Inference of ancestry: Constructing hierarchical reference populations and assigning unknown individuals

Jayne E. Ekins,[1] *Jacob B. Ekins,*[1] *Lara Layton,*[1] *Luke A.D. Hutchison,*[1] *Natalie M. Myres*[1] *and Scott R. Woodward*[1,2]*

[1] Sorenson Molecular Genealogy Foundation, Salt Lake City, UT, USA
[2] Department of Molecular Biology and Microbiology, Brigham Young University, Provo, UT, USA
**Correspondence to*: Tel: +1 801 461 9789; Fax: +1 801 461 9722; Email: scott@smgf.org

Date Received: 11th April 2005

Abstract

The ability to infer personal genetic ancestry is being increasingly utilised in certain medical and forensic situations. Herein, the unsupervised Bayesian clustering algorithms *structure,* is employed to analyse 377 autosomal short tandem repeats typed on 1,056 individuals from the Centre d'Etude du Polymorphisme Humain Human Diversity Panel. Individuals of known geographical origin were hierarchically classified into a framework of increasingly homogeneous clusters to serve as reference populations into which individuals of unknown ancestry can be assigned. The groupings were characterised by the geographical affinities of cluster members and the accuracy of these procedures was verified using several genetic indices. Fine-scale substructure was detectable beyond the broad population level classifications that previously have been explored in this dataset. Metrics indicated that within certain lines, the strongest structuring signals were detected at the leaves of the hierarchy where lineage-specific groupings were identified. The accuracy of unknown assignment was assessed at each level of the hierarchy using a 'leave one out' strategy in which each individual was stripped of cluster membership and then re-assigned using the supervised Bayesian clustering algorithm implemented in GeneClass2. Although most clusters at all levels of resolution experienced highly accurate assignment, a decline was observed in the finer levels due to the mixed membership characteristics of some individuals. The parameters defined by this study allowed for assignment of unknown individuals to genetically defined clusters with measured likelihood. Shared ancestry data can then be inferred for the unknown individual.

Keywords: *population genetics, human population structure, clustering, Bayesian inference, short tandem repeats (STRs)*

Introduction

Hypervariable microsatellite markers, situated across the autosomes, have been shown to produce stronger resolution for high-level differentiation of populations when compared with biallelic markers.[1] Several expanded studies have demonstrated the usefulness and accuracy with which multi-locus microsatellites can define genetic groupings that correspond well with geographical and other proxy designations;[2–7] however, the resolution of such studies has been variable. Using 377 autosomal loci and Bayesian clustering methods, Rosenberg *et al.*[6] demonstrated genetic differentiation among major continents and the ability within certain localities to identify a subpopulation as a single genetic grouping from other geographically adjacent populations.

Although the confounding effects of homoplasy have provided conflicting analyses at times, microsatellites are generally considered to be highly informative genetic markers for high-resolution population differentiation studies. Using extensive multi-locus genotypes from a worldwide population sampling, within-population variance (0.930−0.950)[6] dominates the total variance of the world population. Although it has been demonstrated that homoplastic mutations increase the likelihood of common identical by state alleles among unrelated individuals, thereby reducing variance for the individual, adjusted estimates for within-population variance (0.812−0.854)[8] still exceed between-population variance. This indicates an overall similarity between populations — as defined by current geopolitical or other proxy designations — and strong variance within such populations. Within-population variance is expected to decrease, however, when examining populations based on hierarchical genetic similarities rather than proxy definitions. Further, a systematic hierarchical analysis of the genetic composition of subpopulations would allow groups that have strong genetic homogeneity to be identified and reveal relationships that are probably due to extended familial ties. These relationships may persist across geopolitical borders, but are expected to produce a genetic framework for describing the biological links detected between members of the total dataset.

This framework represents a set of reference populations with which individuals with unknown personal ancestry can be compared and assigned to a most-likely population. Personal origin at a certain level of resolution and potential shared ancestry data can be inferred for the unknown individual by the characterisation of the members of its matching reference cluster. Using the Centre d'Etude du Polymorphisme Humain (CEPH) Human Diversity Panel dataset, we describe and validate its decomposition into fine-scale resolution reference populations to which unknown individuals can be assigned with measured likelihood, revealing relevant ancestral information on a more recent time scale for the unknown individual.

Subjects and methods

The microsatellite data consisted of 377 autosomal loci typed on 1,056 individuals from 52 populations worldwide. Data correlated to samples' geographical origins were obtained from the CEPH Human Diversity Panel resource. This has previously been used to address different hypotheses in several other analyses.[1,6,9–13] Samples were excluded from analysis according to reported instances of mislabelling.[6,14]

Analysis of the dataset was accomplished using *structure v.2.0.*[4] In order to obtain fine-scale resolution clusters, a hierarchical breakdown of samples was performed on each cluster and then subcluster, assembling increasingly homogeneous groupings of individuals until the lowest level could not be decomposed further. The substructure of the dataset was treated as unknown; all runs in *structure* were performed with the 'no pop info' parameter and no proxy designations were applied during the hierarchical analysis of the samples. For all runs, the *structure* algorithm was applied with a burn-in of 10^3 and with 10^5 data collection steps. The admixture and correlated allele frequencies models[15] and an infer-alpha prior of 1 (for use in determining K) were used in all runs. While *structure* accounts for the correlation of alleles due to divergence from a common ancestral population, correlations induced by gene flow across unrelated populations are ignored. Using the current dataset, Fu *et al.*[12] found high correlations among populations which could lead to overestimation of genetic differentiation among the populations. A new mixture model was proposed which estimates correlation explicitly due to shared history and gene flow from which future analyses are likely to benefit. Also within *structure*, missing data are ignored when estimating allele frequencies and admixture proportions for individuals in populations at each update. Admixture proportion estimates are less accurate for individuals having missing data, but the exclusion of such individuals is not recommended unless most of the genotype data are missing.[16] Considering that in this set 98 per cent of the samples had ≥ 90 per cent complete data, it was determined to include the total complement of individuals for this investigation. Alternate methodologies for managing missing data have been explored in the literature. Yang *et al.*[13] found *structure* performed better in a dataset with 3.27 per cent missing data overall, when subjects were restricted to those having only complete data. A separate analysis package, BAPS 2, handles missing allelic points through data augmentation.[11,17] Future applications of the current dataset would probably benefit from similar strategies.

A fundamental difficulty in using unsupervised clustering methods such as *structure* is that the user must specify the number of clusters (*K*) into which to partition the data and then determine which solution best represents the data. To address this in our analysis, a selection method was developed, defining criteria with which to select an appropriate *K* for the data at each level of analysis. The strategy for selecting *K* was two-fold, representing a balance between maximising the posterior probability of the solution while taking into account the similarity of the solutions produced for independent runs with identical input and parameters (see Appendix). Alternate analysis packages offer related Bayesian methods that avoid this difficulty, where allele frequencies, individual assignments and *K* are estimated simultaneously.[11,17,18] Possibilities exist for future comparative analysis of results using the different available methods.

Genetic characterisation of clusters

At each level of analysis, several types of quantitative genetic data were collected as each parent cluster broke into multiple child clusters to demonstrate that partitions applied to the data were productive in assembling close genetically-related individuals and excluding others. Measures collected for this dataset included F_{ST} values, intra- and inter-cluster allele-sharing statistics and average gene diversity (*H*) of clusters.[19] *H* was also used to characterise genetic affinity within small clusters suspected to contain members from extended family groups. Intra-population allele-sharing statistics were collected for each individual i as the mean of the number of shared alleles between i and i_1 to i_n within a cluster. Likewise, inter-population measures were calculated as the mean number of shared alleles between i and i_1 to i_n in all sibling clusters. Inter-population allele-sharing statistics were calculated using sibling clusters only, as individuals in these clusters were more closely related to the cluster of interest than individuals in more distantly differentiated clusters, producing a stricter measure with which to indicate between-cluster allele sharing.

Unknown sample assignment

Following the decomposition of the total dataset, all clusters were fixed at each level of resolution, giving discrete hierarchical population assignments to each individual. Using a 'leave one out' method, a member of the dataset was stripped of its cluster designation and then used as an unknown individual to be assigned to the set of reference populations in order to test the ability to assign individuals to the reference populations with accuracy. Population assignments were made by an analytical Bayesian assignment algorithm implemented in GeneClass2[20] using Rannala and Mountain priors.[21,22] As with the admixture model implemented in *structure*, this algorithm allows for partial assignment of individuals to several populations, based on the differential assignment of chromosomal fragments to clusters. In GeneClass2, the probability of an individual being assigned to a cluster is the product of the probability of membership of each locus under consideration.[22] Success was achieved if the majority of the individual's assignment was to the original cluster into which it was placed by *structure*; failure occurred if it was assigned into a different cluster. Each individual was tested in this manner at each level of resolution to approximate the ability accurately to assign a truly unknown test sample to the cluster likely to have the most shared ancestry.

Results

Geographical/lineage-specific characterisation of reference clusters

The decomposition of the dataset was documented over five rounds of analysis, representing progressively tighter levels of resolution. Using the defined selection criteria to choose *K*, the total dataset initially broke into seven major clusters, representing geographically distinct continental designations and two isolated populations:

1. The Biaka Pygmy lineage from Congo.
2. All other sub-Saharan African individuals.
3. An Oceanic population.
4. Individuals from East Asia.
5. One diverse cluster containing European, North African, Middle Eastern and Pakistani samples.
6. Individuals from the Americas.
7. A lineage representing a linguistic isolate from Pakistan, the Kalash.

Common characteristics observed in the many features of the subsequent decomposition of the seven major lines include the following:

1. Initially broad population differentiations produced clusters that were remarkably homogeneous for sample origin, both by geography and lineage.[2,6]
2. Subsequent clustering attempts deeper in the hierarchy tended to produce clusters in which samples had mixed membership in multiple clusters.
3. Mixed membership was most often observed within densely sampled areas in populations that were geographically intermediate to distant populations.
4. Some highly specific geographical groups and lineages were completely isolated from all other populations at fine levels of resolution in the hierarchy, even if they were geographically proximal to other sampled groups.
5. Purported extended family groups showing strong genetic homogeneity were extracted from among population lineages through the recursive application of the hierarchical analysis (Table 1).

Further characterisation of the geographical and lineage-specific decomposition is presented in the Supplemental Material (see below), in the context of the seven major lines identified after the first round of analysis.

Genetic characterisation of clusters

Data for several genetic indices were collected to document the extent of differentiation of clusters as they were identified through the hierarchy. Figure 1 illustrates mean F_{ST} values measured over the course of the hierarchy for each of the seven major lines. These data points are summary values for all subclusters that branched out from the initial main line defined at the second tier of analysis. In the course of the decomposition, F_{ST} values remained high. Although at certain midpoints the Middle East/European/Pakistani, African and East Asian clusters demonstrated weaker measures of differentiation between the proposed clusters, they experienced their highest F_{ST} ratings towards the end of the hierarchy. Substructure detected at each tier was tested by 10^4 permutation steps ($p < 0.00001$).

H was also measured and summarised as a mean value for the seven main lines at each level in the hierarchy (Figure 2). Generally, these measures illustrated a decrease in the diversity of the composition of the clusters over the course of the decomposition.

Intra- and inter-cluster allele sharing data were collected as each parent cluster divided into *K* child clusters and analysed in relative frequency histograms. Using window-smoothing techniques, the resulting curves were observed as approximately normal (Figure 3). Median within- and between-population measurements for each curve were observed as point estimates indicating the extent of differentiation of the newly defined subpopulation from its sibling clusters. Figure 4 summarises the mean number of pairwise matches observed within and between clusters over the course of the hierarchy for all seven major lines. An increase in intra-cluster allele sharing was observed over the course of the hierarchy for all lines. Inter-cluster allele sharing also increased, but, on average, intra-cluster statistics were always greater.

Observations were also made as to the productivity of population separation by analysing the degree of differentiation of the inter- and intra-cluster distributions. The null hypothesis that the allele sharing distributions for the sampled populations were the same was tested at a level of significance (α) of 0.05. Measured β-values quantified the probability of incorrectly accepting the null hypothesis when allele sharing distributions for the inter- and intra-cluster populations

Table 1. Quantitative description of purported family lineages identified in the CEPH Human Diversity Panel. Recursive hierarchical analysis allowed identification of the tightly related subgroups found in the sampling of the major geographic areas of the world.

World region	Origin	Cluster size	Median no. shared LT[a]	*H*	RASR[b]
Africa	Bantu	2	215.50	0.65782	1.00
	Biaka Pygmy	2	227.50	0.64728	1.00
	Biaka Pygmy	2	215.50	0.65875	1.00
	Biaka Pygmy	2	211.50	0.66667	1.00
	Pygmy	5	210.75	0.62612	0.60
	Mbuti	2	207.00	0.64632	1.00
	San	2	220.00	0.65378	1.00
	Yoruba	2	238.00	0.63951	1.00
	Yoruba	3	207.00	0.66114	1.00
Oceania	Melanesian	4	246.33	0.52982	1.00
	Melanesian	4	233.58	0.53272	1.00
	Melanesian	3	219.18	0.54069	1.00
The Americas	Colombian	2	262.00	0.49862	1.00
	Colombian	2	256.00	0.5105	1.00
	Colombian	2	250.50	0.49907	1.00
	Colombian	2	244.00	0.52927	1.00
	Colombian	2	241.50	0.48823	1.00
	Colombian	3	203.83	0.58459	0.67
	Karitiana	2	254.00	0.51019	0.50
	Karitiana	3	250.50	0.47578	0.67
	Karitiana	2	243.50	0.51824	0.00
	Karitiana	6	222.83	0.51286	1.00
	Karitiana	5	218.10	0.52119	0.80
	Karitiana	5	201.00	0.5734	0.80
	Mayan	2	233.00	0.62134	0.00
	Mayan	4	199.67	0.60093	0.75
	Pima	2	268.00	0.54087	1.00
	Pima	2	246.50	0.5277	0.50
	Pima	4	244.33	0.49139	1.00
	Pima	3	235.00	0.49215	1.00
	Pima	5	233.60	0.51385	0.80

(*continued*)

Table 1. *Continued.*

World region	Origin	Cluster size	Median no. shared LT[a]	*H*	RASR[b]
	Pima	3	211.50	0.54383	0.67
	Surui	3	245.00	0.46837	0.67
	Surui	5	238.40	0.45546	0.80
	Surui	7	234.14	0.41629	1.00
	Surui	5	229.10	0.47286	1.00
Middle East	Bedouin	2	227.00	0.5893	1.00
	Druze	2	240.00	0.55994	1.00
	Druze	2	227.50	0.57143	1.00
	Druze	2	220.50	0.62389	0.50
	Mozabite	2	227.50	0.6185	1.00
	Palestinian	2	233.00	0.62232	1.00
Europe	French	2	210.00	0.62526	0.00
	Orcadian	2	229.50	0.60214	1.00
Pakistan	Balochi	2	224.50	0.6185	0.50
	Kalash	2	230.50	0.5859	0.00
	Sindhi	2	218.50	0.60568	1.00
East Asia	Cambodian	2	204.50	0.60903	0.00
	Lahu	2	233.00	0.59637	0.00
	Lahu	2	218.50	0.6125	0.00
	Naxi	2	234.00	0.5964	0.00
	Oroqen	2	221.50	0.61186	0.00

[a] LT = locus types
[b] RASR = Re-assignment success rate

were distinct. The β-values are an indicator of the extent of genetic divergence of the newly formed sibling clusters. Examination of four of the seven main lines — the Biaka Pygmy, Oceanic, American and Kalash clusters and their subclusters — shows that, on average, there was excellent differentiation of the intra-cluster peak from the inter-cluster peak throughout the hierarchy. The African population demonstrated a fairly high β-value with its initial division, indicating weaker differentiation of the intra- and inter-cluster distributions. The East Asian line demonstrated strong separation from its sibling clusters with the initial division; with succeeding rounds of analysis, however, the mean β-value indicated less genetic distinction between subsequently divided populations. Similarly, genetic differentiation was weak in the Middle Eastern/European/Pakistani main line, but increased over the course of the hierarchy. Subclusters having similarly diverse genetic composition were the result of partitions where β-values were high.

Unknown sample assignment

To approximate the ability to assign unknown individuals to the structured reference clusters, a 'leave one out' method was utilised and results from these tests were quantified (Table 2). In the first three levels of the hierarchy, nearly all samples were re-assigned to their original clusters. In subsequent levels, a drop in success of re-assignment was observed; however a large proportion of clusters in all levels of the hierarchy continued

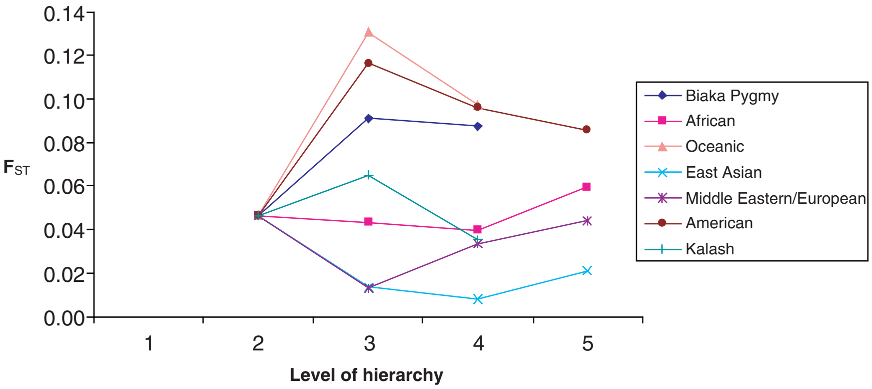


Figure 1. Quantification of hierarchical substructure detected through seven major lines: mean F_{ST}. Lines from Oceania, the Americas and Biaka Pygmy exhibited strong mean F_{ST} statistics at each level of the hierarchy. The remaining lines demonstrated lower F_{ST} values but increased over the progression of the hierarchy, indicating that there were stronger population substructure signals at the finer levels of resolution beyond the initial continental population groupings. The differential termination of the main lines at various levels of resolution was due to differential conclusion of analysis for the main lines in the course of the decomposition.

to perform well (>90 per cent success rate). In observing the rate of success over the course of the hierarchy for the total dataset, it was acknowledged that the mis-assignment of particular individuals might indicate that more stable genetic groupings exist than those defined in the original decomposition of the dataset. When mis-assigned individuals' cluster membership classifications were adjusted to those defined in re-assignment tests by GeneClass2,[20] an overall increase in success rate was noted at all levels of resolution in subsequent re-assignment procedures. This step was termed 'post-clustering adjustment'.

Discussion

A method has been described integrating supervised and unsupervised Bayesian clustering algorithms in which unknown individuals can be assigned to likely reference populations for the purpose of inferring personal ancestry data.

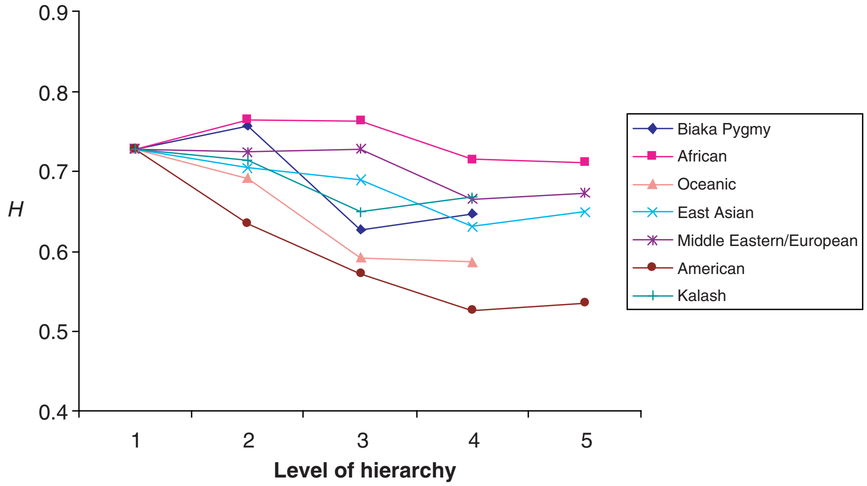


Figure 2. Change in average gene diversity (*H*) over the hierarchical decomposition of the dataset. In general, the average gene diversity for each line decreased with each subsequent level of resolution. This demonstrated the successful assembly of informative clusters, wherein the most closely related individuals were assembled and more diverse individuals were excluded, even at the finest levels of resolution. The differential termination of the main lines at various levels of resolution was due to differential conclusion of analysis for the main lines in the course of the decomposition.

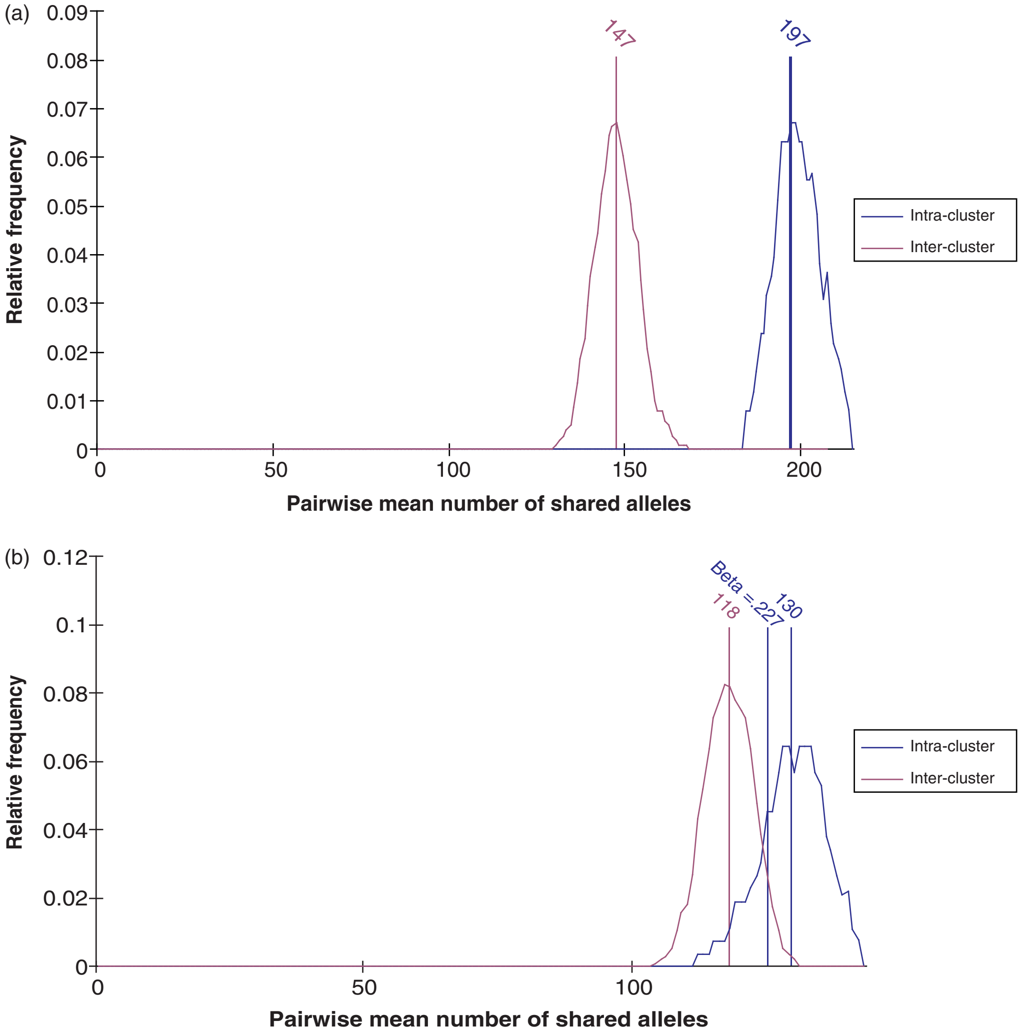


Figure 3. Normalised pairwise mean allele sharing within clusters and among sibling clusters. It is common for substantial overlap to be observed between the distributions, but intra-population measures are typically greater than inter-population measures.[23] (a) Within Karitiana, among other sibling clusters from the Americas. This population subdivision defined by *structure* demonstrated strong population differentiation, even at finer levels of resolution, such that the distributions were completely separated. (b) Within Israel–Negev, among European and other Middle Eastern sibling clusters. Other population divisions defined by *structure* showed more overlap between the distributions, indicating that although there was strong homogeneity within the intra-cluster population, similar alleles were found in sibling clusters, further indicating that the populations were more closely related. The degree of overlap was quantified to describe the relatedness of the samples being partitioned.

Geographical and genetic evaluations confirm the efficacy of the procedure.

Geographical/lineage-specific characterisation of reference clusters

Several trends were observed when cluster composition was analysed by correlating geographical origins to cluster members. In all populations sampled, small groups were identified that exhibited markedly increased levels of homogeneity (Table 1). These clusters were generally identified deep in the hierarchy as they separated from larger population groupings. They are suspected to contain members of extended family groups collected in the sampling, which form a further level of genetic resolution within already highly differentiated populations.

At broader levels of resolution, it was observed that many of the populations that experienced high levels of differentiation were either isolated by distance in the sampling (Yakut, Russian, Mozabite etc) or were populations likely to have experienced prolonged periods of genetic isolation (Basque, Biaka, Orcadian, American etc). This was evident in the

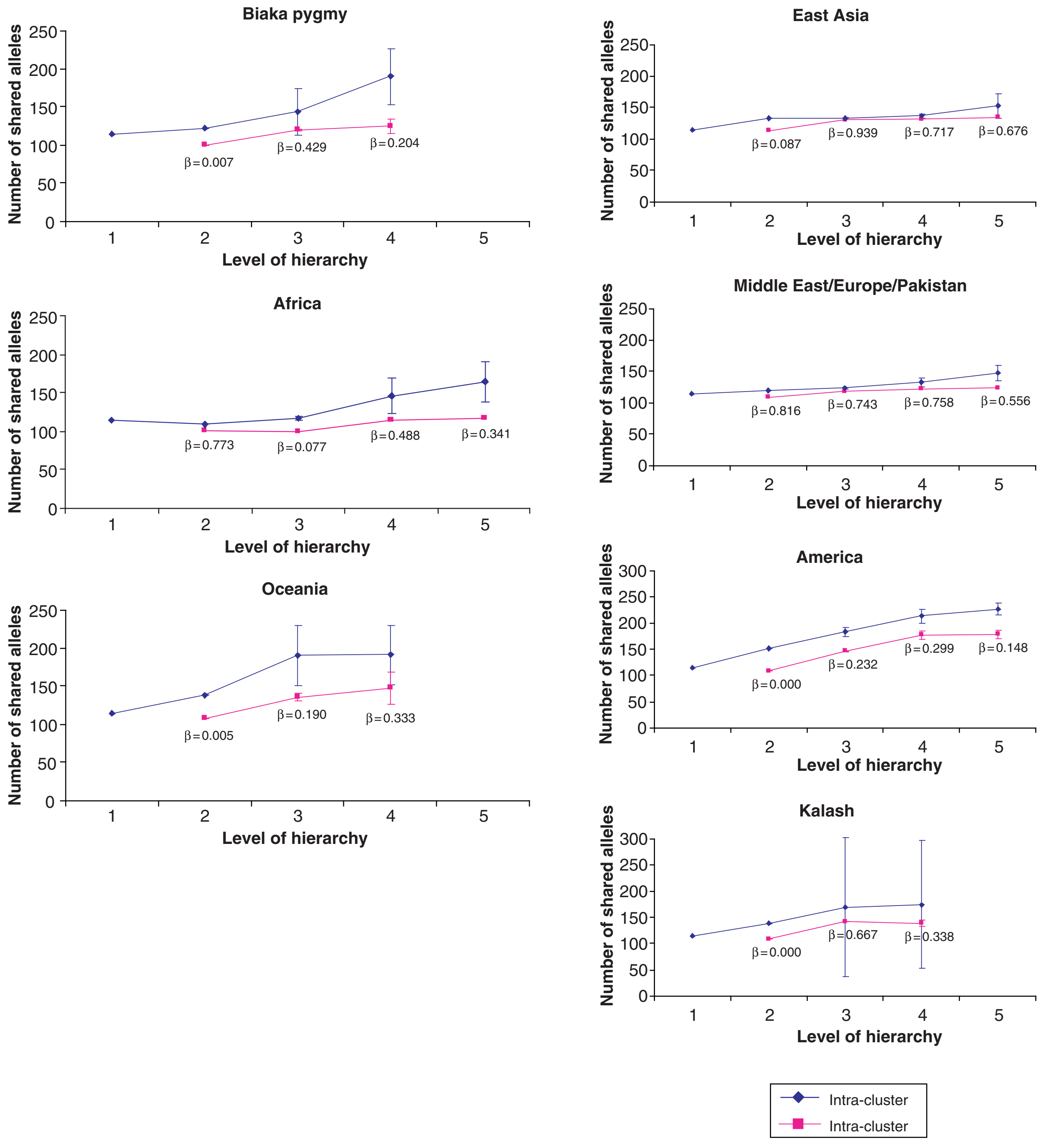


Figure 4. Quantification of hierarchical substructure detected through seven major lines: intra-, inter-cluster allele sharing. In general, mean intra-cluster allele sharing increased over the course of the hierarchy and was always greater than the allele sharing observed among sibling clusters. The 95 per cent confidence intervals for mean allele sharing statistics are also shown and most often did not overlap for inter- and intra-cluster measurements. These data gave another indication of the high productivity of the partitions imposed by *structure*, even at fine levels of resolution. (Note: the margin of error is wide in the Kalash line, due to the small number of clusters ($K = 3$) defined for the sample.)

Table 2. Re-assignment characteristics of unknown samples to reference clusters defined in the hierarchical decomposition of a dataset by *structure*. Re-assignment success rates were high over the first several levels of resolution in the analysis. The decrease in re-assignment success over the course of the hierarchy was primarily attributed to an increase in the number of samples exhibiting mixed membership properties among multiple clusters. Despite the decline in overall success, a large proportion of the defined clusters at each tier demonstrated highly successful re-assignment rates. These high-performance clusters have the potential to provide highly informative assignments for truly unknown individuals in ancestry testing procedures. Additionally, post-clustering adjustments, wherein certain outlier samples were systematically relocated, had the effect of increasing success rates at all levels of resolution.

	Raw re-assignment						Post-clustering adjustment re-assignment			
Level of resolution	**No. clusters**	**Median cluster size**	**RASR**[a]	**No. clusters >90% RASR**	**Criterion**	**R^2**	**No. clusters**	**Median cluster size**	**RASR**	**No. clusters >90% RASR**
1	1	1,056	1.000	—	—	—	—	—	—	—
2	7	86	0.995	6	IMM[b]	0.900	7	86	0.999	6
					Cluster size	0.427				
3	30	19	0.939	21	IMM	0.931	29	18	0.960	23
					Cluster size	0.216				
4	78	4	0.818	40	IMM	0.943	72	3	0.823	39
					Cluster size	0.296				
5	113	3	0.633	47	IMM	0.898	87	3	0.700	42
					Cluster size	0.662				

[a] RASR = Re-assignment success rate
[b] IMM = Individual mixed membership (in multiple clusters)

composition of the final clusters of the four major European groups, where anchoring populations for each cluster consisted of Orcadian (cluster S4D-1), Russian (cluster S4D-2), Basque (cluster S4E-1) and Sardinian (cluster S4E-2) populations (see Supplemental Material). The balance of cluster composition was supplemented by various other populations having geographically intermediate placement in the sampling or less isolated genetic histories. Geographically intermediate populations often exhibited mixed genetic characteristics of geographically distant population anchors. The geographically intermediate Caucasus Adygei were found to be admixed between European and Central Asian populations, displaying transitional genetic characteristics of both of the anchoring populations. This was, to some extent, also observed with Middle Eastern samples, being geographically central to European and Central Asian populations.

Densely sampled Pakistan populations exhibited some population substructure down to fine levels of resolution. Population-specific clusters were anchored by the Burusho, Brahui and Hazara, but otherwise it was difficult to detect clear genetic distinctions among the other tribal groups. The density of sampling when compared with other world regions may prevent the separation of distinct populations observed elsewhere. This situation was also observed in southern China, where sampling was dense. It is possible that the different hierarchical properties observed within populations can be attributed to the incongruent sampling schemes observed within the data, where sampling of closely-related populations detects a genetic gradient that diminishes the ability to resolve populations according to proxy designations. By contrast, some broad population-level groupings have a high degree of within-population substructure due to sampling of populations that are distantly isolated from one another, or due to thorough sampling of population isolates that are likely to have accumulated distinct allele frequencies. Further population structure is observed within some clusters as a result of the sampling of related individuals, causing stratification even at fine levels of population resolution.

Meanwhile, the substructure of other populations, having individuals with mixed membership in multiple populations and other transitional attributes, is more difficult to resolve. This may not be due to intrinsic population characteristics but rather to dense sampling of geographically contiguous populations, likely to have more admixed properties among the sampled members. The sampling from the Americas shows a low degree of within-population variance and high between-population differentiation.[6] Although this has been demonstrated in this dataset previously, the representation of the Americas is fragmentary and would be expected to produce strong quantitative differentiation by virtue of the geographical isolation of the sampled populations and the sampling of population isolates, as well as the existence of multiple family groups within these populations. When contrasted with the dense representation of populations in East or Central Asia, the quantitative measures showing weaker differentiation of subpopulations are put into context.

In some populations, distinct genetic characteristics were detected despite the tight geographical proximity of the samples. Among the three Middle Eastern groups sampled, the Druze (cluster S4A-4) were the most distinct; the Palestinians (cluster S4B-1) and Negev Bedouins (cluster S4A-6) also formed some exclusive associations (see Supplemental Material). Although geographically close, these populations have been genetically isolated from one another as demonstrated in the ability to identify genetic differentiation.

Common trends observed in the decomposition of the sampled world populations suggest population substructure characteristics that may be observed in the hierarchical decomposition of other sample datasets. Initial broad breaks in the hierarchy were successful in isolating genetically disparate populations that were typically separated by large physical distances. Subsequent rounds of analysis further down the hierarchy were attempted on populations that had stronger genetic similarities than in previous tiers. This resulted in divisions that exhibited more overlap and mixed characteristics between the newly formed subclusters. Additionally, mixed membership among subclusters was most often observed in densely sampled areas among populations that were geographically intermediate to distant populations. Population isolates often anchored highly differentiated clusters that contained various portions of more heterogeneous groups. Regions that were represented by distantly sampled populations showed clearer differentiation among subgroups. Through hierarchical analysis, however, it was possible, even within geographically proximal populations, to completely isolate certain populations into their own unique subcluster. Additionally, smaller groups of individuals that show strong genetic homogeneity can be identified from larger population groups, and probably exhibit the presence of closer familial relationships. A hierarchical approach to detecting population structure allowed fine-resolution clusters to be defined, representing more recent relationships from among the dataset. The genetic clusters defined at all levels of resolution can serve as a template with which unknown samples can be compared and assigned for the purposes of ancestry testing. The potential for highly informative assignments exists due to the genetic, geographical and lineage-specific composition of many of the clusters identified in the total dataset.

Genetic characterisation of clusters

Descriptive genetic values identified through the progression of the hierarchy demonstrated that the imposed population partitions were productive in assembling closely-related members and excluding genetically dissimilar individuals from the constructed reference populations.

F_{ST} values were useful in observing trends in substructure detected at the various levels of the hierarchy (Figure 1). All of the main lines were characterised either by high F_{ST} values at all points in the hierarchy or values that increased over the course of the decomposition. Even at the highest levels of resolution, F_{ST} values for the Oceanic and American samplings indicated discrete subpopulations, probably due to increased levels of relatedness among the sampled individuals. Although Middle East/European/Pakistani, African and East Asian clusters demonstrated weaker differentiation between the proposed clusters at certain midpoints in the decomposition, they experienced their highest F_{ST} ratings at the leaves of the hierarchy. This was attributed to the strong genetic diversity seen within the populations in the early cluster divisions. Initial partitions within these lines were productive in imposing separations between geographically sensible populations, even though the within-population diversity was still strong. Continued hierarchical analysis formed groupings that preserved cohesion among strongly related individuals, thus leading to the stronger F_{ST} values detected at the leaves of the hierarchy. This analysis revealed the substructure properties of each of the seven major lines examined in this study and also supported the composition of the subpopulations over the hierarchy, demonstrating, on average, strong F_{ST} values showing high significance levels by permutation tests.

Complementary to the findings indicated by the F_{ST} metrics obtained over the hierarchy, *H* measures also showed a decrease in diversity and an increase in homogeneity within the newly defined subclusters at each level of resolution (Figure 2). As with the other metrics, this provided an empirical validation of the proposed structure of the hierarchy, even at the finest levels of resolution.

Within- and between-cluster allele sharing data, and β-values measuring the differentiation of inter- and intra-cluster distributions, were collected as each parent cluster divided into *K* child clusters as a means of expressing quantitatively the productivity of the imposed partitions (Figure 4). For all lines, the mean number of intra-cluster shared alleles increased over the course of the hierarchy but was less pronounced in the East Asian group. High variance within populations probably contributed to the difficulty in resolving substructure within this population compared with others analysed. On average, however, the mean number of shared alleles within a cluster did increase with new population subdivisions for all seven major lines. This indicated that the subdivisions introduced at each new tier were productive in assembling closely-related individuals and excluding others. As previously indicated, the within-population median is usually greater than the between-population point estimator.[23] On average, this was observed in the analysis. This gives an indication that the individuals within clusters are more genetically similar than individuals found outside of their own cluster, thus supporting the composition of the clusters constructed over the hierarchy.

β-values were also used to gauge the extent of population differentiation detected among newly defined clusters. A total overlap of inter- and intra-cluster distributions indicated that the population division into *K* child clusters produced a new subpopulation definition that was genetically indistinct from its sibling clusters, suggesting that substructure could not be detected in the sampling. A total separation of the distributions represented the achievement of ultimate productivity in defining subpopulations that demonstrated strong separation from sibling clusters, indicating the *K* divisions of the parent cluster were most productive in the defined partitions (Figure 3). Examination of four of the seven main lines — the Biaka Pygmy, Oceania, American and Kalash clusters and their subclusters — showed that, on average, there was excellent separation of the intra-cluster peak from the inter-cluster peak throughout the hierarchy. This demonstrated the presence of strong subpopulations having distinctive allele sharing characteristics from other closely-related and recently differentiated groups. This supported the population partitions determined by *structure* even at the most specific levels of resolution where familial groups were extracted. The African population exhibited a fairly high β-value with its initial division. This was due to strong diversity among samples within populations. Further rounds of analysis helped to differentiate these groups into subpopulations that were more genetically homogenous and excluded others into sibling clusters. The East Asian line demonstrated strong separation from its sibling clusters with the initial division. With succeeding rounds of analysis, however, the mean β-value indicated weaker differentiation between subsequently divided populations. As analysis progressed down the hierarchy for this line, the populations were increasingly similar, making it more difficult to obtain distinct subpopulations with distinct genetic composition. Similar properties were observed in the Middle Eastern/European/Pakistani main line.

These analyses provided quantitative support to the composition of the reference clusters defined over the hierarchy of the decomposition of the sampling. In general, it was observed that the partitions defined were productive in assembling individuals of close genetic similarity and excluding others. In many cases, the strongest genetic substructure of many lines was observed beyond the broad levels of resolution, revealing distinct genetic groupings and previously unrecognised extended family groupings to which unknown individuals have the potential to be assigned. Ancestry for the unknown can then be inferred from a characterisation of the matching cluster contents.

Unknown sample assignment

In performing re-assignment tests using GeneClass2,[20] several trends were observed, revealing characteristics of the dataset and its decomposition. The rate of success was high in early levels of analysis and, as it declined over the course of the hierarchy, an attempt was made correctly to attribute the causes to particular factors. The size of clusters decreased through the progression of the decomposition analysis. To test if small reference cluster size adversely affected the re-assignment tests, correlation coefficients were calculated to assess the relationship of success in re-assignment to cluster size. These indicated only a weak positive association (Table 2). Further examination showed that generally smaller clusters had highly successful re-assignment rates, probably due to decreased diversity within the groups suspected to contain members from a common lineage (Table 1). Among other quantities tested, the strongest correlation was observed between success rate of re-assignment and the degree of mixed membership observed for individuals in clusters. Membership coefficients calculated by *structure* indicated the probability of assignment of an individual to the newly defined child clusters. Individuals having membership in multiple clusters, rather than a strong signal in a single cluster, were more likely to fail on re-assignment.

Marked improvements in re-assignment rates at all levels of resolution were seen subsequent to post-clustering adjustment. Many of the individuals that achieved more stable population cluster definitions had mixed membership signals in the early stages of the hierarchy. Post-clustering adjustment allowed these individuals to be placed in a more stable grouping, as reflected in increased re-assignment success rates; however, some samples were still mis-assigned after this adjustment. As the hierarchy progressed, partitions in clusters were made among increasingly similar groups, creating a situation where individuals tended to have mixed membership in multiple clusters. Thus, when complex relationships existed among the individuals, precise population assignments to a single group were less representative of the properties of the dataset. The degree of mixed membership among multiple clusters for an individual was a good predictor of re-assignment success. The re-assignment characteristics of the various geographical samples were explored. Regional characteristics were seen in individuals that were mis-assigned in this test, both in terms of the frequency of mis-assignment and in the degree of mis-assignment relative to the original cluster placement (see Supplemental Material).

We have outlined a new approach to the challenge of inferring ancestry for individuals of unknown origin. The two-stage method integrates the novel use of an unsupervised algorithm to construct a hierarchical framework of reference clusters with a supervised algorithm to perform cluster assignment of the unknown, as suggested by Baudouin *et al.*[22] Unknown individuals can be assigned to any level of resolution desired, with the potential for high probability assignment to a highly informative cluster defined in the hierarchy. How informative these cluster assignments are varies with the lineage and geographical specificity of cluster members and the confidence in the membership of the cluster definitions. The confidence in cluster composition can be estimated by the stability observed when performing self-assignment procedures. Many clusters were shown to have stable group membership, in that all individuals were successfully re-assigned to the same cluster, where others showed more volatility. The probability of cluster composition was estimated for each cluster in the hierarchical framework by observing the proportion of successful re-assignments when the defined population structure was subjected to self-assignment tests. Together with the likelihood calculated by GeneClass2[20] in the assignment of the unknown, this estimation contributed multiplicatively to the probability of unknown sample assignments to its matching reference cluster. This allowed appropriate weighting of unknown assignments to reference clusters that have less stable composition of cluster members. Confidence scores for assignment of the unknown obtained in this manner were generally comparable to the original probability of assignment estimated by *structure* at the time of the individual's placement into the hierarchical framework. Although some reference clusters will not allow confident assignment of unknowns — due to a low probability of reference cluster membership — many clusters that have the potential for highly confident and informative assignment of unknown samples exist at all levels of resolution in the hierarchy (Table 2).

Conclusion

With the entire hierarchical structure definition as a parameter, unknown individuals have the potential to be assigned to highly resolved cluster definitions that represent specific localities and also likely family groups. The presented level of resolution gives new insight to characteristics of this dataset that were not revealed in previous analyses using different techniques. As in this investigation, Rosenberg *et al.*[6] used *structure* with the admixture model. Using an alternate clustering strategy, Corander *et al.*[11,17] utilised mixture modelling and laid conditions on the geographical sampling of individuals such that subjects from the same site were assigned as a group to the product clusters. At the global and population levels of resolution, much of the broad structure detected was similar in all studies (Table 3). With the hierarchical strategy presented here, further resolution of the Mandenka, Orcadian and Russian populations generally into a single cluster was observed, while the Adygei were not seen as a genetically distinct group as observed elsewhere. Fine-scale substructure was detected beyond broad population level classifications using the current approach, which identified lineage-specific and extended family level groupings, many of which were detected at the leaves of the hierarchy (see Tables 1 and 3). Such specific groupings, quantitatively supported by strong differentiation measures from genetic indices, can serve as candidate reference populations to which unknown individuals can be assigned with measured likelihood and highly informative shared ancestry data can be learned.

It is anticipated that these findings will be of significance to individuals that have an interest in delineating more recent

Table 3. Multi-level structure partitions of genetically distinguishable geography-based groupings identified in the sampling. Geographically-defined groups at global and population levels, in which the strong majority of individuals were uniformly and uniquely assigned to a single cluster, were considered genetically distinguishable from others in the sampling. Lineage level structure was identified when small subgroups of individuals demonstrating elevated indications of relatedness (see Table 1) were found within geographically-defined groups.

Resolution	Geographic grouping
Global level	Africa,[a,b] Oceania,[a,b] the Americas,[a,b] Middle East/Europe,[a,b] East Asia[a,b]
Population level	Basque,[a] Biaka Pygmy,[a] Burusho,[a] Colombian,[a] Druze,[a] Japanese, Kalash,[a] Karitiana,[a,b] Lahu,[a] Mandenka, Maya,[a] Mbuti,[a] Melanesian,[a] Mozabite,[a] Orcadian, Palestinian,[a] Papuan,[a] Pima,[a] Russian, San,[a] Sardinian,[a] Surui,[a,b] Yakuta Adygei[c]
Lineage level (Table 1)	Balochi, Bantu, Bedouin, Biaka Pygmy, Cambodian, Colombian, Druze, French, Kalash, Karitiana, Lahu, Mayan, Mbuti, Melanesian, Mozabite, Naxi, Orcadian, Oroqen, Palestinian, Pima, San, Sindhi, Surui, Yoruba

[a] Structure partition also identified in Rosenberg *et al.*[6]
[b] Structure partition also identified in Corander *et al.*[11]
[c] Structure partition identified in Rosenberg *et al.*[6], but not in current analysis

genetic relationships that surpass the broad population level classifications that are regularly explored. High-resolution genetic groupings and unknown assignment strategies may be valuable in such global applications as disease linkage studies, parameters for extended familial studies, for use as proxy medical classifications to assess epidemiological risk and in forensic applications.[24]

Supplemental material

Geographic characterisation of subclusters

Clusters at each level of analysis were characterised by individual members' lineage and geographical origin. Cluster composition was depicted on a Mercator map projection, with latitude/longitude plot points for each individual represented proportionally on the map. Ellipses representing the geographical coverage of each cluster, weighted by sample density, were constructed and assisted in identifying the populations most heavily represented in the proposed clusters. The length and angles of the major and minor axes were calculated from a covariance matrix as two standard deviations from the calculated geographic centroid, weighted by sample density. Map points were plotted using planiglobe, Online Map Creation (OMC) software.

Biaka Pygmy

The Biaka Pygmy lineage was a highly specific genetic grouping, excluding all other geographically proximal African populations in its initial association. Further rounds of analysis produced four small subgroups containing two to five members that demonstrated increased levels of homogeneity (Table 1).

Sub-Saharan Africa

All other sub-Saharan African individuals, with some mixture from Middle Eastern individuals, clustered together in a main line separate from the Biaka Pygmy. This set was found to be best partitioned at $K = 3$. In the resulting subclusters, the Mbuti Pygmy population (cluster S1A-3) and all individuals of the San lineage (cluster S1A-2) formed two completely isolated genetic clusters (Figure S1A). The Bantu in Kenya and the geographically distant Yoruba and Mandenka belonged to the same genetic cluster at this level in the hierarchy (cluster S1A-1). At the next level of resolution, most of the Mandenka (cluster S1B-2) were distinguished into their own subgroup (Figure S1B). Among the geographical subgroups, five small clusters were identified ranging in size from two to three individuals, all showing decreased gene diversity (Table 1).

Oceania

Oceanic populations (Figure S2) demonstrated a sharp division between the geographically isolated subpopulation of Papua New Guinea (cluster S2-1) and Melanesians on the Bougainville Islands (cluster S2-2). The genetic sampling of these populations is highly structured and allows for an unequivocal division, not only between the major geographical regions, but also within the resulting subgroups. Three small subgroups were identified within the geographical clusters, demonstrating increased homogeneity (Table 1).

The Americas

Likewise, the American sampling was found to be highly structured. Of the five populations represented, the majority of samples were uniformly assigned to clusters with others from the same origin population with little evidence of gene flow (Figure S3). Only the Colombian and Karitiana had a slight proportion of samples clustering with the Maya (cluster S3-1), the rest forming their own exclusive association (cluster S3-2). The Pima population (cluster S3-3) was completely distinct — and even the geographically proximal Karitiana (cluster S3–5) and Surui (cluster S3-4) showed

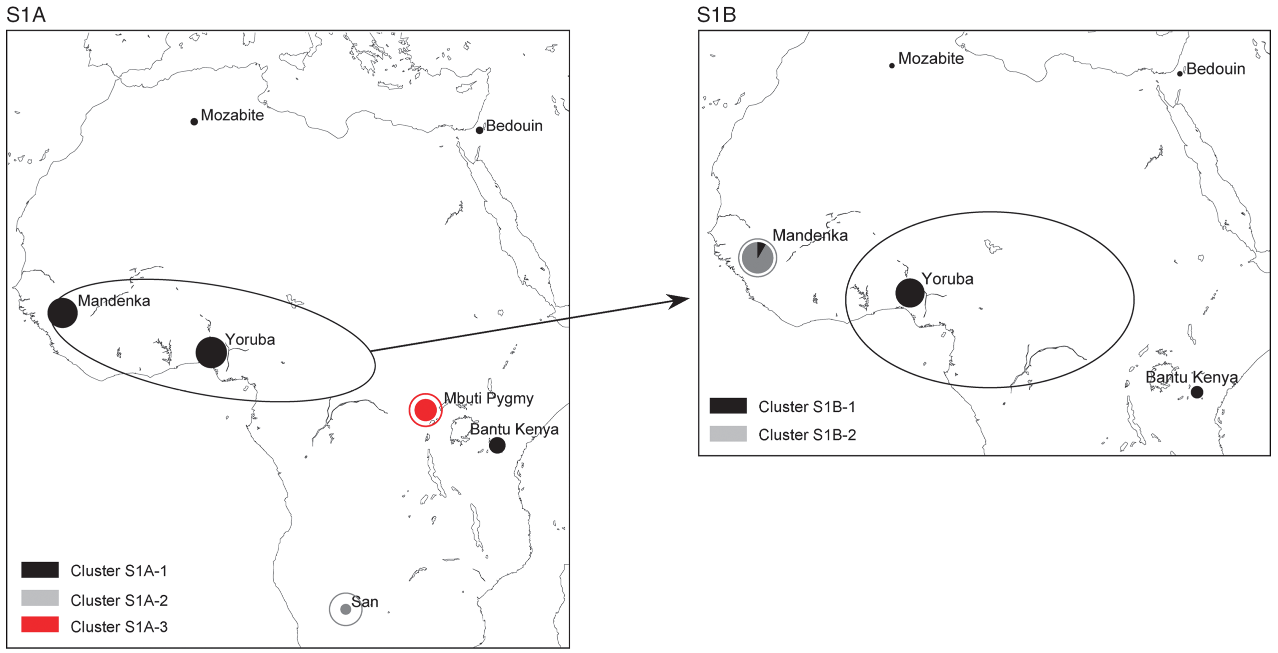


Figure S1. Geographical coverage of hierarchical genetic subgroups detected in Sub-Saharan Africa. (a) *K* = 3 was found as the best representation of substructure at this level in the hierarchy. Clusters were specific to geography and population, such that the San and Mbuti formed distinct clusters, excluding individuals from all other populations. (b) The next level in the hierarchy allowed cluster S1A-1 to be subdivided into *K* = 2 genetic groups, where the majority of Mandenka form an exclusive genetic grouping from the remaining population representatives.

highly differentiated population characteristics. Small subgroups demonstrating decreased diversity were detected within each of the geographical subpopulations, as expected with the explicit familial sampling strategy for these localities.[9] Twenty-four such groups were identified in the American sampling (Table 1).

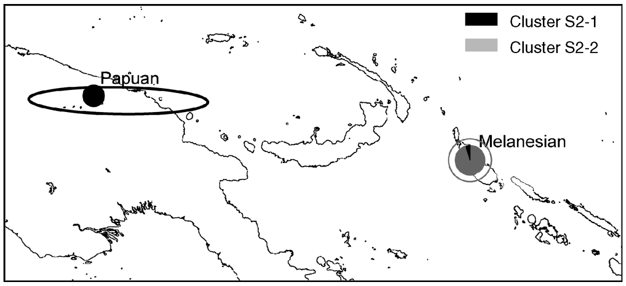


Figure S2. Geographical coverage of hierarchical genetic subgroups detected in Oceania. A near-complete separation of the geographically isolated Papuans and Melanesians was obtained with *K* = 2 subgroups. Further genetically homogeneous subgroups were found in cluster 4-2, probably indicating the presence of extended family groups within the sampling (Table 1).

Middle Eastern/European/Pakistani

The second level of hierarchical analysis (Figure S4A) divided this geographically diverse cluster into six groups: a European group (cluster S4A-1), a Mozabite group (cluster S4A-5), a Pakistani cluster (cluster S4A-2), a group containing the majority of the Druze (cluster S4A-4), a cluster consisting exclusively of half of the Israeli–Negev Bedouins (cluster S4A-6) and the final cluster containing the majority of Palestinians with the balance of the Israeli–Negev (cluster S4A-3). Further subdivision of the Palestinian anchored Middle East cluster (Figure S4B) forms an exclusive subcluster composed of mostly Palestinian individuals (cluster S4B-1).

The next level of analysis of the European samples (Figure S4C) initially produced a subcluster anchored by Russian and Orcadian individuals (cluster S4C-1), and a separate cluster delineated principally by Basque and Sardinian populations (cluster S4C-2). Other groups sampled in Europe tend to be more intermediate in their affinities, having members belonging to both major subclades. Further analysis (Figures S4D, S4E) split the Russian (cluster S4D-1), Orcadian (cluster S4D-2), Basque (cluster S4E-1) and Sardinian (cluster S4E-2) populations into their own clusters, supplemented by various members from other European populations.

A north–south geographical cline in cluster membership was detected in the initial separation of the Pakistani sampling

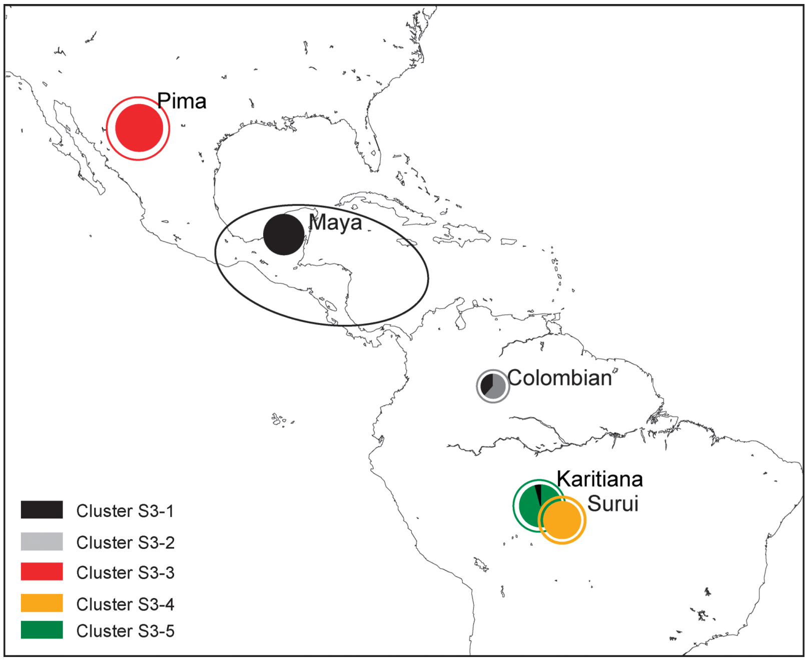


Figure S3. Geographical coverage of hierarchical genetic subgroups detected in the Americas. Clear differentiation is seen among virtually all populations sampled. Strongly homogeneous subgroups were found in all subclusters (Table 1).

into three groups (Figure S4F): a northern group anchored by the Burusho and Hazara (cluster S4F-3); a southern cluster anchored by the Brahui (cluster S4F-2); and a central population that was less exclusive having members from multiple Pakistani populations (cluster S4F-1). Ultimately, in the final levels of the hierarchy (Figures S4G, S4H, S4I), distinct clusters were anchored by the Brahui (cluster S4G-1), the Hazara (cluster S4I-1), and the Burusho (cluster S4I-2), while other clusters exhibited various combinations from other populations.

East Asian

After the first partition, the East Asian cluster (Figure S5A) divided into two subclusters representing a northern region, anchored by the Siberian Yakut (cluster S5A-1), and a southern region (cluster S5A-2). Further breakdown of the clusters in the next tier of analysis (Figure S5B) resulted in the near total isolation of the Yakut population in the north (cluster S5B-1) and divided the southern population into four subgroups (Figure S5C). Among the resulting clusters, the Lahu group was seen in its own exclusive grouping (cluster S5-C3), while a Japanese cluster was observed with some mixed north-easterly samples (cluster S5-C4). A cluster with a western orientation was observed (cluster S5C-2), as well as a cluster directed towards the southern peninsula (cluster S5C-1). The next round of analysis further partitioned the previous clusters into regionally specific clusters (Figures S5D, S5E, S5F), but substantial overlap and mixed membership between the subclusters was observed. Several small subclusters were identified as having decreased levels of diversity (Table 1).

Regional classification analysis of mis-assigned unknown individuals

In order to classify and describe failures in re-assignment for pseudo-unknown individuals, each individual was traced from its original cluster definition to its re-assigned cluster. Figure S6 depicts mis-assigned individuals at each tier, in terms of their degree of mis-assignment as it relates to the average gene diversity of the two clusters involved. The degree of mis-assignment was quantified by the number of tiers that must be regressed in order for the two clusters involved to converge. Initially, it was observed that many mis-assignments took place when an individual from a smaller cluster with more restricted genetic membership was re-assigned to a sibling cluster that had broader, more inclusive genetic statistics. The units on the x-axis quantify the proportional difference of the average genetic diversity index between the originally assigned and mis-assigned clusters. Points observed to the far right of the y-axis (x = 0) were individuals that were re-assigned to clusters that had more genetically diverse members than in its initially defined cluster, while points to the left of the y-axis (x = 0) were re-assigned to clusters with lower than average gene diversity.

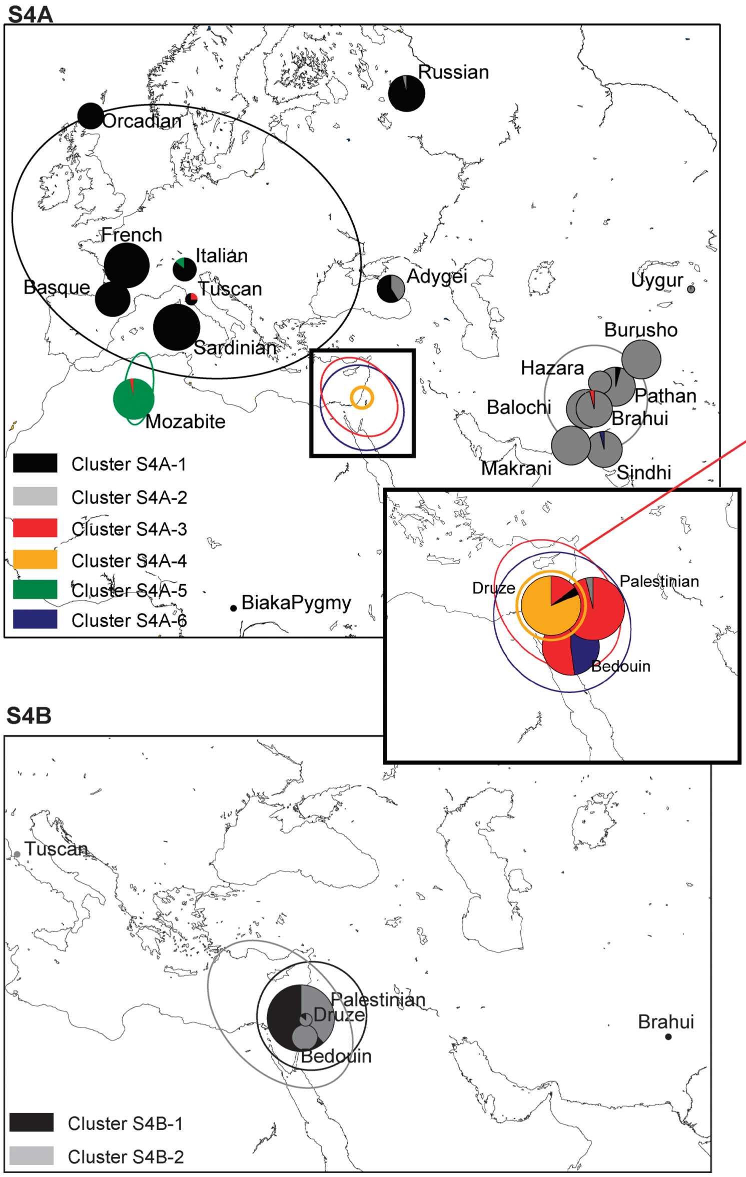


Figure S4. Geographical coverage of hierarchical genetic subgroups detected in Europe, the Middle East and Pakistan. (a) The initial partition defined by *structure* assembled this strongly diverse cluster. Subsequent rounds of decomposition produced subclusters with samples coming from common, geographically sensible origins. (b) Although overlap is seen across Middle Eastern populations, subclustering produced severally lineage-exclusive clusters: Bedouin (S4A-6), Palestinian (S4B-1) and Druze (S4A-4). **(c)** Subclustering of European samples initially indicated a north–south division where Orcadian and Russian (S4C-1) and Sardinian and Basque (6C-2) populations were associated, with various levels of supplementation from other European populations. (d) These associations were later dissolved when Orcadian (6D-1) and Russian (6D-2) samples anchored their own clusters. (e) Basques (S4E-1) and Sardinians (S4E-2) also divided to form their own anchored subclusters. (f) Densely sampled Pakistan produced two clusters where the Brahui (S4F-2), Burusho and Hazara (S4F-3) were uniformly classified. Clinal membership in the $K = 3$ clusters was observed in the remainder of the sampled populations. (g–i) Subdivision of these clusters resulted in mixed membership for most Pakistani populations, but the Brahui (S4G-1), Hazara (S4I-1) and Burusho (S4I-2) essentially maintained their homogeneity in cluster affiliation.

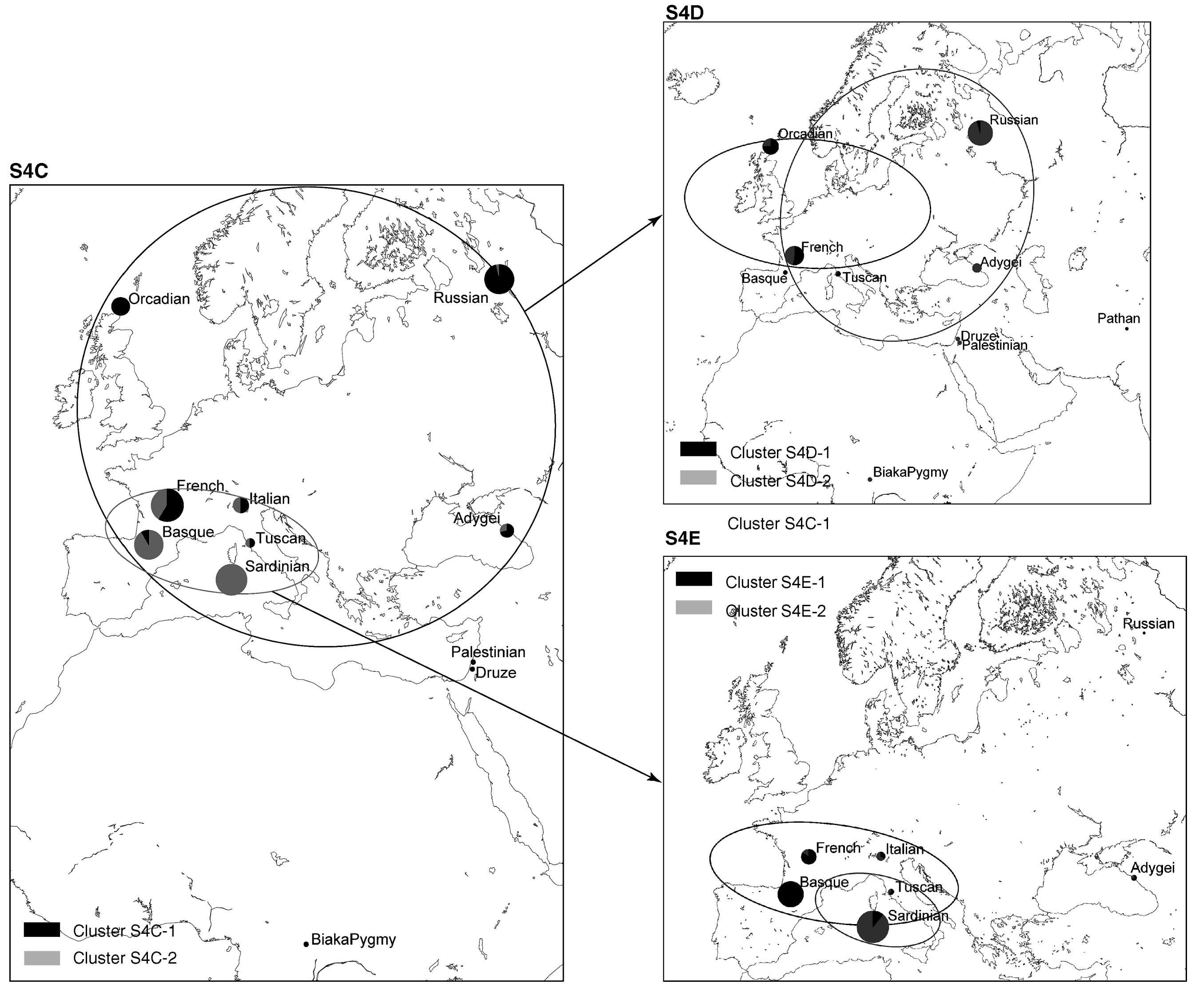


Figure S4. Continued.

World regions displayed distinct re-assignment characteristics over the course of the hierarchy. Clustering of symbols representing different world populations were observed at all levels of resolution. Individuals originating in Pakistan covered the total number of mis-assignments in level two. In subsequent tiers, Pakistanis were mis-assigned to clusters that converged distantly in the hierarchy with the originally assigned clusters. In all tiers, Pakistanis were among the most numerous mis-assigned individuals and were re-assigned to clusters that had lower *H* measures than their original cluster. Mis-assigned individuals were similarly numerous among Middle Easterners. These individuals were mis-assigned to clusters that had both lower and higher *H* indices than the initial clusters. East Asian individuals were also mis-assigned with similar frequency; however, East Asians were more frequently re-assigned to sibling or closer-degree clusters having higher *H* statistics. European individuals were mis-assigned with lower frequency until level 5, where discrepancies were more numerous. Europeans were generally seen to move to sibling or second-degree clusters with similar gene diversity measures. All other regional groups experienced high success rates in the re-assignment process. When mis-assignments were detected, they were most often characterised by an individual being re-assigned to a cluster with a low degree of disparity from the original, which had much wider genetic diversity than its initially designated cluster. African, American, Oceanic and the Pakistani Kalash individuals re-assigned relatively well throughout the hierarchy. Following post-clustering

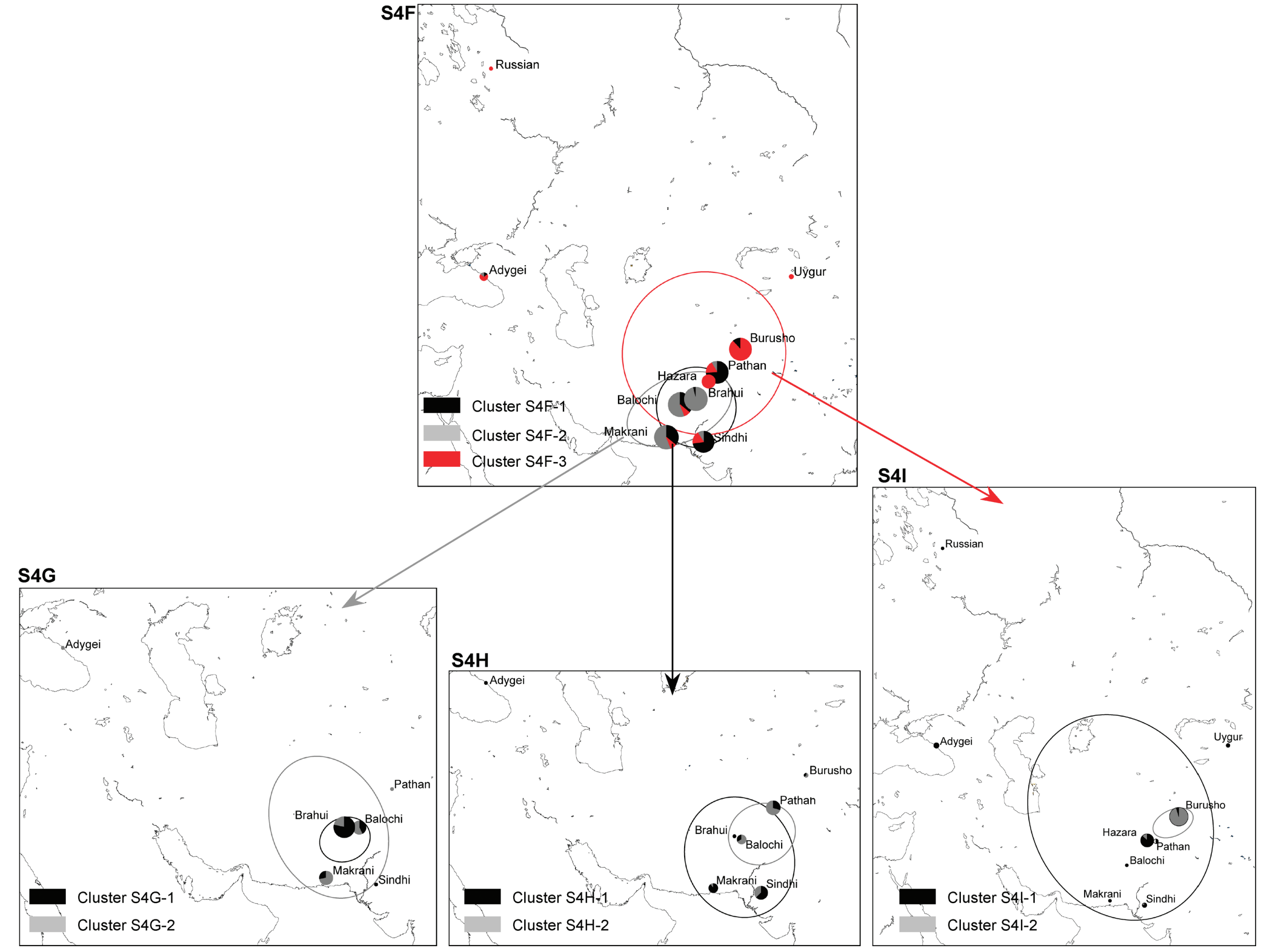


Figure S4. Continued.

adjustments, individuals that were re-assigned and became more stable in their cluster assignments were included in previously high-degree mis-assigned classifications, as seen in Figure S6.

Electronic-database information

Arlequin homepage, http://lgb.unige.ch/arlequin/ (for free download of genetic analysis software Arlequin, used to estimate F_{ST} and gene diversity statistics), CEPH Human Diversity Panel resource, available at http://research.marshfieldclinic.org/genetics/Freq/FreqInfo.htm (for genealogical/geographical origins of subjects in CEPH Human Diversity Panel).

GeneClass2 homepage, available at http://www.montpellier.inra.fr/URLB/GeneClass2/Help/ (for free download of genetic analysis software GeneClass2 used in assignment of unknowns to structured reference clusters) accessed 26th February, 2005. Noah Rosenberg's website, http://www.cmb.usc.edu/people/noahr/diversity.html (for download of CEPH human diversity panel data explored in this study) accessed 20th July, 2004.

Pritchard Lab, available at http://pritch/bsd/uchicago.edu/ (for *structure*, the software used to detect population structure and infer population assignment for individuals) accessed 4th February, 2003.

Acknowledgments

We thank Noah Rosenberg for making the current dataset available and for helpful communications regarding his analysis. We also thank Dennis Drayna, Rick Kittles and two anonymous reviewers for critical review of earlier drafts which helped to improve the quality of this manuscript. Funding for this investigation was provided by the Sorenson Molecular Genealogy Foundation.

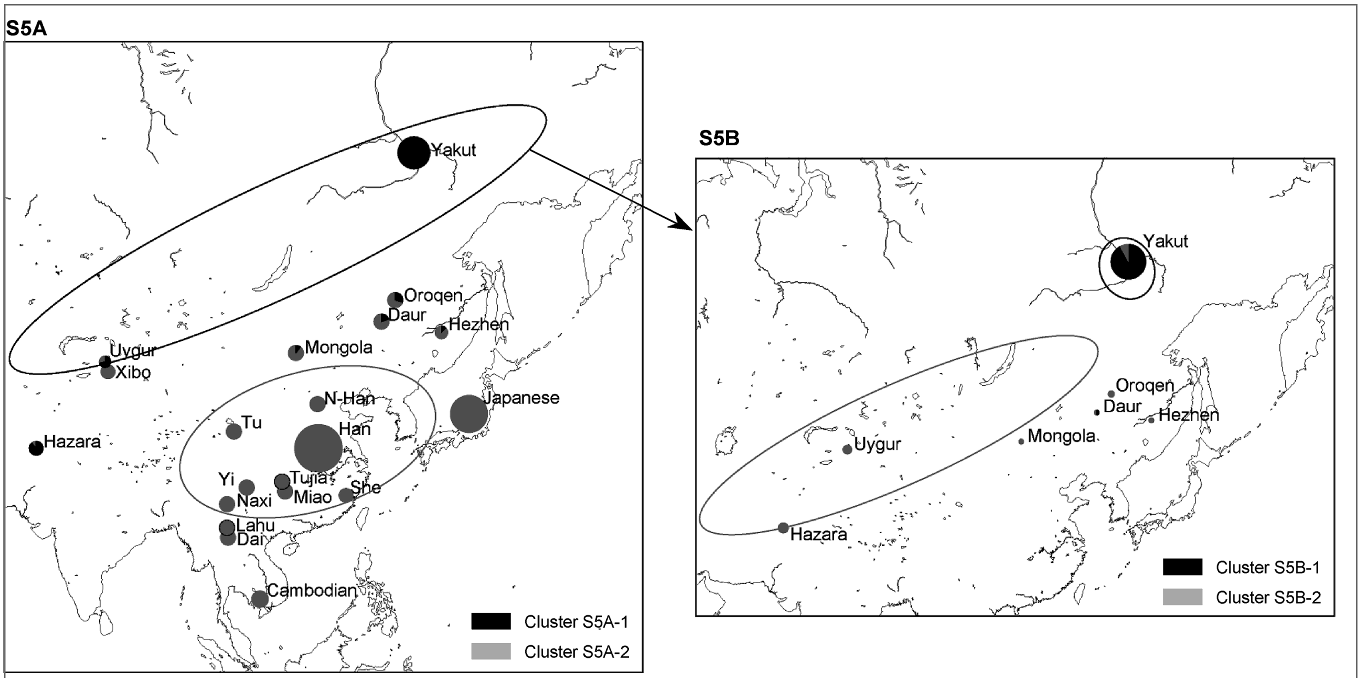


Figure S5. Geographical coverage of hierarchical genetic subgroups detected in East Asia. (a) The initial division at $K = 2$ introduced a north–south partition. (b) Subsequent analysis allowed the Yakut to be nearly completely isolated. (c) Division in the southern cluster (S5A-2) produced four distinct clusters, one with a southern orientation (S5C-1), a westerly group (S5C-2), a completely isolated Lahu cluster (S5C-3) and a strongly anchored Japanese cluster (S5C-4). (d–f) Further analysis allowed the separation of a distinctly Japanese group (S5F-2) and mixed membership clusters in the remainder of the groupings.

References

1. Rosenberg, N.A., Li, L.M., Ward, R. and Pritchard, J.K. (2003), ‘Informativeness of genetic markers for inference of ancestry’, *Am. J. Hum. Genet.* Vol. 73, pp. 1402–1422.
2. 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.
3. Estoup, A., Garnery, L., Solignac, M. and Cornuet, J.M. (1995), ‘Microsatellite variation in honey bee (*Apis mellifera* L.) populations: Hierarchical genetic structure and test of infinite allele and stepwise mutation models’, *Genetics* Vol. 140, pp. 679–695.
4. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000), ‘Inference of population structure using multilocus genotype data’, *Genetics* Vol. 155, pp. 945–959.
5. Rosenberg, N.A., Burke, T., Elo, K. *et al.* (2001), ‘Empirical evaluation of genetic clustering methods using genotypes from 20 chicken breeds’, *Genetics* Vol. 159, pp. 699–713.
6. Rosenberg, N.A., Pritchard, J.K., Weber, J.L. *et al.* (2002), ‘Genetic structure of human populations’, *Science* Vol. 298, pp. 2381–2385.
7. Parker, H.G., Kim, L.V., Sutter, N.B. *et al.* (2004), ‘Genetic structure of purebred domestic dog’, *Science* Vol. 304, pp. 1160–1164.
8. Excoffier, L. and Hamilton, G. (2003), ‘Comment on “Genetic Structure of Human Populations”’, *Science* Vol. 300, p. 1877B.
9. Cann, H.M., de Toma, C., Cazes, L. *et al.* (2002), ‘A human genome diversity panel cell line panel’, *Science* Vol. 296(5566), pp. 261–262.
10. Zhivotovsky, L.A., Rosenberg, N.A. and Feldman, M.W. (2003), ‘Features of evolution and expansions of modern humans, inferred from genomewide microsatellite markers’, *Am. J. Hum. Genet.* Vol. 72, pp. 1181–1186.
11. Corander, J., Waldmann, P., Marttinen, P. *et al.* (2004), ‘BAPS 2: Enhanced possibilities for the analysis of genetic population structure’, *Bioinformatics* Vol. 20, pp. 2363–2369.
12. Fu, R., Dey, K.D. and Holsinger, K.E. (2005), ‘Bayesian models for the analysis of genetic structure when populations are correlated’, *Bioinformatics* Vol. 21, pp. 1516–1529.
13. Yang, B.Z., Hongyu, Z., Kranzler, H.R. *et al.* (2005), ‘Practical population group assignment with selected informative markers: Characteristics and properties of Bayesian clustering via STRUCTURE’, *Genet. Epidemiol.* Vol. 28, pp. 302–312.
14. Mountain, J.L. and Ramakrishnan, U. (2005), ‘Impact of human population history on distributions of individual-level genetic distance’, *Hum. Genom.* Vol. 2, pp. 4–19.
15. Falush, D., Stephens, M. and Pritchard, J.K. (2003), ‘Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies’, *Genetics* Vol. 164, pp. 1567–1587.
16. Pritchard, J.K. and Wen, W. (2004), Documentation for structure software: Version 2, available at http://pritch.bsd.uchicago.edu/software/readme_2_1/readme.html, University of Chicago). Accessed 4th February, 2003.
17. Corander, J., Waldmann, P. and Sillanpaa, M.J. (2003), ‘Bayesian analysis of genetic differentiation between populations’, *Genetics* Vol. 163, pp. 367–374.
18. Dawson, K.J. and Belkhir, K. (2001), ‘A Bayesian approach to the identification of panmictic populations and the assignment of individuals’, *Genet. Res. Camb.* Vol. 78, pp. 59–77.
19. Schneider, S., Roessli, D. and Excoffier, L. (2000), Arlequin: A software for population genetics data analysis. Ver 2.000, Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Switzerland.
20. Piry, S., Alapetite, A., Cornuet, J.-M. *et al.* (2004), ‘GeneClass2: A software for genetic assignment and first-generation migrant detection’, *J. Hered.* Vol. 95, pp. 536–539.

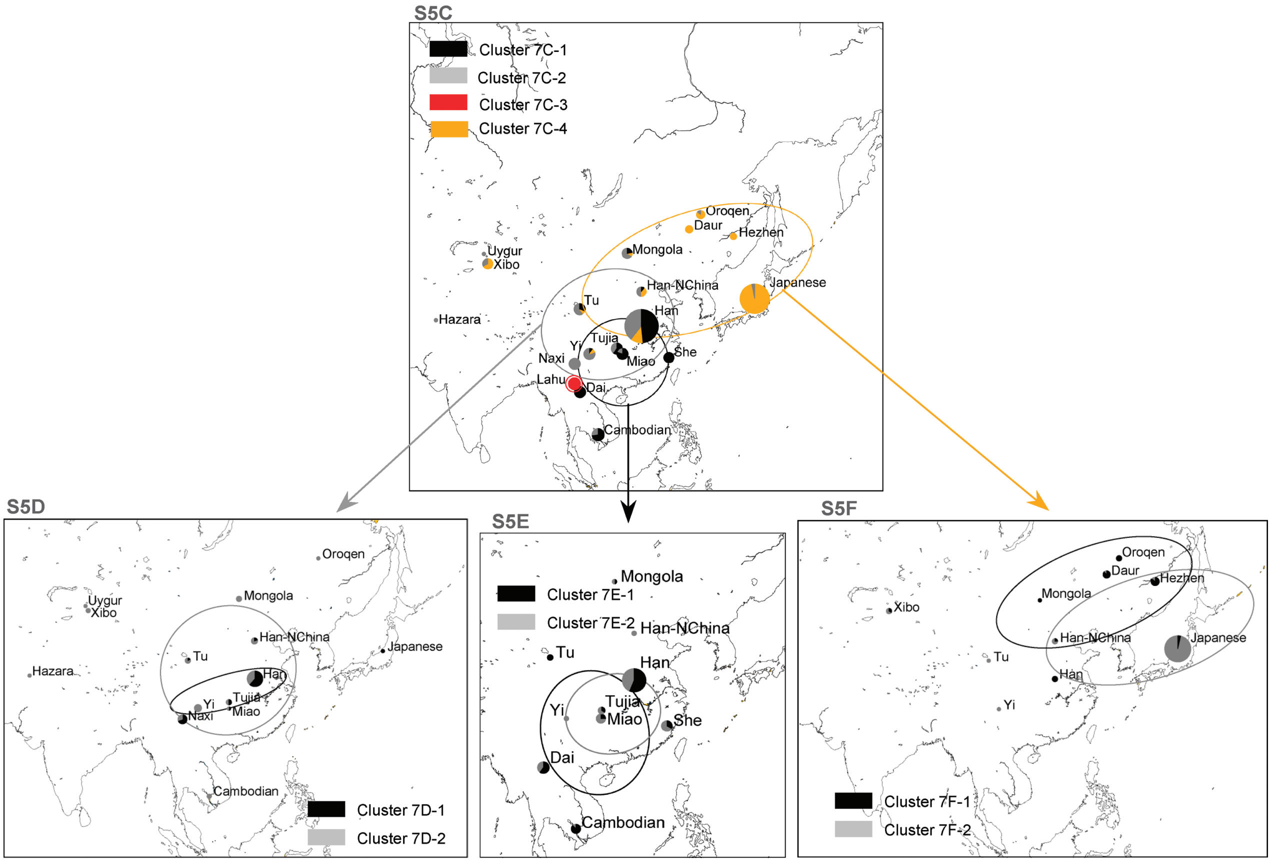


Figure S5. Continued.

21. Rannala, B. and Mountain, J.L. (1997), 'Detecting immigration by using multilocus genotypes', *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 9197–9201.
22. Baudouin, L., Piry, S. and Cornuet, J.M. (2004), 'Analytical Bayesian approach for assigning individuals to populations', *J. Hered.* Vol. 95, pp. 217–224.
23. Bamshad, M., Wooding, S., Salisbury, B.A. and Stephens, J.C. (2004), 'Deconstructing the relationship between genetics and race', *Nat. Rev. Genet.* Vol. 5, pp. 598–609.
24. Frudakis, T., Venkateswarlu, K., Thomas, M.J. *et al.* (2003), 'A classifier for the SNP-based inference of ancestry', *J. For. Sci.* Vol. 48, pp. 771–782.
25. Rosenberg, N.A., Woolf, E., Pritchard, J.K. *et al.* (2001), 'Distinctive genetic signatures in Libyan Jews', *Proc. Natl. Acad. Sci. USA* Vol. 98, pp. 858–863.
26. Wilson, J.F., Weale, M.E., Smith, A.C. *et al.* (2001), 'Population genetic structure of variable drug response', *Nat. Genet.* Vol. 29, pp. 265–269.
27. Hoggart, C.J., Parra, E.J., Shriver, M.D. *et al.* (2003), 'Control of confounding of genetic associations in stratified populations', *Am. J. Hum. Genet.* Vol. 72, pp. 1492–1504.
28. Hao, K., Wang, X., Niu, T. *et al.* (2004), 'A candidate gene association study on preterm delivery: application of high-throughput genotyping technology and advanced statistical methods', *Hum. Mol. Genet.* Vol. 13, pp. 683–691.
29. Hinds, D.A., Stokowski, R.P., Patil, N. *et al.* (2004), 'Matching strategies from genetic association studies in structured populations', *Am. J. Hum. Genet.* Vol. 74, pp. 318–325.
30. Keuhn, R., Haller, H., Schroeder, W. and Rottmann, O. (2004), 'Genetic roots of the red deer (*Cervus elaphus*) population in eastern Switzerland', *J. Hered.* Vol. 95, pp. 136–143.
31. Larson, S.R., Jones, T.A. and Jensen, K.B. (2004), 'Population structure in *Pseudoroegneria spicata* (Poaceae: Triticeae) modeled by Bayesian clustering of AFLP genotypes', *Am. J. Bot.* Vol. 91, pp. 1789–1801.
32. Muir, G., Lowe, A.J., Fleming, C.C. and Vogl, C. (2004), 'High nuclear genetic diversity, high levels of outcrossing and low differentiation among remnant populations of *Quercus petraea* at the margin of its range in Ireland', *Ann. Bot.* Vol. 93, pp. 691–697.
33. Stinchcombe, J.R., Weinig, C., Ungerer, M. *et al.* (2004), 'A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA', *Proc. Natl. Acad. Sci. USA* Vol. 101, pp. 4712–4718.
34. Wirth, T., Wang, X., Linz, B. *et al.* (2004), 'Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: Lessons from Ladakh', *Proc. Natl. Acad. Sci. USA* Vol. 101, pp. 4746–4751.
35. Bamshad, M.J., Wooding, S., Watkins, W.S. *et al.* (2003), 'Human population genetic structure and inference of group membership', *Am. J. Hum. Genet.* Vol. 72, pp. 578–589.
36. Stephens, M. (2000), 'Dealing with label-switching in mixture models', *J.R. Stat. Soc. Ser. B* Vol. 62, pp. 795–890.

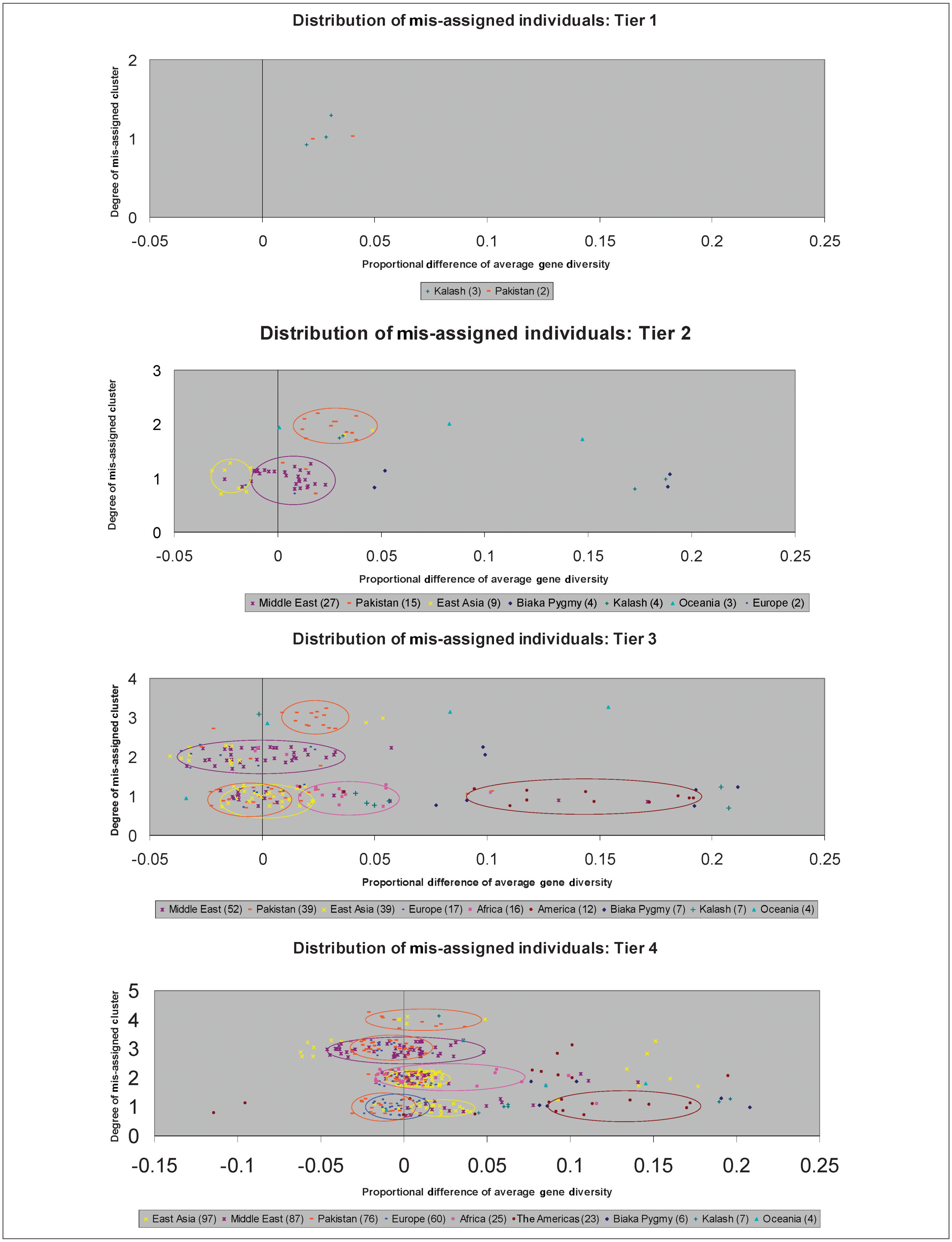

Distribution of mis-assigned individuals: Tier 1
Degree of mis-assigned cluster
0
1
2
-0.05
0
0.05
0.1
0.15
0.2
0.25
Proportional difference of average gene diversity
Kalash (3)
Pakistan (2)
Distribution of mis-assigned individuals: Tier 2
Degree of mis-assigned cluster
0
1
2
3
-0.05
0
0.05
0.1
0.15
0.2
0.25
Proportional difference of average gene diversity
Middle East (27)
Pakistan (15)
East Asia (9)
Biaka Pygmy (4)
Kalash (4)
Oceania (3)
Europe (2)
Distribution of mis-assigned individuals: Tier 3
Degree of mis-assigned cluster
0
1
2
3
4
-0.05
0
0.05
0.1
0.15
0.2
0.25
Proportional difference of average gene diversity
Middle East (52)
Pakistan (39)
East Asia (39)
Europe (17)
Africa (16)
America (12)
Biaka Pygmy (7)
Kalash (7)
Oceania (4)
Distribution of mis-assigned individuals: Tier 4
Degree of mis-assigned cluster
0
1
2
3
4
5
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
Proportional difference of average gene diversity
East Asia (97)
Middle East (87)
Pakistan (76)
Europe (60)
Africa (25)
The Americas (23)
Biaka Pygmy (6)
Kalash (7)
Oceania (4)

Figure S6. Distribution of mis-assigned unknown individuals. Individuals are categorised by degree of mis-assignment (y-axis), and on the continuous variable x-axis quantified by proportional difference in average gene diversity between mis-assigned and original clusters. Degree of mis-assignment was quantified by the number of tiers required to be regressed in order for the two clusters involved to converge.

Appendix

Selection method and criteria

Because the structure of the dataset was unknown, runs were performed on the total data ranging from $K = 1$ to $K = 10$, with $n = 10$ independent runs for each value of K, all with identical parameters. The goal with the initial partition was to identify a value of K that captured the major structure of the data with the intent that finer population structure could be extracted in subsequent rounds of analysis. The strategy for selecting K was two-fold, representing a balance between maximising the posterior probability of the solution and taking into account the similarity of the solutions produced for independent runs with identical input and parameters. As recommended and applied previously,[4,5,15,25–35] a most likely solution can be selected from the series of separate runs resulting from incremented K values. The run in which the posterior probability of the data, $Pr(K)$, is maximised represents the best K for the data.

Due to the complexity of the relationships among individuals in the dataset, *structure* regularly ascertained different solutions arising from identical input and parameters. This observation coincided with previous explanations of the components of the *structure* algorithm, in which it may settle into different modes at the end of the burn-in period and result in different clustering solutions.[4,27,29] Such bad mixing or label switching has been handled or avoided in other contexts.[12,13,36] One approach utilised multiple parallel Markov chain Monte Carlo to identify the set of important candidate partitions and simultaneously estimate K, cluster membership and allele frequencies for resulting clusters. Using the current model, however, similar solutions frequently had similar likelihood values.[5] These similar solutions as an aggregate were defined as a 'major mode' if:

1. The same solution was observed multiple times within the solution space with a threshold of ≥ 0.97 similarity.
2. The independent solutions that constituted the major mode had similar likelihood scores, $Pr(K)$, and the range covered by similar $Pr(K)$ values for the mode of interest did not overlap the range associated with a different mode. This criterion assured the independence and repeatability of the highest likelihood solution, in relation to other observed solutions.

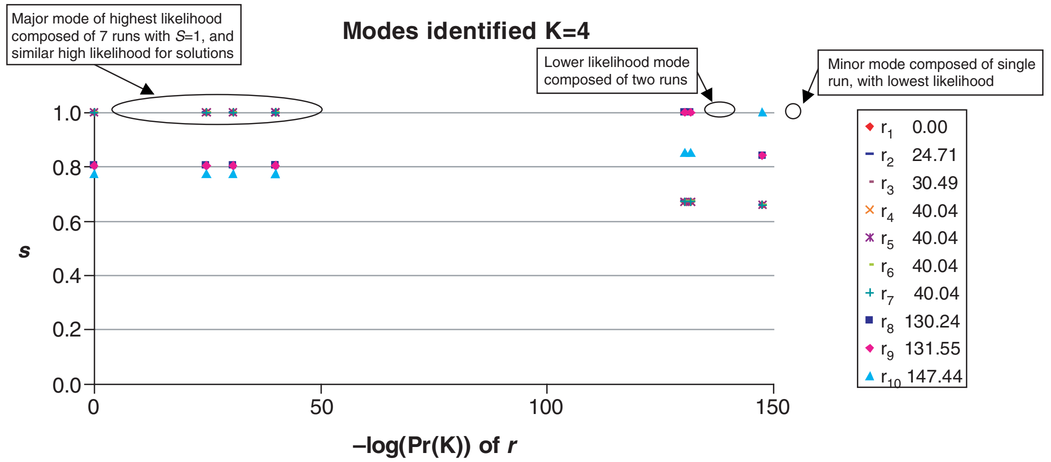
Pairwise similarity of solutions, S, was quantified based on the extent of agreement of sample assignment to clusters among runs. Although many modes were observed in the solution space, a major mode was observed only if it met the criteria specified above. All other solutions were categorised as belonging to a minor mode. Figure A1 illustrates multiple modes found in the solution space of $n = 10$ runs for the same dataset with identical run parameters. The major mode having the highest likelihood $Pr(K)$ was selected as the best solution, as it represented a statistically likely solution and one that was empirically observed to be repeatable throughout the solution space.

This concluded the selection analysis for one set of $n = 10$ runs for a specific K value. The process was repeated for $n = 10$ runs from all of the singly incremented K values. The major mode of the highest likelihood was detected from each set of runs and the $Pr(K|X)$ values were collected. Using the values that represented the best clustering solution at each singly incremented K, the best K was selected from among them using Bayes' rule to find the highest $Pr(K)$ corresponding to the best K with which to analyse the data.

Initially, the process was performed on the total dataset. When the best solution was selected, the dataset was partitioned into K clusters based on the maximum membership coefficient for each sample. The same selection method was applied to each subcluster until the maximum decomposition was reached. This was signalled when analysis of a subcluster identified $K = 1$ as the best solution. At this point, analysis for that particular line was terminated. Figure A2 depicts a flowchart representing the procedure for selecting the best K for a dataset applied hierarchically to detect fine levels of population structure.

Similarity coefficient s

A measure for comparing the differing clustering solutions of independent *structure* runs was previously proposed by Rosenberg *et al.*[6] This algorithm provides a thorough method for determining the extent to which two independently attained solutions agree with one another, and allows for testing between sets where one of the pair is a subset of the full dataset. Because the computations take into account all possible permutations of the membership coefficient matrix over K clusters, the run time has a factorial increase with increasing K. This makes the algorithm intractable for any K greater than seven or eight. In the course of this study, as well as in previous work, values for K ranging from $K = 12$,



Major mode of highest likelihood composed of 7 runs with *S*=1, and similar high likelihood for solutions

Lower likelihood mode composed of two runs

Minor mode composed of single run, with lowest likelihood

Figure A1. Example of modes detected for a subset of the total dataset, $n = 10$ runs at $K = 4$. Three modes were explored by the *structure* algorithm in this dataset. The $Pr(K)$ values for runs r_l to r_n and an $n \times n$ matrix of similarity scores S for the set of runs is calculated. A symbol (specified in the key) is assigned to each run r and plotted on the x-axis at the $-\log_{10}(Pr[K])$ value for runs r_l to r_n, versus S representing the level of similarity calculated between r and r_l to r_n. This method expands upon previous systems for selecting a repeatable K in the sample space. Rosenberg *et al.*[6] also generated an $n \times n$ matrix of similarity values and took the mean of all values in the matrix to assess levels of similarity for runs generated at various K. This method was found to be restrictive, in that certain K values were rejected due to a single disparate run that skewed the overall similarity below the selected similarity threshold (0.97) for acceptance of the K. For the $n \times n$ matrix generated for this dataset, seven of the ten runs produced a similarity score of $S = 1$ relative to one another. The remaining three solutions were ≤ 80 per cent similar to the other seven, and had a much lower likelihood. Screening K based on the mean of all S scores (0.868 versus the minimum defined S threshold 0.97) would disqualify this K from consideration as a viable representation of the structure of the dataset. Because a major mode of high probability and high similarity can be identified within the dataset, this illustrates the selection of a statistically likely and empirically stable substructure of the dataset.

$K = 19$,[5] to $K = 20$ were proposed.[7] Thus, an alternate method was proposed which readily handles large K values and is suitable for exploring certain hypotheses.

It was desirable to produce a monotonic value that can be used to compare sets of independent runs at a given K. In order to characterise this process, it was necessary to define how individuals are classified into clusters from the data obtained from *structure* in the $N \times K$ Q-matrix. Previous investigators have used minimum membership coefficient thresholds to assign individuals to particular clusters within a run, with values ranging from ≥ 0.25 to ≥ 0.75.[5,15] As previously applied, for some analyses it is most useful to dispense with lesser admixture estimates in ancillary clusters and bin an individual completely to a single cluster, using the maximum membership coefficient, as the *best* forced fit for that individual.

Once the cluster for an individual i was defined by its maximum membership coefficient value, it was associated with the other individuals from that run who were also placed into the same cluster, termed set G_1. In a subsequent independent run of *structure*, the maximum membership coefficient of individual i again determined its cluster assignment and the individuals with which it grouped in this second run were defined as the set G_2. The proportion of individuals shared between sets G_1 and G_2 was determined and termed G_m.

$$G_m = G_1 \cap G_2 \tag{1}$$

The count of individuals of this set N_{Gm} quantified the number of individuals from G_1 that individual i also clustered with in G_2. A similarity score for individual i, S_i, measured the proportion of individuals in common that it clustered with across the independent runs, giving a measure of how well individual i adhered to other individuals from its original cluster G_1. N_{G_1} is the count of individuals in set G_1.

$$S_i = \frac{N_{Gm}}{N_{G_1}} \tag{2}$$

A measure of the total similarity, S_t, of the cluster assignments for all individuals across two independent

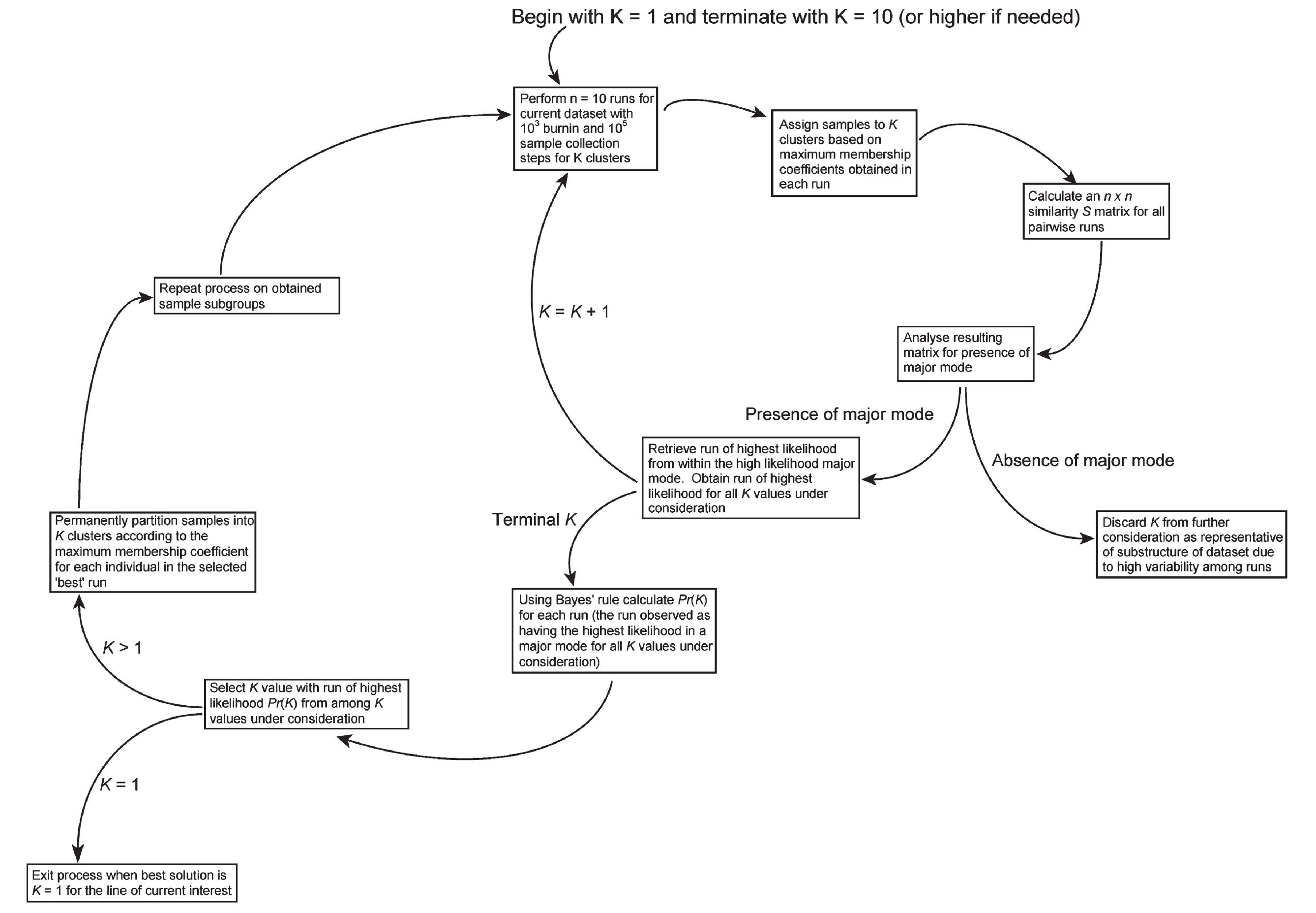

Begin with K = 1 and terminate with K = 10 (or higher if needed)
Perform n = 10 runs for current dataset with 10^3 burnin and 10^5 sample collection steps for K clusters
Assign samples to K clusters based on maximum membership coefficients obtained in each run
Calculate an n x n similarity S matrix for all pairwise runs
Analyse resulting matrix for presence of major mode
Presence of major mode
Absence of major mode
Discard K from further consideration as representative of substructure of dataset due to high variability among runs
Retrieve run of highest likelihood from within the high likelihood major mode. Obtain run of highest likelihood for all K values under consideration
K = K + 1
Terminal K
Using Bayes' rule calculate Pr(K) for each run (the run observed as having the highest likelihood in a major mode for all K values under consideration)
Select K value with run of highest likelihood Pr(K) from among K values under consideration
K > 1
Permanently partition samples into K clusters according to the maximum membership coefficient for each individual in the selected 'best' run
Repeat process on obtained sample subgroups
K = 1
Exit process when best solution is K = 1 for the line of current interest

Figure A2. Flowchart for multi-level determination of *K* and detection of hierarchical fine-level substructure for a dataset. The total dataset was initially decomposed using *structure* into *K* clusters according to the best run, as determined by the selection criteria. This produced *K* subgroups that were subsequently subjected to the decomposition process, and so on recursively until the finest substructure for each subgroup line was detected, as signalled by $K = 1$ for the best solution.

applications of *structure* was estimated from the mean of all values of S_i for the total number of individuals, N_t, in the full dataset.

$$S_t = 1/N_t \sum_{i=1}^{N_t} S_i \tag{3}$$

For comparison, independent runs must meet two criteria for this method to produce meaningful estimates of the similarity of the solutions. The two solutions must comprise the same individuals and the number of clusters, *K*, must be consistent for the two solutions being compared. Additionally, this calculation is directional, resulting in a different value for S_t when N_{G_2} is used as the normalising factor in equation (2). The results are asymmetrical but are generally comparable values. This summary statistic can be used in *post hoc* analysis of *structure* runs to determine the best *K* with which to fit the input data.