Review TRENDS in Genetics Vol.23 No.9

Going the distance: human population genetics in a clinal world

Lori J. Lawson Handley[1,4], Andrea Manica[2], Jérôme Goudet[1,3] and François Balloux[1]

[1] Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK
[2] Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK
[3] Department of Ecology & Evolution, Biophore, University of Lausanne, CH-1015 Lausanne, Switzerland
[4] Present address: Molecular ecology and evolution group, Department of Biological Sciences, The University of Hull, Hull HU6 7RX, UK

Global human genetic variation is greatly influenced by geography, with genetic differentiation between populations increasing with geographic distance and within-population diversity decreasing with distance from Africa. In fact, these 'clines' can explain most of the variation in human populations. Despite this, population genetics inferences often rely on models that do not take geography into account, which could result in misleading conclusions when working at global geographic scales. Geographically explicit approaches have great potential for the study of human population genetics. Here, we discuss the most promising avenues of research in the context of human settlement history and the detection of genomic elements under natural selection. We also review recent technical advances and address the challenges of integrating geography and genetics.

Global patterns of human genetic variation

The influence of geography on patterns of genetic variation was recognized in the first half of the twentieth century, when population genetics was still an emerging discipline

Glossary

Approximate Bayesian Computational (ABC) methods: methods that are based on a combination of Bayesian inference and summary statistics. The summary statistics are used to approximate a posterior distribution, which is less computationally intensive than conventional Monte Carlo Markov Chain (MCMC) methods.
Ascertainment bias: bias resulting from selection and characterization of markers in a subset of samples. For example, if markers are chosen on the basis of their high polymorphism in a particular population, their variability might be artificially inflated in this population relative to the total.
Carrying capacity: the maximum number of individuals that can be sustained by local resources.
Cline: the gradual linear change in a character (e.g. allele frequency, within-population genetic diversity or between-population genetic differentiation) with increasing geographic distance.
Friction: relative difficulty in moving through a landscape.
F_{ST}: the correlation of genes drawn at random from each subpopulation [69]. F_{ST} is commonly used as a measure of the degree of genetic differentiation of subpopulations
HapMap Project: an international project (http://www.hapmap.org) that aims to compare genetic sequences of different individuals; these sequences can then be used by researchers to identify common genetic variants in human populations.
HGDP-CEPH cell line panel: a publicly available collection of DNA samples collected by the Human Genome Diversity Project and housed at the Centre d'etude du Polymorphisme Humain (http://www.cephb.fr/HGDP-CEPH-Panel/). The panel consists of 1064 lymphoblastoid cell lines, representing individuals from 54 globally distributed populations. Recently, atypical (i.e. potentially mislabelled or duplicate samples) and first and second order relatives have been identified in the panel and excluded in the three standardized subsets labelled H1048, H971 and H952 respectively [62]. The H971 subset, used to generate Figure 1 in the main text and Figure Ib in Box 1 excludes two atypical individuals and pairs of first-degree relatives. The panel has been used in several recent studies of global human genetic variation [8,10–12,50,52] and has been genotyped at nearly 1000 autosomal microsatellite markers, as well as biallelic polymorphisms [22,36,68].
Hs: gene diversity as defined by Nei (equation 7.39 in Ref. [70]). *Hs* is an unbiased estimator of the mean expected heterozygosity under Hardy–Weinberg equilibrium.
Island model: a model of population structure defined by Wright (1931) [71], which describes gene flow in a subdivided population. Each subpopulation of size *N* sends out and receives migrants to each of the other subpopulations at the same rate (*m*). This contrasts with a stepping-stone model, in which subpopulations send out and receive migrants to their right and left neighbours only.
Isolation-by-distance (IBD): the tendency for most individuals to migrate between neighbouring populations, which results in a smooth increase in genetic differentiation with increasing geographic distance between populations (i.e. a cline).
Multiregional evolution: a model of modern human evolution that proposes that all human populations living today originated in their various continents with archaic human populations always linked by gene flow [30]. Genetic evidence, for example that global diversity is a subset of that found in Africa, does not support this model and therefore the alternative Unique Origin (or ''Out-of-Africa'') model is generally accepted nowadays.
Multi-dimensional scaling: a type of multivariate analysis that allows information to be displayed in two (or more) dimensions
One-dimensional stepping-stone: a model of population structure in which populations are arranged linearly and exchange individuals with their right and left neighbours.
Pharmacogenetics: the study of genetically determined variations in drug response, efficacy and frequency of adverse reactions.
Procrustes analysis (or least-squares orthogonal mapping): a potentially informative alternative to the Mantel test for comparing two sets of data based on matching corresponding points (landmarks or sampling locations) from each of the two data sets.
Selective sweep: the rapid fixation of an allele under strong directional selection. Neutral variants that are physically linked to the selected allele are also swept to fixation (a phenomenon known as genetic hitchhiking). This results in a region of reduced variation on the chromosome relative to other genomic regions.

Corresponding author: Handley, L.J.L. (L.Lawson-Handley@hull.ac.uk).
Available online 25 July 2007.

 0168-9525/$ – see front matter doi:10.1016/j.tig.2007.07.002

Two-dimensional stepping-stone: extension of the one-dimensional stepping-stone model, in which each local population can exchange migrants with its four neighbours.
Unique Origin Model: also referred to as the 'Out-of-Africa' or 'Recent African Origins' model of modern human evolution. This model proposes that anatomically modern humans arose in Africa ~200 000 years ago as a new species [29]. Approximately 100 000 to 60 000 years ago, a small population in East Africa started expanding and eventually spread across the world, replacing all non-African archaic humans [20]. It is often assumed that there was a major bottleneck associated with the exit from Africa.

(e.g. Refs [1,2]). In the 1970s, studies revealed a simple relationship between the frequency of human blood group polymorphisms and geographic location [3,4]. Further analyses revealed even stronger geographic patterns or 'clines' (see Glossary): genetic distance between populations increased with geographic distance at both continental and global scales [5,6], which is characteristic of so-called 'isolation by distance' (IBD) [7] models. Despite these early studies demonstrating the importance of IBD, few studies have applied IBD models to global-scale questions, such as the colonization of the world by anatomically modern humans.

In recent years, the field of human population genetics has greatly benefited from large-scale surveys of human genetic variation at hundreds of loci. In particular, the paper by Rosenberg and colleagues [8] will probably be remembered as a milestone in the field. For the first time, the entire scientific community could access a dataset of nearly 400 autosomal microsatellite markers typed on the Human Genome Diversity Project (HGDP-CEPH) cell line panel [9], a resource of over 1,000 DNA samples from individuals from more than 50 globally distributed populations. Several recent studies have examined patterns of human genetic variation in the HGDP-CEPH panel and other large datasets, using geographically explicit frameworks (e.g. Refs [10–13]), and such studies support the single-locus based patterns of IBD generated by Cavalli-Sforza and colleagues [5]. As well as their obvious relevance to human settlement history, these results have increasing significance in a medical context, for example in pharmacogenetics, where the genetic structure of a population is used as a predictor of the efficacy of drugs or the likelihood of adverse reactions (e.g. Refs [14–16]). Given these recent developments and their implications, we discuss here the relevance of these global clinal patterns for inference in human population genetics and for studies of genomic regions under natural selection. We also examine the technical prospects, challenges and pitfalls of geographically explicit analyses (Box 1) and address whether clines are compatible with other results that have described human genetic variation as 'clustered' (Box 2).

Human genetic variation is mainly clinal

Several groups have confirmed that the genetic differentiation between pairs of populations correlates exceptionally well with the geographic distance separating them (e.g. Refs [10,11,13,17,18]). Relethford [17]demonstrated a remarkably strong pattern of isolation by distance when correlating geographic distance (great circle distances forced through choke points; Box 1) with either variation in cranial morphology or genetic distance (as measured by F_{ST} and estimated from 14 blood polymorphisms in 32 populations and from microsatellite loci in the HGDP-CEPH panel) [8]. This was later confirmed in a larger microsatellite dataset [10], and a similar pattern was uncovered using a more complex method for computing probable colonization routes along landmasses [11,13] (Figure 1a and Box 1 Figure Ia). These studies revealed that geographic distance explains at least 75% of the variance between human populations [10,11,13] (Figure 1a).

An even more striking feature in the microsatellite data is that geographic distance from East Africa (the probable cradle of anatomically modern humans (e.g. Refs [19,20]) explains an impressive 85% of the smooth decrease in gene diversity (*Hs*) within human populations [10,12,13] (Figure 1b and Box 1 Figure Ib). Perhaps even more remarkably, similar clinal patterns can be recovered for variation in human craniometric measurements [21] and even the gut bacterium *Helicobacter pylori,* suggesting that this species has been commensal with human populations since our initial exit from Africa [13]. In each case, no step decrease(s) in genetic diversity were found that could be interpreted as evidence for genetic discontinuities, even at continental boundaries. Moreover, the global distribution of single nucleotide polymorphism (SNP)-based haplotypes and the extent of linkage disequilibrium within populations are also compatible with a mostly clinal pattern of human genetic variation [22]. Taken together, these results illustrate that modern humans have a recent African origin and that there was essentially continuous gene flow over limited distances (including between Africa and Eurasia) during the colonization process. These clines might seem to contrast with work that has described human genetic variation as 'clustered'; however, the important point that we make in Box 2 is that clinal models explain the great majority of the variance. As we demonstrate in the following sections, models based on a simple diffusion process are therefore useful for interpreting patterns of human settlement history and should thus also be useful as null models when testing for selection.

Population genetics inference in a clinal world

Given the strong patterns of IBD outlined above, it is clearly time to move on from thinking about human population genetics in the traditional island model framework, which is still often implicitly assumed. It has, for example, recently been suggested that the Americas were founded by fewer than 80 effective individuals [23]. Although this deduction was based on a highly sophisticated coalescence model that makes mathematical sense and enabled several parameters to be estimated simultaneously, it assumed, rather artificially, that Asia and America constituted two separate, random mating populations and did not consider the geographical relationship between samples within and between these two areas. The model thus assumes that migrants are randomly drawn from all over Asia, as opposed to coming predominantly from neighbouring populations (which would be consistent with an IBD model). The expected diversity in the simulated source population is therefore likely to be an overestimate, meaning that the

Box 1. Geographically explicit analyses: technical problems and potential solutions

Geographic distance between populations

The first step for a geographically explicit analysis is to establish the distances between populations. The simplest approach is to consider the great circle distance between two locations, which is the distance over the shortest, most direct route on a sphere (i.e. 'as the crow flies'). However, in most instances, it is desirable to take into account barriers, such as oceans or mountain ranges, which prevent the free flow of individuals (and thus genes). If clear 'choke points' are present (such as the Suez and Bering straits for movement from Africa to Asia and from Asia to America, respectively), it is possible to simply force any movement between regions through them but allow completely free movement within the regions by using great circle distances [10,17]. A more sophisticated approach that enables addition of more complex barriers is to build up explicit friction matrices that describe how difficult it is to move from one location to another. For small-scale movement, friction matrices can be computed over projected maps (e.g. using the software PATHMATRIX [26]), but for large-scale movement the distortion is far too great, and distances computed onto a spherical referential are far superior [12,24]. The latter approach has been used to model human migrations across the whole globe, forcing movement to occur only on land with an altitude less than 2000 m [11] (Figure Ia).

Statistical frameworks for inference

Once the appropriate distances between populations have been estimated, we need to make inferences about population genetic differences on the basis of their geographic distances. The oldest and most commonly used framework for geographically explicit analysis is to compute the Mantel correlation [59] between matrices. A matrix of pairwise genetic distances between all possible pairs of populations is correlated with a matrix of geographic distances. The significance of the resulting correlation coefficient (called a Mantel coefficient) is best determined through permutations. A similar, but much less common approach is to use multi-dimensional scaling to generate a synthetic configuration of locations based on the pairwise genetic distances, and then match it to the geographic configuration using Procrustes analysis [60].

In a few instances, when there is a specific hypothesis of directional movement from a source, it is possible to look for geographic patterns in a particular genetic variable. Cavalli-Sforza and colleagues [5] pioneered this approach by creating 'synthetic maps' by plotting Principal Component Scores to infer movement of individuals within continents. This approach can be further formalized by explicitly correlating population estimates (e.g. heterozygosity) with distance from the source, a technique successfully applied to the 'out of Africa hypothesis' [10,12] (see main text). We illustrate how such genetic diversity data can be interpolated in the form of a synthetic map in Figure Ib. General Linearized Models (GLMs) could be adopted if a more sophisticated framework is required, such as when testing for non-linear effects of geographic distance or the presence of discrete clusters of populations [55].

Investigating selection

The geographically explicit frameworks described above can be expanded to look at selection. In a nutshell, any selected gene should show both a signal of ancient demography (i.e. the pattern revealed by neutral loci) as well as its own peculiar signature of selection. The methods above can be used to estimate the strength of the demographic (neutral) pattern, and deviations from that pattern can be interpreted as due to selection. In a pairwise framework, we can use partial Mantel tests, which allow us to investigate the link between pairwise genetic distances and pairwise differences in the selecting force (e.g. difference in maximum temperature between locations) after having accounted for geographic distance. This approach has been used extensively, but concerns have been raised on the validity of *P* values associated with permutation tests [61]. With a specific, directional movement hypothesis, it is possible to use a GLM to fit both distance from the source as well as the selective factors. The latter approach successfully confirmed the link between pathogen richness and diversity at the human leukocyte antigen (HLA) class I genes [39].

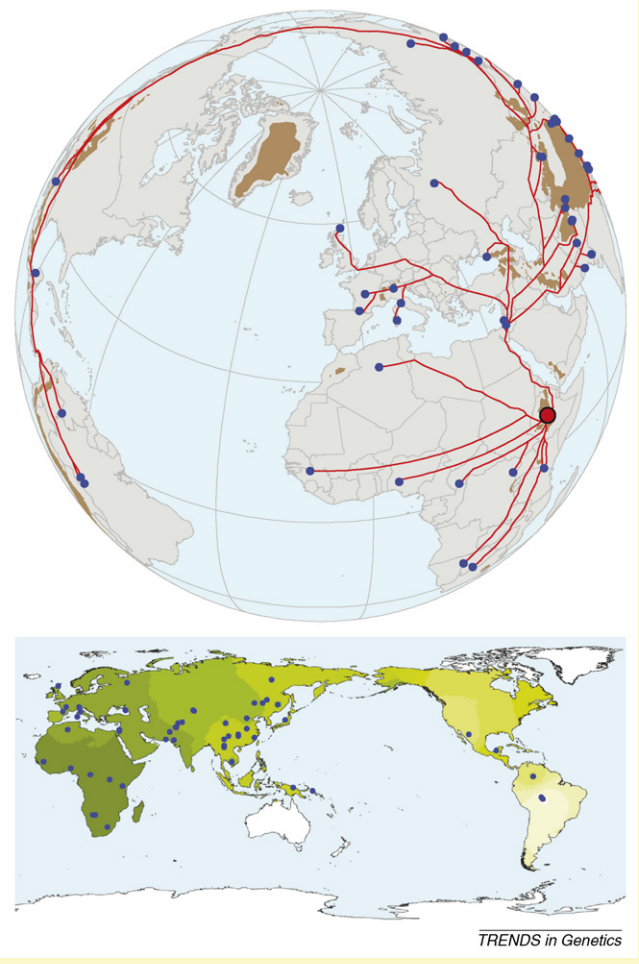

Figure I. **(a)** Estimating geographic distances. The map shows likely colonization routes (red lines) between populations in the HGDP-CEPH panel (blue spots) assuming an origin of modern humans in East Africa (Addis Ababa, red spot). Geographic distances were estimated between populations along the colonization routes using an approach based on graph theory. Routes were forced through landmasses with altitude less than 2000 m (areas over 2000 m are shown by brown shading). Geographic distances from Addis Ababa, as illustrated in this figure, as well as a matrix of pairwise distances between all HGDP-CEPH populations are available as supplementary material online. **(b)** Interpolation of global human genetic diversity. The intensity of the green colour represents the genetic diversity obtained with an inverse distance-weighted (IDW) interpolation method on landmasses using the ArcGIS Spatial Analyst extension. Blue dots represent the 54 populations from the H971 subset of the HGDP-CEPH dataset [62].

true number of colonists needed to explain the diversity observed in the Americas might in fact be much greater.

Several recent papers have based their inference of human settlement history on more realistic, geographically explicit models. The simplest approach takes advantage of the clinal patterns of human genetic variation and the fact that geography explains such a large portion of the variance. For example Ramachandran *et al.* [10] performed a series of simulations to investigate the mode of global colonization of modern humans. They developed an innovative method to locate the most likely starting point of global population expansion, given the pattern of declining heterozygosity with geographic distance in the HGDP-CEPH panel. A model of colonization of the world through a serial founder

Box 2. The 'clines versus clusters' debate

The strong clinal patterns in Figure 1 in the main text seem to be at odds with work that has described human genetic diversity as discontinuous or 'clustered' (e.g. Ref. [8,15]). For instance, using the programme STRUCTURE [63], Rosenberg and colleagues identified six groups of genetically similar individuals ('clusters'), five of which correspond to major geographic regions, suggesting reduced gene flow at continental boundaries [8,10,49].

These two apparently incompatible representations of human genetic diversity led to numerous reanalyses of the HGDP-CEPH datasets and promoted debate on whether human genetic variation could be better described by clusters or clines [11,12,39,49–51,56,64]. STRUCTURE reveals gradients of ancestry proportions even under a model of strict IBD [51,63]. If sampling is heterogeneous (sampling sites are themselves clustered) then the data will reveal genetic clusters that are biologically meaningless [51,63] (see Figure I). Serre and Pääbo [50], investigating this through simulations, argued that the clusters described by Rosenberg *et al.* [8] were caused by the discontinuous nature of the sampling scheme used for the HGDP-CEPH panel and found that, by sampling individuals uniformly across the globe, a picture of continuous, clinal variation emerged. Rosenberg *et al.* [49] subsequently explored several sub-sampling strategies and reached an opposite conclusion: clusters remain even when sampling uniformly across the globe. They suggested these clusters were genuine and attributed their presence to slight discontinuities in the pattern of IBD previously identified [10,11,17], which is consistent with reduced gene flow at geographical barriers such as the Himalayas and Sahara [49,65].

These different representations of human genetic diversity are, however, not mutually exclusive and several authors agree that human genetic diversity can probably be best explained by a synthetic model, in which most of the population differentiation can be explained by IBD, with some discontinuities arising from barriers to dispersal [51,56,66,67]. In other words, human genetic variation might be best explained by a combination of both clines and clusters. However, clusters explain only a minute fraction of the variance [8,49] relative to clines. As mentioned in the main text (Figure 1b), >75% of the total variance of pairwise F_{ST} can be captured by geographic distance alone. Adding information on genetic clusters to this model captures only an extra ~2% of the variance.

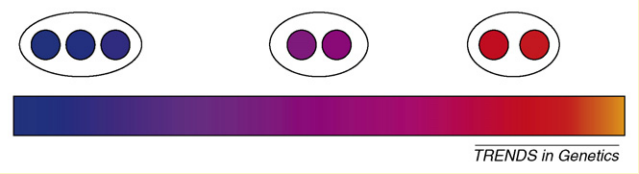

Figure I. Heterogeneous sampling can reveal genetic clusters that are biologically meaningless. The gradation in colour from blue to orange represents a hypothetical situation of strictly continuous variation in allele frequencies. If sampling is heterogeneous (population samples represented here by circles) then the pattern of clinal variation can be mistaken for genetically distinct clusters (black ellipses).

effect (or 'dynamic stepping-stone') starting at a single African origin provided an excellent fit to the observed geographic pattern of heterozygosity. Liu *et al.* [24] also took advantage of these clinal patterns to investigate human settlement history. They used a dynamic population genetics model based on a one-dimensional stepping-stone coupled with an explicit geographical framework (i.e. geographic distances along landmasses from East Africa, avoiding areas with mean altitude >2000 m; Box 1 Figure Ia) to simulate parameters of the colonization process. Parameter values were estimated that provide the best fit to the variance in allele size computed in the HGDP-CEPH dataset. Their results indicate a scenario in which the world was colonized by a founding population of ~1000 effective individuals that started expanding some 56 000 years ago and rapidly colonized new habitats [24].

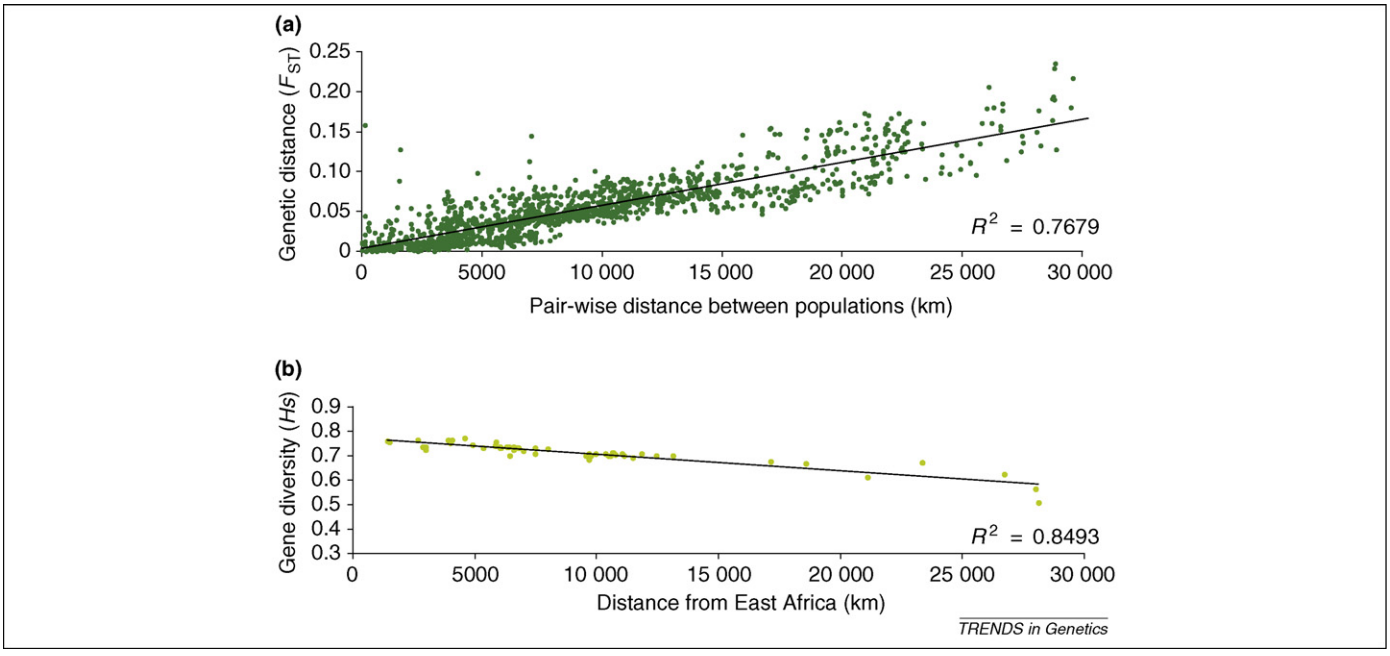

Figure 1. Human genetic variation is mainly clinal. **(a)** Pairwise genetic distance (F_{ST}) between populations in the HGDP-CEPH cell line panel is plotted against pairwise geographic distance. There is a strong, positive, linear relationship between genetic differentiation and geographic distance between populations, which is consistent with IBD. 77% of the variance can be explained by geographic distance between populations. **(b)** Gene diversity (*Hs*) within the HGDP-CEPH populations is plotted against geographic distance from East Africa (Addis Ababa). In this case, geographic distance from East Africa explains nearly 85% of the decrease in genetic variation within populations. Decreasing diversity with distance from East Africa is consistent with African origins for modern humans. In both cases, geographic distances were estimated along likely colonization routes (see Box 1 text and Box 1 Figure Ia) and analyses were based on 783 autosomal microsatellites typed in the H971 standardized subset of the HGDP-CEPH panel [62]. The figures are updated from Ref. [13]. See also Refs [10,12].

A much more complex, geographically explicit framework has recently been adopted to investigate specific questions pertaining to the origins and settlement history of modern humans [25,26]. The approach used in these studies was based on the highly innovative method described by Currat and Excoffier [25] and implemented in the software SPLATCHE [27]. Briefly, geographical information, such as vegetation or topography, is used to define the carrying capacity and relative 'friction' for each 100 km^2 cell (or 'deme'). The friction map is then used in a deterministic forward demographic simulation under a dynamic two-dimensional stepping-stone [28] to generate a database of migration rates and population sizes for each cell. A stochastic coalescent (i.e. backward in time) simulation is then performed to obtain the range of possible patterns of genetic diversity given the demography. Parameters of the demographic expansion can then be evaluated against observed data. Using an early version of this method, which assumed environmental homogeneity, Currat and Excoffier [25] simulated the range expansion of modern humans into Europe under realistic demographic scenarios to investigate potential admixture between colonizing humans and resident Neanderthals. Their simulations indicated that even with only a few admixture events, the contribution of Neanderthal genes to the current human gene pool should be large because new (Neanderthal) genes have a high probability of persistence when entering a progressively expanding (modern human) population compared with those entering a stationary population. However, the complete absence of Neanderthal mtDNA in modern Europeans indicates no (or virtually no) successful admixture events between Neanderthal females and human males.

More recently, Ray *et al.* [26] compared the observed pattern of genetic differentiation between populations (in the HGDP-CEPH panel) to that expected under various models of modern human origins to see whether the observed patterns of genetic diversity enabled them to distinguish between Unique Origin (UO) models (e.g. Ref. [29]) and different possible scenarios of Multiregional Evolution (MRE) (e.g. Ref. [30]; see Ref. [31] for a review of the models). After correcting for a potential ascertainment bias, they inferred that a unique origin in East Africa was the model with the highest likelihood. However, they found that several other UO scenarios (but no MRE model) had high likelihoods, illustrating the limitations of this method, and the authors suggest improvements, for example by implementing Approximate Bayesian Computational (ABC) methods, which have recently shown to be useful for inferring migration rates after spatial expansion [32]. Other large challenges include allowing for temporal changes in population sizes and dispersal rates. Despite the limitations, this approach is a milestone in methods for incorporating spatial data into population genetics models of human evolution.

Testing for selection within a geographically explicit framework

Discovering and describing genomic elements under selection is of great interest because of the close relationship between heritable disease susceptibility and natural selection. As our ancestors colonized the entire world, different populations were exposed to various environments and infectious agents. A textbook example for the signature of differential exposure to a selective pressure is the distribution of the Hb^s mutation causing sickle cell anaemia [33], which is found at high frequency in Africa. It is likely that there has been a catalogue of such regional selective pressures, as suggested by the presence of a large number of selective sweeps [34] in all three populations from the International HapMap Project (Yorubans from Africa, individuals of European descent and individuals of Chinese and Japanese ancestry).

Integrating geographical information into analyses holds great promise for more robust inference of selection. Essentially all approaches used to test for selection are based on the same premise: given that all polymorphisms, whether selected or not, are affected by the past demography, neutral or nearly neutral genomic regions will fit to a general pattern, whereas selected ones will deviate from it (see Ref. [35] for a recent review). Various approaches differ in the exact way they test for a deviation, but they can be broadly classified into two main categories, depending on whether they estimate the general pattern directly from the data ('model-free' tests) or through an explicit underlying demographic model ('model-based' tests). These two approaches are discussed below, and in Box 1 we outline how the geographic frameworks themselves can be expanded to investigate selection.

Model-free methods

Model-free approaches (sometimes referred to as 'outlier approaches') are conceptually simple, as they entail computation of the same statistics for many polymorphisms and flagging outliers to the main distribution as candidate selected genes. Despite their simplicity, these approaches have proven powerful. For example, Young *et al.* [36] investigated the genetic basis of hypertension by comparing the distribution of SNPs in five genes involved in blood pressure regulation with the distribution of the HGDP-CEPH microsatellites and 42 control SNPs. They found two functional SNPs to show a stronger association with latitude (an environmental proxy for hypertension susceptibility) than any of the other markers, and another five functional SNPs to be as closely associated with latitude as the most extreme of the control markers.

However, it is important to bear in mind that the ability of model-free tests to detect true selection is a function of the number of comparable polymorphisms in the dataset, as these define the global empirical distribution from which outliers are detected. For example, the AIDS-resistant 32-base deletion haplotype in the cytokine receptor CCR5 (*CCR5-Δ32*) is confined to a narrow geographic range (northern Europe), where it is found at very high frequency. Although there is evidence that this extraordinary pattern has been generated by selection [37], when compared with a large number of markers from the same chromosome and markers found within 180 immunological genes, *CCR5-Δ32* was found to be unexceptional in its diversity or distribution [38].

The need for a large number of control loci to be typed in the same populations as the polymorphism of interest is an

obvious limitation of most model-free approaches. A possible way to overcome such a limitation is to take advantage of the clinal patterns in neutral markers described in the previous section. Because global patterns are so strong, it is possible to predict the distribution of neutral alleles even in populations that have not yet been typed. This approach was used by Prugnolle *et al.* [39] to investigate balancing selection in human major histocompatibility complex (MHC) class I genes. After correcting for the decline in diversity observed with increasing distance from East Africa, genetic diversity at the MHC was shown to be positively associated with the number of endemic diseases to which a population was exposed. This approach could easily be applied to other polymorphisms for which information has been compiled from disparate sources (such as those covered by the allele frequency database ALFRED, http://alfred.med.yale.edu/alfred/). Although the HapMap data has been useful for detecting selection in many cases [34,40], a far larger number of populations would increase the statistical power to detect selection using geographic approaches. Furthermore, it would enable characterization of associations between regional environmental factors and genotypes. Unfortunately, though, this approach might be limited in its ability to detect regions under selection if the population has been through a bottleneck, or if the selected allele was previously neutral (rather than a new mutation), is recessive and/or is in a region of high recombination [41].

Model-based approaches

Pioneered by R.A. Fisher's 'wave of advance of advantageous genes' model [42], model-based tests rely on a general demographic model, rather than empirical data, to decide whether the distribution of a given gene is compatible with the assumption of neutrality. Estimates from this approach are therefore only as good as the demographic models on which they are based. A particular complication stems from the rapid spatial expansion of anatomically modern humans. It has been shown that for spatially expanding populations, alleles present at the edge of the expansion can reach unusually high frequencies [43,44]. Such local overrepresentation would be characterized as incompatible with neutrality by a static model. This mechanism has been invoked as an alternative to natural selection to explain the distribution of variants of the *microcephalin* gene, which is essential for brain development [45–47]. Returning to the case for selection at *CCR5-Δ32,* Fisher's model was recently adapted to investigate the distribution of this allele in a geographically explicit framework [48]. In contrast to results from the model-free approach of Sabeti and colleagues [38], fitting selection gradients to observed allele frequency data confirmed that strong selection and long-range dispersal have been important in determining the spread of *CCR5-Δ32.* A limitation of this model is that it enables only estimation of the ratio of dispersal to selection and not direct estimation of the selection coefficient (unless the age of the allele is known). The model could also be extended to include genetic drift and also information at linked markers.

Although the case of *CCR5-Δ32* illustrates that both model-free and model-based approaches have their limitations, both methods have enormous potential. Further use of such geographically explicit approaches could greatly benefit our understanding of how natural selection has shaped the human genome and enable us to detect associations between regional environmental factors and genotypes.

Concluding remarks and future directions

The genetic structure of human populations at neutral loci is largely characterized by clinal patterns that are consistent with global-scale IBD. The jury is still out, however, on the exact biological processes that have generated the 5–6 clusters observed in human populations [8,49–51] and their importance for our understanding of the distribution of human genetic variation. Focusing on the simple clinal patterns enables us to infer key parameters of human settlement history using tractable population genetics models and to explore the influence of selection on different regions of the genome.

This is an important and exciting time for human population genetics research, as additional data are becoming increasingly available. For instance, 15 populations from India [52] and 29 from The Americas (A. Ruiz-Linares, personal communication) have recently been typed for the autosomal microsatellites previously used in the HGDP-CEPH panel [10]. As reflected in this review, analyses have so far focused on microsatellites, but there is also a wealth of data being generated from SNPs, in particular thanks to the HapMap initiative. SNP data offers exciting prospects for the analysis of functional regions. Unfortunately, though, the HapMap project intentionally focused on only four populations, which is too small a number to enable meaningful geographically explicit analyses. The situation could change with different initiatives under way that aim to collect genotypes from large association mapping studies [for example the Welcome Trust Case Control Consortium (http://www.wtccc.org.uk/), the Genetic Association Information Network (http://www.fnih.org/GAIN2/home_new.shtml) and the European Genotype Archive (http://www.ebi.ac.uk/ega/)], but it is unlikely that samples from a diverse geographic range will be available in the near future because association studies tend to focus on a limited number of populations.

Progress in the field will not depend just on the availability of more data. Although geographically explicit models are now well within reach, there are several challenges for their further development and application in human population genetics research. There is room for considerable improvement in the geographic models themselves (see Box 1), by the inclusion of many other sources of information. For example, maps of past vegetation could be used to adjust relative parameters of carrying capacity and migration rate in different parts of the world. An improvement to human colonization models could come from the integration of palaeontological and archaeological data, and information on human remains in different geographical areas could be incorporated into the inference procedure to model the wave of advance. It might be possible to integrate indirect information, such as the mass faunal extinction that is thought to have coincided with the spread of anatomically modern humans [53]. More recent major events in human

evolution, such as the re-colonization of northern latitudes after the Ice Ages, could also be taken into account.

Although all these prospects are exciting, it is important to keep in mind that the optimal complexity of models depends on their capacity to capture the pattern in the data while keeping the number of parameters to a minimum. There will always be a trade-off between the complexity of a model and the transparency and reproducibility of results. Expanding the number of parameters in population genetics models without a concomitant increase in the quality and quantity of the data can lead to spurious results owing to model over-fitting (i.e. if the complexity of the model is too great for the number of observations). There is no simple rule for what level of complexity is best, as this will depend crucially on the question being asked and the quality of the data. Searching for genomic regions under natural selection, for instance, might best be approached with simple, robust, geographically explicit models.

We have of course focused this review on inference of patterns in human populations, but geographical frameworks will have an important role to play more generally. For example geographic frameworks could be used to study phylogeographic patterns linked to post-glacial (re-) colonization in natural populations, or the spread of human pathogens or commensals, as demonstrated by the *Helicobacter* study [13] mentioned previously. Finally, we have also concentrated on inference at a global scale, but there are many important regional-scale questions, which could be pursued using spatially explicit frameworks that incorporate information on local geography, culture and language, and we envisage this to be a significant avenue for future development. For these more regional-scale analyses, human population geneticists should pay attention to the field of landscape genetics. This emerging discipline integrates population genetics and landscape ecology in a spatially explicit framework to evaluate how environment and landscape influence genetic structure of populations (see Refs [27,54–57] for important methodological developments and Ref. [58] for a recent review).

Acknowledgements

We thank Mattias Jakobsson, Franck Prugnolle, Toomas Kivisild and the anonymous referees for helpful comments on this manuscript. We also thank the BBSRC for funding.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2007.07.002.

References

1 Wright, S. (1943) An analysis of local variability of flower colour in *Linanthus parryae*. *Genetics* 28, 139–156
2 Dobzhansky, T. and Epling, C. (1944) Contributions to the genetics, taxonomy and ecology of Drosophila pseudoobscura and its relatives. *Carnegie Inst. Washington Publ.* 554, 1–183
3 Mourant, A.E. *et al.* (1976) *The Distribution of Human Blood Groups and Other Polymorphisms.* Oxford University Press
4 Lewontin, R.C. (1972) The apportionment of human diversity. *Evol. Biol.* 6, 381–398
5 Cavalli-Sforza, L. *et al.* (1994) *The History and Geography of Human Genes.* Princeton University Press
6 Cavalli-Sforza, L.L. and Feldman, M.W. (2003) The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* 33, 266–275
7 Wright, S. (1943) Isolation by distance. *Genetics* 28, 114–138
8 Rosenberg, N.A. *et al.* (2002) Genetic structure of human populations. *Science* 298, 2381–2385
9 Cann, H.M. *et al.* (2002) A human genome diversity cell line panel. *Science* 296, 261–262
10 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
11 Manica, A. *et al.* (2005) Geography is a better determinant of human genetic differentiation than ethnicity. *Hum. Genet.* 118, 366–371
12 Prugnolle, F. *et al.* (2005) Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15, R159–R160
13 Linz, B. *et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* 445, 915–918
14 Wilson, J.F. *et al.* (2001) Population genetic structure of variable drug response. *Nat. Genet.* 29, 265–269
15 Jorde, L.B. and Wooding, S.P. (2004) Genetic variation, classification and 'race'. *Nat. Genet.* 36, S28–S33
16 Campbell, C.D. *et al.* (2005) Demonstrating stratification in a European American population. *Nat. Genet.* 37, 868–872
17 Relethford, J.H. (2004) Global patterns of isolation by distance based on genetic and morphological data. *Hum. Biol.* 76, 499–513
18 Bortolini, M.C. *et al.* (2003) Y-chromosome evidence for differing ancient demographic histories in the Americas. *Am. J. Hum. Genet.* 73, 524–539
19 McDougall, I. *et al.* (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* 433, 733–736
20 Mellars, P. (2006) Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonization of Eurasia. *Science* 313, 796–800
21 Manica, A. *et al.* (2007) The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 448, 346–348
22 Conrad, D.F. *et al.* (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* 38, 1251–1260
23 Hey, J. (2005) On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol.* 3, 965–975
24 Liu, H. *et al.* (2006) A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* 79, 230–237
25 Currat, M. and Excoffier, L. (2004) Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biol.* 2, 2264–2274
26 Ray, N. *et al.* (2005) Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res.* 15, 1161–1167
27 Currat, M. *et al.* (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol. Ecol. Notes* 4, 139–142
28 Kimura, M. and Weiss, G.H. (1964) The stepping-stone model of population structure and the decrease of genetic correlation with distance. *Genetics* 49, 561–576
29 Cann, R.L. *et al.* (1987) Mitochondrial-DNA and human-evolution. *Nature* 325, 31–36
30 Wolpoff, M.H. (1989) Multiregional evolution: The fossil alternative to Eden. In *The Human Revolution: Biological Perspectives in the Origins of Modern Humans* (Mellars, P. and Stringer, C.B., eds), pp. 62–108, Princeton University Press
31 Stringer, C. (2002) Modern human origins: progress and prospects. *Philos. Trans. Roy. Soc. Lond. B Biol. Sci.* 357, 563–579
32 Hamilton, G. *et al.* (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* 170, 409–417
33 Allison, A.C. (1954) Protection afforded by sickle-cell trait against subtertian malarial infection. *BMJ* 1, 290–292
34 Voight, B.F. *et al.* (2006) A Map of Recent Positive Selection in the Human Genome. *PLoS Biol.* 4, 446–458
35 Biswas, S. and Akey, J.M. (2006) Genomic insights into positive selection. *Trends Genet.* 22, 437–446

36 Young, J.H. *et al.* (2005) Differential susceptibility to hypertension is due to selection during the Out-of-Africa expansion. *PLoS Genet.* 1, 730–738
37 Slatkin, M. and Bertorelle, G. (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* 158, 865–874
38 Sabeti, P.C. *et al.* (2005) The case for selection at *CCR5-Δ32*. *PLoS Biol.* 3, e378
39 Prugnolle, F. *et al.* (2005) Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* 15, 1022–1027
40 Sabeti, P.C. *et al.* (2006) Positive natural selection in the human lineage. *Science* 312, 1614–1620
41 Teshima, K.M. *et al.* (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16, 702–712
42 Fisher, R.A. (1937) The wave of advance of advantageous genes. *Ann. Eugen.* 7, 353–369
43 Edmonds, C.A. *et al.* (2004) Mutations arising in the wave front of an expanding population. *Proc. Natl. Acad. Sci. U. S. A.* 101, 975–979
44 Klopfstein, S. *et al.* (2006) The fate of mutations surfing on the wave of a range expansion. *Mol. Biol. Evol.* 23, 482–490
45 Evans, P.D. *et al.* (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science* 309, 1717–1720
46 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
47 Yu, F. *et al.* (2007) Comment on "Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in Homo sapiens. *Science* 316, 370b
48 Novembre, J. *et al.* (2005) The geographic spread of the CCR5 Delta 32 HIV-resistance allele. *PLoS Biol.* 3, 1954–1962
49 Rosenberg, N.A. *et al.* (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PloS Genet.* 1, 660–671
50 Serre, D. and Pääbo, S.P. (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* 14, 1679–1685
51 Witherspoon, D.J. *et al.* (2006) Human population genetic structure and diversity inferred from polymorphic L1 (LINE-1) and Alu insertions. *Hum. Hered.* 62, 30–46
52 Rosenberg, N.A. *et al.* (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet.* 2, 2052–2061
53 Burney, D.A. and Flannery, T.F. (2005) Fifty millennia of catastrophic extinctions after human contact. *Trends Ecol. Evol.* 20, 395–401
54 Guillot, G. *et al.* (2005) A spatial statistical model for landscape genetics. *Genetics* 170, 1261–1280
55 Foll, M. and Gaggiotti, O.E. (2006) Identifying the environmental factors that determine the genetic structure of populations. *Genetics* 174, 875–891
56 Francois, O. *et al.* (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* 174, 805–816
57 Miller, M.P. (2005) Alleles In Space (AIS): Computer software for the joint analysis of interindividual spatial and genetic Information. *J. Hered.* 96, 722–724
58 Storfer, A. *et al.* (2007) Putting the "landscape" in landscape genetics. *Heredity* 98, 128–142
59 Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, 209–220
60 Legendre, P. and Legendre, L. (1998) *Numerical Ecology*. Elsevier
61 Raufaste, N. and Rousset, F. (2001) Are partial mantel tests adequate? *Evolution Int. J. Org. Evolution* 55, 1703–1705
62 Rosenberg, N.A. (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.* 70, 841–847
63 Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959
64 Kittles, R.A. and Weiss, K.M. (2003) Race, ancestry, and genes: implications for defining disease risk. *Annu. Rev. Genomics Hum. Genet.* 4, 33–67
65 Gayden, T. *et al.* (2007) The Himalayas as a directional barrier to gene flow. *Am. J. Hum. Genet.* 80, 884–894
66 Shriver, M.D. *et al.* (2005) Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Hum. Genomics* 2, 81–89
67 Nievergelt, C.M. *et al.* (2007) Generalized analysis of molecular variance. *PLoS Genet.* 3, 467–478
68 Butty, V. *et al.* (2007) Signatures of strong population differentiation shape extended haplotypes across the human CD28, CTLA4, and ICOS costimulatory genes. *Proc. Natl. Acad. Sci. U. S. A.* 104, 570–575
69 Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics* 15, 323–354
70 Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, p. 164
71 Wright, S. (1931) Evolution in Mendelian populations. *Genetics* 16, 97–159