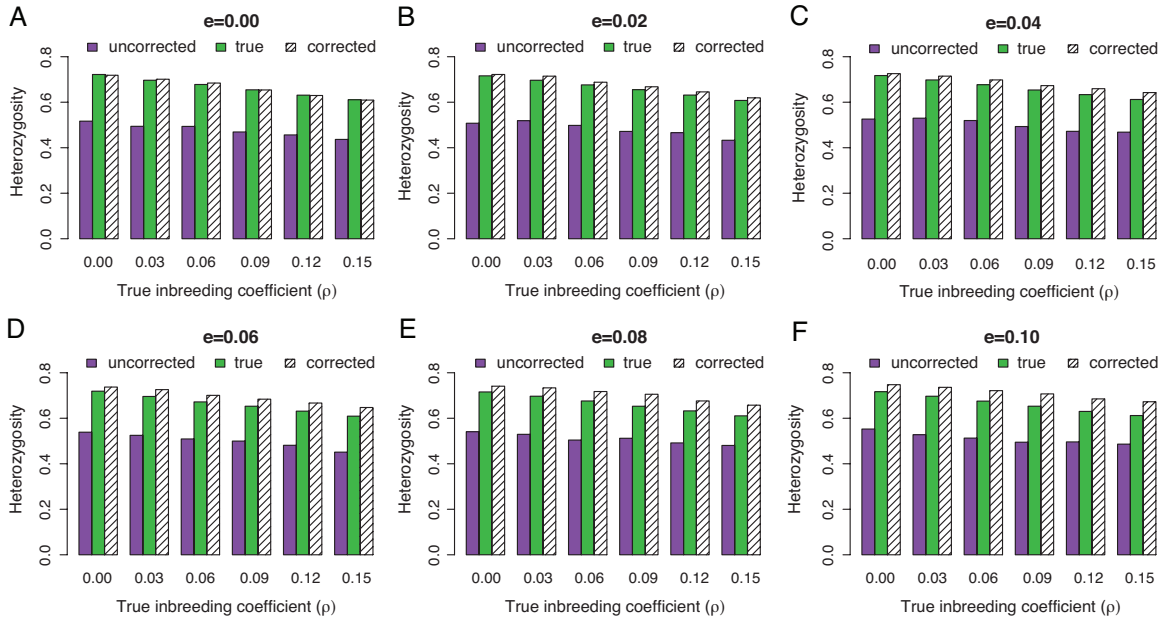

Figure S2.3: Correcting the underestimation of observed heterozygosity for simulated data with genotyping errors other than allelic dropout. In each panel, a purple bar indicates the uncorrected observed heterozygosity averaged across all individuals in a simulated data set after applying allelic dropout; a green bar indicates the "true" observed heterozygosity averaged across all individuals in the same simulated data set before applying allelic dropout and before introducing genotyping errors; and a striped black bar indicates the corrected observed heterozygosity averaged across all individuals and across 100 imputed data sets. The x-axis indicates values of the inbreeding coefficient that were set for different simulations. Different panels correspond to different levels of simulated genotyping errors that come from sources other than allelic dropout. (A) $e = 0$; (B) $e = 0.02$; (C) $e = 0.04$; (D) $e = 0.06$; (E) $e = 0.08$; (F) $e = 0.10$.

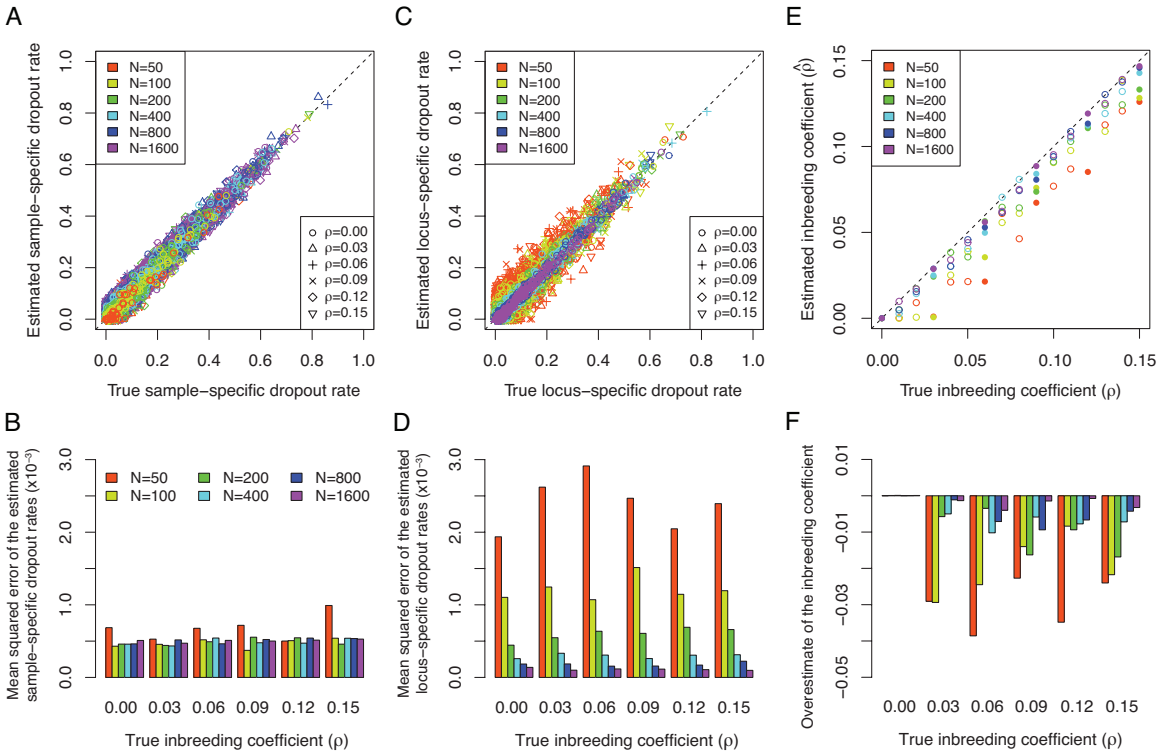

Figure S2.4: Estimated dropout rates and inbreeding coefficients for simulated data with different numbers of individuals and the same number of loci ($L = 250$). Each data set was simulated with no population structure and no genotyping errors other than allelic dropout. (A) Comparison of the estimated sample-specific dropout rates and the assumed true sample-specific dropout rates. (B) Mean squared errors across all the estimated sample-specific dropout rates for each of the 36 data sets shown in panel A. (C) Comparison of the estimated locus-specific dropout rates and the assumed true locus-specific dropout rates. (D) Mean squared errors across all the estimated locus-specific dropout rates for each of the 36 data sets shown in panel C. (E) Comparison of the estimated inbreeding coefficient and the assumed true inbreeding coefficient, in which each point corresponds to one of 96 simulated data sets. The 36 solid points correspond to the simulated data sets shown in the other panels (A, B, C, D, and F). (F) Overestimation of the inbreeding coefficient, calculated by subtracting the assumed true inbreeding coefficient from the estimated inbreeding coefficient, or $\hat{\rho} - \rho$.

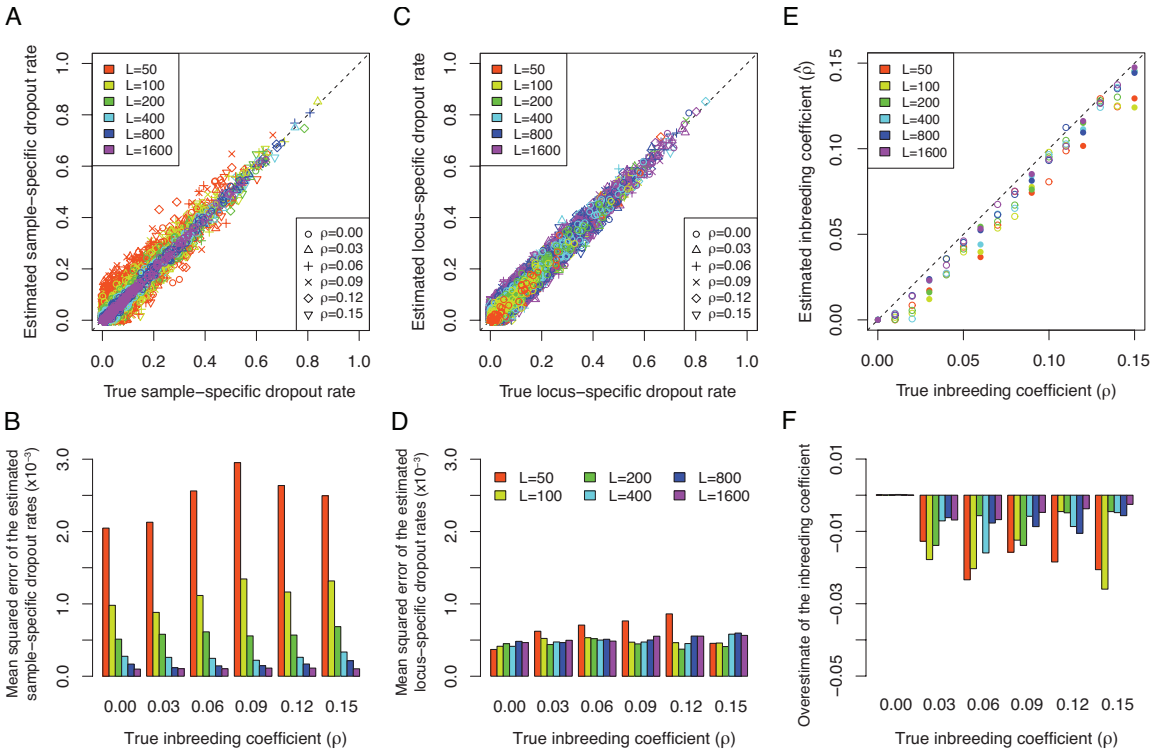

Figure S2.5: Estimated dropout rates and inbreeding coefficients for simulated data with different numbers of loci and the same number of individuals ($N = 250$). The allele frequencies for the loci were sampled with replacement from the MLEs of the Native American data. Each data set was simulated with no population structure and no genotyping errors other than allelic dropout. (A) Comparison of the estimated sample-specific dropout rates and the assumed true sample-specific dropout rates. (B) Mean squared errors across all the estimated sample-specific dropout rates for each of the 36 data sets shown in panel A. (C) Comparison of the estimated locus-specific dropout rates and the assumed true locus-specific dropout rates. (D) Mean squared errors across all the estimated locus-specific dropout rates for each of the 36 data sets shown in panel C. (E) Comparison of the estimated inbreeding coefficient and the assumed true inbreeding coefficient, in which each point corresponds to one of 96 simulated data sets. The 36 solid points correspond to the simulated data sets shown in the other panels (A, B, C, D, and F). (F) Overestimation of the inbreeding coefficient, calculated by subtracting the assumed true inbreeding coefficient from the estimated inbreeding coefficient, or $\hat{\rho} - \rho$.

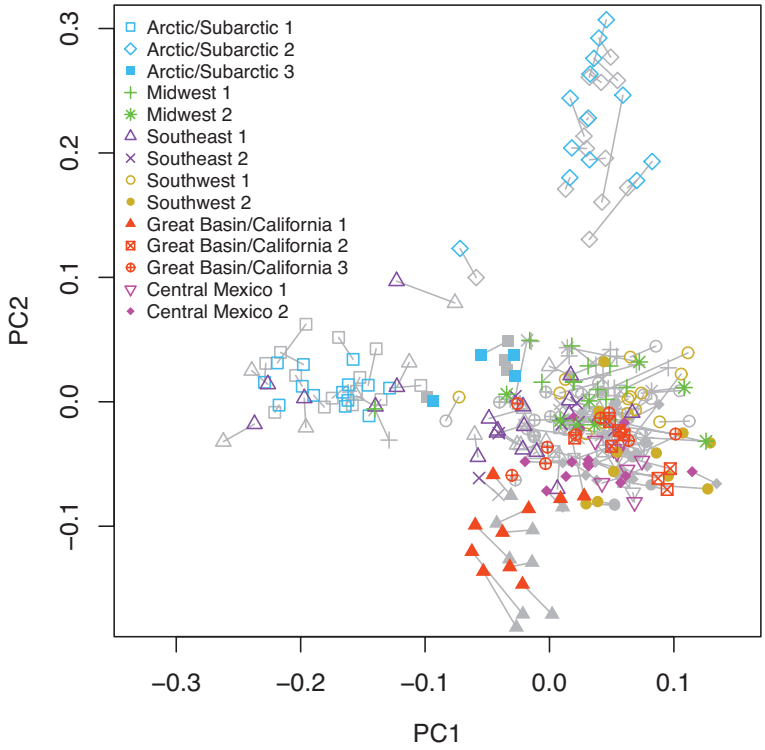

Figure S2.6: Multidimensional scaling (MDS) analysis of the Native American data. The results of MDS analysis on the original microsatellite data are shown by colored points, with the x-axis corresponding to the first principal coordinate and the y-axis corresponding to the second principal coordinate. The results of MDS analysis on one set of imputed microsatellite data are displayed with gray points, Procrustes-transformed to best match the results from the original data (*Wang et al.*, 2010). Each pair of corresponding points is connected by a gray line. The allele-sharing distance matrices calculated from the original data, averaging across loci and ignoring loci for which one or both individuals was missing, and from one set of imputed data (after correcting for allelic dropout) were used as the input to the *cmdscale* function in *R*.

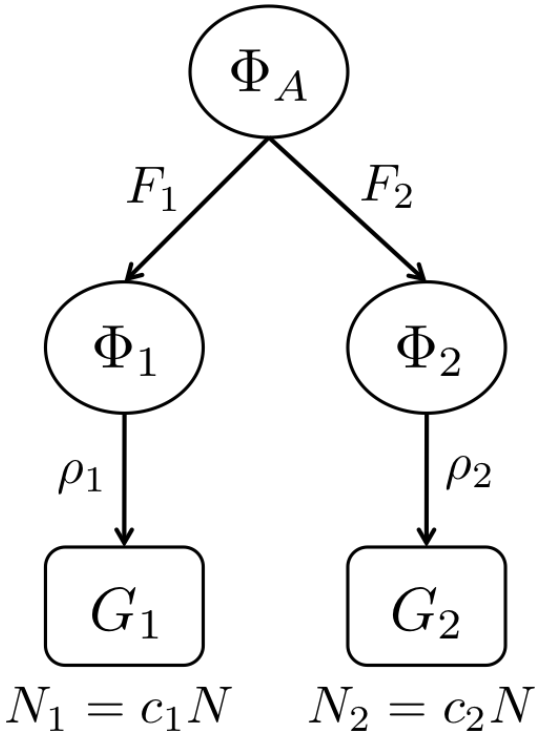

Figure S2.7: Illustration of a structured population with two subpopulations, under the F model. Φ_A denotes the allele frequencies of a common ancestral population of the two subpopulations. Φ_1 and Φ_2 are allele frequencies of the two subpopulations. The F parameter and the inbreeding coefficient for subpopulation j are F_j and ρ_j, respectively ($j = 1, 2$). In the pooled genotype data of N individuals, c_1 is the proportion sampled from subpopulation 1, producing genotype data G_1, $c_2 = 1 - c_1$ is the proportion sampled from subpopulation 2, producing genotype data G_2.

CHAPTER III

Comparing spatial maps of human population-genetic variation using Procrustes analysis

3.1 Introduction

Multivariate analysis techniques such as principal components analysis (PCA) and multidimensional scaling (MDS) are often used with population-genetic data to produce "statistical maps" of sampled individuals or populations (*Menozzi et al.*, 1978; *Zhivotovsky et al.*, 2003; *Patterson et al.*, 2006; *Novembre and Stephens*, 2008). With these techniques, each sampled individual or population is represented as a point in a Euclidean vector space in such a manner that the placement of points carries information about the similarity of the genotypes in the underlying individuals or populations. Applications to population-genetic data of PCA, MDS, and other multivariate techniques have recently been reviewed by *Jombart et al.* (2009).

Many PCA and MDS studies of population-genetic data have posed questions about the relationship of two or more such statistical maps, or about the relationship of a statistical map of population-genetic samples to a map of another type, such as a geographic map. For example: (1) does a statistical map of populations obtained from data match the statistical map predicted by a model (*Novembre and Stephens*, 2008;

McVean, 2009)? (2) Does a statistical map of populations match the geographic map of their sampling locations (*Ramachandran et al.*, 2005; *Heath et al.*, 2008; *Jakkula et al.*, 2008; *Jakobsson et al.*, 2008; *Lao et al.*, 2008; *Novembre et al.*, 2008; *Tian et al.*, 2008; *Chen et al.*, 2009; *Price et al.*, 2009; *Xu et al.*, 2009)? (3) Does a statistical map of individuals in one type of analysis match a statistical map in another type of analysis of the same samples (*Jakobsson et al.*, 2008)? For each of these questions, two maps are paired, typically in two dimensions, so that each data point in one map corresponds to a particular data point in the other map.

Comparisons between two or more such maps involving population-genetic data have generally been assessed in a qualitative manner, by visual evaluation. To provide a sensible quantitative approach for map comparison, we suggest that another technique, namely the Procrustes method (*Dryden and Mardia*, 1998; *Cox and Cox*, 2001; *Gower and Dijksterhuis*, 2004), can be borrowed from multivariate analysis. With this approach, each of two maps is transformed, preserving relative distances among pairs of points within each map. The transformations that maximize a measure of the similarity of the transformed maps are then identified, and the similarity score between the two optimally transformed maps is obtained. A permutation test can then evaluate the probability that a randomly chosen permutation of the points in one of the maps leads to a greater similarity score than that observed for the actual data points (*Jackson*, 1995; *Peres-Neto and Jackson*, 2001).

Here, we illustrate the applications of Procrustes analysis in population genetics, in scenarios that exemplify some of the questions posed above. First, we compare a two-dimensional PCA map on the basis of single-nucleotide polymorphism (SNP) data from European populations to a geographic map of population sampling locations. We next perform a similar computation for worldwide SNP data with a geographic map and an MDS map generated by classical metric multidimensional scaling (hereafter, labeled simply an "MDS map" for brevity). Our third example compares MDS and

PCA maps based on SNP data from different but overlapping worldwide samples. Finally, again using worldwide samples, we compare two-dimensional MDS maps on the basis of copy-number variant (CNV) data to a SNP-based MDS map. These various examples support the view that statistical maps on the basis of SNPs and CNVs in human populations have a high level of agreement with each other and closely reflect geography.

3.2 The Procrustes approach

We briefly review the basic Procrustes technique for the population-genetic context. Details of the approach appear elsewhere (*Dryden and Mardia*, 1998; *Cox and Cox*, 2001; *Gower and Dijksterhuis*, 2004), and our description largely follows *Cox and Cox* (2001). Consider two matrices, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$. $\mathbf{X}$ is $n \times p$, and each row in $\mathbf{X}$ corresponds to one of n points in $\mathbb{R}^p$; $\mathbf{Y}$ is $n \times q$, and each row in $\mathbf{Y}$ corresponds to one of n points in $\mathbb{R}^q$. The points are paired, so that $\mathbf{x}_r$ and $\mathbf{y}_r$ represent coordinate vectors of taxon r in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. The $\mathbf{X}$ and $\mathbf{Y}$ matrices can be viewed as describing two separate sets of coordinates for the same n taxa (two "maps" of the taxa). It is not required that p and q be equal, but in our applications, $p = q = 2$, representing two-dimensional spaces. The "taxa" can be either populations or individuals, depending on the particular case considered.

The Procrustes method aims to find the transformations, f^* and g^*, that minimize a function $d(f(\mathbf{X}), g(\mathbf{Y}))$ over all choices f and g that preserve relative pairwise distances among points in $\mathbf{X}$ and among points in $\mathbf{Y}$, respectively. First, $|p - q|$ columns of zeros are added at the end of the matrix with fewer columns in order to place both sets of points in the same k-dimensional space, with $k = \max(p, q)$. Thus, both $\mathbf{X}$ and $\mathbf{Y}$ become $n \times k$ matrices. Without loss of generality, $g^*(\mathbf{Y}) = \mathbf{Y}$ can be assumed, so that only $\mathbf{X}$ is transformed. The transformation f can be written as $f(\mathbf{x}_r) = \rho \mathbf{A}^T \mathbf{x}_r + \mathbf{b}$, where ρ is a scalar dilation, $\mathbf{A}$ is a $k \times k$ orthogonal matrix

representing a rotation and possibly a reflection, and $\mathbf{b}$ is a $k \times 1$ translation vector.

The objective function d to be minimized is the sum across taxa of squared Euclidean distances between corresponding coordinates of the taxa in the matrices $f(\mathbf{X})$ and $\mathbf{Y}$, or

$$d(f(\mathbf{X}), \mathbf{Y}) = \sum_{r=1}^{n} (\mathbf{y}_r - f(\mathbf{x}_r))^T (\mathbf{y}_r - f(\mathbf{x}_r)). \tag{B1}$$

Let $\mathbf{X}_0$ be an $n \times k$ matrix, with each row equal to $\mathbf{x}_0^T = \sum_{r=1}^{n} \mathbf{x}_r^T / n$. Similarly, let $\mathbf{Y}_0$ be an $n \times k$ matrix with each row equal to $\mathbf{y}_0^T = \sum_{r=1}^{n} \mathbf{y}_r^T / n$. Here, $\mathbf{x}_0^T$ and $\mathbf{y}_0^T$ represent the centroids of the points in $\mathbf{X}$ and $\mathbf{Y}$, respectively. We use $\mathbf{X}_c$ and $\mathbf{Y}_c$ to represent $\mathbf{X}$ and $\mathbf{Y}$ after centering points in the matrices around $\mathbf{x}_0^T$ and $\mathbf{y}_0^T$, respectively. Thus, $\mathbf{X}_c = \mathbf{X} - \mathbf{X}_0$ and $\mathbf{Y}_c = \mathbf{Y} - \mathbf{Y}_0$.

Writing the singular value decomposition of $\mathbf{C} = \mathbf{Y}_c^T \mathbf{X}_c$ as $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are $k \times k$ orthonormal matrices and $\mathbf{\Lambda}$ is a $k \times k$ diagonal matrix of singular values, the solution f^* has

$$\begin{aligned}
\mathbf{A} &= \mathbf{V}\mathbf{U}^T && \text{(B2)} \\
\rho &= \mathrm{tr}(\mathbf{\Lambda}) / \mathrm{tr}(\mathbf{X}_c^T \mathbf{X}_c) && \text{(B3)} \\
\mathbf{b} &= \mathbf{y}_0 - \rho \mathbf{A}^T \mathbf{x}_0 && \text{(B4)} \\
d(f^*(\mathbf{X}), \mathbf{Y}) &= \mathrm{tr}(\mathbf{Y}_c \mathbf{Y}_c^T) - [\mathrm{tr}(\mathbf{\Lambda})]^2 / \mathrm{tr}(\mathbf{X}_c^T \mathbf{X}_c), && \text{(B5)}
\end{aligned}$$

where "tr" represents the trace of a matrix. The solution can be viewed as providing a method for optimally representing $\mathbf{X}$ and $\mathbf{Y}$ on the same coordinate system, so that the sum of squared distances between corresponding points of $\mathbf{X}$ and $\mathbf{Y}$ is minimized. The minimum is $d(f^*(\mathbf{X}), \mathbf{Y})$, which can be scaled by dividing by $\mathrm{tr}(\mathbf{Y}_c^T \mathbf{Y}_c) = \mathrm{tr}(\mathbf{Y}_c \mathbf{Y}_c^T)$ to give the Procrustes statistic

$$D(\mathbf{X}, \mathbf{Y}) = 1 - [\mathrm{tr}(\mathbf{\Lambda})]^2 / [\mathrm{tr}(\mathbf{X}_c^T \mathbf{X}_c) \mathrm{tr}(\mathbf{Y}_c^T \mathbf{Y}_c)]. \tag{B6}$$

Considering all possible $\mathbf{X}$ and $\mathbf{Y}$, this quantity has minimum 0 and maximum 1.

A permutation approach can be used for evaluating the similarity of the two corresponding sets of coordinates (*Jackson*, 1995; *Peres-Neto and Jackson*, 2001). The similarity of $\mathbf{X}$ and $\mathbf{Y}$ is computed as $t(\mathbf{X}, \mathbf{Y}) = \sqrt{1 - D(\mathbf{X}, \mathbf{Y})}$. A permutation distribution of t can be obtained by choosing random permutations $\mathbf{X}'$ of the rows of $\mathbf{X}$ and evaluating the distribution across permutations of $t(\mathbf{X}', \mathbf{Y})$. Using t_0 for the value of t from the unpermuted matrices, $\mathbb{P}[t(\mathbf{X}', \mathbf{Y}) > t_0]$ gives the probability that a random pairing of the taxa in $\mathbf{X}$ and $\mathbf{Y}$ leads to greater similarity than the actual pairing. Each of our permutation tests employed 10,000 permutations.

3.3 Genes and geography in Europe

Novembre et al. (2008) compared a two-dimensional PCA map of European samples, obtained by analyzing 197,146 SNPs in 1,387 individuals from 36 countries, to a geographic map of sampling locations. They examined rotations of the coordinates of the points in the two-dimensional plot of PC1 and PC2, determining the angle of rotation around the origin (PC2, PC1) = (0, 0) that maximized the sum of the correlation with longitude of the first coordinate in the rotated PC space and the correlation with latitude of the second coordinate in the rotated PC space. This analysis found that a 16° counterclockwise rotation of the PCA plot most closely resembled the geographic map. To qualitatively demonstrate the resemblance, their Figure 1a provided a striking juxtaposition of the rotated PCA plot alongside a geographic map of Europe. Similar results have been presented by *Heath et al.* (2008) and *Lao et al.* (2008).

With the Procrustes approach, it is further possible to *superimpose* the *Novembre et al.* (2008) genetic and geographic maps of Europe in a manner that minimizes the sum across countries of squared distances between geographic coordinates and transformed PCA coordinates. For our analysis, the (PC2, PC1) and (longitude, latitude)

coordinates of the samples were kindly shared by J. Novembre. Multiple individuals were sampled per country, with all individuals assumed to have the same geographic coordinates. For each country, from (longitude, latitude) coordinates (λ, ϕ) measured in degrees, we used the Gall-Peters projection, an equal-area projection that preserves distance along the 45°N parallel, to obtain rectangular coordinates $(R\pi\lambda\sqrt{2}/360°, R\sqrt{2}\sin\phi)$, where R represents the radius of the earth. These geographic coordinates are plotted in Figure 3.1A.

For each country, we also obtained the centroid on the *Novembre et al.* (2008) PCA plot of the individuals sampled from the country. Using the 36 pairs of geographic and PCA coordinates, we employed eqs. B2-B4 to identify the optimal transformation for aligning the PCA coordinates with the (Gall-Peters-projected) geographic coordinates. This transformation was then applied to the (PC2, PC1) coordinates of all sampled individuals. Figure 3.1B shows the Procrustes-transformed coordinates of the PCA plot, superimposed on the geographic map of Europe. The centroid of the 36 sets of geographic coordinates and the centroid of the 36 sets of PCA coordinates coincide at 47.539°N 15.498°E, ∼90 km southwest of Vienna, Austria. The rotation applied to the PCA coordinates is 8.860° counterclockwise, reasonably close to the rotation angle of 16° obtained by the method of *Novembre et al.* (2008). Note, however, that beyond the difference due to our use of Procrustes analysis, two differences exist between our analysis and that of *Novembre et al.* (2008). First, we applied a projection to the (longitude, latitude) geographic coordinates, whereas *Novembre et al.* (2008) used unprojected coordinates. When we repeat our Procrustes analysis using unprojected coordinates, we obtain 10.500° for the angle of rotation. Second, in aligning genetic and geographic coordinates, we used centroid coordinates for each country, whereas in the analysis of *Novembre et al.* (2008), coordinates were aligned at the individual level (treating all individuals from the same country as having identical coordinates). When we repeat our analysis using individual coordinates, we

obtain 16.428° for the rotation angle. Further, if we use unprojected geographic coordinates and individual rather than centroid coordinates, as was done by *Novembre et al.* (2008), we obtain a rotation angle of 16.050°, in close agreement with the 16° angle of *Novembre et al.* (2008).

Applying the permutation test with our analysis relying on projected geographic coordinates and population centroids, we find that $t_0 = 0.874$, with $P < 0.0001$ that a random permutation of the labels in the PCA plot produces greater similarity to the geographic coordinates than that seen with the correct labels (Figure 3.2). Thus, the pattern of relative distances among points in the PCA plot has a demonstrably high degree of similarity to the corresponding pattern of relative distances in the geographic map. Through a quantitative assessment of this similarity, our computations confirm the qualitatively striking concordance of genetics and geography reported by *Novembre et al.* (2008).

3.4 Genes and geography worldwide

We next performed an analogous alignment of coordinates computed from genetic data to geographic sampling locations, for samples collected worldwide. In an analysis of 512,762 SNPs in 443 individuals from 29 worldwide human populations, *Jakobsson et al.* (2008) obtained a two-dimensional MDS plot on the basis of an individual-level pairwise allele-sharing genetic distance matrix. Qualitatively, the MDS plot resembled a geographic map of the sampling locations, with the axes corresponding largely to latitude and longitude. This same phenomenon is visible in the work of *Li et al.* (2008) and *Biswas et al.* (2009).

To quantitatively assess the resemblance, we Procrustes-transformed SNP-based MDS coordinates to produce an optimal alignment with geographic coordinates. For this analysis, we used coordinates of an MDS plot based on a population-level genetic distance matrix. We used `microsat` (*Minch et al.*, 1998) to obtain the allele-sharing

genetic distance matrix (*Mountain and Cavalli-Sforza*, 1997) between populations for the data of *Jakobsson et al.* (2008). Classical metric multidimensional scaling was applied to the matrix, using the `cmdscale` command in `R` (*Ihaka and Gentleman*, 1996). For the geographic coordinates, we used (Gall-Peters-projected) latitudes and longitudes from Table S6 of *Jakobsson et al.* (2008).

Figure 3.3A shows the geographic coordinates of the 29 populations, drawn on a world map. Figure 3.3B provides the Procrustes-transformed two-dimensional MDS plot of the genetic data. Although genetic coordinates for some populations are quite distant from the corresponding sampling locations, a geographic pattern in the MDS plot is clear. The value of t_0 for the genetic and geographic coordinates is 0.799 ($P < 0.0001$), considerably exceeding the similarity values for all 10,000 permutations examined for the labels in the MDS plot (Figure 3.4). As was true in the case of Europeans, a formal quantitative comparison supports the qualitative resemblance of genetic coordinates to geographic coordinates.

3.5 MDS and PCA

Our next example considered the similarity of MDS and PCA plots obtained on the basis of SNP data in overlapping worldwide samples. In particular, we compared the individual-level two-dimensional MDS plot of *Jakobsson et al.* (2008) with the corresponding individual-level PCA plot of the first two principal components in *Biswas et al.* (2009). For the MDS plot, we used coordinates from the individual-level SNP-based MDS plot presented by *Jakobsson et al.* (2008), in which MDS was performed on an individual level pairwise allele sharing genetic distance matrix. The PCA coordinates from *Biswas et al.* (2009) were based on the analysis of 643,884 autosomal SNPs and 944 unrelated individuals from 52 populations (*Li et al.*, 2008), and were kindly shared by J. Akey. The SNP genotype matrix for their PCA was normalized using equation (3) in *Patterson et al.* (2006). The datasets underlying

the two plots have considerable overlap, in that 433 individuals are included in both datasets.

We applied Procrustes analysis to the common set of 433 individuals, represented by 433 pairs of points, one each in the MDS and PCA plots. The 433 points in the PCA plot were transformed to produce an optimal alignment with the 433 corresponding points in the MDS plot. The optimal transformation was then applied to all 944 points in the PCA plot.

Figure 3.5A shows the individual-level MDS plot of genetic data, in which 443 individuals from 29 populations are included (*Jakobsson et al.*, 2008). The orientation of this figure was determined by Procrustes transformation, aligning individual-level MDS coordinates to the geographic coordinates of the individuals. Figure 3.5B shows the Procrustes-transformed PCA plot with all 944 individuals from 52 populations included. The two plots are quite similar, with the larger number of points present in the PCA plot filling in gaps visible in the MDS plot. Considering 10,000 permutations of the labels in the PCA plot of the 433 shared points, we find that $t_0 = 0.993$ with $P < 0.0001$. This high value of t_0 indicates a very strong concordance between MDS and PCA in analyzing the data, as is expected given the close relationship of these two techniques (indeed, for a given use of PCA, a certain special case of MDS produces identical results (*Mardia et al.*, 1979)). The example further illustrates how Procrustes analysis can be used to compare two plots in which the sets of points only partially overlap.

3.6 SNPs and CNVs

Our final comparison examined the similarity of MDS plots obtained using different types of markers collected in the same samples. We compared an MDS plot on the basis of 396 copy-number-variable loci reported by *Jakobsson et al.* (2008) to the SNP-based MDS plot in the same worldwide populations. The population-level

CNV genetic distance matrix was obtained as in *Jakobsson et al.* (2008). MDS and Procrustes computations were conducted in the same manner as in the analysis of worldwide SNPs and geography.

The CNV-based and SNP-based MDS plots are qualitatively dissimilar, with the SNP-based plot (Figure 3.3B) resembling the geographic sampling locations (Figure 3.3A), and the CNV-based plot (Figure 3.6A) instead having all except three points located near the center. The similarity statistic between the CNV-based and SNP-based plots reflects this relative discordance ($t_0 = 0.285$, $P = 0.1536$). Removal from the two MDS plots of the three outlier populations — Kalash, Melanesian, and Papuan — followed by reapplication of Procrustes analysis leads to greater qualitative similarity (Figure 3.6B). Although the similarity statistics in Figures 3.6A and 3.6B are not strictly comparable because of the different numbers of points in the two plots, it is noteworthy that upon removal of the outliers, the t statistic between the CNV-based and SNP-based MDS plots increases to $t_0 = 0.400$ ($P = 0.0292$).

The importance of the three outlier populations in determining the nature of the axes in the CNV-based MDS plot is potentially a consequence of high genetic distances in comparisons involving these populations (Table S1 of *Jakobsson et al.* (2008)). These high distances result from high numbers of CNVs detected in the three outlier populations (*Jakobsson et al.*, 2008), which in turn might be a consequence of high values in these populations of a tuning parameter used in the CNV genotyping assays (*Itsara et al.*, 2009). CNV genotypes were obtained using `PennCNV` (*Wang et al.*, 2007a) applied to genome-wide genotyping intensity signals. For a given sample, the variability of genotyping intensity across the genome influences the ability of `PennCNV` to accurately identify CNVs (*Itsara et al.*, 2009; *Wang et al.*, 2007a). The "standard deviation of the log R ratio," henceforth denoted s, provides a measure of this variability, where the log R ratio at a given site considers $\log_2$ of the ratio of the genotyping intensity for one allelic type to the genotyping intensity for the other

allelic type. Higher values of the parameter s lead to greater difficulty in accurate CNV identification by `PennCNV`, systematically giving rise to additional false positive CNV detections.

The Procrustes approach enables us to assess the hypothesis that the dissimilarity of the CNV-based and SNP-based MDS plots in Figures 3.6A and 3.3B ultimately traces to high-s low-quality genotyping assays in outlier populations. We first varied the maximal value of s allowed for samples included in the analysis. Among 443 unrelated individuals studied by *Jakobsson et al.* (2008), the CNV-based MDS plot in Figure 3.6A utilized 405 of these individuals, each with $s < 0.28$. Starting from this set of 405 individuals, we generated nine datasets based on nine values of the upper bound on s for samples included in the analysis. These choices for the cutoff on s were selected at intervals of 0.01 from 0.20 to 0.28 inclusive. The choice of 0.28, used by *Jakobsson et al.* (2008), matches that of Figure 3.6 and is the most permissive, producing a dataset with the most CNVs, but with a potentially greater number of false positive CNV identifications. The choice of 0.20 is the most restrictive, leading to a smaller dataset with fewer samples, but also with fewer false positives. For each choice of cutoff, samples were excluded from the initial collection of 405 individuals if their s values were greater than or equal to the cutoff (no samples had s equal to a two-digit decimal number, so that exclusions of s values strictly greater than the cutoff would have produced the same datasets). Using each set of individuals derived from the original dataset, CNV loci polymorphic in the reduced set were identified, and non-singleton autosomal CNVs were retained for MDS analysis. In some populations, as few as two individuals were retained in the reduced datasets (Table 3.1), but each of the nine datasets included individuals from all populations (Table 3.2). To ensure that all datasets included at least two individuals from each population, we did not consider cutoff choices below 0.20.

MDS analyses of the eight new CNV datasets proceeded using the same methods

as were used in the analysis of the initial $s < 0.28$ dataset. For each CNV dataset, we constructed an allele-sharing population-level genetic distance matrix in the same manner as was done by *Jakobsson et al.* (2008) for the $s < 0.28$ dataset. We then performed MDS and used Procrustes analysis to compare the resulting plots to the SNP-based MDS plot in Figure 3.3B.

Figure 3.7 displays the Procrustes-transformed CNV-based MDS plots based on the nine choices of the cutoff on s. As the cutoff decreases, the resemblance of the MDS plot to the SNP-based MDS plot in Figure 3.3B increases. The smallest values of the cutoff on s lead to MDS plots with a similar triangular structure to the plot obtained with SNPs: populations from Africa lie in the lower left corner, populations from the Middle East and Europe lie near the top, populations from the Americas lie on the right, and a series of populations from Asia lies along an upper edge. The values of t_0 are greatest for the lowest values of the cutoff, and all plots except the $s < 0.28$ plot produce $P < 0.0001$. Figure 3.8 demonstrates that for cutoffs of 0.25 or less, t_0 is quite high, greater even than the value of t_0 for the comparison of SNPs and geography in Figure 3.4. The t_0 similarity statistic is somewhat lower with cutoffs $s < 0.26$ and $s < 0.27$, and it is considerably lower with the original cutoff of $s < 0.28$.

Thus, Procrustes analysis of reduced CNV datasets suggests that CNVs produce similar patterns of population structure to those observed with SNPs. When restricting the CNV dataset to smaller sets of individuals with more reliable CNV detection, the similarity of CNV-based MDS plots to the SNP-based MDS plot increases. The observations support the view that high values of s for certain individuals from the Kalash, Melanesian, and Papuan populations explain the outlier status of these populations in a previous analysis of CNV population structure (*Jakobsson et al.*, 2008). As suggested by *Itsara et al.* (2009), it is likely that high-s individuals contain numerous false-positive CNV genotypes; however, removal of these individuals only reinforces the observation of *Jakobsson et al.* (2008) that a general similarity exists between

CNV-based and SNP-based inferences of population structure.

3.7 Discussion

The Procrustes approach for investigating the concordance of separate sets of spatial positions has been used for diverse biological problems, particularly in the context of morphometric data (*Dryden and Mardia*, 1998; *Bookstein*, 1996; *Adams et al.*, 2004). We suggest that this approach similarly has considerable potential for use with population-genetic data. Our examples quantitatively comparing genes and geography with Procrustes analysis strengthen the evidence for patterns previously identified qualitatively. They support a strong role for geography in predicting patterns of population structure, both in Europe and worldwide. Our Procrustes example with CNV-based and SNP-based MDS plots shows that the similarity of CNV-based inference of human population structure to SNP-based inference is greater than had been reported previously with a permissive cutoff for sample inclusion in CNV analysis.

In agreement with *Itsara et al.* (2009), our Procrustes analysis supports the view that the difference between CNV-based and SNP-based inference in our previous analysis (*Jakobsson et al.*, 2008) was due to use of a permissive cutoff. However, in contrast to the claim of *Itsara et al.* (2009) that there is "limited evidence for stratification of CNVs in geographically distinct human populations," our use of a more restrictive cutoff leads to the conclusion that population structure is detectable on the basis of CNVs, and that the CNV population structure pattern has a strong concordance with that inferred using SNPs. The concordance between CNV-based and SNP-based MDS plots, $t_0 = 0.892$ for the $s < 0.22$ cutoff on the standard deviation of the log R ratio, exceeds the concordance between the SNP-based MDS plot and the geographic coordinates.

We note that many alternatives to the Procrustes approach exist for aligning

sets of points, including methods that are robust to the presence of outliers (*Dryden and Mardia*, 1998; *Rohlf and Slice*, 1990). In addition, the Mantel coefficient (*Mantel*, 1967; *Sokal and Rohlf*, 1995) and the RV coefficient (*Robert and Escoufier*, 1976; *Heo and Gabriel*, 1998) provide alternatives to the Procrustes t statistic for measuring the similarity of pairs of plots. To compare t and the RV coefficient, for each of the CNV-based MDS plots in Figure 3.7, we repeated our comparisons to the SNP-based MDS plot in Figure 3.3B, substituting the RV coefficient in place of the t statistic. The correlation of RV and t across the nine plots was high ($r = 0.994$), and P-values from permutation tests with RV were similar to those with t ($P = 0.2836$ for the $s < 0.28$ plot and $P < 0.0001$ for all other plots). However, while t and the RV coefficient appear to perform similarly, t is perhaps more intuitive in the Procrustes context, as it is a simple function of the the sum of squared Euclidean distances between corresponding points in the two plots when the plots are optimally aligned.

The computations we have performed involve comparisons of genes and geography, comparisons of results from two separate multivariate analysis techniques (PCA and MDS), and comparisons of inferences from separate types of markers. However, the Procrustes approach has several other potential uses in population genetics. The Procrustes t statistic can provide a method for comparing PCA or MDS plots based on observed data to those based on simulations, thereby assisting in evaluating the fit of PCA and MDS patterns in population-genetic data to those that population-genetic models predict. The Procrustes approach also enables the comparison of variant analyses performed with the same multivariate analysis technique, such as in examining MDS plots based on different genetic distances or based on different bootstrap replicates. As in our example comparing PCA results of *Biswas et al.* (2009) and MDS results of *Jakobsson et al.* (2008), Procrustes analysis can be used in integrating separate results on the basis of sample sets that overlap only partially. In our investigation of multiple analyses of CNVs, we based the comparison on similarity

to a reference dataset; if no natural basis exists for selecting a particular dataset as the reference, such as in comparing multiple genetic distances, bootstrap replicates, or repeated simulations, a generalized Procrustes technique can be used, in which results from the various analyses are transformed iteratively until a least-squares sum considering all pairs of configurations cannot be further reduced (*Dryden and Mardia*, 1998; *Gower*, 1975). In all these applications, Procrustes methods can make the results of separate analyses of standard data sets commensurable. Further, Procrustes analysis is applicable to data both in two dimensions and in higher-dimensional spaces for which no simple visual representation alternative exists. Thus, the examples we have considered represent only a small subset of the category of problems in population genetics for which the Procrustes approach might provide an informative tool for data analysis.

3.8 Acknowledgements

We are grateful to J. Akey and J. Novembre for assistance with the data from their papers. We thank T. Jombart and an anonymous reviewer for comments on the manuscript. This work was supported in part by NIH grants R01 GM081441 and T32 GM070449, by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences, by an Alfred P. Sloan Research Fellowship, and by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human Services (project number Z01-AG000932-02).

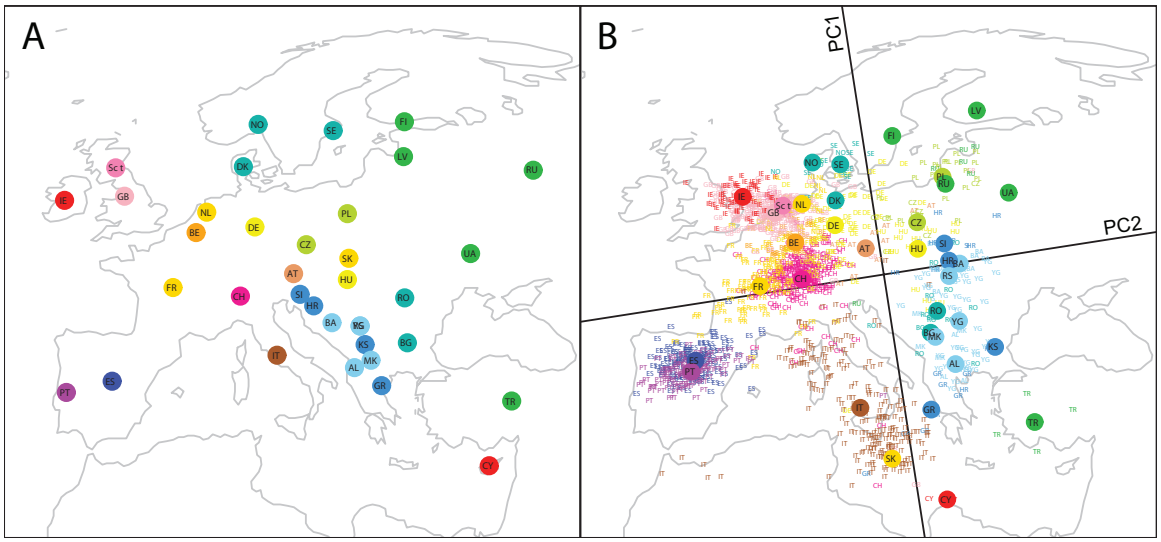

Figure 3.1: Procrustes analysis of genetic and geographic coordinates in Europe, based on data from *Novembre et al.* (2008). (A) Geographic coordinates for 36 countries. (B) Procrustes-transformed plot of the first two principal components of genetic variation. The plot is centered at the geographic centroid of the populations. Individuals are represented by two- and three-letter abbreviations, and circles represent the centroids of the PCA coordinates for individuals from a country. Abbreviations are as follows: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, Great Britain; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NL, Netherlands; NO, Norway; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia; Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia. Population labels follow the color scheme of *Novembre et al.* (2008). The figures are drawn according to the Gall-Peters projection.

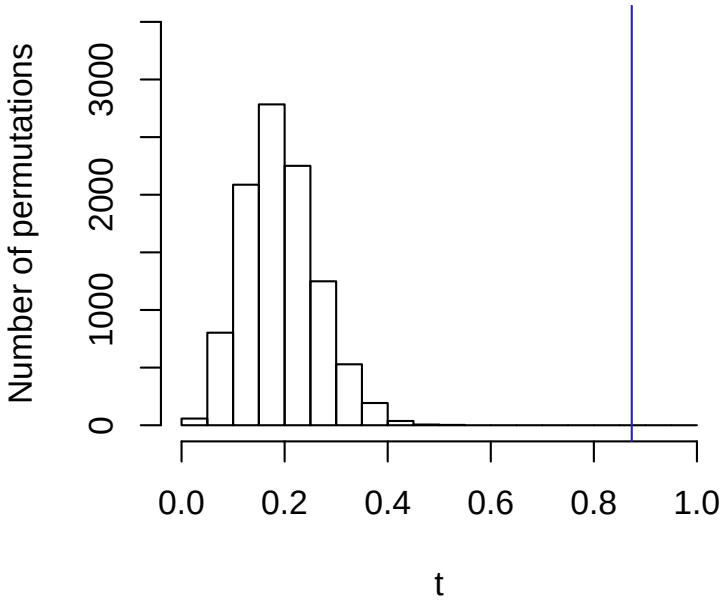

Figure 3.2: Distribution of the permutation test statistic t, comparing a geographic map of sampling locations (Figure 3.1A) and a SNP-based PCA map (Figure 3.1B) in European populations. The value of t_0, the permutation test statistic obtained from the unpermuted data, is represented by the blue vertical line, and it equals 0.874 ($P < 0.0001$).

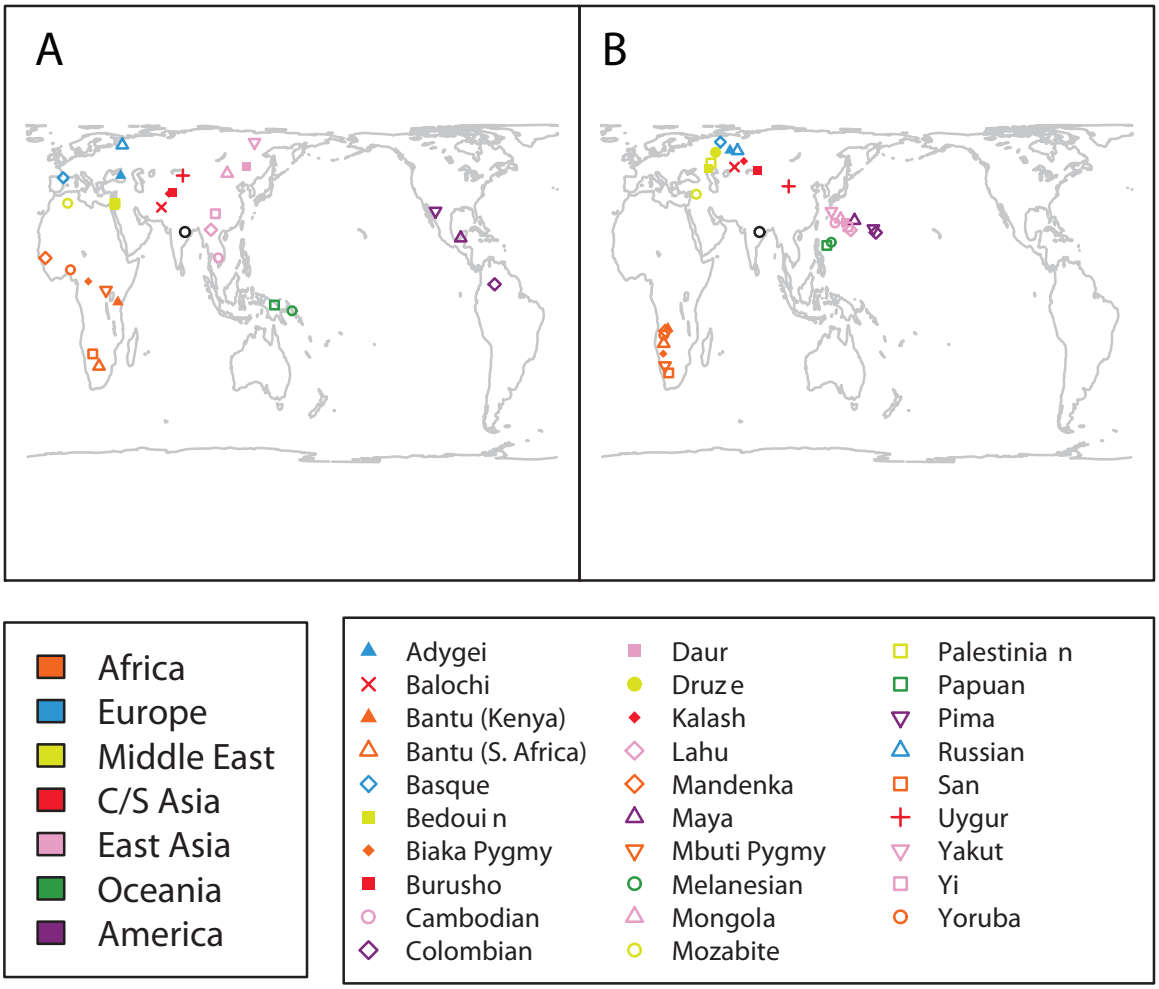

Figure 3.3: Procrustes analysis of genetic and geographic coordinates worldwide, based on data from *Jakobsson et al.* (2008). (A) Geographic coordinates for 29 populations. (B) Procrustes-transformed MDS plot of genetic variation. The figures are drawn according to the Gall-Peters projection. For each graph, the black open circle represents the centroid of the points plotted.

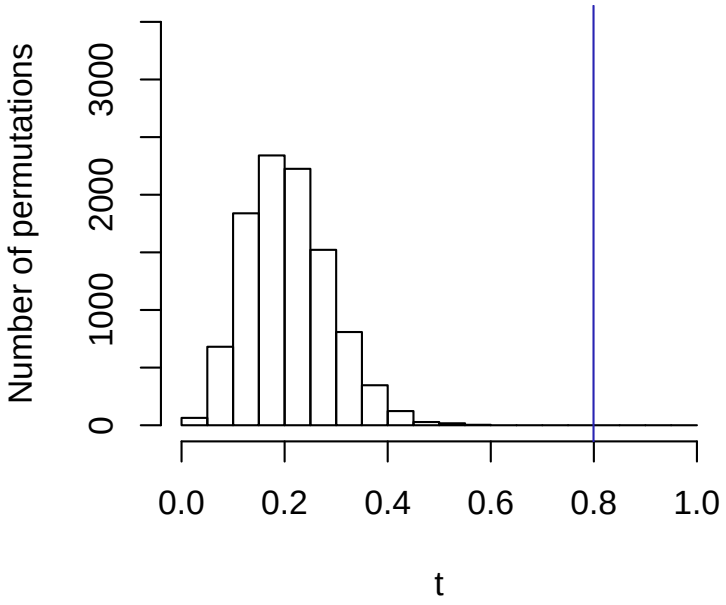

Figure 3.4: Distribution of the permutation test statistic t, comparing a geographic map of sampling locations (Figure 3.3A) and a SNP-based MDS map (Figure 3.3B) in worldwide populations. The value of t_0, the permutation test statistic obtained from the unpermuted data, is represented by the blue vertical line, and it equals 0.799 ($P < 0.0001$).

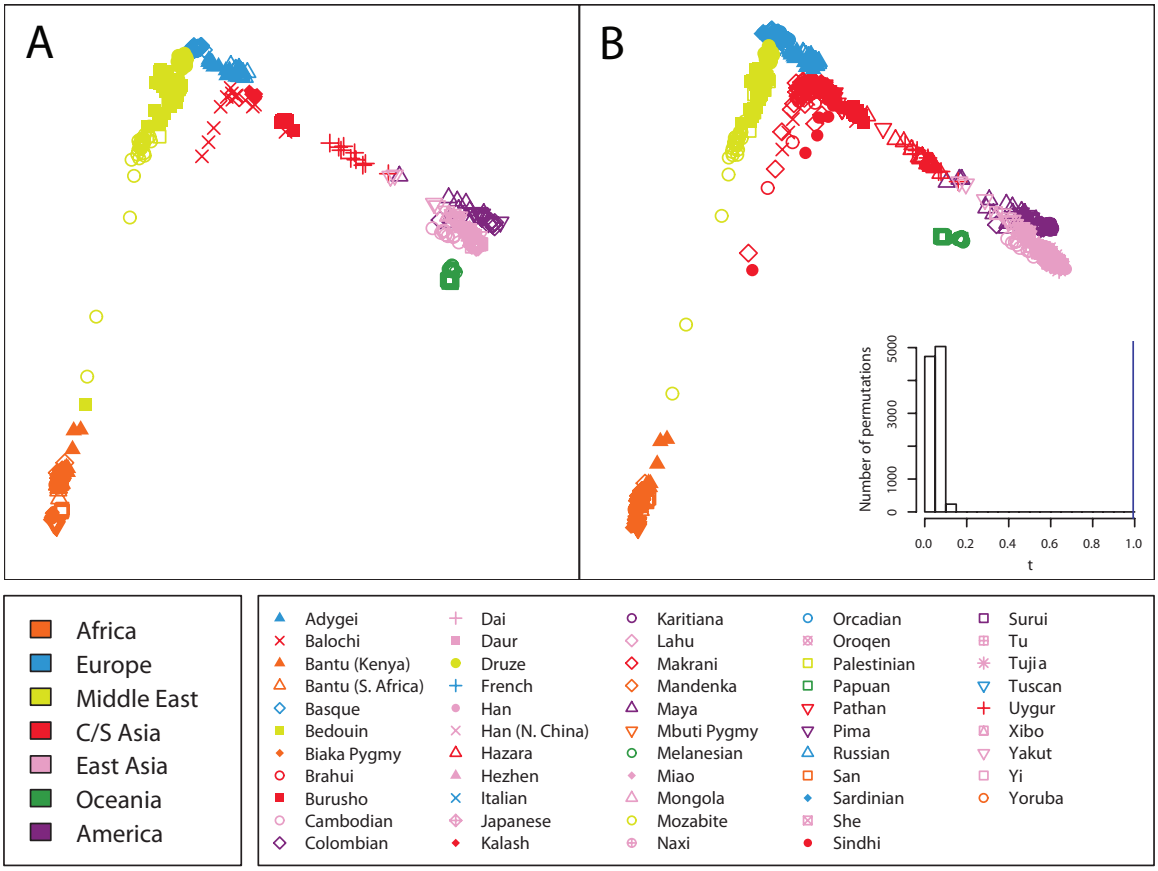

Figure 3.5: Procrustes analysis of genetic coordinates obtained using MDS and PCA. (A) MDS plot of genetic variation for 443 individuals from 29 worldwide populations, based on data from *Jakobsson et al.* (2008). (B) Procrustes-transformed PCA plot of genetic variation for 944 individuals from 52 worldwide populations, based on data from *Biswas et al.* (2009). The Procrustes analysis is based on a subset of 433 individuals included in both datasets. Note that unlike Biswas *et al.*, our plot splits the Han and Han (N. China) groups, so that the 944 individuals are separated into 53 populations rather than 52. A histogram of the t statistic across 10,000 permutations appears in the lower right corner ($t_0 = 0.993$, $P < 0.0001$).

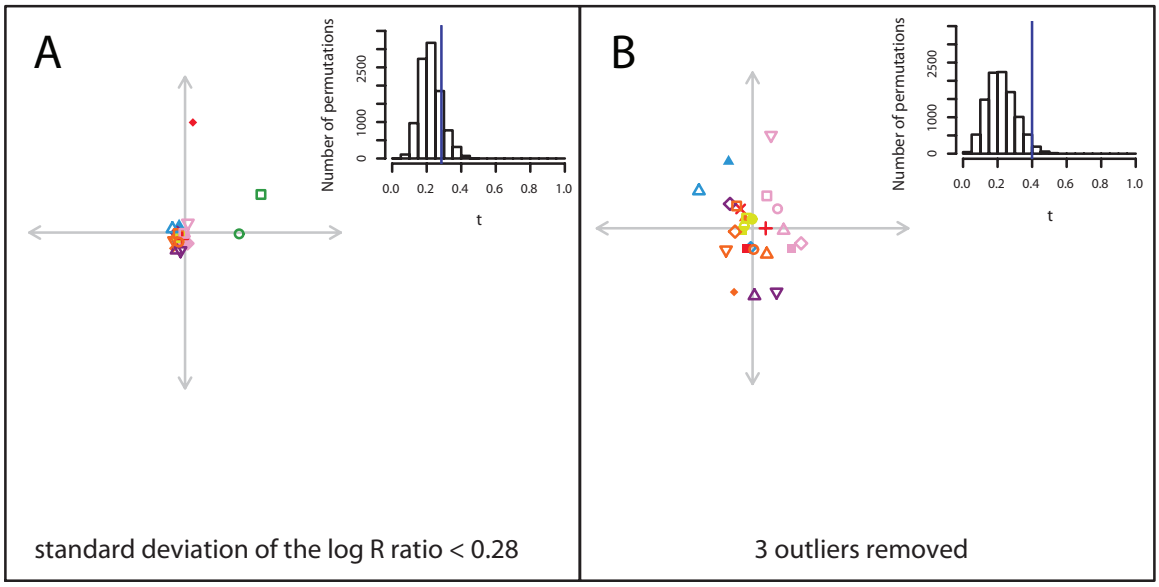

Figure 3.6: Procrustes analysis of CNV-based MDS genetic coordinates. (A) Procrustes-transformed MDS plot for CNV data, aligned to the SNP-based MDS plot in Figure 3.3B. A histogram of the t statistic across 10,000 permutations appears in the upper right corner ($t_0 = 0.285$, $P = 0.1536$). A version of the MDS plot without the Procrustes transformation appeared in Figure S14 of *Jakobsson et al.* (2008). (B) Procrustes-transformed CNV-based MDS plot, excluding three outliers, aligned to the restriction of the SNP-based MDS plot in Figure 3.3B to the 26 non-outlier populations. The three outlier populations are Kalash, Melanesian, and Papuan. A histogram of the t statistic across 10,000 permutations appears in the upper right corner ($t_0 = 0.400$, $P = 0.0292$). The population labels and colors follow those of Figure 3.3, and for each graph, the center of the cross represents the centroid of the points plotted.

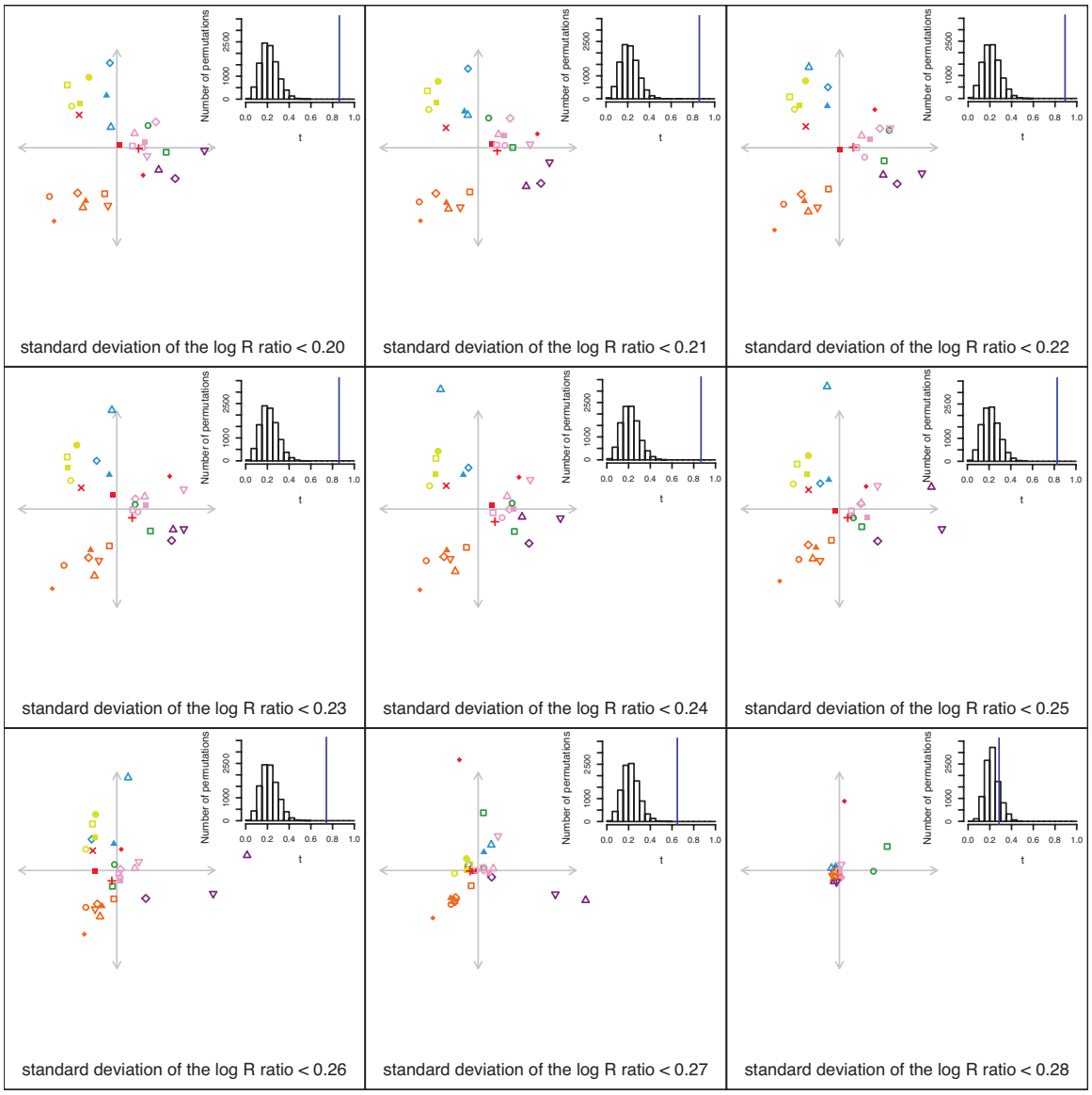

Figure 3.7: Procrustes analysis of CNV-based MDS genetic coordinates, for nine separate choices of the cutoff on s for inclusion of samples in the CNV data. Each graph represents a Procrustes-transformed MDS plot for the CNV data based on a particular choice of the cutoff on s, aligned to the SNP-based MDS plot in Figure 3.3B. The $s < 0.28$ MDS plot is the same as the plot in Figure 3.6A. In increasing order of the cutoff on s, the values of t_0 are 0.862, 0.859, 0.892, 0.860, 0.867, 0.827, 0.742, 0.648, and 0.285. For the cutoff of 0.28, $P = 0.1536$, and for all other cutoffs, $P < 0.0001$. The population labels and colors follow those of Figure 3.3, and for each graph, the center of the cross represents the centroid of the points plotted.

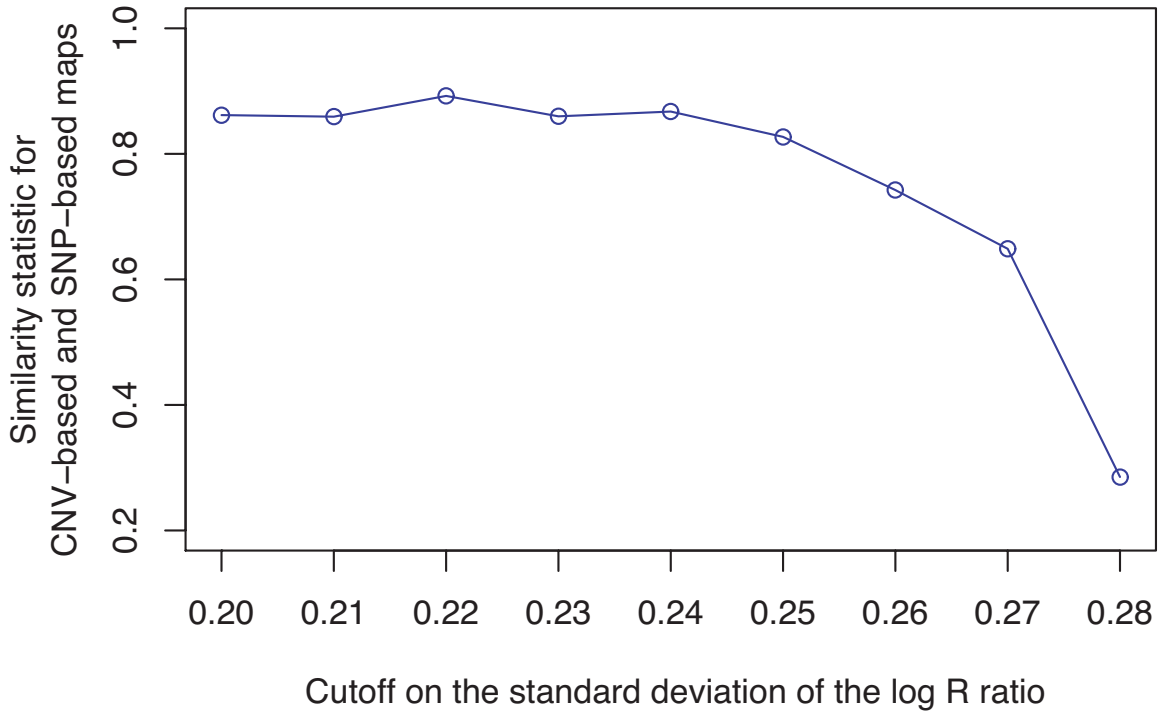

Figure 3.8: Relationship of the t_0 similarity statistic between CNV-based and SNP-based MDS plots and the cutoff on the standard deviation of the log R ratio.

Table 3.1: Sizes of CNV datasets reduced according to cutoffs on the standard deviation of the log R ratio.

Cutoff on the standard deviation of the log R ratio (s)	Number of individuals including relatives	Number of individuals excluding relatives	Smallest sample size across populations when excluding relatives	Number of autosomal non-singleton CNV loci when excluding relatives
0.20	351	320	2	208
0.21	371	340	3	231
0.22	386	355	3	243
0.23	402	370	4	255
0.24	413	379	4	272
0.25	418	384	5	285
0.26	425	389	5	298
0.27	431	395	5	332
0.28	443	405	5	396

Table 3.2: Number of unrelated individuals in each of 29 populations, in CNV datasets reduced according to cutoffs on the standard deviation of the log R ratio.

	Number of unrelated individuals in reduced CNV datasets								
Population	$s < 0.20$	$s < 0.21$	$s < 0.22$	$s < 0.23$	$s < 0.24$	$s < 0.25$	$s < 0.26$	$s < 0.27$	$s < 0.28$
Adygei	9	10	10	12	12	12	12	13	13
Balochi	11	12	13	14	14	14	14	14	14
Bantu (Kenya)	10	10	10	10	10	10	10	11	11
Bantu (S. Africa)	7	7	7	7	7	7	7	7	7
Basque	6	7	11	11	11	11	11	11	11
Bedouin	37	40	40	40	40	40	41	41	41
Biaka Pygmy	19	19	19	21	22	22	22	22	23
Burusho	5	5	5	6	6	6	6	6	6
Cambodian	10	10	10	10	10	10	10	10	10
Colombian	7	7	7	7	7	7	7	7	7
Daur	8	8	8	8	9	9	9	10	10
Druze	31	32	33	33	33	34	34	34	35
Kalash	2	5	5	6	6	6	6	7	12
Lahu	8	8	8	8	8	8	8	8	8
Mandenka	20	20	20	20	21	22	22	22	22
Maya	3	4	4	4	4	7	8	8	8
Mbuti Pygmy	9	10	11	11	12	12	12	12	12
Melanesian	5	5	6	6	6	6	6	6	7
Mongola	6	7	7	9	9	9	9	9	9
Mozabite	26	28	28	28	28	28	28	28	28
Palestinian	19	20	21	22	23	23	23	23	23
Papuan	7	7	7	8	8	8	8	10	12
Pima	2	3	3	4	5	5	5	5	5
Russian	3	5	7	9	12	12	13	13	13
San	5	5	6	6	6	6	6	6	6
Uygur	9	9	9	9	9	9	9	9	9
Yakut	6	7	9	10	10	10	12	12	12
Yi	8	8	9	9	9	9	9	9	9
Yoruba	22	22	22	22	22	22	22	22	22

CHAPTER IV

A quantitative comparison of the similarity between genes and geography in worldwide populations

4.1 Introduction

The geographic structure of human genetic variation has long been of interest for its implications for studying human evolutionary history (*Sokal et al.*, 1991; *Cavalli-Sforza et al.*, 1994; *Barbujani*, 2000; *Cavalli-Sforza and Feldman*, 2003; *Novembre and Ramachandran*, 2011). In recent years, the expansion of population-genetic datasets has contributed to an increase in geographic investigations of human genetic variation, often on the basis of classic multivariate statistical techniques such as PCA and MDS (*Ramachandran et al.*, 2005; *Li et al.*, 2008; *Jakobsson et al.*, 2008; *Novembre et al.*, 2008; *Biswas et al.*, 2009). In PCA, samples are projected onto a series of orthogonal axes (principal components or PCs) that are constructed from a linear combination of genotypic values across genetic markers, such that each PC sequentially maximizes the variance among samples projected on it (*Menozzi et al.*, 1978; *Patterson et al.*, 2006). Classic MDS analyzes a genetic distance matrix between pairs of samples and places the samples in a low-dimensional space in such a way that pairwise Euclidean distances among samples in the low-dimensional space approximate their relative

genetic distances (*Cox and Cox*, 2001). The population structure of genetic variation is often summarized in easily visualized two-dimensional statistical maps obtained from the first two components of PCA or MDS. Especially for large-scale single-nucleotide polymorphism (SNP) data, PCA and MDS are popular because of their computational efficiency and high level of resolution in decomposing the complex structure of human genetic variation (*Patterson et al.*, 2006; *Paschou et al.*, 2007). Generally, results produced by PCA and MDS are very similar to each other (*Wang et al.*, 2010).

Several recent studies have reported detectable similarity between statistical maps of genetic variation and geographic maps of population locations. Such observations are particularly prominent within Europe, where striking similarity between genes and geography is observed both at a continental level (*Novembre et al.*, 2008; *Lao et al.*, 2008; *Heath et al.*, 2008) and in more localized studies such as in Finland (*Jakkula et al.*, 2008; *Hoggart et al.*, 2012), Iceland (*Price et al.*, 2009), and Sweden (*Salmela et al.*, 2011). Analogous but visually less striking observations have also been reported in studies of other geographic regions, including in worldwide samples (*Ramachandran et al.*, 2005; *Li et al.*, 2008; *Jakobsson et al.*, 2008; *Biswas et al.*, 2009; *Xing et al.*, 2009, 2010) and in samples from Asia (*Xing et al.*, 2010; *The HUGO Pan-Asian SNP Consortium*, 2009; *Tian et al.*, 2008), Africa (*Bryc et al.*, 2010b; *Sikora et al.*, 2011), China (*Chen et al.*, 2009; *Xu et al.*, 2009), and Japan (*Yamaguchi-Kabata et al.*, 2008). However, this similarity of genes and geography is in many cases reported in a qualitative sense and has not been assessed systematically across different studies, so that it has been difficult to compare levels of agreement between genes and geography in different regions. Further, different studies have used different sets of genetic markers and different statistical techniques (e.g. PCA and MDS), further complicating comparisons across datasets. Even for studies that used PCA, several versions of this technique have been employed in different studies.

For example, some studies have performed PCA on genotypic matrices (*Novembre et al.*, 2008; *Biswas et al.*, 2009; *Patterson et al.*, 2006; *Price et al.*, 2009), whereas others have applied PCA on pairwise genetic distance matrices (*Li et al.*, 2008; *Xing et al.*, 2009, 2010).

A formal comparison of genes and geography in different regions using a single technique and a common marker set can provide a systematic basis for evaluating the role of geography in explaining the genetic similarity of individuals or populations in different locations. We have previously developed a Procrustes analysis approach to quantify the similarity between statistical maps of genetic variation and geographic maps (*Wang et al.*, 2010). This approach identifies data transformations that minimize the sum of squared Euclidean distances between two sets of coordinates while preserving relative pairwise distances among points within each set. The statistical significance of the similarity between genetic coordinates and geographic coordinates is then examined using a permutation test.

In this study, we apply the Procrustes approach together with PCA to systematically study the geographic structure of human genetic variation across different geographic regions. By compiling data from a variety of published sources (*Novembre et al.*, 2008; *Xing et al.*, 2010; *Bryc et al.*, 2010b; *Pemberton et al.*, 2012; *Simonson et al.*, 2010), we have assembled genome-wide SNP data and geographic coordinates for 149 populations worldwide. Based on a common set of autosomal SNP markers shared by datasets collected from different studies, we evaluate the similarity between genes and geography in examples from Europe, Sub-Saharan Africa, Asia, East Asia, and Central/South Asia, as well as in a worldwide sample. We compare the level of similarity across the various datasets, finding that all show a high level of similarity, and that the highest similarity score appears in Asia. We also examine the dependence of the similarity on the choice of populations included in the analysis and on the number of markers studied. Our results provide information about the importance of

geography in human evolutionary history, and can facilitate statistical methods for inferring the ancestral origin of human individuals from their genotypes.

4.2 Results

We integrated published genome-wide SNP data on 4,257 individuals from 149 worldwide populations, taking data from the Human Genome Diversity Project (HGDP) (*Li et al.*, 2008; *Pemberton et al.*, 2012), International Haplotype Map Project Phase III (HapMap Phase III) (*Pemberton et al.*, 2012; *The International HapMap 3 Consortium*, 2010), and POPRES (*Novembre et al.*, 2008) samples, as well as from several other publications (*Xing et al.*, 2010; *Bryc et al.*, 2010b; *Simonson et al.*, 2010). In our analyses, we focused on the data from 128 populations (Tables S4.1-S4.3). We constructed six datasets for evaluating the geographic structure of genetic variation in different geographic regions: a worldwide sample, continental samples from Europe, Sub-Saharan Africa, and Asia, and subcontinental samples from East Asia and Central/South Asia (Table 4.1).

Our analyses were based on 32,991 autosomal SNP markers that were shared among datasets obtained from different genotyping platforms. We applied PCA on datasets after quality control and removal of PCA outliers (see *Materials and Methods*), and we then used Procrustes analysis to compute the similarity score, denoted as t_0, between the first two PCs of genetic variation and the geographic coordinates of the populations.

We evaluated the statistical significance of the similarity score by permutation. We further examined the robustness of our results using a leave-one-out approach, in which we repeated PCA and Procrustes analysis on datasets with a single population excluded. PCA coordinates obtained from these new datasets were compared to the original PCA coordinates obtained from the whole dataset and to the geographic coordinates, with the respective Procrustes similarity scores denoted as t' and t'' (see

Materials and Methods). These analyses were applied systematically on all datasets.

4.2.1 Worldwide sample

Our worldwide example was based on 938 unrelated individuals from 53 worldwide populations (Fig. 4.1A), taken from the study of *Li et al.* (2008). None of these individuals was found to have >5% missing data or to appear as a PCA outlier.

A PCA plot finds that as in previous studies (*Li et al.*, 2008; *Jakobsson et al.*, 2008; *Biswas et al.*, 2009), samples from the same geographic region (indicated by colors in Fig. 4.1) generally cluster together, and that different clusters align on the PCA plot in a way that qualitatively resembles the geographic map of sampling locations. The first two PCs of our PCA explain 6.22% and 4.72% of the total genetic variation, respectively. These values are considerably less than the values reported by *Li et al.* (2008) in their Fig. S3B, which were 52.3% for PC1 and 27.8% for PC2. The difference can be attributed primarily to the different versions of PCA used in the analyses. We applied PCA on the $N \times L$ genotypic matrix for N individuals and L loci, whereas Li *et al.* applied PCA on an $N \times N$ matrix recording levels of identity-by-state for pairs of individuals (*Li et al.*, 2008). Although the two approaches provide visually similar PCA plots, the values and the interpretation of the proportions of variance explained by each PC differ, as they are based on quite distinct computations.

Using Procrustes analysis, we identified an optimal alignment of the genetic coordinates to the (Gall-Peters-projected) geographic coordinates that involved a rotation of the PCA plot by 31.91° counterclockwise. The genetic coordinates were then superimposed on the geographic map by applying the optimal transformation, thereby highlighting the similarity between genes and geography (Fig. 4.1). This qualitative resemblance is demonstrated by the Procrustes similarity score of $t_0 = 0.705$, which is highly significant in 100,000 permutations ($P < 10^{-5}$). Applying the leave-one-out approach with populations excluded individually, the similarity score between genes

and geography ranges from 0.697 to 0.715, with mean 0.705 and standard deviation 0.003 (Table S4.4). Some populations, such as Native American and Oceanian populations, align in Fig. 4.1B distantly from their geographic locations. In most but not all cases, excluding one of these populations leads to an increase in the Procrustes similarity score.

4.2.2 Europe

Visually striking similarity betweeen PCA plots of genetic variation and a geographic map of Europe has been reported by several studies (*Novembre et al.*, 2008; *Lao et al.*, 2008; *Heath et al.*, 2008). Our analysis was based on nearly the same sample studied by Novembre *et al. Novembre et al.* (2008), containing 1,385 individuals from 37 populations widely spread across Europe (Fig. 4.2A). After excluding five individuals with >5% missing data and two PCA outliers, our final analysis examined 1,378 individuals.

Our PCA plot is very similar to the plot of *Novembre et al.* (2008), with a close correspondence of genes and geography (Fig. 4.2B). One difference is that in the PCA plot of *Novembre et al.* (2008), individuals are more widely spread along PC2 than in our plot. As we applied PCA in the same way as *Novembre et al.* (2008), the difference arises primarily because they employed coordinates given directly by the eigenvectors in PCA, such that PC1 and PC2 were scaled to have the same variance (J. Novembre, personal communication). To simplify the standardization of analyses across datasets, we chose not to scale the PC axes in our analyses, so that the relative amounts of genetic variation explained by each PC are reflected in the PCA plot (see *Material and Methods*). Our PC1 and PC2 explain 0.30% and 0.16% of the total genetic variation respectively, in close agreement with the values of 0.30% and 0.15% reported by *Novembre et al.* (2008).

We used Procrustes analysis to superimpose the PCA plot on the geographic map,

rotating the PCA coordinates 72.66° clockwise (Fig. 4.2). The rotated genetic coordinates of the European samples are spread over a larger distance along longitudinal lines than along latitudinal lines, although the geographic locations of the samples are distributed in the opposite way. This observation reflects the result that the genetic differentiation among Europeans is larger in a north-south direction than in an east-west direction (*Auton et al.*, 2009). The Procrustes similarity between the genetic coordinates and the geographic coordinates is $t_0 = 0.780$ ($P < 10^{-5}$). Excluding populations from the analysis individually, the Procrustes similarity between genes and geography ranges from 0.764 for the analysis without the United Kingdom to 0.810 for the analysis without Italy, with a mean of 0.780 across populations and a standard deviation of 0.007 (Table S4.5). Populations that have a relatively large effect on the similarity score are mostly those with large sample sizes (e.g., Italy, Portugal, Spain, and the United Kingdom). The Russian population is an exception; its sample size is small ($n = 6$), but the genetic coordinates of the Russian sample align poorly with the geographic coordinates (*Novembre et al.*, 2008) (Fig. 4.2). Thus, this population has a relatively large effect on the similarity with geography ($t'' = 0.788$ when excluding Russians, Table S4.5). Excluding Russians has minimal effect on the PCA coordinates for the remaining samples, however, as reflected in the high similarity score between the PCA coordinates before and after excluding the Russian sample ($t' = 1.000$, Table S4.5). Reducing the sizes of large samples also has a relatively small impact; when repeating our analyses on a subset of the data in which 50 individuals are selected randomly from populations with larger samples, t_0 changes slightly to 0.777, and both F_{ST} and the proportions of variance explained by PC1 and PC2 undergo slight increases (Fig. S4.1).

4.2.3 Sub-Saharan Africa

Sub-Saharan Africa is the location of the origin of modern humans and has the highest genetic variation among all continents (*Li et al.*, 2008; *Xing et al.*, 2009; *Bowcock et al.*, 1994; *Rosenberg et al.*, 2002; *Tishkoff et al.*, 2009). Previous studies have found that when isolated hunter-gatherer populations are included in the analysis, PCA plots of genetic variation in Sub-Saharan Africa display low qualitative similarity to the geographic map of sampling locations (*Li et al.*, 2008; *Xing et al.*, 2009; *Henn et al.*, 2011). Bryc *et al.* recently studied 12 populations in West Africa, and revealed a high similarity between a SNP-based PCA map and the corresponding geographic map, when Mbororo Fulani, a nomadic pastoralist population, was excluded from the analysis (*Bryc et al.*, 2010b). By integrating SNP data from multiple sources (*Xing et al.*, 2010; *Bryc et al.*, 2010b; *Pemberton et al.*, 2012), we investigated Sub-Saharan African populations in a broader region than in the analysis of *Bryc et al.* (2010b). We first excluded four hunter-gatherer populations (!Kung, San, Biaka Pygmy, and Mbuti Pygmy) and Mbororo Fulani. After further excluding six individuals with >5% missing data and two PCA outliers, our analyses examined 348 individuals from 23 populations in Sub-Saharan Africa (Fig. 4.3A).

Applying PCA on this combined Sub-Saharan African dataset, we found that PC1 accounts for 1.34% of the total genetic variation, largely separating populations from west to east. PC2 accounts for 0.69% of the total genetic variation and largely separates populations from north to south (Fig. 4.3B). Generally, populations along the west coast of Africa cluster closely with each other, while interior populations form relatively isolated clusters. Bantu-speaking populations tend to cluster with each other, and can be divided into three groups according to their geographic locations: two populations in the west (Fang and Kongo), two in the east (Kenyan Bantus from the HGDP and Luhya), and five in the south (Southern African Bantus from the HGDP, Nguni, Pedi, Sotho/Tswana, and Xhosa). Despite the large geo-

graphic separation among these three groups, their genetic separation in the PCA plot is relatively small (Fig. 4.3B). In particular, Luhya and Kenyan Bantus from the HGDP align between the western Bantu populations and the eastern non-Bantu populations such as Alur and Hema. The Maasai sample, consisting of 30 unrelated individuals randomly selected from the HapMap Phase III (*Pemberton et al.*, 2012; *The International HapMap 3 Consortium*, 2010), forms a cluster distant from the other populations along PC1 (and PC3, results not shown).

Procrustes analysis identifies a rotation angle of 16.11° counterclockwise for the genetic coordinates (Fig. 4.3B), and the similarity score between genes and geography is $t_0 = 0.790$ ($P < 10^{-5}$). Among all populations, Maasai has the largest impact on both the PCA and Procrustes analysis (Table S4.6); as shown in Fig. S4.2, when analyzed without Maasai, the other 22 populations align more closely with geography, and the Procrustes similarity score increases to 0.832 ($P < 10^{-5}$). Excluding any of the populations in South Africa leads to a decrease of the similarity between genes and geography, and the lowest similarity is obtained when excluding the combined Sotho/Tswana sample ($t'' = 0.768$, Table S4.6). This result suggests that the genetic map of Sub-Saharan Africans might look more similar to the geographic map if additional populations from the undersampled southern region of Africa were included.

When hunter-gatherer populations (!Kung, San, Biaka Pygmy, and Mbuti Pygmy) and Mbororo Fulani were included in the analysis, they appeared as isolated clusters on the PCA plots and greatly reduced the similarity between PCA maps and geographic maps (Fig. S4.3, Table S4.7). The similarity score decreased from 0.790 to 0.548 after including all five of these populations in the analysis. This value, however, is still statistically significant, with a P-value of 4.0×10^{-4}; further, if we disregard the hunter-gatherer populations and Mbororo Fulani in Fig. S4.3B and only examine the relative locations of the original 23 populations, we can still find a clear resemblance between genetic and geographic coordinates. Compared to the other 23 populations,

the four hunter-gatherer populations appear as isolated groups at the south, and Mbororo Fulani appears at the north. These observations are clearer in plots with only one among the five outlier populations included at a time (Figs. S4.3C-G), each of which also produces significant similarity scores between genetic and geographic coordinates (Fig. S4.4, Table S4.7).

4.2.4 Asia

Our Asian example included 760 individuals from 44 populations distributed widely across Asia (Fig. 4.4A). Previous studies based on largely overlapping datasets have reported correlations between genetic and geographic distances across Eurasia (*Xing et al.*, 2009, 2010). Our dataset combined data from these studies as well as from *Li et al.* (2008) and *Simonson et al.* (2010), and after excluding 11 PCA outliers, our final dataset for Asia contains 749 individuals.

In our PCA plot (Fig. 4.4B), PC1 largely separates populations on different sides of the Himalayas, accounting for genetic variation in an east-west direction. PC2, on the other hand, distinguishes northern and southern populations. PC1 accounts for 5.42% of the total genetic variation, a much larger value than the 0.85% captured by PC2, reflecting large genetic distances between populations separated by the Himalayas. Interestingly, populations around the Himalayas form a ring shape on the PCA plot, with the Nepalese population from the Himalaya region aligning in the middle. As noted by *Xing et al.* (2010), the Nepalese samples were collected from different subgroups that have different levels of ancestry shared with Central/South Asians and East Asians, and the dispersion of the Nepalese sample is therefore not unexpected. Tibetans, on the northern side of the Himalayas, do not spread over a large area in the plot and are well clustered with other East Asian populations.

One interesting result concerns the Uygur and Kyrgyzstani populations, both of which lie along ancient trade routes between Europe and East Asia. Compared to the

Uygur population, which lies farther to the east, the Kyrgyzstani population clusters closer to East Asian populations, especially to the Yakut and Buryat populations, supporting a view that the Kyrgyzstani group has a proportion of its ancestry in Siberia (*Bregel*, 2003). A third population sampled from near the Uygur and Kyrgyzstani populations is the Xibo population, which clusters clearly with East Asians from northeastern China. This pattern matches the expectation given documentation that this Xibo group moved in 1764 from northeastern China to Xinjiang province (*Du and Yip*, 1993; *Powell et al.*, 2007).

The PCA map of genetic variation in Asia is rotated 5.05° counterclockwise in the Procrustes superposition on the geographic map (Fig. 4.4B). Despite the discontinuity caused by the Himalayas, most populations align in a way that is highly concordant with their geographic locations. This observation is confirmed by a Procrustes similarity score of $t_0 = 0.849$ ($P < 10^{-5}$). Among all populations, the tribal population Irula, which appears south of India as an isolated cluster in Fig. 4.4B, has the largest impact among all populations on the Procrustes similarity with geography (Table S4.8). When excluding Irula, the PCA map aligns more closely with geography, with the Procrustes similarity increasing to 0.871 ($P < 10^{-5}$, Fig. S4.5). This exclusion generates increased separation on the PCA map for some populations. For example, in Fig. S4.5, Iban from Sarawak is more clearly distinguished from other Southeast Asian populations. Overall, the similarity score between genes and geography in Asia is robust to the exclusion of any one population, with the lowest Procrustes similarity score of $t'' = 0.839$ occurring when the Buryat population is excluded (Table S4.8).

4.2.5 East Asia

To further examine populations on either side of the Himalaya Moutains, we performed additional analyses of East Asia and Central/South Asia. We first considered

the East Asian populations in our Asian example. This dataset consists of 341 individuals from 23 populations. After excluding seven PCA outliers, our analyses were based on 334 individuals from 23 East Asian populations (Fig. 4.5A).

Individuals in this East Asian dataset generally align along a curve on the PCA plot. PC1 explains 1.58% of the total genetic variation and largely accounts for a north-south genetic gradient; PC2 explains 0.98% of the genetic variation and mainly separates two Siberian populations (Buryat and Yakut) and three Southeast Asian populations (Cambodians, Iban, and Thai) from the other East Asian populations (Fig. 4.5B). The Tibetan population is also separated by PC2, but on the opposite side to the Siberians and Southeast Asians. Overall, PC1 largely matches geography in the north-south direction, and PC2 shows only a partial similarity to the east-west direction.

The imperfect match between PCA coordinates and geography is reflected by a relatively low Procrustes similarity score of $t_0 = 0.640$, which, however, is still statistically significant with $P = 0.00038$. The optimal transformation rotates the PCA map 67.27° counterclockwise prior to superposition on the geographic map (Fig. 4.5B). Interestingly, excluding populations one at a time, we found that the PCA coordinates were reflected over PC1 when Procrustes-transformed to match the geographic coordinates if either the Iban, Tibetan, or Yakut population was excluded (Fig. S4.6). Such abrupt changes of the Procrustes transformation are consistent with the fact that PC2 matches less closely with geography; a reflection over PC1 has a small effect on the similarity score. The Procrustes similarity score with geography can be substantially increased by excluding Japanese ($t'' = 0.755$, $P < 10^{-5}$); other than the Japanese population, Iban, Thai, and Yakut have the largest effect on the similarity scores both with geography and with the original PCA (Table S4.9).

4.2.6 Central/South Asia

Our last example focused on Central/South Asia, using an initial sample of 372 individuals from 18 populations. Ten individuals were excluded as PCA outliers, leaving 362 individuals from 18 populations for the final analysis (Fig. 4.6A).

The first two components of the PCA anlaysis account for 1.59% and 1.31% of the total genetic variation, respectively. Overall, the PCA pattern for the separate anlaysis of Central/South Asian populations is similar to the pattern for the same set of populations in our analysis of all of Asia (Fig. 4.4). After rotating the PCA coordinates 11.78° counterclockwise, we obtained a Procrustes similarity score of 0.737 ($P < 10^{-5}$) when comparing PCA coordinates to geography (Fig. 4.6B). Most populations from Pakistan cluster closely on the first two PCs except for the Hazara population, which clusters with the Uygur population and aligns distantly from its sampling location. When excluding Hazara, the Procrustes similarity score to geography increases from 0.737 to $t'' = 0.769$, larger than for any other exclusion (Table S4.10). Excluding Irula has the second largest effect on the similarity score to geography, but more interestingly, this exclusion has the largest effect on the PCA coordinates (smallest value for t' in Table S4.10). A closer examination of the PCA results reveals that when Irula is excluded, the Kalash population in Pakistan is separated from the other Pakistani populations and appears as an isolated group in the north (results not shown). This result accords with the identification of this isolated group as distinct in previous studies (*Jakobsson et al.*, 2008; *Rosenberg et al.*, 2002).

4.2.7 Comparison across geographic regions

We have found that significant similarity between genes and geography exists in general at different geographic levels (Table 4.2). The highest similarity score was found in the data from Asia, followed by Sub-Saharan Africa when five outlier populations were excluded, and by Europe. Five of the six analyses had P-values smaller

than 10^{-5}, and only the data from East Asia had a nonzero P-value estimate in 100,000 permutations. When comparing the permutation distributions of the similarity score (Fig. 4.7), however, a difference in the significance levels is evident for the five examples with $P < 10^{-5}$. The worldwide and Asian datasets have similarity scores t_0 considerably exceeding the similarity scores from all 100,000 permutations (Figs. 4.7A and D). By contrast, although the European, Sub-Saharan African, and Central/South Asian datasets have similarity scores higher than that of the worldwide dataset, their similarity scores are closer to the corresponding permutation distributions (Figs. 4.7B, C, and F), indicating relatively high P-values compared to the worldwide data.

To examine the robustness of our results to the number of SNPs analyzed, we repeated our analyses with subsets of randomly selected loci. We found that our Procrustes similarity scores between genes and geography are quite robust as long as enough SNPs (>10,000) are used (Fig. 4.8). Indeed, for the worldwide and Asian datasets, ∼1,000 SNPs are sufficient to obtain a similarity score between genes and geography close to the score obtained using all 32,991 SNPs. For the African, East Asian, and Central/South Asian datasets, the number of SNPs needed increases to ∼4,000. Interestingly, many more SNPs are required for the European dataset to reach a high similarity score between genes and geography. Although the increase of the similarity score for the European dataset becomes slow when the number of SNPs exceeds 10,000, it continues even when the number of SNPs is as high as ∼30,000 (Fig. 4.8). If we use the same 197,146 SNPs as used by *Novembre et al.* (2008), the similarity score between genes and geography for the European example would become 0.799, slightly higher than the value for our Sub-Saharan African example based on 32,991 SNPs. This larger number of SNPs required might reflect a relatively homogeneous population structure in Europe that requires more genetic markers to characterize subtle differentiation.

To explore the relationship between genetic differentiation and the number of SNPs required to produce convergence in the Procrustes similarity, we computed F_{ST} across populations, a measurement of population differentiation, for all of our datasets, on the basis of the 32,991 autosomal SNP markers. We found $F_{ST} = 0.212\%$ for the European dataset, much smaller than the values of 9.704% and 4.706% for the worldwide and Asian datasets. The values of F_{ST} for the Sub-Saharan Africans (without outlier populations), the East Asians, and the Central/South Asians are 1.334%, 1.874% and 2.140%, respectively. As expected, datasets that have less population differentiation, as indicated by smaller F_{ST} values, need more markers to reveal geographic structure in the PCA plot, consistent with a previous finding that the dataset size required for the population structure to be evident in PCA is inversely related to F_{ST} (*Patterson et al.*, 2006). Further, we found F_{ST} and the sum of the proportions of variance explained by PC1 and PC2 to be positively correlated (Pearson correlation $r = 0.996$, Fig. 4.9). This strong linear correlation is not surprising because of the connection between F_{ST} and the proportions of variance: F_{ST} can be computed as the proportion of the variance in an allelic indicator variable contributed by between-population differences (*Weir*, 1996). It has been shown under a two-population model that the proportion of the total variance explained by PC1 is approximately equal to F_{ST} (*McVean*, 2009). Here, we have observed a qualitatively similar relationship.

4.3 Discussion

Both simulation-based and theoretical studies have shown that under spatial models in which migration and gene flow occur in a homogeneous manner over short distances, a similarity between PCA maps of genetic variation and geography is predicted (*McVean*, 2009; *Novembre and Stephens*, 2008). In this study, we have systematically assessed this similarity in different geographic regions using a shared set of autosomal SNPs and a shared statistical approach. We have found that although they gener-

ally explain a relatively small proportion of the total genetic variation, the first two principal components in PCA often produce a map that resembles the geographic distribution of sampling locations. Our results quantitatively demonstrate the general existence in different geographic regions of a considerable similarity between genes and geography, supporting the view that geography, in the form of incremental migration and gene flow primarily with nearby neighbors, plays a strong role in producing human population structure.

One particularly interesting observation concerns our analysis of the Asian dataset. Asia contains the Himalaya region, a strong geographic barrier to gene flow that has generated noticeable genetic differentiation between populations on opposite sides (*Rosenberg*, 2011). Such barrier effects can produce a distortion of PCA maps from those expected under homogeneous isolation-by-distance models (*McVean*, 2009; *Novembre and Stephens*, 2008), leading to a decrease in the similarity to geography. However, although the concordance of a PCA plot with geography is perhaps best known for Europe — which does not have a barrier of comparable importance to the Himalayas — we obtained the unexpected result that in spite of the Himalaya barrier, the Procrustes similarity score t_0 was actually highest in Asia. When further examining the population structure on separate sides of the Himalayas, we found lower similarity scores between genes and geography in our East Asian and Central/South Asian samples. Especially for the East Asian sample, our results indicate weaker correlation between genes and geography in the east-west direction.

To make the similarity scores between genes and geography commensurable for different datasets, we performed our analyses with the same markers and the same statistical approach. However, one aspect of the analysis that is not homogeneous across datasets is the nature of the geographic coordinates. For example, while most of the analyses employed population sampling locations, for the European dataset, coordinates did not necessarily represent sampling locations. Sampling locations may

also vary in the extent to which they represent long-term locations where groups have resided. One example that highlights this issue is the Xibo population, which was sampled in northwestern China, but which clusters genetically with populations in northeastern China (Fig. 4.5). This group is known to have migrated westward from near Shenyang in northeastern China about 250 years ago (*Du and Yip*, 1993; *Powell et al.*, 2007), and if we were to use the coordinates of Shenyang (41.8°N, 123.4°E) for Xibo rather than the sampling location, t_0 would increase from 0.640 to 0.654 for the East Asian dataset, from 0.849 to 0.859 for the Asian dataset, and from 0.705 to 0.709 for the worldwide dataset.

Additional limitations apply to our geographic analysis. In all of the datasets, population-level rather than individual-level coordinates were used, so that all individuals from the same population were assigned to a single geographic location. This approach can potentially obscure substructure within populations. For example, although both the northern and southern Han Chinese groups from the HGDP dataset were assigned to the same location, they can be genetically distinguished from each other, with the northern group clustering closer to the northern populations in China (Fig. 4.5). Use of individual-level coordinates might lead to higher values of the similarity score t_0. Another concern is that the choice of a map projection (including the projection that consists of using unprojected latitudes and longitudes as a rectangular coordinate system) can have different effects in geographic regions at different distances from the equator, as the level of distortion of the surface of the earth varies with the choice of projection. This issue is expected to be of greatest concern in analyses at high latitudes or in datasets with a wide range of latitudes.

We note that theoretical work and simulation studies have found that results from the PCA approach can be sensitive to the sample size distribution over geographic space (*McVean*, 2009; *Novembre and Stephens*, 2008; *Engelhardt and Stephens*, 2010). In most of our analyses excluding one population at a time, patterns in PC1 and PC2

did not differ greatly from analyses in which all populations were included. However, exclusions of genetically distinctive populations, populations that were geographically distant from the center of a dataset, or populations with large sample sizes sometimes had sizeable effects on t_0. In some analyses, particularly in considering the Luhya and Maasai populations from the HapMap, we therefore included only a subset of available individuals in order to reduce the influence of the large sample sizes for these populations. More generally, an analysis of the role of the geographic distribution of the sample can be performed by analysis of subsamples of a full dataset with different levels of geographic unevenness. A previous analysis of population structure inference using *STRUCTURE* for a variety of samples with different geographic distributions did not find a particularly strong role for the geographic dispersion of the sample (*Rosenberg et al.*, 2005), but the issue has not yet been systematically investigated with PCA.

Through a combination of PCA and Procrustes analysis, we have investigated genes and geography using the same standardized approach in different regions. The general observation of a concordance of genes and geography in different regions and at different geographic levels can provide a foundation for refinement of methods for inferring local geographic origin of human individuals from their genotypes (e.g. *Novembre et al.*, 2008; *Hoggart et al.*, 2012; *Yang et al.*, 2012). In addition, our computations illustrate the use of Procrustes analysis in assisting the interpretation of PCA, such as in comparing PCA maps to different types of spatial maps and in assessing the impact of certain populations or individuals on PCA results. Similar applications of PCA and Procrustes approaches can be used to evaluate evolutionary models by comparing PCA maps obtained from observed data to those obtained from simulated data generated by these models. With the incorporation of the Procrustes similarity score for quantifying patterns in PCA, results from PCA can potentially find new uses in additional applications in population-genetic studies.

4.4 Materials and Methods

4.4.1 Genotype data

We examined genome-wide SNP datasets previously reported in several studies (*Novembre et al.*, 2008; *Xing et al.*, 2010; *Bryc et al.*, 2010b; *Pemberton et al.*, 2012; *Simonson et al.*, 2010). The data of *Pemberton et al.* (2012) merged unrelated samples from earlier datasets obtained from the HGDP (*Li et al.*, 2008) and HapMap Phase III (*The International HapMap 3 Consortium*, 2010; *Pemberton et al.*, 2010). Some of the data of *Xing et al.* (2010) were previously reported in an earlier paper of *Xing et al.* (2009).

Because the datasets were genotyped on different genotyping platforms, including Illumina 650K (*Pemberton et al.*, 2012), Illumina Human 1M (*Pemberton et al.*, 2012), Affymetrix 500K (*Novembre et al.*, 2008; *Bryc et al.*, 2010b), Affymetrix NspI 250K (*Xing et al.*, 2010), and Affymetrix 6.0 (*Xing et al.*, 2010; *Pemberton et al.*, 2012; *Simonson et al.*, 2010), we identified a shared set of 32,991 autosomal SNPs included in all five datasets (*Novembre et al.*, 2008; *Xing et al.*, 2010; *Bryc et al.*, 2010b; *Pemberton et al.*, 2012; *Simonson et al.*, 2010). This number was smaller than the maximum possible set of overlapping SNPs shared among these genotyping platforms, because some SNPs were excluded during the quality control procedures of the studies that originally published the data (*Novembre et al.*, 2008; *Xing et al.*, 2010; *Bryc et al.*, 2010b; *Pemberton et al.*, 2012; *Simonson et al.*, 2010). At 6,549 among these 32,991 markers, the datasets from *Novembre et al.* (2008) and *Bryc et al.* (2010b) had genotypes given for opposite strands when compared to the datasets of *Xing et al.* (2010), *Pemberton et al.* (2012), and *Simonson et al.* (2010). In these instances, we converted the genotypes from *Novembre et al.* (2008) and *Bryc et al.* (2010b) to the opposite strand, so that genotypes were consistent across datasets from different sources. In total, we obtained genotype data on 32,991 autosomal SNPs for 4,257

samples from 149 populations worldwide, with dense sampling from Asia, Europe, and Sub-Saharan Africa. In our final dataset, the physical distance between pairs of nearby SNPs has mean 84 kb (median 45 kb).

We next created six datasets at different geographic scales, including a worldwide sample, continental samples for Europe, Sub-Saharan Africa, and Asia, and subcontinental samples from East Asia and Central/South Asia (Fig. S4.7, Table 4.1). For the worldwide example, we included 938 unrelated individuals from 53 populations in the HGDP (*Li et al.*, 2008; *Pemberton et al.*, 2012). For the European sample, we used a set of individuals that was nearly identical to that analyzed by *Novembre et al.* (2008), containing 1,385 individuals from 37 populations defined by ancestral origins. We did not include two French individuals (sample ID 31645 and 32480) that were included by *Novembre et al.* (2008) but that are not found in the release we obtained of the POPRES dataset in the NCBI dbGaP database (*Nelson et al.*, 2008; *Mailman et al.*, 2007). For Sub-Saharan Africa, we integrated data on African populations from three sources (*Xing et al.*, 2010; *Bryc et al.*, 2010b; *Pemberton et al.*, 2012), including 30 unrelated Luhya (LWK) individuals and 30 unrelated Maasai (MKK) individuals, both randomly selected from the HapMap Phase III (*Pemberton et al.*, 2012). Because some populations in Sub-Saharan Africa are known to be genetically distinctive when compared to most other Sub-Saharan Africans (*Li et al.*, 2008; *Jakobsson et al.*, 2008; *Xing et al.*, 2010; *Bryc et al.*, 2010b; *Rosenberg et al.*, 2002; *Tishkoff et al.*, 2009), we created two datasets for Sub-Saharan Africa, one including and the other excluding these distinctive populations (!Kung, San, Biaka Pygmy, Mbuti Pygmy, and Mbororo Fulani). When excluding all five of these populations, we have 356 individuals from 23 Sub-Saharan African populations. Including them, we have 422 individuals from 28 groups. Note that both Pygmy populations that we examined are from the HGDP (*Li et al.*, 2008; *Pemberton et al.*, 2012), and we did not include the Mbuti Pygmy data from *Xing et al.* (2010). Further, we also did not include the Luhya individu-

als from *Xing et al.* (2010); these individuals are a subset of those of the HapMap (*Pemberton et al.*, 2012; *The International HapMap 3 Consortium*, 2010). As in *Xing et al.* (2010), we analyzed three Sotho samples and five Tswana samples together as a single population, labeled as "Sotho/Tswana."

Our sample from Asia has 760 individuals from 44 populations with sampling locations distributed widely across Asia. These data include 27 populations from the HGDP dataset (*Li et al.*, 2008; *Pemberton et al.*, 2012), 16 populations from *Xing et al.* (2010), and one population (Tibetan) from *Simonson et al.* (2010). For populations studied by both *Pemberton et al.* (2012) and *Xing et al.* (2010) (Cambodian, Han Chinese, and Japanese), we only included the HGDP samples from *Pemberton et al.* (2012). Samples for East Asia and Central/South Asia are subsets of the Asian sample. The East Asian sample consists of 341 individuals from 23 populations: 18 populations from the HGDP dataset (*Li et al.*, 2008; *Pemberton et al.*, 2012), 4 populations from *Xing et al.* (2010), and the Tibetan population from *Simonson et al.* (2010). The Central/South Asian sample has 372 individuals from 18 populations in total, including 9 populations each from the HGDP dataset (*Li et al.*, 2008; *Pemberton et al.*, 2012) and the Xing *et al.* dataset (*Xing et al.*, 2010).

We applied two additional processing steps on each dataset to remove samples with high missing data rates and samples that appear to be outliers. First, we removed individuals with more than 5% missing data in the 32,991 SNPs. Next, in each analysis, we used an iterative PCA approach to identify and remove outlier individuals, as outliers can potentially distort PCA maps of genetic variation (*Price et al.*, 2006). After applying PCA on a dataset, individuals greater than 10 standard deviations from the mean PC position on at least one of the top 10 PCs were considered outliers and were removed from the dataset. This procedure was repeated iteratively until no more outliers were detected. For all datasets, only a small proportion of samples were identified as outliers and removed by this procedure (Table 4.1). The

data processing procedures are illustrated in Figs. S4.7-S4.9, and are summarized in Table 4.1. Individuals that were identified as PCA outliers are listed in Table S4.11.

4.4.2 Geographic coordinates

We assigned all individuals from the same population to a single geographic location, as listed in Tables S4.1-S4.3. For the HGDP samples (*Pemberton et al.*, 2012), we used previously reported coordinates as the geographic locations for all populations (Table 1 in *Rosenberg*, 2011). The geographic locations for the European dataset were reported in Table S3 of *Novembre et al.* (2008), and represent countries of origin. The geographic coordinates for the African populations from *Bryc et al.* (2010b) are sampling locations, and we used the values reported by *Tishkoff et al.* (2009) in their Table S1. Geographic coordinates for populations from *Xing et al.* (2010) were kindly provided by J. Xing. For the Tibetan samples, we used the sampling location reported by *Simonson et al.* (2010). For the two HapMap populations included in this study (Luhya and Maasai), we used the sampling locations reported by HapMap (*The International HapMap 3 Consortium*, 2010).

We used longitude and latitude measured in degrees as our geographic coordinates (λ, ϕ) for all datasets except the worldwide dataset. Latitudes in the southern hemisphere and longitudes in the western hemisphere were denoted by negative values. For the worldwide dataset, we shifted the Americas by adding 360° to longitudes smaller than −25°. We then used the Gall-Peters projection, an equal-area projection that preserves distance along the 45°N parallel, to obtain rectangular coordinates $(\pi\lambda\sqrt{2}/360°, \sqrt{2}\sin\phi)$ as our geographic coordinates. For other datasets, we used unprojected longitude-latitude coordinates.

4.4.3 Principal components analysis

We coded the genotype data for each dataset by an $N \times L$ matrix C, in which $C_{i\ell}$ counts the number of copies of a reference allele at locus ℓ of individual i, N is the number of individuals, and L is the number of loci. For autosomal SNPs, $C_{i\ell}$ is 0, 1, 2, or missing. We first ignored missing data and estimated the reference allele frequency among nonmissing genotypes, or $\hat{p}_\ell$. Following the *smartpca* program (*Patterson et al.*, 2006), we standardized the nonmissing entries in C by

$$X_{i\ell} = (C_{i\ell} - 2\hat{p}_\ell)/\sqrt{\hat{p}_\ell(1-\hat{p}_\ell)}, \tag{B1}$$

where X is a matrix with the same dimensions as C. If a locus was monomorphic in a dataset ($\hat{p}_\ell = 0$ or 1), eq. B1 is undefined, and we set all entries in the column of X for this locus to zero. Entries representing missing data were set to zero in X as well.

We performed PCA by applying the function *eigen* in *R* (www.r-project.org) to the $N \times N$ matrix $M = XX^T$ *McVean* (2009). The coordinates of the N individuals on the jth PC are given by $\sigma_j^{1/2}\vec{v}_j$, where σ_j is the jth eigenvalue of M, sorted in decreasing order, and $\vec{v}_j$ is the corresponding eigenvector. The proportion of variance explained by the kth PC is calculated as $\sigma_k / \sum_{j=1}^{J} \sigma_j$, where J is the total number of eigenvectors of M. This quantity measures the variation among individuals along the kth PC direction, relative to the total variance in the standardized genotypic matrix X. In our examples, $L \gg N$, and $J = N - 1$ because X has rank $N - 1$ after standardization (eq. B1).

We note that some studies have used the eigenvectors $\vec{v}_j$ directly as PCs, so that all PCs have equal variance. We follow an alternative convention (*McVean*, 2009; *Hastie et al.*, 2009), reporting PCs using $\sigma_j^{1/2}\vec{v}_j$, so that the proportions of variance explained by each PC are reflected on the PCA plot. In PCA plots superimposed on

geographic maps, because horizontal and vertical axes are plotted on different scales, PC1 and PC2 can appear to not be perpendicular.

4.4.4 Procrustes analysis and permutation test

We applied Procrustes analysis (*Cox and Cox*, 2001; *Wang et al.*, 2010) to compare the individual-level coordinates of the first two components (PC1 and PC2) in the PCA performed on the SNP data to the geographic coordinates. Procrustes analysis minimizes the sum of squared Euclidean distances between two sets of points (two "maps") by transforming one set of points to optimally match the other set, while preserving the relative pairwise distances among all points within maps. Possible transformations include translation, scaling, rotation, and reflection. The similarity between two maps is then quantified by a Procrustes similarity statistic $t_0 = \sqrt{1 - D}$, in which D is the minimum sum of squared Euclidean distances between the two maps across all possible transformations. D, which is given by equation 6 in *Wang et al.* (2010), has been scaled to have minimum 0 and maximum 1. The similarity statistic t_0 therefore also ranges from 0 to 1. In our analyses, we fixed the geographic coordinates and Procrustes-transformed the PCA coordinates in order to superimpose the PCA maps on the geographic maps. In addition to t_0, we also report the rotation angle θ of the PCA map as given by the Procrustes analysis, measured in degrees counterclockwise.

To test the statistical significance of t_0, we used a permutation test. In each permutation, we randomly permuted the population geographic locations, assigning all individuals from the same population to a single geographic location in the permuted dataset. We then applied Procrustes analysis to compute the similarity score t between the PCA coordinates and the randomly permuted geographic coordinates. We calculated the P-value as $\mathbb{P}(t > t_0)$, representing the probability of observing a similarity statistic higher than t_0 under the null hypothesis that no geographic

pattern exists in the population structure. For each dataset, we employed 100,000 permutations for the permutation test.

4.4.5 Analyses with populations excluded individually

We investigated the effect of each population on our PCA and Procrustes analysis using a leave-one-out approach. For each dataset, we excluded one population at a time and repeated PCA to obtain a new set of genetic coordinates (for each population excluded, this PCA started from the same final set of individuals after exclusions owing to missing data and PCA outliers, and we did not repeat the search for outliers). We then performed two Procrustes analyses. In the first one, we compared the new PCA coordinates and the original PCA coordinates obtained before removing any population. This comparison was based on the common set of individuals included in both analyses, and its similarity score was denoted t'. In the second Procrustes analysis, we computed the similarity between the new set of PCA coordinates and the corresponding geographic coordinates, denoting the similarity score by t''.

4.4.6 Subsets of loci

To investigate the effect of the number of markers on our results, we created a series of marker lists by randomly selecting L loci from the 32,991 total loci. These marker lists were selected independently of each other and had $L = 500, 1000, ..., 32500$. We then repeated PCA and Procrustes analysis for each geographic region using genotypes at the loci in each of our marker lists. For Sub-Saharan Africa, we used the dataset that excludes hunter-gatherer populations and the Mbororo Fulani. Given L, the analyses for different geographic regions are based on the same set of markers, so that their results are comparable.

4.4.7 F_{ST} estimation

We calculated F_{ST} in each dataset using Weir and Cockerham's estimator (eq. 10 in *Weir and Cockerham*, 1984) based on all 32,991 loci.

4.5 Acknowledgements

The authors are grateful to Katarzyna Bryc, John Novembre, Trevor Pemberton and Jinchuan Xing for assistance with data from their papers, and John Novembre and two anonymous reviewers for comments on an earlier version of the article. C.W. acknowledges funding support from a Howard Hughes Medical Institute International Student Research Fellowship. This work was supported in part by NIH grants R01 GM081441 and R01 HG005855, and by the Burroughs Wellcome Fund.

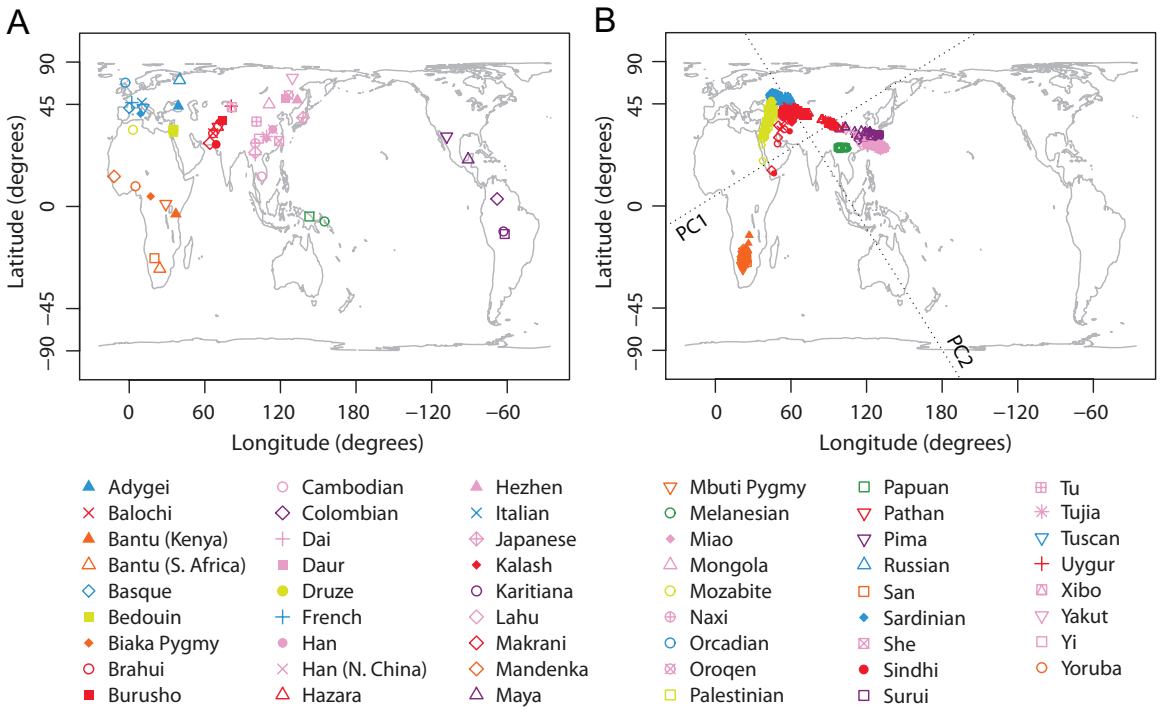

Figure 4.1: Procrustes analysis of genetic and geographic coordinates of worldwide populations. (A) Geographic coordinates of 53 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the Gall-Peters projected coordinates of geographic locations and PC1-PC2 coordinates of 938 individuals. The figures are plotted according to the Gall-Peters projection. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. PC1 and PC2 account for 6.22% and 4.72% of the total variance, respectively. The Procrustes similarity is $t_0 = 0.705$ ($P < 10^{-5}$). The rotation angle of the PCA map is $\theta = 31.91°$.

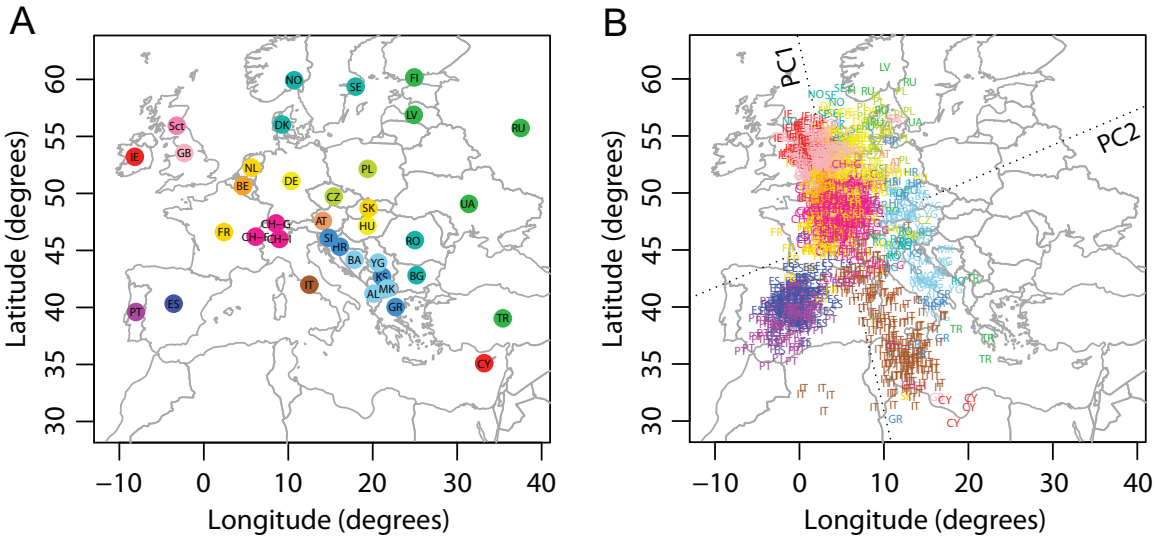

Figure 4.2: Procrustes analysis of genetic and geographic coordinates of European populations. (A) Geographic coordinates of 37 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 1378 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. Abbreviations are as follows: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH-F, Swiss-French; CH-G, Swiss-German; CH-I, Swiss-Italian; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NL, Netherlands; NO, Norway; PL, Poland; PT, Portugal; RO, Romania; RU, Russia; Sct, Scotland; SE, Sweden; SI, Slovenia; TR, Turkey; UA, Ukraine; YG, Serbia and Montenegro. Population labels follow the color scheme of Novembre *et al.* *Novembre et al.* (2008). PC1 and PC2 account for 0.30% and 0.16% of the total variance, respectively. The Procrustes similarity is $t_0 = 0.780$ ($P < 10^{-5}$). The rotation angle of the PCA map is $\theta = -72.66°$.

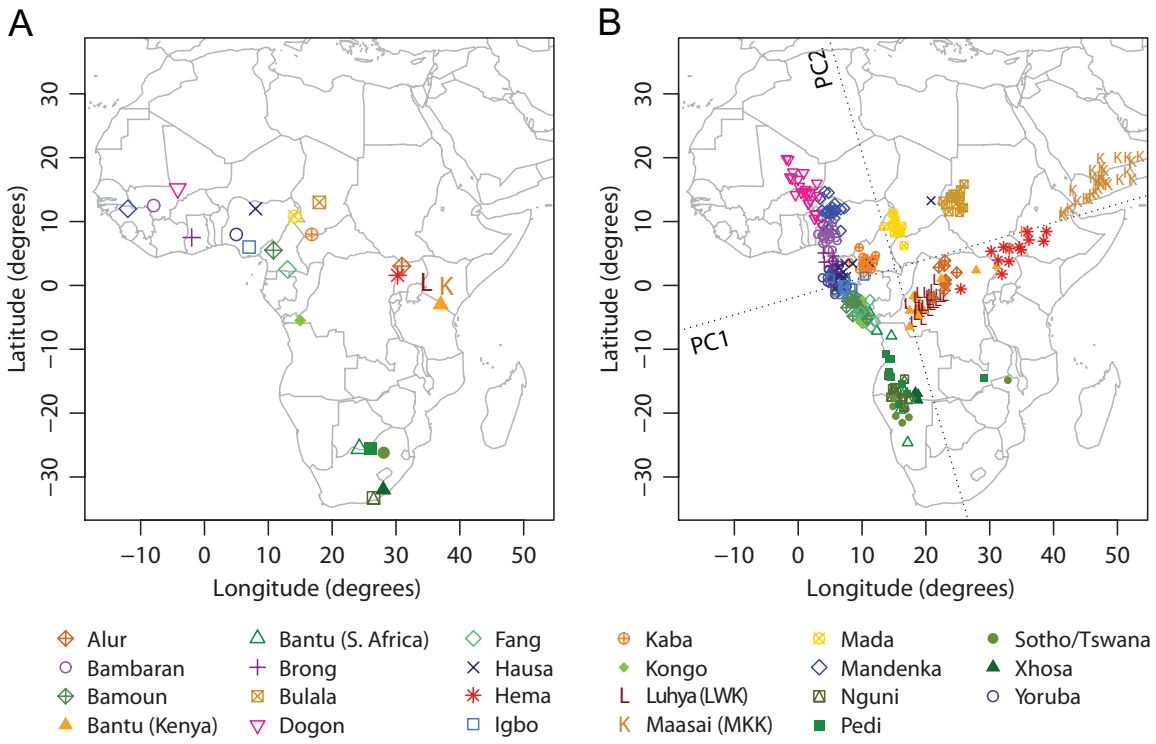

Figure 4.3: Procrustes analysis of genetic and geographic coordinates of Sub-Saharan African populations, excluding hunter-gatherer populations and Mbororo Fulani. (A) Geographic coordinates of 23 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 348 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. PC1 and PC2 account for 1.34% and 0.69% of the total variance, respectively. The Procrustes similarity is $t_0 = 0.790$ ($P < 10^{-5}$). The rotation angle of the PCA map is $\theta = 16.11°$.

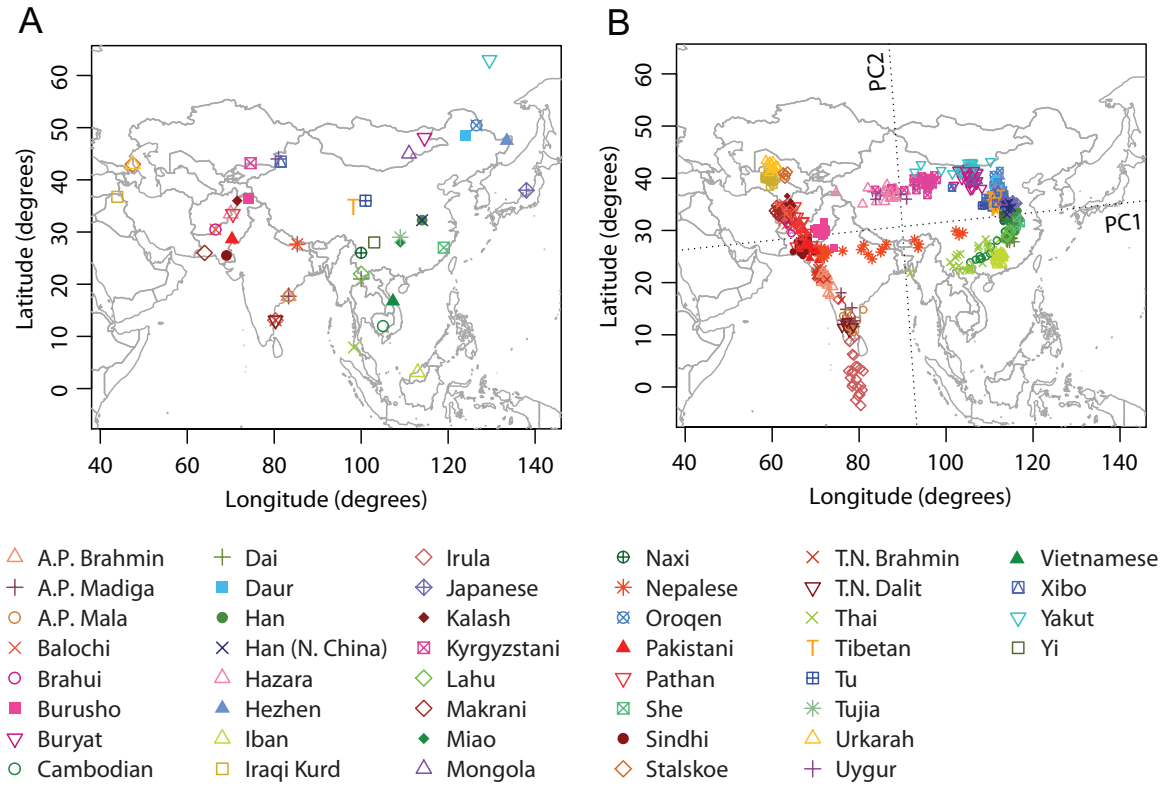

Figure 4.4: Procrustes analysis of genetic and geographic coordinates of Asian populations. (A) Geographic coordinates of 44 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 749 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. PC1 and PC2 account for 5.42% and 0.85% of the total variance, respectively. The Procrustes similarity is $t_0 = 0.849$ ($P < 10^{-5}$). The rotation angle of the PCA map is $\theta = 5.05°$.

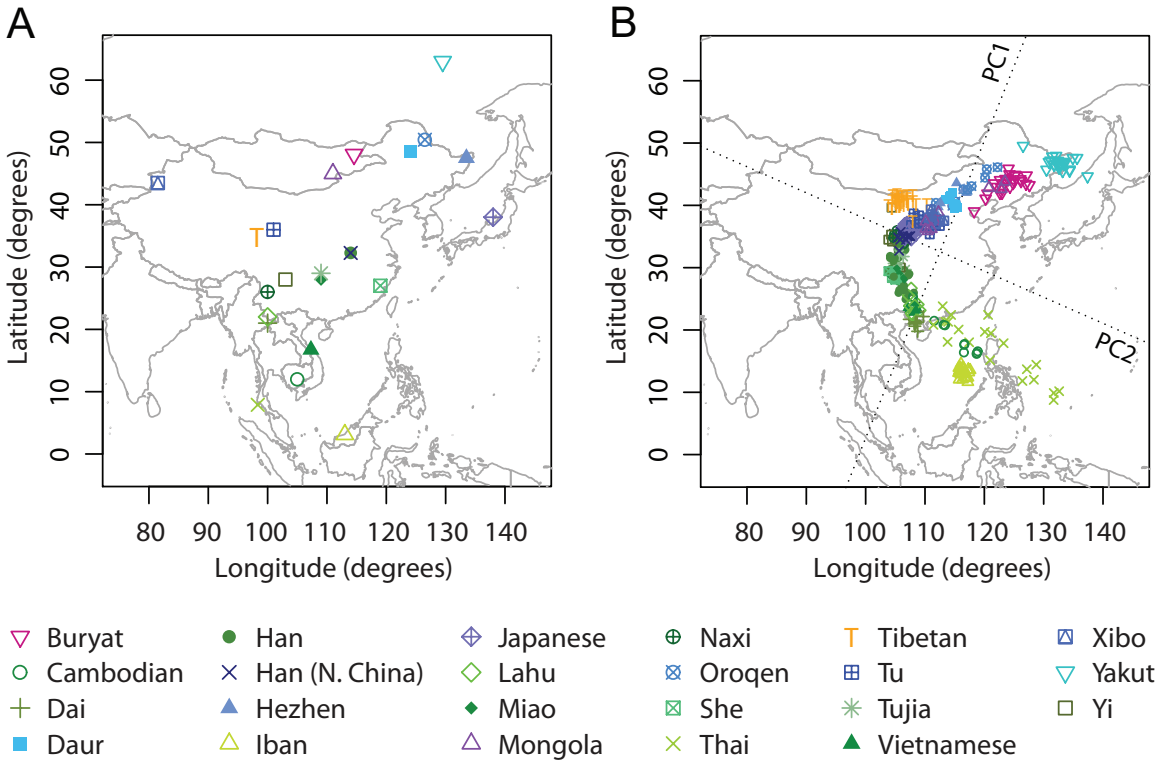

Figure 4.5: Procrustes analysis of genetic and geographic coordinates of East Asian populations. (A) Geographic coordinates of 23 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 334 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. PC1 and PC2 account for 1.58% and 0.98% of the total variance, respectively. The Procrustes similarity statistic is $t_0 = 0.640$ ($P = 0.00038$). The rotation angle of the PCA map is $\theta = 67.27°$.

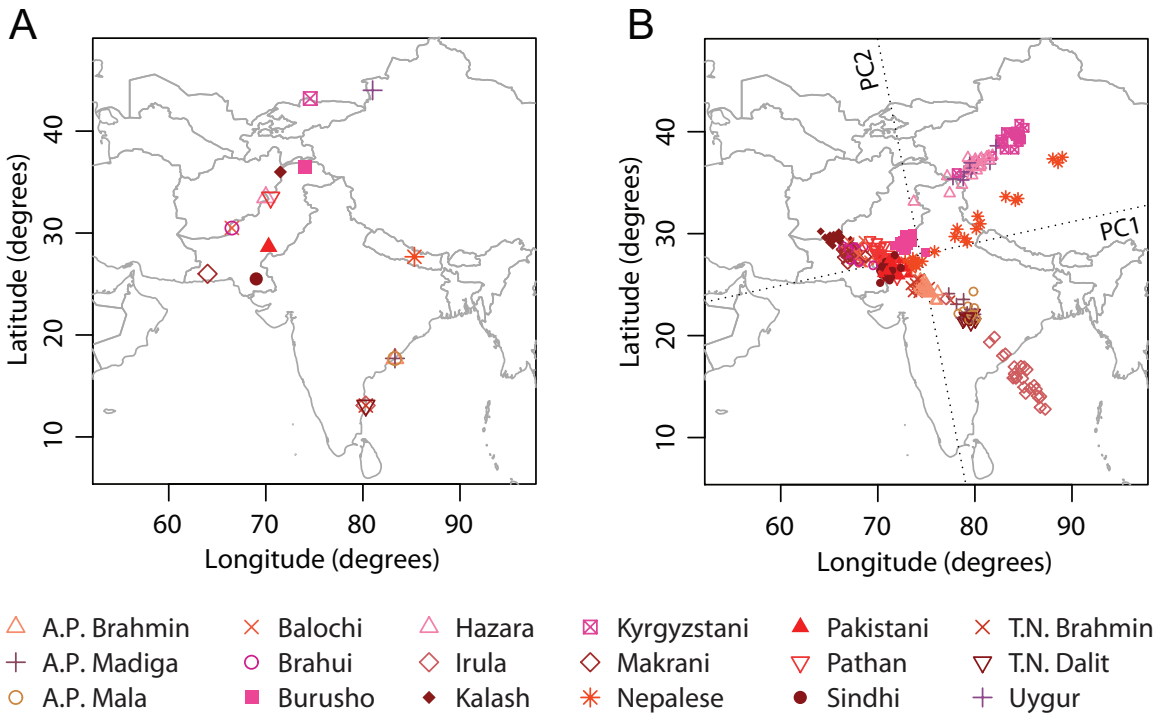

Figure 4.6: Procrustes analysis of genetic and geographic coordinates of Central/South Asian populations. (A) Geographic coordinates of 18 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 362 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. PC1 and PC2 account for 1.59% and 1.31% of the total variance, respectively. The Procrustes similarity statistic is $t_0 = 0.737$ ($P < 10^{-5}$). The rotation angle of the PCA map is $\theta = 11.78°$.

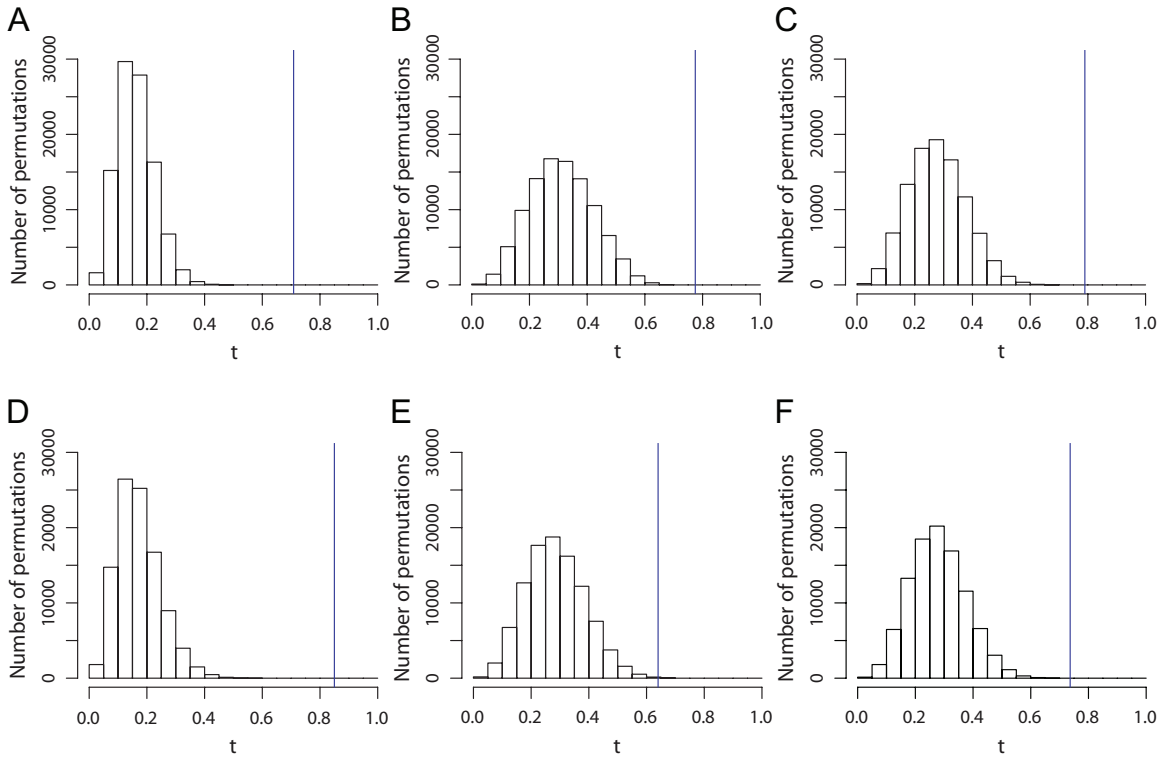

Figure 4.7: Histograms of the Procrustes similarity t of 100,000 permutations for analyses in Figs. 4.1-4.6. The blue vertical lines indicate the value of t_0. (A) The worldwide dataset in Fig. 4.1 ($t_0 = 0.705$, $P < 10^{-5}$). (B) The European dataset in Fig. 4.2 ($t_0 = 0.780$, $P < 10^{-5}$). (C) The Sub-Saharan African dataset in Fig. 4.3 ($t_0 = 0.790$, $P < 10^{-5}$). (D) The Asian dataset in Fig. 4.4 ($t_0 = 0.849$, $P < 10^{-5}$). (E) The East Asian dataset in Fig. 4.5 ($t_0 = 0.640$, $P = 0.00038$). (F) The Central/South dataset in Fig. 4.6 ($t_0 = 0.737$, $P < 10^{-5}$).

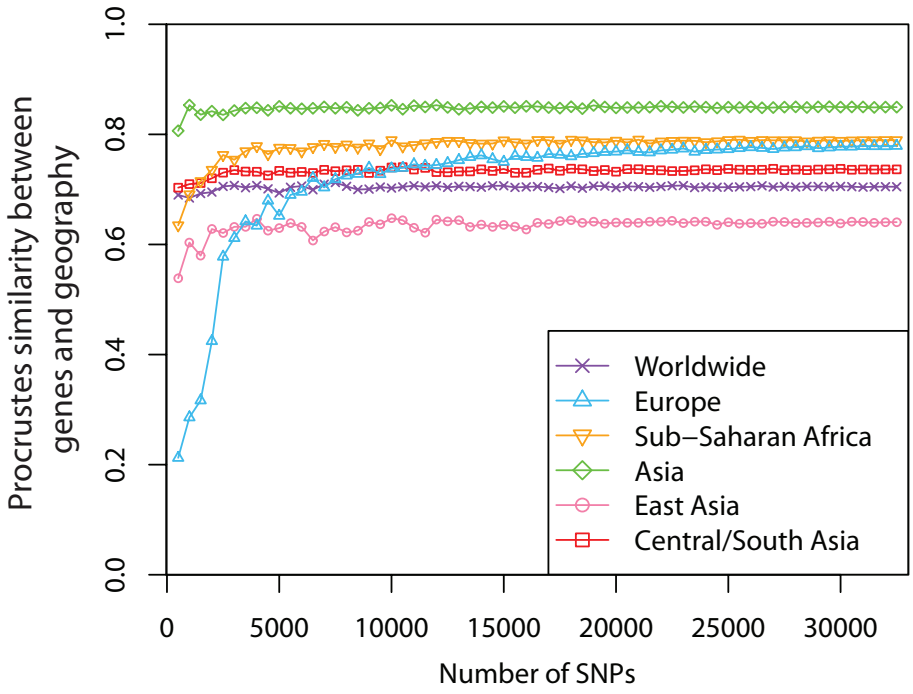

Figure 4.8: Procrustes analyses of genetic and geographic coordinates based on different numbers of loci. The same sets of L randomly selected markers were used to generate PCA maps of genetic variation to compare with geographic maps for different regions. $L = 500, 1000, ..., 32500$.

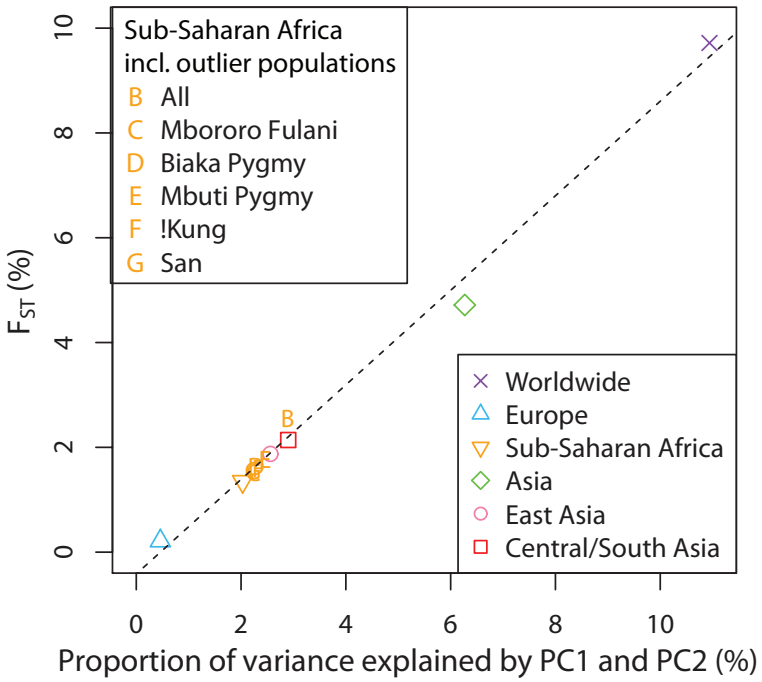

Figure 4.9: Relationship between F_{ST} and the proportion of genetic variation explained by the first two components of the PCA. Both the main analyses of the paper in Table 4.2 and the supplementary analyses of Sub-Saharan Africa, in which certain populations excluded from the main analysis are included, are considered in obtaining the regression line. The values on the x-axis were obtained by summing the proportions of variance explained by PC1 and PC2 (columns 2 and 3 in Table 4.2, columns 6 and 7 in Table S4.7). F_{ST} values were estimated from the same datasets as used in the PCA (column 7 in Table 4.2, column 11 in Table S4.7). The dashed line indicates the linear least squares fit of $y = 0.902x - 0.416$. The Pearson correlation is $r = 0.996$.

Table 4.1: SNP datasets for different geographic regions.

Region	Number of populations	Number of individuals collected	Number of high-missing-data individuals	Number of PCA-outlier individuals	Number of individuals in our analysis	Genotyping platforms	Data sources
Worldwide	53	938	0	0	938	Illumina 650K	*Pemberton et al.*, 2012
Europe	37	1,385	5	2	1,378	Affymetrix 500K	*Novembre et al.*, 2008
Sub-Saharan Africa	23	356	6	2	348	Illumina 650K Illumina Human 1M Affymetrix NspI 250K Affymetrix 500K Affymetrix 6.0	*Xing et al.*, 2010 *Bryc et al.*, 2010b *Pemberton et al.*, 2012
Asia	44	760	0	11	749	Illumina 650K Affymetrix NspI 250K Affymetrix 6.0	*Xing et al.*, 2010 *Pemberton et al.*, 2012 *Simonson et al.*, 2010
East Asia	23	341	0	7	334	Illumina 650K Affymetrix NspI 250K Affymetrix 6.0	*Xing et al.*, 2010 *Pemberton et al.*, 2012 *Simonson et al.*, 2010
Central/South Asia	18	372	0	10	362	Illumina 650K Affymetrix NspI 250K Affymetrix 6.0	*Xing et al.*, 2010 *Pemberton et al.* (2012)

Table 4.2: Summary of the results for datasets from different geographic regions. θ is the rotation angle for the PCA map that optimizes the Procrustes similarity with the geographic map, and it is measured in degrees counterclockwise. P-values are obtained from 100,000 permutations of population labels.

Region	Variance explained by PC1 (%)	Variance explained by PC2 (%)	Geographic map projection	Rotation angle θ (°)	Procrustes similarity t_0	P-value of t_0	F_{ST} (%)
Worldwide	6.22	4.72	Gall-Peters	31.91	0.705	$< 10^{-5}$	9.704
Europe	0.30	0.16	Unprojected	-72.66	0.780	$< 10^{-5}$	0.212
Sub-Saharan Africa	1.34	0.69	Unprojected	16.11	0.790	$< 10^{-5}$	1.334
Asia	5.42	0.85	Unprojected	5.05	0.849	$< 10^{-5}$	4.706
East Asia	1.58	0.98	Unprojected	67.27	0.640	0.00038	1.874
Central/South Asia	1.59	1.31	Unprojected	11.78	0.737	$< 10^{-5}$	2.140

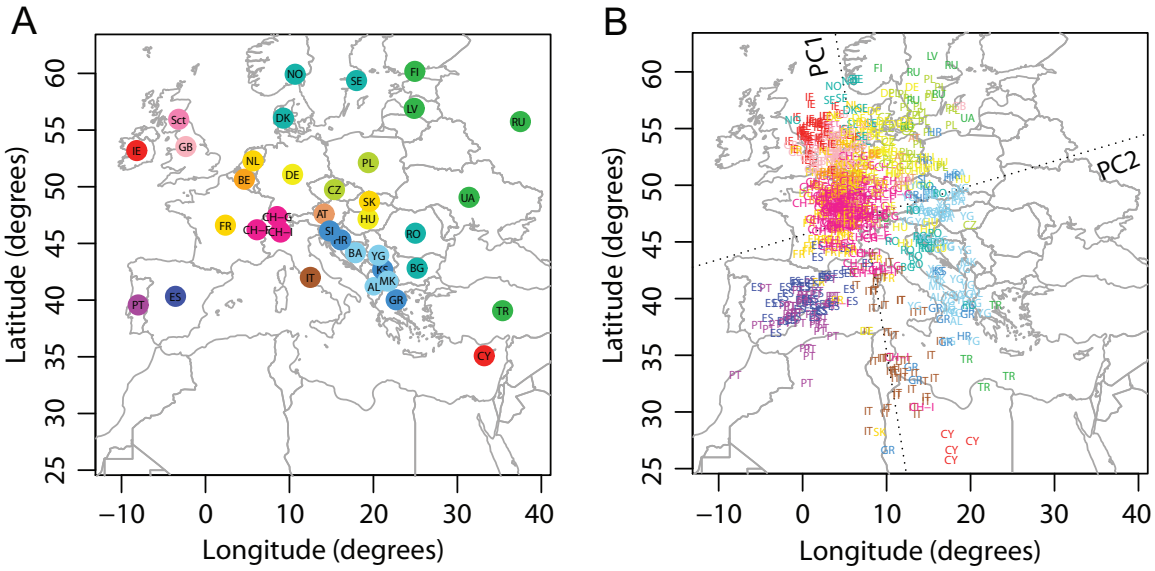

Figure S4.1: Procrustes analysis of genetic and geographic coordinates of European populations, when reducing the maximal sample size to 50. That is, for each population that has sample size $N > 50$ in Fig. 4.2, we reduce the sample size to 50 by randomly excluding $N - 50$ individuals. (A) Geographic coordinates of 37 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 721 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. Population abbreviations can be found in the caption of Fig. 4.2. PC1 and PC2 account for 0.35% and 0.25% of the total variance, respectively. The Procrustes similarity is $t_0 = 0.777$ ($P < 10^{-5}$). The rotation angle of the PCA map is $\theta = -77.75°$. $F_{ST} = 0.230\%$.

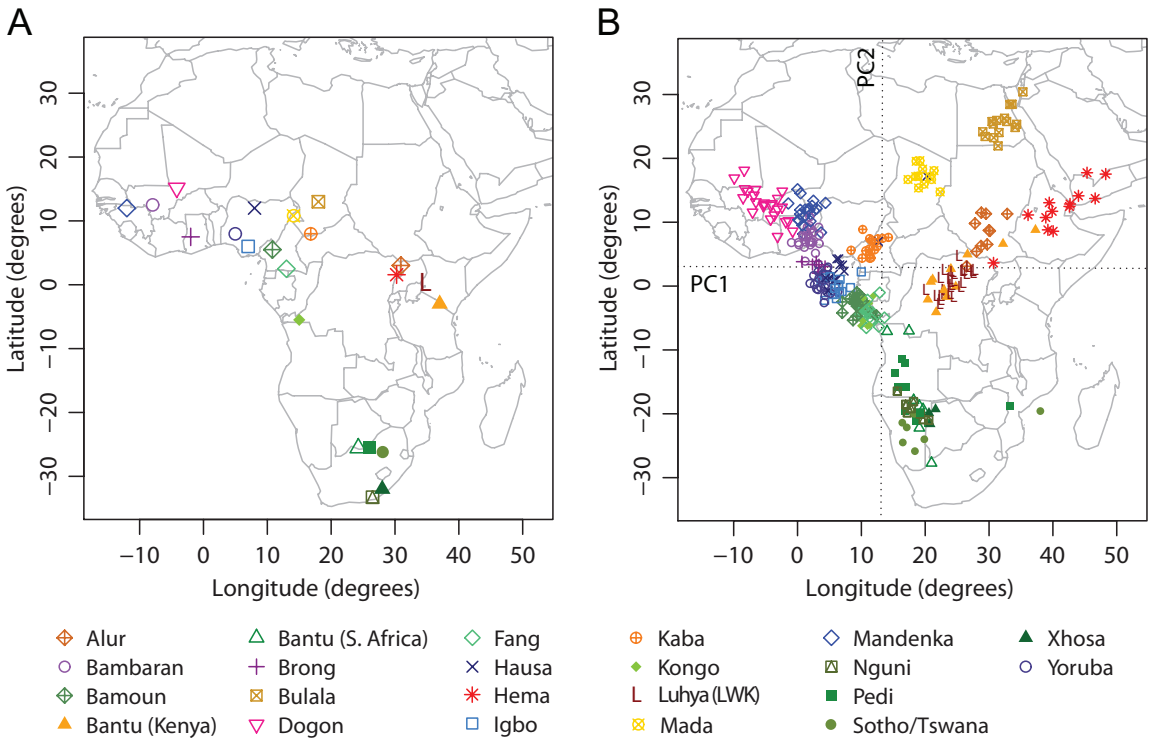

Figure S4.2: Procrustes analysis of genetic and geographic coordinates of Sub-Saharan African populations, excluding Maasai (MKK) as well as Mbororo Fulani and four hunter-gatherer populations. (A) Geographic coordinates of 22 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 318 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. PC1 and PC2 account for 0.89% and 0.75% of the total variance, respectively. The Procrustes similarity statistic is $t_0 = 0.832$ ($P < 10^{-5}$). The rotation angle of the PCA map is $\theta = -0.24°$.

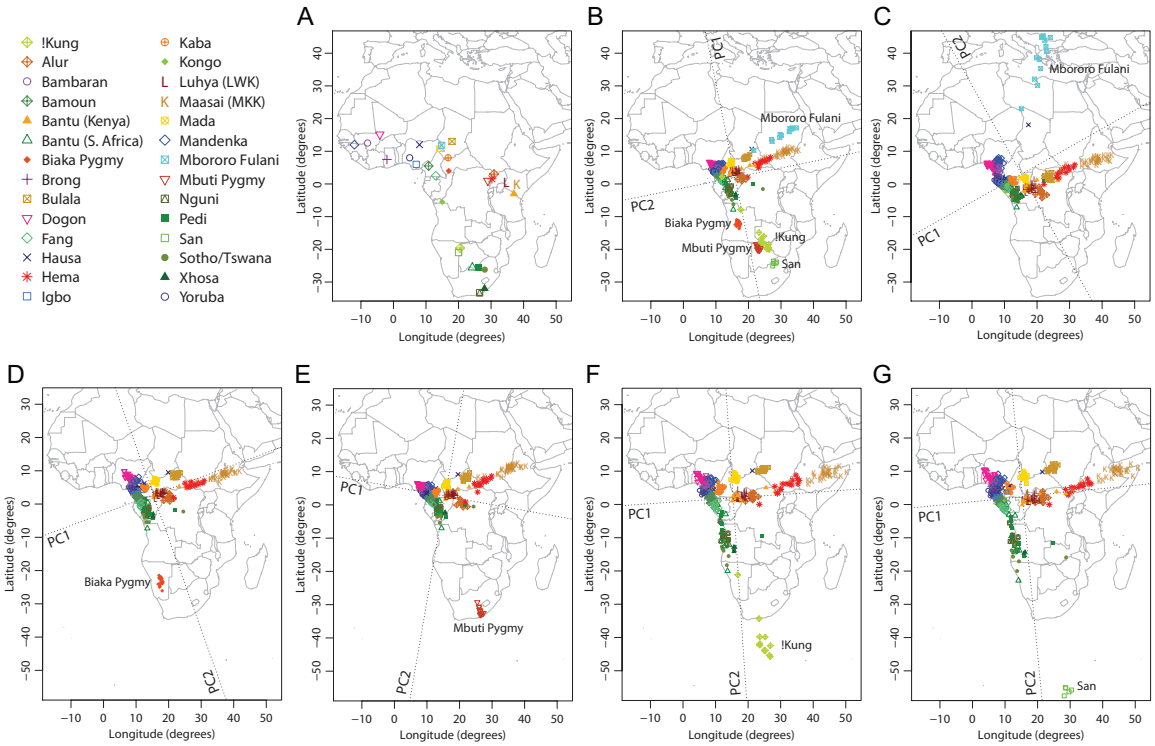

Figure S4.3: Procrustes analysis of genetic and geographic coordinates of Sub-Saharan African populations, including 23 populations in Fig. 4.3 plus Mbororo Fulani and four hunter-gatherer populations (Biaka Pygmy, Mbuti Pygmy, !Kung, and San). (A) Geographic coordinates of all 28 populations. (B-G) Procrustes-transformed PCA plots of genetic variation. (B) All 28 populations. (C) 23 populations and Mbororo Fulani. (D) 23 populations and Biaka Pygmy. (E) 23 populations and Mbuti Pygmy. (F) 23 populations and !Kung. (G) 23 populations and San. Results are summarized in Table S4.7.

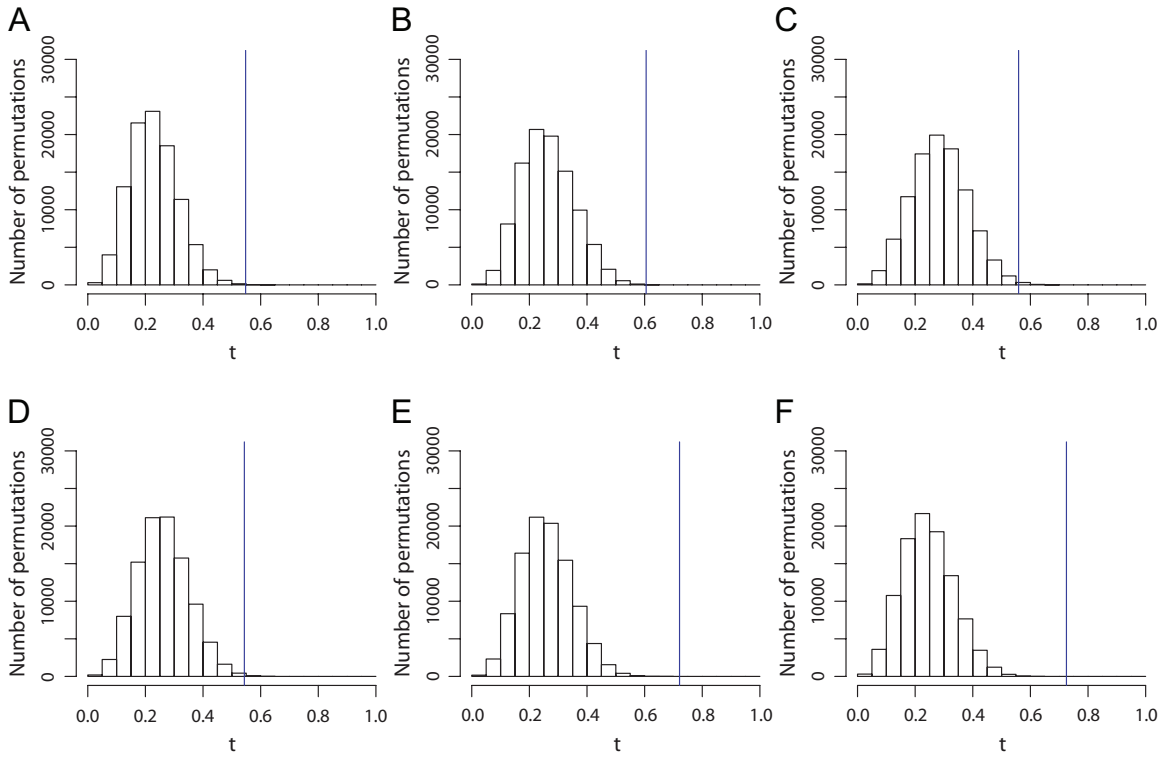

Figure S4.4: Histograms of the Procrustes similarity t of 100,000 permutations for the Sub-Saharan African examples in Fig. S4.3. The blue vertical lines indicate the value of t_0. (A) All 28 populations (corresponding to Fig. S4.3B, $t_0 = 0.548$, $P = 0.00040$). (B) 23 populations and Mbororo Fulani (Fig. S4.3C, $t_0 = 0.605$, $P = 0.00005$). (C) 23 populations and Biaka Pygmy (Fig. S4.3D, $t_0 = 0.559$, $P = 0.00278$). (D) 23 populations and Mbuti Pygmy (Fig. S4.3E, $t_0 = 0.543$, $P = 0.00120$). (E) 23 populations and !Kung (Fig. S4.3F, $t_0 = 0.721$, $P < 10^{-5}$). (F) 23 populations and San (Fig. S4.3G, $t_0 = 0.725$, $P < 10^{-5}$).

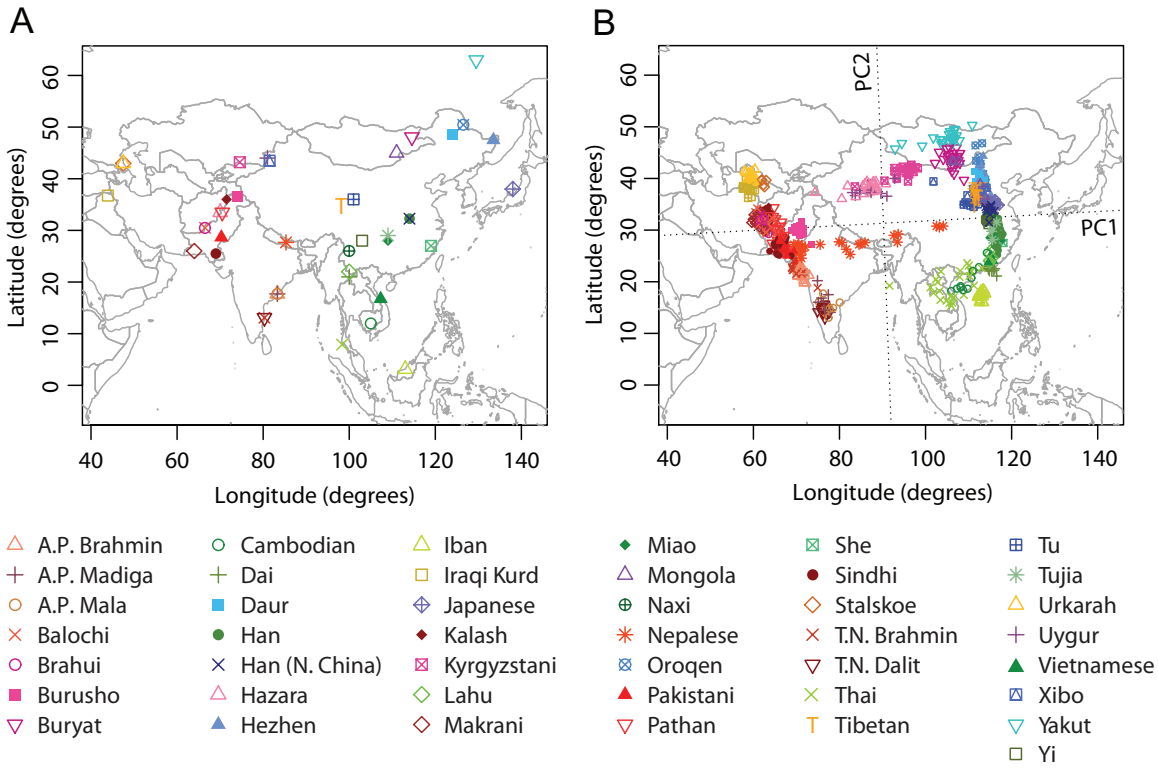

Figure S4.5: Procrustes analysis of genetic and geographic coordinates of Asian populations, excluding Irula. (A) Geographic coordinates of 43 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 725 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. PC1 and PC2 account for 5.55% and 0.74% of the total variance, respectively. The Procrustes similarity statistic is $t_0 = 0.871$ ($P < 10^{-5}$). The rotation angle of the PCA map is $\theta = 2.61°$.

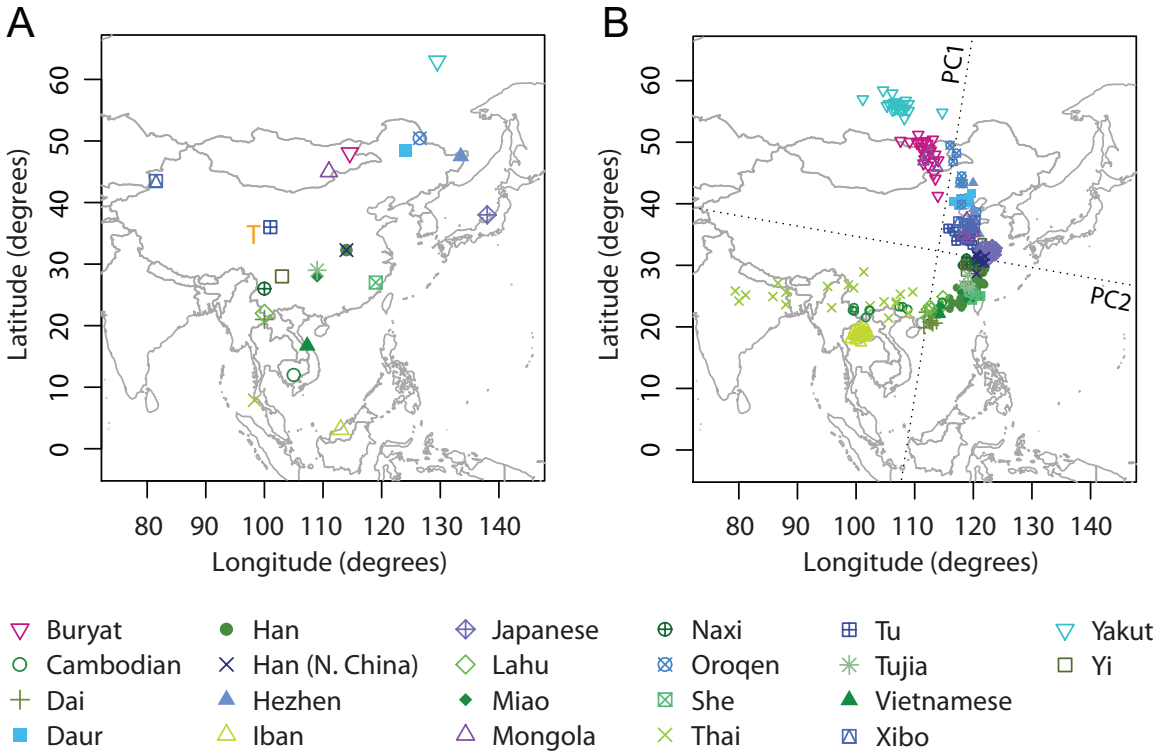

Figure S4.6: Procrustes analysis of genetic and geographic coordinates of East Asian populations, excluding Tibetans. (A) Geographic coordinates of 22 populations. (B) Procrustes-transformed PCA plot of genetic variation. The Procrustes analysis is based on the unprojected latitude-longitude coordinates and PC1-PC2 coordinates of 303 individuals. PC1 and PC2 are indicated by dotted lines, crossing over the centroid of all individuals. PC1 and PC2 account for 1.72% and 1.02% of the total variance, respectively. The Procrustes similarity statistic is $t_0 = 0.655$ ($P = 0.00025$). The rotation angle of the PCA map is $\theta = 80.44°$.

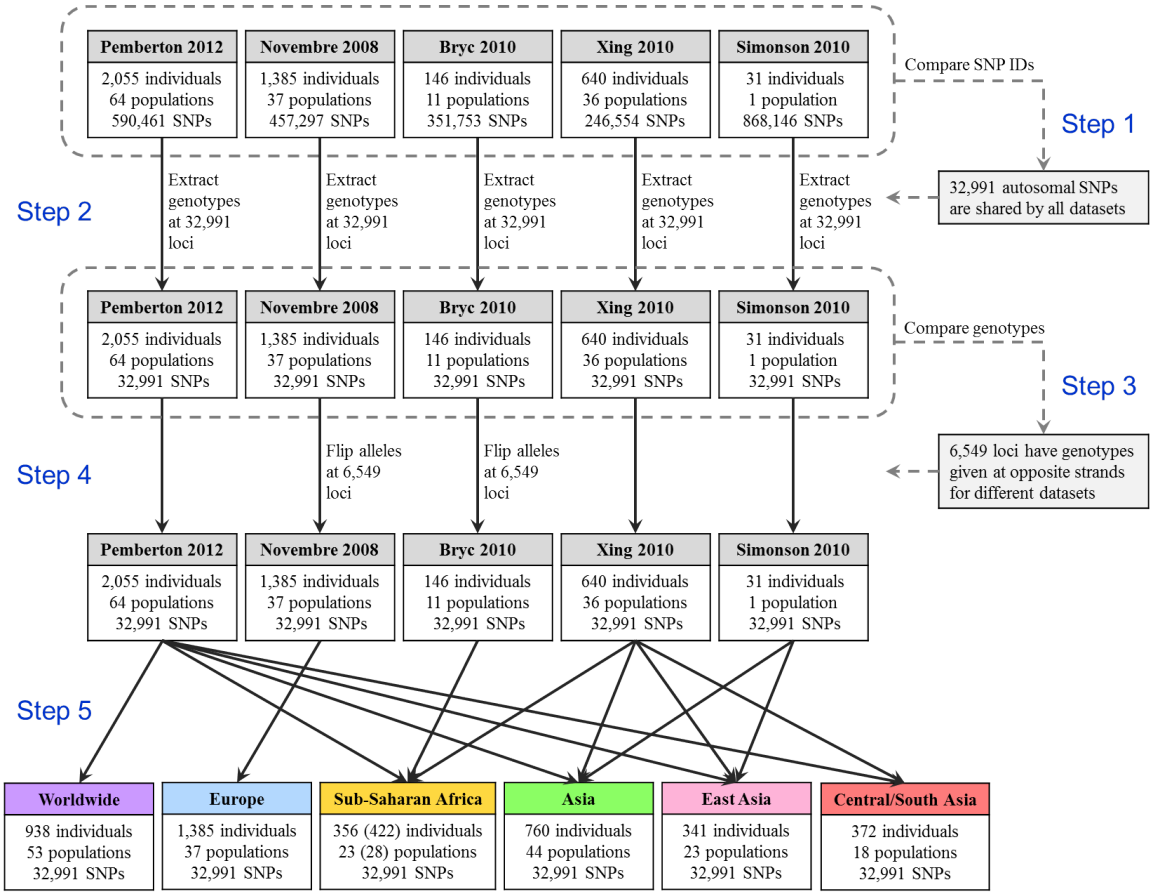

Figure S4.7: Data preparation procedure for creating datasets for different geographic regions.

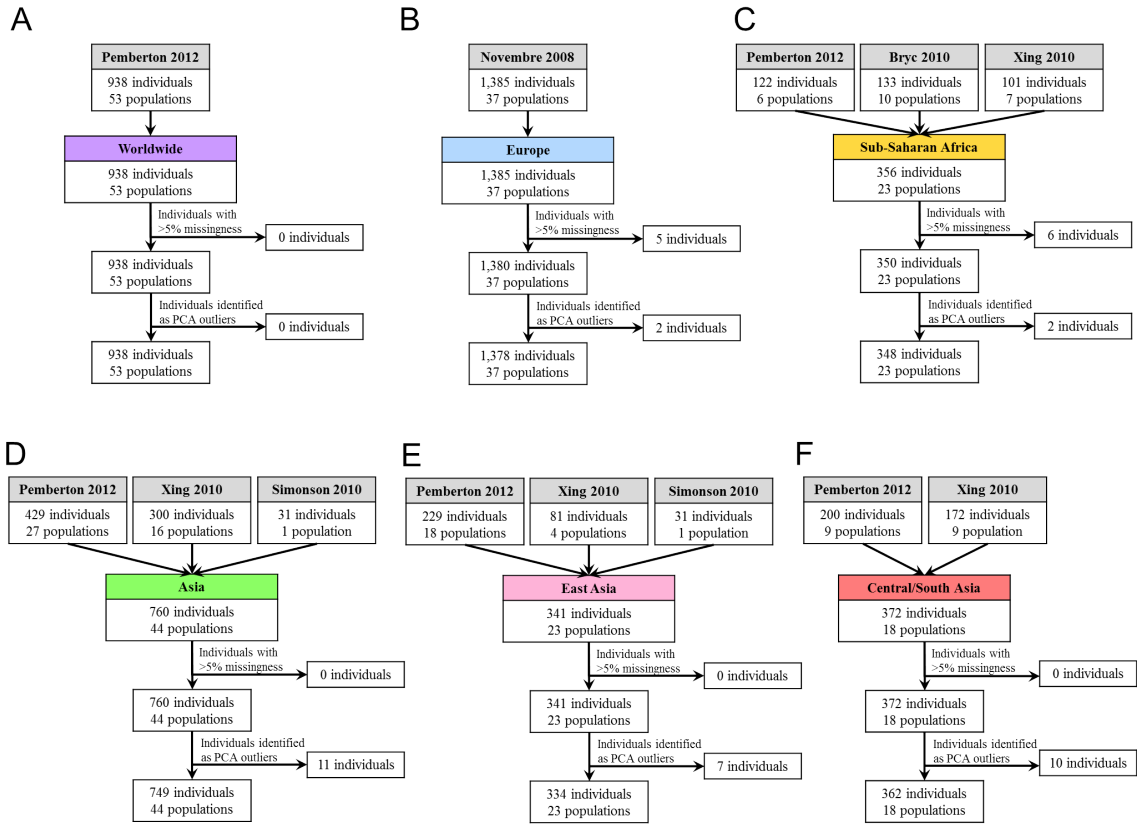

Figure S4.8: Data-processing procedures for datasets from different geographic regions. (A) The worldwide dataset in Fig. 4.1. (B) The European dataset in Fig. 4.2. (C) The Sub-Saharan African dataset in Fig. 4.3 (excluding Mbororo Fulani and four hunter-gatherer populations). (D) The Asian dataset in Fig. 4.4. (E) The East Asian dataset in Fig. 4.5. (F) The Central/South Asian dataset in Fig. 4.6.

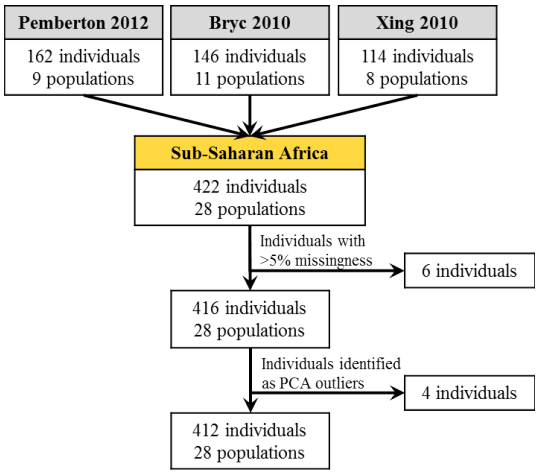

Figure S4.9: Data-processing procedure for the supplementary example of Sub-Saharan Africa when including Mbororo Fulani and four hunter-gatherer populations (Biaka Pygmy, Mbuti Pygmy, !Kung, and San). Similar procedures (not shown) were also used to prepare datasets for the analyses in Figs. S4.3C-G, in each of which only one outlier population was included.

Table S4.1: Populations included in this study (Part I).

Population	Latitude (degrees)	Longitude (degrees)	Source of coordinates	Sample size	High-missing-data samples	Genotyping platform	Source of SNP data
Adygei	44	39	*Rosenberg* (2011)	17	0	Illumina 650K	*Pemberton et al.* (2012)
Balochi	30.5	66.5	*Rosenberg* (2011)	24	0	Illumina 650K	*Pemberton et al.* (2012)
Bantu (Kenya)	-3	37	*Rosenberg* (2011)	11	0	Illumina 650K	*Pemberton et al.* (2012)
Bantu (S. Africa)	-25.6	24.3	*Rosenberg* (2011)	8	0	Illumina 650K	*Pemberton et al.* (2012)
Basque	43	0	*Rosenberg* (2011)	24	0	Illumina 650K	*Pemberton et al.* (2012)
Bedouin	31	35	*Rosenberg* (2011)	45	0	Illumina 650K	*Pemberton et al.* (2012)
Biaka Pygmy	4	17	*Rosenberg* (2011)	22	0	Illumina 650K	*Pemberton et al.* (2012)
Brahui	30.5	66.5	*Rosenberg* (2011)	25	0	Illumina 650K	*Pemberton et al.* (2012)
Burusho	36.5	74	*Rosenberg* (2011)	25	0	Illumina 650K	*Pemberton et al.* (2012)
Cambodian	12	105	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Colombian	3	-68	*Rosenberg* (2011)	7	0	Illumina 650K	*Pemberton et al.* (2012)
Dai	21	100	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Daur	48.5	124	*Rosenberg* (2011)	9	0	Illumina 650K	*Pemberton et al.* (2012)
Druze	32	35	*Rosenberg* (2011)	42	0	Illumina 650K	*Pemberton et al.* (2012)
French	46	2	*Rosenberg* (2011)	28	0	Illumina 650K	*Pemberton et al.* (2012)
Han	32.3	114	*Rosenberg* (2011)	34	0	Illumina 650K	*Pemberton et al.* (2012)
Han (N. China)	32.3	114	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Hazara	33.5	70	*Rosenberg* (2011)	22	0	Illumina 650K	*Pemberton et al.* (2012)
Hezhen	47.5	133.5	*Rosenberg* (2011)	9	0	Illumina 650K	*Pemberton et al.* (2012)
Italian	46	10	*Rosenberg* (2011)	12	0	Illumina 650K	*Pemberton et al.* (2012)
Japanese	38	138	*Rosenberg* (2011)	28	0	Illumina 650K	*Pemberton et al.* (2012)
Kalash	36.0	71.5	*Rosenberg* (2011)	23	0	Illumina 650K	*Pemberton et al.* (2012)
Karitiana	-10	-63	*Rosenberg* (2011)	13	0	Illumina 650K	*Pemberton et al.* (2012)
Lahu	22	100	*Rosenberg* (2011)	8	0	Illumina 650K	*Pemberton et al.* (2012)
Makrani	26	64	*Rosenberg* (2011)	25	0	Illumina 650K	*Pemberton et al.* (2012)
Mandenka	12	-12	*Rosenberg* (2011)	22	0	Illumina 650K	*Pemberton et al.* (2012)
Maya	19	-91	*Rosenberg* (2011)	21	0	Illumina 650K	*Pemberton et al.* (2012)
Mbuti Pygmy	1	29	*Rosenberg* (2011)	13	0	Illumina 650K	*Pemberton et al.* (2012)
Melanesian	-6	155	*Rosenberg* (2011)	11	0	Illumina 650K	*Pemberton et al.* (2012)
Miao	28	109	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Mongola	45	111	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Mozabite	32	3	*Rosenberg* (2011)	27	0	Illumina 650K	*Pemberton et al.* (2012)
Naxi	26	100	*Rosenberg* (2011)	8	0	Illumina 650K	*Pemberton et al.* (2012)
Ocadian	59	-3	*Rosenberg* (2011)	15	0	Illumina 650K	*Pemberton et al.* (2012)
Oroqen	50.4	126.5	*Rosenberg* (2011)	9	0	Illumina 650K	*Pemberton et al.* (2012)
Palestinian	32	35	*Rosenberg* (2011)	46	0	Illumina 650K	*Pemberton et al.* (2012)
Papuan	-4	143	*Rosenberg* (2011)	17	0	Illumina 650K	*Pemberton et al.* (2012)
Pathan	33.5	70.5	*Rosenberg* (2011)	22	0	Illumina 650K	*Pemberton et al.* (2012)
Pima	29	-108	*Rosenberg* (2011)	14	0	Illumina 650K	*Pemberton et al.* (2012)
Russian	61	40	*Rosenberg* (2011)	25	0	Illumina 650K	*Pemberton et al.* (2012)
San	-21	20	*Rosenberg* (2011)	5	0	Illumina 650K	*Pemberton et al.* (2012)
Sardinian	40	9	*Rosenberg* (2011)	28	0	Illumina 650K	*Pemberton et al.* (2012)
She	27	119	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Sindhi	25.5	69	*Rosenberg* (2011)	24	0	Illumina 650K	*Pemberton et al.* (2012)
Surui	-11	-62	*Rosenberg* (2011)	8	0	Illumina 650K	*Pemberton et al.* (2012)

Table S4.2: Populations included in this study (Part II).

Population	Latitude (degrees)	Longitude (degrees)	Source of coordinates	Sample size	High-missing-data samples	Genotyping platform	Source of SNP data
Tu	36	101	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Tujia	29	109	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Tuscan	43	11	*Rosenberg* (2011)	7	0	Illumina 650K	*Pemberton et al.* (2012)
Uygur	44	81	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Xibo	43.5	81.5	*Rosenberg* (2011)	9	0	Illumina 650K	*Pemberton et al.* (2012)
Yakut	63.0	129.5	*Rosenberg* (2011)	25	0	Illumina 650K	*Pemberton et al.* (2012)
Yi	28	103	*Rosenberg* (2011)	10	0	Illumina 650K	*Pemberton et al.* (2012)
Yoruba	8	5	*Rosenberg* (2011)	21	0	Illumina 650K	*Pemberton et al.* (2012)
Luhya (LWK)	0.6	34.8	HapMap3 (2010)	30	0	HapMap3 rel2	*Pemberton et al.* (2012)
Maasai (MKK)	0	37.9	HapMap3 (2010)	30	0	HapMap3 rel2	*Pemberton et al.* (2012)
Albania (AL)	41.2	20.1	*Novembre et al.* (2008)	3	0	Affymetrix 500K	*Novembre et al.* (2008)
Austria (AT)	47.6	14.1	*Novembre et al.* (2008)	14	0	Affymetrix 500K	*Novembre et al.* (2008)
Bosnia-Herzegovina (BA)	44.2	17.9	*Novembre et al.* (2008)	9	0	Affymetrix 500K	*Novembre et al.* (2008)
Belgium (BE)	50.7	4.61	*Novembre et al.* (2008)	43	1	Affymetrix 500K	*Novembre et al.* (2008)
Bulgaria (BG)	42.8	25.2	*Novembre et al.* (2008)	2	0	Affymetrix 500K	*Novembre et al.* (2008)
Swiss-French (CH-F)	46.2	6.15	*Novembre et al.* (2008)	125	0	Affymetrix 500K	*Novembre et al.* (2008)
Swiss-German (CH-G)	47.4	8.55	*Novembre et al.* (2008)	84	2	Affymetrix 500K	*Novembre et al.* (2008)
Swiss-Italian (CH-I)	46	8.95	*Novembre et al.* (2008)	13	0	Affymetrix 500K	*Novembre et al.* (2008)
Cyprus (CY)	35.1	33.2	*Novembre et al.* (2008)	4	0	Affymetrix 500K	*Novembre et al.* (2008)
Czech Republic (CZ)	49.7	15.4	*Novembre et al.* (2008)	11	0	Affymetrix 500K	*Novembre et al.* (2008)
Germany (DE)	51.1	10.4	*Novembre et al.* (2008)	71	2	Affymetrix 500K	*Novembre et al.* (2008)
Denmark (DK)	56.1	9.25	*Novembre et al.* (2008)	1	0	Affymetrix 500K	*Novembre et al.* (2008)
Spain (ES)	40.3	-3.57	*Novembre et al.* (2008)	136	0	Affymetrix 500K	*Novembre et al.* (2008)
Finland (FI)	60.2	24.9	*Novembre et al.* (2008)	1	0	Affymetrix 500K	*Novembre et al.* (2008)
France (FR)	46.6	2.39	*Novembre et al.* (2008)	89	0	Affymetrix 500K	*Novembre et al.* (2008)
United Kingdom (GB)	53.5	-2.33	*Novembre et al.* (2008)	200	1	Affymetrix 500K	*Novembre et al.* (2008)
Greece (GR)	40	22.7	*Novembre et al.* (2008)	8	0	Affymetrix 500K	*Novembre et al.* (2008)
Croatia (HR)	45.3	16.1	*Novembre et al.* (2008)	8	0	Affymetrix 500K	*Novembre et al.* (2008)
Hungary (HU)	47.2	19.4	*Novembre et al.* (2008)	19	0	Affymetrix 500K	*Novembre et al.* (2008)
Ireland (IE)	53.2	-8.18	*Novembre et al.* (2008)	61	1	Affymetrix 500K	*Novembre et al.* (2008)
Italy (IT)	42	12.5	*Novembre et al.* (2008)	219	0	Affymetrix 500K	*Novembre et al.* (2008)
Kosovo (KS)	42.7	21.1	*Novembre et al.* (2008)	2	0	Affymetrix 500K	*Novembre et al.* (2008)
Latvia (LV)	56.9	24.9	*Novembre et al.* (2008)	1	0	Affymetrix 500K	*Novembre et al.* (2008)
Macedonia (MK)	41.7	21.7	*Novembre et al.* (2008)	4	0	Affymetrix 500K	*Novembre et al.* (2008)
Netherlands (NL)	52.3	5.67	*Novembre et al.* (2008)	17	0	Affymetrix 500K	*Novembre et al.* (2008)
Norway (NO)	59.9	10.7	*Novembre et al.* (2008)	3	0	Affymetrix 500K	*Novembre et al.* (2008)
Poland (PL)	52.1	19.4	*Novembre et al.* (2008)	22	0	Affymetrix 500K	*Novembre et al.* (2008)
Portugal (PT)	39.6	-8.02	*Novembre et al.* (2008)	128	0	Affymetrix 500K	*Novembre et al.* (2008)
Romania (RO)	45.9	25	*Novembre et al.* (2008)	14	0	Affymetrix 500K	*Novembre et al.* (2008)
Russia (RU)	55.8	37.5	*Novembre et al.* (2008)	6	0	Affymetrix 500K	*Novembre et al.* (2008)
Scotland (Sct)	56	-3.2	*Novembre et al.* (2008)	5	0	Affymetrix 500K	*Novembre et al.* (2008)
Sweden (SE)	59.4	18	*Novembre et al.* (2008)	10	0	Affymetrix 500K	*Novembre et al.* (2008)
Slovenia (SI)	46.1	14.8	*Novembre et al.* (2008)	2	0	Affymetrix 500K	*Novembre et al.* (2008)
Slovakia (SK)	48.7	19.5	*Novembre et al.* (2008)	1	0	Affymetrix 500K	*Novembre et al.* (2008)
Turkey (TR)	39.1	35.4	*Novembre et al.* (2008)	4	0	Affymetrix 500K	*Novembre et al.* (2008)

Table S4.3: Populations included in this study (Part III).

Population	Latitude (degrees)	Longitude (degrees)	Source of coordinates	Sample size	High-missing-data samples	Genotyping platform	Source of SNP data
Ukraine (UA)	49.1	31.4	*Novembre et al.* (2008)	1	0	Affymetrix 500K	*Novembre et al.* (2008)
Serbia-Montenegro (YG)	43.9	20.6	*Novembre et al.* (2008)	44	0	Affymetrix 500K	*Novembre et al.* (2008)
Bamoun	5.5	10.8	*Tishkoff et al.* (2009)	20	2	Affymetrix 500K	*Bryc et al.* (2010b)
Brong	7.5	-2.0	*Tishkoff et al.* (2009)	8	1	Affymetrix 500K	*Bryc et al.* (2010b)
Bulala	13.0	18.0	*Tishkoff et al.* (2009)	15	0	Affymetrix 500K	*Bryc et al.* (2010b)
Fang	2.5	13.0	*Tishkoff et al.* (2009)	18	1	Affymetrix 500K	*Bryc et al.* (2010b)
Hausa	12.0	8.0	*Tishkoff et al.* (2009)	13	2	Affymetrix 500K	*Bryc et al.* (2010b)
Igbo	6.0	7.0	*Tishkoff et al.* (2009)	17	4	Affymetrix 500K	*Bryc et al.* (2010b)
Kaba	8.0	16.8	*Tishkoff et al.* (2009)	16	0	Affymetrix 500K	*Bryc et al.* (2010b)
Kongo	-5.5	15.0	*Tishkoff et al.* (2009)	9	0	Affymetrix 500K	*Bryc et al.* (2010b)
Mbororo Fulani	11.8	14.8	*Tishkoff et al.* (2009)	13	2	Affymetrix 500K	*Bryc et al.* (2010b)
Mada	10.8	14.1	*Tishkoff et al.* (2009)	12	0	Affymetrix 500K	*Bryc et al.* (2010b)
Xhosa	-32.0	28.0	*Tishkoff et al.* (2009)	5	2	Affymetrix 500K	*Bryc et al.* (2010b)
!Kung	-19.6	20.5	J. Xing	13	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Alur	-3.0	30.9	J. Xing	10	0	Affymetrix NspI 250K	*Xing et al.* (2010)
A.P. Brahmin	17.7	83.3	J. Xing	25	0	Affymetrix NspI 250K	*Xing et al.* (2010)
A.P. Madiga	17.7	83.3	J. Xing	10	0	Affymetrix NspI 250K	*Xing et al.* (2010)
A.P. Mala	17.7	83.3	J. Xing	11	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Bambaran	12.5	-8.0	J. Xing	25	0	Affymetrix 6.0	*Xing et al.* (2010)
Buryat	48.1	114.6	J. Xing	25	0	Affymetrix 6.0	*Xing et al.* (2010)
Dogon	15.1	-4.2	J. Xing	24	0	Affymetrix 6.0	*Xing et al.* (2010)
Hema	1.6	30.3	J. Xing	15	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Iban	3.1	113.0	J. Xing	25	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Iraqi Kurd	36.7	43.9	J. Xing	24	0	Affymetrix 6.0	*Xing et al.* (2010)
Irula	13.1	80.3	J. Xing	24	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Kyrgyzstani	43.2	74.6	J. Xing	25	0	Affymetrix 6.0	*Xing et al.* (2010)
Nepalese	27.7	85.3	J. Xing	25	0	Affymetrix 6.0	*Xing et al.* (2010)
Nguni	-33.3	26.5	J. Xing	9	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Pakistani	28.6	70.3	J. Xing	25	0	Affymetrix 6.0	*Xing et al.* (2010)
Pedi	-25.5	26.1	J. Xing	10	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Sotho/Tswana	-26.2	28.1	J. Xing	8	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Stalskoe	43.0	47.5	J. Xing	5	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Thai	7.9	98.3	J. Xing	24	0	Affymetrix 6.0	*Xing et al.* (2010)
T.N. Brahmin	13.1	80.3	J. Xing	14	0	Affymetrix NspI 250K	*Xing et al.* (2010)
T.N. Dalit	13.1	80.3	J. Xing	13	0	Affymetrix 6.0	*Xing et al.* (2010)
Urkarah	43.0	47.5	J. Xing	18	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Vietnamese	16.7	107.3	J. Xing	7	0	Affymetrix NspI 250K	*Xing et al.* (2010)
Tibetan	34.9	98.2	*Simonson et al.* (2010)	31	0	Affymetrix 6.0	*Simonson et al.* (2010)

Table S4.4: Change of the Procrustes similarity when excluding one population from the worldwide example. The Procrustes similarity between genetic coordinates and geographic coordinates is $t_0 = 0.705$ in the original analysis (Fig. 4.1).

Population excluded	Number of individuals excluded	Similarity to original PCA t'	Similarity to geography t''	$t'' - t_0$
Han	34	1.000	0.715	0.010
Maya	21	1.000	0.713	0.008
Karitiana	13	1.000	0.710	0.005
Xibo	9	1.000	0.710	0.005
Dai	10	1.000	0.708	0.003
Yi	10	1.000	0.708	0.003
Tujia	10	1.000	0.708	0.003
Miao	10	1.000	0.708	0.003
Tu	10	1.000	0.707	0.002
Naxi	8	1.000	0.707	0.002
Lahu	8	1.000	0.707	0.002
Surui	8	1.000	0.707	0.002
Sindhi	24	1.000	0.707	0.002
Makrani	25	1.000	0.707	0.002
Mongola	10	1.000	0.707	0.002
Yakut	25	1.000	0.707	0.002
Han (N. China)	10	1.000	0.707	0.002
She	10	1.000	0.707	0.002
Hazara	22	1.000	0.707	0.002
Brahui	25	1.000	0.707	0.002
Cambodian	10	1.000	0.707	0.002
Papuan	17	1.000	0.707	0.002
Japanese	28	1.000	0.707	0.002
Balochi	24	1.000	0.707	0.002
Daur	9	1.000	0.706	0.001
Colombian	7	1.000	0.706	0.001
Oroqen	9	1.000	0.706	0.001
Melanesian	11	1.000	0.706	0.001
Pathan	22	1.000	0.706	0.001
Kalash	23	1.000	0.706	0.001
Hezhen	9	1.000	0.706	0.001
Mandenka	22	1.000	0.705	0.000
Uygur	10	1.000	0.705	0.000
Burusho	25	1.000	0.705	0.000
Yoruba	21	1.000	0.704	-0.001
Tuscan	7	1.000	0.704	-0.001
Druze	42	1.000	0.704	-0.001
Adygei	17	1.000	0.704	-0.001
Biaka Pygmy	22	1.000	0.703	-0.002
Italian	12	1.000	0.703	-0.002
Mbuti Pygmy	13	1.000	0.703	-0.002
Orcadian	15	1.000	0.703	-0.002
Basque	24	1.000	0.702	-0.003
Russian	25	1.000	0.702	-0.003
French	28	1.000	0.702	-0.003
Palestinian	46	1.000	0.701	-0.004
Bantu (Kenya)	11	1.000	0.701	-0.004
Bedouin	45	1.000	0.701	-0.004
Sardinian	28	1.000	0.701	-0.004
San	5	1.000	0.700	-0.005
Pima	14	1.000	0.700	-0.005
Mozabite	27	1.000	0.699	-0.006
Bantu (S. Africa)	8	1.000	0.697	-0.008

Table S4.5: Change of the Procrustes similarity when excluding one population from the European example. The Procrustes similarity between genetic coordinates and geographic coordinates is $t_0 = 0.780$ in the original analysis (Fig. 4.2).

Population excluded	Number of individuals excluded	Similarity to original PCA t'	Similarity to geography t''	$t'' - t_0$
Italy (IT)	219	0.986	0.810	0.030
Russia (RU)	6	1.000	0.788	0.008
Swiss-French (CH-F)	125	1.000	0.785	0.005
Swiss-German (CH-G)	84	1.000	0.785	0.005
Germany (DE)	69	1.000	0.783	0.003
France (FR)	89	1.000	0.783	0.003
Sweden (SE)	10	1.000	0.782	0.002
Swiss-Italian (CH-I)	13	1.000	0.781	0.001
Austria (AT)	14	1.000	0.781	0.001
Slovakia (SK)	1	1.000	0.780	0.000
Hungary (HU)	19	1.000	0.780	0.000
Romania (RO)	14	1.000	0.780	0.000
Finland (FI)	1	1.000	0.780	0.000
Ukraine (UA)	1	1.000	0.780	0.000
Bulgaria (BG)	2	1.000	0.780	0.000
Slovenia (SI)	2	1.000	0.779	-0.001
Denmark (DK)	1	1.000	0.779	-0.001
Latvia (LV)	1	1.000	0.779	-0.001
Norway (NO)	3	1.000	0.779	-0.001
Poland (PL)	22	0.999	0.779	-0.001
Turkey (TR)	4	1.000	0.779	-0.001
Croatia (HR)	8	1.000	0.779	-0.001
Kosovo (KS)	2	1.000	0.779	-0.001
Belgium (BE)	42	1.000	0.779	-0.001
Czech Republic (CZ)	11	1.000	0.779	-0.001
Cyprus (CY)	4	1.000	0.779	-0.001
Scotland (Sct)	5	1.000	0.779	-0.001
Netherlands (NL)	17	1.000	0.779	-0.001
Macedonia (MK)	4	1.000	0.779	-0.001
Albania (AL)	3	1.000	0.779	-0.001
Bosnia-Herzegovina (BA)	9	1.000	0.779	-0.001
Greece (GR)	8	1.000	0.778	-0.002
Ireland (IE)	60	0.999	0.776	-0.004
Serbia-Montenegro (YG)	44	0.998	0.772	-0.008
Spain (ES)	136	0.994	0.770	-0.010
Portugal (PT)	126	0.990	0.769	-0.011
United Kingdom (GB)	199	0.998	0.764	-0.016

Table S4.6: Change of the Procrustes similarity when excluding one population from the Sub-Saharan African example. The Procrustes similarity between genetic coordinates and geographic coordinates is $t_0 = 0.790$ in the original analysis (Fig. 4.3).

Population excluded	Number of individuals excluded	Similarity to original PCA t'	Similarity to geography t''	$t'' - t_0$
Maasai (MKK)	30	0.980	0.832	0.042
Luhya (LWK)	30	0.999	0.808	0.018
Bamoun	18	1.000	0.797	0.007
Bantu (Kenya)	11	1.000	0.797	0.007
Fang	17	1.000	0.796	0.006
Mandenka	22	0.999	0.795	0.005
Kaba	16	1.000	0.794	0.004
Hausa	13	1.000	0.794	0.004
Igbo	17	1.000	0.791	0.001
Kongo	9	1.000	0.791	0.001
Yoruba	21	1.000	0.791	0.001
Alur	10	1.000	0.789	-0.001
Brong	7	1.000	0.788	-0.002
Dogon	24	0.995	0.788	-0.002
Bambaran	25	0.999	0.786	-0.004
Mada	12	1.000	0.785	-0.005
Hema	13	1.000	0.784	-0.006
Xhosa	3	1.000	0.783	-0.007
Bantu (S. Africa)	8	0.999	0.781	-0.009
Bulala	15	0.999	0.780	-0.010
Pedi	10	0.998	0.775	-0.015
Nguni	9	0.998	0.774	-0.016
Sotho/Tswana	8	0.997	0.768	-0.022

Table S4.7: Summary of the results for Sub-Saharan Africa when all or one of five additional African populations are included (corresponding to Fig. S4.3). θ is the rotation angle for the PCA map that optimizes the Procrustes similarity with the geographic map, and it is measured in degrees counterclockwise. P-values are obtained from 100,000 permutations of population labels.

Populations added	Panel in Fig. S4.3	Number of populations	Number of individuals collected	Number of individuals in the analysis	Variance explained by PC1 (%)	Variance explained by PC2 (%)	Rotation angle θ (°)	Procrustes similarity t_0	P-value of t_0	F_{ST} (%)
All	B	28	422	412	1.68	1.21	-78.47	0.548	0.00040	2.567
Mbororo Fulani	C	24	369	361	1.40	0.84	29.25	0.605	0.00005	1.518
Biaka Pygmy	D	24	378	369	1.26	1.03	20.01	0.559	0.00278	1.652
Mbuti Pygmy	E	24	369	359	1.27	1.19	-10.05	0.543	0.00120	1.781
!Kung	F	24	369	361	1.29	1.04	3.89	0.721	$< 10^{-5}$	1.616
San	G	24	361	354	1.31	0.92	5.66	0.725	$< 10^{-5}$	1.578

Table S4.8: Change of the Procrustes similarity when excluding one population from the Asian example. The Procrustes similarity between genetic coordinates and geographic coordinates is $t_0 = 0.849$ in the original analysis (Fig. 4.4).

Population excluded	Number of individuals excluded	Similarity to original PCA t'	Similarity to geography t''	$t'' - t_0$
Irula	24	0.993	0.871	0.022
Xibo	9	1.000	0.857	0.008
Tibetan	31	1.000	0.854	0.005
Kyrgyzstani	25	1.000	0.854	0.005
A.P. Brahmin	25	1.000	0.854	0.005
Nepalese	25	1.000	0.853	0.004
Yakut	25	0.999	0.853	0.004
T.N. Dalit	13	1.000	0.853	0.004
A.P. Mala	11	1.000	0.852	0.003
Hazara	22	1.000	0.852	0.003
A.P. Madiga	10	1.000	0.852	0.003
Naxi	8	1.000	0.852	0.003
T.N. Brahmin	14	1.000	0.851	0.002
Lahu	8	1.000	0.851	0.002
Yi	10	1.000	0.851	0.002
Dai	10	1.000	0.850	0.001
Tu	10	1.000	0.850	0.001
Thai	24	1.000	0.850	0.001
Uygur	10	1.000	0.849	0.000
Vietnamese	7	1.000	0.849	0.000
Tujia	10	1.000	0.849	0.000
Miao	10	1.000	0.849	0.000
Kalash	23	1.000	0.849	0.000
Stalskoe	5	1.000	0.849	0.000
Burusho	25	1.000	0.848	-0.001
Han (N. China)	10	1.000	0.848	-0.001
Iban	25	0.999	0.848	-0.001
Cambodian	10	1.000	0.848	-0.001
Pathan	22	1.000	0.847	-0.002
Hezhen	9	1.000	0.847	-0.002
She	10	1.000	0.847	-0.002
Mongola	10	1.000	0.847	-0.002
Makrani	20	1.000	0.847	-0.002
Balochi	22	1.000	0.847	-0.002
Japanese	28	1.000	0.847	-0.002
Brahui	23	1.000	0.847	-0.002
Daur	9	1.000	0.846	-0.003
Pakistani	25	1.000	0.846	-0.003
Sindhi	22	1.000	0.846	-0.003
Oroqen	9	1.000	0.846	-0.003
Urkarah	18	1.000	0.845	-0.004
Iraqi Kurd	24	1.000	0.845	-0.004
Han	34	1.000	0.844	-0.005
Buryat	25	1.000	0.839	-0.010

Table S4.9: Change of the Procrustes similarity when excluding one population from the East Asian example. The Procrustes similarity between genetic coordinates and geographic coordinates is $t_0 = 0.640$ in the original analysis (Fig. 4.5).

Population excluded	Number of individuals excluded	Similarity to original PCA t'	Similarity to geography t''	$t'' - t_0$
Japanese	28	0.999	0.755	0.115
Thai	20	0.994	0.691	0.051
Han	34	0.999	0.673	0.033
Xibo	8	1.000	0.655	0.015
Tibetan	31	0.996	0.655	0.015
She	10	1.000	0.654	0.014
Hezhen	9	1.000	0.645	0.005
Han (N. China)	10	1.000	0.645	0.005
Miao	10	1.000	0.642	0.002
Tujia	10	1.000	0.642	0.002
Mongola	10	1.000	0.640	0.000
Dai	10	1.000	0.637	-0.003
Vietnamese	7	1.000	0.637	-0.003
Tu	10	1.000	0.637	-0.003
Lahu	8	1.000	0.636	-0.004
Daur	9	1.000	0.636	-0.004
Cambodian	10	1.000	0.635	-0.005
Buryat	25	0.999	0.635	-0.005
Naxi	8	1.000	0.634	-0.006
Yi	10	1.000	0.631	-0.009
Oroqen	9	1.000	0.631	-0.009
Yakut	23	0.988	0.577	-0.063
Iban	25	0.993	0.561	-0.079

Table S4.10: Change of the Procrustes similarity when excluding one population from the Central/South Asian example. The Procrustes similarity between genetic coordinates and geographic coordinates is $t_0 = 0.737$ in the original analysis (Fig. 4.6).

Population excluded	Number of individuals excluded	Similarity to original PCA t'	Similarity to geography t''	$t'' - t_0$
Hazara	22	1.000	0.769	0.032
Kalash	23	1.000	0.754	0.017
A.P. Brahmin	25	1.000	0.749	0.012
T.N. Brahmin	14	1.000	0.748	0.011
Nepalese	25	1.000	0.747	0.010
Burusho	25	1.000	0.747	0.010
Pathan	22	1.000	0.740	0.003
Pakistani	25	1.000	0.736	-0.001
Sindhi	22	1.000	0.732	-0.005
A.P. Madiga	10	1.000	0.724	-0.013
Uygur	10	1.000	0.723	-0.014
A.P. Mala	11	1.000	0.721	-0.016
Kyrgyzstani	25	0.992	0.720	-0.017
Balochi	23	0.999	0.720	-0.017
T.N. Dalit	13	0.999	0.720	-0.017
Brahui	23	0.999	0.718	-0.019
Makrani	20	0.999	0.718	-0.019
Irula	24	0.979	0.717	-0.020

Table S4.11: Samples identified as PCA outliers in the analyses for different geographic regions. Note that AFH7 and AFH10, which appeared as PCA outliers in most of the Sub-Saharan African examples, are likely to be relatives based on allele-sharing analysis (results not shown).

Analysis	Region	Number of PCA outliers	Sample ID of PCA outliers	Population
Fig. 4.1	Worldwide	0	-	-
Fig. 4.2	Europe	2	POPR26466 POPR48136	Portugal (PT) Portugal (PT)
Fig. 4.3	Sub-Saharan Africa	2	AFH7 AFH10	Hema Hema
Fig. 4.4	Asia	11	HGDP00057 HGDP00060 HGDP00013 HGDP00029 HGDP00130 HGDP00139 HGDP00150 HGDP00154 HGDP00157 HGDP00173 HGDP00175	Balochi Balochi Brahui Brahui Makrani Makrani Makrani Makrani Makrani Sindhi Sindhi
Fig. 4.5	East Asia	7	F066579 F066599 F066607 F066612 HGDP01243 HGDP00949 HGDP00953	Thai Thai Thai Thai Xibo Yakut Yakut
Fig. 4.6	Central/South Asia	10	HGDP00057 HGDP00013 HGDP00029 HGDP00130 HGDP00150 HGDP00151 HGDP00154 HGDP00157 HGDP00173 HGDP00175	Balochi Brahui Brahui Makrani Makrani Makrani Makrani Makrani Sindhi Sindhi
Fig. S4.3B	Sub-Saharan Africa	4	AFH7 AFH10 NA21417 NA21596	Hema Hema Maasai (MKK) Maasai (MKK)
Fig. S4.3C	Sub-Saharan Africa	2	AFH7 AFH10	Hema Hema
Fig. S4.3D	Sub-Saharan Africa	3	AFH7 AFH10 NA21417	Hema Hema Maasai (MKK)
Fig. S4.3E	Sub-Saharan Africa	4	AFH7 AFH10 NA21417 NA21596	Hema Hema Maasai (MKK) Maasai (MKK)
Fig. S4.3F	Sub-Saharan Africa	2	NA21417 TSW25	Maasai (MKK) Sotho/Tswana
Fig. S4.3G	Sub-Saharan Africa	1	NA21417	Maasai (MKK)

CHAPTER V

Conclusion

Research presented in this dissertation has centered around development of statistical methods for population-genetic studies based on large-scale genetic data. I have explored two main topics: correcting for allelic dropout in microsatellite genotypes (Chapter II), and assessing the geographic structure of human genetic variation (Chapters III and IV). These two topics represent statistical challenges at two different stages of studies using large-scale genetic data: upstream data quality control and downstream data analysis. I have developed methods to accurately estimate allelic dropout rates from a single set of microsatellite genotypes, to prepare imputed data sets to circumvent allelic dropout in downstream microsatellite-based analyses, and to quantitatively compare and interpret spatial maps of population-genetic variation from separate analyses. In addition, based on a systematic analysis of genome-wide autosomal SNP variation in worldwide populations, I have provided a quantitative assessment of the similarity between genes and geography in different geographic regions around the world.

In Chapter II (*Wang et al.*, 2012a), I developed a maximum likelihood method together with an EM algorithm to correct for allelic dropout in microsatellite data. Unlike most existing methods that rely on replicate genotypes, my method was designed to estimate allelic dropout rates from a single nonreplicated set of microsatellite

genotypes. The method is based on a general allele frequency model, which allows different model assumptions — with or without Hardy-Weinberg equilibrium — and different sources of allelic dropout caused by both sample-specific and locus-specific factors. Based on extensive simulations, I showed that my estimates of sample-specific and locus-specific dropout rates are accurate and fairly robust to some violations of model assumptions, such as existence of population structure and genotyping errors from sources other than allelic dropout. Further, I devised an empirical Bayesian approach to both impute the missing data and replace some homozygous genotypes that might be mistakenly reported due to allelic dropout. Multiple imputed data sets generated by this strategy can then be used in downstream analyses to account for the genotype uncertainty that allelic dropout introduces. As an example, I demonstrated the strategy by showing that estimation of the observed heterozygosity from imputed data sets can effectively correct for a downward bias caused by allelic dropout. The method will be useful for a large number of ecologists and geneticists who often analyze microsatellite data genotyped from poor-quality samples. In particular, I have applied the method to improve a Native American data set that will be used for further studies by our collaborators. To assist other researchers in their work, I have implemented the method in a publicly available software program called *MicroDrop*.

In Chapter III (*Wang et al.*, 2010), I described a Procrustes analysis approach for quantitatively assessing the similarity of population-genetic and geographic maps. I confirmed in two scenarios, one using SNP data from European populations and the other using SNP data from worldwide populations, that a measurably high level of concordance exists between statistical maps of population-genetic variation and geographic maps of sampling locations. My third example involved comparing results from two partially overlapping worldwide samples, verifying the concordance of SNP analyses based on PCA and MDS. Further, I showed that statistical maps of worldwide CNVs generally accord with statistical maps of SNP variation, especially when

CNV analysis is limited to samples with the highest-quality data. These examples highlighted the potential of the Procrustes-based quantitative approach for comparing and interpreting statistical maps generated by PCA and MDS. In particular, one nice property of the Procrustes approach is that it makes the spatial maps generated by separate analyses commensurable, such that the maps can be presented in the same coordinate system. My examples illustrated this feature by superimposing statistical maps of genetic variation on the geographic maps to highlight the similarity between genes and geography. Using this feature, Procrustes analysis can have many other applications in population-genetic studies, especially when a common set of markers is not possible for all studied samples. In a recent study of ancient human DNA samples (*Skoglund et al.*, 2012), the Procrustes approach was used to integrate PCA results from separate analyses on each ancient sample, in which large amounts of missing genotypes were present due to poor quality of the DNA. Similarly, this approach can be applied to infer the ancestral origins of admixed individuals based on DNA segments from a certain ancestral group. The data, after excluding DNA segments from different ancestral groups, will have a lot of missing genotypes present at different genomic locations for different individuals, resulting in a missing data problem similar to that of the ancient DNA samples.

Chapter IV (*Wang et al.*, 2012b) presents the first systematic quantitative analysis of spatial patterns of human genetic variation based on PCA and Procrustes analysis. By systematically assessing the similarity between genes and geography in different locations using a shared set of autosomal SNPs, I showed that significant similarity scores can be obtained in different geographic regions and at different geographic levels. Surprisingly, the highest similarity score appeared in Asia rather than in Europe, where the qualitative similarity is perhaps best known in the literature (e.g. *Heath et al.*, 2008; *Lao et al.*, 2008; *Novembre et al.*, 2008). This unexpected result was found even though the Himalaya Mountains, as strong geographic barriers,

have generated noticeable genetic differentiation between Central/South Asians and East Asians (*Rosenberg*, 2011). Further, by examining the dependence of the Procruste similarity on the number of markers studied, I found that the number of SNPs required for convergence of PCA is inversely related to the level of population differentiation in the sample, as measured by F_{ST}. Together, the results quantitatively demonstrate the general existence of similarity between genes and geography, providing a systematic basis for evaluating the role of geography in human evolutionary history. In addition, the results suggest that using appropriate statistical methods, we can infer human individuals' ancestral geographic locations with high accuracy based on large amounts of genetic data.

In the era of genomics, fast accumulation of genetic data has brought exciting opportunities to learn about human evolutionary history and to dissect the genetic basis of complex diseases. This dissertation contributes two novel statistical tools to analyze large-scale genetic data from diverse populations, as well as a systematic discussion of the similarity between genes and geography across the globe. Results from this dissertation provide biological insights on the geographic structure of human genetic variation in worldwide populations. These results, together with the statistical tools, can benefit studies in many areas that rely on genetic variation data, including population genetics, evolutionary biology, molecular ecology, and medical genetics. Ideas presented in this dissertation can facilitate development of related statistical methods with applications in genetics to advance our knowledge on human evolution and genetic diseases.

BIBLIOGRAPHY

BIBLIOGRAPHY

Adams, D. C., F. J. Rohlf, and D. E. Slice (2004), Geometric morphometrics: ten years of progress following the 'revolution', *Ital. J. Zool.*, *71*, 5–16.

Auton, A., et al. (2009), Global distribution of genomic diversity underscores rich complex history of continental human populations, *Genome Res.*, *19*, 795–803.

Barbujani, G. (2000), Geographic patterns: how to identify them and why, *Hum. Biol*, *72*, 133–153.

Biswas, S., L. B. Scheinfeldt, and J. M. Akey (2009), Genome-wide insights into the patterns and determinants of fine-scale population structure in humans, *Am. J. Hum. Genet.*, *84*, 641–650.

Bonin, A., E. Bellemain, P. B. Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet (2004), How to track and assess genotyping errors in population genetics studies, *Mol. Ecol.*, *13*, 3261–3273.

Bookstein, F. L. (1996), Biometrics, biomathematics and the morphometric synthesis, *Bull. Math. Biol.*, *58*, 313–365.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza (1994), High resolution of human evolutionary trees with polymorphic microsatellites, *Nature*, *368*, 455–457.

Bregel, Y. (2003), *An Historical Atlas of Central Asia*, Brill, Boston.

Broquet, T., and E. Petit (2004), Quantifying genotyping errors in noninvasive population genetics, *Mol. Ecol.*, *13*, 3601–3608.

Broquet, T., N. Ménard, and E. Petit (2007), Noninvasive population genetics: a review of sample source, diet, fragment length and microsatellite motif effects on amplification success and genotyping error rates, *Conserv. Genet.*, *8*, 249–260.

Bryc, K., C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C. D. Bustamante, and H. Ostrer (2010a), Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations, *Proc. Natl. Acad. Sci. USA*, *107 (Suppl)*, 8954–8961.

Bryc, K., et al. (2010b), Genome-wide patterns of population structure and admixture in West Africans and African Americans, *Proc. Natl. Acad. Sci. USA*, *107*, 786–791.

Buchan, J. C., E. A. Archie, R. C. van Horn, C. J. Moss, and S. C. Alberts (2005), Locus effects and sources of error in noninvasive genotyping, *Mol. Ecol. Notes*, *5*, 680–683.

Casella, G., and R. L. Berger (2001), *Statistical Inference*, 2nd ed., Duxbury, Pacific Grove, CA.

Cavalli-Sforza, L. L., and M. W. Feldman (2003), The application of molecular genetic approaches to the study of human evolution, *Nat. Genet.*, *33 (Suppl)*, 266–275.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza (1994), *The History and Geography of Human Genes*, Princeton University Press, Princeton.

Chen, J., et al. (2009), Genetic structure of the Han Chinese population revealed by genome-wide SNP variation, *Am. J. Hum. Genet.*, *85*, 775–785.

Cox, T. F., and M. A. A. Cox (2001), *Multidimensional Scaling*, 2nd ed., Chapman & Hall, Boca Raton.

Dakin, E., and J. C. Avise (2004), Microsatellite null alleles in parentage analysis, *Heredity*, *93*, 504–509.

DeGiorgio, M., M. Jakobsson, and N. A. Rosenberg (2009), Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa, *Proc. Natl. Acad. Sci. USA*, *106*, 16,057–16,062.

Dryden, I. L., and K. V. Mardia (1998), *Statistical Shape Analysis*, Wiley, New York.

Du, R., and V. F. Yip (1993), *Ethnic Groups in China*, Science Press, Beijing.

Ellegren, H. (2000), Heterogeneous mutation processes in human microsatellite DNA sequences, *Nature Genet.*, *24*, 400–402.

Engelhardt, B. E., and M. Stephens (2010), Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis, *PLoS Genet.*, *6*, e1001,117.

Falush, D., M. Stephens, and J. K. Pritchard (2003), Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, *Genetics*, *164*, 1567–1587.

Fernando, P., T. N. C. Vidya, C. Rajapakse, A. Dangolla, and D. J. Melnick (2003), Reliable noninvasive genotyping: fantasy or reality?, *J. Hered.*, *94*, 115–123.

Friedlaender, J. S., et al. (2008), The genetic structure of Pacific Islanders, *PLoS Genet.*, *4*, e19.

Gagneux, P., C. Boesch, and D. S. Woodruff (1997), Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair, *Mol. Ecol.*, *6*, 861–868.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman (1995a), An evalutation of genetic distances for use with microsatellite loci, *Genetics*, *139*, 463–471.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman (1995b), Genetic absolute dating based on microsatellites and the origin of modern humans, *Proc. Natl. Acad. Sci. USA*, *92*, 6723–6727.

Gower, J. C. (1975), Generalized Procrustes analysis, *Psychometrika*, *40*, 33–51.

Gower, J. C., and G. B. Dijksterhuis (2004), *Procrustes Problems*, Oxford University Press.

Hadfield, J. D., D. S. Richardson, and T. Burke (2006), Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework, *Mol. Ecol.*, *15*, 3715–3730.

Hartl, D. L., and A. G. Clark (1997), *Principles of Population Genetics*, 3rd ed., Sinauer, Sunderland, MA.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York.

Heath, S. C., et al. (2008), Investigation of the fine structure of European populations with applications to disease association studies, *Eur. J. Hum. Genet.*, *16*, 1413–1429.

Henn, B. M., et al. (2011), Hunter-gatherer genomic diversity suggests a southern African origin for modern humans, *Proc. Natl. Acad. Sci. USA*, *108*, 5154–5162.

Heo, M., and K. R. Gabriel (1998), A permutation test of association between configurations by means of the RV coefficient, *Commun. Stat. Simul. Comp.*, *27*, 843–856.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio (2009), Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *Proc. Natl. Acad. Sci. USA*, *106*, 9362–9367.

Hirschhorn, J. N., and M. J. Daly (2005), Genome-wide association studies for common diseases and complex traits, *Nature Rev. Genet.*, *6*, 95–108.

Hoffman, J. I., and W. Amos (2005), Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion, *Mol. Ecol.*, *14*, 599–612.

Hoggart, C. J., P. F. O'Reilly, M. Kaakinen, W. Zhang, J. C. Chambers, J. S. Kooner, L. J. Coin, and M. R. Jarvelin (2012), Fine-scale estimation of location of birth from genome-wide single-nucleotide polymorphism data, *Genetics*, *190*, 669–677.

Holsinger, K. E., and B. S. Weir (2009), Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}., *Nat. Rev. Genet.*, *10*, 639–650.

Ihaka, R., and R. Gentleman (1996), R: a language for data analysis and graphics, *J. Comput. Graph. Stat.*, *5*, 299–314.

Itsara, A., et al. (2009), Population analysis of large copy number variants and hotspots of human genetic disease, *Am. J. Hum. Genet.*, *84*, 148–161.

Jackson, D. A. (1995), PROTEST: a Procrustean randomization test of community environment, *Ecoscience*, *2*, 297–303.

Jakkula, E., et al. (2008), The genome-wide patterns of variation expose significant substructure in a founder population, *Am. J. Hum. Genet.*, *83*, 787–794.

Jakobsson, M., et al. (2008), Genotype, haplotype and copy-number variation in worldwide human populations, *Nature*, *451*, 998–1003.

Johnson, P. C. D., and D. T. Haydon (2007), Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data, *Genetics*, *175*, 827–842.

Jombart, T., D. Pontier, and A.-B. Dufour (2009), Genetic markers in the playground of multivariate analysis, *Heredity*, *102*, 330–341.

Jorde, L. B., A. R. Rogers, M. Bamshad, W. S. Watkins, P. Krakowiak, S. Sung, J. Kere, and H. C. Harpending (1997), Microsatellite diversity and the demographic history of modern humans, *Proc. Natl. Acad. Sci. USA*, *94*, 3100–3103.

Jorde, L. B., et al. (1995), Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data, *Am. J. Hum. Genet.*, *57*, 523–538.

Kimmel, M., R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins, and L. B. Jorde (1998), Signatures of population expansion in microsatellite repeat data, *Genetics*, *148*, 1921–1930.

Lai, Y., and F. Sun (2004), Sampling distribution for microsatellites amplified by PCR: mean field approximation and its applications to genotyping, *J. Theor. Biol.*, *228*, 185–194.

Lange, K. (2002), *Mathematical and Statistical Methods for Genetic Analysis*, 2nd ed., Springer, New York.

Lao, O., et al. (2008), Correlation between genetic and geographic structure in Europe, *Curr. Biol.*, *18*, 1241–1248.

Li, J. Z., et al. (2008), Worldwide human relationships inferred from genome-wide patterns of variation, *Science*, *319*, 1100–1104.

Little, R. J. A., and D. B. Rubin (2002), *Statistical Analysis with Missing Data*, 2nd ed., John Wiley & Sons, Hoboken, NJ.

Mailman, M. D., et al. (2007), The NCBI dbGaP database of genotypes and phenotypes., *Nat. Genet.*, *39*, 1181–1186.

Mantel, N. (1967), The detection of disease clustering and a generalized regression approach, *Cancer Res.*, *27*, 209–220.

Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly (2004), The effects of human population structure on large genetic association studies, *Nature Genet.*, *36*, 512–517.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979), *Multivariate Analysis*, Academic Press, London.

McVean, G. (2009), A genealogical interpretation of principal components analysis, *PLoS Genet.*, *5*, e1000,686.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza (1978), Synthetic maps of human gene frequencies in Europeans, *Science*, *201*, 786–792.

Miller, C. R., P. Joyce, and L. P. Waits (2002), Assessing allelic dropout and genotype reliability using maximum likelihood, *Genetics*, *160*, 357–366.

Minch, E., A. Ruiz Linares, D. B. Goldstein, M. W. Feldman, and L. L. Cavalli-Sforza (1998), MICROSAT (version 1.5d): a program for calculating statistics on microsatellite data, Department of Genetics, Stanford University, Stanford, CA.

Morin, P. A., K. E. Chambers, C. Boesch, and L. Vigilant (2001), Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*), *Mol. Ecol.*, *10*, 1835–1844.

Mountain, J. L., and L. L. Cavalli-Sforza (1997), Multilocus genotypes, a tree of individuals, and human evolutionary history, *Am. J. Hum. Genet.*, *61*, 705–718.

Navidi, W., N. Arnheim, and M. S. Waterman (1992), A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations, *Am. J. Hum. Genet.*, *50*, 347–359.

Nelson, M. R., et al. (2008), The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research., *Am. J. Hum. Genet.*, *83*, 347–358.

Novembre, J., and S. Ramachandran (2011), Perspectives on human population structure at the cusp of the sequencing era, *Annu. Rev. Genomics Hum Genet.*, *12*, 245–274.

Novembre, J., and M. Stephens (2008), Interpreting principal component analyses of spatial population genetic variation, *Nature Genet.*, *40*, 646–649.

Novembre, J., et al. (2008), Genes mirror geography within Europe, *Nature*, *456*, 98–101.

Paschou, P., E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas (2007), PCA-correlated SNPs for structure identification in worldwide human populations, *PLoS Genet.*, *3*, 1672–1686.

Patterson, N., A. L. Price, and D. Reich (2006), Population structure and eigenanalysis, *PLoS Genet.*, *2*, 2074–2093.

Payseur, B. A., P. Jing, and R. J. Haasl (2011), A genomic portrait of human microsatellite variation, *Mol. Biol. Evol.*, *28*, 303–312.

Pemberton, J. M., J. Slate, D. R. Bancroft, and J. A. Barrett (1995), Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies, *Mol. Ecol.*, *4*, 249–252.

Pemberton, T. J., C. Wang, J. Z. Li, and N. A. Rosenberg (2010), Inference of unexpected genetic relatedness among individuals in HapMap Phase III, *Am. J. Hum. Genet.*, *87*, 457–464.

Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li (2012), Genomic patterns of homozygosity in worldwide human populations, *Am. J. Hum. Genet.*, *91*, 275–292.

Pena, S. D. J., and R. Chakraborty (1994), Paternity testing in the DNA era, *Trends Genet.*, *10*, 204–209.

Peres-Neto, P. R., and D. A. Jackson (2001), How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test, *Oecologia*, *129*, 169–178.

Pickrell, J. K., et al. (2009), Signals of recent positive selection in a worldwide sample of human populations, *Genome Res.*, *19*, 826–837.

Pompanon, F., A. Bonin, E. Bellemain, and P. Taberlet (2005), Genotyping errors: causes, consequences and solutions, *Nat. Rev. Genet.*, *6*, 847–859.

Powell, G. T., H. Yang, C. Tyler-Smith, and Y. Xue (2007), The population history of the Xibe in northern China: a comparison of autosomal, mtDNA and Y-chromosomal analyses of migration and gene flow, *Forensic Sci. Int. Genet.*, *1*, 115–119.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006), Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genet.*, *38*, 904–909.

Price, A. L., A. Helgason, S. Palsson, H. Stefansson, D. St. Clair, O. A. Andreassen, D. Reich, A. Kong, and K. Stefansson (2009), The impact of divergence time on the nature of population structure: an example from Iceland, *PLoS Genet.*, *5*, e1000,505.

Pritchard, J. K., and N. A. Rosenberg (1999), Use of unlinked genetic markers to detect population stratification in association studies, *Am. J. Hum. Genet.*, *65*, 220–228.

Prugnolle, F., A. Manica, and F. Balloux (2005), Geography predicts neutral genetic diversity of human populations, *Curr. Biol.*, *15*, R159–R160.

Ramachandran, S., and N. A. Rosenberg (2011), A test of the influence of continental axes of orientation on patterns of human gene flow, *Am. J. Phys. Anthropol.*, *146*, 515–529.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza (2005), Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa, *Proc. Natl. Acad. Sci. USA*, *102*, 15,942–15,947.

Robert, P., and Y. Escoufier (1976), A unifying tool for linear multivariate statistical methods: the RV-coefficient, *Appl. Stat.*, *25*, 257–265.

Rohlf, F. J., and D. Slice (1990), Extensions of the Procrustes method for the optimal superimposition of landmarks, *Syst. Zool.*, *39*, 40–59.

Rosenberg, N. A. (2011), A population-genetic perspective on the similarities and differences among worldwide human populations, *Hum. Biol.*, *83*, 659–684.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman (2002), Genetic structure of human populations, *Science*, *298*, 2381–2385.

Rosenberg, N. A., L. M. Li, R. Ward, and J. K. Pritchard (2003), Informativeness of genetic markers for inference of ancestry, *Am. J. Hum. Genet.*, *73*, 1402–1422.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman (2005), Clines, clusters, and the effect of study design on the inference of human population structure, *PLoS Genet.*, *1*, 660–671.

Salmela, E., et al. (2011), Swedish population substructure revealed by genome-wide single nucleotide polymorphism data, *PLoS One*, *9*, e16,747.

Schlötterer, C. (2004), The evolution of molecular markers — just a matter of fashion?, *Nat. Rev. Genet.*, *5*, 63–69.

Seeb, J. E., G. Carvalho, L. Hauser, K. Naish, S. Roberts, and L. W. Seeb (2011), Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms, *Mol. Ecol. Resour.*, *Suppl 1*, 1–8.

Sefc, K. M., R. B. Payne, and M. D. Sorenson (2003), Microsatellite amplification from museum feather samples: effects of fragment size and template concentration on genotyping errors, *Auk*, *120*, 982–989.

Sikora, M., H. Laayouni, F. Calafell, D. Comas, and J. Bertranpetit (2011), A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations, *Eur. J. Hum. Genet.*, *19*, 84–88.

Simonson, T., et al. (2010), Genetic evidence for high-altitude adaptation in Tibet., *Science*, *329*, 72–75.

Skoglund, P., H. Malmström, M. Raghavan, J. Storå, P. Hall, E. Willerslev, M. T. Gilbert, A. Götherström, and M. Jakobsson (2012), Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe, *Science*, *336*, 466–469.

Slatkin, M. (1995), A measure of population subdivision based on microsatellite allele frequencies, *Genetics*, *139*, 457–462.

Sokal, R. R., and F. J. Rohlf (1995), *Biometry*, 3rd ed., Freeman, New York.

Sokal, R. R., N. L. Oden, and C. Wilson (1991), Genetic evidence for the spread of agriculture in Europe by demic diffusion, *Nature*, *351*, 143–145.

Subramanian, S., R. K. Mishra, and L. Singh (2003), Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genome regions, *Genome Biol.*, *4*, R13.

Sun, J. X., J. C. Mullikin, N. Patterson, and D. E. Reich (2009), Microsatellites are molecular clocks that support accurate inferences about history, *Mol. Biol. Evol.*, *26*, 1017–1027.

Taberlet, P., and G. Luikart (1999), Non-invasive genetic sampling and individual identification, *Biol. J. Linn. Soc.*, *68*, 41–55.

Taberlet, P., S. Griffin, B. Goossens, S. Questiau, V. Manceau, N. Escaravage, L. P. Waits, and J. Bouvet (1996), Reliable genotyping of samples with very low DNA quantities using PCR, *Nucleic Acids Res.*, *24*, 3189–3194.

Taberlet, P., L. P. Waits, and G. Luikart (1999), Noninvasive genetic sampling: look before you leap, *Trends Ecol. Evol.*, *14*, 323–327.

The 1000 Genome Project Consortium (2010), A map of human genome variation from population-scale sequencing, *Nature*, *467*, 1061–1073.

The HUGO Pan-Asian SNP Consortium (2009), Mapping human genetic diversity in Asia, *Science*, *326*, 1541–1545.

The International HapMap 3 Consortium (2010), Integrating common and rare genetic variation in diverse human populations, *Nature*, *467*, 52–58.

The International HapMap Consortium (2003), The International HapMap Project, *Nature*, *426*, 789–796.

The International Human Genome Sequencing Consortium (2001), Initial sequencing and analysis of the human genome, *Nature*, *409*, 860–921.

Tian, C., R. Kosoy, A. Lee, M. Ransom, J. W. Belmont, P. K. Gregersen, and M. F. Seldin (2008), Analysis of East Asia genetic substructure using genome-wide SNP arrays, *PLoS One*, *3*, e3862.

Tishkoff, S. A., et al. (2009), The genetic structure and history of Africans and African Americans, *Science*, *324*, 1035–1044.

Urquhart, A., C. P. Kimpton, T. J. Downes, and P. Gill (1994), Variation in Short Tandem Repeat sequences — a survey of twelve microsatellite loci for use as forensic identification markers, *Int. J. Legal. Med*, *107*, 13–20.

Wang, C., Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy, A. B. Singleton, and N. A. Rosenberg (2010), Comparing spatial maps of human population-genetic variation using Procrustes analysis., *Stat. Appl. Genet. Mol. Biol.*, *9*, Article 13.

Wang, C., K. B. Schroeder, and N. A. Rosenberg (2012a), A maximum likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes, *Genetics*, *doi: 10.1534/genetics.112.139519*.

Wang, C., S. Zöllner, and N. A. Rosenberg (2012b), A quantitative comparison of the similarity between genes and geography in worldwide human populations, *PLoS Genet.*, *8*, e1002,886.

Wang, J. (2004), Sibship reconstruction from genetic data with typing errors, *Genetics*, *166*, 1963–1979.

Wang, K., M. Li, D. Hadley, R. Liu, J. Glessner, S. F. A. Grant, H. Hakonarson, and M. Bucan (2007a), PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data, *Genome Res.*, *17*, 1665–1674.

Wang, S., et al. (2007b), Genetic variation and population structure in Native Americans, *PLoS Genet.*, *3*, 2049–2067.

Wasser, S. K., C. Mailand, R. Booth, B. Mutayoba, E. Kisamo, B. Clark, and M. Stephens (2007), Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban, *Proc. Natl. Acad. Sci. USA*, *104*, 4228–4233.

Weir, B. S. (1996), *Genetic Data Analysis II*, Sinauer, Sunderland, MA.

Weir, B. S., and C. C. Cockerham (1984), Estimating *F*-statistics for the analysis of population structure, *Evolution*, *38*, 1358–1370.

Weissenbach, J., G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vaysseix, and M. Lathrop (1992), A second-generation linkage map of the human genome, *Nature*, *409*, 928–933.

Wright, J. A., R. J. Barker, M. R. Schofield, A. C. Frantz, A. E. Byrom, and D. M. Gleeson (2009), Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples, *Biometrics*, *65*, 833–840.

Xing, J., et al. (2009), Fine-scaled human genetic structure revealed by SNP microarrays, *Genome Res.*, *19*, 815–825.

Xing, J., et al. (2010), Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping., *Genomics*, *96*, 199–210.

Xu, S., et al. (2009), Genomic dissection of population substructure of Han Chinese and its implication in association studies, *Am. J. Hum. Genet.*, *85*, 762–774.

Yamaguchi-Kabata, Y., K. Nakazono, A. Takahashi, S. Saito, N. Hosono, M. Kubo, Y. Nakamura, and N. Kamatani (2008), Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies, *Am. J. Hum. Genet.*, *83*, 445–456.

Yang, W. Y., J. Novembre, E. Eskin, and E. Halperin (2012), A model-based approach for analysis of spatial structure in genetic data, *Nat. Genet.*, *44*, 725–731.

Zhivotovsky, L. A., N. A. Rosenberg, and M. W. Feldman (2003), Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers, *Am. J. Hum. Genet.*, *72*, 1171–1186.