Perspectives on Human Population Structure at the Cusp of the Sequencing Era

John Novembre[1,*] and Sohini Ramachandran[2,*]

[1]Department of Ecology and Evolutionary Biology and Interdepartmental Program on Bioinformatics, University of California, Los Angeles, California 90403; email: jnovembre@ucla.edu

[2]Department of Ecology and Evolutionary Biology and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island 02912; email: sramachandran@brown.edu

Annu. Rev. Genomics Hum. Genet. 2011. 12:245–74

First published online as a Review in Advance on July 21, 2011

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev-genom-090810-183123

1527-8204/11/0922-0245$20.00

*Both authors contributed equally.

Keywords

single-nucleotide polymorphisms, genetic drift, migration, dispersal, genome, resequencing, high-throughput technology

Abstract

Human groups show structured levels of genetic similarity as a consequence of factors such as geographical subdivision and genetic drift. Surveying this structure gives us a scientific perspective on human origins, sheds light on evolutionary processes that shape both human adaptation and disease, and is integral to effectively carrying out the mission of global medical genetics and personalized medicine. Surveys of population structure have been ongoing for decades, but in the past three years, single-nucleotide-polymorphism (SNP) array technology has provided unprecedented detail on human population structure at global and regional scales. These studies have confirmed well-known relationships between distantly related populations and uncovered previously unresolvable relationships among closely related human groups. SNPs represent the first dense genome-wide markers, and as such, their analysis has raised many challenges and insights relevant to the study of population genetics with whole-genome sequences. Here we draw on the lessons from these studies to anticipate the directions that will be most fruitful to pursue during the emerging whole-genome sequencing era.

Single-nucleotide polymorphism (SNP): genetic variation in which the observed DNA base pair at a single nucleotide site varies across individuals, typically with only two observed alleles

Haplotype: the set of alleles found at multiple loci along a chromosome or in a haploid gamete

Admixture: the outcome of matings between individuals from two or more genetically differentiated populations

SNP array: an efficient tool for genotyping SNPs; the genotyped SNPs are typically variants with a minor allele frequency >5% in a discovery panel and perform well as tag SNPs

Linkage disequilibrium (LD): the nonrandom association of alleles on haplotypes; LD implies that the allelic state at one locus can be used to predict the unobserved state at another

Tag SNPs: SNPs that because of LD patterns can serve as proxies for other SNPs

1. INTRODUCTION

Human beings have long been preoccupied with their past, and have created numerous complementary disciplines of research to investigate it. Archaeology, biological and cultural anthropology, history, linguistics, and demography all arise from our fascination with characterizing the history of our species. Genetics has increasingly played an important role in this integrative approach to understanding human evolution, especially as technological advances have made it possible to describe the distribution of genetic variation across individuals with increased resolution. Population structure is the presence of variation in levels of genetic similarity within a population as a consequence of factors such as geographical subdivision and finite population size. The departures from random mating that have generated population structure are indicative of historical demographic events and must be understood for many applied uses of population-genetic data. For instance, population structure needs to be considered when searching for the genetic underpinnings of common disease (to avoid spurious associations with noncausal loci) and can help explain variation in the distribution of some genetic disorders across ethnic groups.

A decade has passed since the reference sequence of the human genome was generated (63, 149), and more than 1.4 million single-nucleotide polymorphisms (SNPs) were identified at that time (137), ushering in a new phase of human genetics in which the signatures of population-genetic forces could be studied on a genome-wide scale (60, 61). The distribution of SNPs and haplotypes across geographic regions has provided evidence of past population size changes, admixture events, and barriers to migration, as well as human adaptation to new environments, modes of subsistence, and cultural practices over the past 100,000 years.

Here we review many insights gained about human population structure using SNP array technology. Our understanding of human evolution is deeper than before but still not complete, and as whole-genome sequencing becomes widespread, our field will be empowered to return to some of the oldest questions posed by human geneticists. Nielsen (99) estimated that 30,000 human genomes will be completely sequenced by the end of 2011, and the field of human population genetics stands to gain a great deal from these efforts. It is therefore useful to pause and reflect on the opportunities that have arisen and the challenges that have been posed by the genome-scale observations SNP arrays made possible.

From our own experiences with data interpretation, we feel that an in-depth understanding of human population genetics is possible only if one understands the capabilities and limitations of the inference methods. Moreover, the methodological challenges encountered in analyzing SNP-scale data prefigure many (though certainly not all) of the challenges that will arise in large-scale resequencing data, so we feel a review of existing methods is crucial and instructive at the cusp of the next era of human population genetics. We begin by reviewing the methods for generating and analyzing SNP genotype data, and then move to the insights gained from these methods (Section 4).

2. HIGH-THROUGHPUT SNP GENOTYPING

The development of high-throughput SNP arrays was initiated in the biomedical human genetics community with the hope of using observed SNP genotypes to comb the genome for regions harboring loci affecting complex disease traits—i.e., genome-wide association studies (GWAS). For the approach to work, it required identifying a small set of tag SNPs that, owing to linkage disequilibrium (LD), could serve as proxies for nearby common variants found in human populations. The International HapMap Consortium (59) took on this task, discovering variants and describing patterns of LD in a sparse sampling of the globe, and in doing so catalyzed the development of affordable SNP arrays. In particular, the HapMap Project greatly advanced our

knowledge of LD patterns and recombination rate variation in the human genome (27). There appear to be recombination hot spots spaced roughly every 100 kb, and the common variants found between two hot spots can often be surveyed indirectly with a small number of so-called tag SNPs. Thus, using SNP arrays with hundreds of thousands of markers, patterns of common variation could, for the first time, be assayed at a genomic scale. Moreover, the relatively cheap costs of SNP arrays have led to their rapid application in studies of human population structure (**Table 1**).

The large number of markers on a SNP array offers several advantages. First, because of the coalescent process, there are diminishing returns in the information gained for population-genetic inference by sampling more individuals as opposed to more loci (see **Figure 1**; 42, 115). Because each independent locus has its own coalescent history, the increase in information from additional sampled loci tapers off only once the spacing of sampled loci along the genome is smaller than the scale of LD in the population.

Second, because gene trees are highly stochastic given a particular demographic model (e.g., 134), it is paramount to base inferences regarding population structure on many loci. The patterns at a single locus can be misleading and need not reflect the history of population divergences. For instance, deep divergence between neighboring geographic locations can be inferred from a single locus when in fact gene flow has been homogenous across time and space (64). Under models of divergence followed by migration, two alleles from different populations may coalesce more recently than the time their respective populations diverged (132, 152). As a result, gene genealogies reconstructed from a single locus may not reflect the historical relationships between the populations under comparison, particularly for closely related populations. An appreciation for the stochasticity of single gene trees is one reason why population geneticists familiar with coalescent theory are often wary of detailed demographic inference based on single gene trees, such as occurs in mitochondrial DNA (mtDNA) and Y chromosome studies, and classical phylogeographic studies more generally. [That said, the mtDNA and Y chromosome provided us the first broad sketch of human population structure and continue to be foundational, especially for the study of sex-specific migratory processes (see 159).]

Despite the advantages listed earlier, paradoxically, one of the great efficiencies of SNP arrays—that they interrogate only polymorphic loci—is also one of its greatest hindrances for parameter inference. Array manufacturers must choose SNPs that have been identified in discovery panels (such SNPs tend to have higher minor allele frequencies than random SNPs), and SNPs are spaced to tag common variation across the genome. These two ascertainment biases are not hugely problematic for GWAS that aim to identify common SNPs underlying disease, but are major problems for applying many statistical methods in population genetics that assume a random subset of all variants is being surveyed. One of the great promises of the sequencing era is that it will largely remove the biases introduced by SNP array design. Although this will be largely true, the various quality-control filters currently applied to next-generation sequencing data induce a subtler form of ascertainment bias that can pose challenges for analysis (e.g., 69).

Coalescent process: the process, looking backward in time, by which two or more sampled alleles' ancestral lineages trace back to a common ancestor, often represented by a gene genealogy (gene tree)

Minor allele frequency: the frequency of the minor (i.e., less common) allele

Ascertainment bias: bias because a study's design induces a nonrandom sample of observations; with SNP arrays, bias is introduced via SNP discovery and selection of tag SNPs

3. METHODS FOR INVESTIGATING POPULATION STRUCTURE

3.1. Descriptive Methods for Large-Scale SNP Data

Descriptive approaches for large-scale SNP data fall into two classes: approaches based on discrete population admixture models and approaches based on multidimensional statistics such as principal components analysis (PCA). When using either approach for investigating population structure, two major initial questions are (*a*) whether there are detectable subgroups within the data set under consideration,

Table 1 Survey of published SNP studies from 2008 to 2010 with relevance to human population structure

Region	Sample description	Sample size	Platform	References
Global	HGDP	938	Illumina 650	84
Global	HGDP (29 groups)	485	Illumina 550	66
Global	HGDP	971	Mixed (2,380 SNPs across 211 genes)	14
Global	East Asian, South Asian, European, and African American	3,845	Affymetrix 500K	6, 98
Global	23 groups	383	Affymetrix NspI array (~250K)	162
Global	13 groups	296	Affymetrix 6.0	161
Global	HGDP	944	Illumina 650	12
Global	HapMap 3	1,397	Mixed	62
Africa	12 West African groups, African Americans	203/365	Affymetrix 500K	16
Africa	Gambia	2,340	Affymetrix 500K	67
Africa	Malawi	226	Illumina 650	70
Africa	Nigerians and African Americans	1,188/743	Affymetrix 6.0	74
Africa	South Africa Coloured	60	Affymetrix 6.0	110
Africa	South Africa Coloured	959	Affymetrix 500K	35
Europe	Spain	825	Affymetrix NspI array (~250K)	46
Europe	Spanish and Basque	300	Illumina 300 (followed by targeted genotyping of 109 AIMs)	79
Europe	13 countries	5,847	Illumina 300	50
Europe	Finnish	1,395	Illumina 300/370	65
Europe	23 populations	2,514	Affymetrix 500K	80
Europe	Northern Europe	999	Affymetrix 250K/500K	81
Europe	Nordic populations	5,000	Mixed	82
Europe	Northern Europeans	2,099	Illumina 300	90
Europe	5 European groups	4,110	Affymetrix 5.0/6.0	94
Europe	Northeastern European	1564	Illumina 370	97
Europe	38 populations	3,192	Affymetrix 500K	103
Europe	Ireland and Britain	3,367	Affymetrix 5.0/6.0	107
Europe	3 islands in Scotland, 2 villages in Croatia, and 3 valleys in Italy	36/157/57	Illumina 300	106
Europe	Iceland	877	Illumina 300	119
Europe	Basque	83	Mixed	130
Europe	Finnish	4,763	Illumina 370	136
Europe	12 regions of the United Kingdom	14,000	Affymetrix 500K	158
Europe	Ashkenazi Jewish	471	Affymetrix 6.0	15
Europe and Asia	7 groups	237	Affymetrix 6.0	5
Europe and Asia	14 Jewish and 69 Old World non-Jewish groups	121/1,166	Illumina 610/660	10
European American	Sample derived from U.S.-based GWAS study	4,198	Illumina 300	118

Table 1 (*Continued*)

Region	Sample description	Sample size	Platform	References
European American	Sample derived from U.S.-based GWAS study	>2,000	Illumina 300	146
Asia	73 Asian groups	1,928	Affymetrix 50K	56
Asia	10 provinces of China	8,200	Illumina 610	25
Asia	16 North Asian groups (including 10 populations from Korea) and 3 South Asian groups	240	Affymetrix 50K	71
Asia	7 areas of Japan	7,003	Perlegen (270K SNPs)	166
Asia	Uyghur	26	Affymetrix 50K	164
Asia	19 East Asian groups	100	Illumina 300	145
Asia	Han (26 regions of China)	2,700	Affymetrix 6.0	165
Asia	Austroasiatic Indian and Burmese	41	Illumina 610	24
Asia	India	132	Affymetrix 6.0	129
Asia	Chinese, Malay, and Indian	292	Affymetrix 6.0, Illumina 1M	144
Asia	Qatar	168	Affymetrix 500K	58
Oceania	Polynesia, Borneo, Papua New Guinea, Fiji	100	Affymetrix 6.0	160
Australian Aborigine	Riverine region of New South Wales	38	Affymetrix 6.0	89
Latin America	5 groups	212	Illumina 610/Affymetrix 500	17
Latin America	6 mestizo groups in Mexico and 1 indigenous group	330	Affymetrix 100K	140

List compiled December 2010. Abbreviations: AIM, ancestry-informative marker; GWAS, genome-wide association study; HGDP, Human Genome Diversity Panel; SNP, single-nucleotide polymorphism.

and (*b*) how individuals derive their ancestry from across such groups.

3.1.1. Admixture inference. The Structure method introduced in 2000 by Pritchard et al. (123) has been a standard in the field to address these problems and introduced the discrete population admixture model. This model views each individual's genotype data as being drawn from multiple clusters—each cluster is defined by genotypic frequencies—and the method estimates the proportion of an individual's alleles drawn from each cluster. These clusters may or may not correspond to ancestral populations from which the sampled individuals derived; thus, caution is necessary when interpreting the results (133). Although it has been extremely influential (over 5,000 citations as of December 2010), Structure uses a Markov chain Monte Carlo method that is computationally challenging to apply to genome-wide SNP data. Recently, alternative, computationally fast maximum-likelihood-based estimation approaches (Frappe, ADMIXTURE) have been developed to infer genome-wide ancestry using SNP array data (3, 143).

An extension to Structure by Falush et al. (41) allows individual ancestry proportions to be inferred site by site, known as local ancestry estimates. In local ancestry inference models, each individual's chromosome is modeled as a mixture of haplotypes from distinct parental populations. One of the more refined implementations, HAPMIX (122; **Figure 2**), is based on the Li & Stephens model (85) and infers local ancestry using unphased data from the admixed individual and phased haplotypes sampled from at most two reference populations. A more computationally efficient alternative is LAMP (109, 138), which solves ancestry

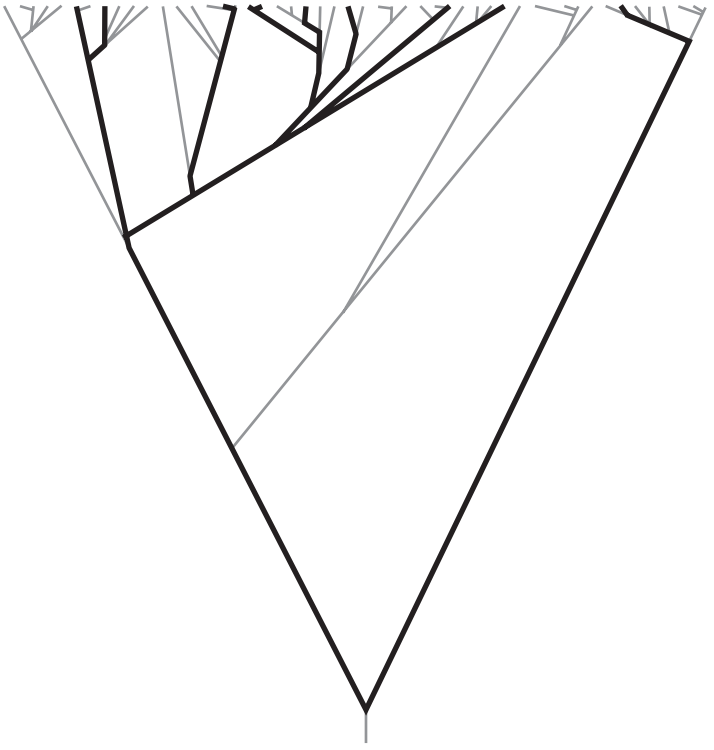


Figure 1

A coalescent genealogy demonstrating the marginal gains in information from increasing the number of individuals sampled. The genealogy depicts relationships among 40 sampled chromosomes. The genealogy of the first 10 sampled chromosomes is marked in bold. Note how the additional chromosomes sampled mainly represent lineages that are closely related to previously sampled lineages, and thus provide little novel information about coalescence rates in the past. Figure reproduced from Reference 42, figure 3.

using an expectation-maximization algorithm in user-defined windows whose size must be chosen carefully depending on the levels of differentiation between parental populations and how recently admixture took place. We expect future extensions of these methods to include modeling admixture between more than two populations and estimating ancestry with incomplete sampling of ancestral populations.

In place of estimating admixture at the individual level, there are also approaches to infer admixture at the population level using SNP data. Reich et al. (129) proposed novel methods to infer population-level genome-wide proportions of ancestry and to estimate the extent of genetic drift along branches in the phylogenetic tree describing the parental and admixed populations. The methods are based on 3-Population and 4-Population tests (129), as well as a general model-fitting procedure based on squared allele-frequency differences among populations. Using these methods, the authors found indications of a potential north–south gradient of admixture in India and uncovered evidence that samples from India are well modeled as a mixture of "Ancestral North Indians" (ANI, related to Europeans) and "Ancestral South Indians"(ASI) (**Figure 3**).

Most applications of admixture methods have focused on admixture events that occurred between distantly related populations (e.g., African Americans as an admixture of European and African ancestry, as in 16). As geographic sampling of individuals has become more refined and as sequencing technologies improve, admixture will be studied at smaller geographic scales in which gene flow rates are high and potentially vary through time (for an example of this challenge, see 17).

3.1.2. Multidimensional summary statistics.

An alternative to inference with the admixture model has been to apply methods from multidimensional statistics, in particular PCA. One motivation for this is that, for populations studied at fine geographic scales, models of discrete populations (even with admixture) are inappropriate. PCA methods may more flexibly describe continuous forms of population structure. Population-based PCA approaches were first used with allele frequencies at classical markers (22, 92), but recently PCA has been applied to individual-level SNP genotype data by coding SNP genotypes as ordered integers and normalizing the values (e.g., 120). With sufficient data, PCA will produce clusters representing discrete populations, and admixed individuals will occur at positions between their source populations that are determined by their relative ancestry from their source populations (91, 111, 120). In the presence of homogeneous isolation-by-distance patterns, the first two principal components (PCs) will represent perpendicular gradients in geographic space

(91, 104), such that if individuals are plotted along PC1 and PC2 coordinates they will be arranged according to their geographic positions (e.g., center panels of **Figure 4**).

The chief drawbacks of PCA are that sampling density affects the results (91, 104), large regions of LD should be pruned from the data before running PCA (103, 146, 158), and interpreting the PCs can be complicated [for instance, when interpreting PCs as signatures of population expansions (20, 43, 104, 105)]. Several interesting developments related to PCA are theoretical results that relate PCA coordinates to expected pairwise coalescent times (91), methods to identify transformations that maximize similarity between a PC and geographic map (155), methods to infer ancestry blocks using PC coordinates computed along a chromosome (16), and more flexible alternatives to PCA based on sparse factor analysis (40) and mixture models (e.g., 73).

The use of descriptive methods to detect population structure is at an exciting juncture given how rapidly sample sizes and the number of loci are increasing in human data sets. Patterson et al. (111) argued that if the product of the sample size and the number of loci is greater than a threshold determined by F_{ST} (a measure of population differentiation) between populations, then population structure will be evident in the sample. Although the threshold was derived for PCA-based methods and is an approximation, presumably a similar result also holds for admixture-based methods (see also 37 and 91). As a consequence, historians or anthropologists attempting to understand very fine-scale historical questions (34) can expect that, with large numbers of markers and individuals, subtle relationships among populations might soon be discerned.

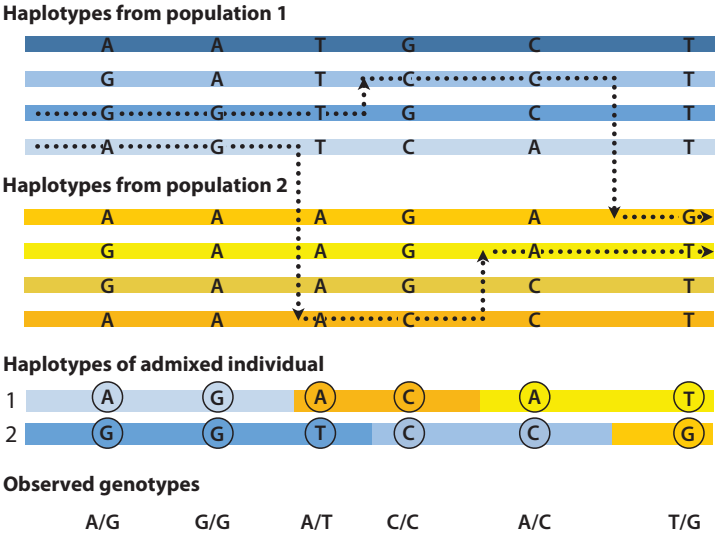


Figure 2

Schematic of a hidden Markov model used for local ancestry inference. The haplotypes of an admixed individual are generated as a mosaic of haplotypes that are copied from the reference panel (population 1 haplotypes are shown in shades of *blue*; population 2 haplotypes are shown in shades of *yellow*). Genotypes in the admixed individual are observed, and the unknown ancestry of the haplotypes is the goal of inference (i.e., the *blue/yellow* status at each point along the admixed haplotype). The copying paths for both of the admixed individual's haplotypes are shown with dotted lines and reflect the ancestry of each segment. Not represented in the figure are cases with genotyping errors, alleles in the admixed individual that are not found in the reference panel, and miscopying events that model coalescent events predating population divergence. Figure derived from Reference 121, figure 1.

3.2. Demographic Inference with SNP Array Data

The descriptive methods that have evolved to explore fine-scale SNP-based data sets provide insights into geographically local and global genetic structure. A central challenge of conducting inference in formal demographic models (with parameters such as divergence times and migration rates) has been accounting for ascertainment bias in SNP data.

Several authors have presented the problems ascertainment bias induces as well as possible corrections (2, 26, 38, 75, 88, 100, 127, 131, 153). For example, Albrechtsen et al. (2) showed that ascertainment bias has little effect on some statistics (e.g., F_{ST}), but a substantial effect on others (such as PCA) depending on the populations used for SNP discovery. Correction methods require knowledge of the SNP selection process, which has typically been complex, heterogeneous, and thus effectively unknown.

To help address ascertainment bias, some studies propose a model for the SNP selection process, then fit the basic parameters

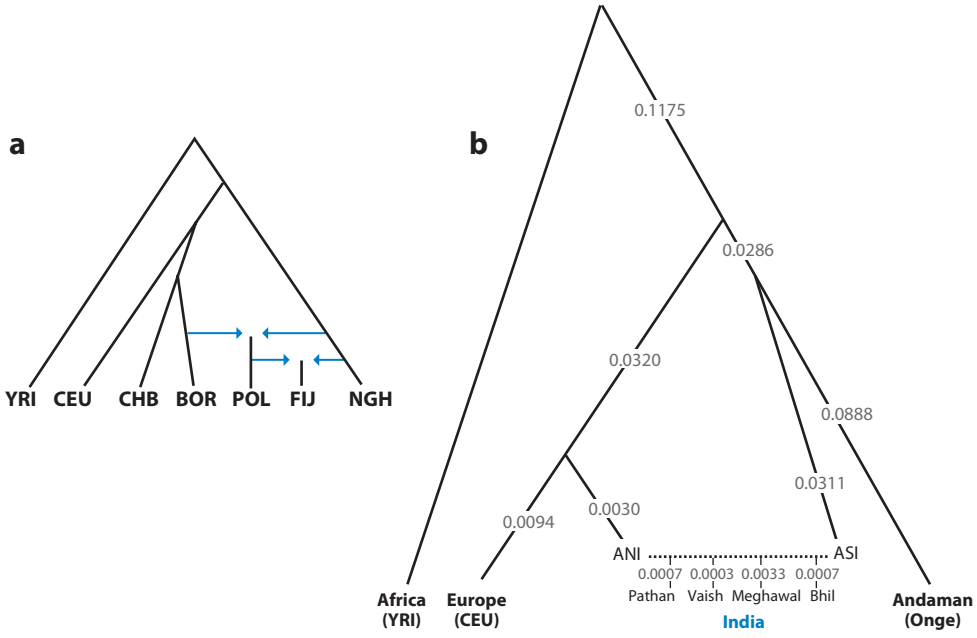


Figure 3

Two population histories with admixture inferred from single-nucleotide polymorphism (SNP) data. (*a*) An out-of-Africa demographic model with an Asian admixture scenario for Polynesians. All Eurasian populations appear to descend from a single ancestral population. Polynesians (POL) appear to be admixed between ancestors of Borneo (BOR) and New Guinea Highlands (NGH) populations, and Fijians (FIJ) show evidence of additional NGH ancestry. Other abbreviations: CEU, Utah residents with ancestry from northern and western Europe (CEPH collection); CHB, Han Chinese (Beijing, China); YRI, Yoruban (Ibadan, Nigeria). Figure from Reference 160. (*b*) Indian populations can be modeled as admixtures of Ancestral North Indian (ANI) and Ancestral South Indian (ASI) populations, where the ASI populations are most closely related to Andamanese populations. The authors' estimates of genetic drift on each lineage are shown. Figure from Reference 129.

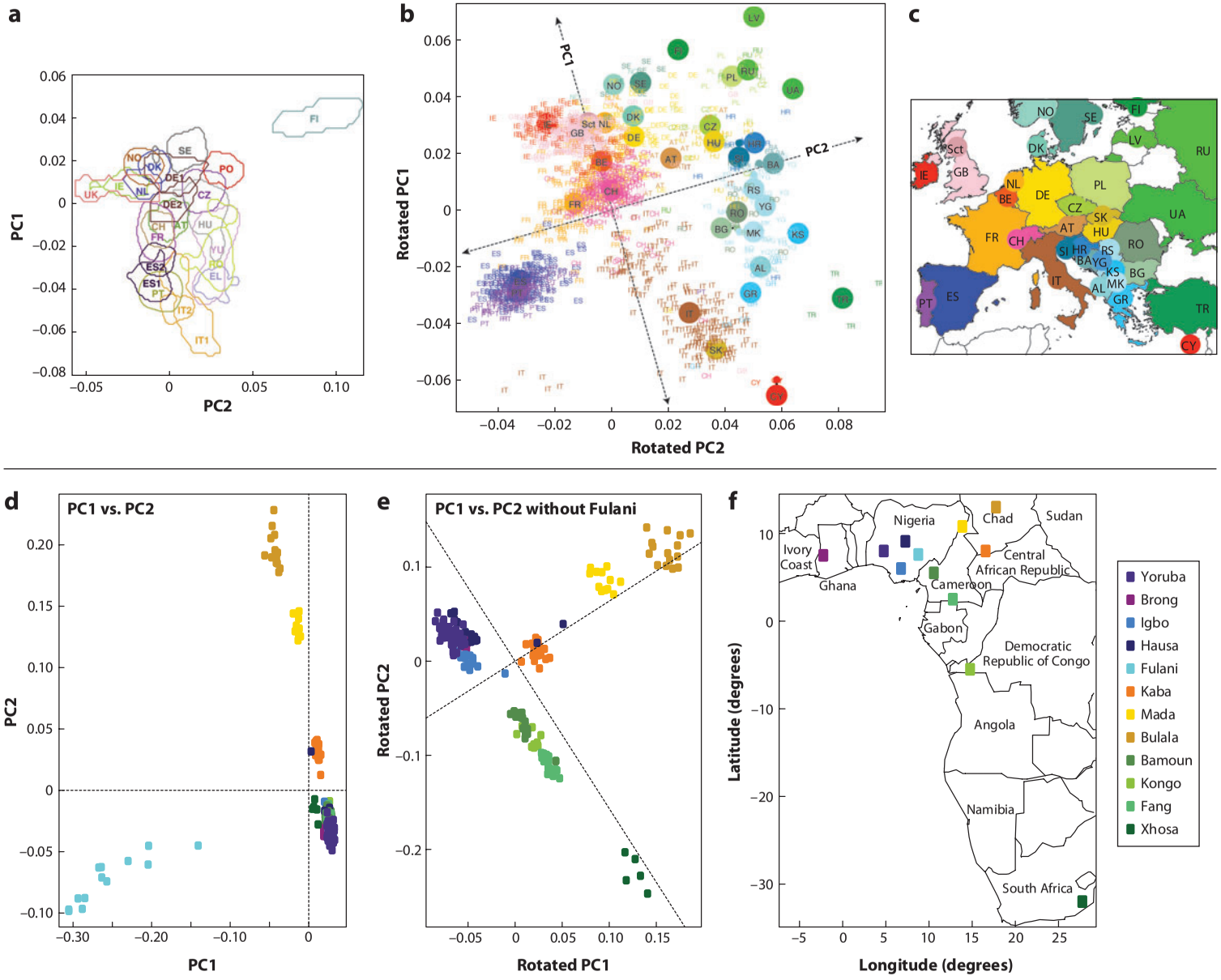
Demographic parameters: parameters that define a population's size and structure over time (e.g., divergence times, rates of migration and growth, and the severity/duration of past bottlenecks)

of the model and account for the resulting ascertainment bias in any demographic inferences (2, 75, 88, 160). For example, Wollstein et al. (160) perform a correction for ascertainment on the Affymetrix 6.0 SNP microarray in their study of Oceania (**Figure 3*a***). The authors modeled the discovery depth for each population and incorporated this into an approximate Bayesian computation framework for inferring demographic parameters. The authors found that their postcorrection inferences were similar to demographic inferences made with the resequenced ENCODE (Encyclopedia of DNA Elements) regions.

Another approach has been to base inferences on patterns of haplotype variation, as these are less sensitive to ascertainment bias (28, 86, 95). It has been shown that past

Figure 4

Population isolates, genes, and geography for European (*top*) and West African (*bottom*) samples. Each row shows two principal component (PC) analysis plots paired with a geographic map. In each case, the leftmost plot (*a* and *d*) shows evidence of an isolated population [Finnish (FI) in the case of Europe, Fulani in the case of West Africa]; the central plot (*b* and *e*) shows how, in analyses without the isolate, the remaining variation is explained clearly by geography; and the rightmost plot (*c* and *f*) provides sampling locations. Figures from References 17, 67, 80, and 103.



demographic history can be inferred from genome-wide SNP data using summary statistics such as the number of observed haplotypes and the frequency of the most common haplotype in windows across the genome (86). Full likelihood-based approaches for haplotypes are challenging—one approximate approach investigated by Davison et al. (31), based on the copying-with-recombination approach of Li & Stephens (85), is promising but produced biased estimators that require data-set-specific simulations to correct.

The copying-with-recombination model of Li & Stephens (85) has also been used to study human colonization history assuming a serial founder model (51). In the serial founder model (33, 57, 126), new daughter populations are formed by subsampling previously founded populations in an expansion. The model's expectation is that haplotype diversity will decrease and LD will increase with distance from the origin of the expansion (32; **Figure 5**). Thus, one expects more recently colonized regions of the world to produce samples that can be modeled as copies of haplotypes found in regions of the world colonized further in the past. The method of Reference 51 uses this idea to infer a plausible order in which the world was colonized, and finds support for the out-of-Africa model.

Using haplotypes is not without its challenges, though: haplotype-based inferences require modeling recombination across the genome accurately, and the ascertainment bias introduced by selecting SNPs preferentially based on physical distance or LD patterns will not produce an accurate picture of haplotypic variation [although Lohmueller et al. (86) carefully accounted for the bias introduced by SNP discovery].

Although analyzing sequence similarity in a pairwise fashion has been routine in population genetics (e.g., computation of average pairwise nucleotide differences), genome-wide markers have opened up to inspection a new dimension of pairwise patterns of genetic variation. One can use SNP data to identify the lengths of regions of pairwise haplotype identity, known as shared haplotype tracts. These can be considered in pairs of haplotypes found within a single individual (i.e., runs of homozygosity that may aid the mapping of disease genes in inbred populations) or between individuals (in which case shared haplotype tracts indicate relatedness, and long shared tracts can imply recent common ancestry).

Methods to perform demographic inference using shared haplotype tracts are in their infancy. Reich et al. (129) used a related genomic summary, the decay of allele sharing between individuals as a function of physical distance between SNPs, to estimate the age of founder events for Indian populations. The results suggest that many subgroups were founded more than 30 generations ago and that strong endogamy has existed since then, allowing the signatures of founder events to still be detectable in extant individuals hundreds of years later. This history of endogamous marriage predicts a high rate of recessive diseases, and the use of runs of homozygosity to identify mutations underlying Mendelian disease may be successful in India as it has been in Finland (e.g., Meckel syndrome in 142). Extended haplotype sharing has also been studied in relation to the diseases and demographic history of Ashkenazi Jews (5, 15).

Shared haplotype tracts have the potential to clarify whether migration occurred once between groups, at a constant rate, or at a variable rate over time. Pool & Nielsen (117) developed

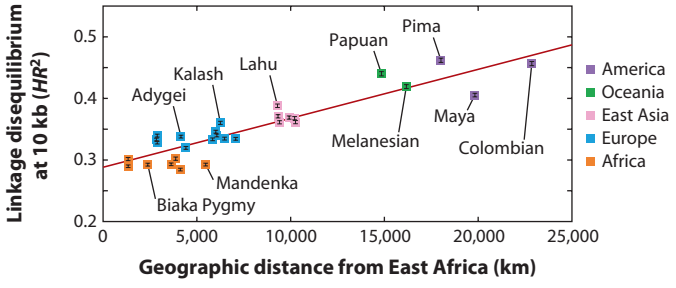


Figure 5

Most likely as a result of serial founder effects during the expansion outward from an origin in East Africa, humans have decreased haplotype diversity and in turn increased linkage disequilibrium as a function of geographic distance from East Africa. Figure reproduced from Reference 66, figure 2*c*.

a method to test demographic hypotheses related to historical changes in migration rate. Their method uses shared haplotype tracts between source and admixed populations, along with the simplifying assumption that migrant tracts do not recombine together, and uses an excess of long shared tracts between groups as evidence of a recent increase in migration rate.

As sequencing data enhances our ability to identify breakpoints at the end of shared haplotype tracts, we expect to gain more information about demographic processes through the information held in the spatial arrangement of shared SNPs in the genome.

3.3. The Power of Sequencing Data to Estimate Demographic Parameters

Most quantitative population-genetic models that are capable of inferring demographic parameters rely on observations of the allele-frequency spectrum without ascertainment bias (26, 100). Thus, the increase in genotype data generated by SNP array technology has come at the expense of understanding human evolutionary history through the estimation of demographic parameters. The sequencing era will allow population geneticists to move beyond the limitations of ascertainment bias to harness genomic-scale data for the inference of detailed demographic parameters—and will require overcoming new challenges.

Fitting formal demographic models to data sampled from many human populations means estimating a large number of parameters that represent historical processes and events (for an example, see **Figure 6**). Because many of the most advanced population-genetic inference methods are framed in terms of the coalescent genealogy underlying the sampled chromosomes—and that genealogy is not observed directly—most methods use approximate methods, such as Markov chain Monte Carlo, to sum over all possible genealogies consistent with the data. Examples of such approaches are isolation-with-migration models (8, 53–55, 101, 157) and the models in the LAMARC, MIGRATE, and BEAST software packages (9, 36, 78). Obtaining accurate approximations with these methods for even a handful of loci is challenging when the sample size reaches the scale of hundreds of individuals.

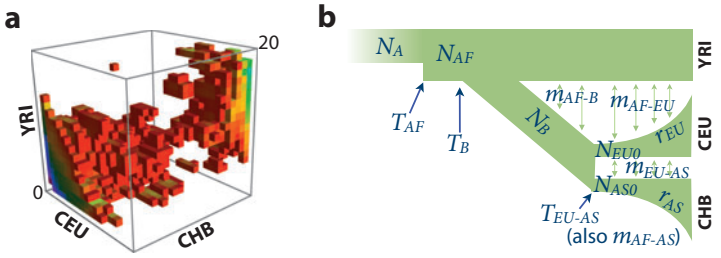


Figure 6

(*a*) An example of a three-dimensional site frequency spectrum based on individuals from three HapMap populations (60), and (*b*) a schematic of the demographic model relating the populations' joint history. In panel *a*, warmer colors (e.g., *red*, *orange*, and *yellow*) code smaller numbers of observations relative to the cooler colors (e.g., *blue*, *green*, and *purple*) for the three-way joint allele frequency represented by that position in space. Note how the distribution is concentrated in the lower left rear of the cube, a region that represents rare, derived alleles. This joint frequency spectrum was used to fit 14 free parameters to the demographic model depicted in panel *b*, including effective population sizes at various points in the joint history, times of divergence, migration rates, and recent population grown rates. For example, N_A is an ancestral population size, N_B is the effective size of individuals in a bottleneck, T_{AF} represents a time at which growth in Africa began, and m_{EU-AS} denotes the migration rate between Europe and Asia. Abbreviations: CEU, Utah residents with ancestry from northern and western Europe (CEPH collection); CHB, Han Chinese (Beijing, China); YRI, Yoruban (Ibadan, Nigeria). Figure from Reference 48, figure 2*a*,*b*.

New approaches to analyze DNA sequence polymorphisms across many loci and multiple populations were recently offered by Gutenkunst et al. (48) and Garrigan (45), who summarized genetic variation by the joint allele-frequency spectrum across populations and implemented composite likelihood methods. Gutenkunst et al.'s (48) method offers a promising, computationally efficient alternative to the coalescent approach via a diffusion-based method, ∂a∂i. This method allows the modeling of three simultaneous populations (see **Figure 6**) and produces results consistent with growth rates, bottlenecks, and migration rates inferred by previous analyses (1, 88, 139, 151).

The composite likelihood approaches assume polymorphic sites are independent, which could constrain the approach's applicability to

Allele-frequency spectrum: the counts of polymorphisms at all possible observed frequencies in a sample; for unlinked sites, it is a sufficient statistic for demographic inference

large-scale genomic data sets. Gutenkunst et al. (48) address this concern by using a bootstrap procedure accounting for LD to estimate confidence intervals for parameters, noting that LD between neutral loci does not affect the expectation of the allele-frequency spectrum, but rather the variance.

To maximize the information gleaned from whole-genome data, methods incorporating LD into inference are needed. The Cosi method (139) models LD patterns and the allele-frequency spectra across three populations; this method is quite computationally intensive owing to its coalescent-based approach and cannot provide confidence intervals around best-fit model parameters. Because the allele-frequency spectrum does not uniquely determine past population history (95), variation at linked loci can provide additional information to infer changes in population size. Thus the sensitivity of modeling techniques to LD remains an important area for investigation.

As advances in next-generation sequencing make the sequencing of whole genomes economical, we expect there will be a return to model-based approaches to estimate demographic parameters. Here we have noted only some of the recent developments in these quantitative approaches. The methods used to model the joint histories of divergence and gene flow among multiple populations that will be developed in the sequencing era will make use of increasingly vast data resources, while helping geneticists return to some of the oldest questions regarding modern human evolution.

4. INSIGHTS INTO HUMAN POPULATION STRUCTURE FROM SNP STUDIES

SNP studies build on a long legacy of studies in human population genetics, archaeology, and linguistics that have produced a sketch of the events that shape how human populations are structured today (**Figure 7**, 21; also see 68 and references therein). The major outline is of the emergence of modern *Homo sapiens* in Africa 100,000 years ago; then, out of a possibly subdivided population within Africa, humans began to expand outside of Africa to colonize the rest of the globe, undergoing serial founder effects as they did so, and eventually settling distant locations such as the Americas (15,000–35,000 years ago) and most recently the remotest islands of Oceania (2,000–3,000 years ago). In many cases recent migrations have led to contact between distantly related populations and given rise to major admixture events—for example, as occurred with the ancestors of Latinos in the Americas. The details beyond this coarse-scale resolution have been the subject of considerable uncertainty and debate, and SNP-based studies combined with recent sampling strategies are providing increased data resolution with which to resolve these and other questions about human history.

Summarizing the results of these recent SNP-based studies is challenging because human population structure is fractal-like—there are fascinating patterns to behold at the macroscopic scale of global structure, yet one can continually focus in to continental scales, subcontinental scales, and onward in an expanding effort to learn about the outlines of human history in finer and finer detail (**Figure 8**). Recent studies even address population structure at the extreme of neighboring villages (106). It is important to recall that human genetic groups are extremely similar to one another on numerous metrics (e.g., pairwise sequence diversity, proportion of variation found within versus between populations). It is only by pooling information across multiple loci that SNP data sets can resolve subtle patterns of population differentiation that are indicative of the human past (37, 83).

4.1. Global-Scale Studies

While the HapMap provided insights into common variation and LD in three populations, two studies using the Human Genome Diversity Panel (HGDP; 52 populations) marked the advent of SNP-based studies of human population

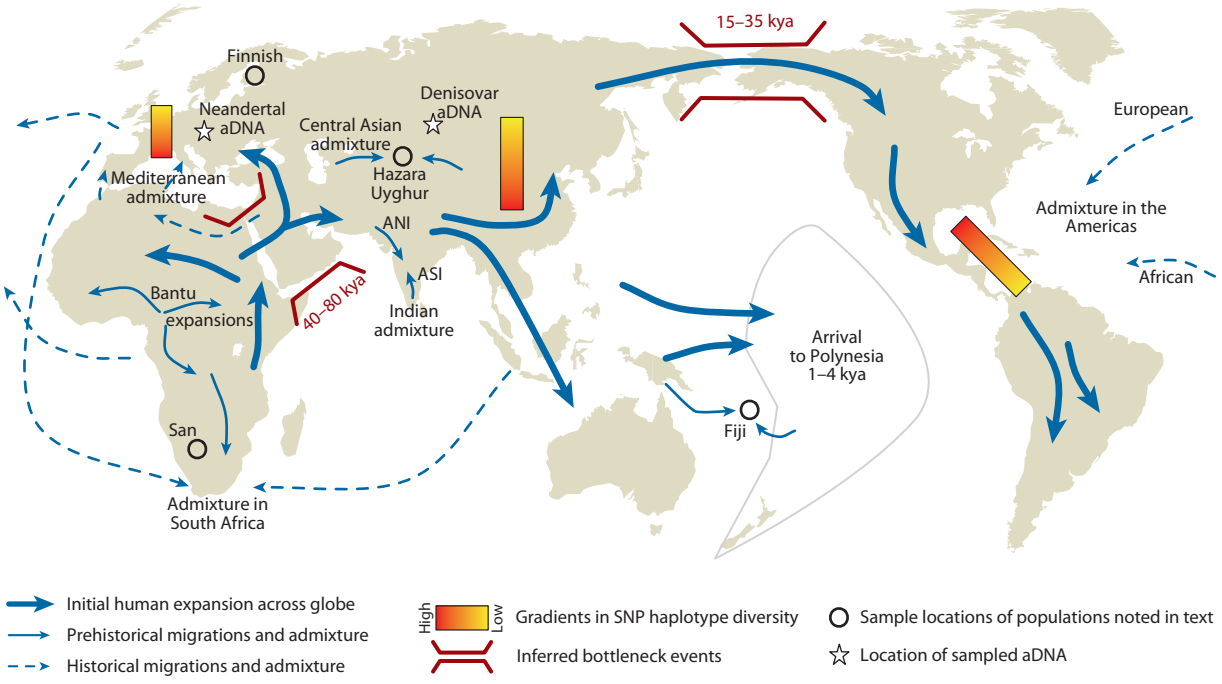


Figure 7

A schematic of human demographic history, highlighting hypotheses investigated by recent single-nucleotide polymorphism (SNP) studies discussed in this review. Numbers indicate the estimated number of years before present at which migrations took place across continents. The distinction between prehistorical and historical is in some cases approximate—the timing and duration of these events is an ongoing area of study. Abbreviations: aDNA, ancestral DNA; ANI, Ancestral North Indian; ASI, Ancestral South Indian.

structure (66, 84). These are part of a long line of papers using the HGDP (18) to showcase the power of new technologies for advancing our knowledge of population structure (28, 29, 126, 135), and have been built upon by other global studies (6, 12, 14, 161, 162). These studies found patterns that support serial founder effects during the expansion outward from Africa. Support for this model is not new (124, 126), but novel observations of genome-wide patterns of haplotype homozygosity, patterns of pairwise LD (**Figure 5**), and copy number variation now give strong support to the serial founder effect model of out-of-Africa expansion. Fine-scale population structure within continents was also observed in these studies, but given the sampling design of the HGDP, detailed and conclusive insights into structure within global regions require larger, regionally oriented sampling designs found in the studies discussed below.

4.2. Africa

Dense sampling of African populations has been relatively rare in population-genetic studies undertaken thus far. Recent efforts reveal that population structure in sub-Saharan Africa correlates with geography, language family, and mode of subsistence (e.g., hunting and gathering versus herding) (16, 52, 147). Population structure in West Africa is determined largely by language group and geographic distance, bearing a strong signature of the Bantu expansion approximately 4,000 years ago (16, 70). Hunter-gatherer populations studied so far remain highly differentiated from each other, and tend to have the lowest levels of genome-wide LD observed across African populations (52).

SNP studies make clear that haplotype diversity is high in Africa (for example, 4, 66, 84). This creates problems when using tag SNPs discovered in Europeans, and even tag SNPs

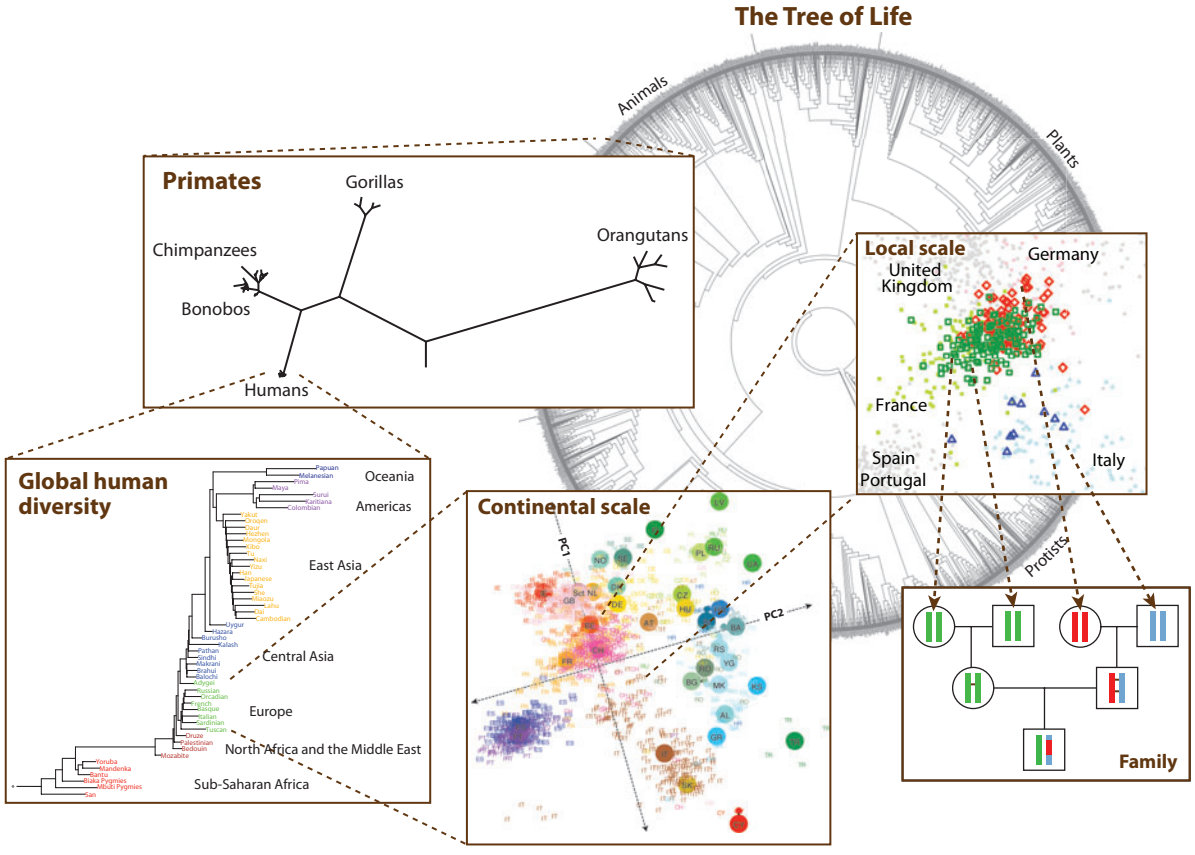


Figure 8

The fractal-like nature of human population structure. Figure sources: the Tree of Life (112; also see **http://www.zo.utexas.edu/faculty/antisense/DownloadfilesToL.html**), primates (72), global human diversity (84), continental scale (103), local scale (103), family (author illustration).

selected from the Yoruban HapMap population, for GWAS in other populations. It was estimated that, to achieve the same statistical power in a GWAS, one would need 1.5 million SNPs in Africans compared with 0.6 million in Europeans (60). Even with the same causal variant in both populations, it may be more difficult to replicate a disease association in African populations (74). The sequencing era allows for techniques that can overcome this problem—namely, multilocus imputation and economical targeted resequencing [both used illustratively in a study of malaria and hemoglobin S by Jallow et al. (67)]—but understanding the genetic underpinnings of disease in Africans and admixed individuals with African ancestry will first require concerted effort to discover variation in potentially highly differentiated populations. Imputation will be particularly important to assess African differentiation at a fine genomic scale; the positive trade-off for association studies is that low LD means identifying the causal variant influencing phenotypes is easier than in non-African populations (67).

An understanding of African population structure also contributes to increased understanding of the evolution of various admixed populations throughout the world. Bryc et al. (16) deconvoluted ancestry along chromosomes in African Americans, finding variation in the estimated proportion of European ancestry (median 18.5%) and African ancestry

derived from groups similar to non-Bantu Niger-Kordofanian populations. A population with mixed ancestry within Africa that has been recently studied is the Cape Mixed Ancestry population, also referred to as "South African Coloureds." Trade and colonization led to the establishment of this population, which makes up 9% of South Africa's population, and these individuals draw ancestry from groups similar to the Khoisan and Bantu Africans, as well as from Europe and Asia (35, 110, 147). As with many recently admixed populations, there are signatures of sex-biased migration found in their genomes: mtDNA analyses have shown that the maternal contribution to the population is strongly Khoisan (125), while the non-African contribution to X-chromosomal ancestry is inferred to derive from Indonesia, a result of Indonesians being brought to South Africa by the Dutch (110).

Many open questions remain regarding the evolution of populations in Africa; we mention two in particular here. First, we know little about what population structure existed in Africa at the time of the out-of-Africa migration. Subdivision of these ancestral populations affects inferences of divergence times between Africans and non-Africans (48), as well as inferences about our interactions with other hominids (see Section 4.7). Second, although studies of African populations find a strong signature of homogenization due to the Bantu migration 4,000 years ago (16, 147), less is known about the demographic histories of hunter-gatherer populations, although they share common ancestry separately from other African populations (52). Whole-genome sequences will bring a deeper understanding of African population structure, shedding light on the origins of modern humans and on the roles of cultural traits in shaping human genetic variation.

4.3. Europe

Given the concentration of biomedical research that takes place in European countries, the first geographic area to receive detailed attention using SNPs was Europe. Within the span of a year, several groups used PCA-based approaches to produce multiple "genetic maps" of Europe—showing dramatically how the genetic structure of European populations is predicted by geography (**Figure 4**) (50, 80, 90, 103, 146). The scale of detectable geographic structure is such that it can be detected within small geographic areas, such as Switzerland (103), Finland (136), Iceland (119), and rural regions of Europe, even among geographically proximal villages (106).

There are notable departures from the general geographic patterns—Finnish, Sardinian, Basque, and European Jewish individuals all show unique patterns of variation—in many cases supporting hypotheses of bottlenecks and reduced gene flow between these populations and others in Europe (**Figure 4**) (5, 10, 15, 80, 136, 148; also see 14, 44, 46, 79, 130). Some populations that are not particularly differentiated are of interest—for example, Hungarians, though speakers of a Finno-Ugric language, are genetically much like their Indo-European-speaking geographic neighbors (e.g., 103), supporting models in which the Hungarian language was imposed by elites and providing an interesting example of how linguistic and genetic patterns sometimes misalign.

Much of the study of European population genetics has centered on understanding whether there are directional clines to the patterns of variation. Some of this work was conducted originally using PCA-based methods applied to populations (22, 92)—in this light, it is interesting to note how the direction of the PC1 gradient in the various SNP studies widely differs (e.g., north-northwest–south-southeast in Reference 103 versus nearly west–east in Reference 50). This inconsistency is not unexpected given the sensitivity of PCA to the spatial distribution of the sample (see above). More reliable indicators of directionality in the patterns of variation come from observations of decreasing south–north gradients in haplotype diversity in Europe (6, 80).

Such gradients in haplotype diversity are expected under several models that would produce a larger effective population size (N_e) for southern Europe populations. Relevant models might include features such as the initial human colonization northward into Europe, subsequent expansions/contractions associated with the Pleistocene, the onset of the Neolithic Era in Europe, and/or substantial trans-Mediterranean gene flow with northern Africa. Recent results have shown that the highest haplotype diversity in European samples is found in the Iberian peninsula, and this region shares more haplotypes with Yoruban populations than does any other European region (6). These results, along with those of a recent analysis based on 3-Population and 4-Population tests and the chromosomal scale of allelic correlations (93), suggest a history of trans-Mediterranean gene flow at least partially contributes to the higher levels of diversity found in southern Europe. These studies are probably the first of several that will move beyond coarse descriptions of how genetics and geography relate in Europe to more detailed inference of past demography.

4.4. Asia

Several large SNP studies of East Asian populations (25, 60, 71, 144, 145, 165, 166) and South Asian populations (24, 129) have been undertaken. Many of these studies observe fine-scale geographic structure—for example, within the Japanese islands (166), Korea (71), India (129), and Polynesia (160). At a broader scale, these studies give insights into the history of Asian populations, but the results are difficult to summarize and interpret because Asia encompasses such a broad geographic area and complex network of populations. Despite this complexity, a few clear patterns have emerged from these major studies.

The patterns of variation observed in the HUGO Pan-Asian SNP Consortium data (56) argue for a single southern route during the initial colonization of southeastern and eastern Asia, followed by northward expansions counter-clockwise around the Tibetan plateau. A strong correlation of haplotype diversity was found in East Asia with latitude, and this has been interpreted as the outcome of serial founder effects as populations of humans moved northward from an initial presence in southern Asia.

In the HGDP SNP data, East Asian and Southeast Asian populations cluster distinctly from central Asian populations (66, 84). Further, the alignment of numerous allele-frequency clines at putatively selected loci (29) suggests that the western regions of the Tibetan plateau are an important area of genetic diversity in Eurasia. For instance, the Uyghurs and Hazara show evidence of descending from an admixture event ~2,500 years ago between ancestral populations that gave rise to East Asian/Southeast Asian and central Asian/European populations (66, 84, 163, 164).

Studies of populations from the Indian subcontinent show substantial structure at fine scales (6, 24, 129, 161, 162), and Reich et al. (129) argued that the data are best explained by models in which Ancestral North Indian (ANI) and Ancestral South Indian (ASI) populations contribute to the diversity of modern Indian populations. Although it is not clear, the ANI population likely was made up of descendants of central Asians who then expanded into India; the placement of the HGDP central Asian populations on the extreme of the ANI/ASI cline (129) and affinities between central Asians and northern Indians suggest this may be true (129, 161, 162). The ASI population's closest living relatives in the study were Andaman Island populations (Onge), who themselves appear more recently diverged from Southeast Asian groups (Dai) than Near Oceanian groups (Papuans). The history of the ASI population is not clear and may require integrating patterns of variation observed in Southeast Asia, Australia, and Oceania to gain more insight.

4.5. Australia and Oceania

Australian and Oceanian populations have been largely overlooked, with only two relatively

small studies published recently (89, 160). The single study from Australia (89), based on 38 sampled Australian aboriginal individuals, shows Australian samples clustering with indigenous groups of Near Oceania and evidence for recent admixture with Europeans. A major question is how the ancestors of Australians and Near Oceanians are related to neighboring Asian populations. A SNP study has concluded that Near Oceanian groups, such as New Guinea Highlanders, diverged from Eurasian populations fairly recently, approximately 27,000 years ago (**Figure 3**; 160), suggesting a single migration to East Asia, Near Oceania, and Australia followed by a recent divergence of Near Oceania/Australian populations from East Asians.

Moving east to Polynesia, a SNP study by Wollstein et al. (160) supported a model in which Polynesians descend from an admixture event roughly 3,000 years ago between ancestors of modern Austronesian speakers and Papuan speakers. One exception is Fiji, which appears to have a unique signature of additional Papuan admixture (500 years ago) after Polynesian settlement (**Figure 3**).

Despite these advances in understanding Australian and Oceanian populations, there are still many open questions regarding how these populations relate to their nearest neighbors, and given that modern humans may have followed a route along the coast of the Indian Ocean to first occupy Asia, Australia, and Oceania, it is difficult to consider the studies of these regions in isolation of each other. The challenge in addressing these questions is the need to integrate patterns of variation across a vast geographic area, while current studies cover smaller scales and only partially overlap. Whole-genome sequencing will not necessarily make the logistical and analysis challenges to achieving such a synthesis more tractable, but it will provide new insights. For example, the very recent genome sequencing of an ancient specimen from Siberia (Denisovar genome) observed a unique signature of admixture in Near Oceanian populations that is suggestive of prehistorical population movements and structure around the time of admixture with Denisovars (128).

4.6. Americas

In considering patterns of variation in the Americas, there are two major facets: the description of Amerindian populations, and the description of admixed populations in the Americas. Two recent studies focus on the description of admixed Latino groups and highlight their wide range of admixture both within and between populations (17, 140). Ancestry varies within Mexico (140) and across different regions of the Americas. For example, populations from Mexico and Ecuador have more Amerindian ancestry, whereas populations from the Caribbean (Puerto Ricans and Dominicans) have much more African ancestry (17).

The increased resolution of SNP-based data allowed Bryc et al. (17) to assign continental-level ancestry to segments of each individual's genome. Strikingly, they found that across admixed Latino individuals, European segments are most similar to southern European variation, African segments are most similar to Yoruban (West African) variation, and Amerindian segments tend to vary between populations, with the Mexico sample showing a large contribution from an ancestral population similar to the current-day Nahua, one of the largest Amerindian populations in the region. They also were able to contrast X chromosome versus autosome patterns of ancestry with patterns from mtDNA versus Y chromosome haplotypes to show evidence for a bias in matings toward pairings of males with more European ancestry and females of more indigenous ancestry. These fine-scale patterns are in agreement with several aspects of the historical record, and suggest that genetic variation can reveal historical processes. Neither SNP study revisits the question of the original peopling of the Americas, though the HGDP SNP studies [and microsatellite studies (156)] show that Amerindian genetic diversity is low, consistent with a serial founder effect (66, 84). To fully advance studies of human genetic variation,

further studies are needed to describe patterns of Amerindian diversity and to survey the wide range of admixture patterns in the Americas.

4.7. Archaic Admixture

There is broad consensus that anatomically modern humans emerged from Africa and replaced other hominins. What is not clear is the extent to which *Homo sapiens* admixed with archaic humans during this process. If small amounts of admixture took place, it would produce deep divergence and unique patterns of population structure in a small fraction of loci in the human genome.

Motivated by the HGDP SNP data, DeGiorgio et al. (32) conducted a simulation study and were able to reject a model where modern humans serially expand into and mix with a collection of preexisting archaic populations. Such a model of archaic persistence produces an increase in heterozygosity with distance from the origin of the modern human expansion as well as a decrease in LD with distance from the expansion, neither of which are observed in SNP data from globally distributed human populations. Instead, DeGiorgio et al. (32) found that global SNP patterns are consistent with either a serial founder model without an archaic contribution or one with a small amount of archaic ancestry (5% in their study).

The development of high-throughput DNA sequencing technologies has allowed for the genome-wide sequencing of nuclear DNA from ancient specimens, culminating in a draft sequence of the Neandertal genome in 2010 (47). These data suggest that between 1% and 4% of Eurasian genomes are derived from Neandertals, supporting the idea that the vast majority of genetic variants outside of Africa came from Africa during the modern human diaspora, but also supporting archaic admixture with modern humans before the divergence between Eurasian peoples. Supporting evidence comes from sequence-based studies such as those of Plagnol & Wall (114) and Wall et al. (154), who used sequence data to estimate a 14% archaic contribution to European samples (argued to be from Neandertals) and a 1.5% admixture proportion in the East Asian samples studied (argued to be from *Homo erectus* and/or *Homo floresiensis*). An interesting finding of Green et al. (47) is that Neandertals do not appear to be more closely related to Europeans than to East Asians. This may be due to the Neolithic expansion erasing signatures of earlier admixture between archaic hominins and *Homo sapiens* in Europe, or due to ancient substructure in Africa before the modern human expansion. The technology that produced the Neandertal genome creates exciting opportunities to explore these alternative hypotheses with studies of additional preagricultural human and hominin fossils.

4.8. Expanding Beyond the Autosomes

The generation of genome-wide data sets has led to new appreciation of X-linked genetic variation across human populations. Comparing inferences made across all the marker systems available in the genome provides important insights into human evolutionary history. One recent example is a study by Blum & Jakobsson (13), who found that an ancient bottleneck in the Middle Pleistocene, possibly from an ancestral structured population, can reconcile the distributions of variation found on the mitochondrion, autosomes, and X chromosome from the human expansion out of Africa.

The X chromosome's female-dominated life history and hemizygosity in males make it an ideal system for comparison with the autosomes to study sex-specific demographic differences, differences in mutation rate and patterns of LD between males and females, and the behavior of both adaptive and purifying selection (**Table 2**). For example, Bryc et al. (16) observed elevated levels of African ancestry on X-chromosomal blocks relative to autosomes, providing evidence for sex-biased gene flow. In general, though, theoretical expectations for X-chromosomal versus autosomal variation differ markedly under different evolutionary scenarios. Selection on standing variation, likely experienced by organisms moving into

Table 2 Comparison of X-chromosomal and autosomal genetic markers

	X chromosome versus autosomes	Reference
Mutation	Autosomal rate: 0.025 per megabase per generation. The X-linked rate is lower, such that an estimate of the ratio of the male mutation rate to the female mutation rate is approximately 3 (although a range has been observed for this ratio).	96
Nucleotide diversity (π/D)	Autosomal loci: 0.083 in HapMap CEU samples. X-linked loci: 0.046 in HapMap CEU samples.	76, 49
r, centimorgans per megabase	Sex-averaged rate on autosomes: 1.13. Female rate of recombination on the X chromosome: 1.14. Sex-averaged-rate on the X chromosome: 0.75 ($\approx 2/3 \times 1.14$, due to absence of recombination in males).	77, table 1
Purifying selection	Selection will affect larger regions on the X chromosome than on autosomes because of differences in linkage disequilibrium; this dynamic also depends on the dominance of deleterious mutations, as X-linked recessive alleles are exposed in males.	150
Background selection	Background selection will purge more variation from the autosomes than from the X chromosome, because deleterious alleles can reach higher frequencies on the autosomes.	11
Positive selection	Adaptation will lower X-linked diversity more than autosomal diversity because of the smaller N_e of the X chromosome.	11
Selection on standing genetic variation	The X chromosome will have a slower rate of adaptation than autosomes if selection acts on standing alleles previously at mutation-selection balance, because of the lower copy number of X chromosomes relative to autosomes.	108
Bottlenecks	X-linked variation will be reduced more than autosomal variation during a bottleneck, but recovery of variation will be quicker on the X chromosome with increased growth rates.	116
Life-history traits	Differing rates across the sexes in adult mortality, development time, and age at reproduction are unlikely to cause a change in the relative diversity of the X chromosome to autosomes, but a high variance of reproductive success for males could, in extreme cases, lead to $N_{e,X} > N_{e,Aut}$.	23

Abbreviation: CEU, Utah residents with ancestry from northern and western Europe (CEPH collection).

new environments, will not reduce X-linked variation as much as autosomal variation (108). However, bottlenecks will strongly reduce X-chromosomal diversity (116), and bottlenecks preceded movement into new environments for human populations under a serial founder model.

Scans for selection on the X chromosome have not been as frequently conducted as on the autosomes, because of the small number of markers available compared with the pooled 22 autosomes and the different threshold for significance needed because of the X's mode of inheritance. An analysis of 16,300 X-linked

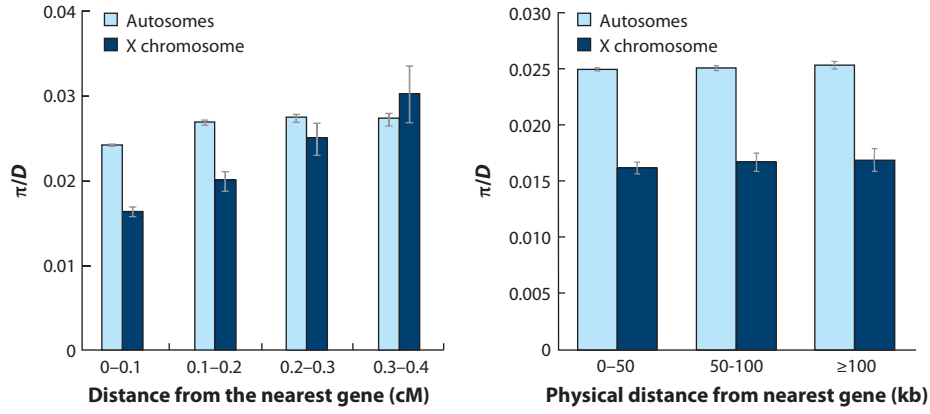


Figure 9

Diversity (π/D) as a function of (*a*) genetic and (*b*) physical distance from genes. Values are shown as means ± standard error. Figure from Reference 49, figure 1 and supplementary figure 3.

SNPs in the HGDP populations found that, after accounting for different effects of drift on the X chromosome and the autosomes, there have been proportionally more events affecting the X chromosome that cause significant allele-frequency changes between continents (19); this could be driven by the effect of both positive selection and serial bottlenecks.

Further, reduced diversity on the X chromosome is correlated with genetic distance from genes (49); in fact, X-linked diversity surpasses autosomal diversity in some cases far from genic regions (**Figure 9**). Although the hemizygous state of the X chromosome in males could explain in part this correlation of diversity and genetic distance along that chromosome, the trend is likely the product of both selection near genes and a genome-wide effect of a higher variance in reproductive success for males compared with females.

Another recent study emphasized that different estimators of effective sex ratios (e.g., estimations based on sequence diversity versus those based on F_{ST}) detect biases in the sex ratio at different timescales (39). The authors found that estimates based on population structure (F_{ST}-based estimates) are sensitive to biases in the sex ratio between two populations after they have diverged. The ability to decouple the competing signatures of different evolutionary forces such as population size changes, gene flow, and selection will be aided by using the X chromosome, where haplotypes are easily accessible, allowing the whole-genome era to truly take advantage of data sources across the genome.

5. DISCUSSION

Here we focus on inference of human population structure and demographic history; for related reviews of the impact of population structure on selected loci and GWAS, see References 102 and 122. It is impressive that just five years after the first HapMap publication, over 50 SNP genotyping studies relevant to human population structure have been published (**Table 1**). Collectively, we estimate these studies have genotyped over 85,000 unique individuals from hundreds of populations across the globe. This rapid application of SNP technology is perhaps a premonition of how quickly whole-genome sequencing will be applied once costs become more affordable. The 1000 Genomes Project and other whole-genome and exome sequencing initiatives underway will help overcome the major challenges involved in curating and analyzing such large data sets, and will hopefully facilitate the rapid, widespread use of genome-wide sequencing.

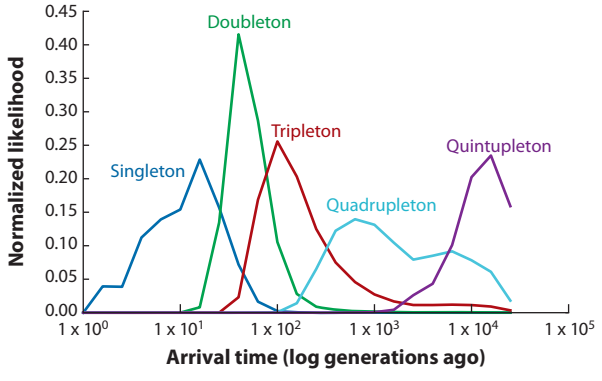
Given the both exciting and daunting scale of the data we are about to be immersed in, there are many future research directions to choose from. Here we suggest four avenues that can begin to unlock the potential of whole-genome data. First, the ability of large sequence data sets to discover rare variants and describe their geographic distribution will elucidate more recently generated human population structure. Second, the ability to carefully identify shared haplotype tracts and the development of novel theory to interpret such tract lengths will yield insights into demographic history and structure. Third, neutral demographic processes can be more deeply understood by studying the geographic distribution of variants that appear to be under recent selection. Finally, we expect more progress to be made in studying modifiers to gene flow such as topographical barriers and historical contacts between populations.

Rare variants have the potential to unravel much more recent gene flow patterns than the common variants that have been assayed by SNP technologies. The geographic distribution of a variant is determined by the patterns of gene flow during the time span since the allele arose. Rare variants are on average much younger than more common variants—using human population parameters, it has been estimated that alleles found in 1 out of 1,000 individuals have an average age of 22 generations, or roughly 550 years (142; **Figure 10**). Whereas common variants give us a picture of gene flow patterns time-averaged over several millennia, rare variants will allow us to investigate patterns on the timescale of centuries (**Figure 11**). Thus, human population genetics will be able to shift its focus from ancient population movements that colonized the globe to events during the timeframe of the historical record, potentially enhancing the intersection of fields studying human history. For example, a recent resequencing study of two genes in 14,000 individuals has shown how historical-era, superexponential rates of population growth can be inferred from rare variants (30).

Rare variant: a base pair polymorphism with a frequency between 1% and 5% in the populations studied

As described above, the use of genome-wide markers has allowed us to conceptualize genetic similarity in a new form. Rather than considering the pairwise nucleotide sequence divergence at kilobase scales (as was the custom with sequence loci), whole-genome data (whether SNP or complete sequence) makes it possible

a



b

Frequency	Generations	Years
1 in 100,000	11	275
1 in 10,000	16	400
1 in 1,000	22	550
1 in 100	32	800

Figure 10

The expected age of rare variants. (*a*) Normalized likelihoods for the arrival time (i.e., expected age) of a singleton (*blue*), doubleton (*green*), tripleton (*red*), quadrupleton (*cyan*), and quintupleton (*purple*), based on a sample of approximately 15,000 individuals and a model of superexponential population growth. Figure from Reference 30. (*b*) Theoretical calculations for the expected age of alleles in a two-stage model configured to represent rapid human growth ($r = 0.4$ during the last 25 generations, 0.001 more ancestrally). Table based on Reference 141.

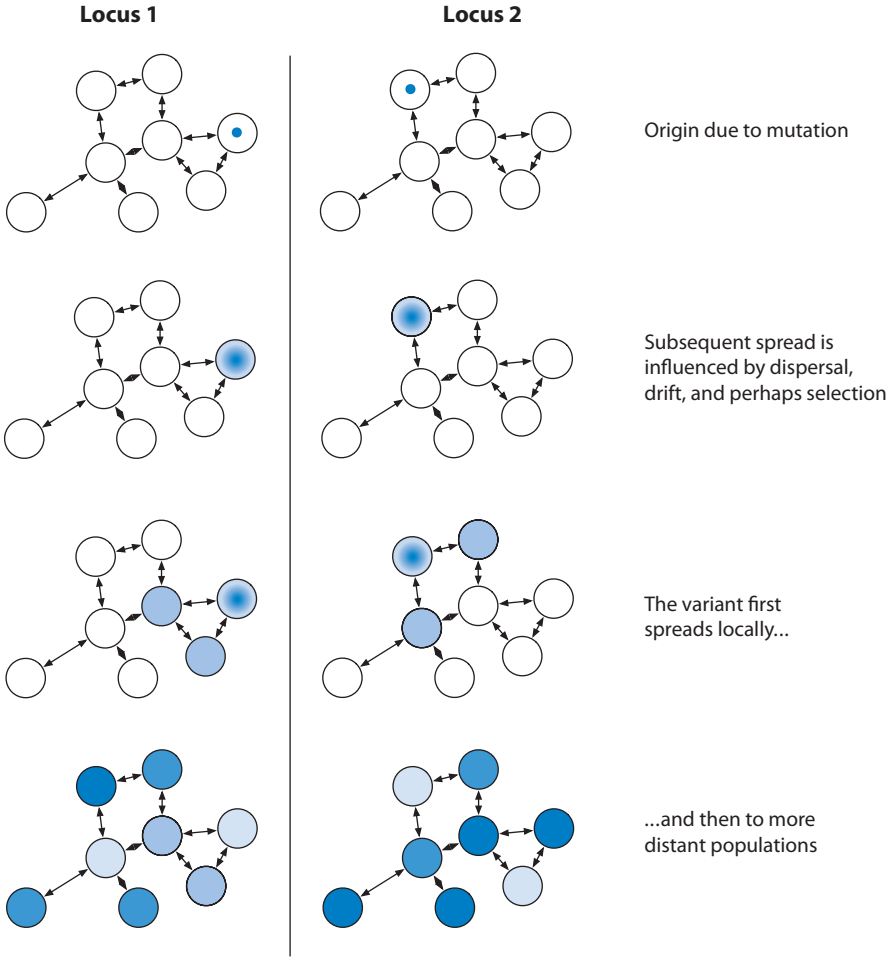


Figure 11

The expected geographic distribution of rare variants. Rare variants are typically due to novel mutations that spread first locally and then to more distant populations. During the early phases of their spread, they are particularly indicative of dispersal patterns between populations. Circles indicate subpopulations, arrows represent gene flow, and shades of blue indicate allele frequencies (with darker shades of blue indicating higher allele frequencies).

to identify shared haplotype tracts at 100 kb to megabase scales. These tracts will be detected more reliably with complete sequence data, and the distribution of their lengths will provide insight into recent population structure. The lengths of shared haplotype tracts are reduced by recombination, and therefore long tracts can be informative of recent common ancestry—patterns of ancestry that can be difficult to detect by analyzing patterns of pairwise nucleotide sequence divergence alone. As mentioned previously, methods for analyzing such data are in their infancy, but we expect development in this area to expand and provide exciting insights.

We have intentionally focused our review and discussion on genome-wide SNP patterns rather than on patterns at putatively selected loci. Nonetheless, we expect that the careful analysis of the geographic distribution of selected loci will provide novel insights into neutral processes that generate population

structure. Analyses of putatively selected loci in the HGDP suggest that selection in humans has been weak enough that most selected loci are strongly influenced by patterns of dispersal and migration (29, 113). Selected loci are on average younger than neutral alleles of the same frequency, suggesting the potential for unique insights into dispersal and migration processes for more recent timescales. For example, SNP markers with high F_{ST} between Asia and Europe (loci that likely have been adaptively diverged) show clines in central Asia indicative of secondary contact between ancestral populations (29).

In many of the empirical studies of human population structure reviewed here, genes mirror geography closely, yet these same SNP studies reveal slight distortions from this gene-geography correlation, suggestive of historical and standing modifiers of gene flow. Some of these might align with major topographic features. Genotypic similarity could also arise from shared cultural traits and historical relationships; for example, in West Africa, Bantu populations and other Niger-Kordofanian populations cluster separately based on genotypic data (16). Two methods to infer genetic boundaries from allele-frequency distributions are wombling and Monmonier's algorithm [both reviewed by Barbujani (7) and Manel et al. (87)], which search for zones where the slope of various allele-frequency surfaces is high; methods like these might increasingly be used to uncover the genetic signatures of recent historical events.

As population genomics brings the recent history of human population structure into focus, we expect to see new signatures that previous analyses could not detect. For example, there will be increasing cases where assuming infrequent gene flow between populations postdivergence is no longer compatible with the data. At the other extreme, we expect there to be increasing cases where slightly diverged subpopulations become distinguishable. As a result, we will have the ability to observe admixture among less-differentiated populations, and perhaps as a result a larger fraction of human genomes will be recognized as admixed mosaics. We also expect that quantifying the abundance of rare variants will lead to a new perspective on human similarity. It has long been appreciated that, relative to other primates, humans are very similar to each other genetically (e.g., 72). For example, a pair of human sequences is more similar on average than those of two chimpanzees or two orangutans. If the exceptional population growth of humans has led to an excess of rare variants across the genome (30), it may help us realize that variation in humans is disproportionately unique—variants in humans are more likely to be private to a single person than variants in another primate. Interestingly, then, among primates, humans may be uncommonly similar to one another and yet, in another sense, uniquely unique.

One motivation for studying human population structure is that it fulfills a fascination to understand our origins; however, the global and fine-scale structure of human genetic variation is also an important framework for building new personalized approaches to medicine. In addition, humans are now the model species for purely observational population genetics. The studies reviewed here reveal the complications and opportunities for learning about the evolutionary history of other organisms.

As a result of so much progress in human population genetics, there is increasing and well-warranted public interest in human genetic variation. Aligning with previous authors (e.g., 37, 83), we feel that, while human ancestral origins are fascinating, an individual's origins should not form the basis of political judgments or categorizations; nor, in this era of direct-to-consumer ancestry testing, should they form the basis of notions of self-worth or exclusivity. Instead, we hope and expect that, as whole-genome sequencing becomes a mainstream technique for population-genetic research, we will reach new frontiers and circle back to old questions, achieving more wisdom in keeping with the name we gave ourselves—*Homo sapiens*.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank Chaolong Wang, Darren Kessner, Daniel Wegmann, Aravinda Chakravarti, and one anonymous reviewer for comments on a draft of this manuscript.

LITERATURE CITED

1. Adams A, Hudson R. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168:1699–712
2. Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* 27:2534–47
3. Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1655–64
4. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58
5. Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, et al. 2010. Abraham's children in the genome era: Major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern ancestry. *Am. J. Hum. Genet.* 86:850–59
6. Auton A, Bryc K, Boyko A, Lohmueller K, Novembre J, et al. 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* 19:795–803
7. Barbujani G. 2000. Geographic patterns: how to identify them and why. *Hum. Biol.* 72:133–53
8. Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* 17:1505–19
9. Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. USA* 98:4563–68
10. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, et al. 2010. The genome-wide structure of the Jewish people. *Nature* 466:238–42
11. Betancourt A, Kim Y, Orr H. 2004. A pseudohitchhiking model of X versus autosomal diversity. *Genetics* 168:2261–69
12. Biswas S, Scheinfeldt LB, Akey JM. 2009. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am. J. Hum. Genet.* 84:641–50
13. Blum M, Jakobsson M. 2011. Deep divergences of human gene trees and models of human origins. *Mol. Biol. Evol.* 28:889–98
14. Bosch E, Laayouni H, Morcillo-Suarez C, Casals F, Moreno-Estrada A, et al. 2009. Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: Most population isolates do not show increased LD. *BMC Genomics* 10:338
15. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST. 2010. Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proc. Natl. Acad. Sci. USA* 107:16222–27
16. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, et al. 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. USA* 107:786–91
17. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, et al. 2010. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc. Natl. Acad. Sci. USA* 107(Suppl. 2):8954–61
18. Cann H, de Toma C, Cazes L, Legrand M, Morel V, et al. 2002. A human genome diversity cell line panel. *Science* 296:261–62

19. Casto A, Li J, Absher D, Myers R, Ramachandran S, Feldman M. 2010. Characterization of X-linked SNP genotypic variation in globally distributed human populations. *Genome Biol.* 11:R10
20. Cavalli-Sforza LL. 2010. Interview with Luigi Luca Cavalli-Sforza: past research and directions for future investigations in human population genetics. *Hum. Biol.* 82:245–66
21. Cavalli-Sforza LL, Feldman MW. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* 33(Suppl.):266–75
22. Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes*. Princeton, NJ: Princeton Univ. Press
23. Charlesworth B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. *Genet. Res. Camb.* 77:153–66
24. Chaubey G, Metspalu M, Choi Y, Mägi R, Romero IG, et al. 2011. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. *Mol. Biol. Evol.* 28:1013–24
25. Chen J, Zheng H, Bei J-X, Sun L, Jia W-H, et al. 2009. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am. J. Hum. Genet.* 85:775–85
26. Clark AG, Hubisz M, Bustamante C, Williamson S, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15:1496–502
27. Clark AG, Wang X, Matise T. 2010. Contrasting methods of quantifying fine structure of human recombination. *Annu. Rev. Genomics Hum. Genet.* 11:45–64
28. Conrad D, Jakobsson M, Coop G, Wen X, Wall J, et al. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* 38:1251–60
29. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al. 2009. The role of geography in human adaptation. *PLoS Genet.* 5:e1000500
30. Coventry A, Bull-Otterson L, Liu X, Clark A, Maxwell T, et al. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* 1:131
31. Davison D, Pritchard J, Coop G. 2009. An approximate likelihood for genetic data under a model with recombination and population splitting. *Theor. Popul. Biol.* 75:331–45
32. DeGiorgio M, Jakobsson M, Rosenberg N. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci. USA* 106:16057–62
33. Deshpande O, Batzoglou S, Feldman M, Cavalli-Sforza L. 2009. A serial founder effect model for human settlement out of Africa. *Proc. R. Soc. B Biol. Sci.* 276:291–300
34. Destro-Bisol G, Jobling MA, Rocha J, Novembre J, Richards MB, et al. 2010. Molecular anthropology in the genomic era. *J. Anthropol. Sci.* 88:93–112
35. De Wit E, Delport W, Rugamika CE, Meintjes A, Möller M, et al. 2010. Genome-wide analysis of the structure of the South African Coloured population in the Western Cape. *Hum. Genet.* 128:145–53
36. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214
37. Edwards AWF. 2003. Human genetic diversity: Lewontin's fallacy. *Bioessays* 25:798–801
38. Eller E. 2009 (2001). Effects of ascertainment bias on recovering human demographic history. *Hum. Biol.* 81:735–51
39. Emery LS, Felsenstein J, Akey JM. 2010. Estimators of the human effective sex ratio detect sex biases on different timescales. *Am. J. Hum. Genet.* 87:848–56
40. Engelhardt BE, Stephens M. 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* 6:e1001117
41. Falush D, Stephens M, Pritchard J. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–87
42. Felsenstein J. 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.* 23:691–700
43. François O, Currat M, Ray N, Han E, Excoffier L, Novembre J. 2010. Principal component analysis under population genetic models of range expansion and admixture. *Mol. Biol. Evol.* 27:1257–68
44. Garagnani P, Laayouni H, González-Neira A, Sikora M, Luiselli D, et al. 2009. Isolated populations as treasure troves in genetic epidemiology: the case of the Basques. *Eur. J. Hum. Genet.* 17:1490–94

45. Garrigan D. 2009. Composite likelihood estimation of demographic parameters. *BMC Genet.* 10:72
46. Gayán J, Galan JJ, González-Pérez A, Sáez ME, Martínez-Larrad MT, et al. 2010. Genetic structure of the Spanish population. *BMC Genomics* 11:326
47. Green R, Krause J, Briggs A, Maricic T, Stenzel U, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328:710–22
48. Gutenkunst R, Hernandez R, Williamson S, Bustamante C. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695
49. Hammer M, Woerner A, Mendez F, Watkins J, Cox M, Wall J. 2010. The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nat. Genet.* 42:830–31
50. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, et al. 2008. Investigation of the fine structure of European populations with applications to disease association studies. *Eur. J. Hum. Genet.* 16:1413–29
51. Hellenthal G, Auton A, Falush D. 2008. Inferring human colonization history using a copying model. *PLoS Genet.* 4:e1000078
52. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl. Acad. Sci. USA* 108:5154–62
53. Hey J. 2010. Isolation with migration models for more than two populations. *Mol. Biol. Evol.* 27:905–20
54. Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–60
55. Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. USA* 104:2785–90
56. HUGO Pan-Asian SNP Consort. 2009. Mapping human genetic diversity in Asia. *Science* 326:1541–45
57. Hunley K, Healy M, Long J. 2009. The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. *Am. J. Phys. Anthropol.* 139:35–46
58. Hunter-Zinck H, Musharoff S, Salit J, Al-Ali KA, Chouchane L, et al. 2010. Population genetic structure of the people of Qatar. *Am. J. Hum. Genet.* 87:17–25
59. Int. HapMap Consort. 2003. The International HapMap Project. *Nature* 426:789–96
60. Int. HapMap Consort. 2005. A haplotype map of the human genome. *Nature* 437:1299–320
61. Int. HapMap Consort. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–61
62. Int. HapMap 3 Consort. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58
63. Int. Hum. Genome Seq. Consort. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
64. Irwin DE. 2002. Phylogeographic breaks without geographic barriers to gene flow. *Evolution* 56:2383–94
65. Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, et al. 2008. The genome-wide patterns of variation expose significant substructure in a founder population. *Am. J. Hum. Genet.* 83:787–94
66. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003
67. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, et al. 2009. Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.* 41:657–65
68. Jobling M, Hurles M, Tyler-Smith C. 2004. *Human Evolutionary Genetics*. New York: Garland Sci.
69. Johnson PLF, Slatkin M. 2008. Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* 25:199–206
70. Joubert BR, North KE, Wang Y, Mwapasa V, Franceschini N, et al. 2010. Comparison of genome-wide variation between Malawians and African ancestry HapMap populations. *J. Hum. Genet.* 55:366–74
71. Jung J, Kang H, Cho YS, Oh JH, Ryu MH, et al. 2010. Gene flow between the Korean peninsula and its neighboring countries. *PLoS One* 5:e11855
72. Kaessmann H, Wiebe V, Weiss G, Pääbo S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat. Genet.* 27:155–56
73. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42:348–54

74. Kang SJ, Chiang CWK, Palmer CD, Tayo BO, Lettre G, et al. 2010. Genome-wide association of anthropometric traits in African- and African-derived populations. *Hum. Mol. Genet.* 19:2725–38
75. Keinan A, Mullikin J, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.* 39:1251–55
76. Keinan A, Mullikin J, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat. Genet.* 41:66–70
77. Kong A, Gudbjartsson D, Sainz J, Jonsdottir G, Gudjonsson S, et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* 31:241–47
78. Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22:768–70
79. Laayouni H, Calafell F, Bertranpetit J. 2010. A genome-wide survey does not show the genetic distinctiveness of Basques. *Hum. Genet.* 127:455–58
80. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, et al. 2008. Correlation between genetic and geographic structure in Europe. *Curr. Biol.* 18:1241–48
81. Lappalainen T, Salmela E, Andersen PM, Dahlman-Wright K, Sistonen P, et al. 2010. Genomic landscape of positive natural selection in Northern European populations. *Eur. J. Hum. Genet.* 18:471–78
82. Leu M, Humphreys K, Surakka I, Rehnberg E, Muilu J, et al. 2010. NordicDB: a Nordic pool and portal for genome-wide control data. *Eur. J. Hum. Genet.* 18:1322–26
83. Lewontin RC. 1972. The apportionment of human diversity. *Evol. Biol.* 6:381–98
84. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–4
85. Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165:2213–33
86. Lohmueller KE, Bustamante CD, Clark AG. 2009. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* 182:217–31
87. Manel S, Schwartz M, Luikart G, Taberlet P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* 18:189–97
88. Marth G, Czabarka E, Murvai J, Sherry S. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166:351–72
89. McEvoy BP, Lind JM, Wang ET, Moyzis RK, Visscher PM, et al. 2010. Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry. *Am. J. Hum. Genet.* 87:297–305
90. McEvoy BP, Montgomery GW, McRae AF, Ripatti S, Perola M, et al. 2009. Geographical structure and differential natural selection among North European populations. *Genome Res.* 19:804–14
91. McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet.* 5:e1000686
92. Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* 201:786–92
93. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, et al. 2011. The history of African gene flow into southern Europeans, Levantines, and Jews. *PLoS Genet.* 7:e1001373
94. Moskvina V, Smith M, Ivanov D, Blackwood D, St. Clair D, et al. 2010. Genetic differences between five European populations. *Hum. Hered.* 70:141–49
95. Myers S, Fefferman C, Patterson N. 2008. Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* 73:342–48
96. Nachman M, Crowell S. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156:297–304
97. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, et al. 2009. Genetic structure of Europeans: a view from the North-East. *PLoS One* 4:e5472
98. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, et al. 2008. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.* 83:347–58
99. Nielsen R. 2010. In search of rare human variants. *Nature* 467:1050–51

100. Nielsen R, Hubisz M, Clark A. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168:2373–82
101. Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158:885–96
102. Novembre J, Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. *Nat. Rev. Genet.* 10:745–55
103. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. 2008. Genes mirror geography within Europe. *Nature* 456:98–101
104. Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* 40:646–49
105. Novembre J, Stephens M. 2010. Response to Cavalli-Sforza interview [*Human Biology* 82(3):245–266 (June 2010)]. *Hum. Biol.* 82:469–70
106. O'Dushlaine C, McQuillan R, Weale ME, Crouch DJM, Johansson A, et al. 2010. Genes predict village of origin in rural Europe. *Eur. J. Hum. Genet.* 18:1269–70
107. O'Dushlaine CT, Morris D, Moskvina V, Kirov G, Int. Schizophr. Consort., et al. 2010. Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* 18:1248–54
108. Orr H, Betancourt A. 2001. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* 157:875–84
109. Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. 2009. Inference of locus-specific ancestry in closely related populations. *Bioinformatics* 25:i213–21
110. Patterson N, Petersen DC, Van Der Ross RE, Sudoyo H, Glashoff RH, et al. 2010. Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.* 19:411–19
111. Patterson N, Price A, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2:e190
112. Pennisi E. 2003. Modernizing the tree of life. *Science* 300:1692–97
113. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–37
114. Plagnol V, Wall J. 2006. Possible ancestral structure in human populations. *PLoS Genet.* 2:e105
115. Pluzhnikov A, Donnelly P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144:1247–62
116. Pool J, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. *Evolution* 61:3001–6
117. Pool J, Nielsen R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* 181:711–19
118. Price AL, Butler J, Patterson N, Capelli C, Pascali VL, et al. 2008. Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* 4:e236
119. Price AL, Helgason A, Palsson S, Stefansson H, St. Clair D, et al. 2009. The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet.* 5:e1000505
120. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–9
121. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, et al. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 5:e1000519
122. Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* 11:459–63
123. Pritchard J, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–59
124. Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15:R159–60
125. Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, et al. 2010. Strong maternal Khoisan contribution to the South African Coloured population: a case of gender-biased admixture. *Am. J. Hum. Genet.* 86:611–20
126. Ramachandran S, Deshpande O, Roseman C, Rosenberg N, Feldman M, Cavalli-Sforza L. 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* 102:15942–47

127. Ramírez-Soriano A, Nielsen R. 2009. Correcting estimators of theta and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics* 181:701–10
128. Reich D, Green RE, Kircher M, Krause J, Patterson N, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1053–60
129. Reich D, Thangaraj K, Patterson N, Price A, Singh L. 2009. Reconstructing Indian population history. *Nature* 461:489–94
130. Rodríguez-Ezpeleta N, Alvarez-Busto J, Imaz L, Regueiro M, Azcárate MN, et al. 2010. High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations. *Hum. Genet.* 128:113–17
131. Romero IG, Manica A, Goudet J, Handley LL, Balloux F. 2009. How accurate is the current picture of human genetic variation? *Heredity* 102:120–26
132. Rosenberg NA, Feldman MW. 2002. The relationship between coalescence times and population divergence times. In *Modern Developments in Theoretical Population Genetics*, ed. M Slatkin, M Veuille, pp. 130–64. Oxford: Oxford Univ. Press
133. 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
134. Rosenberg NA, Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3:380–90
135. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. 2002. Genetic structure of human populations. *Science* 298:2381–85
136. Sabatti C, Service SK, Hartikainen A-L, Pouta A, Ripatti S, et al. 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 41:35–46
137. Sachidanandam R, Weissman D, Schmidt S, Kakol J, Stein L, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–33
138. Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008. Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* 82:290–303
139. Schaffner S, Foo C, Gabriel S, Reich D, Daly M, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–83
140. Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, et al. 2009. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proc. Natl. Acad. Sci. USA* 106:8611–16
141. Slatkin M. 2000. Allele age and a test for selection on rare alleles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 355:1663–68
142. Tallila J, Jakkula E, Peltonen L, Salonen R, Kestilä M. 2008. Identification of CC2D2A as a Meckel syndrome gene adds an important piece to the ciliopathy puzzle. *Am. J. Hum. Genet.* 82:1361–67
143. Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* 28:289–301
144. Teo Y-Y, Sim X, Ong RTH, Tan AKS, Chen J, et al. 2009. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.* 19:2154–62
145. Tian C, Kosoy R, Lee A, Ransom M, Belmont JW, et al. 2008. Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PLoS One* 3:e3862
146. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, et al. 2008. Analysis and application of European genetic substructure using 300K SNP information. *PLoS Genet.* 4:e4
147. Tishkoff S, Reed F, Friedlaender F, Ehret C, Ranciaro A, et al. 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035–44
148. Veeramah KR, Tönjes A, Kovacs P, Gross A, Wegmann D, et al. 2011. Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *Eur. J. Hum. Genet.* In press
149. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
150. Vicoso B, Charlesworth B. 2009. Recombination rates may affect the ratio of X to autosomal non-coding polymorphism in African populations of *Drosophila melanogaster*. *Genetics* 181:1699–701

151. Voight B, Adams A, Frisse L, Qian Y, Hudson R, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci. USA* 102:18508–13
152. Wakeley J. 1996. Pairwise differences under a general model of population subdivision. *J. Genet.* 75:81–89
153. Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K. 2001. The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *Am. J. Hum. Genet.* 69:1332–47
154. Wall J, Lohmueller K, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* 26:1823–27
155. Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, et al. 2010. Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat. Appl. Genet. Mol. Biol.* 9:13
156. Wang S, Lewis CM Jr, Jakobsson M, Ramachandran S, Ray N, et al. 2007. Genetic variation and population structure in Native Americans. *PLoS Genet.* 3:e185
157. Wang Y, Hey J. 2010. Estimating divergence parameters with small samples from a large number of loci. *Genetics* 184:363–79
158. Wellcome Trust Case Control Consort. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–78
159. Wilkins J. 2006. Unraveling male and female histories from human genetic data. *Curr. Opin. Genet. Dev.* 16:611–17
160. Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, et al. 2010. Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* 20:1983–92
161. Xing J, Watkins WS, Shlien A, Walker E, Huff CD, et al. 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* 96:199–210
162. Xing J, Watkins WS, Witherspoon DJ, Zhang Y, Guthery SL, et al. 2009. Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Res.* 19:815–25
163. Xu S, Huang W, Qian J, Jin L. 2008. Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *Am. J. Hum. Genet.* 82:883–94
164. 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.* 83:322–36
165. Xu S, Yin X, Li S, Jin W, Lou H, et al. 2009. Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am. J. Hum. Genet.* 85:762–74
166. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, et al. 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.* 83:445–56

Annual Review of Genomics and Human Genetics

Volume 12, 2011

Contents

Putting Medical Genetics into Practice
Malcolm A. Ferguson-Smith 1

Copy Number and SNP Arrays in Clinical Diagnostics
Christian P. Schaaf, Joanna Wiszniewska, and Arthur L. Beaudet 25

Copy-Number Variations, Noncoding Sequences, and Human Phenotypes
Eva Klopocki and Stefan Mundlos 53

The Genetics of Atrial Fibrillation: From the Bench to the Bedside
Junjie Xiao, Dandan Liang, and Yi-Han Chen 73

The Genetics of Innocence: Analysis of 194 U.S. DNA Exonerations
Greg Hampikian, Emily West, and Olga Akselrod 97

Genetics of Schizophrenia: New Findings and Challenges
Pablo V. Gejman, Alan R. Sanders, and Kenneth S. Kendler 121

Genetics of Speech and Language Disorders
Changsoo Kang and Dennis Drayna 145

Genomic Approaches to Deconstruct Pluripotency
Yuin-Han Loh, Lin Yang, Jimmy Chen Yang, Hu Li, James J. Collins, and George Q. Daley 165

LINE-1 Elements in Structural Variation and Disease
Christine R. Beck, José Luis Garcia-Perez, Richard M. Badge, and John V. Moran 187

Personalized Medicine: Progress and Promise
Isaac S. Chan and Geoffrey S. Ginsburg 217

Perspectives on Human Population Structure at the Cusp of the Sequencing Era
John Novembre and Sohini Ramachandran 245

Rapid Turnover of Functional Sequence in Human and Other Genomes
Chris P. Ponting, Christoffer Nellåker, and Stephen Meader 275

Recent Advances in the Genetics of Parkinson's Disease
Ian Martin, Valina L. Dawson, and Ted M. Dawson 301

Regulatory Variation Within and Between Species
Wei Zheng, Tara A. Gianoulis, Konrad J. Karczewski, Hongyu Zhao, and Michael Snyder 327

The Repatterning of Eukaryotic Genomes by Random Genetic Drift
Michael Lynch, Louis-Marie Bobay, Francesco Catania, Jean-François Gout, and Mina Rho 347

RNA-Mediated Epigenetic Programming of Genome Rearrangements
Mariusz Nowacki, Keerthi Shetty, and Laura F. Landweber 367

Transitions Between Sex-Determining Systems in Reptiles and Amphibians
Stephen D. Sarre, Tariq Ezaz, and Arthur Georges 391

Unraveling the Genetics of Cancer: Genome Sequencing and Beyond
Kit Man Wong, Thomas J. Hudson, and John D. McPherson 407

Indexes

Cumulative Index of Contributing Authors, Volumes 3–12 431

Cumulative Index of Chapter Titles, Volumes 3–12 435

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://genom.annualreviews.org/errata.shtml