# Impact of human population history on distributions of individual-level genetic distance

*Joanna L. Mountain*[1,2*] *and Uma Ramakrishnan*[3]

[1] Department of Anthropological Sciences, Stanford University, Stanford, CA 94305-2117, USA
[2] Department of Genetics, Stanford University, Stanford, CA 94305-5120, USA
[3] National Centre for Biological Sciences, GKVK Campus, Bellary Road, Bangalore 560065, India
*Correspondence to: Tel: +1 650 725 5009; Fax: +1 650 725 9996; E-mail: mountain@stanford.edu

*Date received (in revised form): 2nd December 2004*

## Abstract

Summaries of human genomic variation shed light on human evolution and provide a framework for biomedical research. Variation is often summarised in terms of one or a few statistics (eg $F_{ST}$ and gene diversity). Now that multilocus genotypes for hundreds of autosomal loci are available for thousands of individuals, new approaches are applicable. Recently, trees of individuals and other clustering approaches have demonstrated the power of an individual-focused analysis. We propose analysing the distributions of genetic distances between individuals. Each distribution, or common ancestry profile (CAP), is unique to an individual, and does not require *a priori* assignment of individuals to populations. Here, we consider a range of models of population history and, using coalescent simulation, reveal the potential insights gained from a set of CAPs. Information lies in the shapes of individual profiles — sometimes captured by variance of individual CAPs — and the variation across profiles. Analysis of short tandem repeat genotype data for over 1,000 individuals from 52 populations is consistent with dramatic differences in population histories across human groups.

***Keywords:*** *human population genetic structure, genetic similarity, short tandem repeats (STRs), multilocus genotypes*

## Introduction

The collective human gene pool, consisting of the genomes of all living people, has much to reveal regarding human population history. Until recently, surveys of human genetic variation have been sparse, in that hundreds or thousands of individuals have been studied for a small number of genetic regions (eg blood groups, Human Lymphocyte Antigens (HLA), mitochrondrial DNA, Y chromosome[1–3]) and a few individuals have been studied for a large fraction of the genome (eg through the Human Genome Project). In the past few years, however, larger sets of individuals have been studied for hundreds of genetic regions[4] and, concomitantly, new data analysis tools have been developed.[5] With new data and new tools, we are rapidly gaining a more precise understanding of how genetically similar individuals are, and of how that similarity corresponds to other dimensions of human variation.

### Summaries of human genetic variation

Most differences between genomes take the form of single nucleotide polymorphisms (SNPs) rather than DNA insertions, deletions or multiplications.[6] For the autosomes, two DNA sequences chosen at random appear to differ at an average of about one per 1,000–1,500 nucleotide sites.[7–9] This level of diversity corresponds to between 2 and 3.2 million nucleotide differences between individual genomes and is about one order of magnitude lower than the diversity detected within *Drosophila* (fruitfly) populations.[7]

Numerous studies have indicated that the number of differences between human genomes varies greatly depending on the pair of genomes considered. The most striking and consistent pattern is the higher level of genetic diversity in Africa than in other regions and the relatively low levels of diversity in the Americas. Zhao and colleagues, in examining a 10 kilobase (kb) non-coding region, found an average of 8.5 differences between African samples and an average of 8.2 differences between non-African samples.[8] Yu and colleagues found a somewhat lower level of nucleotide diversity ($\pi$) of 0.076 per cent among Africans and 0.047 per cent among non-Africans.[9] As indicated in the summary of short tandem repeat (STR) data by Rosenberg *et al.*, diversity within African groups (average heterozygosity = 0.774) tends to be slightly higher than diversity within Middle Eastern (0.756), European (0.751) and Central and South Asian (0.752) populations.[4]

Those groups are, in turn, somewhat more diverse than are the East Asian populations (heterozygosity = 0.723), which, in their turn, are more diverse than the Oceanic (0.683) and Native American (0.599) populations.[4] All differences in heterozygosity for pairs of continents are significant at $p < 0.00001$, except for Europe versus the Middle East ($p = 0.0058$), Europe versus Central/South Asia ($p = 0.7182$) and the Middle East versus Central/South Asia ($p = 0.0554$) (Noah Rosenberg, personal communication).

Human genetic variation is often summarised in terms of hierarchical population genetic structure. In 1972, Lewontin estimated, using blood group and protein polymorphism data, that about 6.3 per cent of genetic variation was explained by differences among seven groups that he termed 'races'.[10] Differences between members of the same population accounted for 85.4 per cent of the total genetic variation. The remaining 8.3 per cent was accounted for by the variation between populations, within each of the seven 'races'.[10] In recent years, geneticists have replicated Lewontin's finding using independent regions of the genome: most estimates of $F_{ST}$ (between-group variation) have ranged from 0.05–0.15.[4,11–14] These estimates indicate that two individuals affiliated with different racially or ethnically identified groups are only slightly more likely to differ at a given neutrally evolving locus than are two individuals affiliated with the same group. A large proportion of human genetic variation is found within racially, ethnically or linguistically identified groups. Notable exceptions, reflecting smaller effective population sizes, include the mitochondrial genome and Y chromosome SNPs, with recent estimates of between-group variation ranging from 0.3 to 0.4.[11,15]

Although human genetic variation has often been summarised using single statistics such as $F_{ST}$; such single statistics are an inadequate and potentially misleading summary of our species' diversity.[16,17] $F_{ST}$ is most straightforwardly interpreted if the underlying population history is of a single population that instantaneously divides into a number of equally sized, panmictic subpopulations, each of which remains at the same size throughout the subsequent time. Human history is far from fitting such a model. Genetic distances, often represented in the form of population trees,[18] provide a more detailed representation of structure.[1] Recently, Long and Kittles used a sequential model-fitting approach to infer structure, generating a tree relating a set of human populations.[16] The latter study highlights the hierarchical and uneven structure of human genetic variation (see their Figure 2D).

## A focus on the individual

Although a combination of heterozygosity and genetic distance estimates for a set of populations may provide a fairly accurate summary of genetic variation, these statistics describe variation within the data most completely when population histories are relatively simple. One way to summarise population genetic structure in greater detail is to focus on individuals rather than on populations. For some species, summarisation at the individual level may reveal substructure that is hidden by population level summaries. In addition, the focus on the individual takes away the emphasis on the group labels. This change of emphasis can be particularly important when individuals have multiple group affinities. Finally, individual-based approaches have the potential to provide information about within- and between-group variations simultaneously.

Several research groups have used trees of individuals to summarise genetic variation.[19–21] Such trees provide much greater detail than do population trees, but are limited to a relatively small number of individuals and are not readily summarised. A number of algorithms, including those implemented via the *immanc*,[22] *BayesAss*[23] and *structure*[5] software, allow one to assign, with a given probability, an individual (or portion of its genome) to a particular population. The *BayesAss* approach is valuable in estimating migration rates, inbreeding coefficients and recent immigrant ancestry simultaneously, but does require *a priori* assignment of individuals to populations. The *structure* algorithm identifies clusters of genetically similar individuals without *a priori* population assignment. The approach provides information regarding within-group variation, to the extent that individuals are inferred to be members of multiple clusters. The combination of *structure* analysis and *distruct*,[24] a program that generates a graphical representation of population structure, provides a valuable exploratory tool. The many available approaches to estimation of relatedness between individuals also focus directly on individuals, but are most successfully applied in the context of relatively large, random mating populations.[25–29]

## Common ancestry profiles

We introduce an exploratory, individual-focused approach that complements population level analyses, trees of individuals, relatedness estimation and assignment/clustering algorithms. Distributions of genetic distances between individuals, here termed common ancestry profiles (CAPs), emphasise the shared ancestry of all members of a species and provide a detailed description of genetic variation without the need for *a priori* assignment of individuals to populations. Like *distruct*, the approach provides a visual representation of genetic variation, thereby constituting an exploratory data analysis tool. The profiles enable us to visualise how genetically similar an individual is to others in the context of linguistic, social or geographic variation. In addition, the approach brings together genealogical and population level perspectives.

The total set of genetic distances among individuals can be partitioned in a manner analogous to a partitioning of variance: individual heterozygosity (the fraction of an individual's loci that is heterozygous) represents within-individual variation; comparisons among individuals of a population

represent within-population, between-individual variation; comparisons of individuals of different populations of a region represent within-region, between-population variation; and comparisons of individuals of different regions represent between-region variation. The CAPs can, therefore, provide a graphical display of an often misinterpreted breakdown of total genetic variation into its components.

As illustrated in Figure 1, in practice CAPs can vary quite dramatically across individuals. The overall profile for the individual affiliated with the Pima population (Figure 1a) is more skewed than that of the French individual (Figure 1b) because the Pima individual is genetically much more similar to other Pima individuals than to non-Pima individuals (Figure 1c), while the French individual is, on average, almost as similar to non-French individuals as to other French individuals (Figure 1d).

Going one step further, by combining individual profiles across all individuals of each population, we see variation across 52 previously described populations (Figure 2).[4,30] Several of the population samples from the Americas, as well as the Melanesian sample, reveal relatively broad distributions, even though individuals known to be closely related to others in the sample set have been removed. The Colombian and Melanesian profiles, for instance, reveal a number of pairs of genetically very similar individuals. Several of these pairs represent first- or second-degree relatives according to *relpair*[31] analysis.

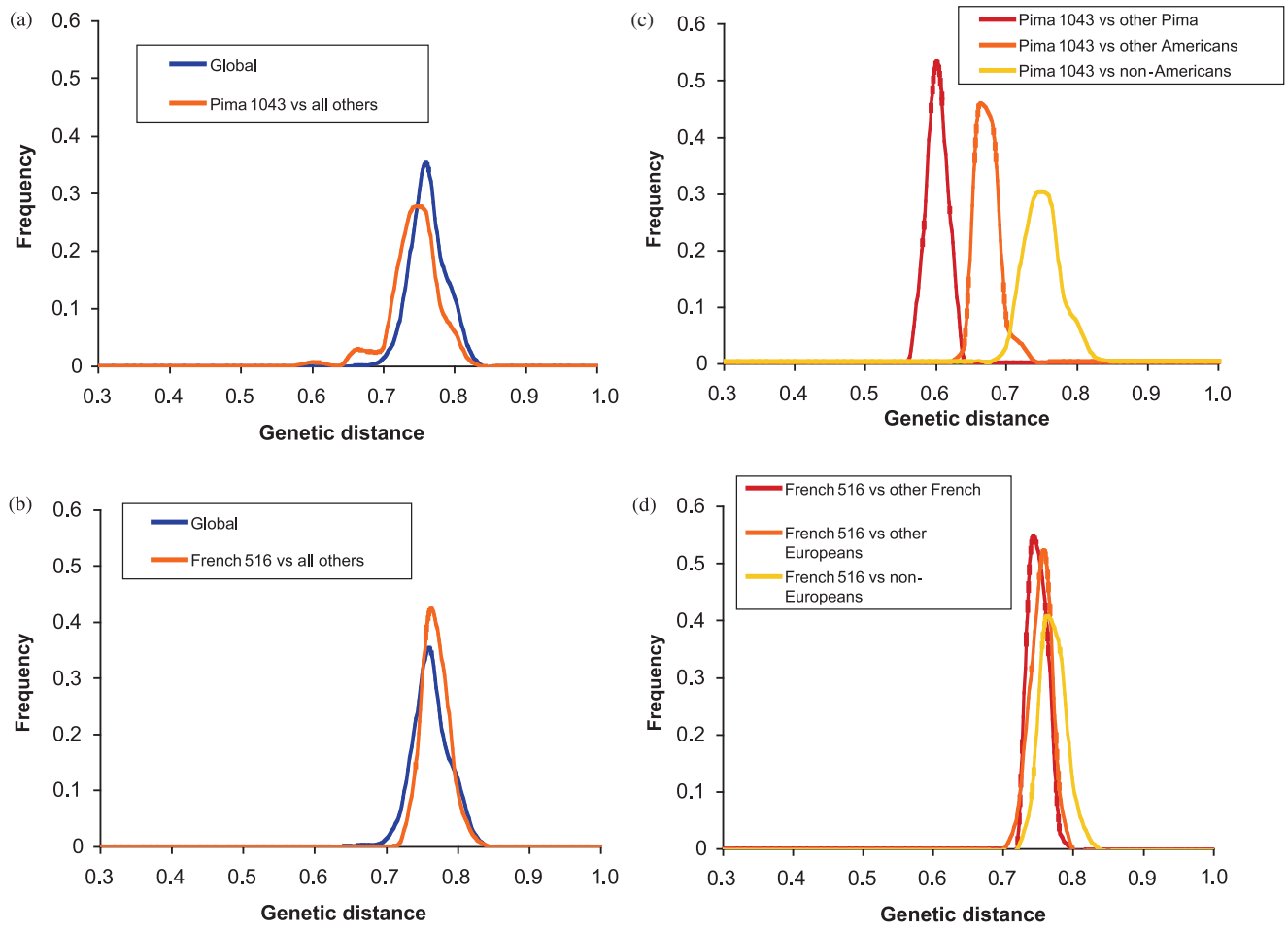


**Figure 1.** Common ancestry profiles for two individuals, based on genotype data for 377 short tandem repeat (STR) loci.[4] Distribution of genetic distance estimates for all possible pairs drawn from 1,013 individuals of the CEPH-HGDP STR dataset (overall) and for all pairs including individual Pima1043 or French 516. (a) Pima 1043 vs all other individuals; (b) French 516 vs all other individuals; (c) Pima 1043 vs three sets of individuals: other Pima, other non-Pima Americans and all non-Americans of CEPH-HGDP set; (d) French 516 vs three sets of individuals: other French, non-French Europeans and non-Europeans. Genetic distance for a pair of individuals is defined as the probability with which two alleles, one drawn randomly from each of the two individuals, differ in state, averaged across loci. Forty-three individuals (13 duplicates and 30 close relatives) excluded from original Rosenberg dataset.[4]

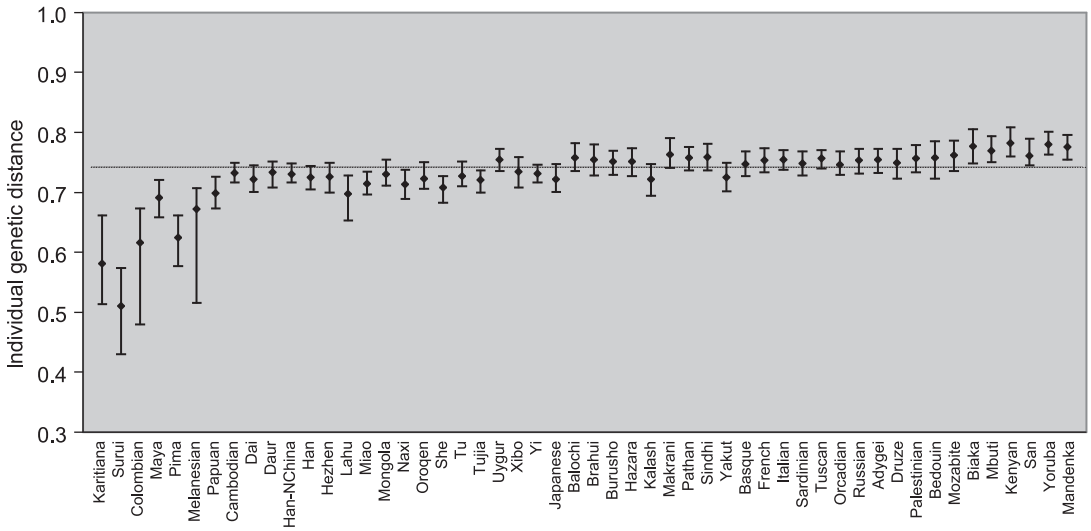


**Figure 2.** Summary of common ancestry profiles for 52 human populations. Mean genetic distance ($\hat{d}_{xy}$) among individuals within each of 52 human populations of the CEPH-HGDP panel, with range indicated by the 5th to 95th percentiles. Genetic distance of individuals *x* and *y* reflects the probability that two short tandem repeat alleles drawn, one from *x* and one from *y* at a particular locus, differ in state. The horizontal dotted line indicates average genetic distance (0.74) for all within-population comparisons.

A CAP can be informative by indicating possible duplicated samples, by indicating closely related individuals within a dataset or as a graphical display of the partitioning of variance. The power of the approach is greatest, however, in cases where we have expectations for the shape of a profile. For a random mating population with all sampled individuals distantly related, a Central Limit Theorem argument leads to the expectation of a normal distribution for an individual CAP. Expectations under more realistic models of population history, however, are essential if we are to accurately interpret a set of profiles.

In order to facilitate interpretation of CAPs, we have considered a set of simple models of population history, simulating genetic variation in the context of those models using a coalescent approach. Through simulation, we have explored the impact of population isolation and of gene flow on CAPs, and the potential for inferring recent or long-term gene flow from a set of CAPs. In light of these simulations, we have evaluated a set of individual and group CAPs derived from STR variation at 377 loci.

# Methods

## CAPs

A CAP consists of the distribution of genetic distance for a set of pairs of individuals. Although any measure of dissimilarity between individuals might be applied, we focus on $d_{xy}$ — the probability of non-identity in state for two alleles, one from individual *x* and one from individual *y*, chosen at random from a particular genomic location. Here, the probability of identity in state is simply the probability that two alleles are of identical type.[32] $d_{xy}$ is equivalent to $1 - s_{xy}$ — the probability of identity in state of two alleles, one from individual *x* and one from individual *y*. Under simple models of population history, $s_{xy} = rp + (1 - r)p^2$, where $p$ is the allele frequency in a base population and $r$ is a relatedness measure or 'probability of identity-by-descent'.[32] In many cases, however, population history is more complex or we are interested in identity-by-descent of individuals in the base population. We therefore proceed without consideration of a base population and estimate $s_{xy}$ (and hence $d_{xy}$) directly from multilocus genotypes. $d_{xy}$ is estimated as:

$$\hat{d}_{xy} = 1 - \hat{s}_{xy},$$

where $\hat{s}_{xy}$, is calculated as follows for each locus.

$\hat{s}_{xy} = 1$ for (genotype x = ii, genotype y = ii);

0.5 for (x = ij and y = ij) or (x = ii and y = ij) or vice versa;

0.25 for (x = ij and y = ik);

0 for (x = ii and y = jj), and

0 for x = ij and y = kl

where i,j,k,l represent distinct alleles. The similarity estimate meets the criteria for transformation to a distance, in that it is everywhere non-negative definite.[33]

Estimates are averaged across loci to generate an approximation of distance for a pair of individuals. The distance

measure ranges in value from 0–1.0, with 0 indicating the comparison of two individuals identical and homozygous at all loci and 1.0 indicating no overlap of alleles. $\hat{d}_{xx}$ (the distance between an individual and him- or herself) for individuals heterozygous at all loci is 0.5. More generally, the distance between an individual and him- or herself (or a monozygotic twin) is $\hat{d}_{xy} = 0.5h$, where $h$ is the fraction of loci heterozygous in that individual.

An individual CAP consists of the distribution of estimates for a single focal individual compared with a set of other individuals. We represent individual CAPs as binned relative frequencies, with the range divided into a set of equal-sized bins. The set of comparison individuals may be a geographically global sample, a set of cases or controls in a medical context or any other set of interest. A group CAP consists of genetic distance estimates between all pairs of a set of individuals. Conditional on the genotype of the individual, an individual CAP consists of a set of independent distances. Conditional on the genotypes for a group of individuals, a CAP for that group consists of a set of non-independent distances.

## Models of population history

We considered a range of models to explore the impact of sample size and population history on CAPs. The basic model included two populations of effective size 1,000 that diverged 2,000 generations ago. We investigated the sensitivity of the CAPs and summary statistics to: (a) sample size (n = 25, 50 and 100 individuals per population); (b) time of population divergence (t = 1,000, 2,000 and 5,000 generations in the past); and (c) rate and timing of gene flow following divergence. We investigated both continuous gene flow (continuous gene flow following divergence at the rate of 0.5 or 2.0 migrants per generation) and recent gene flow. The recent gene flow model represents population divergence 2,000 generations ago followed by isolation for 1,900 generations, and then gene flow at a rate of 2.0 migrants per generation during the past 100 generations. We investigated both symmetrical and asymmetrical gene flow models. Results presented here are for asymmetrical models unless otherwise stated.

## Coalescent simulations

Using coalescent-based simulation[34] of the above population histories, we generated genotypes for each of $n$ sampled individuals per population. We assumed a single-step mutation model with a range constraint in order to best approximate evolution at STR, or microsatellite, loci. Five hundred unlinked loci were modelled for each sampled individual. We assumed a mutation rate of 0.0005/generation/locus, on the order of published estimates of effective mutation rate for STR loci.[35] Given an average of 10.8 (±0.2) alleles for 377 human STR loci,[36] we assumed a range constraint of 15 repeat alleles; that is, stepwise mutation generated novel alleles until a total of 15 alleles had been generated, at which point mutation generated only new copies of existing alleles.

## CAP summary statistics

CAPs of simulated and empirical data were analysed similarly for pairs of populations. An individual 'overall' CAP consists of the distribution of genetic distances between a focal individual in the reference population and all other individuals of the two populations. An individual 'within' CAP consists of the distribution of distances between the focal individual and the $n_0-1$ other individuals of the reference population. An individual 'between' CAP consists of comparisons between the focal individual and the $n_1$ individuals of the non-reference population. CAPs were generated for each sampled individual of a reference population. The following summary statistics were calculated using all such CAPs of the reference population sample: average and standard deviation (across individuals) of the average $\hat{d}_{xy}$ between a focal individual and others, and average and standard deviation (across individuals) of the standard deviation of $\hat{d}_{xy}$ for each individual. The average $\hat{d}_{xy}$ captures the central tendency of the distributions, while its standard deviation indicates variation across individuals in that tendency. The summaries of the standard deviations of individual profiles capture the spread of individual profiles and the variation across individuals in that spread. We calculated a raggedness statistic, $r$,[37] for each profile:

$$r = \sum_{i=2}^{d} (x_i - x_{i-1})^2,$$

where $d$ is the number of bins and $x_i$ is the weight in bin $i$. We also calculated expected heterozygosity for each population sample and for the combined (overall) sample of $n_0 + n_1$ individuals for comparison with mean individual genetic distance. Summary statistics were calculated for four sets of CAPs: (a) 'overall', (b) 'within', (c) 'between' and (d) 'cryptic'—an individual drawn at random from the reference and non-reference population is compared with other individuals of a random sample. Individual CAPs are presented as binned relative frequencies, with the range divided into 100 equal-sized bins.

For models with gene flow, we summarised the distributions in greater detail by calculating the average weight (summarising over individuals) in three particular genetic distance bins. These genetic distance bins correspond to: (1) average genetic distance when reference individuals are compared with other individuals within the reference population; (2) average genetic distance when reference individuals are compared with other individuals from the non-reference population; and (3) the mid-point between these two bins, which corresponds to the average genetic distance when reference individuals from a population are compared with individuals within the reference population with mixed ancestry.

### Data analysis

We analysed the CEPH-HGDP[30] multilocus STR genotype data generated by Rosenberg and colleagues in collaboration with the Marshfield Genotyping Service.[4] That dataset includes 377 STR loci tested in 1,056 individuals from 52 human populations. Although many more populations have been typed for a small number of genetic markers, including the classical markers (eg blood groups), mtDNA and the Y chromosome, the Rosenberg STR dataset remains the richest published set in terms of the number of individuals typed for a relatively large number of markers. Each individual in the dataset is associated with a population (identified in a variety of ways in the contexts of a number of separate research projects) and a geographical region.[4] We use those labels when referring to particular individuals by population or region.

We eliminated data for 13 individuals representing duplicate samples (see Table 1). Thirty individuals from four populations from the Americas are known to be closely related to other individuals in the sample (see Table 2). We carried out analyses both with and without these related individuals. We report results for the reduced dataset unless otherwise stated. For each of the 1,013 individuals, we generated the distribution of $\hat{d}_{xy}$ for that individual paired with all other persons of that individual's population. For two individuals (Pima 1043 and French 516) we generated distributions for three comparison groups: all other individuals of the same population; all individuals of a different population in the same geographical region; and all individuals of other geographical regions. A group CAP, including all between-individual distances for a given set of individuals, was generated for each of the 52 populations and for the full set of 1,013 individuals.

We considered four pairs of populations in greater detail in light of the simulation results: two pairs of geographically proximate populations and two pairs of geographically distant populations. In each case, we calculated summary statistics and CAPs for an individual versus: (1) other individuals of his or her local (reference) population; (2) other individuals in the comparison (non-reference) population; and (3) all other individuals in both the local and comparison population. Individual CAPs are presented as binned relative frequencies and are summarised as described above.

## Results

### Simulations

*CAPs and summary statistics: Basic population structure model*

We illustrate the simulation results with CAPs for ten individuals generated under the basic population model (Figure 3). These CAPs represent three categories of comparison: 'overall' (an individual from a reference population is compared with others in the reference and non-reference populations); 'within' (an individual from a reference population is compared with others in the reference population); and 'between'

**Table 1.** List of pairs of CEPH-HGDP samples[30] determined via common ancestry profile analysis of short tandem repeat data[4] to be duplicates. Population identification codes (IDs) drawn from Noah Rosenberg (http://www.cmb.usc.edu/people/noahr//diversitycodes.txt).

| Duplicate pair | 1st sample ID | 2nd sample ID | Population ID(s) | Population name(s) |
|---|---|---|---|---|
| 1 | 1022 | 813 | 601 | Han |
| 2 | 1235 | 1233 | 608 | Hezhen |
| 3 | 1025 | 762 | 684 | Japanese |
| 4 | 220 | 111 | 58/54 | Pathan, Hazara |
| 5 | 1154 | 149 | 27 | Italian-Bergamo |
| 6 | 589 | 583 | 37 | Druze |
| 7 | 652 | 650 | 36 | Bedouin |
| 8 | 659 | 658 | 71 | Melanesian |
| 9 | 826 | 657 | 71 | Melanesian |
| 10 | 979 | 660 | 71 | Melanesian |
| 11 | 981 | 472 | 488 | Biaka |
| 12 | 1087 | 452 | 488 | Biaka |
| 13 | 1092 | 457 | 488 | Biaka |

**Table 2.** List of individuals removed from analysis because of known close relationship (within two degrees) to another individual included in CEPH-HGDP short tandem repeat dataset.[4,30] Additional pairs of individuals indicated as possible close relatives by common ancestry profile analysis were not removed. Population identification codes (IDs) drawn from Noah Rosenberg (http://www.cmb.usc.edu/people/noahr//diversitycodes.txt).

| Sample ID | Population ID | Population name | Country |
|---|---|---|---|
| 995 | 82 | Karitiana | Brazil |
| 998 | 82 | Karitiana | Brazil |
| 999 | 82 | Karitiana | Brazil |
| 1004 | 82 | Karitiana | Brazil |
| 1006 | 82 | Karitiana | Brazil |
| 1008 | 82 | Karitiana | Brazil |
| 1011 | 82 | Karitiana | Brazil |
| 1012 | 82 | Karitiana | Brazil |
| 1014 | 82 | Karitiana | Brazil |
| 1016 | 82 | Karitiana | Brazil |
| 1017 | 82 | Karitiana | Brazil |
| 1018 | 82 | Karitiana | Brazil |
| 830 | 83 | Surui | Brazil |
| 833 | 83 | Surui | Brazil |
| 839 | 83 | Surui | Brazil |
| 840 | 83 | Surui | Brazil |
| 841 | 83 | Surui | Brazil |
| 842 | 83 | Surui | Brazil |
| 850 | 83 | Surui | Brazil |
| 858 | 83 | Surui | Brazil |
| 878 | 86 | Maya | Mexico |
| 1039 | 87 | Pima | Mexico |
| 1040 | 87 | Pima | Mexico |
| 1042 | 87 | Pima | Mexico |
| 1045 | 87 | Pima | Mexico |
| 1046 | 87 | Pima | Mexico |
| 1049 | 87 | Pima | Mexico |
| 1050 | 87 | Pima | Mexico |
| 1055 | 87 | Pima | Mexico |
| 1061 | 87 | Pima | Mexico |

(an individual from the reference population is compared with individuals from the non-reference population). The overall CAPs have two peaks (Figure 3a). Figures 3b and 3c reveal the components underlying those two peaks: lower genetic distance for within-population comparisons (Figure 3c) and a higher genetic distance for between-population comparisons (Figure 3b). Table 3 reveals that average genetic distance across individuals is highest for the between-group CAPs (0.537), intermediate for the overall CAPs (0.501) and lowest for within-population CAPs (0.469). The overall CAPs have the highest standard deviations (0.036), indicating higher within-CAP variance than for the within-population and between-population CAPs (0.01 and 0.015, respectively). Heterozygosity estimates were highest for the overall category. Average raggedness, which increases with rapid changes in bin frequencies, was highest for the within-population comparisons, despite the smoothness of those distributions (Figure 3c). The raggedness statistic fails to capture the multimodal nature of the overall CAPs.

*Impact of sample size.* As indicated in Table 3, reducing the sample size from 100 to 25 individuals per population does not significantly change the average or standard deviation of individual CAPs, consistent with the average being linear in the data. Raggedness decreases with sample size for all comparison groups, although within-population CAPs are the most ragged for all sample sizes.

*Impact of divergence time.* Table 3 indicates the impact of population divergence time on individual CAPs. As expected, the average genetic distance for between-population comparisons increases with earlier population divergence. Earlier divergence therefore leads to greater separation between the within-population and between-population genetic distance peaks of a CAP.

*Impact of gene flow.* A summary of CAPs for populations with asymmetric gene flow is presented in Table 4. We illustrate the simulation results with example CAPs for sets of ten individuals (Figure 4). These CAPs are 'cryptic', in that any simulated population structure is ignored and a subsample of individuals is drawn without consideration of population affiliation. Overall, the average genetic distance increases with increasing gene flow, as does raggedness. The standard deviation of individual genetic distance distributions decreases, as does the sample heterozygosity, with increasing gene flow. These results reflect the appearance of weight in the central region of the CAP distribution (Figure 4b), between the average genetic distance for the within-population comparisons and the average genetic distance for the between-population comparisons. Intermediate peaks correspond to comparisons between focal individuals in the reference population (which receives gene flow) and migrant individuals (individuals with a high proportion of immigrant ancestry) of the reference population. The weight in this intermediate portion of the distribution (averaged over 50 randomly selected CAPs) increases with gene flow (Table 4).

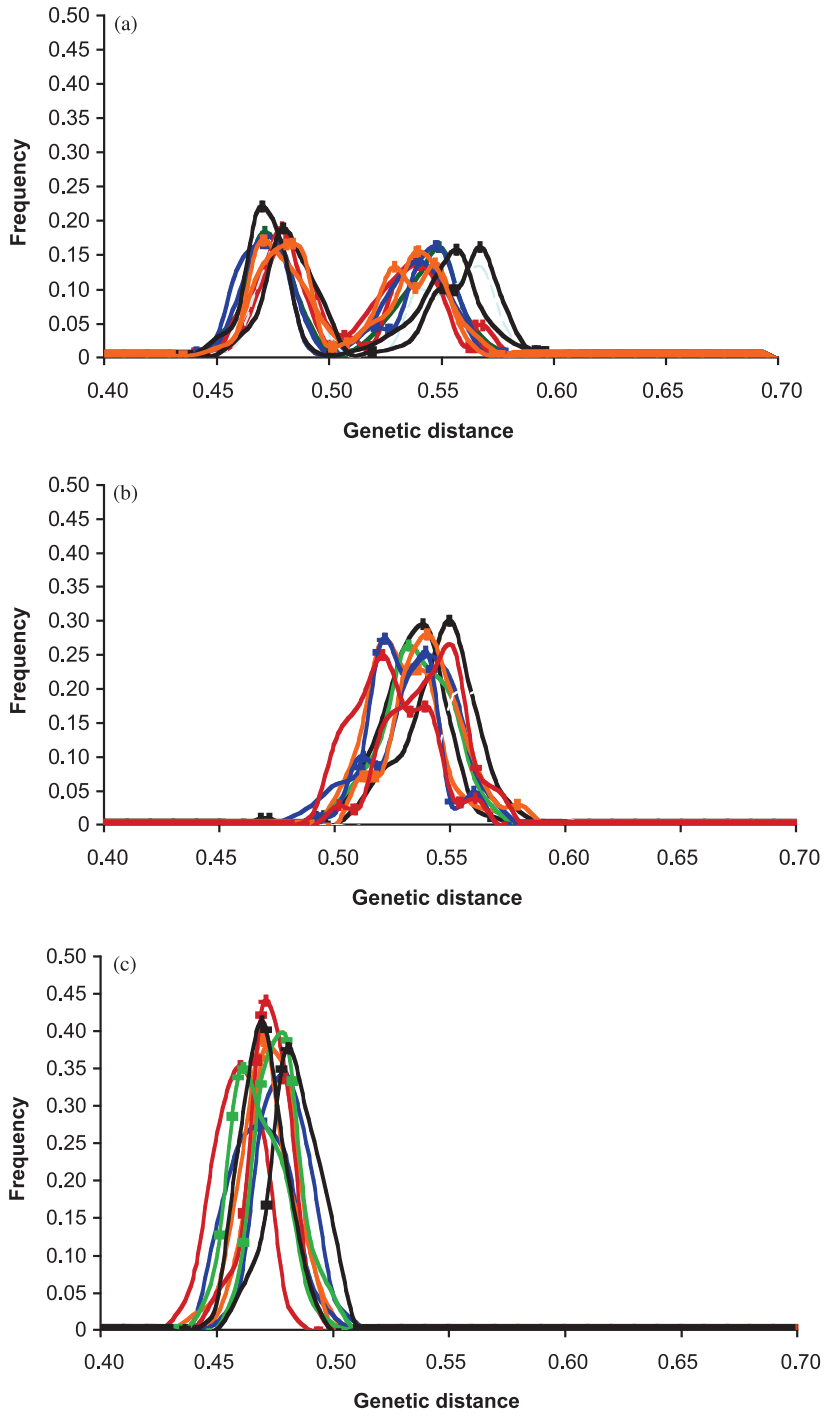


**Figure 3.** Ten examples each of simulated common ancestry profiles (CAPs) comparing an individual to: (a) all other individuals in two populations ('overall'); (b) all other individuals in the same population ('between'); and (c) all others in a different population ('within'). CAPs derived from coalescent simulations of two populations of effective size 1,000 that diverged 2,000 generations ago, generating 500 short tandem repeat loci (mutation rate: 0.0005/locus/generation; range constraint: 15, stepwise mutation model).

Additionally, as expected, the frequency of high genetic distances decreases with gene flow. Recent migration following a relatively long period of population isolation leads to a slight increase in both average and standard deviation of the individual CAPs. Summary statistics for within-population CAPs (for the reference population that receives immigrants) reveal

**Table 3.** Summaries of individual common ancestry profiles (CAPs) derived from data simulated via two-population models. Effective population size: 1,000 individuals per population. Statistics calculated across all individuals of simulated sample. Standard deviations included in parentheses.

| Pairs[a] | Time[b] | n[c] | Average[d] | Standard deviation[e] | Ht[f] | Raggedness[g] |
|---|---|---|---|---|---|---|
| Overall | 5,000 | 100 | 0.508 (0.007) | 0.045 (0.004) | 0.699 | 0.043 |
| | 2,000 | 100 | 0.501 (0.007) | 0.036 (0.003) | 0.672 | 0.046 |
| | 1,000 | 100 | 0.497 (0.007) | 0.024 (0.003) | 0.664 | 0.039 |
| | 2,000 | 25 | 0.489 (0.008) | 0.043 (0.005) | 0.640 | 0.084 |
| Within population | 5,000 | 100 | 0.470 (0.007) | 0.011 (0.001) | 0.568 | 0.125 |
| | 2,000 | 100 | 0.469 (0.006) | 0.010 (0.001) | 0.568 | 0.133 |
| | 1,000 | 100 | 0.479 (0.006) | 0.009 (0.001) | 0.572 | 0.125 |
| | 2,000 | 25 | 0.465 (0.007) | 0.010 (0.001) | 0.552 | 0.208 |
| Between population | 5,000 | 100 | 0.553 (0.009) | 0.017 (0.001) | | 0.050 |
| | 2,000 | 100 | 0.537 (0.009) | 0.015 (0.001) | | 0.056 |
| | 1,000 | 100 | 0.522 (0.008) | 0.013 (0.001) | | 0.074 |
| | 2,000 | 25 | 0.538 (0.010) | 0.014 (0.002) | | 0.134 |
| Cryptic | 5,000 | 100 | 0.510 (0.008) | 0.044 (0.004) | 0.693 | 0.056 |
| | 2,000 | 100 | 0.504 (0.008) | 0.031 (0.003) | 0.671 | 0.057 |
| | 1,000 | 100 | 0.500 (0.007) | 0.024 (0.003) | 0.648 | 0.055 |
| | 2,000 | 25 | 0.498 (0.009) | 0.041 (0.005) | 0.653 | 0.104 |

[a] Overall — all individuals of two-population sample compared; Within — individuals of same population compared; Between — individuals of different populations compared; Cryptic — random subset of 100 individuals compared.
[b] Number of generations since two populations diverged.
[c] Number of individuals sampled per population.
[d] Average genetic distance for a set of pairs of individuals; standard deviation reflects variation in that average across individuals.
[e] Standard deviation of individual CAPs, averaged across individuals.
[f] Heterozygosity $= 1 - \Sigma p_i$ where $p_i$ is the population frequency of allele i. Averaged across loci.
[g] Raggedness calculated according to Harpending.[37]

**Table 4.** Impact of gene flow on individual common ancestry profiles (CAPs) derived from coalescent simulations. Time of divergence of two populations — 2,000 generations; effective population size ($N_e$) — 1,000 individuals; sample size — 100 individuals per population. Standard deviations included in parentheses.

| Migration model[a] | Average[b] | Standard deviation[c] | Ht[d] | Raggedness[e] | Peak 1[f] | Peak2[f] | Peak 3[f] |
|---|---|---|---|---|---|---|---|
| $N_em = 0$ | 0.504 (0.008) | 0.031 (0.003) | 0.671 | 0.057 | 0.086 | 0.039 | 0.096 |
| $N_em = 0.5$ | 0.514 (0.010) | 0.027 (0.007) | 0.663 | 0.059 | 0.092 | 0.081 | 0.114 |
| $N_em = 2.0$ | 0.521 (0.013) | 0.018 (0.008) | 0.642 | 0.083 | 0.082 | 0.142 | 0.107 |
| CIRM | 0.509 (0.008) | 0.034 (0.004) | 0.667 | 0.059 | 0.116 | 0.038 | 0.110 |

[a] Rate of migration from population 2 into population 1. CIRM: **c**omplete **i**solation (1,900 generations) followed by **r**ecent **m**igration ($N_em = 2.0$).
[b] Average genetic distance for a set of pairs of individuals.
[c] Standard deviation of individual CAPs, averaged across individuals.
[d] Heterozygosity $= 1 - \Sigma p_i$ where $p_i$ is the population frequency of allele *i*, averaged across all alleles at all loci.
[e] Raggedness calculated according to Harpending.[37]
[f] Average weight of distribution (across individuals) in each of three sets of bins corresponding to peak at lower genetic distance (1), peak at higher genetic distance (3) and mid-point between these two peaks (2). See text for further details.

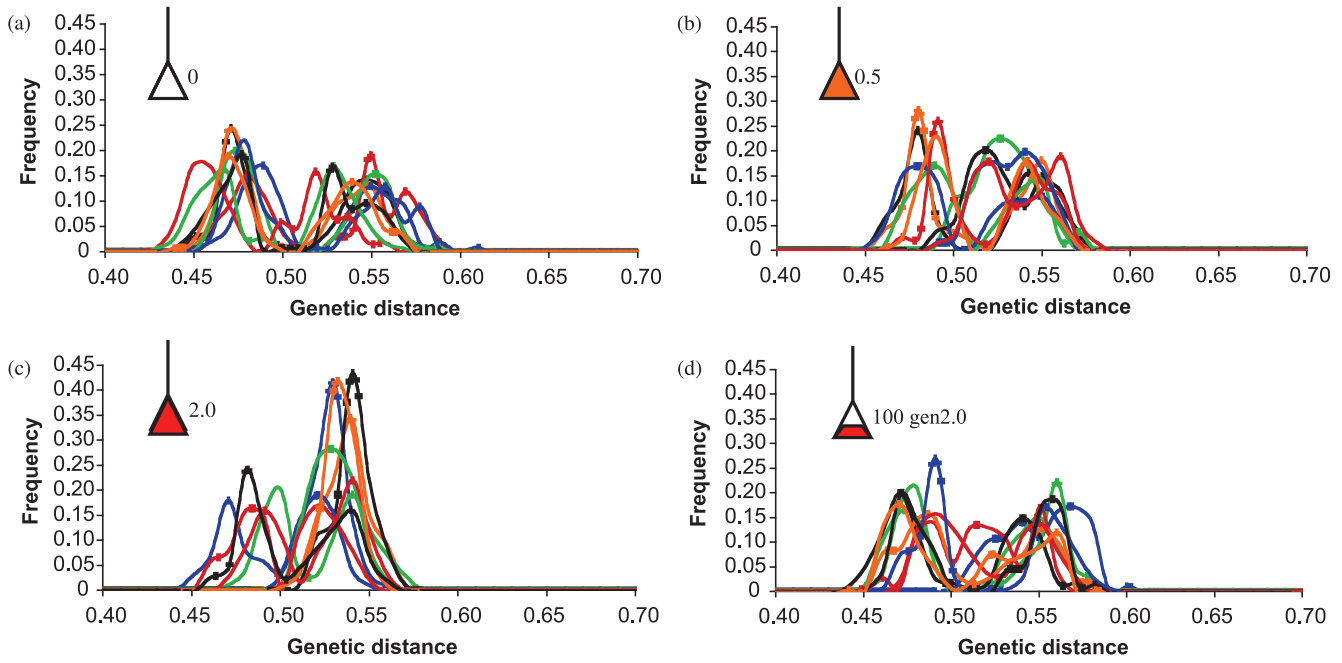


**Figure 4.** Ten examples of common ancestry profiles (CAPs) generated under each of four models of population history. Each 'cryptic' comparison set is based on 100 samples randomly selected from 200 possible samples in both populations, as might be realistic in the case of cryptic population structure. CAPs derived from coalescent simulations of two populations of effective size 1,000 that diverged 2,000 generations ago given: (a) complete isolation; (b) continuous gene flow at the rate of 0.5 migrants per generation; (c) continuous gene flow at the rate of 2.0 migrants per generation; and (d) gene flow over the past 100 generations at the rate of 2.0 migrant per generation, following 1,900 generations of isolation. Gene flow is asymmetrical. CAPs derived from simulated data for 500 short tandem repeat loci (mutation rate = 0.0005/locus/generation, range constraint = 15, stepwise mutation model).

higher average genetic distance for models with gene flow and higher standard deviation of genetic distance for models with recent migration. Additionally, migration leads to CAPs with multiple peaks for within-population comparisons.

## Data analysis

The overall CAP (based on genetic distance, $\hat{d}_{xy}$) for 1,013 individuals (512,578 pairs) is leptokurtotic and slightly positively skewed (Figure 1a, 'global'), with a median of 0.771 (mean of 0.772) and 5th to 95th percentile range of 0.732–0.816.

Two individual CAPs (Pima 1043 and French 516, each versus all other individuals in the dataset) illustrate the potential for variation across individual CAPs (Figures 1a and 1b). The overall distribution for the French individual (Figure 1b) is approximately normal, reflecting the overlap of the different CAPs shown in Figure 1d. The CAP for the Pima individual, however, is less symmetrical. The first peak from the left in the overall Pima 1043 CAP (Figure 1a) represents the comparison of Pima 1043 with other Pima, the second peak reflects comparison with non-Pima individuals in the Americas and the third small peak represents comparison with individuals outside the Americas (Figure 1c).

Group, or population, CAPs for 52 human populations are summarised in Figure 2. The distributions for the indigenous populations of the Americas and Oceania have the highest variances: pairs of individuals from the samples of populations in those regions have the broadest range of similarity estimates (Figure 2). All population samples from regions outside of the Americas and Oceania have similar levels of between-individual variation in terms of both mean and variance. There is a geographical trend, however, in that the genetic distance estimates for pairs of individuals from Africa are highest, followed by pairs from the Middle East and Europe. Pairs within East Asian populations tend to be slightly more similar to one another than pairs within African, Middle Eastern, European or Central/South Asian populations. Note that these distributions are dependent on the population labelling of individuals. We can compare the mean genetic distance across all pairs ($d_t = 0.772$) to the mean genetic distance between individuals within populations ($d_p = 0.740$) to obtain an estimate of between-population variation; $(d_t - d_p)/d_t = 0.041$ is an example of a ratio of differences recently discussed at length by Rousset.[32] The estimate is analogous to a standard $F_{ST}$, except that here within-individual variation is not considered.

Table 5 reports the largest genetic distance for any two individuals from each pair of geographical regions. The two most genetically dissimilar individuals in the dataset ($\hat{d}_{xy} = 0.861$) are an individual from Africa (Mbuti) and one from the

**Table 5.** Maximum genetic distance ($\hat{d}_{xy}$) between any pair of individuals drawn from each pair of geographical regions.

| | Africa | Mid East | Eur | C/S Asia | E Asia | Oc | Amer |
|---|---|---|---|---|---|---|---|
| Africa | 0.846 | 0.853 | 0.853 | 0.853 | 0.853 | 0.847 | 0.861 |
| Middle East | | 0.823 | 0.820 | 0.828 | 0.829 | 0.823 | 0.824 |
| Europe | | | 0.803 | 0.812 | 0.822 | 0.823 | 0.823 |
| Central/South Asia | | | | 0.811 | 0.818 | 0.821 | 0.822 |
| East Asia | | | | | 0.786 | 0.817 | 0.805 |
| Oceania | | | | | | 0.760 | 0.803 |
| Americas | | | | | | | 0.749 |

Americas (Pima). The two most different individuals in Africa (a Yoruba/Mbuti comparison with $\hat{d}_{xy} = 0.846$) are more different than any two individuals outside of Africa (a Han/Druze comparison with $\hat{d}_{xy} = 0.825$), consistent with our understanding of the high level of genetic diversity and population substructure within Africa. Mean genetic distance can be directly compared with degree of relationship in a small number of cases. CAPs of individuals in 19 populations were consistent with a relationship of degree 1 (siblings or parent–offspring pairs). Genetic distance ($\hat{d}_{xy}$) varied dramatically across these putative first-degree relative pairs (0.630–0.411). In fact, the two most dissimilar Surui individuals ($\hat{d}_{xy} = 0.419$) in the sample were estimated to be more similar than two putative first-degree relative pairs in African populations (one pair of Mbuti individuals and one pair of San individuals).

CAPs (Figure 5) and summary statistics (Table 6) vary across the Surui/Karitiana, Burusho/Kalash, Pima/Mbuti and Papuan/Biaka comparisons. Average within-population genetic distances are highest for the Biaka and Mbuti, intermediate for the Kalash, Burusho, Pima and Papuan and lowest for the Surui and Karitiana. The overall CAPs are bimodal for the Surui/Karitiana comparison and the Pima/Mbuti comparison. By contrast, the Burusho/Kalash comparison is unimodal, except for a small peak representing comparisons between more distant individuals. The Papuan/Biaka comparison is intermediate with two overlapping peaks.

## Discussion

CAPs are novel, graphical representations of within- and between-group variation from the perspective of the individual. Like population or individual trees and other clustering algorithms, CAPs provide insight into population genetic structure. Through simulation, we have generated expectations for CAPs for two-population models and evaluated the sensitivity of those expectations to sample size, divergence time and gene flow between the two populations. Simulations demonstrated that, for simple population histories, sample size has little influence on summary statistics characterising the distributions. This finding is particularly relevant for studies where population structure is cryptic, so that sample sizes of subpopulations are unknown. Sensitivity to sample size was considered in the context of complete isolation between populations. It is likely that more complex models including gene flow would lead to greater sensitivity to sample size.

The simulation study of the impact of divergence time on CAPs revealed that the average between-population genetic distance differs from the average within-population genetic distance to a greater extent for populations that diverged earlier in time. This finding is consistent with expectations for the change in $F_{ST}$ or population genetic distance over time. Ongoing gene flow has a different impact on CAPs than does recent gene flow following isolation: such recent gene flow leads to much broader distributions. Simulations presented above focused on two-population models, including unidirectional gene flow. Overall CAPs (where individuals in a reference population are compared with individuals in both the reference and the non-reference populations) generated given such models consist of three categories of genetic distance estimates. If the focal individual (for whom the CAP is generated) is an individual in the recipient population, these genetic distances correspond to the following comparisons: the focal individual versus individuals of the reference population with little immigrant ancestry; the focal individual versus individuals of mixed ancestry (in the reference population); and, finally, the focal individual versus individuals of the non-reference population. The magnitude and positions of the peaks resulting from these comparisons change as the amount of gene flow increases (Figure 4), suggesting that CAPs are informative regarding the rate of gene flow between populations. The more recent the population divergence, the lower the difference between the 'within' and 'between' genetic distances and, consequently, the less potential for recognizing gene flow. In situations where populations have diverged very recently, using a larger number of markers reduces the variance of a CAP and may, therefore, provide additional insight.

Simulations were designed to explore the impact of sample size and population processes on CAPs generated from STR

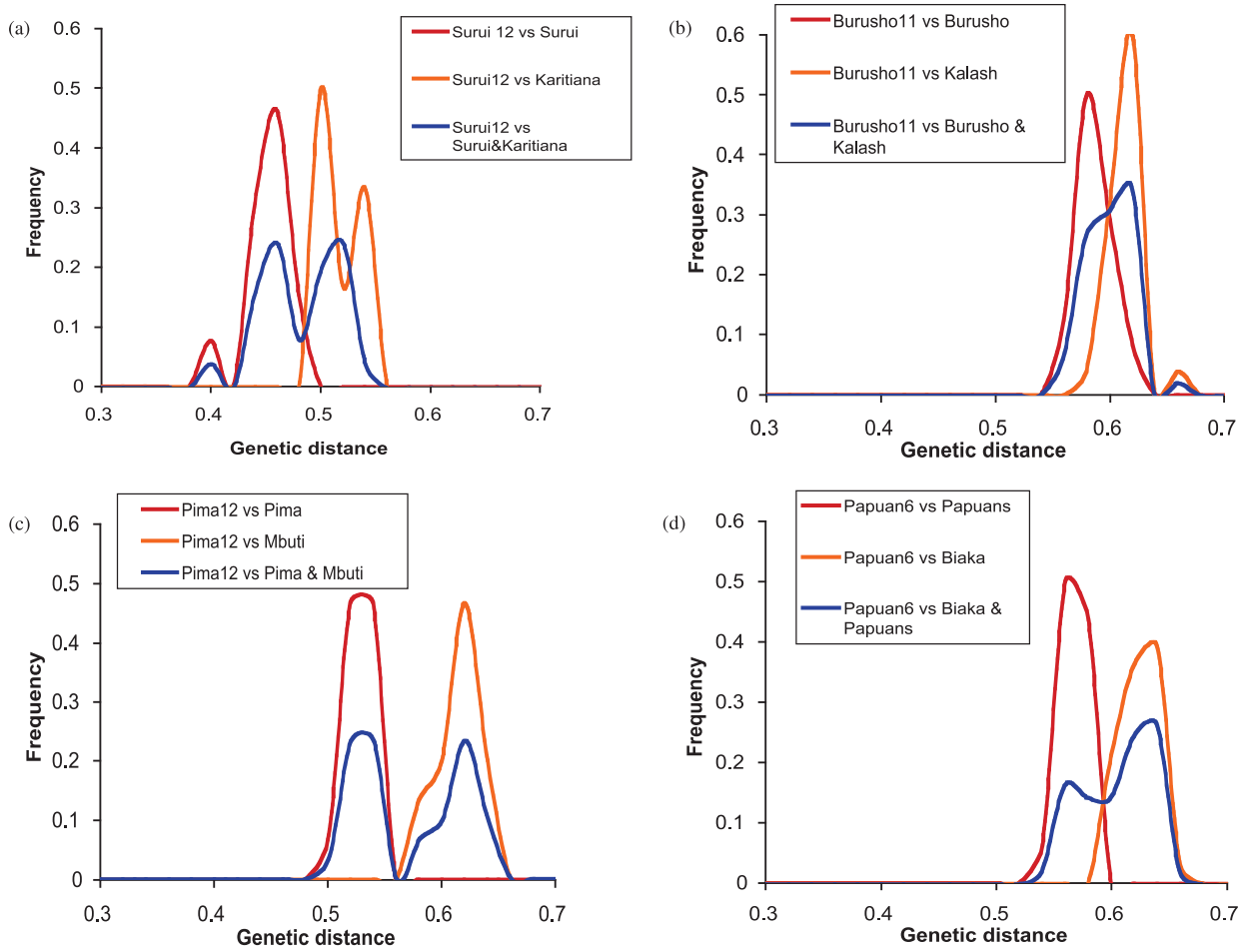


**Figure 5.** Common ancestry profiles (CAPs) for four individuals in the context of four pairs of populations, including geographically proximate populations, (a) Surui/Karitiana and (b) Burusho/Kalash, and geographically distant populations, (c) Pima/Mbuti and (d) Papuan/Biaka. Each figure illustrates a 'within', 'between' and 'overall' CAP for a focal individual. For example, the Surui/Karatiana comparison illustrates: (1) a Surui individual versus other Surui; (2) a Surui individual versus all Karitiana individuals; and (3) a Surui individual versus all Karitiana and all other Surui individuals.

multilocus genotypes. CAPs generated from SNP multilocus genotypes might differ from the STR-based profiles. Given the high heterozygosity of STR polymorphisms relative to SNPs, CAPs based on STRs are more likely to reach 'saturation' than are those based on SNPs. That is, the divergence between individuals is likely to approach an upper limit that depends on the mutation rate and range constraint, as well as population history. CAPs based on tens of thousands of SNPs may be more informative if recurrent mutation is rare. SNPs, however, are more likely to be subject to an ascertainment bias than are STRs. Given the impact of ascertainment bias on estimates of heterozygosity[38] and the correlation between heterozygosity and individual genetic distance (Table 3), such bias is very likely to influence CAPs.

Simulations presented in this paper assume randomly mating populations; however, there is extensive evidence for non-random mating, consanguinity and complex social structure (including matriarchy and patriarchy) in many human populations.[39–41] Given the potential for such demographic and sociocultural processes to influence individual genetic distances in real populations, models including more complex mating systems deserve further investigation. Other demographic factors, including population growth and population bottlenecks, are also likely to influence the shape of CAPs. Further simulations are required to assess the impact on CAPs of such demographic processes.

## CAPs for 1,013 human individuals

CAPs generated from CEPH-HGDP STR multilocus genotypes are consistent with known patterns of human genetic variation.[16] The overall CAP (based on the individual genetic

**Table 6.** Summaries of common ancestry profiles (CAPs) for four population pairs. Surui/Karitiana (Rondonia, Brazil) and Burusho/Kalash (Pakistan) are pairs of geographically proximate populations. Pima (North America)/Mbuti (Central Africa) and Papuan (Oceania)/Biaka (Central Africa) are pairs of geographically distant populations. Standard deviations are included in parentheses. For the 'between' and 'overall' comparisons, focal individuals are always drawn from the first population (ie Surui, Burusho, Pima and Papuan, respectively). Short tandem repeat data drawn from Rosenberg *et al.*[4]

| Surui vs Karitiana $n_1$=14, $n_2$=12 | Average | Standard deviation | Ht | Raggedness |
|---|---|---|---|---|
| Surui | 0.430 (0.011) | 0.017 (0.003) | 0.492 | 0.106 |
| Karitiana | 0.475 (0.013) | 0.016 (0.004) | 0.553 | 0.148 |
| Surui vs Karitiana | 0.492 (0.096) | 0.015 (0.004) | | 0.625 |
| Surui and Karitiana | 0.453 (0.015) | 0.033 (0.007) | 0.586 | 0.080 |
| Burusho, Kalash $n_1 = 25$, $n_2 = 25$ | | | | |
| Burusho | 0.577 (0.008) | 0.014 (0.002) | 0.703 | 0.263 |
| Kalash | 0.599 (0.008) | 0.013 (0.002) | 0.732 | 0.277 |
| Burusho vs Kalash | 0.598 (0.008) | 0.013 (0.002) | | 0.128 |
| Burusho and Kalash | 0.582 (0.009) | 0.019 (0.002) | 0.736 | 0.089 |
| Pima, Mbuti $n_1 = 16$, $n_2 = 15$ | | | | |
| Pima | 0.513 (0.008) | 0.014 (0.004) | 0.603 | 0.097 |
| Mbuti | 0.611 (0.004) | 0.012 (0.002) | 0.739 | 0.163 |
| Pima vs Mbuti | 0.608 (0.004) | 0.015 (0.003) | | 0.071 |
| Pima and Mbuti | 0.565 (0.025) | 0.040 (0.015) | 0.740 | 0.042 |
| Papuan, Biaka $n_1 = 33$, $n_2 = 17$ | | | | |
| Papuan | 0.547 (0.008) | 0.014 (0.002) | 0.673 | 0.117 |
| Biaka | 0.614 (0.008) | 0.016 (0.003) | 0.759 | 0.086 |
| Biaka vs Papuan | 0.605 (0.009) | 0.015 (0.003) | | 0.086 |
| Biaka and Papuan | 0.591 (0.014) | 0.025 (0.007) | 0.768 | 0.048 |

distance measure, $\hat{d}_{xy}$) for humans is slightly skewed in a positive direction (Figure 1a, 'global'). In light of the simulations, we can conclude that this positive skewness reflects subdivision within the species. If mating is random with respect to genomes, the variance of $\hat{d}$ is expected to be low. That is, most pairs of individuals are similarly divergent. Higher levels of substructure correspond to higher CAP variances. The concentration of genetic distance in a relatively narrow range (Figure 1a, 'global') is consistent with a generally low level of human population substructure (low $F_{ST}$); for pairs of individuals separated by more than three generations (ie most pairs), the genetic distance is very close to the overall average. Exceptions are in the lower tail of the distribution that includes pairs of closely related individuals. These exceptions include pairs of individuals in small populations that have undergone substantial random genetic drift, for instance during the peopling of the Americas. Heterozygosity is relatively low in indigenous populations of the Americas,[35] and two 'unrelated' individuals from such a population are far more similar than are two individuals chosen at random from anywhere else in the world. $F_{ST}$ estimates, because they reflect an average difference between groups, mask some of the between-population variation.[16] The analyses presented here highlight the variation not captured by summary statistics.

The highest genetic distance value overall is 0.861, for a pair of individuals including one affiliated with the Mbuti population and one affiliated with the Pima population. These individuals are also among the most geographically distant from one another if we measure geographical distance along a migration pathway out of Africa, east through Eurasia and then into the Americas. Population subdivision within Africa has been so high that the two most genetically dissimilar individuals in Africa are more dissimilar than any two individuals outside of Africa, but not so high that those two

individuals are the most dissimilar overall. As indicated in Table 5, the region with the greatest divergence between individuals (Africa) is also the region with highest heterozygosity. The pairs of individuals with the largest genetic distances vary depending on the distance metric (results not shown).

Many of the population samples included in the CEPH-HGDP panel were included for anthropological interest. These populations are often small, more isolated than most ethnic/linguistic groups and considered to be the indigenous peoples of a region. They can be considered valuable with regard to understanding human genetic variation, in that they probably represent the extremes in terms of effective size and degree of isolation and, therefore, individual genetic distance.

In some cases, population profiles indicate deviations from simple models of population history. Profiles for several populations of the Americas and Oceania are much broader than those of other regions (Figure 2), possibly reflecting population substructure. As noted above, samples for the CEPH-HGDP cell line panel are distributed with indication that the Karitiana, Surui, Mayan and Pima samples include relative pairs. The data analyses described above do not include known relative pairs; however, reanalysis including sets of closely related individuals led to more highly skewed profiles (results not shown).

CAPs analysis revealed that the CEPH-HGDP sample set includes 13 duplicate samples. Such detection of duplicate samples is best carried out using a distance measure that gives a distance of 0 between two genetically identical (or almost identical, if occasional genotyping errors have occurred) individuals.

The four population pairs considered in detail illustrate the diversity of human CAPs. The CAPs (Figure 5), summarised in Table 6, can be interpreted in light of the simulations. Average genetic distances are consistent with high effective size for both the Biaka and the Mbuti, intermediate effective sizes for the Kalash, Burusho, Pima and the Papua New Guineans and low effective sizes for the Surui and the Karitiana. Figure 5 reveals older divergence for the geographically distant population pairs (given the difference between the average 'within' and 'between' genetic distances) compared with geographically proximate population pairs. For geographically distant population pairs, unimodal, distinct CAPs for the 'within' and 'between' comparisons indicate lack of gene flow. Overlapping 'within' and 'between' CAPs for geographically proximate populations are consistent with more recent divergence of these groups. The Kalash and Burusho, for example, seem to have similar effective sizes and to have diverged relatively recently. The second peak of the 'overall' comparison corresponds to a high genetic distance, indicating the presence of some particularly distant Burusho individuals. The Karitiana 'within' CAP is also bimodal, with one peak having a lower than average genetic distance. This peak could correspond to comparisons between local Karitiana and other local Karitiana (and the other peak corresponds to local Karitiana compared with Karitiana with recent migrant ancestry). Alternatively, the first peak may reflect inbreeding within the Karitiana population. The other two comparisons, Pima versus Mbuti and Papuan versus Biaka, reveal very high levels of variability in the African populations, consistent with previous analyses of these and other data.

## Applications

The pattern of human population genomic variation is relevant in a number of research and education contexts. As noted above, the pattern reflects — and therefore may provide insight into — population history. In medical genetics, knowledge of any genetic substructure of a set of probands may inform research decisions. In forensics, an understanding of patterns of genetic variation is becoming increasingly relevant as institutions attempt to infer racial or ethnic affiliation of individuals using DNA data.[42] In secondary and undergraduate education, the discussion of race and genetics has typically been highly superficial. As publicly available data regarding genetic information accrue, a basic understanding of human population genetic variation becomes an increasingly important component of public education.

When the research goal is to take into account cryptic population subdivision, as in case-control studies, genomic controls[43–45] or clustering approaches (eg *structure*) are appropriate; however, these approaches may not always reveal a small fraction of individuals that stands out from the rest in terms of genetic distance. The *structure* approach, for instance, is sensitive to sample size.[4] The CAPs approach may more readily reveal anomalies such as duplicate samples and closely related pairs of individuals.

CAPs vary both within and across geographically and socially defined groups. The profiles indicate that some population labels serve as better proxies for genetic similarity than do others. That is, some linguistic or social groups consist of individuals much more genetically similar to one another than to individuals of other groups, while other groups do not. Emphasis on absolute description of variation can be valuable, in that continuity of the measurement is naturally emphasised. The individual-based CAPs approach also emphasises the shared ancestry of all humans: all pairs of individuals fall into the continuum of genetic distance. Finally, although the CAPs approach does not require *a priori* information regarding an individual's affiliation with one group or another, the approach does allow us to explore hypotheses regarding the correspondence between genetic and non-genetic dimensions of human variation.

CAPs can be considered as genomic versions of pairwise difference distributions for single DNA sequence loci.[46,47] These genomic profiles enable us to consider both within- and between-group variation simultaneously and to complement traditional summary statistics in revealing differences among

individuals in the variances of individual CAPs (eg Figures 1 and 5). Although most information regarding population genetic structure is captured in a sufficiently hierarchical analysis of variance,[16] CAPs reveal, in addition, information at the genealogical level. While CAPs are no replacement for traditional population genetic summary statistics, direct estimation of gene flow (eg *BayesAss*[23]) or direct inference of degree of relationship (eg *relpair*[31]), they serve as a valuable exploratory tool and as an independent check of estimates derived using other methods.

## Acknowledgments

We thank Alec Knight, Neil Risch, Noah Rosenberg and an anonymous reviewer for helpful discussion and suggestions regarding a previous version of this manuscript. Research was supported in part by NIH grant GM28428 and NSF grant BCS 9905574 to J.L.M.

## References

1. Cavalli-Sforza, L.L., Piazza, A. and Menozzi, P. (1994), *History and Geography of Human Genes*, Princeton University Press, Princeton, NJ.
2. Salas, A., Richards, M., De la Fe, T. *et al.* (2002), 'The making of the African mtDNA landscape', *Am. J. Hum. Genet.* Vol. 71, pp. 1082–1111.
3. Underhill, P.A., Passarino, G., Lin, A.A. *et al.* (2001), 'The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations', *Ann. Hum. Genet.* Vol. 65, pp. 43–62.
4. Rosenberg, N.A., Pritchard, J.K., Weber, J.L. *et al.* (2002), 'Genetic structure of human populations', *Science* Vol. 298, pp. 2381–2385.
5. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000), 'Inference of population structure using multilocus genotype data', *Genetics* Vol. 155, pp. 945–959.
6. Venter, J.C., Adams, M.D., Myers, E.W. *et al.* (2001), 'The sequence of the human genome', *Science* Vol. 291, pp. 1304–1351.
7. Li, W.H. and Sadler, L.A. (1991), 'Low nucleotide diversity in man', *Genetics* Vol. 129, pp. 513–523.
8. Zhao, Z., Jin, L., Fu, Y.X. *et al.* (2000), 'Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22', *Proc. Natl. Acad. Sci. USA* Vol. 97, pp. 11354–11358.
9. Yu, N., Zhao, Z., Fu, Y.X. *et al.* (2001), 'Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1', *Mol. Biol. Evol.* Vol. 18, pp. 214–222.
10. Lewontin, R. (1972), 'The apportionment of human diversity', In: Hecht, M.K. and Steere, W.S. (eds.), *Evolutionary Biology*, Vol.6, Plenum, New York, NY.
11. Jorde, L.B., Watkins, W.S., Bamshad, M.J. *et al.* (2000), 'The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data', *Am. J. Hum. Genet.* Vol. 66, pp. 979–988.
12. Barbujani, G., Magagni, A., Minch, E. and Cavalli-Sforza, L.L. (1997), 'An apportionment of human DNA diversity', *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 4516–4519.
13. Romualdi, C., Balding, D., Nasidze, I.S. *et al.* (2002), 'Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms', *Genome Res.* Vol. 12, pp. 602–612.
14. Excoffier, L. and Hamilton, G. (2003), 'Comment on genetic structure of human populations', *Science* Vol. 300, pp. 1877, author reply 1877.
15. Wilder, J.A., Kingan, S.B., Mobasher, Z. *et al.* (2004), 'Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males', *Nat. Genet.* Vol. 36, pp. 1122–1125.
16. Long, J.C. and Kittles, R.A. (2003), 'Human genetic diversity and the nonexistence of biological races', *Hum. Biol.* Vol. 75, pp. 449–471.
17. Edwards, A.W. (2003), 'Human genetic diversity: Lewontins fallacy', *Bioessays* Vol. 25, pp. 798–801.
18. Cavalli-Sforza, L.L. and Edwards, A.W. (1967), 'Phylogenetic analysis. Models and estimation procedures', *Am. J. Hum. Genet.* Vol. 19, pp. 233–257.
19. Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J. *et al.* (1994), 'High resolution of human evolutionary trees with polymorphic microsatellites', *Nature* Vol. 368, pp. 455–457.
20. Mountain, J.L. and Cavalli-Sforza, L.L. (1997), 'Multilocus genotypes, a tree of individuals, and human evolutionary history', *Am. J. Hum. Genet.* Vol. 61, pp. 705–718.
21. Shriver, M.D., Kennedy, G.C., Parra, E.J. *et al.* (2004), 'The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs', *Hum. Genomics* Vol. 1, pp. 274–286.
22. Rannala, B. and Mountain, J.L. (1997), 'Detecting immigration by using multilocus genotypes', *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 9197–9201.
23. Wilson, G.A. and Rannala, B. (2003), 'Bayesian inference of recent migration rates using multilocus genotypes', *Genetics* Vol. 163, pp. 1177–1191.
24. Rosenberg, N.A. (2004), 'Distruct: A program for the graphical display of population structure', *Mol. Ecol. Notes* Vol. 4, pp. 137–138.
25. Queller, D.C. and Goodnight, K.F. (1989), 'Estimating relatedness using genetic markers', *Evolution* Vol. 43, pp. 258–275.
26. Lynch, M. and Ritland, K. (1999), 'Estimation of pairwise relatedness with molecular markers', *Genetics* Vol. 152, pp. 1753–1766.
27. Li, C.C., Weeks, D.E. and Chakravarti, A. (1993), 'Similarity of DNA fingerprints due to chance and relatedness', *Hum. Hered.* Vol. 43, pp. 45–52.
28. Weeks, D.E., Young, A. and Li, C.C. (1995), 'DNA profile match probabilities in a subdivided population: When can subdivision be ignored?', *Proc. Natl. Acad. Sci. USA* Vol. 92, pp. 12031–12035.
29. Milligan, B.G. (2003), 'Maximum-likelihood estimation of relatedness', *Genetics* Vol. 163, pp. 1153–1167.
30. Cann, H.M., de Toma, C., Cazes, L. *et al.* (2002), 'A human genome diversity cell line panel', *Science* Vol. 296, pp. 261–262.
31. Epstein, M.P., Duren, W.L. and Boehnke, M. (2000), 'Improved inference of relationship for pairs of individuals', *Am. J. Hum. Genet.* Vol. 67, pp. 1219–1231.
32. Rousset, F. (2002), 'Inbreeding and relatedness coefficients: What do they measure?', *Heredity* Vol. 88, pp. 371–380.
33. Johnson, R.A. and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ.
34. Excoffier, L., Novembre, J. and Schneider, S. (2000), 'SIMCOAL: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography', *J. Hered.* Vol. 91, pp. 506–509.
35. Zhivotovsky, L.A., Rosenberg, N.A. and Feldman, M.W. (2003), 'Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers', *Am. J. Hum. Genet.* Vol. 72, pp. 1171–1186.
36. Ramakrishnan, U. and Mountain, J.L. (2004), 'Precision and accuracy of divergence time estimates from STR and SNPSTR variation', *Mol. Biol. Evol.* Vol. 21, pp. 1960–1971.
37. Harpending, H.C. (1994), 'Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution', *Hum. Biol.* Vol. 66, pp. 591–600.
38. Mountain, J.L. and Cavalli-Sforza, L.L. (1994), 'Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms', *Proc. Natl. Acad. Sci. USA* Vol. 91, pp. 6515–6519.
39. Storz, J.F., Ramakrishnan, U. and Alberts, S.C. (2001), 'Determinants of effective population size for loci with different modes of inheritance', *J. Hered.* Vol. 92, pp. 497–502.
40. Hussain, R. and Bittles, A.H. (1998), 'The prevalence and demographic characteristics of consanguineous marriages in Pakistan', *J. Biosoc. Sci.* Vol. 30, pp. 261–275.

41. Bittles, A.H. (2002), 'Endogamy, consanguinity and community genetics', *J. Genet.* Vol. 81, pp. 91–98.
42. Cho, M.K. and Sankar, P. (2004), 'Forensic genetics and ethical, legal and social implications beyond the clinic', *Nat. Genet.* Vol. 36, pp. S8– 12.
43. Devlin, B. and Roeder, K. (1999), 'Genomic control for association studies', *Biometrics* Vol. 55, pp. 997–1004.
44. Pritchard, J.K. and Rosenberg, N.A. (1999), 'Use of unlinked genetic markers to detect population stratification in association studies', *Am. J. Hum. Genet.* Vol. 65, pp. 220–228.
45. Reich, D.E. and Goldstein, D.B. (2001), 'Detecting association in a case-control study while correcting for population stratification', *Genet. Epidemiol.* Vol. 20, pp. 4–16.
46. Slatkin, M. and Hudson, R.R. (1991), 'Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations', *Genetics* Vol. 129, pp. 555–562.
47. Rogers, A.R. and Harpending, H. (1992), 'Population growth makes waves in the distribution of pairwise genetic differences', *Mol. Biol. Evol.* Vol. 9, pp. 552–569.