Hum Genet (2012) 131:601–613
DOI 10.1007/s00439-011-1104-8

ORIGINAL INVESTIGATION

Recombination networks as genetic markers in a human variation study of the Old World

Asif Javed · Marta Melé · Marc Pybus · Pierre Zalloua · Marc Haber · David Comas · Mihai G. Netea · Oleg Balanovsky · Elena Balanovska · Li Jin · Yajun Yang · GaneshPrasad ArunKumar · Ramasamy Pitchappan · Jaume Bertranpetit · Francesc Calafell · Laxmi Parida · The Genographic Consortium

Received: 6 July 2011 / Accepted: 3 October 2011 / Published online: 18 October 2011

Abstract We have analyzed human genetic diversity in 33 Old World populations including 23 populations obtained through Genographic Project studies. A set of 1,536 SNPs in five X chromosome regions were genotyped in 1,288 individuals (mostly males). We use a novel analysis employing subARG network construction with recombining chromosomal segments. Here, a subARG is constructed independently for each of five gene-free regions across the X chromosome, and the results are aggregated across them. For PCA, MDS and ancestry inference with STRUCTURE, the subARG is processed to obtain feature vectors of samples and pairwise distances between samples. The observed population structure, estimated from the five short X chromosomal segments, supports genome-wide frequency-based analyses: African populations show higher genetic diversity, and the general trend of shared variation is seen across the globe from Africa through Middle East, Europe, Central Asia, Southeast Asia, and East Asia in broad patterns. The recombinational analysis was also compared with established methods based on SNPs and haplotypes. For haplotypes, we also employed a fixed-length approach based on information-content optimization. Our recombinational analysis suggested a southern migration route out of Africa, and it also supports a single, rapid human expansion from Africa to East Asia through South Asia.

A. Javed and M. Melé are joint first authors.

Members of the Genographic Consortium are provided in the "Appendix".

Electronic supplementary material The online version of this article (doi:10.1007/s00439-011-1104-8) contains supplementary material, which is available to authorized users.

A. Javed · L. Parida (✉)
Computational Biology Center, IBM T J Watson Research,
Yorktown, USA
e-mail: parida@us.ibm.com

M. Melé · M. Pybus · D. Comas · J. Bertranpetit ·
F. Calafell (✉)
IBE, Institut de Biologia Evolutiva (CSIC-UPF), Barcelona,
Catalonia, Spain
e-mail: francesc.calafell@upf.edu

P. Zalloua · M. Haber
School of Medicine, Lebanese American University,
Beirut, Lebanon

M. G. Netea
Department of Medicine and Nijmegen Institute for Infection,
Inflammation, and Immunity, Radboud University Nijmegen
Medical Center, Nijmegen, The Netherlands

O. Balanovsky · E. Balanovska
Research Centre for Medical Genetics, Moscow, Russia

O. Balanovsky
Vavilov Institute for General Genetics, Moscow, Russia

L. Jin · Y. Yang
MOE Key Laboratory of Contemporary Anthropology, School
of Life Sciences and Institutes of Biomedical Sciences,
Fudan University, Shanghai 200433, China

G. ArunKumar · R. Pitchappan
School of Biological Sciences, Madurai Kamaraj University,
Madurai 625021, India

R. Pitchappan
Chettinad Academy of Research and Education,
Chettinad Health City, Rajiv Gandhi Salai, Kelambakkam,
Chennai, India

Introduction

The reconstruction of human population history is an ongoing endeavor that has been made possible, thanks to the availability of a wealth of data on human genetic diversity. Such data are generated in two main forms: complete sequences and allele frequencies of previously ascertained polymorphisms. Contiguous SNPs contain a finer degree of information than that represented by independent markers. The joint history of haplotypes, including recombination events that are part of their shared history, leave footprints in neighboring markers. On average, the recombination rate among contiguous nucleotides is of the same order of magnitude as that of the mutation rate per nucleotide (Kong et al. 2010; Durbin et al. 2010). Thus, recombination has likely played a major role in shaping the current haplotype phylogeny.

Methods that can detect recombination events and incorporate them into the reconstruction of haplotype phylogenies are needed. We provided a first step in this direction in our previous paper, in which we used a model-based algorithm (IRiS) to detect recombination events in extant haplotypes (Melé et al. 2010). The IRiS parameters were previously validated in a third-party simulation environment [cosi; (Schaffner et al. 2005)]. The use of a consensus across multiple parameter sets gave individual recombination events, along with its extant descendants, with high confidence.

We now extend this protocol to additionally extract the donor haplotypes contributing to each individual recombination in an effort to reconstruct the ancestral recombination graph (ARG) from this information. The ARG is necessarily not fully resolved (hence we call it a subARG) because not all genetic events can be reconstructed. In this context, we provided a mathematical description of reconstructability in Parida et al. (2011) which showed evidence of vanishing reconstructability of the ARG with depth. Therefore, we used only the top few recombinations of high confidence, and only reconstructed the high probability nodes in the ARG. This conservative approach to subARG construction has sufficient fidelity to the true ARG under simulation conditions.

Furthermore, since the ARG is being constructed around recombinations, its resolution relies on the frequency of recombination events in the underlying region. In this regard, traces of older recombinations are continuously being overwritten by newer ones in each successive generation. This is even truer near hotspots where only the most recent events can be defined with any certainty. Conversely, rare recombinations, in cold spots, may bring together ancient haplotypes that reveal the deeper past (Fisher 1954; Baird 2006).

Haplotypes are the product of a number of evolutionary forces, including mutation, selection, drift and recombination. Their use as the tool of choice in human population genetics was pioneered by Ken Kidd (Tishkoff et al. 1996). Since then, multiple studies have taken advantage of their increased sensitivity to population history (Tishkoff et al. 1998; Mateu et al. 2001, 2002; Conrad et al. 2006), which was specifically compared to that of SNPs (Jakobsson et al. 2008). In that study, it was found that at the high density considered, unphased SNPs provided considerable population structure information, although haplotype data can contribute an additional informative component for population structure analysis.

Commercial genomewide SNP chips are comprised of mainly tagSNPs. Each tagSNP represents its neighboring correlated markers by proxy, and the correlation between genotyped markers is low. IRiS detects recombination events by exploiting SNP patterns that define a haplotype. Thus, the detection of these events improves with higher correlation among neighboring markers (Melé et al. 2010). This fact limits the applicability of off-the-shelf genomewide assays to detect specific recombination events.

To achieve dense physical coverage, we used a customized assay consisting of 1,536 SNPs focusing on five short regions on the X chromosome; all together the regions comprise about 2 Mb of sequence. The markers were chosen to achieve high SNP density, independent of the underlying linkage disequilibrium structure, for an unbiased assessment of recombinations. The choice of the X chromosome has multiple benefits. First, in an ideal population containing equal number of males and females, the number of X chromosomes is 3/4 that of any autosome. Thus, the effective population size of X chromosome is lower, which, in turn, means that genetic drift has a higher impact on it (Hammer et al. 2010). Second, the genomic regions being analyzed in this study recombine only in females since they lie outside the pseudoautosomal region. The frequency of recombination events in the phylogeny of these regions is about a third less per generation than an autosomal counterpart with comparable recombination rate. Fewer recent events allow traces of older recombination to survive longer in the haplotypes, thus allowing a deeper look into the past. Third, males carry only one X chromosome. The haplotype of each male chromosome can therefore be directly determined from the genotype. Although our detection of recombinations has been shown to work well in high-quality phased autosomal data (Melé et al. 2010), by preferentially choosing male samples one potential source of error is removed.

In this study, 23 worldwide human populations were sampled by members of the Genographic Consortium and DNAs provided for this study (Fig. 1). Samples from an additional ten populations from the HapMap project (http://hapmap.ncbi.nlm.nih.gov) were also analyzed. In total, 1,318 chromosomes from 33 populations contributed

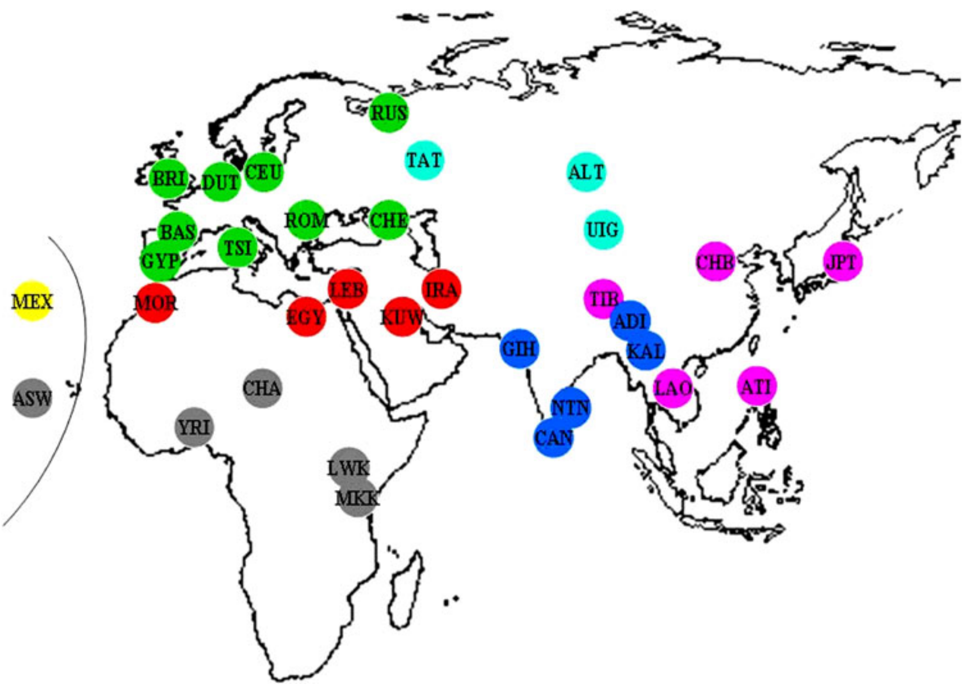


Fig. 1 Geographic distribution of the sampled populations in the Old World

to our study. With these samples, we show the advantages of analyzing in parallel the subARG, SNPs, and haplotypes, and add an additional dimension (recombination) to the usual approaches in human population genetics.

Subjects and methods

DNA samples and data preparation

DNA samples were collected based on geographic distribution from the Old World (Fig. 1, Supplemental Tables 1 and 2). Samples from 23 populations obtained by members of the Genographic Consortium were provided for recombination-based analysis. Ample information was available first-hand for the cultural, linguistic, and historical aspects that could help in interpret the results. Mitochondrial DNA sequences and haplotype and haplogroups in the Y chromosome have been studied for most of the samples. Informed consent was obtained from all study subjects, under approval from ethics committees at all institutions where investigators collecting samples work. Samples from an additional six populations included in HapMap phase 3 were obtained from the Coriell Cell Repository. In total, 1,288 samples were genotyped for 1,536 markers. The data for four HapMap phase 2 populations (release 21) were downloaded from the project website (http://hapmap.ncbi.nlm.nih.gov) and added to our analysis.

To reduce the impact of phasing errors, the study focused on the X chromosome, and male samples were preferred over those of females. Five regions were identified on the X chromosome with high SNP density that were at least 50 Kb away from known genes, copy number variations and segmental duplications (Supplemental Table 3). SNPs were selected based on HapMap phase 2 release 24, and genotyping was performed using the Illumina GoldenGate custom Oligos array for 1,536 SNPs. The average and median distance between SNPs were 1,623 and 804 bp, respectively.

After genotyping, SNPs with more than 15% of missing data, as well as those having a cluster of heterozygous positions in male samples (80 SNPs), were removed. Samples with more than 10% missing markers (123 samples), or male samples with more than 3 heterozygous positions, were also removed (14 samples). The remaining heterozygous markers in male samples were recoded as missing and imputed. Markers that were monomorphic across all populations were also removed (201 SNPs). The final data set resulted in 1,255 SNPs being genotyped in 1,318 samples (from 1,269 males and 49 females) belonging to 33 worldwide populations (Supplemental Table 1). None of the 22 internal replicas showed inconsistencies. Missing values were imputed using fastPHASE (Scheet and Stephens 2006) and the female samples were phased using PHASE 2.1 (Stephens and Scheet 2005), using the direct data of haplotypes given by males.

Identifying recombinations

Melé et al. (2010) developed a method to identify the breakpoint position and the extant descendants of the recombinant carrying the breakpoint junction. A consensus among sliding windows of varying sizes is used to accentuate high confidence recombinations. The method was validated by coalescent simulations under realistic demographic parameters using cosi (Schaffner et al. 2005).

In the current manuscript, we extend this methodology by exploiting the haplotypic pattern at the junction. Neighboring patterns, at each of the two sides of the inferred breakpoint, carry the *donor* haplotypes contributing to the recombination. Agreement across multiple runs identifies the current sequences which are the descendants of these donors. Thus, every recombination defines a trichotomy of contributing extant sequences: the direct descendants of the recombinant which carry both donor haplotypes, and the two sets of sequences that share ancestry on only one side (either upstream or downstream) of the breakpoint. The events creating donor haplotypes are ancestral to the recombination event itself; their descendants may not be the progeny of the immediate parent of the recombination (Fig. 2).

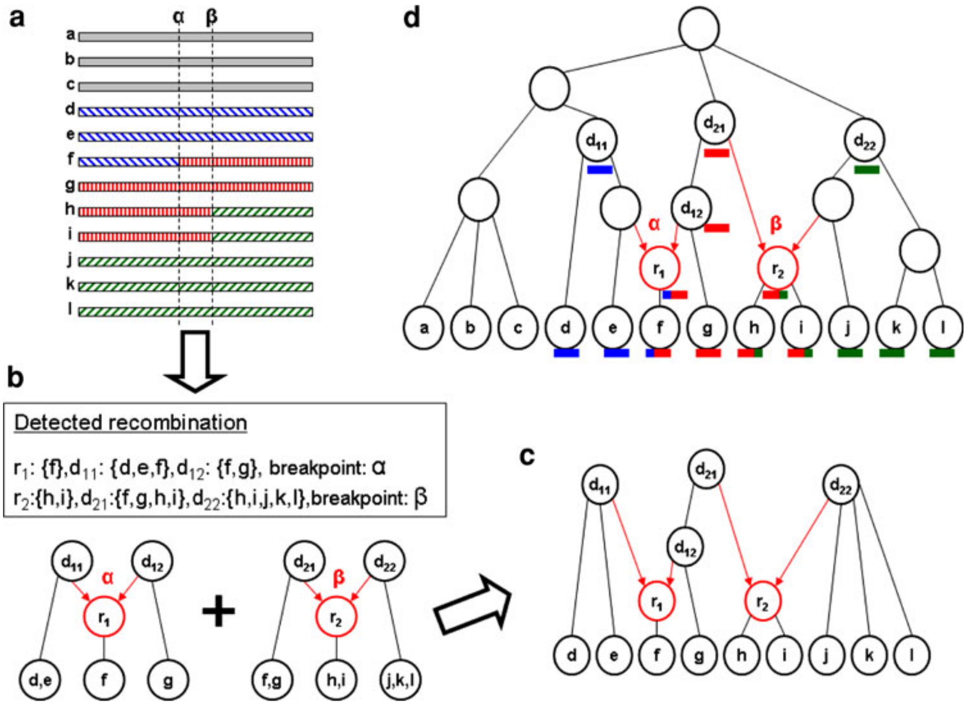
Constructing the subARG

Each detected recombination event, in combination with the donor haplotypes, defines a local network with three to five nodes in a subset of the samples carrying the neighboring segments around the breakpoint location. If *r* is the total number of recombinations detected, then, in the second step, we reconcile these *r* networks to produce a single consensus network, under a parsimony model, i.e., the number of newly generated nodes is minimized. The integrity of the network is maintained by appropriately assigning segments to the new nodes such that a sample belongs to exactly one leaf node. Note that the problem of node minimization has a unique solution, but that the number of nodes could be very large, and very distant segments could be joined, which is genetically implausible. Hence, we parameterize the ARG with a positive integer *d*, which controls the extent of refinement (or size) of the network. We call this a subARG, whose size is a function of *d*.

An implementation of the recombination detection and subARG construction algorithms has been released in the public domain (https://researcher.ibm.com/researcher/view_project.php?id=2303) in the software package IRiS (Javed et al. 2011), which in addition to the executables also contains a detailed user manual describing the file formats.

Validation in coalescent simulations

The subARG is validated by coalescent simulations using the software cosi (Schaffner et al. 2005). A hundred independent coalescent ARGs are created based on randomly generated recombination profiles representing LD patterns at different genomic regions of length 200 kbp, using best fit model parameters as estimated in Schaffner et al. (2005). Only those with 5,000 nodes or less are retained for the use in validation. The simulation allows for the complete genealogy of each extant sequence, along



Detected recombination

r_1: {f},d_{11}: {d,e,f},d_{12}: {f,g}, breakpoint: α

r_2:{h,i},d_{21}:{f,g,h,i},d_{22}:{h,i,j,k,l},breakpoint: β

Fig. 2 A diagram illustrating a simple example of subARG construction. **a** Depicts the haplotypes at current samples. Two neighboring recombinations r_1 and r_2, are detected. Let d_{11} and d_{12} carry the detected left and right donor haplotypes of r_1, respectively. Similarly d_{21} and d_{22} carry the detected left and right donor haplotypes of r_2, respectively. Note that these nodes may not be an immediate parent of a recombinant (see "Subjects and methods"). The local topologies inferred from the recombinations are depicted in (**b**) and combined to construct a subARG in (**c**). The true ARG is shown in (**d**) to highlight the fidelity of the subARG

with genomic segments borne by each node in the ARG, to be known.

To validate the topology of a reconstructed subARG, we compute its fidelity to its corresponding cosi ARG. We compare both the nodes and the edges of the subARG to that of the corresponding cosi ARG. To measure the equivalence of the nodes, we compare the descendant sets using the Jaccard index (J), which is a single value that combines precision as well as recall. We create an appropriate threshold for J based on its (estimated) null distribution to detect the false positive nodes in the subARG. The coverage of the subARG is defined as the ratio of the number of true positive nodes to the total number of nodes in the cosi ARG.

It should be noted here that multiple nodes in a neighboring topology could yield nearly identical sets of descendants, rendering them indistinguishable from a reconstructability perspective. To resolve this multiplicity, as well as to validate the connectivities, each edge $v_i v_{\mathrm{j}}$ in the subARG is checked for the existence of a path in the mapped nodes of v_i and v_j in the cosi ARG restricted to the segment borne by v_j. If such a path does not exist, then edge $v_i v_j$ is a false positive; otherwise it is a true positive. The connectivity measure is the ratio of true positive edges to the total number of edges in the subARG. On average, 88% of subARG edges map to a path in the corresponding cosi ARG.

Actually, in a tree, the descendants of a node v are rather straightforward to compute: chromosome sample u is a descendant of node v if there exists a path from v to u, in the genealogical tree. However, in an ARG, a path is valid only if all of the edges in the path belong to at least one marginal tree in the ARG. In the evaluation process, we maintain the integrity of the paths by tracking only the segments carried by the subARG nodes as we traverse edges of the true ARG. To quantify the coverage of the cosi ARG, the focus is placed only on the internal nodes. Figure 3b shows the coverage achieved among non-leaf nodes in each simulation.

To summarize, we validate the subARG topology along three mutually independent directions. These include: (a) the precision and recall of every true positive node in the subARG; (b) the node coverage of the cosi ARG defined as the ratio of the number of true positive nodes of the subARG to the total number of nodes in the cosi ARG; and (c) the connectivity ratio of the subARG with respect to the cosi ARG. In isolation, each of the three measures could be substantially improved at the cost of the other two through trivial changes to the reconstructed subARG. However, the three measures together approximate fairly the fidelity of the subARG to the true ARG.

Estimating the age of the subARG nodes

The age of an allele correlates with its frequency in extant samples (Watterson and Guess 1977). In the subARG, an estimate of the age of a node can be used to compute the distance between two chromosomes as the age of their last common ancestor (LCA) node in the subARG. We applied the expected age of the node (in units of Ne generations) based on the estimate in Kimura and Ohta (1969):

$$E(\text{age}) = \frac{-2p}{1-p}\ln(p)$$

where p is the relative frequency of the extant descendants of a node. The performance of the estimate is evaluated via coalescent simulations where the true age of each node is known. The difference between the true age and the estimated age was computed for each node during topology validation and is depicted in Fig. 3c, d.

The genetic distance between a pair of individuals can be defined in terms of the depth of their LCA. However, in an ARG, due to recombinations, the neighboring segment may not be co-inherited from the same ancestor; hence, there exist multiple possible LCAs. Furthermore, to ensure that the LCA carries common ancestral material of the two nodes, the LCA and the two nodes must all lie on at least one of the marginal trees of the recombining segment. Therefore, we average the age of all the LCAs across all the non-recombining segments. Since the subARG is not completely resolved, depending on the extant samples and the confidence in the reconstructed recombinations, the relationship between every pair of samples may not be defined at each segment; such segments do not contribute to the average.

A new haplotype definition

Defining haplotype length in population genetic studies is not straightforward. Longer haplotypes retain larger diversity and provide higher differentiation among samples, to the extreme case in which each individual is different. The other extreme is a purely uncorrelated SNP-based analysis where the information about neighboring markers is omitted. We define a measure of informativity for haplotypes that incorporates both the number of different haplotypes and their average frequencies:

$$I_L = \frac{1}{\text{ncol}}\sum_{i=1}^{\text{ncol}} h_i \sum_{i=1}^{\text{ncol}} \frac{N}{h_i}$$

where ncol is the number of columns obtained when dividing each sequence into windows of L SNPs, h_i is the number of different haplotypes found in each column i, and N is the number of sequences. Informativity was calculated for

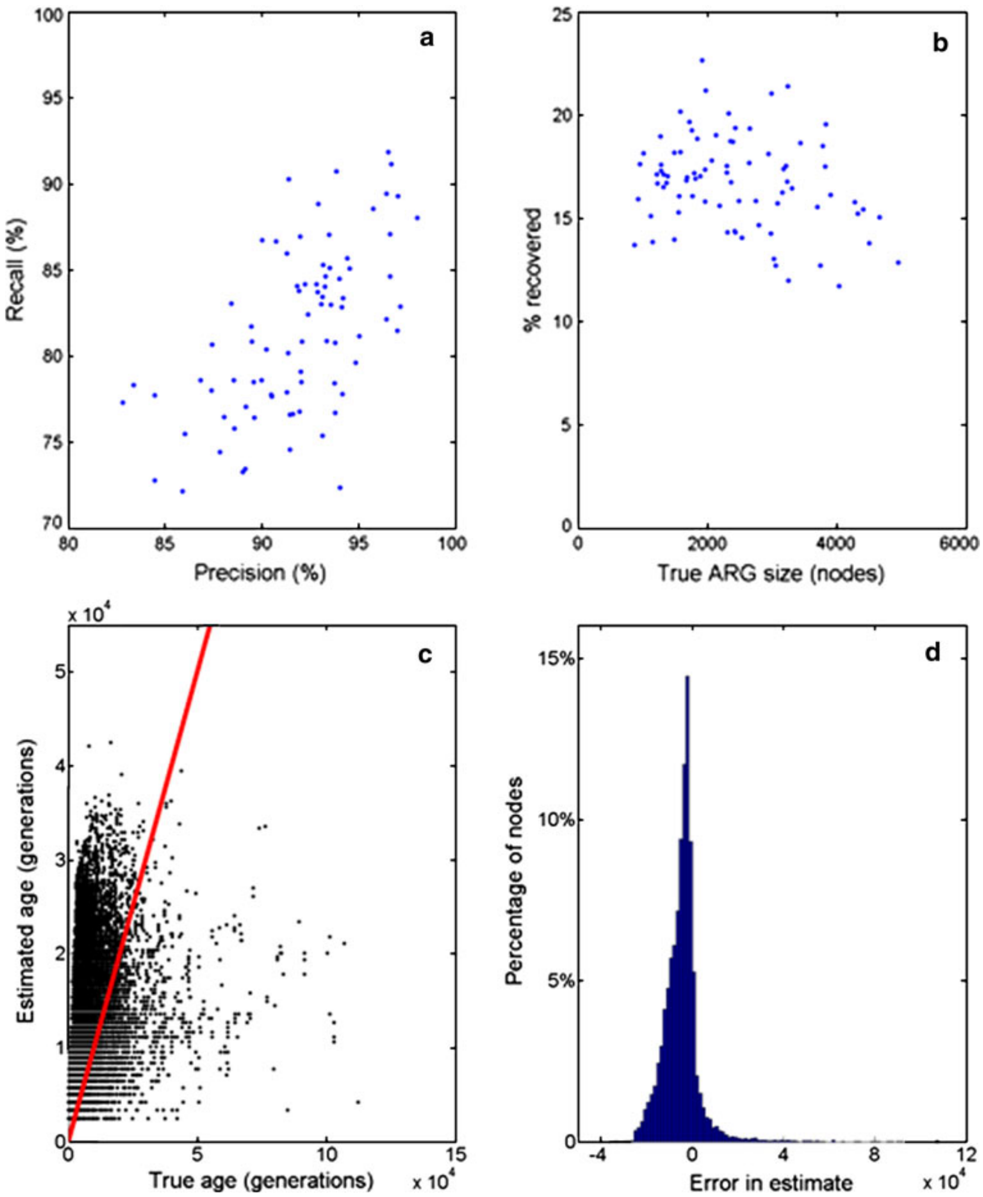


Fig. 3 Each data point in subplot **a** represents the precision and recall among descendant sets across all the subARG nodes in a coalescent simulation. Each data point in the scatter plot **b** represents the percentage of internal true ARG nodes recovered in the subARG. **c**, **d** represent the performance of the age estimate. One hundred coalescent simulations were conducted using cosi and ARGs of 5,000 nodes or less were retained. The true age of a subARG node is the age of the node to which it maps in the true ARG. **c** the plot of the true age versus the estimate. Each point in the scatter plot represents a node in the simulations. The line represents perfect prediction. **d** Plots the histogram of prediction error. It is defined as the difference between true node age and the estimate

lengths 5, 10, 20, 30, 40, 60, 80, 100, and 120 SNPs. For each length *L*, five different informativity values were calculated by changing the starting position of the first window from position 1 and increasing it in L/5 steps; all five values were subsequently averaged. The highest average informativity value laid between lengths 30 and 40 SNPs and, thus, we performed the same calculation for lengths 30, 32, 34, 36, 38 and 40 taking as starting positions all even positions (Supplemental Figure 3). The haplotype length having the highest average informativity was 38 SNPs, and this was the length used in all subsequent analyses.

Genetic diversity analysis

We applied dimensionality reduction methods and a Bayesian population structure algorithm to our data set in four different ways, based on SNPs, the subARG, haplotypes defined as in Jakobsson et al. (2008), and fixed-length haplotypes (see above). Jakobsson et al. (2008) define 20 unique haplotype clusters at each marker. For every individual at each SNP, the probability of assignment to each of the 20 clusters is estimated using fastPHASE (Scheet and Stephens 2006). Eigensoft (Patterson et al. 2006) was used to apply PCA to SNP and haplotype data, with the LD correction parameter. Since Eigensoft requires biallelic data, every multiallele haplotype column was split into multiple biallelic columns, as suggested in Patterson et al. (2006). To compute the PCA for the subARG data, the graph relations need to be encoded as a matrix. This matrix is constructed by representing each non-leaf node as a column with the corresponding sample values indicating the progeny of this node; i.e., matrix entry (i, j) is 1 if the sample i is descendant of the node *j*; otherwise it is 0. The distance between pairs of individuals from different populations were averaged (allele and haplotype sharing percentages were used for allele and haplotype data, respectively) and represented with multidimensional scaling (MDS) using Matlab.

Bayesian clustering was performed using Structure (Falush et al. 2003; Pritchard et al. 2000). A subset of uncorrelated tagging SNPs was selected using Haploview (Barrett et al. 2005) for SNP-based analysis. For fixed-length haplotypes, we applied Bayesian clustering on five different data sets with varying starting positions. Jakobsson et al. (2008) defined haplotypes by a probabilistic cluster assignment across 20 clusters for each sample at every marker. The probability matrix was sampled to create 10 different data sets. The subARG-based matrix, as encoded for PCA, was also analyzed by this method. For each methodology and number of clusters, CLUMPP (Jakobsson and Rosenberg 2007) was used to reconcile multiple replicates in a manner similar to Wang et al. (2007) and the results were displayed using *distruct* (Rosenberg et al. 2002). All runs had a burn-in period of 50,000 iterations followed by 50,000 iterations of sampling; the admixture model (with $K = 2$ to $K = 4$) was used.

Correlations with geography and language

We adopted the language classification compiled by the Ethnologue (http://www.ethnologue.org) (Supplemental Table 4). Recently migrated (European Americans) or admixed (African Americans, Mexicans) populations were not considered. Furthermore, languages that do not share a relative do not contribute to the comparative analysis and, hence, were not considered (Basque, Chechen, Ati, Lao, Japanese). Linguistic distances were estimated according to the method defined by Excoffier et al. (1991). Populations belonging to different linguistic families are assigned a distance of 4. Within the same family, populations having distinct language stock or group were assigned distances 3 or 2, respectively. Different languages within the same group were assigned a distance of 1. Parayar and Cape Nadar both speak Tamil and were therefore assigned a distance of 0. In addition, pairwise great circle distance was computed between these populations using Cairo Egypt (30N, 31E) and Istanbul Turkey (41N, 28E) as waypoints to Asia and Europe, respectively.

Results

We have analyzed a final data set consisting of the genotypes of 1,255 SNPs in five regions of the X chromosome in 1,318 (mostly male) samples belonging to 33 human populations; genotype data are available at (https://researcher.ibm.com/researcher/view_project.php?id=2303) We have devised a new method to analyze SNP data based on reconstructing a subARG, which has been validated before applying it to the current data set.

subARG validation

We validated the subARG construction method by means of coalescent simulations with a standard human demography (Schaffner et al. 2005). Figure 3a shows a scatter plot of the precision and recall achieved, in the descendant sets, for each simulation. Overall, across all of the ARGs, the descendant sets exhibited 92% precision and 81% recall. The average precision and recall values were virtually identical for both the nodes generated in recombination detection, and those subsequently computed in the subARG. On average, 17% of the non-leaf true ARG nodes were recovered in each subARG. In this regard, a true node in the subARG may not be called as such if a false positive node exists downstream of it. That is, the recall may be underestimated for deeper nodes, resulting in a globally conservative recall statistic.

Figure 3c and d depicts the performance of the age estimate; the relationship with allele frequency is shown in Supplemental Figure 1. The true age of a subARG node is the age of the node to which it maps in the true ARG. In general, the average estimate tends to be higher to compensate for the long tail of the age distribution. The actual age distribution of the nodes recovered from cosi simulations can be seen in Supplemental Figure 2.

Genetic diversity

The average distance between populations based on four types of analysis (subARG, SNPs, probabilistic haplotypes, fixed-length haplotypes) is represented by means of MDS in Fig. 4. All four types of analysis gave similar results, with Sub-saharan African populations being clearly separated from the other populations. The remaining populations were sorted in a gradient from Europe through South Asia to East Asia. North African populations were consistently slightly closer to Sub-saharan Africans, whereas the Spanish Gypsies showed affinities with South Asians. The northeast Indian Adi clustered with East Asian populations, as did the isolated Philippine Negrito Ati. The third MDS dimension separated Indian populations from the remaining groups, although only in the subARG analysis.

When expanded to consider inter-individual distances and plotted by means of PCA (Fig. 5), some small differences emerged among the four methods. A clearer pattern of population relationships was observed in the subARG-based plot, with a tighter clustering of Sub-saharan

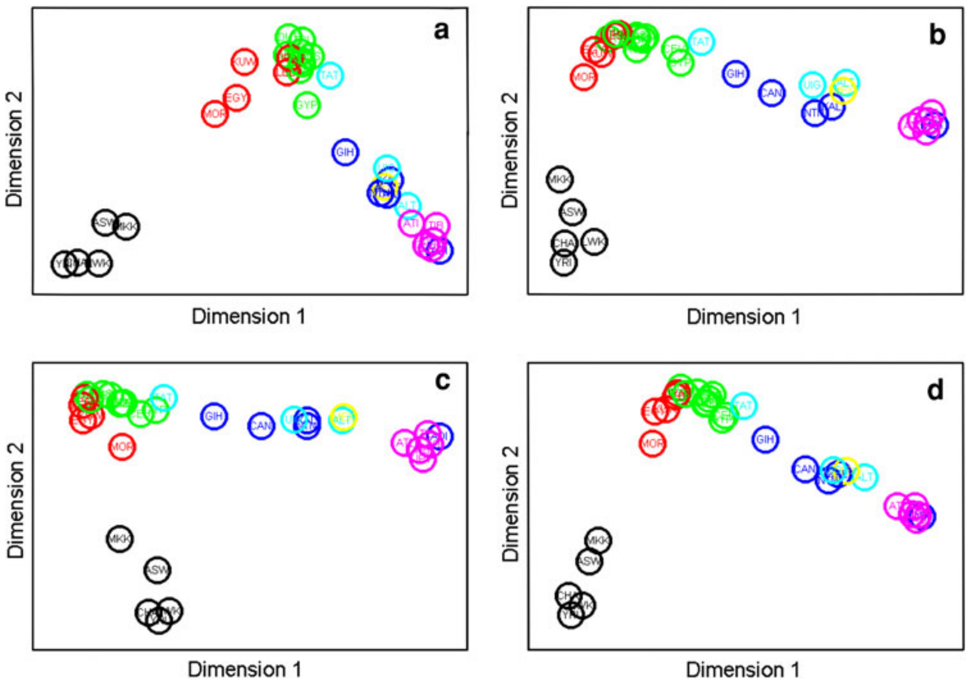


Fig. 4 MDS plots of pairwise population distances estimated from the **a** subARG, **b** allele sharing at SNPs, **c** probabilistic haplotypes (Jakobsson et al. 2008), and **d** fixed-length haplotypes. Population abbreviations are provided in Supplemental Table 1

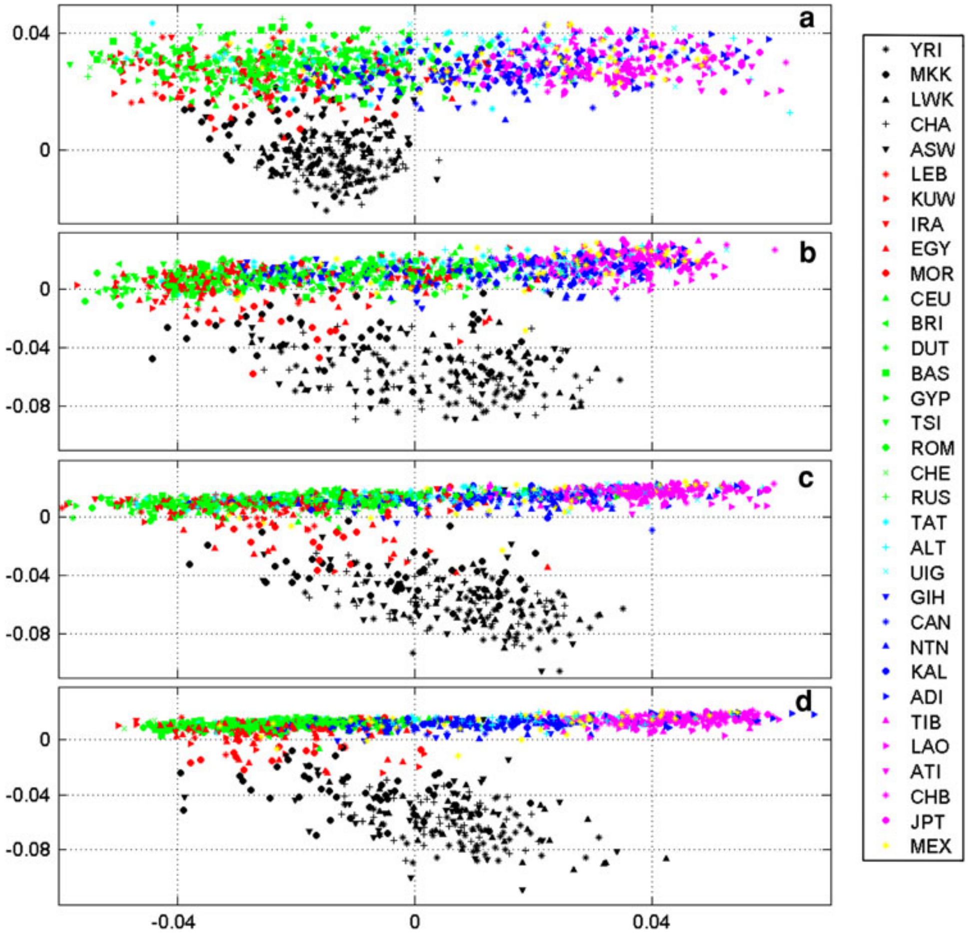


Fig. 5 Principal component analysis computed from **a** subARG, **b** SNPs, **c** haplotypes defined as in Jakobsson et al. (2008), and **d** fixed-length haplotype-based matrices

individuals and a clearer boundary between Sub-saharan individuals and others. By contrast, the boundaries within Eurasia were clearer in the fixed-length haplotype plot. It is notable that PCA plots (particularly the subARG-based one) represent the pattern resembling the geographical map of the Old World, where African, European and East Asian populations find their place according to their actual geographic position (similar results were obtained (Novembre et al. 2008) for the genetic variation within Europe).

The Bayesian clustering algorithm implemented in STRUCTURE produced similar results for the subARG, SNPs, and probabilistic haplotypes (Fig. 6). At $K = 3$, a fuzzy classification of individuals into Sub-saharan African, European, and East Asian components emerged; at $K = 4$, the European component appeared to be randomly split into two. However, for fixed-length haplotypes (Fig. 6), most individuals were largely assigned to a single component, and, at $K = 4$, the new component clearly mapped to South Asia. In Sub-saharan Africa, the European component appeared in some Maasai and African-American individuals; conversely, some Egyptians and Moroccans contained Sub-saharan African haplotypes. The South Asian component was quite apparent in Spanish Gypsies, and reached its highest frequencies in South Indians (Cape Nadar and Parayar); the European component was frequent in the northwest Indian Gujarati, while the East Asian component appeared in the northeast Indian Kalita.

The degree of concordance in the subARG population distances and SNP and haplotype-based measures was quantified using Mantel tests. The pairwise population distances in the subARG correlated with SNPs ($r = 0.44$), haplotypes based on Jakobsson et al.'s definition ($r = 0.40$), and haplotypes based on fixed window size ($r = 0.25$), each with $p < 10^{-6}$. To estimate the apportionment of genetic diversity within and across continental groups and populations, an AMOVA was performed using Arlequin (Excoffier and Lischer 2010). The results indicate that 88–94% of the variance observed is found within populations. The highest discrimination between continental groups was reached with SNPs (9.4%), followed by the subARG (7.24%), probabilistic haplotypes (5.13%), and fixed-length haplotypes (4.62%). The correlation between genetic, linguistic, and geographic distances is shown in Table 1. The results indicated that geography has played a more important role than linguistics in shaping the genetic differentiation of the study populations (Belle and Barbujani 2007).

Discussion

Recombination, SNPs and haplotypes

The combination of SNPs into haplotypes adds a new dimension to studies of genetic diversity. Besides mutation,

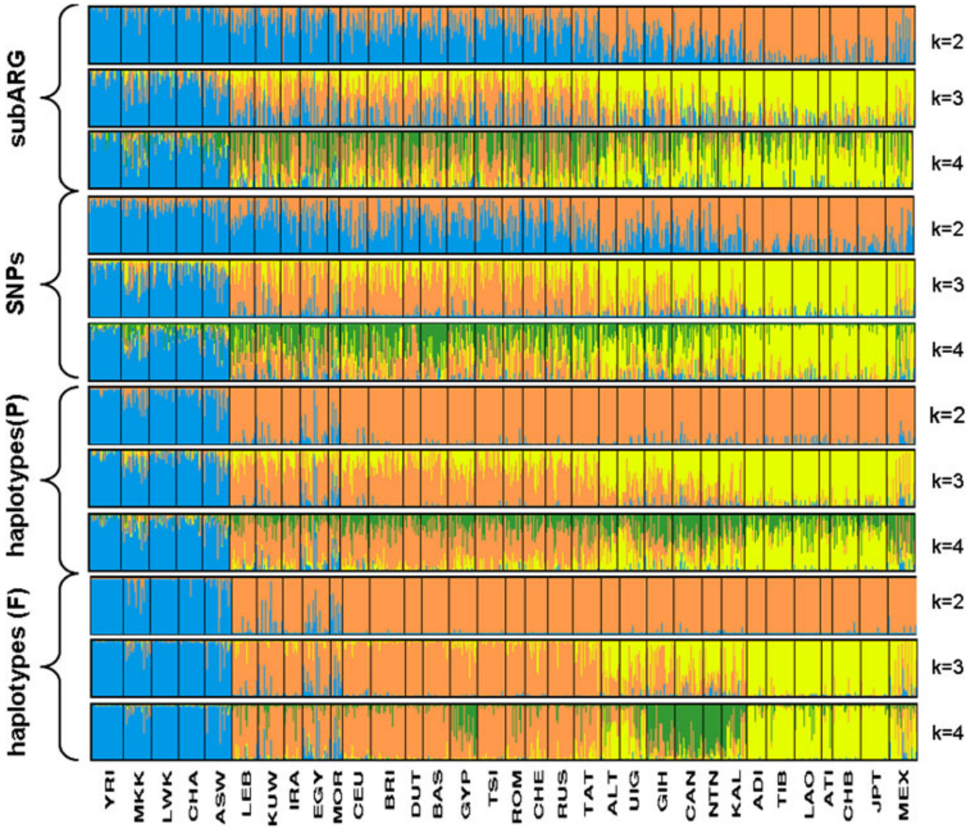


Fig. 6 STRUCTURE plots based on the subARG, SNPs, probabilistic haplotypes (haplotypes P) defined as in Jakobsson et al. (2008), and fixed-length haplotypes (haplotypes F). In each case, plots with $K = 2$ through $K = 4$ are shown

Table 1 Correlation between genetic, linguistic and geographic population distances

Genetic distance measure	Linguistic correlation		Geographic correlation		Linguistic correlation (after correction)		Geographic correlation (after correction)	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*
SNPs	0.32	$<10^{-3}$	0.63	$<10^{-3}$	0.24	$<10^{-3}$	0.60	$<10^{-3}$
Probabilistic haplotypes	0.32	$<10^{-3}$	0.56	$<10^{-3}$	0.23	$<10^{-3}$	0.54	$<10^{-3}$
Fixed-length haplotypes	0.30	$<10^{-3}$	0.50	$<10^{-3}$	0.22	$<10^{-3}$	0.46	$<10^{-3}$
subARG	0.26	0.002	0.55	$<10^{-3}$	0.16	0.011	0.52	$<10^{-3}$

recombination has to be reckoned with as a diversity-generating process. For the first time, we have attempted to reconstruct the phylogenetic structure driven by recombination using the ARG. A complete reconstruction is demonstrably unfeasible, but we have shown that we can provide a subARG with great precision and recall. The subARG provides a new genetic analysis tool which, as we will discuss below, complements existing methods. It is easily scalable to much larger genomic regions than those we analyzed here, and, although it is not suitable for tag SNP-based whole genome arrays, it is perfectly applicable to whole genome resequencing data, where variation is presented regardless of the underlying LD structure, that is, independently of the past recombinational history.

We have explicitly incorporated recombination into our genetic analysis, and have performed a parallel study based on SNPs, the subARG, and haplotypes. The latter naturally combine the effects of mutation, recombination, and demography, but their definition is problematic. The diversity observed in a haplotype depends on the number of SNPs conjoined to form it. If the window size is too small, then the correlation between neighboring loci is lost, and the results revert to those of SNP-based methods; conversely, too large of a window makes each individual unique, and frequency- and sharing-based methods become meaningless. Jakobsson et al. (2008) provided a solution to this conundrum by probabilistically assigning each SNP in each individual to one of 20 haplotype clusters. In the present paper, we apply their method, but also suggest an alternative heuristic in which haplotype length is fixed but optimized for information content, by balancing haplotype numbers and frequencies.

These methods appear complementary to each other. While fixed-length haplotypes provided much clearer individual assignments in STRUCTURE and were the only method to yield a South Asian component, the subARG defined better the separation of Sub-saharan African individuals in the PCA. The latter result may be a consequence of the fact that fixed-length haplotypes had to be optimized for the global sample. Given the lower LD in Sub-saharan Africa, the fixed length that we used was probably too large for that continental region. Analyses that are biased by frequency range will tend to work better with biallelic rather than with multiallelic markers, given than the former are less constrained in their frequencies that the latter. This may explain why SNPs showed larger fractions of the genetic diversity attributed to differences between continental groups.

Old World genetic diversity

This study has been performed with a large number of populations from the Old World, and represents one of the largest surveys of human genetic variation. Although our data set overlap in geographic coverage with other sets such as HGDP, it contains particular features such as the representation of India as well as singular populations such as Gypsies and the Ati, among others.

When analyzing the apportionment of variation to populations and continents, the explicit analysis of recombination in the subARG or its implicit representation in the haplotypes showed that the amount of variation found within populations is larger than that found by SNPs, and, conversely, the variation explained by differences between continental groups is smaller. This result is expected if the recombination events captured are more recent than the mutational events creating the SNPs. Such a finding could be due to the palimpsestic nature of recombination and to the biases implicit in its detection. In general, the results given by recombination are expected to be related to more recent historical events (Melé et al. 2010); the combination of mutation and recombination that is provided by haplotypes may thus capture a wider timeframe of population history.

Independently of the framework used to analyze the data, the genetic structure of the populations correlated with geographic distance, while linguistic classification failed to account for genetic differentiation once the effect of geography was removed. Even though the recombination events detected tend to be recent, most of them may be ancient enough to go beyond the inception of the major linguistic branches, and, thus, may not adequately reflect fast linguistic changes.

Our results are consistent with the out of Africa hypothesis, as African populations are the most differentiated and most internally diverse from compared to other populations in both analyses. Within Africa, both the Maasai and the African-Americans seem to have some West Eurasian or Middle Eastern component, as seen with all the types of analyses employed in this study. In genomewide studies, the Maasai exhibit an East-African specific component (Tishkoff et al. 2009; Henn et al. 2011). In African-Americans, this finding could be explained by their known recent admixture, which may be slightly underestimated by the X chromosome. A male-mediated European admixture would result in a 1:2 ratio of European to African X chromosomes being transmitted. The West Eurasian component in the Maasai can be explained by them being the descendants of populations ancestral to non-Africans and/or gene flow from non-Africans into Africa. In addition, some Middle Eastern and North African populations, such as the Moroccan, Egyptian, and Kuwaiti populations, show discernable traces of African admixture in the STRUCTURE plots, with all methods of analysis employed (subARG/SNPs/haplotypes).

In Europe, the most outstanding result is the clear demonstration of the Indian origin and West Eurasian admixture of Gypsies, which had been shown before using unilinearly transmitted markers [(Mendizabal et al. 2011) and references therein]. This could most clearly be seen with the STRUCTURE plots of fixed-length haplotypes. In the Central Asian continuum of genetic variation, the Tatars showed the smallest East Asian contribution, which was higher in the more easterly located Uighur and Altaians.

In the Bayesian clustering analysis, a component that mostly occurred in Indian populations was revealed only when optimal fixed-length haplotypes were used. Our data set contained two populations from southern India, where this component reached its highest frequencies. Thus, it is possible that this component captures a predominantly south Indian dimension of genetic variation. These high frequencies would explain why this component appears somewhat diluted in the northwest Indian, Indo-European speaking Gujarati, as well as in the northeastern Indian Kalita, which shows West and East Eurasian genetic contributions, respectively. The fact that the Indian component appears in the Lao of SE Asia and in the Ati Negritos of the Philippines may imply that this component may have captured some of the contribution of the southern route out of Africa. Nevertheless, the Ati were clearly linked to East Asian populations, as was shown with unilinear markers (Delfin et al. 2011; Gunnarsdottir et al. 2011). This lack of distinctiveness has strong implications for the peopling of Asia. Traditionally (Cavalli-Sforza et al. 1994), Negritos (together with Melanesians and Australian Aborigines) were regarded as populations directly descending from a first wave of anatomically modern humans that emigrated out of Africa, while other Asians were thought to derive from a more recent migration. Our results support a single migration wave out of Africa into Asia, and a *maturation phase* in South Asia (Dennell and Roebroeks 2005; Macaulay et al. 2005) prior to expansion into regions to the east.

The admixed nature of the Mexican population was also revealed. The lack of Native American reference samples could explain why the majority component was East Asian; sex-biased gene flow may have led to an overestimate of this component. The diversity in individual histories, with varying degrees of Native American versus European ancestry, is apparent in both the Bayesian cluster results (which also reveal, in some individuals, an African contribution) and the wide area occupied by Mexicans in the PCA graph (Fig. 5).

This study has demonstrated that a combination of methods, including a recombination-based approach, allow the extraction of a large amount of genetic information from genomic data. Having analyzed only 0.075% of the genome, we have been able to recover many of the patterns seen with much larger data sets using different sets of unilineal and autosomal markers.

Acknowledgments This research is part of the Genographic Project, funded by National Geographic and IBM. Additional funding was provided by the Spanish Ministry of Science and Innovation projects BFU2007-63657, BFU2007-63171, and BFU2010-19443; MM was supported by grant AP2006-03268, Generalitat de Catalunya; OB was supported by Russian Foundation for Basic Research (grants 10-06-00451, 10-04-01603) and by the Presidium RAS Programme “Molecular and Cell Biology”. We are grateful to Mònica Vallés, UPF, for excellent technical support. Genotyping and bioinformatic services were provided respectively by CEGEN (Centro Nacional de Genotipado) and INB (National Bioinformatics Insitute), Spain. HapMap phase III population samples were obtained from the Coriell Cell Repository.

Appendix

The Genographic Consortium includes: Syama Adhikarla (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Christina J. Adler (University of Adelaide, South Australia, Australia), Danielle A. Badro (Lebanese American University, Chouran, Beirut, Lebanon), Andrew C. Clarke (University of Otago, Dunedin, New Zealand), Alan Cooper (University of Adelaide, South Australia, Australia), Clio S. I. Der Sarkissian (University of Adelaide, South Australia, Australia), Matthew C. Dulik (University of Pennsylvania, Philadelphia, Pennsylvania, USA), Christoff J. Erasmus (National Health Laboratory Service, Johannesburg, South Africa), Jill B. Gaieski (University of Pennsylvania, Philadelphia, Pennsylvania, USA), Wolfgang Haak (University of Adelaide, South Australia, Australia), Angela Hobbs (National Health Laboratory

Service, Johannesburg, South Africa), Matthew E. Kaplan (University of Arizona, Tucson, Arizona, USA), Shilin Li (Fudan University, Shanghai, China), Begoña Martínez-Cruz (Universitat Pompeu Fabra, Barcelona, Spain), Elizabeth A. Matisoo-Smith (University of Otago, Dunedin, New Zealand), Nirav C. Merchant (University of Arizona, Tucson, Arizona, USA), R. John Mitchell (La Trobe University, Melbourne, Victoria, Australia), Amanda C. Owings (University of Pennsylvania, Philadelphia, Pennsylvania, USA), Daniel E. Platt (IBM, Yorktown Heights, NY, USA), Lluis Quintana-Murci (Institut Pasteur, Paris, France), Colin Renfrew (University of Cambridge, Cambridge, UK), Daniela R. Lacerda (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil), Ajay K. Royyuru (IBM, Yorktown Heights, NY, USA), Fabrício R. Santos (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil), Theodore G. Schurr (University of Pennsylvania, Philadelphia, Pennsylvania, USA), Himla Soodyall (National Health Laboratory Service, Johannesburg, South Africa), David F. Soria Hernanz (National Geographic Society, Washington, DC, USA), Pandikumar Swamikrishnan (IBM, Somers, NY, USA), Chris Tyler-Smith (The Wellcome Trust Sanger Institute, Hinxton, UK), Kavitha Valampuri John (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Arun Varatharajan Santhakumari (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Pedro Paulo Vieira (Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil), R. Spencer Wells (National Geographic Society, Washington, DC, USA), Janet S. Ziegle (Applied Biosystems, Foster City, CA, USA).

References

Baird SJ (2006) Phylogenetics: Fisher's markers of admixture. Heredity 97(2):81–83

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263–265

Belle EM, Barbujani G (2007) Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. Am J Phys Anthropol 133(4):1137–1146

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) History and geography of human genes. Princeton University Press, Princeton

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38(11):1251–1260

Delfin F, Salvador JM, Calacal GC, Perdigon HB, Tabbada KA, Villamor LP, Halos SC, Gunnarsdottir E, Myles S, Hughes DA, Xu S, Jin L, Lao O, Kayser M, Hurles ME, Stoneking M, De Ungria MC (2011) The Y-chromosome landscape of the Philippines: extensive heterogeneity and varying genetic affinities of Negrito and non-Negrito groups. Eur J Hum Genet 19(2):224–230

Dennell R, Roebroeks W (2005) An Asian perspective on early human dispersal from Africa. Nature 438(7071):1099–1104

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567

Excoffier L, Harding RM, Sokal RR, Pellegrini B, Sanchez-Mazas A (1991) Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. Hum Biol 63(3):273–307

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4):1567–1587

Fisher RA (1954) A fuller theory of "Junctions" in inbreeding. Heredity 8:187–197

Gunnarsdottir ED, Li M, Bauchet M, Finstermeier K, Stoneking M (2011) High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res 21(1):1–11

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD (2010) The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet 42(10):830–831

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodríguez-Botigué L, Ramachandran S, Hon L, Brisbin A, Lin AA, Underhill PA, Comas D, Kidd KK, Norman PJ, Parham P, Bustamante CD, Mountain JL, Feldman MW (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci USA 108(13):5154–5162

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23(14):1801–1806

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181):998–1003

Javed A, Pybus M, Mele M, Utro F, Bertranpetit J, Calafell F, Parida L (2011) IRiS: construction of ARG networks at genomic scales. Bioinformatics 27(17):2448–2450

Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. Genetics 61(3):763–771

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, Gudjonsson SA, Frigge ML, Helgason A, Thorsteinsdottir U, Stefansson K (2010) Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467(7319):1099–1103

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308(5724):1034–1036

Mateu E, Calafell F, Lao O, Bonne-Tamir B, Kidd JR, Pakstis A, Kidd KK, Bertranpetit J (2001) Worldwide genetic analysis of the CFTR region. Am J Hum Genet 68(1):103–117

Mateu E, Perez-Lezaun A, Martinez-Arias R, Andres A, Valles M, Bertranpetit J, Calafell F (2002) PKLR- GBA region shows almost complete linkage disequilibrium over 70 kb in a set of worldwide populations. Hum Genet 110(6):532–544

Melé M, Javed A, Pybus M, Calafell F, Parida L, Bertranpetit J (2010) A new method to reconstruct recombination events at a genomic scale. PLoS Comput Biol 6(11):e1001010

Mendizabal I, Valente C, Gusmao A, Alves C, Gomes V, Goios A, Parson W, Calafell F, Alvarez L, Amorim A, Gusmao L, Comas D, Prata MJ (2011) Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. PLoS One 6(1):e15988

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456:98–101

Parida L, Palamara PF, Javed A (2011) A minimal descriptor of an ancestral recombinations graph. BMC Bioinformatics 12(Suppl 1):S6

Patterson N, Price AL, Reich D (2006) Population structure and eigen analysis. PLoS Genet 2(12):e190

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298(5602):2381–2385

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576–1583

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4):629–644

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76(3):449–462

Tishkoff SA, Dietzch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonné-Tamir B, Santachiara-Benerecetti S, Moral P, Krings M, Pääbo S, Watson E, Risch N, Jenkins T, Kidd KK (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am J Hum Genet 62(6):1389–1402

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM (2009) The genetic structure and history of Africans and African Americans. Science 324(5930):1035–1044

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, Mazzotti G, Poletti G, Hill K, Hurtado AM, Labuda D, Klitz W, Barrantes R, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Llop E, Rothhammer F, Excoffier L, Feldman MW, Rosenberg NA, Ruiz-Linares A (2007) Genetic variation and population structure in native Americans. PLoS Genet 3(11):e185

Watterson GA, Guess HA (1977) Is the most frequent allele the oldest? Theor Popul Biol 11:141–160