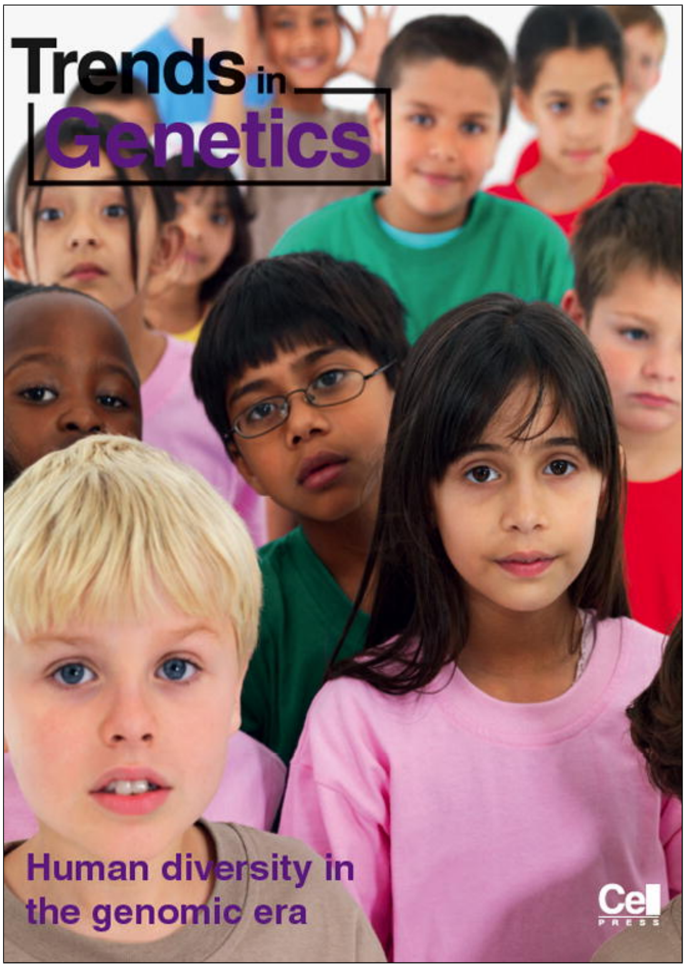

Trends in Genetics
Human diversity in the genomic era
Cell PRESS

Feature Review

Human genome diversity: frequently asked questions

Guido Barbujani and Vincenza Colonna

Department of Biology and Evolution, University of Ferrara, via Borsari 46, 44121 Ferrara, Italy

Despite our relatively large population size, humans are genetically less variable than other primates. Many allele frequencies and statistical descriptors of genome diversity form broad gradients, tracing the main expansion from Africa, local migrations, and sometimes adaptation. However, this continuous variation is discordant across loci, and principally seems to reflect different blends of common and often cosmopolitan alleles rather than the presence of distinct gene pools in different regions of the world. The elusive structure of human populations could lead to spurious associations if the effects of shared ancestry are not properly dealt with; indeed, this is among the causes (although not the only one) of the difficulties encountered in discovering the loci responsible for quantitative traits and complex diseases. However, the rapidly growing body of data on our genomic diversity has already cast new light on human population history and is now revealing intricate biological relationships among individuals and populations of our species.

Some classical questions

Ten years have passed since the White House press release announcing the completion of the first survey of the entire human genome (http://www.ornl.gov/sci/techresources/Human_Genome/project/clinton1.shtml). In that text 'a new era in molecular medicine' was foreseen, characterized by 'new ways to prevent, diagnose, treat and cure disease.' With new complete individual genomes being published almost on a monthly basis, the US National Center for Biotechnology Information (NCBI) reference sequence (in fact, an assemblage of DNAs from five donors) can now be compared with the complete DNA information from sixteen individuals of three continents [1–10]. The technology is developing rapidly and studies of >1 million polymorphisms have appeared by the hundreds.

All this progress notwithstanding, the new era has hardly begun. Many alleles associated with increased risk of disease development have been identified, but turning the rapidly growing wealth of genomic data into a sound picture of the bases of phenotype variation is a different story [11]. For most single-gene diseases, or simple traits such as lactose tolerance [12], we know the nucleotide substitutions accounting for the largest share of the observed variation and we are beginning to understand how sequence variants at modifier genes affect the severity of the symptoms (for example, in sickle-cell anemia [13]). But most diseases are multifactorial, caused by tens or hundreds of genes, often with small phenotypic effects, and by other factors in the environment. Dealing with such levels of complexity requires a large amount of data (and these data are, or will soon be, available), but especially a conceptual framework able to account for thousands of interactions among genetic and nongenetic factors (and we are only beginning to develop it). On a more positive note, however, data on human genome diversity have become so abundant that it makes sense to try to address anew a few classical questions, including some we shall review in this paper (Box 1).

How much human genetic variation?

This is really two different questions – namely how much polymorphism there is in our species' DNA, and how different are the populations of our species. As for the first, most of us are familiar with the notion that more than 98% of the nucleotides in the human and chimpanzee (*Pan troglodytes*) genomes are identical. That notion comes from a classical study of the temperature at which hybrid human–chimp DNA strands disassociate, suggesting that the two species differ, in fact, at 1.76% of their DNA sites [14]. Recent work has essentially confirmed this picture.

Glossary

Balancing selection: a selective process by which multiple alleles are maintained in the population, generally (but not necessarily) by mechanisms of heterozygote advantage.
CEPH: Centre d'Étude du Polymorphisme Humain (Paris, France).
Cline: a gradient of allele frequencies, or of measures of genetic diversity, in the geographical space.
CNV: copy-number variation, due to the presence in different genomes of different numbers of copies of a certain DNA region, the size of which can range from kilobases to several megabases.
Founder effect: the loss of genetic diversity, the random fluctuation of allele frequencies, and the increase of linkage disequilibrium, that occur when a population develops from a small number of founders.
Genetic drift: the random fluctuation of allele frequencies due to the sampling of alleles in populations of small size. This leads to loss of genetic variation within populations, and to genetic divergence of different populations.
HGDP: Human Genome Diversity Panel, a panel of cell lines established at the CEPH (http://www.cephb.fr/HGDP-CEPH-Panel/), comprising 1064 lymphoblastoid cell lines from 54 populations of the world.
Isolation by distance: the asymptotic decline of genetic similarity with geographic distance due to the fact that, on average, rates of gene flow decrease at increasing distances between subpopulations.
Linkage disequilibrium: non-random association of alleles on chromosomes or in gametes.
SNP: single nucleotide polymorphism.
STR: short tandem repeat, or microsatellite.

Corresponding author: Barbujani, G. (g.barbujani@unife.it).

0168-9525/$ – see front matter © 2010 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2010.04.002

Box 1. Analysis of population structure

The structure of a population, namely the genetic relationships among its individuals and/or subpopulations, depends on the relative weights of drift, gene flow, and selection. When reproductive barriers (geographic, behavioral or cultural) subdivide a territory, subpopulations diverge genetically because of drift. Divergence is faster when subpopulations are small or isolated, and is slower when they are large or connected by high rates of gene flow. Different models of natural selection could oppose, or accelerate, the divergence process. The resulting distribution of allele frequencies can be summarized by point statistics (F_{ST} among them) or in more complex manners.

Differences in the mechanisms of hereditary transmission among different components of the genome have an effect upon the information one can obtain from the data. Polymorphism of the Y-chromosome and of the mitochondrial DNA (mtDNA), transmitted by one parent only, evolves through the accumulation of mutations. Because these genome districts are essentially unaffected by recombination they can be used to trace paternal and maternal lineages and build evolutionary trees up to the most recent common ancestor (MRCA). Analysis of these markers is informative on gene flow, and especially on differences between the migratory behavior of males and females, i.e. the phenomena of matrilocality and patrilocality.

MtDNA and the Y-chromosome, however, represent <2% of the genome. To obtain a comprehensive view of the patterns of human diversity there is no alternative to the study of autosomal markers, especially when adaptation cannot be ruled out *a priori*. The results obtained for different kinds of markers cannot be mechanically overlapped because markers differ in their effective population size (*Ne*), and hence in the expected impact of genetic drift upon them. Indeed, in each couple there are four copies of each autosome, but only three X-chromosomes, and one copy of the Y-chromosome and of mtDNA. All other factors being equal, the expected age of the MRCA to a gene genealogy is 2*Ne* generations. Therefore, mitochondrial and Y-chromosome diversity is not only expected to be more heavily affected by genetic drift but also to reflect more recent demographic and evolutionary events than autosomal variation.

The comparison of human and chimpanzee genomes identified 35 million single-nucleotide changes (besides millions of chromosomal rearrangements and insertion/deletion events) [15]. Over an estimated genome length close to 3 billion nucleotides, that figure translates into a rate of single-nucleotide substitution equal to 1.23%. Because 1.06% of these changes appear to be fixed between species, the remaining 0.17% represents the fraction of the human genome occupied by single nucleotide polymorphisms (SNPs). The main genetic differences between humans and other primate species do not seem to depend on point mutations, but on gain or loss of entire genes [16].

Another way to estimate the level of polymorphism in our species is to compare variation between complete or almost complete sequences of human genomes. There is strong concordance for the rate of single nucleotide polymorphism, estimated to be close to 0.1% of the total genome length [1–10]. To mention one example, 4 053 781 SNPs were identified in a Namibian Khoisan's genome [8]. Further studies will doubtless expand the list of polymorphic sites, and hence the estimate based on individual genomes is going to increase and will probably approach the value inferred from human–chimp comparisons.

Box 2. Statistical descriptors of population structure

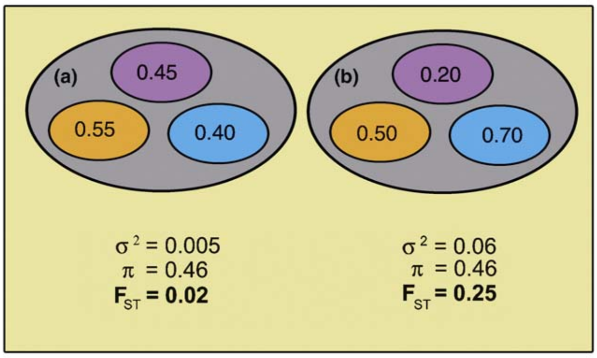
F_{ST}: Wright's fixation index (F_{ST}) can be estimated as the standardized variance of allele frequencies among subpopulations: $F_{ST} = \sigma^2 / \pi(1-\pi)$ with σ^2 and π being the variance and mean, respectively, of the allele frequency. F_{ST} ranges from 0, when all subpopulations are identical, to 1, when different alleles are fixed in different subpopulations. Consider the example in Figure I; the mean allele frequency is the same in both cases, 0.46, but the variance is 10 times as large in B as in A, and so is F_{ST} (Ref. [123] for extensive review). Some F_{ST} estimates for the global human population are given in Table 1.

Principal component analysis: especially when considering many loci, summarizing population structure by means of one or a few indices could imply significant loss of information. By contrast, many loci might contain the same redundant information. Principal component (PC) analysis transforms a number of correlated allele frequencies in a smaller number of uncorrelated synthetic variables, or principal components. The procedure is most useful when much of the information in the data can be summarized by the first few PCs accounting for the greatest fraction of the overall variation [124]. In most applications the first 2 or 3 PCs are plotted in a Cartesian graph or superimposed on a geographic map where gradients and other geographical patterns are easier to recognize.

Model-based genetic clustering: PC analysis emphasizes continuous variation in the geographical space. A popular way to describe discontinuities in population structure is based on an algorithm, STRUCTURE [125], assigning individual genotypes to an arbitrary number of groups or clusters, *k*. Independent analyses are run for a set of *k* values, results are compared across analyses, and coefficients of membership in each inferred cluster are calculated whenever there is evidence of admixture between different clusters.

In practice, each individual genotype (X) is associated to the elements of a Q vector representing the individual's probabilities to belong to each of the *k* clusters. The probability Pr(X|Z, P, Q) of the genotype, given Q, the cluster Z and the allele frequencies in that cluster P, is obtained by constructing a Monte Carlo Markov Chain, with the stationary distribution Pr(Z, P, Q|X). An estimate of the parameters is obtained from their posterior distribution.

Values in the Q vector are graphically summarized as a vertical bars in which each element (i.e. the probability of that genotype to belong to that cluster) is represented by a different color (Figure 2). Substantial genetic structuring is detected when most individuals tend to be assigned to a single cluster (that is, one element of the Q vector is close to 1 and the others are 0). By contrast, there is no genetic structure when individual membership is equally distributed into the *k* clusters, and each Q element is close to1/*k*.



TRENDS in Genetics

Figure I. Schematic representation of allele frequencies in two subdivided populations.

Table 1. Genomic estimates of F_{ST} for the global human population

Number of markers	Samples	F_{ST}	Reference
599 356 SNPs	209 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.13	[22]
1 034 741 SNPs	71 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.10	[22]
1 007 329 SNPs	269 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.12	[23]
443 434 SNPs	3845 individuals distributed worldwide	0.052	[24]
2 841 354 SNPs	210 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.11	[25]
243 855 SNPs	554 individuals from 27 worldwide populations	0.123	[27]
100 *Alu* insertions	710 individuals from 23 worldwide populations	0.095	[69]

Moving to the second question, differences between populations are often summarized by another popular figure, $F_{ST} = 0.15$ (Box 2), and this means that they account for roughly 15% of the species' genetic variance [17–19]. The remaining 85% represents the average difference between members of the same population. One way to envisage these figures is to say that the expected genetic difference between unrelated individuals from distant continents exceeds by 15% the expected difference between members of the same community [20]. Different loci vary greatly in their F_{ST}s, both over the entire genome and in separate analyses of single chromosomes [21,22]. More recent work suggests that the human species' F_{ST} could actually be lower, between 0.05 and 0.13 [23–27] for autosomal SNPs (Table 1), in other words between one-third and one-half of that observed in gorilla (*Gorilla gorilla*; $F_{ST} = 0.38$ [28]) and between Western and Eastern chimpanzee ($F_{ST} = 0.32$ [15]) despite humans occupying a much broader geographic area [29]. In short, not only do humans show the lowest species diversity among primates [30] but are also subdivided into populations that are more closely related than any other primate species, with the possible exception of bonobos (*Pan paniscus*) [18].

As for other kinds of polymorphisms, the F_{ST} for 67 autosomal deletions, insertions, duplications and more complex rearrangements of the genome (collectively referred to as copy number variations, CNVs) in a small set of populations was found to be 0.11 [31], very close indeed to the values estimated for SNPs. Similar levels of population differentiation, around 0.09 or 0.10, were inferred from studies of *Alu* insertions [27,31]. The limited degree of differentiation among human populations does not suggest a history of long-term isolation and differentiation, but rather that genome variation was mostly shaped by our comparatively recent origin from a small number of founders [32,33] who dispersed to colonize the whole planet [34,35].

How is global genetic variation patterned?

Today most people live in cities where individuals of different, and sometimes very different, origins have immigrated recently. To understand the main evolutionary processes it is better to focus on samples of populations that can be assumed not to have changed much in the last few centuries. Genetic variation between these populations is patterned in geographical space and, to a good approximation, can be described as clinal. Starting from the first analysis of the ABO blood group seventy years ago [36], broad gradients of allele frequencies have been repeatedly observed, both at the local level and over large geographical regions. Two recent studies of 783, presumably neutral, short tandem repeat (STR) loci from the Centre d'Étude du Polymorphisme Humain (CEPH) panel [37] showed that there is indeed a strong relationship between geography and various measures of genetic diversity at the worldwide scale [38,39]. Geographic distances between populations (calculated along obligate waypoints that represent likely migration routes within landmasses) proved to predict nicely the respective F_{ST}s [39]; geographic distances from an arbitrary point in East Africa show a high negative correlation with measures of internal population diversity, such as gene diversity and average coalescence time [38]. The main outliers, showing excess genetic divergence from the bulk of the dataset, were populations of South America [39], known to have evolved in extreme isolation, and therefore strongly subjected to drift [40,41]. The best fit was obtained by assuming that the expansion originated in Africa, from a place close to the gulf of Guinea [39], an area where, however, data were lacking.

Several similar studies at a global scale gave broadly consistent results [26,42,43], with the additional observation that linkage disequilibrium increases at increasing distances from Africa [26,42–45]. The crucial role of Africa in human evolution is highlighted by many findings, including (i) the nucleotide differences between two Namibian Khoisans are greater than between European and Asian individuals [8], (ii) genetic differences between African populations are on average greater than those between Africans and Eurasians [46,47], (iii) the alleles found outside Africa are often a subset of the African allele pool [45,48], and (iv) continent-specific alleles or haplotypes are rare in general, but are far more common in Africa than in any other continent [44,49].

Leaving aside the question whether anatomically archaic human forms did [32,50,51] or did not [32,52] leave a small contribution to the contemporary gene pool, geographic distances account for some 75% of the genetic variance between human populations [39]. This strongly suggests that genetic diversity has largely been shaped by phenomena occurring in geographic space – that is, demographic expansions. The peculiar genetic features of Africa point to it as the source of the expansion of a rather small group of founders, probably around 56 000 years ago [38]. Similar and independent estimates of the time of expansion come from a comparison of diversity in populations of humans and of their bacterium *Helicobacter pylori* [53], and from a simulation study in which 50 loci were sequenced in small samples of African, Asian and native American individuals and then compared with the predictions of three demographic models [32]. The model of expanding Africans replacing all pre-existing human forms of Eurasia received overwhelming support over alternative models; extensive analyses based on methods of Approximate Bayesian Computation placed the exit from Africa between 40 000 and 71 000 years ago, and estimated the effective founding population size between 60 and 1220 individuals (intervals of 95% highest posterior density) [32].

In addition, when genetic data are jointly analyzed with cranial measures, Central and Southern Africa appear to represent the putative expansion origins best accounting for the observed patterns [54,55]. In synthesis, neutral genetic markers, and the set of anatomical traits, i.e. cranial measures, which are considered to be less sensitive to natural selection [56], show parallel patterns of variation. Because migration affects all genes equally, whereas selection acts upon specific genome regions, this result seems to imply that traits deviating from the general clinal distribution, both genetic and morphological, could reflect the action of natural selection on local populations.

Serial founder effects or isolation by distance?

There is substantial agreement on the basic pattern of our global biological diversity, but subtle differences as to its interpretation. The observed correlations between genetics and geography are either attributed to a series of expansions accompanied by founder effects [38,39] or to isolation by distance [26], two processes sometimes considered to be equivalent [57]. However, they are not, even though neither includes the effects of selection. In the former case, drift and migration are concentrated, respectively, in episodes of population shrinking and colonization of new territories that are known to generate extensive genetic gradients [58]. Conversely, isolation by distance [59] results from long-term interaction of drift and continuous migration between population units. If $N_e m$, the product of effective population size (N_e) and migration rate (m), is large, the effects of migration prevail and populations tend to converge genetically; if $N_e\ m$ is small, genetic drift plays the predominant evolutionary role, so that populations tend to diverge. Because short-range migrational exchanges are generally more intense between close than between distant populations, m, and thus $N_e\ m$, tend to decrease at increasing spatial distances. In this way, under isolation by distance, genetic similarity between populations declines asymptotically with the geographic distance between them. Therefore, isolation by distance does not generate broad clines but instead produces a patchy distribution of allele frequencies [60] because genetic exchanges are negligible beyond a certain distance. As a consequence, the shape and span of the clines should, in principle, allow one to determine the process that generated them.

In practice, simulation studies based on explicit geographical models have reached different conclusions. One study [61] found substantial agreement between the data and the predictions of a model in which the parameters of a series of founder effects were estimated from the data. By contrast, Hunley and colleagues [35] found that diversity in the 783 loci of the CEPH database is not fully consistent with either isolation by distance (predicting a monotonic decline of genetic similarity with distance) or serial founder effects (predicting a number of population splits of approximately equal effect upon genetic diversity). The closest match between observed and simulated data was found for a model called ‘nested populations’ in which major founder effects occur as humans enter the main geographical regions, with relatively small founder effects and short-range gene flow during the expansions within these regions [35]. Figure 1 is an attempt to summarize early human population history, and its genetic consequences, in a coherent, if necessarily oversimplified, picture, from the time when anatomically modern humans were restricted to Africa, to their dispersal in the Old World and Australia.

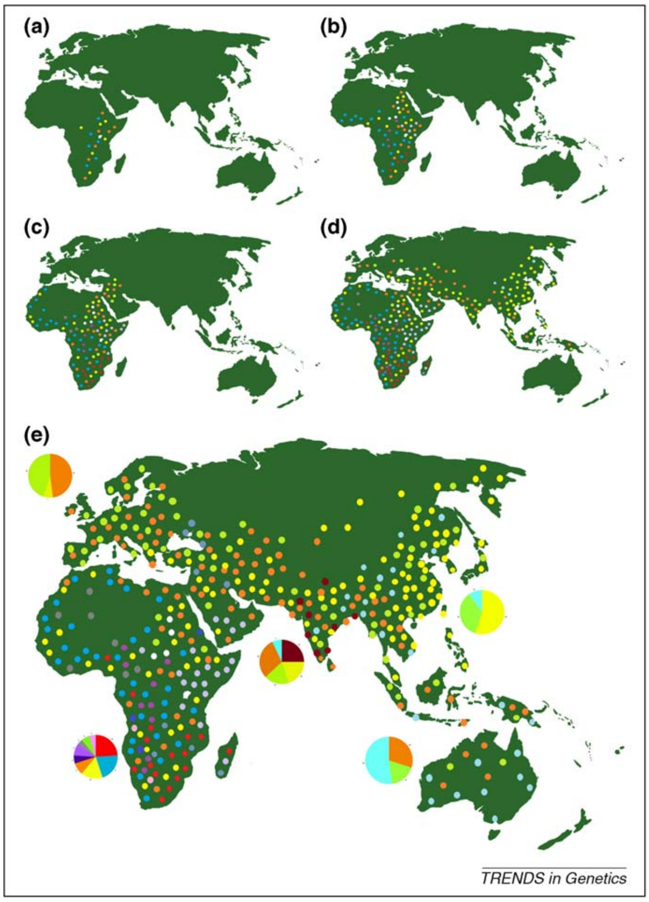


Figure 1. A schematic view of the evolution of human biodiversity. Dots of different colors represent different genotypes, pie charts in panel E represent allele frequencies in five regions at the end of the process. Approximate dates for the five panels: **(a,b)**, >60 000 years before present (BP); **(c)**, 60 000 years BP; **(d)**, 40 000 years BP; **(e)**, 30 000 BP. A broader set of images is available at this site: http://web.unife.it/progetti/genetica/Guido/index.php?lng=it&p=11.

Which are the main human groups?

Is it accurate to assign individuals to discrete geographical groups, thus envisaging our species as essentially discontinuous at the genetic level, or in this way do we misrepresent some aspects of human biodiversity? Since 1972 [19] many independent studies have established that differences between continental populations are small, accounting for less than 10% of the global species variance [20]. That figure holds also for loci under balancing selection, such as the human leukocyte antigens (HLA; 7%) [62], and has been recently confirmed in the analysis of 624 000 SNPs (9%) [26].

By contrast, 10% is little but not zero, and variation in the genome does not comply with the predictions of a pure isolation-by-distance model. Therefore, perhaps reasonably well-defined genetic groups exist in our species, and identifying them would be of potential relevance for biomedical research and clinical practice. To be useful, however, this exercise should produce a consistent list of biological groups, independent of the markers studied.

A popular approach to this question is based on an algorithm, STRUCTURE (Box 2), that assigns individual genotypes to an arbitrary number of groups, k. In the first, and most influential, worldwide analysis based on STRUCTURE, Rosenberg and colleagues [40] typed 377 STRs in the CEPH dataset and recognized six clusters, five of them corresponding to continents or subcontinents, and the sixth to a genetic isolate in Pakistan, the Kalash. In general, individuals of the same population fell consistently in the same cluster, or shared similar membership coefficients in two clusters. The authors concluded that self-reported ancestry contains information on DNA diversity, and hence that an objective clustering of genotypes is possible despite the low between-population variances, if large amounts of data are considered.

In fact, clustering is certainly possible, but is not consistent across studies. In a subsequent paper based on a larger sample of microsatellites [63] the same authors rejected the claim that the distribution of samples in space itself accounts for the apparent population differentiation [43], but failed to confirm the Kalash as a separate unit; instead, the native American populations were this time split in two clusters [63]. The Kalash resurfaced as a distinct group when 15 Indian populations were added to the analysis, leading to the identification of 7 clusters, with most populations of Eurasia now showing multiple memberships [64]. In these studies, all African genotypes formed a single group, in contrast to broadly replicated evidence of deep population subdivision within Africa [35,45,47]. However, when the CEPH 377-marker dataset was analyzed by a different method that searches for zones of sharp genetic change or genetic boundaries [65], Africa appeared subdivided in four groups, and each American population formed an independent group, giving a total of 11 [66].

Previous work based on discriminant analysis had already shown that different clusters emerge when Y-chromosome markers or *Alu* insertions are considered [67]. In a study of more than 500 000 SNPs, STRUCTURE indicated different clusterings if the SNPs were individually analyzed or if they were combined to form haplotypes; in turn, both inferred clusterings were inconsistent with those inferred from CNVs in the same individuals [44]. An independent study of more than 400 000 SNPs in over 3000 individuals [24] identified five clusters that overlap only in part with those listed in previous studies. Finally, it has been repeatedly observed that the genotypes of individuals collected in discrete areas of the world form clear continental clusters for low values of k, but subcontinental structuring emerges at finer levels of analysis [26,27, 40,44].

In practice, when the number of markers is large, many dissimilarities are detected, and a fraction of these are likely to achieve statistical significance. However, minor differences in the markers considered, in the sample distribution, or in the method of analysis, can lead to different clusterings. This should not come as a surprise; more than 80% of human SNP alleles are cosmopolitan: that is, they are present at different frequencies in all continents [44], and the differences between populations and groups represent only a small fraction of the species' diversity. In addition, not all polymorphisms, including most *Alu* insertions and CNV, show the gradients prevailing among SNPs and STRs [44,67–69], and hence should not be expected to reveal an identical population structure. As a consequence, we can cluster people based on any set of polymorphisms, but there is no guarantee that the same clustering will be observed when considering other polymorphisms in the same individuals.

What about skin color?

Wouldn't it be better to simply classify people according to skin color? As a matter of fact, no. With an estimated 70 loci affecting pigmentation, and different metabolic pathways leading to the production of the two main pigments, eumelanin and pheomelanin, skin phenotypes present the challenges of all complex traits [70]. The basic color depends on the proportion of the two main pigments, the size of melanosomes, and their location in the epidermis. Variation has been related to mechanisms of sexual [71] or natural [72] selection, but a general preference for light-skinned partners, predicted by models of sexual selection, is not supported by data [73]. The strong correlation between melanin levels and average UV radiation (UVR) intensity is probably due to geographically variable selection, because melanin protects against excess UVR, but hinders vitamin D synthesis when UVR is low [74]. People living at the same latitude and subjected to similar selective regimes (e.g. Europeans and East Asians) have similar skin color, but that color is the product of convergent evolution in which different mutations determine similar phenotypes [75]. Therefore, clusters of people with similar skin colors would include individuals with very different origins and genotypes.

How is human genetic variation patterned in India and Europe?

In short, so far there has been little success in attempts to define human biological groups on genetic grounds. Exploring variation in specific geographical regions in greater detail might help us understand why. Compare, for instance, the results of two studies of genetic structure in Europe (10 000 SNPs [76]) and South India (240 000 SNPs [27]). In both cases, most individuals showed evidence of multiple origins: that is, membership in more than one cluster. However, the European structure seems to largely depend on the choice of k, with various patterns emerging, only some of them clinal in the geographical space. The clearest picture was obtained for $k = 3$, but higher k values suggested the existence of more complicated structure, probably reflecting extensive historical gene flow in Europe. Conversely, tribal and caste (non-tribal) Indian populations could be distinguished for $k = 2$, and for $k = 3$ lower and upper caste individuals were separated in distinct clusters [27].

India is perhaps the best example of how genetic stratification might arise in response to various kinds of reproductive barriers, social in this case. That stratification might not be very strong, and indeed previous analyses of smaller datasets had failed to detect it [64,77]. However, the point is that, when investigated at the proper scale, population structure appears very complex in India, different but still complex in Europe, and equally reluctant to be described in simple terms in other regions that we do not have the space to consider here. Complexity is the rule, not the exception; if archaeology and demographic history were not sufficient, genomic evidence is there to demonstrate it. Specifically, in India genetic variation seems to reflect not only geography [77,78], language [79] and the layers of the caste system [78,80], but also admixture between rather distinct gene pools, one of them more ancient, and the other associated with the origins of agriculture [81], or with (and this might or might not be the same thing) the arrival of the first Indo-European speakers [80,82]. The average differences between Tamil Nadu and Andhra Pradesh populations, 500 km away, appear to be 1/7 to 1/8 of the differences existing between castes of the same region [83], a clear example of how fragmentation along cultural, religious or social boundaries contributes to maintaining extensive variation within populations.

Clear differences between tribal and caste populations were observed at the Y-chromosome level and much less for mitochondrial DNA [84]. This result was not replicated by Watkins and collaborators [83] who instead found a general correlation of social and genetic distances, and regardless of whether the latter were inferred from autosomal, mitochondrial, or Y-chromosome variation (Box 1). Therefore, the extent to which the rigid endogamy prescribed by the caste system results in long-lasting reproductive isolation requires further evaluation. However, it is logical to speculate that in the approximately three millennia elapsed from the origin of the caste system, reproductive behavior was not the same everywhere, and hence the strength of the isolation (and its genetic effects) varied in different areas of the subcontinent.

Methods of principal component (PC) analysis (Box 1) were recently put under severe scrutiny [85], but in Europe they seem able to capture elusive aspects of population structure that can escape model-based analyses, such as those run with STRUCTURE. Two parallel studies, based on half a million SNPs and on partly overlapping samples (661 in common over 2514 [86] and 3192 [87] samples), showed that despite smaller between-population differences than in India, a clinal structure exists in Europe, with increased internal diversity and lower linkage disequilibrium in the South [86,87]. By plotting the individual genotypes in a coordinate system based on the first two PCs, the authors obtained maps broadly mirroring the individuals' geographic origins (at least for 1387 of the initial 3192 subjects, i.e. those who did not have grandparents from different regions). Based on these PCs, individuals could be assigned to a likely place of origin, and this proved to be not too distant ($\leq$400 km on average) from their actual grandparental origin [87]. However, the first two PCs jointly accounted for 0.45% of the variation in the data [87], implying that 99.55% of the variation in the genome is distributed otherwise. In interpreting studies of human population structure one should keep in mind both the existence of patterns in the data, and the fact that these patterns often explain only a small proportion of the data variance.

Selection, or isolation and migration?

Finding evidence of selection in genetic data has been, and still is, a very complicated task. No matter how sophisticated, all available methods suffer from the fact that almost any genetic pattern expected under selection can also be produced by a combination of events in demographic history. To cite only a recent example, a model of population subdivision in the ancestral African population before the expansion [88] proved able to account for patterns of variation initially interpreted as a strong signal of positive selection outside Africa [89].

Popular strategies to detect the effects of natural selection are based the classical notion [90] that selection affects specific loci whereas the expected impact of migration and drift is the same all over the genome. Therefore, stabilizing and diversifying selection should result in lower and higher than average population diversity, respectively. Extremely high F_{ST} values in or near coding regions should thus reveal phenomena of local adaptation. By applying this methodology, patterns suggesting positive selection in Africa, Western Eurasia or East Asia were detected [21,25] and lists of loci potentially contributing to disease-related variation were compiled [25].

A drawback of this approach lies in the fact that it takes time for both neutral and selected alleles to spread from their population of origin. If the new allele is not eliminated by chance, the dynamics of F_{ST} will depend on population sizes and rates of gene flow until the equilibrium distribution is reached. One way to bypass this problem is by more accurate modeling of neutral expectations, taking demography and migration into account [91,92]. Another is to resort to comparisons across species [93]. In this case one looks for genome regions where the human branch of the evolutionary tree contains an increased number of substitutions, an expected consequence of positive selection [94]. This type of approach tends to identify regulatory regions or transcription-associated regions in the vicinity of genes [95,96]. Another common strategy is to infer positive selection from the extension of stretches of DNA in strong linkage disequilibrium [97], and this approach has proved informative in a number of studies (e.g. Refs [98–101]).

For the purposes of this paper, the key question is whether patterns of global diversity reflect geography, and therefore can be explained by a neutral model of human dispersal from Africa. The answer seems to be a cautious 'yes'. Indeed, whole genome surveys and analyses of variation at candidate loci point to genetic differences among populations, associated with differences in a variety of factors in the environment including pathogens, climate and diet. Sometimes there is a nice correspondence between the expectations of a model of positive selection and the genetic patterns identified at specific loci, such as that

Box 3. A schematic history of human classification

Early attempts to classify the main biological groups of humankind can be traced back to the Bible or to the Egyptian book of the dead, but taxonomies with a scientific basis were published only from the 18th century. A common misconception is that major subdivisions of humankind could be safely identified by the analysis of morphological traits such as skin color or cranial measures. On the contrary, no two racial catalogs proposed are entirely consistent (Table I) and the number of items has increased with time. As European and North American explorers encountered populations previously unknown to them, and in the impossibility to fit them into the existing catalogs, the lists of races were repeatedly expanded, sometimes exceeding the 100 mark. Many such catalogs included taxonomic categories of higher or lower order, i.e. super-races or subraces.

Many police services define the ethnicity of a person according to lists resembling these racial catalogs (the US and UK lists are reported in Table I). Clearly, these are not multi-purpose descriptions of human biological diversity but are instead sets of groups that, for reasons that have little to do with biology, have been of interest to the police (for example, the Irish in the UK). This is also shown by the different number and definition of items in either list. Note that before 2005 another list was used in the UK, comprising 5 groups (European, Middle Easterner, Indian subcontinent, Afro-Caribbean, Southeast Asian) with only one of them (Afro-Caribbean) roughly corresponding to one of the 6 US groups (African American).

Table I. A scheme of the main racial catalogs compiled from Refs [118,126–128]

Author (year)	Number of races	Races or groups proposed
Linnaeus (1735)	6	Europaeus, Asiaticus, Afer, Americanus, Ferus, Monstruosus
Buffon (1749)	6	European, Laplander, Tatar, South Asian, Ethiopian, American
Kant (1775)	4	White, Black, Hun (or Kalmuck, or Mongol), Hindustani
Blumenbach (1795)	5	Caucasian, Mongolian, Ethiopian, American, Malay
Cuvier (1828)	3	Caucasoid, Negroid, Mongoloid
Deniker (1900)	29	Ten of which European[a]
Museum of Natural History, Chicago (1933)	105	http://en.wikipedia.org/wiki/Malvina_Hoffman[a]
Von Eickstedt (1937)	38	[a]
Garn and Coon (1955)	>30	Caucasian, Northeastern Asian, African, North American, South American, Micronesian/Melanesian, Polynesian, Pitcairn islanders, Tristan da Cunha, Cowrie-shell Miao, Lolos, Tasmanians, British colored, plus 'an indeterminate number' bringing the total to more than 30
Biasutti (1959)	53	[a]
Coon (1962)	5	Caucasoid, Mongoloid, Capoid, Congoid, Australoid
US Office of Management and Budget (1997)[b]	6	American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, White
Metropolitan Police Service, London (2005)[c]	16	W1 White British, W2 White Irish, W9 Other White background, M1 Mixed White and Black Caribbean, M2 Mixed White and Black African, M3 Mixed White and Asian, M9 Other Mixed background, A1 Asian Indian, A2 Asian Pakistani, A3 Asian Bangladeshi, A9 Any other Asian background, B1 Black Caribbean, B2 Black African, B9 Other Black background, O1 Chinese, O9 Other Ethnic group

[a] Not reported in the original sources or too many to list here.

[b] http://www.whitehouse.gov/omb/rewrite/fedreg/ombdir15.html.

[c] http://www.gmp.police.uk/mainsite/0/58334145D4842BCF802573FE004CEBCA/$file/Ethnicity%20Classifications.pdf.

determining lactase persistence in adults [102]. But when there is polymorphism selection is generally weak, and so the fate and distribution of even non-neutral alleles mainly appear to reflect neutral processes and population history rather than their selective regime [21,103].

How does all this relate to the debate on biological races in humans?

The US National Human Genome Research Institute stated that although genetic differences among human groups are small, these differences nevertheless can be used to situate many individuals within broad, geographically based groupings [104]. As we have seen, this is true, but it is also true that such groups are highly unstable; no consensus has ever been reached on their number and definition. We do not know if groups are difficult to tell apart because admixture has blurred their boundaries, or if these boundaries never really existed. However, we know that there are many contrasting catalogs of human genetic groups, and our inability to agree on one is part of a broader historical failure to compile the catalog of human races (Box 3). Starting in the 18th century with Linnaeus, such catalogs have contained anything between 2 and 200 items [73], an incongruence that Charles Darwin noticed, concluding that human races graduate into each other, and it is hardly possible to discover distinctive characters between them.

The genetic data accumulated since Darwin's time have not changed the substance of the question. For humans to be divided sensibly into groups, genetic changes in distinct traits must occur together at the group boundaries; this is not the case. Discrete genetic groups form in isolation, and hence we must conclude that there has not been enough isolation in our species' history [34]. As a consequence of these processes our genomes appear to be mosaics (Figure 2), with ancestry from many parts of the globe, and African diversity largely encompassing the diversity observed in other continents [8,35,47].

Careful analyses have demonstrated that definitions of racial and/or ethnic variables in biomedical research are inconsistent, and are based on mixtures of biological, social and economical criteria [105]. It is unclear whether there

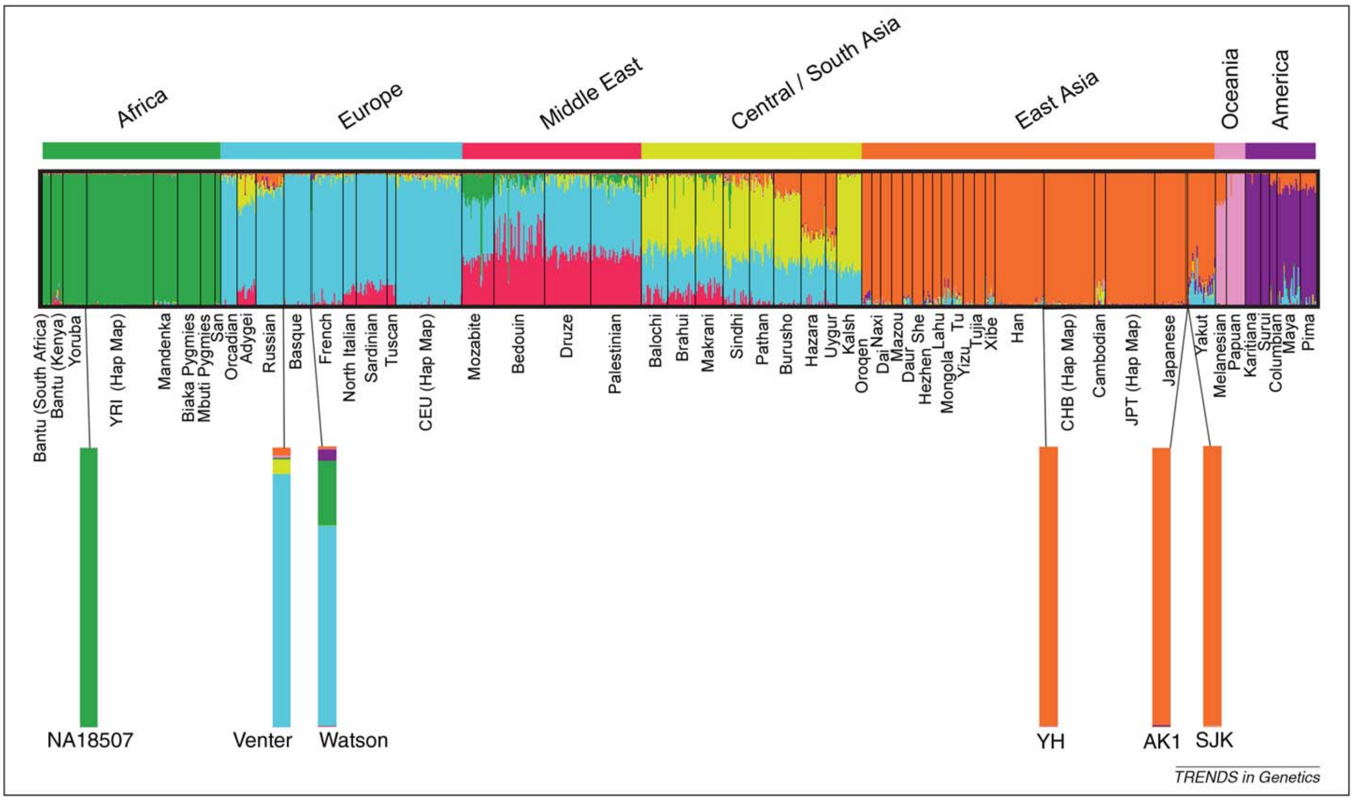


Figure 2. Six fully sequenced human genomes in the context of worldwide variation. The program STRUCTURE (Box 2) was used to assign individuals from the HGDP-CEPH and HapMap panels and six individual genomes to seven genetic clusters. Each individual's genotype is represented by a thin bar in the middle panel and is divided into colors corresponding to the inferred ancestry from the seven genetic clusters ($k = 7$) of the upper panel. At the bottom are the genomes of six individuals, namely a Yoruba (NA18507), two US citizens of European origin (Venter and Watson), a Han Chinese (YH) and two Koreans (AK1, SJK). Reproduced with permission from Ref. [122]. CEU, US residents of Western European ancestry, Utah, USA; CHB, Han Chinese, Beijing, China; YRI, Yoruba, Nigeria; 000; JPT, Japanese, Tokyo, Japan.

might be practical advantages in describing humans as if they were divided into biological races, even though we know they are not [106], but the burden of proof is now on those who say so.

However, in some countries, race is an important factor affecting human interactions and social policies, and this will not vanish just because some scientists say it should. In a sense, races do exist, but only in the sense that the labels we apply to ourselves, and to others, can have practical consequences even if they do not correspond to empirically identifiable biological realities [107]. However, what matters for future research is whether by racial labeling we can approximate what is in a person's genome, and this does not often appear to be the case. For instance, Europeans and Asians appear to be clearly separated in Figure 2, and yet Watson's and Venter's complete genome sequences share more SNPs with a Korean subject (1 824 482 and 1 736 340, respectively) than with each other (1 715 851) [1]. This does not mean that Europeans in general are genetically closer to random Koreans than to each other, but instead highlights the limitations of such coarse categorizations. Populations are indeed structured in the geographical space but, when it comes to predicting individual DNA features, labels such as 'European', 'Asian' and the like are misleading because members of the same group, Watson and Venter in this case, can be very different. Indeed, the only safe way to know what is in a person's DNA is to study that person's DNA, and this is now both feasible and cheap.

What does it mean for the study of the genetic bases of human disease?

In Mendelian disorders, single gene alterations explain almost all occurrences of disease. Conversely, many common disorders are caused by predisposing alleles of multiple genes, each contributing to the individual's risk, and by interactions with factors in the environment.

Under the 'common disease – common variants' hypothesis [108], causative variants of complex traits are posited to have a frequency higher than 5% and exert small additive or multiplicative effects on the phenotype. One has generally no prior information on the position of the causative loci in the genome, and these are sought by comparing the allelic state of many SNPs of known position in affected individuals and in controls. Population structure comes into play at the choice of the control sample because if cases and controls have different ancestries they might differ not only at the loci responsible for the disease, but at many other loci as well. Such stratification, if not properly considered, can result in many false positives [109,110].

Under the alternative 'rare allele' hypothesis [111] the frequency of causative variants is in the range of 0.1–1% and the effect size is expected to be greater than for common variants [112], although not as great as in Mendelian phenotypes. Rare alleles are difficult to detect based on the commonly available microarrays that have been designed to trace common allelic variants [113]. Therefore, the first step is the identification of candidate regions (often by studies of gene expression [114]), followed by resequencing of these regions in cases and controls. Within this framework, potential differences between cases and controls that are unrelated to the trait of interest become less crucial in the identification of loci to be sequenced. The problem persists for the subsequent step where case and control sequences are matched because (i) different alleles of the same locus can produce the same phenotype in different populations (e.g. for skin color), and (ii) rare alleles are generally young and therefore tend to be over-represented in specific populations [111,115].

In reality, both rare and common alleles could contribute to the onset of the same disease, and many intermediate possibilities exist between the two extreme scenarios [113,116]. However, if as seems likely, the genetic factors underlying complex traits are mostly represented by rare alleles [116,117], our current inability to define once and for all the main genetic groups of humankind will not severely hamper our attempts to map the loci involved in the hereditary transmission of disease and disease risk.

What would be an ideal sample of the human species?

Imagine that we have the possibility to collect a sample of 100 human genomes for a multi-purpose description of human genetic diversity. What would be the ideal composition of this sample? One possibility is to superimpose a conveniently spaced grid on the Earth's surface, sampling one individual at each node of the grid. Another is to concentrate the efforts on relatively isolated populations; the latter view prevailed in the Human Genome Diversity Project (HGDP), and the CEPH dataset also partly reflects this strategy. We note that this approach leads one to under-represent the most densely populated areas, and hence to underestimate variation within populations [118]. DNAs sampled at random in, say, Santiago de Chile and Los Angeles, would be much more variable than a collection of sequences of the native Mapuche and Chumash people. If, instead, we wanted a sample representing the current composition of humankind, 18 subjects out of 100 would be Chinese Han and 17 would be from India (versus 4.3% and 0% in the CEPH dataset) whereas native populations from Oceania and the Americas would hardly be sampled (versus 3.7% and 9.6% in the CEPH dataset).

By studying people whose ancestors lived in genetic isolation, in practice one seeks to reconstruct the genetic structure of humankind as it was before the massive migratory movements of the last few centuries. However, no sampling scheme is error-free, and the question becomes what is the best way to reduce the likely error. A solution could lie in the selection of the subjects according to linguistic criteria [119], a time-honored approach [120,121] that emphasizes the role of groups defined by at least one objectively recognizable feature – the language they speak. If the name of the game is to take a picture of the human DNA tree as it was before recent migration messed things up, linguistic affiliations offer a reasonably good approximation. A useful precaution is never to forget classical anthropological wisdom, and to consider in the analysis only people speaking the language of their likely ancestors.

Concluding remarks

In ten years time many questions we cannot currently address will probably find an answer. And, perhaps, new empirical or theoretical findings will show that some of the answers we are now happy about can be improved. For now, it seems evident that the enormous financial and scientific efforts dedicated to the study of the genetic bases of human pathologies have produced significant success in many specific topics, but not yet the general advancement that was expected. Progress has actually been faster in fundamental than in applied science, and our comprehension of the main patterns of human diversity, and of the underlying evolutionary processes, has greatly improved.

Alas, complex traits are complex. The DNA sequence of an individual is a text of which we understand the alphabet (the four bases) and the grammar (the single gene function), but very little of the syntax. To gain a grasp of the syntactic rules (i.e. how genes interact with one another and with many factors in the environment), the development of more sophisticated technical tools will be less crucial than devising new models to make sense of the abundant empirical information that is available.

Acknowledgements

Figure 1 is based on material published at Ken Kidd's website: http://info.med.yale.edu/genetics/kkidd/point.html; Mark Jobling and John Novembre critically read a previous version of this manuscript; François Balloux, Seong-Jin Kim and Chris Tyler-Smith gave us permission to reproduce previously published material; Giorgio Bertorelle, Douglas J. Futuyma, Pascal Gagneux, Oscar Lao, Lorena Madrigal and Mariano Rocchi provided several helpful suggestions; we thank them all.

References

1 Ahn, S.M. *et al.* (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* 19, 1622–1629

2 Kim, J.I. *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature* 460, 1011–1015

3 McKernan, K.J. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19, 1527–1541

4 Bentley, D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59

5 Drmanac, R. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327, 78–81

6 Levy, S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.* 5, e254

7 Lupski, J.R. *et al.* (2010) Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* 362, 1181–1191

8 Schuster, S.C. *et al.* (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463, 943–947

9 Wang, J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature* 456, 60–65

10 Wheeler, D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876

11 Goldstein, D.B. (2009) Common genetic variation and human traits. *N. Engl. J. Med.* 360, 1696–1698

12 Tishkoff, S.A. *et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39, 31–40

13 Lettre, G. *et al.* (2008) DNA polymorphisms at the *BCL11A*, *HBS1L-MYB*, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl. Acad. Sci. U. S. A.* 105, 11869–11874

14 Sibley, C.G. and Ahlquist, J.E. (1984) The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. *J. Mol. Evol.* 20, 2–15

15 The Chimpanzee Sequencingand Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87

16 Hahn, M.W. *et al.* (2007) Accelerated rate of gene gain and loss in primates. *Genetics* 177, 1941–1949

17 Barbujani, G. *et al.* (1997) An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. U. S. A.* 94, 4516–4519

18 Fischer, A. *et al.* (2006) Demographic history and genetic differentiation in apes. *Curr. Biol.* 16, 1133–1138

19 Lewontin, R.C. (1972) The apportionment of human diversity. *Evol. Biol.* 6, 381–398

20 Barbujani, G. (2005) Human races: classifying people vs understanding diversity. *Curr. Genomics* 6, 215–226

21 Coop, G. *et al.* (2009) The role of geography in human adaptation. *PLoS Genet.* 5, e1000500

22 Weir, B.S. *et al.* (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* 15, 1468–1476

23 International Hap Map Consortium (2005) A haplotype map of the human genome. *Nature* 437, 1299–1320

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

25 Barreiro, L.B. *et al.* (2008) Natural selection has driven population differentiation in modern humans. *Nat. Genet.* 40, 340–345

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

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

28 Thalmann, O. *et al.* (2007) The complex evolutionary history of gorillas: insights from genomic data. *Mol. Biol. Evol.* 24, 146–158

29 Stone, A.C. *et al.* (2002) High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proc. Natl. Acad. Sci. U. S. A.* 99, 43–48

30 Kaessmann, H. *et al.* (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat. Genet.* 27, 155–156

31 Redon, R. *et al.* (2006) Global variation in copy number in the human genome. *Nature* 444, 444–454

32 Fagundes, N.J. *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104, 17614–17619

33 Garrigan, D. *et al.* (2007) Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. *Genetics* 177, 2195–2207

34 Cox, M.P. *et al.* (2008) Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. *BMC Genet.* 9, 76

35 Hunley, K.L. *et al.* (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

36 Haldane, J.B.S. (1940) The blood-group frequencies of European peoples and racial origins. *Hum. Biol.* 12, 457–480

37 Cann, H.M. *et al.* (2002) A human genome diversity cell line panel. *Science* 296, 261–262

38 Liu, H. *et al.* (2006) A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* 79, 230–237

39 Ramachandran, S. *et al.* (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. U. S. A.* 102, 15942–15947

40 Rosenberg, N.A. *et al.* (2002) Genetic structure of human populations. *Science* 298, 2381–2385

41 Wang, S. *et al.* (2007) Genetic variation and population structure in native Americans. *PLoS Genet.* 3, e185

42 Conrad, D.F. *et al.* (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* 38, 1251–1260

43 Serre, D. and Paabo, S. (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* 14, 1679–1685

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

45 Tishkoff, S.A. *et al.* (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271, 1380–1387

46 Ramachandran, S. *et al.* (2004) Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. *Hum. Genomics* 1, 87–97

47 Yu, N. *et al.* (2002) Larger genetic differences within africans than between Africans and Eurasians. *Genetics* 161, 269–274

48 Watkins, W.S. *et al.* (2001) Patterns of ancestral human diversity: an analysis of *Alu*-insertion and restriction-site polymorphisms. *Am. J. Hum. Genet.* 68, 738–752

49 Sabbagh, A. *et al.* (2008) Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. *BMC Genet.* 9, 21

50 Garrigan, D. and Hammer, M.F. (2006) Reconstructing human origins in the genomic era. *Nat. Rev. Genet.* 7, 669–680

51 Relethford, J.H. (2008) Genetic evidence and the modern human origins debate. *Heredity* 100, 555–563

52 DeGiorgio, M. *et al.* (2009) Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci. U. S. A.* 106, 16057–16062

53 Linz, B. *et al.* (2007) An African origin for the intimate association between humans and Helicobacter pylori. *Nature* 445, 915–918

54 Betti, L. *et al.* (2009) Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proc. Biol. Sci.* 276, 809–814

55 Manica, A. *et al.* (2007) The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 448, 346–348

56 Betti, L. *et al.* (2009) The relative role of drift and selection in shaping the human skull. *Am. J. Phys. Anthropol.* 141, 76–82

57 Terreros, M.C. *et al.* (2009) Insights on human evolution: an analysis of *Alu* insertion polymorphisms. *J. Hum. Genet.* 54, 603–611

58 Barbujani, G. *et al.* (1995) Indo-European origins: a computer-simulation test of five hypotheses. *Am. J. Phys. Anthropol.* 96, 109–132

59 Wright, S. (1931) Evolution in Mendelian populations. *Genetics* 16, 97–159

60 Relethford, J.H. (2004) Global patterns of isolation by distance based on genetic and morphological data. *Hum. Biol.* 76, 499–513

61 Deshpande, O. *et al.* (2009) A serial founder effect model for human settlement out of Africa. *Proc. Biol. Sci.* 276, 291–300

62 Meyer, D. *et al.* (2006) Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. *Genetics* 173, 2121–2142

63 Rosenberg, N.A. *et al.* (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* 1, e70

64 Rosenberg, N.A. *et al.* (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet.* 2, e215

65 Manni, F. *et al.* (2004) Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. *Hum. Biol.* 76, 173–190

66 Barbujani, G. and Belle, E.M. (2006) Genomic boundaries between human populations. *Hum. Hered.* 61, 15–21

67 Romualdi, C. *et al.* (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res.* 12, 602–612

68 Antunez-de-Mayolo, G. *et al.* (2002) Phylogenetics of worldwide human populations as determined by polymorphic *Alu* insertions. *Electrophoresis* 23, 3346–3356

69 Watkins, W.S. *et al.* (2003) Genetic variation among world populations: inferences from 100 *Alu* insertion polymorphisms. *Genome Res.* 13, 1607–1618

70 Parra, E.J. (2007) Human pigmentation variation: evolution, genetic basis, and implications for public health. *Am. J. Phys. Anthropol. Suppl.* 45, 85–105

71 Aoki, K. (2002) Sexual selection as a cause of human skin colour variation: Darwin's hypothesis revisited. *Ann. Hum. Biol.* 29, 589–608

72 Izagirre, N. *et al.* (2006) A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Mol. Biol. Evol.* 23, 1697–1706

73 Madrigal, L. and Kelly, W. (2007) Human skin-color sexual dimorphism: a test of the sexual selection hypothesis. *Am. J. Phys. Anthropol.* 132, 470–482

74 Chaplin, G. (2004) Geographic distribution of environmental factors influencing human skin coloration. *Am. J. Phys. Anthropol.* 125, 292–302

75 Norton, H.L. *et al.* (2007) Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol. Evol.* 24, 710–722

76 Bauchet, M. *et al.* (2007) Measuring European population stratification with microarray genotype data. *Am. J. Hum. Genet.* 80, 948–956

77 Kashyap, V.K. *et al.* (2006) Genetic structure of Indian populations based on fifteen autosomal microsatellite loci. *BMC Genet.* 7, 28

78 Zerjal, T. *et al.* (2007) Y-chromosomal insights into the genetic impact of the caste system in India. *Hum. Genet.* 121, 137–144

79 Indian Genome Variation Consortium (2008) Genetic landscape of the people of India: a canvas for disease gene exploration. *J. Genet.* 87, 3–20

80 Reich, D. *et al.* (2009) Reconstructing Indian population history. *Nature* 461, 489–494

81 Cordaux, R. *et al.* (2004) Independent origins of Indian caste and tribal paternal lineages. *Curr. Biol.* 14, 231–235

82 Zhao, Z. *et al.* (2009) Presence of three different paternal lineages among North Indians: a study of 560 Y chromosomes. *Ann. Hum. Biol.* 36, 46–59

83 Watkins, W.S. *et al.* (2008) Genetic variation in South Indian castes: evidence from Y-chromosome, mitochondrial, and autosomal polymorphisms. *BMC Genet.* 9, 86

84 Thanseem, I. *et al.* (2006) Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet.* 7, 42

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

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

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

88 Currat, M. *et al.* (2006) Comment on 'Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens' and 'Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans'. *Science* 313, 172; author reply 172

89 Mekel-Bobrov, N. *et al.* (2005) Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. *Science* 309, 1720–1722

90 Cavalli-Sforza, L.L. (1966) Population structure and human evolution. *Proc. R. Soc. Lond. B Biol. Sci.* 164, 362–379

91 Excoffier, L. *et al.* (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* 103, 285–298

92 Nielsen, R. *et al.* (2009) Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* 19, 838–849

93 Pollard, K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121

94 Bush, E.C. and Lahn, B.T. (2008) A genome-wide screen for noncoding elements important in primate evolution. *BMC Evol. Biol.* 8, 17

95 Bird, C.P. *et al.* (2007) Fast-evolving noncoding sequences in the human genome. *Genome Biol.* 8, R118

96 Kudaravalli, S. *et al.* (2009) Gene expression levels are a target of recent natural selection in the human genome. *Mol. Biol. Evol.* 26, 649–658

97 Sabeti, P.C. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918

98 Hawks, J. *et al.* (2007) Recent acceleration of human adaptive evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104, 20753–20758

99 Lopez Herraez, D. *et al.* (2009) Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* 4, e7888

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

101 Voight, B.F. *et al.* (2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72

102 Gerbault, P. *et al.* (2009) Impact of selection and demography on the diffusion of lactase persistence. *PLoS One* 4, e6369

103 Balaresque, P.L. *et al.* (2007) Challenges in human genetic diversity: demographic history and adaptation. *Hum. Mol. Genet.* 16 (Spec No. 2), R134–139

104 Race, E. and the Genetics Working Group (2005) The use of racial, ethnic, and ancestral categories in human genetics research. *Am. J. Hum. Genet.* 77, 519–532

105 Hunt, L.M. and Megyesi, M.S. (2008) The ambiguous meanings of the racial/ethnic categories routinely used in human genetics research. *Soc. Sci. Med.* 66, 349–361

106 Foster, M.W. and Sharp, R.R. (2002) Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome Res.* 12, 844–850

107 Glasgow, J. (2009) *A Theory of Race*, Routledge

108 Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.* 17, 502–510

109 Marchini, J. *et al.* (2004) The effects of human population structure on large genetic association studies. *Nat. Genet.* 36, 512–517

110 Voight, B.F. and Pritchard, J.K. (2005) Confounding from cryptic relatedness in case–control association studies. *PLoS Genet.* 1, e32

111 Bodmer, W. and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* 40, 695–701

112 Gorlov, I.P. *et al.* (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 82, 100–112

113 Manolio, T.A. *et al.* (2008) A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* 118, 1590–1605

114 Hardy, J. and Singleton, A. (2009) Genomewide association studies and human disease. *N. Engl. J. Med.* 360, 1759–1768

115 Keen-Kim, D. *et al.* (2006) Overrepresentation of rare variants in a specific ethnic group may confuse interpretation of association analyses. *Hum. Mol. Genet.* 15, 3324–3328

116 Manolio, T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature* 461, 747–753

117 Maher, B. (2008) The case of the missing heritability. *Nature* 456, 18–21

118 Madrigal, L. and Barbujani, G. (2007) Partitioning of genetic variation in human populations and the concept of race. In *Anthropological Genetics* (Crawford, M.H., ed.), pp. 19–37, Cambridge University Press

119 Ingman, M. *et al.* (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408, 708–713

120 Cavalli-Sforza, L.L. *et al.* (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. U. S. A.* 85, 6002–6006

121 Sokal, R.R. (1988) Genetic, geographic, and linguistic distances in Europe. *Proc. Natl. Acad. Sci. U. S. A.* 85, 1722–1726

122 Yngvadottir, B. *et al.* (2009) The promise and reality of personal genomics. *Genome Biol.* 10, 237

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

124 Patterson, N. *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.* 2, e190

125 Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959

126 Bernasconi, R. and Lott, T.L. (2000) *The Idea of Race*, Hackett

127 Cohen, C. (1991) Les races humaines en histoire des sciences. In *Aux Origines d'Homo Sapiens* (Hublin, J.J. and Tillier, A.M., eds), pp. 9–56, Presses Universitaires de France

128 Garn, S.M. and Coon, C.S. (1955) On the number of races of mankind. *Am. Anthropol.* 57, 996–1001