Genetics and Genomics of Human Population Structure

20

Sohini Ramachandran, Hua Tang, Ryan N. Gutenkunst, and Carlos D. Bustamante

Abstract Recent developments in sequencing technology have created a flood of new data on human genetic variation, and this data has yielded new insights into human population structure. Here we review what both early and more recent studies have taught us about human population structure and history. Early studies showed that most human genetic variation occurs within populations rather than between them, and that genetically related populations often cluster geographically. Recent studies based on much larger data sets have recapitulated these observations, but have also demonstrated that high-density genotyping allows individuals to be reliably assigned to their population of origin. In fact, for admixed individuals, even the ancestry of particular genomic regions can often be reliably inferred. Recent studies have also offered detailed information about the composition of specific populations from around the world, revealing how history has shaped their genetic makeup. We also briefly review quantitative models of human genetic history, including the role natural selection has played in shaping human genetic variation.

Contents

S. Ramachandran (✉)
Society of Fellows, Harvard University, 78 Mount Auburn Street, Cambridge, MA 02138, USA
e-mail: sramach@fas.harvard.edu

H. Tang
Department of Genetics, Stanford Medical School, Stanford, CA, USA
e-mail: huatang@stanford.edu

R.N. Gutenkunst
Theoretical Biology and Biophysics, and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, USA
e-mail: ryang@lanl.gov

C.D. Bustamante
Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA
e-mail: cdb28@cornell.edu

M.R. Speicher et al. (eds.), *Vogel and Motulsky's Human Genetics: Problems and Approaches,*
DOI 10.1007/978-3-540-37654-5_20, © Springer-Verlag Berlin Heidelberg 2010

20

20.1 Introduction

Technological developments arising from the International Human Genome Sequencing and the International Haplotype Map (The International HapMap Consortium, 2003, 2005, 2007) [20–22] projects are transforming the study of human population genetics by dramatically reducing the cost of sequencing and genotyping. For example, as of early 2009, it costs about U.S. $500 per sample to genotype a million variable DNA sites (i.e., SNPs) and structural variants in the human genome and between $50,000 and $100,000 to sequence a human genome de novo. Recalling that the first human genome cost on the order of $1 billion dollars to sequence, this is a 10^4 gain in efficiency over less than a decade. Furthermore, by the time this book is published the costs we quote above may be reduced by another factor of two or three. In the next 5–10 years, therefore, we will likely see hundreds of thousands (if not millions) of human genomes sequenced, and the vast majority of variation within and among human populations cataloged and analyzed to answer fundamental questions in human and medical genomics.

The purpose of this chapter is to lay the groundwork for thinking about how we will begin to make use of this tremendous abundance of data. While these data will dwarf all that has come before, we will see that many of the questions we wish to answer are actually quite old – some as old as the field of human genetics, itself.

20.1.1 Evolutionary Forces Shaping Human Genetic Variation

Quantifying patterns of human genetic variation serves several important roles in genetics. First, it helps us understand human history and often gives us insights into time periods that have left no written record. For example, global patterns of human genetic variation suggest an African origin of modern humans approximately 150,000–200,000 years ago and are consistent with a "serial" founder model (see Sect. 20.5.1) for subsequent colonization and peopling of the world. Second, it helps us understand human *evolutionary* history. For example, patterns of human genetic variation allow us to delineate what genomic changes are unique to our species (i.e., shared by all humans to the exclusion of other apes), and which may be shared ancestrally (or recurrently) with other species. Likewise, patterns of human genetic variation can give us insight into regions of the human genome that may have experienced recent positive, negative, or balancing selection (see Nielsen et al. [39] for a recent review).

Understanding patterns of human genetic variation is also fundamental for the proper design of medical genomic studies, since population structure can often be a confounding variable in genome-wide association mapping. As the density of markers queried for association with disease increases and we begin to look at rare variants that may show limited geographic distributions, quantifying population structure at ever finer scales will be critical to the interpretation and analysis of experiments which aim to correlate patterns of genetic and phenotypic variation. In order to properly set the stage for our discussion, we will briefly review some key concepts from population genetics, anthropology, and genetics that may be unfamiliar to some readers.

The evolutionary dynamics of natural populations (be they human, plant, animal, or otherwise) are governed by a confluence of different evolutionary forces.

Chief among these is *mutation*, which is the ultimate source of variation. As this book illustrates, the process of mutation is a heterogeneous category of changes in DNA that come about through myriad pathways and ultimately induce changes ranging from single base pair alterations (i.e., single nucleotide polymorphisms or SNPs) to small insertion and deletions to large-scale structural rearrangments or even the addition or deletion of whole chromosomes. Most of the variation we will discuss in this chapter will be of the "small scale" variety, with a particular emphasis on understanding patterns of microsatellite, SNP, and haplotype variation.

We limit ourselves to these marker types largely due to practicality: assaying SNP and microsatellite variation has become standardized, and there are now a plethora of studies – such as those cited later on in this chapter – that have undertaken surveys using these markers across diverse human populations. Our hope is that, as the world of personalized genomics becomes a reality, large and micro-scale structural variation becomes cataloged and standardized in similar ways.

The second key force shaping patterns of human genetic variation is *genetic drift*. As you will recall from Chap. 16, genetic drift is a stochastic force that apportions variation by randomly subsampling variation from one generation to the next. Traditionally, we model genetic drift as simple binomial sampling of alleles. That is, if we consider a biallelic locus under no selection and represent the frequency of an allele *A* at time *t* in a given population of size *2N* as x_t, the frequency in the next generation (x_{t+1}) is binomially distributed with probability of success x_t and sample size *2N*. (It turns out this binomial distribution can be generalized and there is a rich treatment of this subject in theoretical population genetics.) This random sampling from generation to generation induces what is known as a "random walk," such that the collection of allele frequencies from the start of the population history until the current time ($x_0, x_1, \ldots, x_t$) as well as the distribution of long-term average frequencies across different sites can be modeled using a litany of theoretical tools.

For our purposes, we will focus on several qualitative impacts of this neutral evolutionary model. First, for a given population, the dynamics of genetic drift will be governed by the magnitude of *2N*, so that populations with a large number of individuals will "drift" more slowly or take smaller steps in frequency space from generation to generation than small populations. This model also suggests that if we were to follow lines-of-descent (i.e., the number of offspring left some time in the future by a given lineage today) with no difference in average offspring number among lineages, then the probability of a given lineage eventually overtaking the population is simply given by its current frequency. (For example, a lineage or allele at 20% frequency has a 20% chance of eventually getting fixed in the population, and an 80% chance of eventually getting lost.) Likewise, the model predicts that frequency is often a good proxy for age (at least for neutral alleles) so that a mutation at 25% frequency in the population is very likely to be older than a mutation at 5% frequency. For this reason, the distribution of SNP frequencies or the so-called allele-frequency spectrum contains a fair amount of information regarding the history of the population. Mathematically, we would define this quantity using an equation such as the following for a population with sample size of n_i individuals:

$$Y_i := \{\text{the number of SNPs where the sample frequency is } i/(2n_i)\}. \quad (20.1)$$

We will return to Y_i later in the chapter and discuss methods for inferring demographic history and selection from these frequencies. (Note: in the equation above we are assuming directionality as to which allele is the ancestral form and which is the derived. In practice, we infer this information from comparative genomic data, ideally, with correction for multiple mutations occurring at the site. See [17, 18] for a discussion of this problem).

The third force that will affect patterns of human genetic variation is *migration* or, more generally, *demographic history*. By this we mean that a given population (certainly for humans) is unlikely to reproduce as a fully endogamous unit. Rather, there is some probability every generation that new migrants from other populations may enter and contribute to the gene pool of the next generation. We also know that a given population is unlikely to remain the exact same size from generation to generation; it may increase or decrease in size, or go through boom/bust cycles. The number of demographic models one can construct is staggering, but certain general properties of models are described below.

For example, populations that have a closely shared evolutionary history – say they are exchanging migrants often or split from a common ancestral population a short time ago – will show a strong and positive correlation in allele frequencies both over time (i.e., the two populations' x_t values for a specific SNP will be correlated over time) as well as across the genome (i.e., the observed Y_i values will be correlated). We can also define a quantity such as the "joint allele frequency spectrum" to help us quantify this correlation and gauge the impact of different evolutionary forces on sets of populations. Mathematically, for a pair of populations *i* and *j* with sample size n_i and n_j this might take the form of a quantity Y_{ij} such that:

$$Y_{ij} = \{\text{the number of SNPs where the sample frequency is } i/(2n_i) \text{ in population } i \text{ and } j/(2n_j) \text{ in population } j\}. \quad (20.2)$$

As we will see throughout this chapter, the allele frequency spectrum both of a single population (i.e., Y_i) and for a pair (Y_{ij}) or more ($Y_{ijkl\ldots}$) contains a fair amount of information regarding the evolutionary history of the populations in question. Many of the commonly used statistics in population genetics such as Wright's F-statistics, defined in Sect. 20.2.1, are

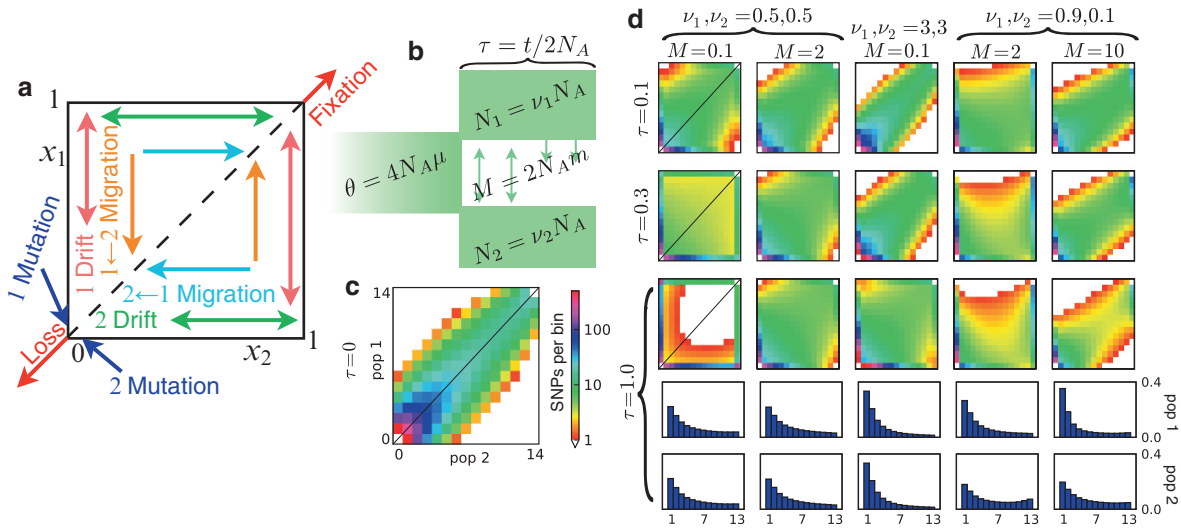

Fig. 20.1 Frequency spectrum gallery (adapted from Gutenkunst et al., manuscript submitted). (**a**) Impact of different evolutionary forces on shared patterns of genetic variation for a pair of populations, as defined by the density of alleles at relative frequencies x_1 and x_2 in populations 1 and 2. (**b**) Graphical description of an evolutionary model in which a pair of populations diverge and continue to exchange migrants. Specifically, an equilibrium population of effective size N_A diverges into two populations $2N_A\,\tau$ generations ago. Populations 1 and 2 have effective sizes $\nu_1 N_A$ and $\nu_2 N_A$, respectively. Migration is symmetric at $m = M/(2N_A)$ per generation, and the scaled mutation rate $\theta = 1{,}000$. (**c**) The allele frequency spectrum (AFS) at $\tau = 0$. Each entry is colored according to the logarithm of the number of SNPs with a given pairwise sample frequency, ranging from 0 to 14 copies of an allele in each population. (**d**) The AFS at various times for various demographic parameters, on the same scale as **c**. From the two-dimensional spectra, note that increased migration leads to more correlated SNP frequencies, and differences in population size lead to asymmetric genetic drift and thus an asymmetric AFS. For $\tau = 1$, the single-population spectra are also shown, where the scale is the fraction of polymorphisms observed at a given sample frequency. In these, note in particular that when populations experience growth, the spectrum is skewed toward rare alleles, particularly for the middle scenario.

summaries of these quantities. In Fig. 20.1 we show how different demographic forces acting on a population can impact their marginal (Y_i) and joint (Y_{ij}) site-frequency spectra.

The fourth force which contributes to the distribution of human genomic variation is natural selection. As was discussed in the chapter on population genetics (Chap. 16), selection works to decrease the frequency of deleterious alleles, increase the frequency of positively selected variants, and stabilize the frequency of variants subject to balancing selection. In human populations, it appears that selection is a much weaker force than genetic drift or demographic history in shaping global patterns of genomic variation. Nonetheless, there are some clear examples of positive and balancing selection on the human genome which have been recently reviewed (see [39]). Here, we will discuss selection briefly and mostly in light of selection against deleterious alleles, since this is the most prevalent form of selection operating on the human genome (see Sect. 20.6.2).

20.2 Quantifying Population Structure

In this section, we introduce several methods for quantifying and detecting population structure. We begin by introducing the classic *F*-statistics, which measure the degree of genetic differentiation among pre-defined and discrete subpopulations. We will then focus on model-based clustering methods, which aim to characterize latent and possibly nondiscrete population structure.

20.2.1 F_{ST} and Genetic Distance

Nonrandom mating in a population with substructure has two consequences: first, preferential mating between individuals from the same subpopulation is a form of *inbreeding*, and has the effect of reducing genetic diversity (measured as, say, heterozygosity) in the overall population; second, as the subpopulations

experience independent genetic drift, allele frequencies at genetic markers tend to diverge. Originally introduced by Wright in 1921 [69] to quantify the inbreeding effect of population substructure, F_{ST} has become one of the most widely used measures of genetic differentiation between predefined subpopulations.

Consider the simple setting, in which a population consists of several subpopulations. F_{ST} is defined as the decrease in heterozygosity among subpopulations (H_S), relative to the heterozygosity in the total population (H_T):

$$F_{ST} = \frac{H_T - H_S}{H_T}, \tag{20.3}$$

where H_S is the *expected* heterozygosity, computed under the assumption that mating is random within each subpopulation (Hardy-Weinberg equilibrium), while H_T is analogously computed assuming random mating in the entire population without population structure.

Alternatively, F_{ST} is often loosely interpreted as the proportion of variance in allele frequencies at a locus that is explained by the subpopulation level of organization. For example, suppose the frequency of an allele is 0 and 1 in two subpopulations, respectively, then $F_{ST}=1$, meaning the variance in allele frequency is completely explained by the population division. Under this framework, F_{ST} at a biallelic single nucleotide polymorphism (SNP) marker can be computed based on the allele frequencies:

$$F_{ST} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})}, \tag{20.4}$$

where σ_p^2 is the variance of allele frequencies among subpopulations and $\bar{p}$ denotes the average allele frequency in the pooled population. It can be shown that (20.3) and (20.4) are mathematically equivalent for biallelic markers, but (20.4) is often computationally more convenient.

F_{ST} is often taken as a genetic distance measure, with higher values of F_{ST} reflecting a greater level of genetic divergence. However, both (20.3) and (20.4) define F_{ST} for a specific locus; F_{ST} can vary considerably from locus to locus. Moreover, a locus that is under population- or environment-specific selection can also exhibit unusually high F_{ST}. For example, across globally-distributed human populations, functional polymorphisms in genes related to skin pigmentation show unusually high levels of F_{ST} (i.e., population differentiation) as compared to the genome-wide distribution (see [45]). To reduce the variance across the markers and the bias due to a small number of strongly selected loci, when F_{ST} is reported as an index for genetic distance among subpopulations, it is often calculated by averaging both the numerator and the denominator in (20.3) or (20.4) across loci.

When one is interested in quantifying the degree of substructure among predefined populations, F_{ST} is a simple and useful measure of genetic distance. However, it is often the case that we are interested in using the genetic data itself to define the populations. In particular, if we are interested in detecting cryptic or hidden population structure, then we need to resort to other approaches (see Sect. 20.4). One method for detecting latent population structure, principal component analysis (PCA), was introduced in Sect. 6.4.4. In the next section, we explain a complementary approach, which defines subpopulations based on statistical genetic models for the data.

20.2.2 Model-Based Clustering Algorithms

Cluster analysis refers to a large family of approaches, whose goal is to simultaneously define subsets (called clusters) and to assign observational units into these clusters, so that members in the same cluster are similar by some criteria. For a comprehensive survey of clustering approaches, readers are referred to Mardia et al. [32] or Hastie et al. [15].

In the context of inferring genetic structure, the data usually consist of individuals genotyped at multiple genetic markers (e.g., restriction fragment length polymorphisms RFLPs, microsatellites, or SNPs). In the discrete population model, all alleles in an individual are assumed to be drawn randomly from one of the subpopulations, according to a set of allele frequencies that are specific to each subpopulation. The goal of the analysis is to simultaneously estimate subpopulation allele frequencies and group membership (i.e., which individuals are drawn from which subpopulation). However, for many human populations, there is often no single group from which individuals derive their ancestry. That is, recent migration gives rise to

genetically admixed individuals, whose genomes represent a mixture of alleles from multiple "ancestral" populations (see Sect. 20.5).

Mathematically, this means that an individual may have partial membership in more than one cluster. These clusters are biologically interpreted as ancestral populations for the admixed individuals. For example, African Americans in the United States are a recently admixed group, deriving ancestry from European and West African ancestral populations [65]. Under the admixture model, an African American individual's population membership is characterized by the *individual ancestry* (IA) proportion, which is a vector representing the probability that a randomly selected allele from this individual originates from a European (or alternatively, an African) ancestor.

Under either the discrete or the admixture model, individuals' memberships (or IA values) are jointly inferred with the allele frequencies in each subpopulation, using either maximum likelihood or Bayesian methods. We begin by explaining the maximum likelihood approach for the discrete subpopulation model, as this model illustrates the principles that underlie most of the model-based approaches [63]. Let $G_i^m = (a(i,m), b(i,m))$ denote the genotype of individual i at marker m, with $a(i,m)$ and $b(i,m)$ being the unordered pair of alleles. Let $Z_i \in (1,\ldots, k)$ indicate the subpopulation membership for individual i, and $P = \{p_{m_l}^k\}$ be the frequency of allele l at marker m in population k. Under the assumption that genotypes among markers are independent conditioning on an individual's membership, and that all markers are in Hardy-Weinberg equilibrium within each subpopulation, the likelihood function, treating Z and P as parameters, is simply the product of the probability of observing each allele:

$$L(P,Z;G) \propto \prod_i \prod_m p_{m_{a(i,m)}}^{z_i} p_{m_{b(i,m)}}^{z_i}. \qquad (20.5)$$

For the admixture model, one can substitute Z_i by $(Z_{i,m}^a, Z_{i,m}^b)$, the population origin of each allele, and model $Z_{i,m}^a$ and $Z_{i,m}^b$ as independent draws from the multinomial probability vectors of individual ancestry. The inference of population structure amounts to the inference on Z_i, or the genome-wide average of $(Z_{i,m}^a, Z_{i,m}^b)$.

In the maximum likelihood approach, the expectation maximization (EM) algorithm can be used to find the maximum likelihood estimates for the parameter values, (P,Z) [57, 63, 70, 74]. Alternatively, Bayesian approaches incorporate prior distributions into the likelihood, in order to evaluate the posterior distribution. The Bayesian methods offer a flexible framework for incorporating more complex population history models. For example, one of the widely used Bayesian programs, STRUCTURE, includes useful features such as modeling linkage among loci, and the ability to model correlated allele frequencies between evolutionarily related ancestral populations [14, 49].

20.2.3 Characterizing Locus-Specific Ancestry

For admixed populations, methods described in the preceding section can be used to infer individual ancestry, which represents the genome-wide average ancestry proportions in an individual. If admixture has occurred recently, the genome of an admixed individual resembles a mosaic of fairly long chromosomal blocks derived from one of the ancestral populations. With high-density genotype data, it is now feasible to delineate these ancestry blocks with relatively high accuracy. Figure 20.2 illustrates how ancestry blocks can be reconstructed. While numerous statistical methods have been proposed (e.g., [60, 62]), it is important to realize that the source of information underlying all methods is the different allele and haplotype frequencies among the ancestral populations. As such, the accuracy with which one can infer locus-specific ancestry depends on the genetic divergence between the ancestral populations. The distribution of the ancestry blocks also depends on the admixing history: ancient admixing events result in smaller ancestry fragments, while recent admixing events give rise to extended blocks. With any method, the ability to identify a switch in ancestral state deteriorates when the blocks are very small. Therefore, the accuracy of locus-specific ancestry inference depends on (at least) two aspects of the population history: the divergence between the ancestral populations, and the time of the admixing events. Simulation studies using HapMap data suggest that current high-density genotype data harbor sufficient information for accurate ancestry inference for African-Americans or Hispanics [62]. Locus-specific ancestry can provide information regarding the population history of an admixed

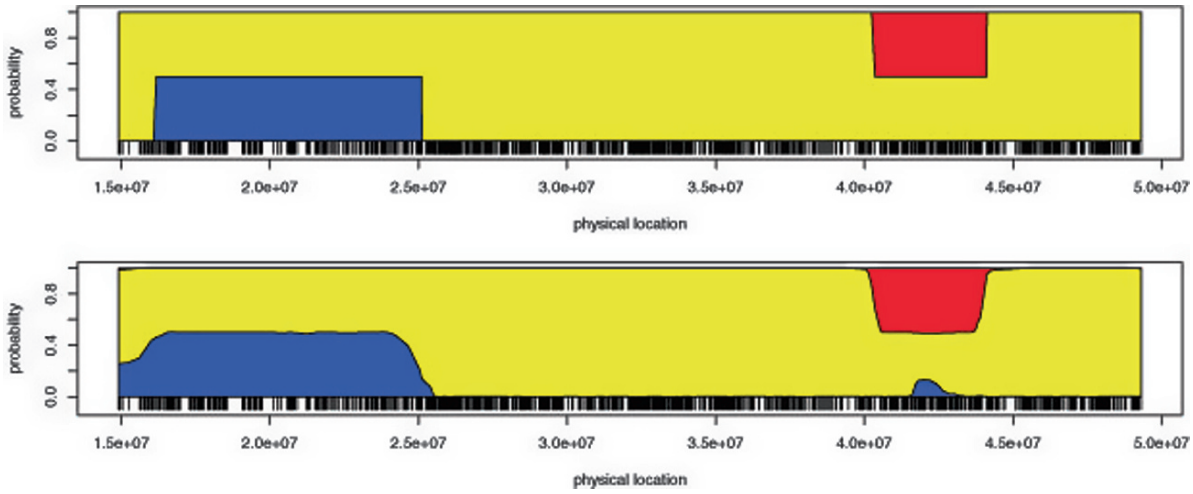

Fig. 20.2 Estimating ancestry along a chromosome. The top panel shows ancestry blocks along a simulated chromosome (*red*: African, *blue*: European; *yellow*: Asian). The bottom panel shows the reconstructed ancestry using high-density SNP markers, which are indicated by the black ticks at the bottom of each panel

population, as well as the finer-scale genetic structure within admixed groups. These are topics we will discuss in greater detail in Sect. 20.5.

20.3 Global Patterns of Human Population Structure

In this and the subsequent section, we begin a detailed exploration of empirical studies of human population genetic structure. First we explore major studies and datasets, now paradigms in the field of human population genetics, that compare human genetic variation at the level of multiple continents; the title of each subsection in this section is the title of a major paper or book in human population genetics. More recent studies of high-density genotyping data reveal patterns in genetic variation at fine geographic scales, as will be discussed after this section's historical perspective is presented.

20.3.1 The Apportionment of Human Diversity

Most studies of human population genetics begin by citing a seminal 1972 paper by Richard Lewontin bearing the title of this subsection [29]. Given the central role this work has played in our field, we will begin by discussing it briefly and return to its conclusions throughout the chapter. In this paper, Lewontin summarized patterns of variation across 17 polymorphic human loci (including classical blood groups such as ABO and M/N as well as enzymes which exhibit electrophoretic variation) genotyped in individuals across classically defined "races" (Caucasian, African, Mongoloid, South Asian Aborigines, Amerinds, Oceanians, Australian Aborigines [29]). A key conclusion of the paper is that 85.4% of the total genetic variation observed occurred within each group. That is, he reported that the vast majority of genetic differences are found within populations rather than between them. In this paper and his book *The Genetic Basis of Evolutionary Change* [30], Lewontin concluded that genetic variation, therefore, provided no basis for human racial classifications.

Lewontin's argument is an important one, and separates studying the geographic distribution of genetic variation in humans from searching for a biological basis to racial classification. His finding has been reproduced in study after study up through the present: two random individuals from any one group (which could be a continent or even a local population) are almost as different as any two random individuals from the entire world (see proportion of variation within populations in Table 20.1 and [20]).

An important point to realize is that Lewontin's calculation (and later work that confirms his finding) are based on the *F*-statistics introduced in Sect. 20.2.1 (see

Table 20.1 In this analysis of molecular variance, the total genetic variation observed is partitioned by that explained within populations in the same sample, among populations within regions, and among regions (from [53], reprinted with permission from AAAS)

			Variance components and 95% confidence intervals (%)		
Sample	Number of regions	Number of populations	Within populations	Among populations within regions	Among regions
World	1	52	94.6 (94.3, 94.8)	5.4 (5.2, 5.7)	
World	5	52	93.2 (92.9, 93.5)	2.5 (2.4, 2.6)	4.3 (4.0, 4.7)
World	7	52	94.1 (93.8, 94.3)	2.4 (2.3, 2.5)	3.6 (3.3, 3.9)
World-B97	5	14	89.8 (89.3, 90.2)	5.0 (4.8, 5.3)	5.2 (4.7, 5.7)
Africa	1	6	96.9 (96.7, 97.1)	3.1 (2.9, 3.3)	
Eurasia	1	21	98.5 (98.4, 98.6)	1.5 (1.4, 1.6)	
Eurasia	3	21	98.3 (98.2, 98.4)	1.2 (1.1, 1.3)	0.5 (0.4, 0.6)
Europe	1	8	99.3 (99.1, 99.4)	0.7 (0.6, 0.9)	
Middle East	1	4	98.7 (98.6, 98.8)	13 (1.2, 1.4)	
Central/South Asia	1	9	98.6 (98.5, 98.8)	1.4 (1.2, 1.5)	
East Asia	1	18	98.7 (98.6, 98.9)	1.3 (1.1,1.4)	
Oceania	1	2	93.6 (92.8, 94.3)	6.4 (5.7, 7.2)	
America	1	5	88.4 (87.7, 89.0)	11.6 (11.0,12.3)	

[67] for a discussion) averaged across single genetic loci. While it is an undeniable mathematical fact that the amount of genetic variation observed within groups is much larger than the differences among groups, this does not mean that genetic data do not contain discernable information regarding genetic ancestry. In fact, we will see that minute differences in allele frequencies across loci when compounded across the whole of the genome actually contain a great deal of information regarding ancestry. Given current technology, for example, it is feasible to accurately identify individuals from populations that differ by as little as 1% in F_{ST} if enough markers are genotyped. (See discussion below for a detailed treatment of the subject.) It is also important to note that when one looks at correlations in allelic variation across loci, self-identified populations and populations inferred for human subjects using genetic data correspond closely [12, 53].

20.3.2 The History and Geography of Human Genes

For more than 40 years, Luigi Luca Cavalli-Sforza and colleagues have worked to document and interpret patterns of human genetic variation. Along the way they have developed and perfected many of the statistical methods used to visualize and quantify patterns of variation and interpret their findings in light of human history and evolution. Their canonical book, *The History and Geography of Human Genes*, summarizes much of what they have learned about the pattern and process of human genetic variation across 1,800 indigenous populations.

Before delving into their findings, it is important to define two important concepts that permeate their work and those of the field a whole. The first is *treeness*, a concept introduced by Cavalli-Sforza and Piazza [5] to summarize population structure across multilocus data. Statistically, we can think of *treeness* as a way of summarizing "block structures" seen in matrices of pairwise genetic distances between populations. Specifically, block structures emerge when populations descended from a common ancestor are grouped together in these matrices, since closely related populations (say sister populations) will show similar levels of differentiation to a distant pair of closely related populations, the matrix will appear to show nearly duplicated rows and columns (or "blocks" of relatedness). By summarizing the blocks as arising from bifurcating trees, one can in theory build up a history of the population splitting events that gave rise to the sampled groups. It is important to emphasize that population trees are somewhat different from traditional phylogenetic (or species) trees since they are summarizing a reticulated history with often a great deal of gene flow among the terminal branches. The second concept that is important to discuss is the technique of principal component analysis (PCA). As we have

already seen in Sect. 6.4.4, PCA is a general tool for exploratory data analysis that has found wide application in genetics. Cavalli-Sforza and colleagues were among the first to use PCA of population allele frequency matrices to identify major axes of variation in the data and interpret these axes in light of human history, as we will discuss below. One important distinction to emphasize is that much of the PCA work they carried out was done at the *population* level while much of the PCA that is carried out today is done at the individual level. (That is, PCA analysis of genotype value matrices where the entries are "0," "1," or "2" depending on how many copies of the "A" allele vs. the "a" allele, a given individual carries at a locus).

Using PCA Cavalli-Sforza and colleagues deeply explored human population genetics structure in *The History and Geography of Human Genes*. (A representative example of the PCA plots they generated is given in Fig. 20.3, which summarizes major axes of variation across the sampled populations they studied). A key emphasis of their work was on understanding how or whether language presented barriers to gene flow (i.e., quantifying how much of nonrandom mating in human populations is attributable to language) (see Fig. 20.4). The idea that languages and genes may evolve at similar rates and that a similarity in linguistic markers between two languages may likely reflect a recent shared genetic history among speakers of those languages remains controversial in the field of linguistics. However, Cavalli-Sforza et al. [8] underscored that human evolutionary genetics studies can rely on data and results from other fields – such as anthropology, archaeology, and linguistics – to synthesize inferences about human history.

The book by Cavalli-Sforza and colleagues is known for its numerous *synthetic maps* [34]. Synthetic

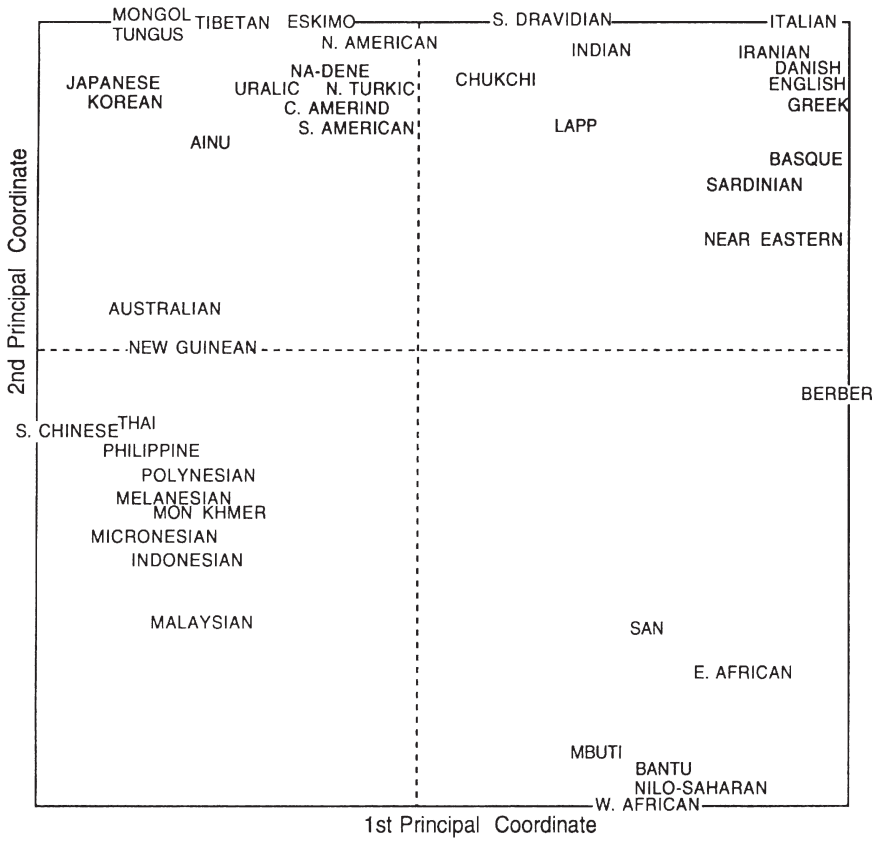

Fig. 20.3 Principal component map of 42 population studies by [6]. The first two PCs summarize 27% and 16% of the variation, respectively. Africans cluster in the lower right quadrant, with Europeans in the upper right, Southeast Asians in the lower left, Northeast Asians and Americans in the upper left. The first PC separates Africans and Europeans from the rest; the authors propose that the first PC does not separate Africans from non-Africans because there are only 6 African populations compared to 36 other populations. From [7]. From Cavalli-Sforza L., The History and Geography of Human Genes, copyright 1994 Princeton University Press. Reprinted by permission of Princeton University Press

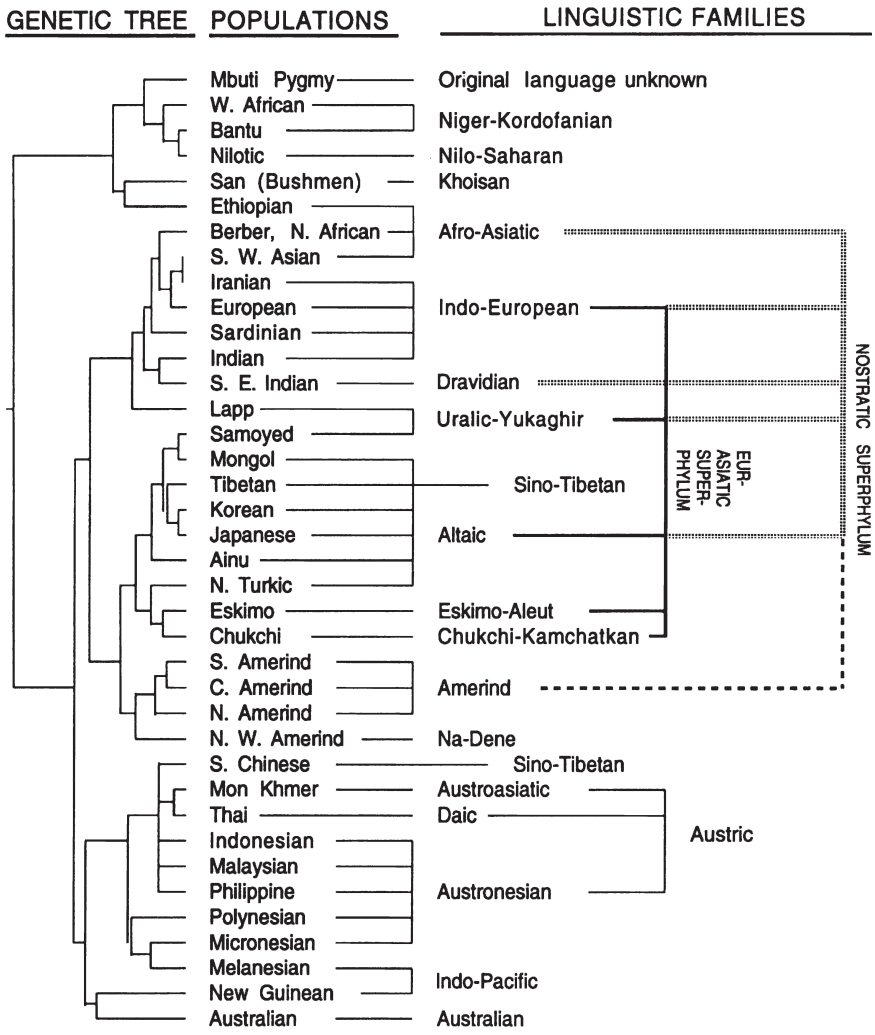

Fig. 20.4 Aligning the genetic tree with linguistic families and superfamilies, as in [8]. From [7]. From Cavalli-Sforza L., The History and Geography of Human Genes, copyright 1994 Princeton University Press. Reprinted by permission of Princeton University Press

maps overlap a single principle component onto a geographic map and are interpreted as revealing migration routes taken through the frequency and geographic spread of the allelic variant (see last three columns of Fig. 20.5). Many synthetic maps reveal north-south and east-west gradients in genetic variation, which might be interpreted as linking variation in a particular gene to climate or ecology. Recent work shows that PCA is expected to reveal axes of genetic variation that are orthogonal [41] and that multiple interpretations may be consistent with a given PCA representation of genetic variation.

20.3.3 Genetic Structure of Human Populations

An important and influential resource for studying human genetic variation has been the Human Genome Diversity Panel. Spearheaded by Howard Cann, Luca Cavalli-Sforza, and Jim Weber [4] (see Box 20.1), the HGDP is a collection of immortalized lymphoblastoid cell lines of over 1,000 individuals from 51 populations. By creating a renewable resource of DNA, the panel has afforded deep inferences on human evolutionary history, especially genomic signatures of

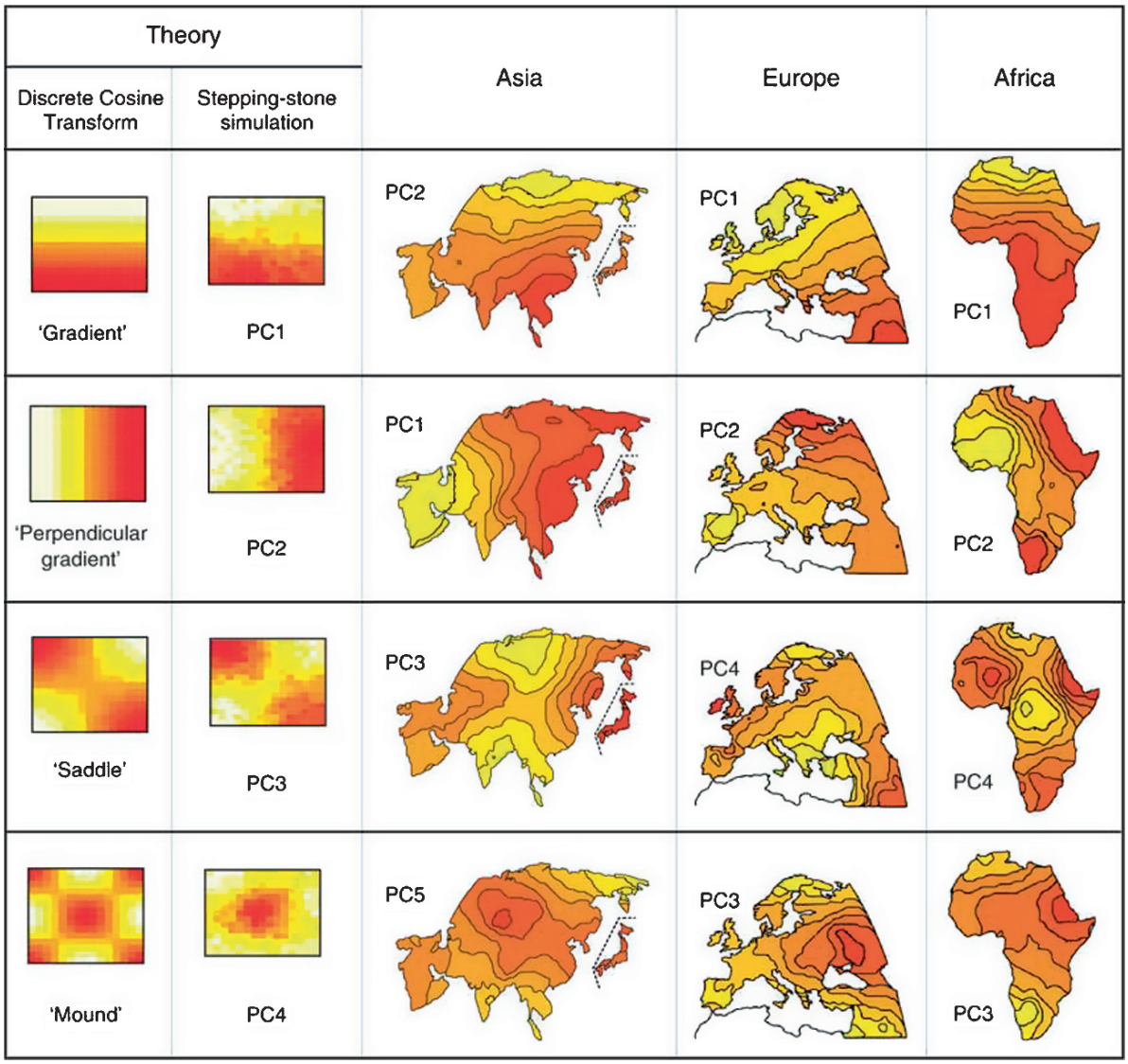

Fig. 20.5 Comparing synthetic maps from [34] with theoretical and empirical expectations. Menozzi et al. [34] performed principal component analyses on frequencies of 38 genes in various populations; the last three columns of this figure depict their original results. In the panel displaying PC1 in Europe, for example, the frequencies of certain alleles decrease (shown by more yellow colors) away from the Middle East; authors state that this pattern parallels the arrival of agriculture, which originated in the Middle East and then spread northward to Europe [6]. The Menozzi et al. [34] figures have been arranged to correspond with the shapes seen in the first two columns, which are based on theoretical and simulation results from [41]. Novembre and Stephens [41] simulated populations evenly-spaced in two-dimensional habitats with homogeneous migration rates aross time and space. Their PCA of these simulations found large-scale orthogonal gradients and "saddle" and "mound" patterns (see first column) when visualizing principal components even under this homogenous migration scheme. The second column displays PCA results from their simulations. The first column shows common structures seen in covariance matrices of population allele frequencies where genetic similarity decreases with geographic distance in a two-dimensional habitat, known as a stepping stone model. The regularity with which they observe these patterns runs counter to Menozzi et al.'s [34] claim that their PCA results are indicative of specific migration events.

historical relationships between populations, by providing a means to genotype (and ultimately sequence) genomes from diverse human populations. The first study of genetic variation in the HGDP scored polymorphism across 377 autosomal microsatellite loci in the panel [53] and recapitulated Lewontin's [29] result

that the vast majority of human population genetic variation is found within local populations. However, the study also demonstrated individuals could be assigned to their continent of origin, and in some cases their population of origin using the model-based clustering algorithm STRUCTURE [49]. The authors reported "it was only in the accumulation of small allele-frequency differences across many loci that population structure was identified."

In a follow-up study (see Fig. 20.6), Rosenberg et al. [52] genotyped 993 total markers in the HGDP and demonstrated increased resolution of population structure as a result of increasing the amount of genetic data used. In particular, when the method is asked to identify two clusters (i.e., $K=2$), the authors found that STRUCTURE differentiates between indigenous American (purple cluster) and African (orange cluster) populations, with other populations having a gradient of membership in the African cluster that drops off with geographic distance from Africa. As the number of clusters K used in the STRUCTURE analysis increased, correlations in genotype data within continents of origin allowed Eurasia, East Asia, and Oceania to be identified as separate clusters as well. The structure that is identified is that of differences between continents, with a few notable exceptions. For example, the orange Africa cluster membership in the Mozabites reflects the gene flow this Middle Eastern population has had with Africa, due to the samples' location in North Africa. Similarly, membership in the blue Eurasian cluster in the Maya reflects gene flow with Europe during colonization that this American population experienced to a greater extent than other American populations in the HGDP.

The genotyping of 650,000 SNPs in these populations [31] allowed the detection of individual ancestry and population substructure with very high resolution within continents as well as across them (more examples of analyses with dense SNP maps will be discussed in Sect. 20.4). Li et al. [31] were further able to examine the distribution of ancestral alleles (nucleotides observed in chimpanzee) in HGDP populations by genotyping two chimpanzee samples at the same markers. The ancestral allele-frequency spectrum across loci can yield clues to the history of individual populations, because we expect populations with a small effective size and/or populations that have experienced a bottleneck to have more pronounced genetic drift, which can result in a relatively rapid increase in derived allele frequencies compared to populations with larger effective sizes or populations that have experienced expansion.

20.3.4 A Haplotype Map of the Human Genome

A comprehensive search for genetic causes of common diseases, such as type II diabetes or macular degeneration, requires examining genetic differences between a large number of affected individuals (i.e., cases) and matched controls. Key to facilitating this effort is knowledge about patterns of linkage disequilibrium (LD) or nonrandom association among SNPs in the genome. Such correlations between causal mutations and their haplotypes have long been used in human genetic research of disease (e.g., in studies of the HLA region, and in the identification of causes of Mendelian disorders such as cystic fibrosis).

The International HapMap Project was formally initiated in October 2002 as a means of systematically describing patterns of linkage disequilibrium in the human genome in order to catalyze medical genetic research into the heritable basis of common disease. It also represented the beginning of a paradigm shift in both the amount of data and types of questions that could be answered by human population geneticists. The stated goal of the project was to "determine common patterns of DNA sequence variation in the human genome" [21], with the goal of typing over one million SNPs in 270 individuals including 60 *trios* (samples of two parents and one of their biological children). The project has far surpassed that goal, as seen in Box 20.1.

The HapMap data provide insight into LD patterns across three populations. Chief among the concepts developed around the project is the notion of *tag SNPs* or representative SNPs in a region that can serve as "proxies" for other SNPs. That is, using tag SNPs means genetic variation can be efficiently queried for association with disease without genotyping every SNP in a given chromosomal region (therefore drastically reducing the cost of carrying out a genome-wide association study).

Tag selection methods exploit redundancy among SNPs; however, since the HapMap initially sampled only three populations, an issue for association studies is whether tag SNPs chosen from the HapMap dataset

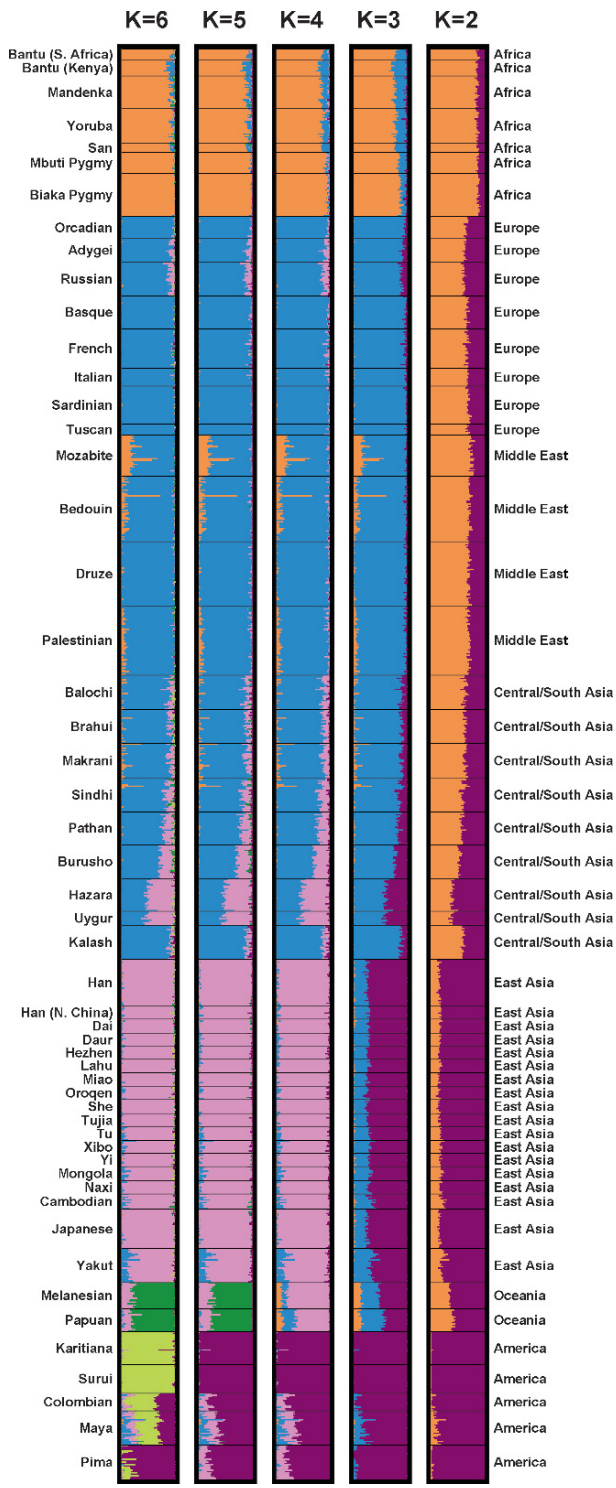

Fig. 20.6 Inferred population structure based on 1,048 individuals and 993 markers. Each individual is represented by a thin line partitioned into *K* colored segments that represent the individuals estimated membership fractions in *K* clusters. Black lines separate populations, whose names are to the left of the figure, with continent listed on the right of the figure. The value of *K* indicates how many clusters STRUCTURE was assuming existed in the dataset for a particular set of runs for the method. From [52]

20

Box 20.1 Examples of publicly available human population genetic datasets

Description of three major datasets used by the human population genetics community. These are not the only large datasets available for research, but illustrate how much data are being generated to better understand human genetic variation, genetic signatures of human history, and the genetic underpinnings of disease.

Dataset Name	Initial reports of data	Amount of data generated
Human Genome Diversity Project	[4] (2002)	lymphoblastoid cell lines from 1,064 individuals in 51 populations
	[53] (2002)	377 autosomal microsatellites
	[52] (2005)	993 markers (microsatellites and insertion/deletion polymorphisms)
	[31] (2008)	650,000 SNPs in 938 of these individuals
International HapMap Project	[21] (2003 paper)	270 people across three populations (30 trios from Yoruba people of Ibadan, Nigeria; 45 unrelated individuals from Tokyo, Japan; 45 unrelated individuals from Beijing, China; 30 United States trios with northern and western European ancestry)
	[22] (2005 paper)	1 million SNPs in these individuals (1 SNP per 5 kilobases)
	[23] (2007 paper)	An additional 2.1 million SNPS (1 SNP per kilobase)
	HapMap Phase III, draft 2 reported online in January 2009	an additional 1.5 million SNPs and an increase to 1184 individuals (populations added: Chinese from Denver; Gujurati from Houston; Luhya from Webuye, Kenya; Mexican ancestry from Los Angeles; Maasai from Kinyawa, Kenya; Toscans from Italy; African ancestry from Southwest USA)
1,000 Genomes Project	www.1000genomes.org	Sequencing the genomes of approximately 2,000 people from around the world

adequately capture patterns of variation in other populations. Conrad et al. [10] showed that the portability of tag SNPs from HapMap to HGDP populations was quite good within large geographic regions such as continents. These dense genotype data reveal other important patterns resulting from continental population structure, such as an increase in LD with distance from Africa, reflecting that African lineages have smaller preserved blocks of LD due to increased time for recombination events to break up correlations (also seen in the HGDP by Conrad et al. [10]).

Data from the initial HapMap project do not enable much inference about evolutionary relationships between populations, so the genotyping of individuals from additional populations has become a priority in human population genetics. As the cost of SNP genotyping lowers, studies allow for across- and within-continental pictures of population structure to emerge. Dense genotype data from multiple populations allow the inference of both continental differentiation and the fine-scale study of within-region relationships among individuals. It is these finer-scale patterns that we will explore in the next section.

20.4 The Genetic Structure of Human Populations Within Continents and Countries

Large-scale human population genetic studies like the Human Genome Diversity Panel and International HapMap Project discussed in Sect. 20.3 initially had to choose between sampling densely geographically and

sampling densely genomically. In just the last 2–3 years, improvements in genotyping technologies have allowed studies to report analyses of hundreds of thousands of SNPs genotyped in individuals from many populations. These datasets reveal the genetic signatures of historical events like migrations and conquests in more detail than geneticists could have hoped for when the field began. Here we explore how the history of Eurasia, the Americas, and Africa has shaped patterns of genetic variation of its inhabitants. (Note: the reason we have chosen to start with a discussion of Eurasia is simply that these are the populations that, to date, have been studied most intensively genetically. We believe the next few years will bring fine-scale studies of population structure across global human populations and strongly advocate these studies be undertaken, particularly in parts of the world currently understudied.)

20.4.1 Genetic Differentiation in Eurasia

Instead of grouping individuals into populations a priori (as early population genetic analyses necessitated), today we can let the data speak for themselves and tell us which individuals naturally cluster together based on genetic distance. A convenient means of accomplishing this is undertaking PCA on individual genotype scores (i.e., the "0," "1," "2" matrices mentioned above). Often when this is done, individuals from the same population tend to cluster together in PCA space. In fact, many PCA plots of globally distributed population structure seem to resemble geographical maps of the world with individuals from contiguous geographic regions clustering near each other in PCA space and revealing a close relationship between geographic distance and genetic differentiation (see Figs. 6.4 and 20.7).

Specifically, multicontinental studies of genomic diversity often find a clustering of populations according to their respective continents in the first few principal components, followed by differentiation between regions within continents. When sampling is dense, principal components can often serve as proxies for geographic axes [41, 42], separating Northern from Southern populations or Eastern from Western. For example, multiple studies observe North-to-South clines in European genetic differentiation, as seen using haplotype diversity in Fig. 20.8. Principal components also reveal evidence of genetic admixture between populations that can often be interpreted based on historical events such as colonization or slave trade; these signatures of admixture are discussed at length in Sect. 20.6.

As Fig. 6.4 shows, the European geographic map is an efficient summary of the first two principal components – or, put another way, dimensions – of European genetic variation. Novembre et al. [42] and Heath et al. [16] also showed that individual genotypes, despite low differentiation among populations in Europe as measured by F_{ST} (see Table 20.1), can be used to predict an individual's geographic origin within a few hundred kilometers (when that individual's geographic origin is representative of their ancestry). Likewise, several recent studies using high-density genotyping arrays have demonstrated the ability to reliably distinguish individuals of Ashkenazi Jewish ancestry from those without Ashkenazi Jewish ancestry in both European and European-American populations [36, 47, 64]. The ability to detect fine-scale geographic structure will only improve as whole-genome sequencing data become available; the studies discussed here are based on SNPs whose minor allele frequency is usually greater than 5%. Newer sequencing technologies will call lower frequency alleles more accurately, and low frequency alleles likely reflect recent mutations and may account for much differentiation between neighboring populations.

The larger mean heterozygosity and smaller mean linkage disequilibrium observed in Southern Europe compared to Northern Europen might be explained by an expansion in Europe from the South to the North [28]. South-to-North movement in Europe occurred during the first Paleolithic settlement of the continent by anatomically modern humans, and during the Neolothic expansion [2]. Thus we might expect to see a genetic signature of such movement, although another important controversy in historical anthropology is whether technologies such as agriculture traveled via demic diffusion (the movement of people and their genes) or cultural diffusion (the spread of technologies without a concomitant genetic signature) (see, for example, [46]).

Genetic variation in specific countries has been studied as well, such as that of Finland [24]. Studying a specific population may give insights into inbreeding or homozygosity patterns, as well as the genetic signatures of founder effects or multiple waves of migration

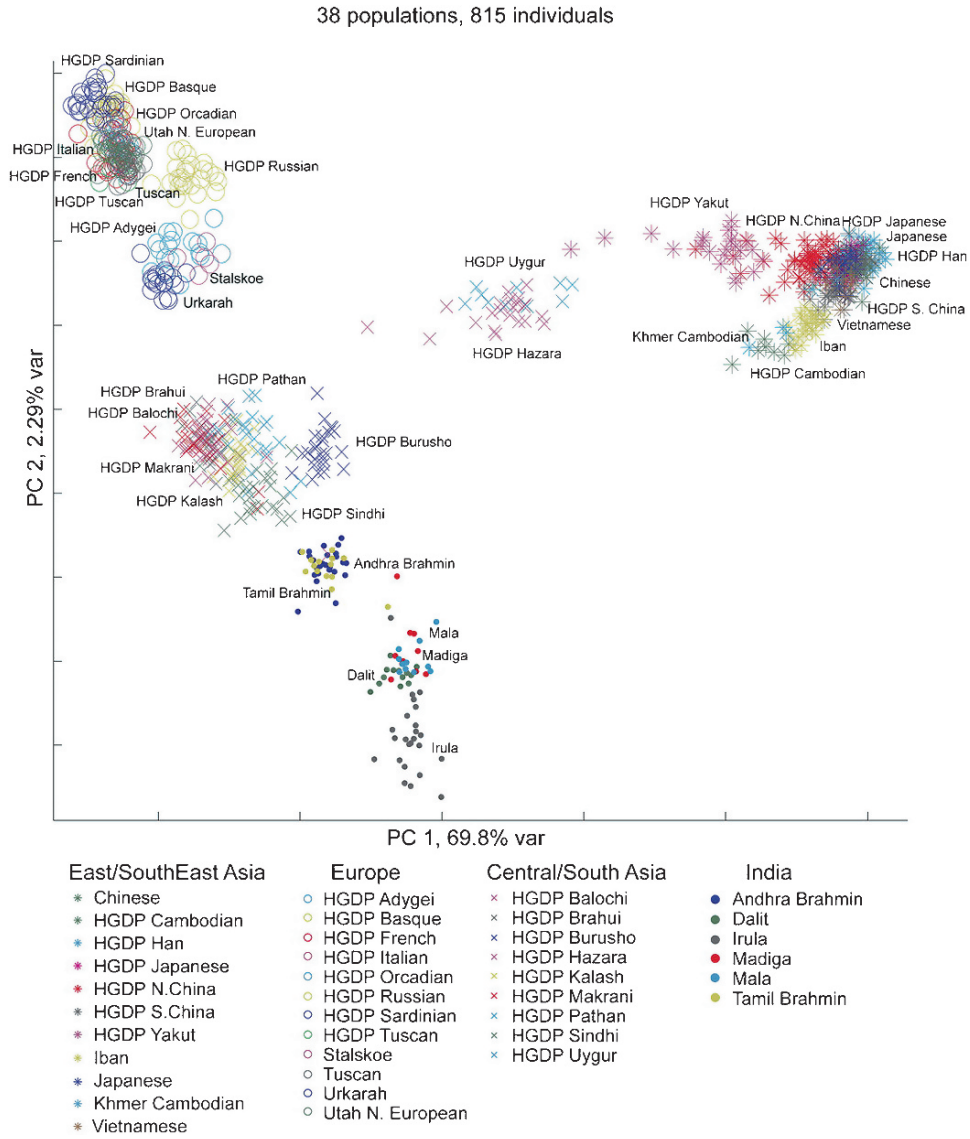

Fig. 20.7 PCA for 815 Eurasian individuals using nearly 50,000 SNPs. Individuals group closely with others in their self-identified population of origin, and the populations are differentiated in a way that mirrors a geographic map of Eurasia, much like the close relationship seen in Fig. 6.4 between genes and geography in Europe. HGDP denotes populations from the Human Genome Diversity Panel [4]. From [71]

in, for example, linkage disequilibrium patterns or admixture blocks. Jakkula et al. [24] found genetic signatures supporting multiple historical bottlenecks resulting from consecutive founder effects, in keeping with Finland's history of two major migration waves (a western one from 4,000 years ago, and a southern and western one from 2,000 years ago). A study of 7,003 Japanese individuals also shows that local regions in Honshu Island, the largest island of Japan, are genetically differentiated despite frequent migration within Japan during the last century [73].

Single-population studies are of great interest, as populations that experienced bottlenecks and subsequent low levels of immigrations (like the Finnish, Askenazi Jewish, or Icelandic peoples) may see a rise in Mendelian disease frequencies, and it has been proposed that gene mapping for complex traits may be easier in these populations than others. The Finns, for

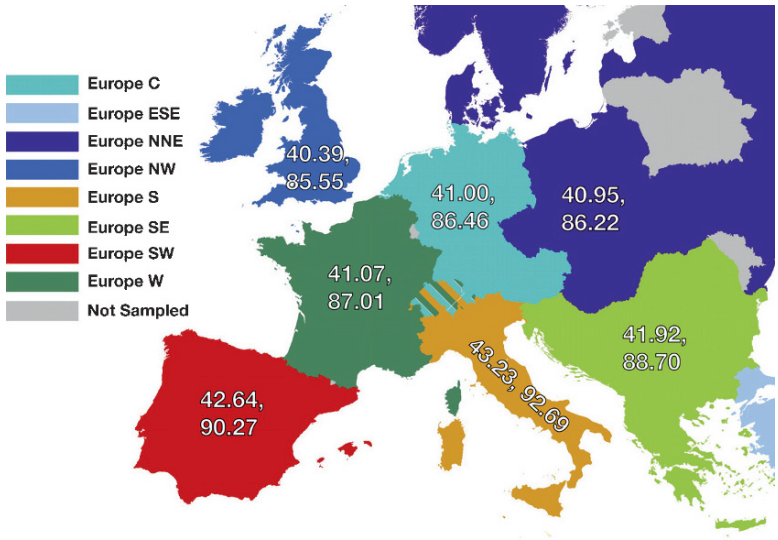

Fig. 20.8 Haplotype diversity within Europe. Two numbers are shown in each region; the first shows the mean number of distinct haplotypes in a region in a genomic window containing 10 SNPs, the second reflects haplotype diversity in a 25-SNP window. The authors find that haplotype diversity, as reflected by the numbers displayed, is higher in southern Europeans countries than in northern European countries, indicating that southern populations have larger effective sizes than northern ones and that the original peopling of Europe happened with migrations from the South to the North. From [1]

example, exhibit a substantial degree of homozygosity due to their population history, although not the amount of homozygosity seen in cultures with consanguineous marriages [24]. In such a population, the tagging of recessive variants in complex disorders may be done with common SNPs; indeed, the use of homozygous segments to identify rare alleles associated with Mendelian mutations has been successful in Finland (Meckel syndrome) [61]. Interestingly, Ashkenazi Jewish populations exhibit very similar patterns of linkage disequilibrium (and, in fact, *less* LD) than the CEPH European populations genotyped as part of HapMap, but approximately 20% higher levels of homozygosity [43]. Studies such as these address the power of genome-wide association datasets to demonstrate history-related stratification even within apparently homogeneous genetic populations, and shed light on the importance of rare variants in fine-scale genomic studies.

20.4.2 Genetic Variation in Native American Populations

During an expansion from a parent population via serial bottlenecks, sometimes called a "serial founder effect," linkage disequilibrium will increase and heterozygosity will decrease with distance from the origin of the expansion in the absence of selection [50]. Large linkage disequilibrium blocks were observed in the five Native American populations genotyped in the HGDP [10]. However, additional population samples from the Americas are important to help us understand the peopling of the Americas via the Bering Strait and what signature colonization, in this case by Europe, might have on genetic data. Wang et al. [66] studied 24 newly sampled populations of Native Americans from Canada, Meso- and South America.

The study found lower heterozygosity at microsatellite loci in indigenous Americans than in the non-American HGDP populations, and also observed a greater variance in heterozygosity among American populations. This could be a signature of a quick initial peopling of the Americas, followed by subsequent isolation of populations in the continent. A rapid coastal migration followed by a slower inland migration was supported by a higher level of genetic diversity in western South America compared to eastern South America. Wang et al. [66] also tested for correlations in differences between linguistic stocks or families and genetic distance between populations, finding that genetic distance and linguistic distance are more highly

correlated within linguistic families than between families.

The study found support for an East Asian origin for Native American genetic variation, with relatively higher similarity to East Asian genetic variation in North Americans than South Americans, and also observed a private allele in the Native American samples. Wang et al. [66] showcase how a variety of hypotheses regarding demographic history can be tested with genetic data, when aligned with linguistic and archaeological data.

20.4.3 The Genetic Structure of African Populations

Africa and African populations play an important role in human evolutionary history given the African origin for anatomically modern humans and the amount of our genomic variation shaped by the out-of-Africa bottleneck [11]. However, it is important to recognize that African populations have been evolving since the human diaspora. Tishkoff et al. [65] sampled 121 African populations at over 1,000 microsatellite loci to study the demographic history of Africans as inferred from genetic data.

The investigators identified 14 ancestral clusters in Africa; these clusters approximately correspond to linguistic families, self-identified ethnicities, and/or cultural practices such as hunting-and-gathering. There was also a great deal of mixed ancestry in most populations, a signature of recent migrations in the African continent.

Three hunter-gatherer populations in the study were among the five most genetically diverse populations in the African sample, and African and Middle Eastern populations were found to share a number of alleles not observed elsewhere. Within Africa, the most private alleles were seen in click-speaking populations.

The spatial distribution of heterozygosity was used to pinpoint the origin of the modern human migration within the African continent in the same manner as by Ramachandran et al. [50]. Tishkoff et al.'s [65] analysis places this origin in southwestern Africa near the border of Namibia and Angola, which corresponds to the current San homeland. This lends support to the San being a genetically ancient population, although perhaps their current geographic origin does not reflect their ancestors' geographic location 100,000 years ago. As geographic analyses become more refined, genetic variation appears to be more clinal than clustered [50]. This is because, at within-continental geographic distances, migration levels may be high and levels of admixture across populations will increase. We cannot study population genetics within continents without understanding recent genetic admixture, the subject of the next section.

20.5 Recent Genetic Admixture

The ultimate cause for population structure is nonrandom mating. For example, if individuals from geographically distant populations are less likely to mate than individuals from the same population, over time, discernible differences in allele frequencies will accumulate (as explained in Sect. 20.1). Compounding these differences across the genome provides power for reliably differentiating individuals from different populations even if, overall, the degree of population differentiation is low [31, 42]. On the other hand, migration facilitates gene flow. Recent global exploration and colonization have led to a rapid increase in gene flow among individuals from different continents. Their offspring are referred to as admixed and we can mathematically model chromosomes from admixed individuals as mosaics of segments derived from different ancestral populations. This section summarizes variation in several admixed populations, with the goal of illustrating the power genetic data have to shed light on a population's recent history.

20.5.1 Populations of the Americas

The population history and genetic structure among the Native American groups was surveyed in Sect. 20.4.2. The first significant wave of European influence arrived in the New World with Christopher Columbus' voyages of 1492–1504. During the sixteenth to nineteenth century, between 9.4 and 12 million Africans (mostly from West and Central Africa) were transported to the New World through the transatlantic slave trade. The arrival of Europeans and Africans in the New World gave rise to numerous

admixed populations. In this section, we focus on two such groups: African Americans and Hispanics.

According to the 2007 U.S. Census, 41 million U.S. residents self-identify as having some degree of direct African ancestry (i.e., identify as "black" or African-American). Studies of genetic variation among African-Americans suggest that, on average, 80% of their genetic ancestry is West African, although individual ancestry proportions vary substantially as do the average ancestry proportions for different sampling localities within the United States [44]. Based on chromosomal block lengths, the admixing time between Europeans and Africans is estimated to have occurred 7–14 generations ago [14, 74]. At 25 years per generation, this places admixture as occurring between 175 and 350 years ago. Historical records indicate that the largest sources of the African slaves were the coastline in West and West Central Africa. However, locating the precise African ancestral populations for the African Americans has been challenging, in particular due to the lack of genetic data in geographically and ethnically diverse African populations. The recent study by Tishkoff et al. [65], discussed in Sect. 20.4.3, fills in this gap by genotyping over 2,000 Africans from 113 populations; in the near future, genetic data will likely be used to characterize admixture patterns within the African component of the African Americans.

Hispanics derive their ancestry from European, African, and Native American individuals. The term Hispanic describes populations that share a common language and cultural heritage, including Mexicans, Puerto Ricans, and Cubans, to name a few. However, these groups do not constitute a uniform ethnicity with a similar genetic background. At present, Hispanics represent the largest and fastest-growing minority population in the United States. Although genetic studies characterizing the population structure in the Hispanic population have been limited both in the marker density and in subgroup representation, evidence is mounting that individual ancestry proportions vary tremendously among subgroups that are identified as Hispanic. At a population level, Puerto Ricans have higher African ancestry compared to the Mexicans [56]; even within Mexico, the Native American ancestry proportions vary among States, ranging from 35% in Sonora in the North to 65% in Guerrero in center-Pacific [59].

The history of admixed populations is clearly discernible in patterns of genetic variation as summarized by PCA. Consider, for example, the GlaxoSmithKline POPRES sample [1, 37, 42] consisting of 3,875 individuals of varying ethnic backgrounds from over 80 countries genotyped on the Affymetrix 500 K platform. In Fig. 20.9, we reproduce key results from Nelson et al. [37] on PCA analysis of PopRes and the core HapMap populations (i.e., Yoruba from Ibadan, Nigeria, CEPH with ancestry from Northern Europe, Japanese from Tokyo, and Han Chinese from Beijing). This sample contains representation from several major continental

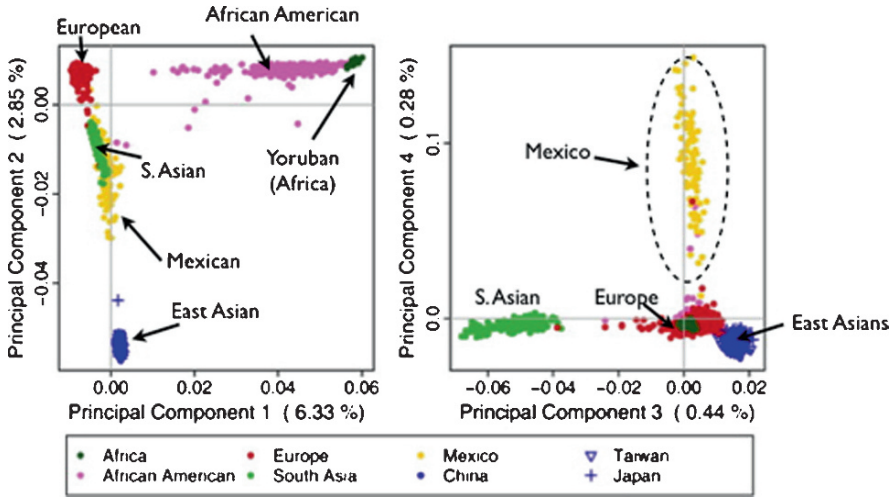

Fig. 20.9 Genetic structure in the PopRes data. Subject scores are colored by continental and/or ethnic origin (see legend). Percent of variation explained by each component is given in parentheses on each axis label. Reprinted from [37], with permission from Elsevier

populations as well as a large sample of African-Americans from the United States and Hispanics from Mexico. We note that the first principal component can be interpreted as an "African-American admixture" principal component (or, equivalently, Africa vs. Europe + Asia); principal component two corresponds to "East Asia vs. Europe"; principal component three corresponds to an "East Asia vs. South Asia" axis of variation; and principal component 4 (PC4) to a "Mexican admixture" axis. Importantly, individuals of admixed ancestry appear on the PCA map as in between the centroids of their putative ancestral populations.

An important feature of this analysis is that along the "Mexican admixture" PC, individuals show varying degrees of admixture between Europeans and a presumably "unsampled" population which likely corresponds to Native Americans. Analyzing just the European, East Asian, and Mexican samples (Fig. 20.10), we find that PC1 is an East Asia vs. Europe principal component and PC2 separates the East Asian sample from the (unsampled) Native American sample (so that the least admixed individuals are furthest away from the East Asian samples along PC2). This suggests that there is substantial genetic differentiation between East Asian and Native American populations so that the former is likely a poor proxy for the later (and vice versa).

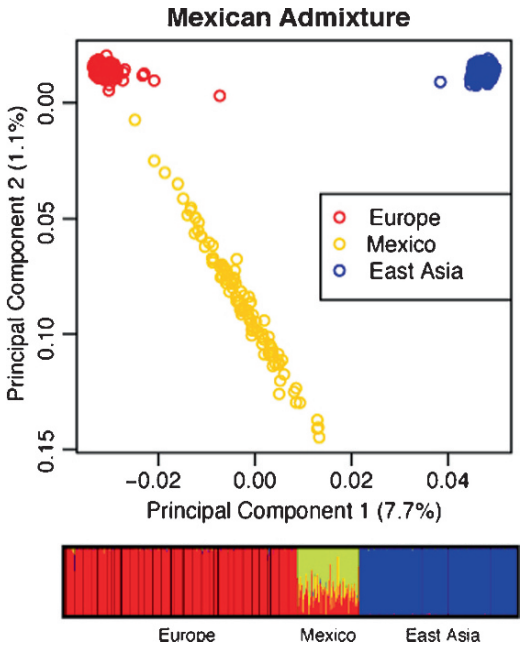

Fig. 20.10 Principal component and STRUCTURE analysis of the PopRes Mexicans from Guadalajara, Europeans from throughout Europe, and East Asians from Japan and China. From [1]

20.5.2 Admixture Around the World

Genetic admixture is a worldwide phenomenon and is not limited to the North America. An example of an admixed population in the Eurasian continent is the Uyghur population living in the Xinjiang province in western China. Because of its proximity to the Silk Road – the historically important trade route connecting East Asia with the West, Central Asia, and the Mediterranean world – the Uyghur population derives ancestry from East Asian, European and the Middle East ancestral populations [31]. Using high-density SNP markers, a recent genetic study estimated approximately equal ancestral contributions from the European and East Asian populations to the Uyghur population [72]. In central Algeria in Northern Africa, the Mozabites originated from a Berber ethnic group in the Middle East that had close cultural contact with diverse African and European populations. Not surprisingly, SNP analysis of the Mozabite individuals in HGDP detects substantial European, African, and Middle Eastern ancestry (see Fig. 20.6).

Owing to the availability of high-throughput genotyping technologies, our ability to detect finer-scale population structure and more ancient admixture has dramatically increased during the past few years. The 600,000 SNP markers typed in the HGDP samples revealed genetic structure that was not detected previously using 300 microsatellites: most prominently, the detection of the Middle Eastern populations as a separate cluster [31, 54]. Moreover, the SNP data suggest that the impact of genetic mixture is more profound than has been previously implicated. Many Middle Eastern individuals appear admixed, perhaps because of the continuous migration in this area. Finally, in the analysis of the European populations, while the genetic structure matches geography, there is undeniable continuity between populations. Thus, it is more appropriate to consider the structure within the European continent as continuous clines rather than as a discrete cluster. In summary, the increasing amount of the genetic data will allow us to characterize population structure at a higher resolution within continents.

20.6 Quantiative Modeling of Human Genomic Diversity

The primarily qualitative studies described above have given great insight into both the global and local patterns of human genetic history. Quantitative models can offer additional insights; for example, we can use quantitative models to ask how severe particular bottlenecks were, or exactly when populations diverged.

Much quantitative modeling relies on resequencing data. SNP genotyping chips provide a genome-wide picture of variation at low cost, but they can be difficult or impossible to use for quantitative inference. Primarily, this is because the sites assayed on a chip are not a random sample of the genome; they are typically chosen because they are known to be polymorphic in some smaller "discovery" population. This ascertainment process biases the resulting data [9], particularly the allele frequency spectrum. Although this bias can, in some cases, be controlled for (e.g., [25]), it is unfeasible in general [38]. In Fig. 20.11, for example, we report the joint and marginal allele frequency spectra (AFS) for the PopRes and HapMap populations. We note that the joint allele frequency spectra reproduce qualitative patterns of the PCA analysis in Fig. 20.9, such as a stronger correlation in allele frequency and lower F_{ST} for closely related populations (e.g., East Asia from PopRes and JPT+CHB from HapMap or Europe from PopRes and CEPH from HapMap). Nonetheless, the marginal (or one dimensional) spectra are quite skewed toward intermediate frequency alleles as a result of the ascertainment bias in the Affymetrix 500K chip, which favored middle frequency variants for use in GWAS. The goal of this section is to describe how (unbiased) AFS data can be used for quantitative demographic inference of both demographic history and selection.

20.6.1 Demographic History

The demographic history of a set of populations encompasses the order and timing of any divergence or admixture events, as well as changes in population sizes and rates of gene flow over time. In principle, the greatest statistical power for inferring such a model from genetic data would arise from calculating the full likelihood of the data given the model [19]. However, at present such calculations are very difficult at the genomic scale. Thus many methods for inferring demographic events rely on modeling summaries of the data. The allele frequency spectrum is a particularly popular summary. As seen in Fig. 20.1, the frequency spectrum encodes substantial information about demographic history. (Although Myers et al. [35] have shown that it does not alone uniquely determine demographic history.) For example, the center column of part D of Fig. 20.1 shows that the one-dimensional (i.e., single population) allele frequency spectrum is skewed toward rare alleles in situations of population growth, while the right two columns show that asymmetric population sizes yield an asymmetric AFS.

An early study by Marth et al. [33] introduced an analytic method for calculating the allele frequency spectrum for a single population with piecewise constant population size. Using this method to fit models for several global populations revealed signatures of ancient population growth in African-Americans (presumably occurring in their African ancestors), and bottlenecks in the history of both European-American and East-Asian populations. These historical events have been well supported by subsequent genetic studies.

Considering the joint history of multiple populations substantially complicates the models, as divergence and gene flow must be incorporated. Consequently, the computational methods become more demanding. In a ground-breaking study, Schaffner et al. [58] used extensive coalescent simulations to replicate both summaries of the allele frequency spectrum and patterns of LD for West African, European, and East Asian populations, developing the first quantitative model for their joint genetic history. The computationally intensive nature of their analysis, however, precluded them from statistically assessing the confidence of their inferences or testing multiple models.

Recent theoretical and computational advances in the simulation of the frequency spectrum with diffusion theory have enabled more comprehensive statistical characterization of such models (Gutenkunst et al., in press). Figure 20.12 shows an illustrative model of human history, and the resulting expected frequency spectra. Within this model, parameters such as divergence times, migration rates, admixture proportions, and bottleneck sizes have been quantitatively inferred.

Models have also reached further back in time, before the emergence of modern humans. In particular, a recent analysis by Fagundes et al. [13] compared several models of early human history, including the possibility of interbreeding with other hominids. The analysis showed

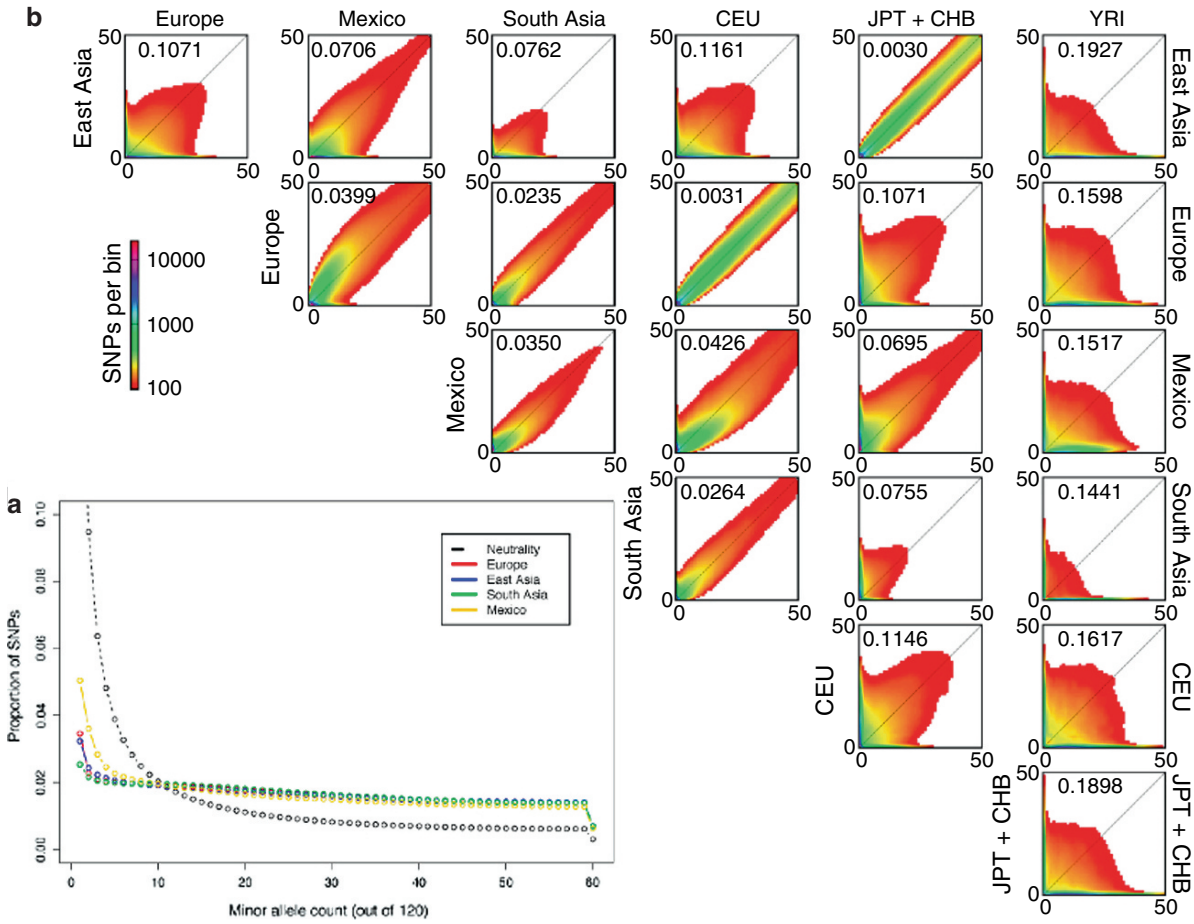

Fig. 20.11 Frequency spectra of the POPRES populations and HapMap samples (*CEU*: CEPH Utah residents with ancestry from northern and western Europe; *CHB* Han Chinese in Beijing, China; *JPT* Japanese in Tokyo, Japan; *YRI* Yoruba in Ibadan, Nigeria). (**a**) Minor Allele Frequency Spectra for the four sub-continental populations. The spectrum expected under neutrality is also shown in black. (**b**) Two-dimensional joint frequency spectra for each pairwise sub-continental population comparison. Colors represent the number of SNPs within each bin. Entries in the spectra containing less than 100 SNPs are shown in white. Autosomal estimates of F_{ST} for each comparison are shown in the upper left hand corner of each figure. From [1]

that a model in which modern humans simply replaced other hominids was best supported by the data.

20.6.2 Quantitative Models of Selection

Quantitative demographic models also play an important role in the search for evidence of selection acting on the genome. In particular, scans for selection seek genomic regions with unusual patterns of genetic variation, and demographic models define the null expectation of how unusual a region must be to be statistically significant when testing the hypothesis that it is under selection [40].

Beyond the search for unusual patterns of genetic variation, quantitative modeling has also given insight into the general signatures left by selection on the human genome. For example, an early analysis of the allele frequency spectrum for different classes of polymorphism revealed strong negative selection on mutations that change amino acid sequence. Furthermore, within those mutations computational algorithms such as Polyphen can predict which changes are most damaging [68].

More recently, the distribution of the selective effects of new mutations has been inferred from the allele frequency spectrum [3]. The selection coefficient s of a mutation is defined as the relative reproductive advantage conferred upon carriers of the mutation. As seen in

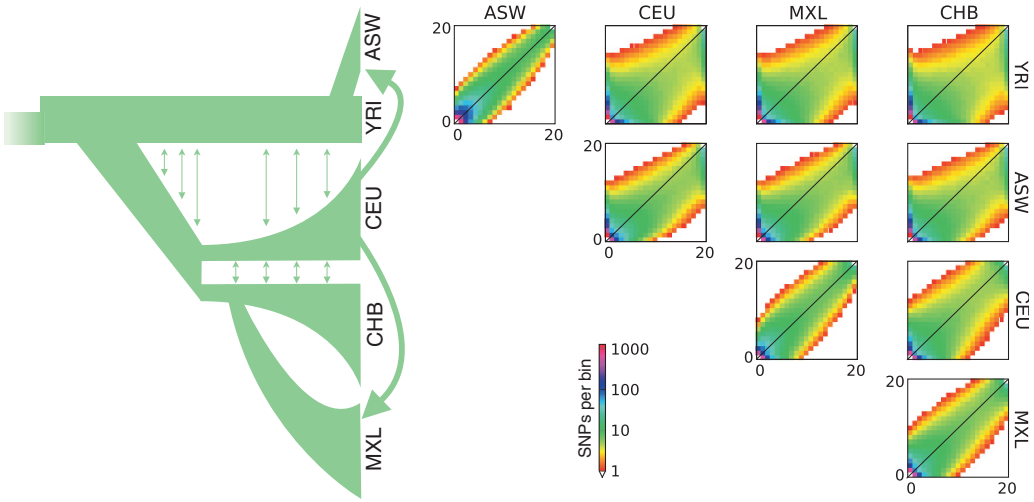

Fig. 20.12 Illustrative model of human expansion out of Africa and across the globe. The model includes African-American (*ASW*), West African (*YRI*), European (*CEU*), East Asian (*CHB*), and Mexican (*MXL*) populations. Using quantitative estimates for divergence times, population sizes, migration rates, and admixture proportions, the expected frequency spectrum under the model can be calculated using either diffusion or coalescent theory. Similar to Fig. 20.1.d., the resulting marginal spectra are shown for each pair of populations. Each spectrum shows distinct signatures of genetic history. For example, the recent European admixture into African-American and Mexican populations results in very highly correlated allele frequencies between populations pairs CEU-ASW (2nd row, 1st column) and CEU-MXL (3rd row, 1st column). Further, the Out-of-Africa bottleneck means that 2D spectra between African and non-African populations are asymmetric. When observed in real data, it is these sorts of signatures that guide quantitative modeling of human history

Fig. 20.13, the frequency spectra of synonymous and nonsynonymous variants differ dramatically. After correcting for demographic history using the synonymous mutations, it was found that the distribution of negative selection coefficients on newly arising amino-acid changing mutations possesses a very long tail. Roughly a third of amino acid substitutions are nearly-neutral ($|s|<0.01\%$), another third are moderately deleterious ($0.01\%<|s|<0.1\%$), and nearly all the remainder are highly deleterious or lethal ($|s|>1\%$). Knowledge of this distribution lets one calculate that very few of the fixed differences between human and chimp are selectively deleterious so that most are neutral or nearly, and that 10–20% of them result from positive selection. As the flood of data from the next generation of sequencing endeavors becomes available (e.g., the 1,000 Genomes Project and associated enterprises), we expect these preliminary estimates to be further refined, along with a quantitative understanding of human demographic history.

Acknowledgments We thank Dr. Sean Myles for helpful comments on an earlier draft of this chapter. Research is supported by GM073059 (to HT) and CDB was supported by NIH R01GM83606)

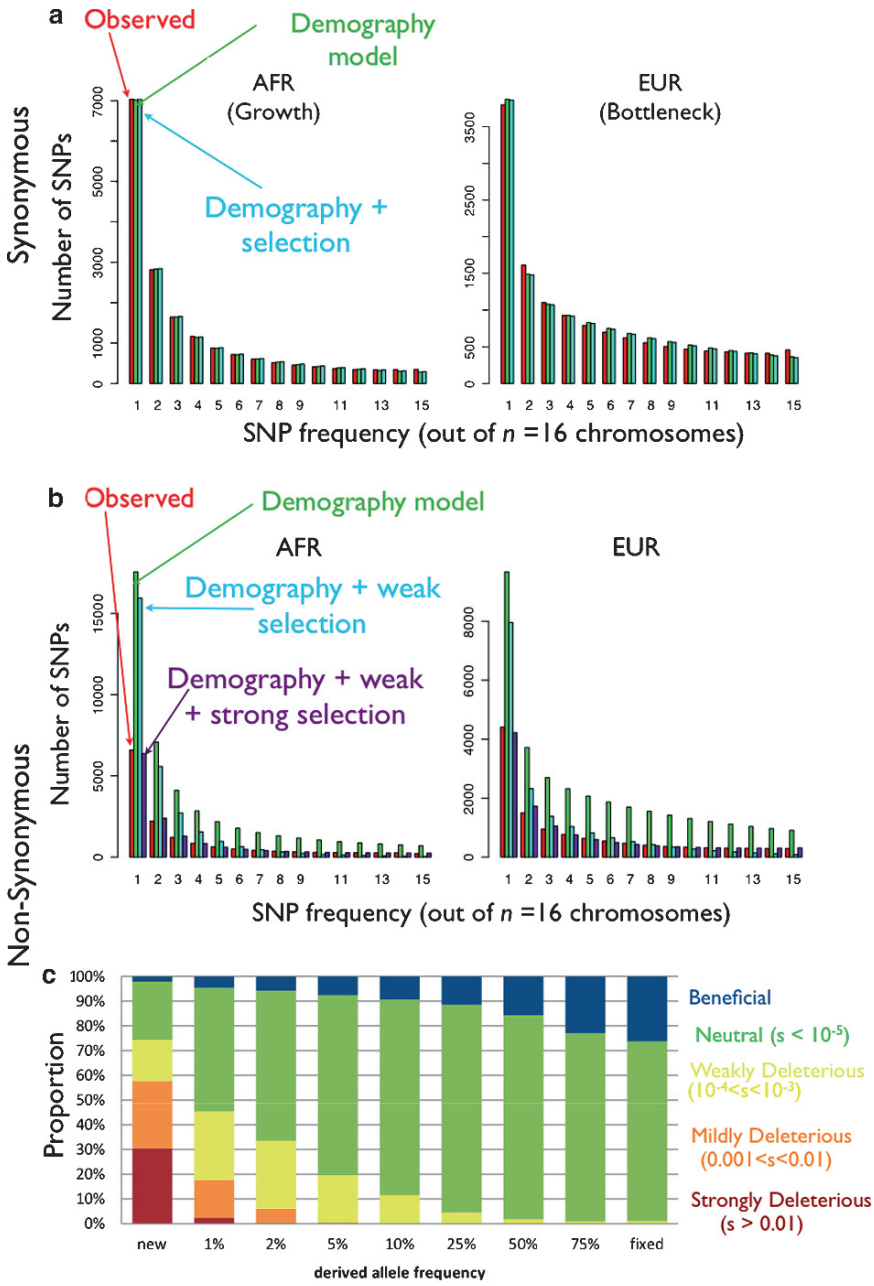

Fig. 20.13 Analysis of site-frequency spectra from over 30,000 coding SNPs found by resequencing of 11,000 genes in 20 European-Americans (*EUR*) and 15 African-Americans (*AFR*) yields estimates of demographic model and distribution of fitness effects of newly arising mutations and SNPs [3]. (**a**) Comparison of observed and predicted SFS for synonymous sites. Predictions are from two different types of models: a demographic model with growth in the AFR and a bottleneck in EUR (*green*) and for a model with weak negative selection on silent sites (*blue*). (**b**) Analogous comparison for nonsynonymous SNPs (nsSNPs) demonstrates that strong purifying selection, weak negative selection, and demographic history are all needed to accurately model the observed distribution of nsSNPs. (**c**) Estimated distribution of fitness effects for newly arising mutations in the human genome as well as SNPs at different population frequencies

References

1. Auton A, Bryc K, Boyko A, Lohmueller K, Novembre J, Reynolds A, Indap A, Wright M, Degenhardt J, Gutenkunst R, King K, Nelson M, Bustamante CD (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res 19:795–803
2. Belle EM, Landry PA, Barbujani G (2006) Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. Proc Biol Sci 273:1595–1602
3. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4:e1000083
4. Cann HM et al (2002) A human genome diversity cell line panel. Science 296:261–262
5. Cavalli-Sforza LL, Piazza A (1975) Analysis of evolution: evolutionary rates, independence and treeness. Theor Popul Biol 8:127–165
6. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ
7. Cavalli-Sforza LL, Menozzi P, Piazza A (1996) The history and geography of human genes. Princeton University Press, Princeton, NJ Abridged Paperback edition
8. Cavalli-Sforza LL, Menozzi P, Piazza A, Mountain J (1998) Reconstruction of human evolution; bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 85:6002–6006
9. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502
10. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38(11):1251–1260
11. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK (2009) The role of geography in human adaptation. PLoS Genetics 5: e1000500
12. Edwards AWF (2003) Human genetic diversity: Lewontin's fallacy. Bioessays 25:798–801
13. Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L (2007) Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA 104(45):17614–17619
14. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4):1567–1587
15. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, Berlin
16. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, Foretova L, Georges M, Janout V, Kabesch M, Krokan HE, Elvestad MB, Lissowska J, Mates D, Rudnai P, Skorpen F, Schreiber S, Soria JM, Syvänen A-C, Meneton P, Herçberg S, Galan P, Szeszenia-Dabrowska N, Zaridze D, Génin E, Cardon LR, Lathrop M (2008) Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 16:1413–1429
17. Hernandez RD, Williamson SH, Zhu L, Bustamante CD (2007) Context dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol Biol Evol 24(10):2196–2202
18. Hernandez RD, Williamson SH, Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of selection. Mol Biol Evol 24(8): 1792–1800
19. Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence times, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167:747–760
20. The Human Genome. Nature 2001;409:following p 812. (series of articles in Nature on the draft genome sequence)
21. The International HapMap Consortium (2003) The International HapMap project. Nature 426:789–796
22. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
23. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861
24. Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, Pedersen NL, deFaire U, Järvelin M-R, Saharinen J, Freimer N, Ripatti S, Purcell S, Collins A, Daly MJ, Palotie A, Peltonen L (2008) The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet 83:787–794
25. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet 39:1251–1255
26. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR (2009) Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci USA 106(10):3871–3876
27. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048
28. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabraci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann H-E, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18:1241–1248
29. Lewontin RC (1972) The apportionment of human diversity. In: Dobzhansky T, Hecht MK, Steere WC (eds) Evolutionary biology 6. Appleton-Century-Crofts, New York, pp 381–398
30. Lewontin RC (1974) The genetic basis of evolutionary change. Columbia University Press, New York
31. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 391:1100–1104

20

32. Mardia KV, Kent JT, Bibby JM (1980) Multivariate analysis. Academic, London
33. Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signatures of differential demographic history in three large world populations. Genetics 166:351–372
34. Menozzi P, Piazza A, Cavalli-Sforza LL (1978) Synthetic maps of human gene frequencies in Europe. Science 201:786–792
35. Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum? Theor Popul Biol 73:342–348
36. Need AC, Kasperaviciute D, Cirulli ET, Goldstein DB (2009) A genome-wide genetic signature of Jewish ancestry perfectly separates individuals with and without full Jewish ancestry in a large random sample of European Americans. Genome Biol 10(1):R7
37. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai Eric H (2008) The population reference sample (POPRES): a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83(3): 347–358
38. Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168:2373–2382
39. Nielsen R, Hellmann I, Hubisz M, Bustamante MCD, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8(11):857–868
40. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Hu X, Boyko A, Indap A, Bustamante CD, Clark AG (2009) Darwinian and demographic forces affecting human protein coding genes. Genome Res 19:838–849
41. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40:646–649
42. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KA, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456:98–101
43. Olshen AB, Gold B, Lohmueller KE, Struewing JP, Satagopan J, Stefanov SA, Eskin E, Kirchhoff T, Lautenberger JA, Klein RJ, Friedman E, Norton L, Ellis NA, Viale A, Lee CS, Borgen PI, Clark AG, Offit K, Boyd J (2008) Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping. BMC Genet 9:14
44. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63(6):1839–1851
45. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19(5):826–837
46. Pinhasi R, Fort J, Ammerman AJ (2005) Tracing the origin and spread of agriculture in Europe. PloS Biol 3:e410
47. Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, Ruiz-Linares A, Groop L, Saetta AA, Korkolopoulou P, Seligsohn U, Waliszewska A, Schirmer C, Ardlie K, Ramos A, Nemesh J, Arbeitman L, Goldstein DB, Reich D, Hirschhorn JN (2008) Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 4(1):e236
48. Pritchard JK, Rosenberg NA (1998) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228
49. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
50. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (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 102:15942–15947
51. Ramachandran S, Rosenberg NA, Feldman MW, Wakeley J (2008) Population differentiation and migration: coalescence times in a two-sex island model for autosomal and X-linked loci. Theor Popul Biol 74:291–301
52. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1:e70
53. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385
54. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H-C, Szpiech AZ, Degnan JH, Wang K, Guerreiro R, Bras JM, Scymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003
55. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837
56. Salari K, Choudhry S, Tang H, Naqvi M, Lind D, Avila PC, Coyle NE, Ung N, Nazario S, Casal J, Torres-Palacios A, Clark S, Phong A, Gomez I, Matallana H, Pérez-Stable EJ, Shriver MD, Kwok PY, Sheppard D, Rodriguez-Cintron W, Risch NJ, Burchard EG, Ziv E (2005) Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. Genet Epidemiol 29(1):76–86
57. Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68(2):466–477
58. Schaffner SF (2004) The X chromosome in population genetics. Nat Rev Genet 5:43–51
59. Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, Contreras A, Balam-Ortiz E, del Bosque-Plata L, Velazquez-Fernandez D, Lara C, Goya R, Hernandez-Lemus E, Davila C, Barrientos E, March S, Jimenez-Sanchez G (2009) Analysis of genomic diversity in Mexican Mestizo populations to

develop genomic medicine in Mexico. Proc Natl Acad Sci USA 106(21):8611–8616
60. Sundquist A, Fratkin E, Do CB, Batzoglou S (2008) Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Res 18(4):676–682
61. Tallila J, Jakkula E, Peltonen L, Salonen R, Kestila M (2008) Identification of CC2D2A as a Meckel syndrome gene adds an important piece to the ciliopathy puzzle. Am J Hum Genet 82(6):1361–1367
62. Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. Am J Hum Genet 79(1):1–12
63. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 28(4):289–301
64. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L, Gregersen PK, Seldin MF (2008) Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4(1):e4
65. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044
66. Wang S, Lewis CM Jr, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C (2007) Genetic variation and population structure in Native Americans. PloS Genet 3:e185
67. Weir B (1996) Genetic data analysis II. Sinauer Press, Sunderland, MA
68. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R et al (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA 102:7882–7887
69. Wright S (1921) Systems of mating. I. The biometric relations between offspring and parent. Genetics 6:111–123
70. Wu B, Liu N, Zhao H (2006) PSMIX: an R package for population stratification inference via maximum likelihood method. BMC Bioinformatics 7:317
71. Xing J, Watkins WS, Witherspoon DJ, Zhang Y, Guthery SL, Thara R, Mowry BJ, Bulayeva K, Weiss RB, Jorde LB (2009) Fine-scaled human genetic structure revealed by SNP microarrays. Genome Res 19:815–825
72. Xu S, Jin L (2008) A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. Am J Hum Genet 83(3):322–336
73. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N (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
74. Zhu X, Zhang S, Tang H, Cooper R (2006) A classical likelihood based approach for admixture mapping using EM algorithm. Hum Genet 120(3):431–445
75. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (in press) Inferring the joint demographic history of multiple populations from multidimensional SNP data PLoS Genetics; arXiv:0909.0925