J Mol Evol (2008) 67:465–487
DOI 10.1007/s00239-008-9148-7

PCA and Clustering Reveal Alternate mtDNA Phylogeny of N and M Clades

G. Alexe · R. Vijaya Satya · M. Seiler · D. Platt · T. Bhanot · S. Hui · M. Tanaka · A. J. Levine · G. Bhanot

Received: 16 February 2008 / Accepted: 14 July 2008 / Published online: 15 October 2008

Abstract Phylogenetic trees based on mtDNA polymorphisms are often used to infer the history of recent human migrations. However, there is no consensus on which method to use. Most methods make strong assumptions which may bias the choice of polymorphisms and result in computational complexity which limits the analysis to a few samples/polymorphisms. For example, parsimony minimizes the number of mutations, which biases the results to minimizing homoplasy events. Such biases may miss the global structure of the polymorphisms altogether, with the risk of identifying a "common" polymorphism as ancient without an internal check on whether it either is homoplasic or is identified as ancient because of sampling bias (from oversampling the population with the polymorphism). A signature of this problem is that different methods applied to the same data or the same method applied to different datasets results in different tree topologies. When the results of such analyses are combined, the consensus trees have a low internal branch consensus. We determine human mtDNA phylogeny from 1737 complete sequences using a new, direct method based on principal component analysis (PCA) and unsupervised consensus ensemble clustering. PCA identifies polymorphisms representing robust variations in the data and consensus ensemble clustering creates stable haplogroup clusters. The tree is obtained from the bifurcating network obtained when the data are split into $k = 2,3,4,\ldots,k_{\max}$ clusters, with equal sampling from each haplogroup. Our method assumes only that the data can be clustered into groups based on mutations, is fast, is stable to sample perturbation,

G. Alexe and R. Vijaya Satya—Joint first authors.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9148-7) contains supplementary material, which is available to authorized users.

G. Alexe
The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

G. Alexe · A. J. Levine · G. Bhanot (✉)
Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540, USA
e-mail: gyanbhanot@gmail.com

R. Vijaya Satya
School of Computer Science, University of Central Florida, Orlando, FL 32816, USA

M. Seiler · S. Hui · G. Bhanot
BioMaPS Institute, Rutgers University, Piscataway, NJ 08854, USA

D. Platt
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

T. Bhanot
Graduate Program in Microbiology & Molecular Genetics, Rutgers University, Piscataway, NJ 08854, USA

M. Tanaka
Tokyo Metropolitan Institute of Gerontology, 35-2 Sakae-cho, Itabashi-ku, Tokyo 173-0015, Japan

G. Bhanot
Department of Physics and Department of Molecular Biology & Biochemistry, Rutgers University, Piscataway, NJ 08854, USA

G. Bhanot
Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08903, USA

uses all significant polymorphisms in the data, works for arbitrary sample sizes, and avoids sample choice and haplogroup size bias. The internal branches of our tree have a 90% consensus accuracy. In conclusion, our tree recreates the standard phylogeny of the N, M, L0/L1, L2, and L3 clades, confirming the African origin of modern humans and showing that the M and N clades arose in almost coincident migrations. However, the N clade haplogroups split along an East-West geographic divide, with a "European R clade" containing the haplogroups H, V, H/V, J, T, and U and a "Eurasian N subclade" including haplogroups B, R5, F, A, N9, I, W, and X. The haplogroup pairs (N9a, N9b) and (M7a, M7b) within N and M are placed in nonnearest locations in agreement with their expected large TMRCA from studies of their migrations into Japan. For comparison, we also construct consensus maximum likelihood, parsimony, neighbor joining, and UPGMA-based trees using the same polymorphisms and show that these methods give consistent results only for the clade tree. For recent branches, the consensus accuracy for these methods is in the range of 1–20%. From a comparison of our haplogroups to two chimp and one bonobo sequences, and assuming a chimp-human coalescent time of 5 million years before present, we find a human mtDNA TMRCA of 206,000 ± 14,000 years before present.

Keywords mtDNA phylogeny · Principal component analysis · Unsupervised consensus ensemble clustering · Clade tree · Homoplasy · Time to most recent common ancestor

Introduction

Although it is agreed that modern humans emerged from Africa between 50 and 70 KYBP (thousand years before present) (Cann et al. 1987), details about their dispersal and migratory routes are still unclear (Harpending et al. 2005; Stringer 2001). Population movements are often inferred from trees based on mtDNA and Y chromosome polymorphisms. However, it is well known that each method often gives different trees depending on which samples and polymorphisms are used, and different methods give different trees for the same samples and polymorphisms. This is because of implicit assumptions in each method which bias it toward certain tree topologies, e.g., maximum parsimony (MP) (Densmore 2001; Stewart 1993; Yang 1996) minimizes the number of polymorphisms, maximum likelihood (ML) (Hasegawa et al. 1991; Jin et al. 2006; Minh et al. 2005; Saitou 1990; Sanderson 1994; Sullivan 2005; Yang 1997) optimizes a likelihood function (Drummond and Rodrigo 2000; Kumar and Gadagkar 2000; Ota and Li 2000; Pearson et al. 1999; Saitou and Nei 1987; Studier and Keppler 1988; Tamura et al. 2004), etc. Verifying the robustness of the inferred trees is compounded by the computational complexity of these methods, which limits them to small sample sizes and few polymorphisms (Felsenstein 1996). Estimating their accuracy using bootstrap analysis on samples and polymorphisms often shows poor consensus on internal branches (see below). Sampling biases in the data also skew the results in unknown ways.

In this paper, we develop a new method using principal component analysis (PCA) (Jolliffe 2002) and consensus ensemble clustering that avoids these problems. The procedure we follow is described below.

Materials and Methods

PCA is used to divide the data into clusters at multiple scales and identify robust polymorphisms that distinguish them. The PCA eigenvectors define linear combinations of polymorphisms which represent a clustering hierarchy of the samples. The number of clusters and sample membership in these clusters as discovered by PCA is not affected by sampling bias because multiple instances of similar samples will cluster close together. Thus, PCA is a simple and visual way to separate samples into clades/haplogroups in the presence of unknown sampling bias. One expects that ancient polymorphisms, representing clusters that have drifted far apart in mtSNP space, will appear on leading eigenvectors. More recent splits will be represented by nonleading eigenvalues and eigenvectors. A simple method we use to exploit this hierarchy of clusters is to first use PCA to separate samples into clades and then, recursively repeating the analysis within each cluster, stratify further into subclades, haplogroups, subhaplogroups, etc.

The polymorphisms representing structure at each stage in the hierarchy can be read off from the high absolute value coefficients of the leading PCA eigenvectors. Using these polymorphisms, the samples are divided into an optimum number of haplogroups using consensus ensemble clustering (Kaufmann and Rousserw 1990; Monti et al. 2003; Strehl and Ghosh 2002). This method produces robust haplogroups by averaging over several clustering methods (probabilistic, agglomerative, and hierarchical) and many bootstrapped datasets. The optimum number of clusters ($k_{\max}$) is inferred using statistical measures (silhouette score (Kaufmann and Rousserw 1990) and gap statistics (Tibshirani et al. 2001)). The robustness and statistical significance of the clusters and polymorphisms used are validated by averaging over 10 2:1 training/test bootstrapped datasets on which we repeat the complete clustering analysis. We require a polymorphism to have a classification accuracy of at least 90% in the training/test validations to be included in the analysis. We also require a 90% agreement across data perturbation and clustering methods for a pair of samples to be assigned to the same haplogroup.

When we create $k + 1$ clusters, we do not use the previous clustering into k clusters. Nevertheless, we find that the $k + 1$ clusters consist of $k - 1$ clusters from the previous clustering plus two more obtained from the splitting of one of the clusters at level k. These recursive bifurcations as $k \rightarrow k + 1$ are strong, inherent features of the data which we interpret as branch splits on the tree. The polymorphisms on branches are identified by a frequency analysis at each split.

Using equal sampling from each haplogroup cluster (to minimize sampling bias) and repeating the consensus clustering, the phylogeny is inferred from the sequence of splits as the number of clusters increases. The bifurcating network obtained in the reclustering defines an unrooted tree as the number of clusters (k) increases from 2 to $k_{\max}$. The root of the tree is identified in two ways: (i) as the internal node equidistant (i.e., with the same average number of mutations) from the leaf haplogroup clusters—this can be identified by simulating Poisson dynamics on the branch polymorphisms; and (ii) as the internal node closest to an outgroup (we use a consensus chimpanzee sequence).

Our method uses the global, "natural" structure of all polymorphisms in the data to infer phylogeny. It is insensitive to data perturbations, is robust against recent polymorphisms, and avoids sampling bias as much as possible. It was extensively validated on synthetic data generated by numerical simulation.

Results

Identifying Clades, Inferring Old Polymorphisms, and Building the Clade Tree

One thousand seven hundred thirty-seven complete mtDNA sequences were downloaded in April 2006 from the public databases http://www.mitomap.org/ and http://www.genpat.uu.se/mtDB/ and pairwise aligned with the Revised Cambridge Reference Sequence (rCRS) (see http://www.mitomap.org/mitoseq.html) using the algorithm Stretcher (Myers and Miller 1998) in Emboss (http://emboss.sourceforge.net). In addition, we used two chimp and one bonobo sequences in our analysis from NCBI (NCBI references: chimp 1, D38113; chimp 2,: X93335; bonobo, D38116). The data had 3177 mtSNPs, of which 90.5% were transitions, 4% transversions, and 5.5% multiallelic. This creates a 1737 × 3177 matrix $M_{i,j}$ whose entries correspond to polymorphisms relative to rCRS. PCA of the matrix M identified 166 eigenvalues representing 85% of the variation in the data corresponding to 410 polymorphisms identified from the top 25% coefficients by absolute value in their eigenvectors. Supplementary Table STI lists the mtDNA sequences at all polymorphic loci for all samples used in our study and the haplogroup assignments by NCBI and from the methods described here.

Figure 1 shows the projection of the samples on the first two principal components, which represents 20% of the total variation. The five clades N, L0/L1, L2, L3, and M (Bandelt et al. 2006) are clearly visible. These were separated into clade clusters using consensus hierarchical clustering on the 410 polymorphisms, averaging over 50 bootstrapped datasets and cutting and combining the consensus hierarchical tree at the fifth level. This gave an agreement matrix of size 1737 × 1737 for the fraction of datasets where two samples were in the same cluster. Using simulated annealing (Cerny 1985; Kirkpatrick et al. 1983) this matrix was resorted into a diagonal block form to identify the clade clusters, whose identity was largely confirmed by the sample labels in the database. In the cases where our results differed from the database labels, we found that the labels were incorrect and we updated the database entries.

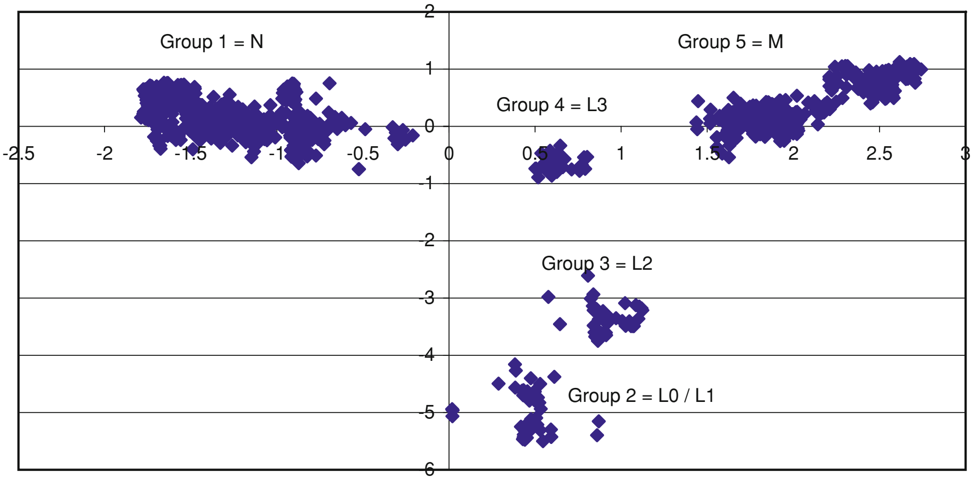


Fig. 1 Results of PCA on all 1737 samples and all 3177 polymorphic loci. The first two principal components account for 20% of the variation in the data and readily split the samples into the five major clades

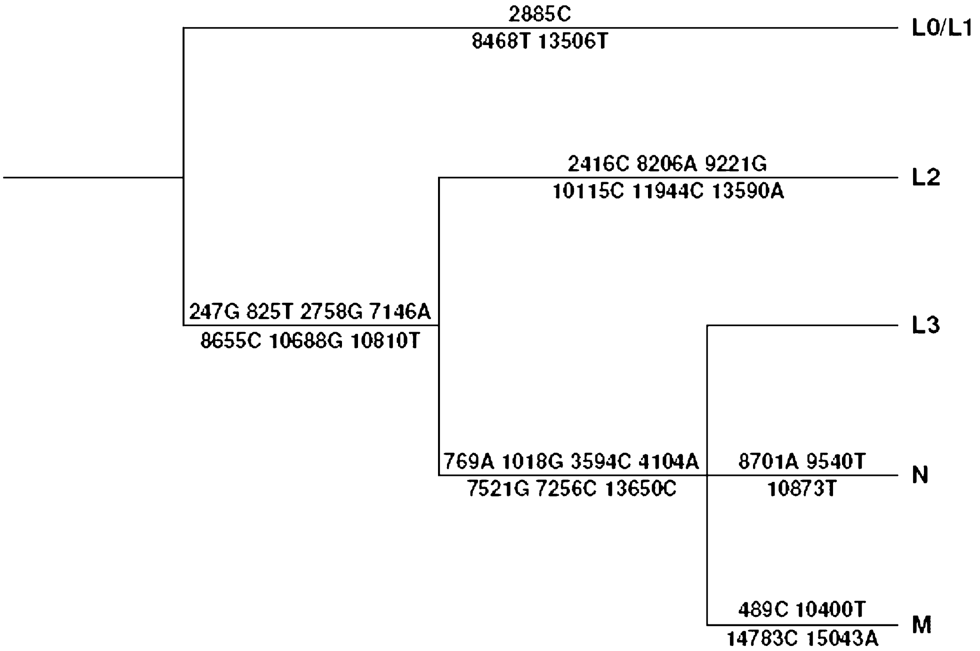


Fig. 2 Consensus clustering tree for the five clades using 30 distinctive mtSNPs with a frequency >90% or a frequency <10% in exactly one clade and the opposite signature in the other clades. Rooting used a consensus chimpanzee as outgroup. The M/N/L3 split is shown as a trifurcation. The time difference between the two migrations cannot be resolved because (a) the mtSNPs distinguishing the L3 clade from L2 and L0/L1 are identically expressed in M and N and (b) the number of robust mtSNPs identified by PCA and ensemble clustering for the M- and N-clade samples are almost equal. Note that the polymorphisms 8701A, 9540T, and 10873T, shown here as defining polymorphisms for the N clade, are wild type with respect to rCRS

A frequency analysis comparing all 3177 loci across the clades seen in PCA identified 34 polymorphisms with a frequency of >95% in one clade and <5% in at least one other clade (see Supplementary Table STII). Sampling equally and repeatedly from each clade, we built a consensus cluster bifurcating tree using these 34 mtSNPs and rooted it with respect to the consensus chimpanzee sequence (D38113 and X93335 from http://www.ncbi.nlm.nih.gov/entrez). This tree is shown in Fig. 2 and is labeled with 30 polymorphisms which have a unique signature for the clades (i.e., either a frequency >90% or a frequency <10% in one of clade and the opposite signature in all other clades). In agreement with the "standard mtDNA tree" ((Bandelt et al. 2006; Ingman et al. 2000); http://www.mitomap.org/mitomap-phylogeny.pdf). Figure 2 confirms that the oldest clade is L0/L1, followed by L2, followed by M/N/L3. Although our results suggest that the M and N clades resulted from two distinct migrations, we are unable to resolve the trifurcation in the M/N/L3 split in Fig. 2. We find that the time interval between these two migrations is too small to be resolved because (a) the markers distinguishing the L3 clade from L2 and L0/L1 are identically expressed in M and N and (b) the number of robust mtSNPs accumulated by the M and N clades since the "out of Africa" event are almost the same (382 vs 373; see below).

PCA for the N Clade Shows B Closer to A Than to T/J/H/V

Repeating the PCA using all 3177 polymorphisms on only the samples in the N clade gives the results shown in Fig. 3. The haplogroup labels shown are the consensus labels of their NCBI assignments. In the pc1-pc2 projection, the leftmost clusters are the T and J haplogroups. The cluster on the extreme right contains a mixture of V and H samples. The cluster in the middle is a mixture of samples from A, B, W, F, I, and X. These four clusters remain well separated in higher PC projections as well. We also verified the stability of these four clusters by repeating the analysis on 10 random 2:1 split training/test datasets created from the N-clade samples. The [A,B] haplogroups always remained in the middle cluster and the distance between [B,J] or [B,T] was consistently and significantly larger than that between [A,B]. This strongly suggests that haplogroups B and A separated from each other more recently than did either B from J or B from T. It also suggests that the so called "R subclade" of the N clade (Bandelt et al. 2006), usually identified by the synonymous polymorphism at locus 12705, is too heterogeneous over the N clade to have originated in a single founder event.

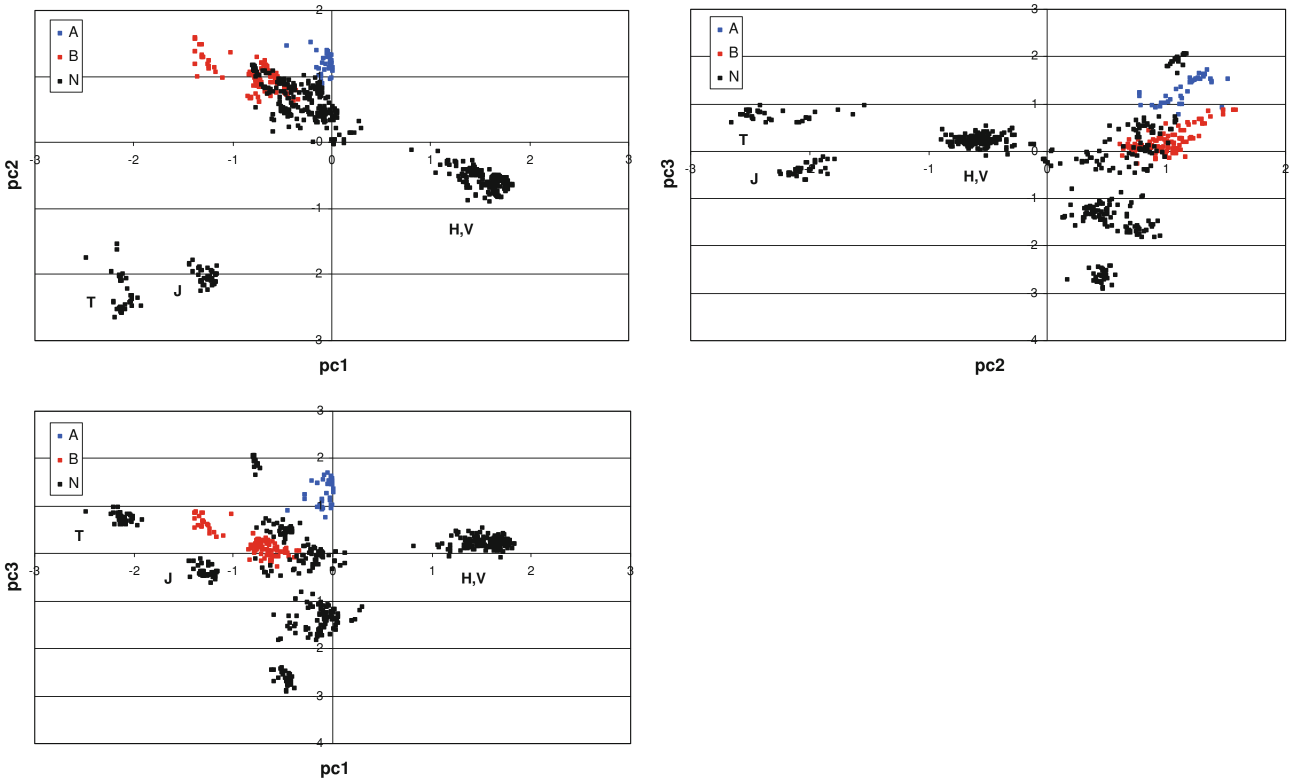


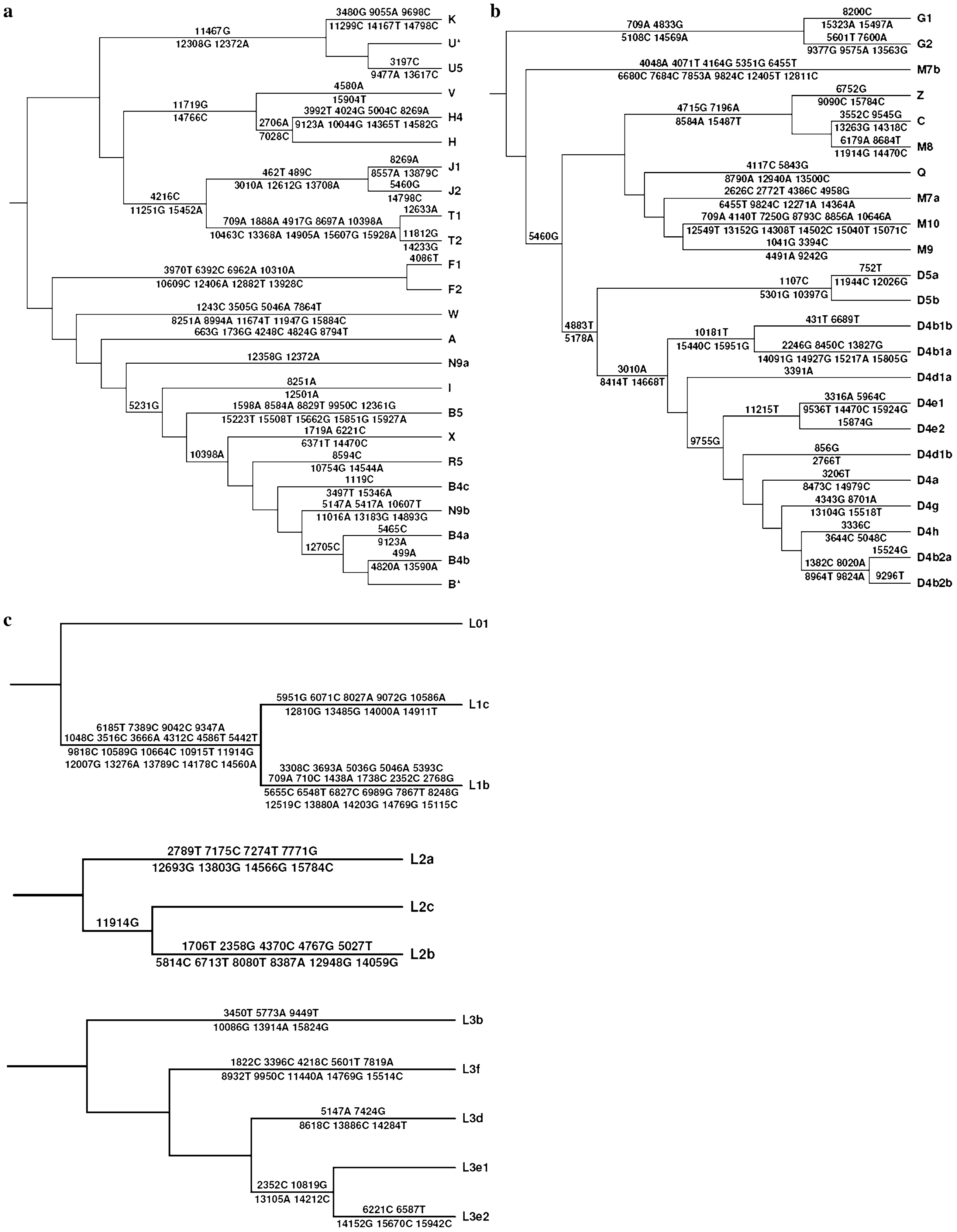
Fig. 3 N-clade PCA projections. The haplogroup labels for the cluster samples are from NCBI. Note that the distance between B and A is much smaller than that between B and any of T/J/H/U/V. This suggests that the B and A split is more recent than the split of B from any of T/J/H/U/V. There is no evidence for the so-called "R subclade," which would include B with T/J/H/U/V. The relative positions of these haplogroups also remain the same in other PCA projections

The N and M Subtrees from PCA and Clustering

Separate PCA of the samples in the N and M clades in Fig. 1 identified 373 and 382 "useful" polymorphisms respectively representing 85% of the variation in each clade. Unsupervised consensus ensemble *k*-clustering and 10-fold training/test validation identified 24 and 23 robust haplogroups in N and M, respectively. Of the 373 polymorphisms in N, 140 had either $\geq$ 90% or $\leq$ 10% frequency in at least one or more clusters in N. Of these, 93 were distinctive, i.e., specific to only one cluster. For M, the corresponding numbers were 141 of 382, with 93 distinctive for the non-D branch and 78 for the D branch. These polymorphisms are listed in Supplementary Table STIII. The near-equality of the numbers of robust polymorphisms for the N and M clades (373/382 and 140/141) suggests that the N- and M-clade migrations "Out of Africa" were almost coincident.

Comparing the cluster samples to their haplogroup labels, we were able to assign haplogroups to 13 previously unlabeled samples in the data. Eighty-five of 1015 samples in N and 122 of 585 samples in M had fluctuating membership in the clusters and were assigned to "Bulk_N*" and "Bulk_M*" clusters, respectively. These samples belong to subhaplogroups that are too poorly sampled to meet our criteria of robustness under bootstrap. Supplementary Table STI gives the list of all sample labels, their source, and their haplogroup assignments in the NCBI listing and using our methods.

We selected equal numbers of samples repeatedly from each of the 24 clusters for N and 23 clusters for M, repeated the *k*-clustering, created the consensus bifurcating network, and found the internal root which was equidistant from the leaves using Poisson statistics on the branch polymorphisms to construct the clade subtrees for N and M. The root for each clade was also verified by ensuring that it was the internal node which was closest to the consensus chimp sequence. Details of the clustering methods used are given in Appendix A. The resulting trees are shown in Figs. 4a and b. In the M clade, we found 13 D and 10 non-D haplogroups. Of these, we identified several subgroups of D4, some of which have been identified previously in the literature (Kong et al. 2006; Tanaka et al. 2004). We label these as D4b1a, D4b1b, D4b2a, D4b2b, D4d1a, and D4d1b, extending their nomenclature in the literature. As a validation exercise, we tested the classification accuracy of the 1015 samples in N and the 585 samples in M using the branch polymorphisms in Fig. 4a and b. We found that, except for samples in BULK_N* and


a
3480G 9055A 9698C
11299C 14167T 14798C
K
11467G
12308G 12372A
U'
3197C
9477A 13617C
U5
4580A
15904T
V
11719G
14766C
3992T 4024G 5004C 8269A
2706A
9123A 10044G 14365T 14582G
H4
7028C
H
8269A
J1
462T 489C
3010A 12612G 13708A
8557A 13879C
5460G
J2
4216C
11251G 15452A
14798C
12633A
T1
709A 1888A 4917G 8697A 10398A
10463C 13368A 14905A 15607G 15928A
11812G
T2
14233G
4086T
F1
3970T 6392C 6962A 10310A
10609C 12406A 12882T 13928C
F2
1243C 3505G 5046A 7864T
W
8251A 8994A 11674T 11947G 15884C
663G 1736G 4248C 4824G 8794T
A
12358G 12372A
N9a
8251A
I
12501A
5231G
1598A 8584A 8829T 9950C 12361G
B5
15223T 15508T 15662G 15851G 15927A
1719A 6221C
X
6371T 14470C
10398A
8594C
R5
10754G 14544A
1119C
B4c
3497T 15346A
5147A 5417A 10607T
N9b
11016A 13183G 14893G
5465C
B4a
12705C
9123A
499A
B4b
4820A 13590A
B'
b
8200C
G1
709A 4833G
5108C 14569A
15323A 15497A
5601T 7600A
G2
9377G 9575A 13563G
4048A 4071T 4164G 5351G 6455T
M7b
6680C 7684C 7853A 9824C 12405T 12811C
6752G
Z
9090C 15784C
4715G 7196A
8584A 15487T
3552C 9545G
C
13263G 14318C
6179A 8684T
M8
11914G 14470C
4117C 5843G
Q
8790A 12940A 13500C
2626C 2772T 4386C 4958G
M7a
6455T 9824C 12271A 14364A
709A 4140T 7250G 8793C 8856A 10646A
M10
12549T 13152G 14308T 14502C 15040T 15071C
5460G
1041G 3394C
M9
4491A 9242G
752T
D5a
1107C
11944C 12026G
5301G 10397G
D5b
431T 6689T
D4b1b
4883T
5178A
10181T
15440C 15951G
2246G 8450C 13827G
D4b1a
14091G 14927G 15217A 15805G
3010A
8414T 14668T
3391A
D4d1a
3316A 5964C
D4e1
11215T
9536T 14470C 15924G
15874G
D4e2
9755G
856G
D4d1b
2766T
3206T
D4a
8473C 14979C
4343G 8701A
D4g
13104G 15518T
3336C
D4h
3644C 5048C
15524G
1382C 8020A
D4b2a
8964T 9824A
9296T
D4b2b
c
L01
5951G 6071C 8027A 9072G 10586A
L1c
12810G 13485G 14000A 14911T
6185T 7389C 9042C 9347A
1048C 3516C 3666A 4312C 4586T 5442T
9818C 10589G 10664C 10915T 11914G
12007G 13276A 13789C 14178C 14560A
3308C 3693A 5036G 5046A 5393C
709A 710C 1438A 1738C 2352C 2768G
L1b
5655C 6548T 6827C 6989G 7867T 8248G
12519C 13880A 14203G 14769G 15115C
2789T 7175C 7274T 7771G
L2a
12693G 13803G 14566G 15784C
L2c
11914G
1706T 2358G 4370C 4767G 5027T
L2b
5814C 6713T 8080T 8387A 12948G 14059G
3450T 5773A 9449T
L3b
10086G 13914A 15824G
1822C 3396C 4218C 5601T 7819A
L3f
8932T 9950C 11440A 14769G 15514C
5147A 7424G
L3d
8618C 13886C 14284T
L3e1
2352C 10819G
13105A 14212C
6221C 6587T
L3e2
14152G 15670C 15942C

◀**Fig. 4** N- and M-clade subtrees from PCA and consensus ensemble clustering. Using equal sampling from each haplogroup, we obtained a bifurcating network by splitting the data into 2,3,…$k_{\max}$ clusters. This network was rooted by (**a**) using Poisson dynamics to find the internal branch equidistant from the leaves and (**b**) finding the internal branch closest to an outgroup sequence (consensus L0/L1). All internal branches shown have 90% consensus accuracy under sample bootstrap. At each split, the mutations labeling the branches are also at least 90% accurate, i.e., they are expressed in more than 90% of samples in one branch and less than 10% of samples in the other. The labeling on branches used an L0/L1 consensus sequence as outgroup. Note that this means that 14766C and 7028C, used here to label V/H/H4 and H/H4, respectively, are wild type with respect to rCRS. (**c**) Consensus clustering subtrees for the L clades L1, L2, and L3. The bootstrap accuracy of every internal branch exceeds 90% accuracy. Only mutations which have one signature (e.g., either frequency >90% or frequency <10%) in one branch and the opposite signature in the other branch are shown. These subtrees were rooted using an L0/L1 consensus sequence as the outgroup

BULK_M*, each sample assigned to a haplogroup exhibited more than 95% of the polymorphisms from the root to the leaf in the tree.

The L Subtree

The L0/L1, L2, and L3 samples analyzed by PCA showed 3, 3, and 5 clusters, respectively, and had 9, 15, and 15 eigenvectors and 235, 167, and 133 mtSNPs, respectively, which represented 85% variation in the data. The L subtrees found in our analysis are shown in Fig. 4c. The African sampling in our dataset is too sparse to allow any additional detail on the L phylogeny, given our stringent criteria for robustness to data perturbation and minimum haplogroup size.

Discussion

We have described a new method to create mtDNA phylogeny using PCA and consensus ensemble clustering but without any additional simplifying assumptions. Our method first identifies all the polymorphisms in the data which distinguish the clusters and then uses them to

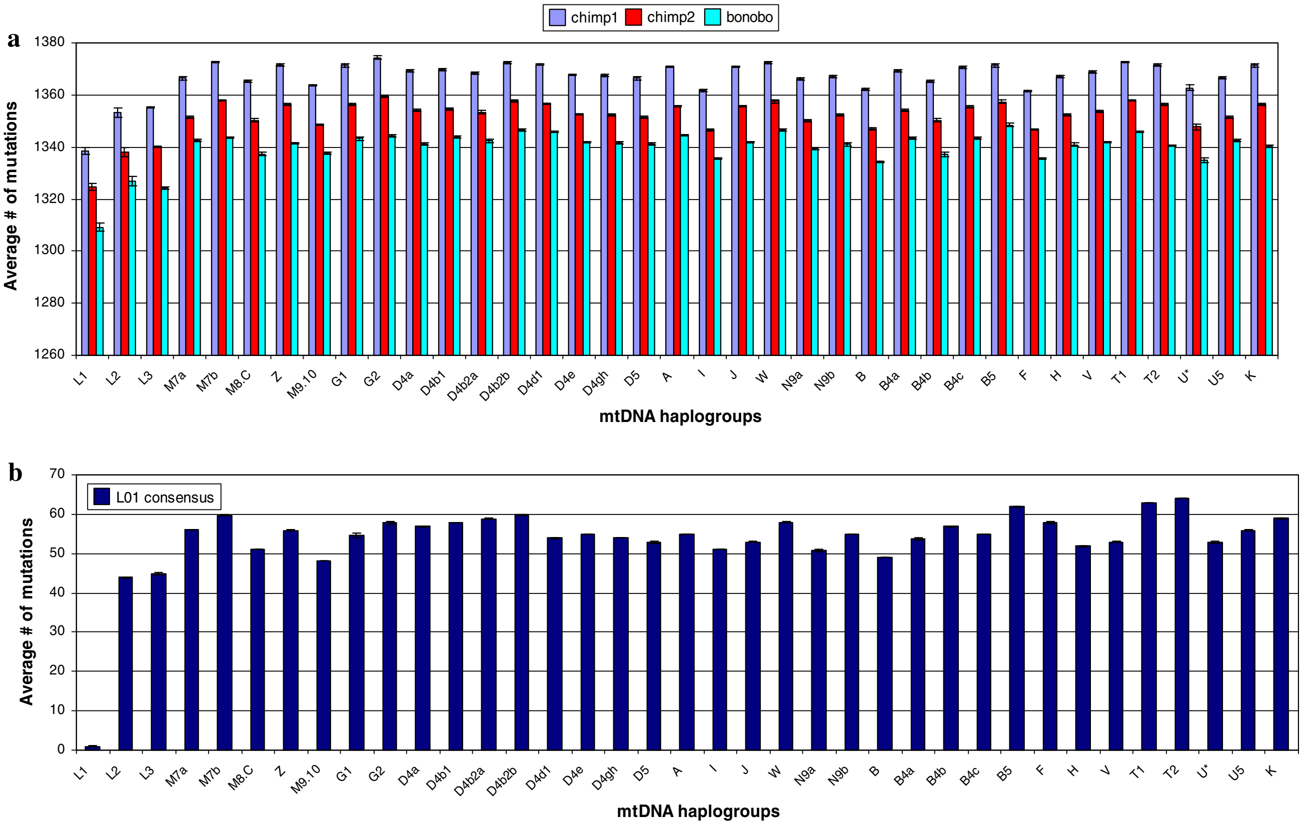
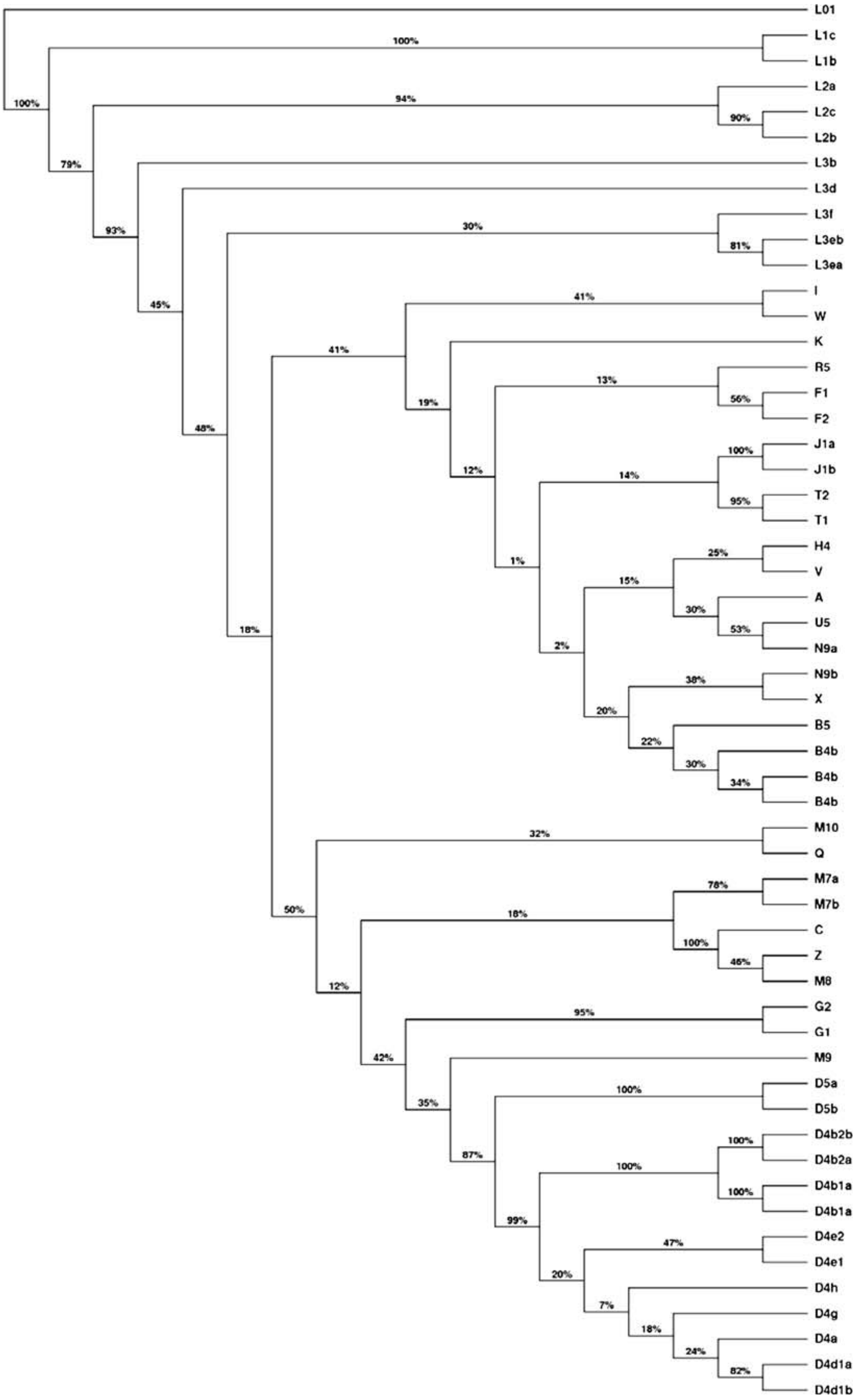


Fig. 5 (**a**) Histogram of the average number of coding-region mtSNPs between human haplogroups and chimp/bonobo. (**b**) Histogram of the number of coding-region mtSNPs between haplogroups and L01 consensus as defined in Supplementary Table STV. Some of the haplogroups were collapsed to internal nodes to ensure at least 15 samples per haplogroup. We retain only those mutations which have ≥ 90% consensus across repeated bootstrap sampling from the haplogroups. Excluding the L1 haplogroups, we estimate that the average number of mtSNPs from the present time to coalescence for the human tree was 27.8 ± 1.9. The numbers of mtSNPs between human-chimp and human-bonobo are 1351.9 ± 6.5 and 1339.81 ± 7.07, respectively. Assuming a chimp-human coalescence time of 5 million years, this gives a TMRCA for the human tree of 206 ± 14 KYBP and a mutation rate in the coding region for human mtDNA of 0.0027 ± 0.0003 mutation/generation, assuming a generation time of 20 years


L01
100%
L1c
L1b
L2a
100%
94%
L2c
90%
L2b
L3b
79%
L3d
L3f
30%
L3eb
93%
81%
L3ea
I
41%
45%
W
K
41%
R5
13%
F1
56%
19%
F2
48%
J1a
100%
14%
J1b
12%
T2
95%
T1
H4
25%
1%
V
15%
A
30%
U5
53%
18%
N9a
2%
N9b
38%
X
20%
B5
22%
B4b
30%
B4b
34%
B4b
M10
32%
Q
M7a
78%
M7b
50%
18%
C
100%
Z
46%
M8
12%
G2
95%
G1
M9
42%
D5a
100%
35%
D5b
D4b2b
100%
87%
D4b2a
100%
D4b1a
100%
D4b1a
D4e2
99%
47%
D4e1
20%
D4h
7%
D4g
18%
D4a
24%
D4d1a
82%
D4d1b

◀ **Fig. 6** The consensus maximum likelihood tree obtained using 869 mtSNPs from the union of all polymorphisms identified by PCA and clustering for the clades and M, N, and L haplogroups. The tree shown is the consensus over trees from 100 datasets, each of which was created by selecting one sample randomly from each of the 55 haplogroups shown. The branch labels on the consensus maximum likelihood tree are a measure of the reliability of the branch. This is estimated as the fraction of cases when the branch splits the downstream haplogroups into the sets shown in the tree over the sampled datasets. The best consensus is shown on all branches. Ancient branches, corresponding to the clade tree in Fig. 2, are reliably reproduced, as are some recent branches. The middle branches have a lower reliability, with branch accuracies of <10% in many cases. This makes the overall reliability of the maximum likelihood tree very low, as it is highly sensitive to sample bootstrap

robustly classify the samples into haplogroup clusters which are stable under perturbation in both polymorphisms and sample bootstrap. Reclustering into $k = 1,2,\ldots,k_{\max}$ clusters with equal sampling from each haplogroup directly yields the tree. We root the tree both with respect to an outgroup as well as using Poisson statistics on the branch polymorphisms to find the internal node which is equidistant from the leaves. We compare these two methods and give additional details about them in Appendix C. Although both these methods give the same root for the mtDNA tree, we prefer the method using Poisson statistics because it finds the best root over all possible evolutionary scenarios, which gives some confidence on the reliability of the identified root.

Our method includes *all* polymorphisms in the data, is fast, is robust to data and sample perturbations, and is based on mild assumptions required for the clustering analysis. The underlying assumption in our method is that sequences that share a specific mutation are more related than those that do not (this is the basis for clustering) and that the more mutations they share, the more closely related they are. Whereas this does impose a restriction on the phylogeny, it is a mild assumption (than say minimizing the total number of mutations) and it does not restrict the possibility of ancient mutations involved in homoplasy events.

A striking observation we made was that the haplogroup clusters emerged naturally as a bifurcating network. Two clusters at each k were derived from a split of one of the clusters at the previous (smaller) k, while the other clusters remained the same. The tree we infer is consistent with a recent African origin of modern humans and the almost-coincident emergence of the M and N clades out of Africa. A major result is that our N clade tree does not have the standard “R clade.” Instead, it has a “European R subclade” with haplogroups T/J/U/V/H/K and a “Eurasian N subclade” with the Asian/Eurasian haplogroups B, F, and R5 (usually placed in the “R clade” (Bandelt et al. 2006)) in proximity to A/B/W/I/X. The clear division between “European R subclade” groups and “Eurasian N subclade” groups suggests that the N clade may have split into distinct West/East migrating subgroups after emerging from Africa and settled in their current geographic locations with relatively little East/West mixing.

Supplementary Table STIV gives the frequency of the allelic forms of all polymorphisms in every haplogroup identified by our analysis as well as the allelic form for rCRS, Chimp1, Chimp2, Bonobo, and a “Consensus L01” sequence (which was obtained from the samples closest to the root of the clade tree which came from haplogroups L0 and L1a). To compute a time to coalescence of the human tree, we collapsed smaller haplogroups in the tree to nodes containing at least 15 samples. We then compared the samples in each haplogroup to the consensus L01 sequence, to one another, and to the chimp sequences using bootstrap sampling and identified all polymorphisms which were identified in more than 90% of the bootstrap samplings. Figure 5a shows the number of coding-region mtSNPs identified for each haplogroup relative to chimp/bonobo, from which we estimate the number of robust coding region mutations between human-chimp and human-bonobo to be 1351.9 ± 6.5 and 1339.81 ± 7.07, respectively. Excluding the L1 haplogroups, we estimate that the average number of mtSNPs from the present time to coalescence for the human tree is 27.8 ± 1.9. Assuming a chimp-human coalescence time of 5 million years and a generation time of 20 years, this gives a TMRCA for the human tree of 206 ± 14 KYBP and a mutation rate in the coding region for human mtDNA of 0.0027 ± 0.0003 mutation per generation. Our list of robust mtSNPs, supported by 90% bootstrapping, selects markers and TMRCA for the clades to be somewhat older but generally consistent with previous estimates (Ingman et al. 2000; Jobling et al. 2004; Parsons and Heflich 1998).

Our tree is in agreement with the so-called “standard” tree ((Bandelt et al. 2006); http://www.mitomap.org/mitomap-phylogeny.pdf) at the clade level but disagrees at the subclade level. To understand the source of this discrepancy, we built consensus trees for the five clades as well as for each of the subclades using four different tree building methods: maximum parsimony, maximum likelihood, neighbor joining (NJ), and UPGMA as implemented in the software package Phylip of J. Felsenstein (http://evolution.genetics.washington.edu/phylip.html; see Appendix B for details).

To build the consensus clade trees, we created 100 datasets, each consisting of one sample randomly chosen from each clade. Using the same polymorphisms used for Fig. 2, we obtained 100 trees for each phylogeny method and combined them into a single consensus tree. We found that all the consensus clade trees had the same topology (which agreed with Fig. 2) and were also unable to resolve the trifurcation in the M/N/L3 split.

Complete consensus trees for all haplogroups for each of the four methods were built in a similar way, using the union of all polymorphisms identified by PCA for the clades, M, N, and L (see Appendix B for details). We found that the accuracy of the internal branches of these trees was highly variable, with many internal branch accuracies in the 1–30% range. As an example of this, we show the consensus maximum likelihood tree in Fig. 6. This analysis shows that the standard techniques used to infer phylogeny are very sensitive to data perturbation and unreliable except at the level of the clade tree. Although some haplogroup cluster topologies remain consistently the same for all methods (mostly near the leaves of the consensus trees), there is a lot of variability in the internal branches across the methods. Even for a single method, the accuracy of many internal branches is quite low. This should be contrasted with our tree (Figs. 2 and 4a–c), where every branch split is constructed to be at least 90% accurate.

The "R clade" is usually defined by a synonymous mutation at locus 12705. In our analysis, this polymorphism is only one among many old (and homoplasic) polymorphisms. We find instead (Fig. 3) that the haplogroups in the so-called "R clade" are too widely separated, and some are too close to non-R-clade haplogroups, to have arisen from a single founder event defined by this polymorphism. Possible reasons for the identification of a homogeneous "R clade" in earlier analysis are (a) a sampling bias toward specific European and Asian haplogroups J, T, H, V, B, and F, all of which have the mutation 12705C → T, (b) tree construction using human-aided identification of polymorphisms, and (c) a strong focus on minimizing homoplasy events using methods (e.g., parsimony) which are less sensitive to the global structure of polymorphisms in the data.

There are many other homoplasy mutations (like 12705) in the data. For instance, there is a mutation at locus 5417 which appears in both N9a and N9b but is not present in any other N haplogroups (Tanaka et al. 2004; Kong et al. 2006). Parsimony methods identify this mutation as a founder mutation and, consequently, place N9a and N9b on adjacent leaves of the N-clade subtree (Bandelt et al. 2006). Indeed the names of these groups are themselves derived from such a placement. However, from a study of the Jomon and Yayoi people in Japan (Shinoda 2005), the N9a haplogroup is believed to have entered Japan from South China through Korea along an eastern route, while the N9b haplogroup came into Japan via a northern route. According to this analysis, these two haplogroups diverged between 15,000 and 20,000 years ago and so should not be adjacent in the tree. Our tree agrees with this analysis and places these two groups relatively far apart based on many global and robust differences in their sequences.

Another example of homoplasy is in the M clade subtree, where polymorphisms at loci 6455 and 9824 are found in haplogroups M7a and M7b. Once again, parsimony methods place these haplogroups in close proximity despite the fact that M7b is found mostly in Korea and came there from an Eastern migration from China, while M7a is most frequent in Ryukyuans on the Okinawa islands of Japan (Tanaka et al. 2004) and is believed to have entered Japan via a southern route. It is believed that these two haplogroups diverged from a common ancestor more than 20,000 years ago (Tanaka and Ozawa 1994) and have been geographically isolated for a long time. Consequently they should not be close together in the M subtree, in agreement with our placements of these groups (Fig. 4b). The overall analysis of homoplasy events in our mtDNA tree, comparisons with other trees, and their relation to selection and population geography/history will be addressed in a separate publication.

Because of our use of complete mtDNA sequences and a data perturbation-independent protocol, our clustering techniques provide a "gold standard" for assignment of samples to haplogroups. They should be useful for the identification of new haplogroups. In Supplementary Table STV, we list the allelic state of the polymorphisms in Supplementary Table STIV as used to label the trees and assign haplogroups. The allelic forms shown for each haplogroup are expressed in more than 90% of the samples in each haplogroup where they are shown. Supplementary Table STVI gives the characteristic mtSNPs to assign samples to haplogroups. For completeness, we give the frequency at each of the 3177 polymorphic loci in each haplogroup in Supplementary Table STVII and the aligned sequences for rCRS and Chimp 1, Chimp 2, and Bonobo which were used in this paper in Supplementary Table STVIII. The L01 reference sequence is defined at all polymorphic loci relative to rCRS in Supplementary Table STVII. At all loci other than these, L01 is identical to rCRS. Finally, Supplementary Fig. SFI gives the full labeled tree using consensus L01 as outgroup.

Our method identified a phylogeny for the N and M clades that better reflects the current geographic location and known migratory history of some of their haplogroups, which are usually grouped in unlikely ways by methods such as parsimony. Supplementary Fig. SFII compares the migration paths and current location of the N-clade haplogroups for a parsimony-based tree (which has an R clade) with those for the tree derived in this paper. The first figure in Supplementary Fig. SFII shows that splitting into R and non-R groups requires a back-migration event to explain the presence of R5/F/B haplogroups in Asia. The second figure shows haplogroups as they are labeled by our tree, which has no R clade and splits the N clade cleanly into Eastern/Western haplogroups, eliminating the need for a back-migration event to explain the presence of R5/F/B in Asia.

Instead of building the tree from a few polymorphisms using methods with strong assumptions, our method identifies and uses all informative polymorphisms and builds the tree from the natural order of bifurcations inherent in the data. Finally, instead of the poor bootstrap support on internal branches which is a common feature of methods such as parsimony and maximum likelihood, our method is robust and has high internal branch bootstrap support. The flexibility and robustness of our method make it suitable for the analysis of datasets with mixed distributions, high variability, unknown sample size bias, and unknown haplogroup structure.

Since the method described in the paper is quite complex, we have developed a software suite in Python with C extensions, described in Appendix D, which reproduces the results in Figs. 1 and 2 and Supplementary Table STII. The source code can be downloaded from https://biomaps.rutgers.edu/wiki/upload/9/93/MtDNA_utility.tar.gz.

This software starts with the 1737 aligned sequences used in the paper, creates the mutation matrix, performs PCA, identifies the subset of mutations that represent 85% of the variation, identifies the clade clusters based on PCA, and, using equal numbers of samples from each clade, uses consensus clustering based on k-means to divide the data into $k = 2, 3, 4, 5, 6$ groups, thereby reproducing the network representing the clade tree in Fig. 2. The 410 mtSNPs used in this analysis are listed in Supplementary Table STIX.

Finally, as a validation exercise for our method, we performed a simulation which mimics the observed origin and migration of the M and N clades from a parent population in Africa. We started with 10,000 individuals, broken up into mtDNA groups of 20–50 individuals related by descent from a common ancestor within the previous 200 generations. Assuming a constant mutation rate across the coding region, we implemented a neutral evolutionary model with a fixed population size (implemented by choosing each individual to have a Poisson-distributed number of offspring with Poisson parameter unity) for 5000 generations ($\sim$100,000 years, assuming a generation time of 20 years). At this point, we simulated two simultaneous migration events of 1000 individuals each (corresponding to the M and N clades), which then expanded into two groups of 10,000 individuals by a 5% increase in birth rate before their respective populations stabilized. The three groups (one original plus two migrant) then evolved under neutral dynamics with approximately stable populations for an additional 3000 generations ($\sim$60,000 years) to the present time.

When we looked at the resulting polymorphisms in the data, we found that they separated into two distinct sets. One set contained old polymorphisms which were fixed by drift and found in large groups of individuals. The other set, which was much larger, contained polymorphisms not yet fixed by drift which were sites of a large number of homoplasy events. We found that our method, because of the bootstrap analysis, was able to distinguish these two sets of polymorphisms and could construct the correct tree with good internal consensus. However, other methods (such as parsimony) often confused recent and ancient homoplasic polymorphisms. An interesting observation was that most of the contribution to the homoplasy events was from the last 1000–1500 generations, which corresponds to about 20,000–30,000 years. This coincides nicely with the East-West split in the N clade and suggests that mtDNA phylogeny using standard methods such as parsimony and maximum likelihood, because of their sensitivity to homoplasy events, gives accurate trees only at the level of the clade tree. This observation also agrees with our reported analysis of real mtDNA sequences.

Acknowledgments G.B. and M.T. thank Dr. K. Shinoda for insight into the Eastern migrations of the N and M haplogroups and Dr. Cabrera for a critical reading of an early version of the manuscript. G.A. and G.B. acknowledge discussions with many colleagues at IBM Research, where this study was initiated in 2005, and at the Aspen Center for Physics in 2007, where it was concluded.

Appendix A: Method for Finding Robust Clusters Using Consensus Ensemble Clustering

Unsupervised clustering algorithms divide data into meaningful groups or clusters such that the intracluster similarity is maximized and the intercluster similarity is minimized [A1]. Clustering is an NP-hard problem. However, many heuristic methods exist and they can be categorized into hierarchical, partitioning, and grid-based methods. We apply all these methods in an unsupervised way to the data, i.e., without assuming a predefined label on the objects to be classified. Unsupervised clustering is known to produce unstable solutions which are sensitive to various data parameters and/or perturbations and to the clustering techniques used [A2]. A relatively recent solution which corrects for this instability is consensus ensemble clustering [A2, A3]: Given several methods of clustering data, find a combination of these methods which is of better quality.

This problem can be broken up into two parts: a method that generates a collection of clustering solutions and a consensus function that combines them to produce a single output clustering of the data. There is an implicit assumption in this that combining the results of several clustering techniques will give groupings that are more reliable and less biased to a particular technique. This has been demonstrated in supervised classification schemes where it was shown that multiple solutions may reduce the variance of the error and, at the same time, increase the robustness of the result [A4, A5]. The ensemble clustering technique was introduced in [2], and the effects of consensus clustering were described in several subsequent studies [A4, A6, A7].

In our study the challenge posed to the ensemble consensus clustering approach was to identify meaningful clusters which were stable and robust both to perturbations of the data and to the choice of clustering methods used. This goal was approached in two ways.

1. If the method was stochastic, we reduced the effect of the stochastic variation by applying the method repeatedly and taking an appropriate average.
2. To reduce the sensitivity to random variation in the data, we applied each clustering method to multiple sample datasets obtained by bootstrapping in both the mutations used in the clustering and in the samples.

Clustering High-Dimensional Data

To correct for the fact that many mutations are only on individual samples and merely add noise to the data and the fact that many mutations travel together, we cluster on subspaces of attributes rather than on the entire space. The subset of attributes (mtSNPs) on which data are clustered may have an important influence on the clustering solution. Since mtDNA data are high dimensional, we restricted the clustering procedures to those attributes which were determined to be "discriminatory" through an initial principal component analysis (PCA).

The details of our clustering method are as follows.

Step 1

For each $k = 2,\ldots,50$, we created k clusters on resampled and random projected datasets based on individual clustering methods. We generated 150 datasets as follows: 50 datasets were created by bootstrapping the samples, 50 datasets by projecting the data onto subsets of mtSNPS bootstrapped from the data, and 50 additional datasets by first projecting the data on a bootstrapped subset of mtSNPS and then bootstrapping samples on the resulting dataset. We then applied representative methods for each major class of known clustering techniques. We discuss these briefly below.

Partitioning Relocation Methods

These methods divide data into several subsets and use certain greedy heuristics in the form of iterative optimization to reassign points between the k clusters. The optimization is applied to an objective function defined on unique cluster representatives (e.g., centroid, medoid), which is usually a dissimilarity measure.

We applied the following algorithms.

i. Partition around medoids (PAM) [A8]: PAM is an iterative optimization that relocates the points between perspective clusters by renominating the points as potential medoids.
ii. CLARA [A8]: This method uses several samples of the data and subjects each of them to PAM. The dataset is then reassigned to the resulting medoids and the best system of medoids is retained.
iii. K-means [A9]: To each cluster, this method associates the mean (centroid) of its points and uses as the objective function the sum of distances between a point and its centroid.
iv. Graph partitioning [A10]: In this method the points (samples) are associated with vertices in a graph and each point is connected to the closest neighbor. The resulting graph is then split into k-clusters by applying a min-cut approach.

Clusters produced by centroid methods (k-means, PAM, CLARA) work by identifying samples into clusters if they form a spheroid shape. Thus, they are suitable for clustering datasets with uniform and relatively low variation among samples. Graph partitioning methods produce clusters in which samples are added in if they are "close" to at least one sample in the candidate cluster. Thus graph partitioning approaches can successfully identify clusters with unequal variance along the feature coordinates (i.e., they can find a "long" shape).

Agglomerative Methods

These methods build the clusters gradually by trying to establish a hierarchical order [A1]. One starts by assigning each sample to its own cluster and then recursively merging two or more most similar clusters until a stopping criterion is fulfilled. The similarity between clusters is usually computed based on a linkage metric which reflects the connectivity and similarity between the clusters. In our study we applied hierarchical clustering techniques based on the following metrics.

Average linkage metric: Computes the distance between two clusters as the average of the distances between the pairs of points in these clusters.

Complete linkage metric: Computes the distance between two clusters as the maximum distance between the pairs of points in the two clusters.

Single linkage metric: Computes the distance between two clusters as the minimum distance between the pairs of points in the two clusters.

McQuitty metric: Computes the distance between two clusters as the average distance between the subclusters of the two clusters.

Centroid metric: Computes the distance between two clusters as the distance between the centroids of the two clusters.

Ward metric: Computes the distance between two clusters as the distance between the centroids of the two clusters averaged to the reciprocal mean of the sizes of the two clusters.

In addition, we applied a *hybrid-biased agglomerative method* (*bagglo*) which combines partitioning clustering with the agglomerative hierarchical approach [A11]. For n samples, we start with an initial partition into $n^{1/2}$ clusters and augment the original feature space by adding $n^{1/2}$ dimensions corresponding to the initial clusters. The agglomerative clustering approach is then applied to this augmented dataset.

Methods Based on Probability

In these methods, data are considered to be a sample independently drawn from a mixture model of several probability distributions and the clusters are associated with the area around the mean of each distribution. It is assumed that each point is assigned to a unique cluster. The probabilistic clustering method optimizes the log-likelihood of the data to be drawn from a given mixture model.

In our approach we applied the *expectation maximization* (EM) method [A12, A13]. EM is a two-step procedure which starts with estimating for each point the probability of belonging to a certain cluster. In the second step EM finds an approximation to the mixture model by maximizing the log-likelihood in an iterative way until the convergence to an optimal solution is reached.

Entropy-Based-Clustering (ENCLUST)

This method [A14] starts by dividing the interval associated with each attribute into one-dimensional bins (cells) and retaining only the cells with a high density. The iterative step consists in creating cells of higher dimensions by joining the cells with low dimension and retaining only those cells which have the entropy below a certain threshold as optimal for clustering.

Clustering on Subsets of Attributes (COSA)

This method [A15] uses an idea similar to that in ENCLUST by preferentially clustering on subsets of attributes with low variability across the samples in the clusters.

Self-Organizing Maps (SOM)

This method [A16] is both a data visualization and a clustering technique which reduces the dimensions of data through the use of self-organizing neural networks. The way SOM reduces dimensions is by producing a map of usually one or two dimensions which plot the similarities of the data by grouping similar data points together.

Step 2

Each method was applied 50 times with different parameter initialization on the full dataset and once on each of the 150 datasets obtained as described in Step 1. Based on the 200 clustering results, we constructed an agreement matrix for each method whose entries m_{ij} represent the fraction of times the pair of samples (i, j) occurred in the same cluster of the total number of times the pair of samples was selected in the 200 datasets.

Using $d_{ij} = 1 - m_{ij}$ as the distance between the samples (i, j), we apply simulated annealing [A17] to find k “consensus” clusters which achieve the maximum value for the average internal similarity and the average external dissimilarity. The cost function used for the simulated annealing is given below.

$$\frac{\sum_{l} \sqrt{\sum_{i,j\in C_l} d_{ij}}}{\sum_{l} n_l \frac{\sum_{i\in C_l, j\in S} d_{ij}}{\sqrt{\sum_{i,j\in C_l} d_{ij}}}} \tag{1}$$

where d_{ij} represents the distance between the samples (i,j), $C_1,\ldots,C_k$ are the k clusters to be determined, n_l is the size of cluster C_l, $l = 1,.., k$, and S is the set of samples.

At the end of this step, each method will give us its best clustering into k clusters.

Step 3

For each k, we combine the results from Step 2 by using an agreement matrix to create a consensus of all the individual clustering techniques and, once again, use simulated annealing to optimize the clustering.

In comparison with the traditional methods which use a single clustering technique, the consensus ensemble clustering approach, in combination with PCA, has better average performance across datasets and a lower sensitivity to noise, outliers, and sampling variation.

References A1

A1. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

A2. Strehl A, Ghosh J (2002) Cluster ensembles: a knowledge reuse framework for combining partitionings. In: Eighteenth National Conference on Artificial Intelligence. Edmonton, Alberta, Canada, July 28–August 1, 2002, pp 93–98

A3. Monti S, Tamayo P, Mesirov PJ, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learn J 52(1–2):91–118

A4. Alexe G, Alexe S, Crama Y, Foldes S, Hammer PL, Simeone B (2004) Consensus algorithms for the generation of all maximal bicliques. Discrete Appl Math 145(1):11–21

A5. Prodromidis AL, Stolfo SJ (1999) A comparative evaluation of meta-learning strategies over large and distributed data sets. Workshop on Meta-learning. Sixteenth International Conference on Machine Learning

A6. Topchy A, Jain AK, Punch W (2005) Clustering ensembles: models of consensus and weak partitions. IEEE Trans Pattern Anal Machine Intel 27:1866–881

A7. Topchy A, Minaei-Bidgoli B, Jain AK, Punch WF (2004) Adaptive clustering ensembles. In: 17th International Conference on Pattern Recognition (ICPR'04): 2004, pp 272–275

A8. Kaufmann L, Rousserw PJ (1990) Finding groups in data: an introduction to cluster analysis, 1st edn. Wiley, New York

A9. Hartigan JA (1975) Clustering algorithms. Wiley, New York

A10. Karypis G, Kumar V (1995) Multilevel graph partitioning schemes. In: Proceedings, 24th International Conference on Parallel Processing. CRC Press, New York, pp 113–122

A11. Rasmussen M, Karypis G (2004) gCLUTO: an interactive clustering, visualization, and analysis system. UMN-CS 2004. TR-04-021

A12. Dempster AP, Laird NM, Rubin DB (1977) R: maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B 39:1–38

A13. Fraley B, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631

A14. Cheng CH, Fu AW, Zhang Y (1999) Entropy-based subspace clustering for mining numerical data. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD-99: 1999. San Diego, CA

A15. Friedman JH, Meulmany JJ (2004) Clustering objects on subsets of attributes. J Roy Stat Soc Ser B 66:1–25

A16. Kohonen T (2001) Self-organizing maps, vol 30. Springer, New York

A17. Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. Science 220(4598): 671–680

Appendix B: Building Consensus Trees Using Maximum Parsimony, Maximum Likelihood, Neighbor Joining, and UPGMA

Preprocessing

There were a total of 1737 complete mtDNA sequences, classified into 61 haplogroups. We removed samples classified into the six bulk clusters: B*, BULK_D*, BULK_M*, BULK_N*, H*, and U*. The number of samples was reduced to 1222 and the number of polymorphisms was reduced to 2321.

Building the Consensus Clade Trees

We built one consensus tree for the five clades (L0/L1, L2, L3, M, and N) for each of the four methods: maximum parsimony (MP), maximum likelihood (ML), neighbor joining (NJ), and unweighted pair group method with arithmetic mean (UPGMA).

We created 100 data sets, each consisting of one sample picked randomly from each clade. The polymorphisms used were the same 34 mtSNPs identified by the cluster analysis on the clades (Supplementary Table STII).

Procedure

We used the software package Phylip [B.1] to construct the trees. Table 1 lists the software used in Phylip. For each of the four methods, one tree was created for each of the datasets. A consensus tree (see Appendix B Fig. 7a–d) was

Table 1 The four Phylip programs used in this analysis

Tree building method	Phylip program
Maximum parsimony	Dnapars
Maximum likelihood	Dnaml
Neighbor joining	Neighbor
UPGMA	neighbor (toggle option N to choose UPGMA method)

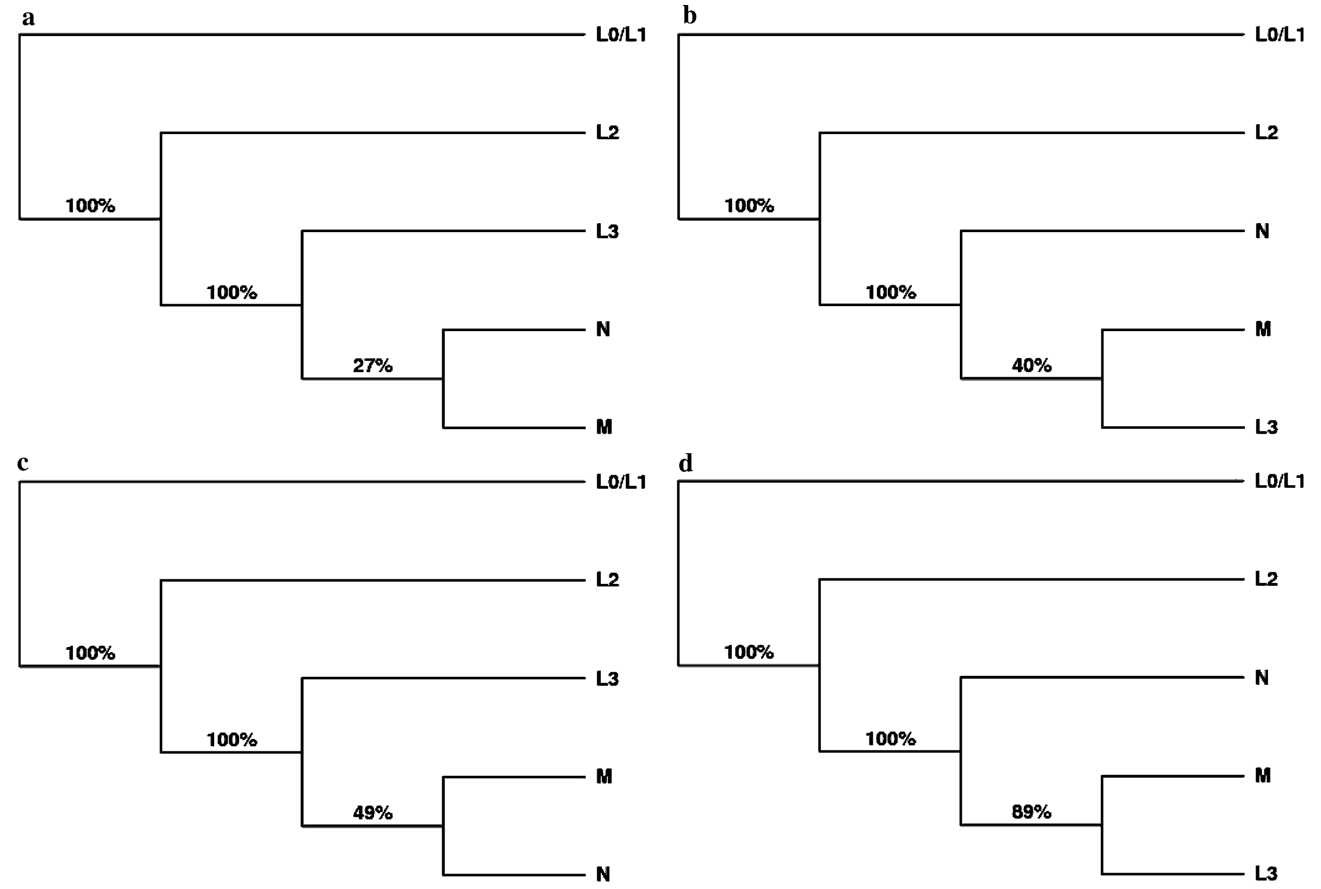


Fig. 7 (**a**) The reliable (high bootstrap agreement) and unreliable (low bootstrap agreement) parts of the MP clade tree are estimated from the consensus fraction on branches across bootstrap replicates in this consensus tree. The tree was constructed using 100 bootstrap datasets, each with one sample from each clade using the 34 mtSNPs listed in Supplementary Table STII and building the consensus tree from them. The older branches have 100% consensus across bootstrap replicates and may be considered robust and reliable. However, the sequence in which the L3/M/N clades split is not determined to any reliable accuracy. In 67% of the bootstrap results, the MP algorithm reported this split as a trifurcation, (**b**) The reliable (high bootstrap agreement) and unreliable (low bootstrap agreement) parts of the ML clade tree are estimated from the consensus fraction on branches across bootstrap replicates in this consensus tree. Once again, the oldest branches are reliable, while the L3/M/N clade split is not. (**c**) The reliable (high bootstrap agreement) and unreliable (low bootstrap agreement) parts of the NJ clade tree are estimated from the consensus fraction on branches across bootstrap replicates in this consensus tree. Once again, the oldest branches are reliable, while the L3/M/N clade split is not. (**d**) The reliable (high bootstrap agreement) and unreliable (low bootstrap agreement) parts of the UPGMA clade tree are estimated from the consensus fraction on branches across bootstrap replicates in this consensus tree. Although the oldest branches are more accurate than the L3/M/N split, UPGMA is the only one of the four methods used here which reliably suggests that the N migration preceded the M migration. However, as the other methods do not resolve this split in a reliable way, we can only conclude that the L3/M/N should be shown as a trifurcation

obtained for each method using the consensus program in Phylip by combining the 100 trees. The branches were labeled with the percentage of cases when the split appeared as shown in the figures. The algorithm automatically chooses the branch percentage which is maximum.

Discussion

While the accuracy of the L0/L1 and L2 branches was always 100%, the percentage of times the split in the M/N/L3 clades appeared over all possibilities is reported in Table 2.

Table 2 Frequencies for the four ways of resolving the M/N/L3 trifurcation for each method: MP, ML, NJ, and UPGMA

	(L3,(M,N)) (%)	((L3,M),N) (%)	((L3,N),M) (%)	(L3,M,N) (%)
Maximum parsimony	27	2	4	67
Maximum likelihood	29	40	31	0
Neighbor joining	49	46	5	0
UPGMA	0	89	11	0

Building Complete Consensus Trees

We built a consensus tree for all 55 haplogroups with each of the four phylogenetic tree building methods: MP, ML, NJ, and UPGMA.

We created 100 datasets by selecting one sample from each of the 55 haplogroups that remained after eliminating the six bulk clusters: B*, BULK_D*, BULK_M*, BULK_N*, H*, and U*. There were 869 polymorphisms left from the combined set of polymorphisms identified by the global PCA as well as the PCA for N, M, L. All these were used to build 100 trees for each of MP, ML, NJ, and UPGMA. From these, consensus trees were obtained for each of the methods. They are shown in Fig. 8a–d. The MP algorithm generated many trees with the same optimum weight for the same dataset. All trees with the same weight were first combined for each dataset before they were combined across datasets. We used the tree plotting software TreeGraph [B.2] to draw the trees.

References B

B.1. Felsenstein J (2004) PHYLIP: Phylogeny Inference Package. ed 3.6. University of Washington, Seattle

B.2. Müller J, Müller K (2004) TreeGraph: automated drawing of complex tree figures using an extensible tree description format. Mol Ecol Notes 4:786–788

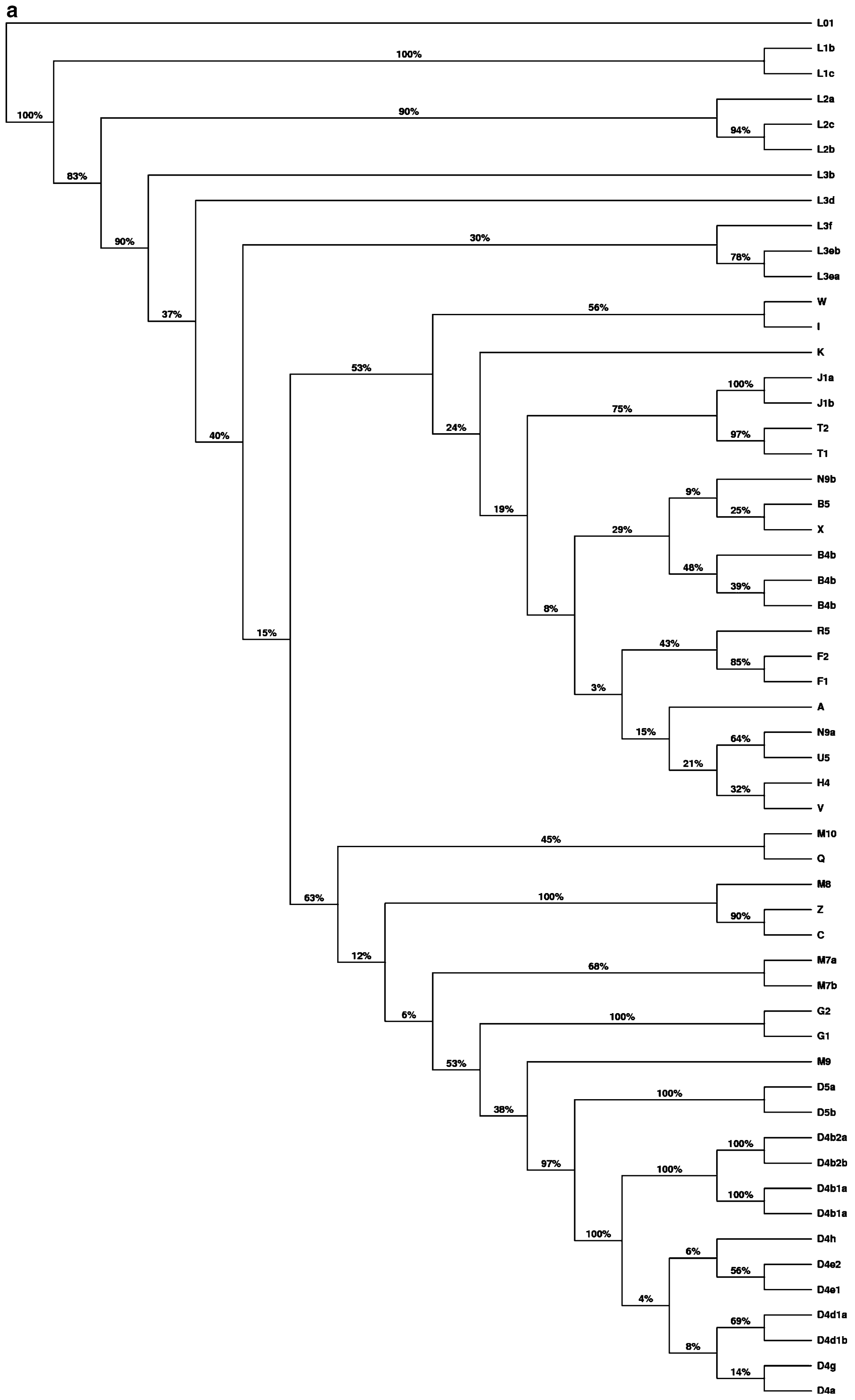
Appendix C: Methods to Select the Root of the Network

In this paper, we have used two methods to root the networks identified by our procedure. The first method used the mtDNA sequence of an “outgroup” species (such as chimpanzee or bonobo). The root was identified as the internal node in the network, which minimizes the number of loci at which the mtDNA sequence at the internal node was different from the outgroup sequence for a robust subset of mtSNPs. The second defined the root as the internal node that was equidistant from the leaves across all possible trees, assuming that the number of polymorphisms on the tree was one instantiation of a Poisson process on the internal branches.

Method I: Rooting with Respect to an Outgroup

The first method is standard and depends only on the availability of an appropriate outgroup sequence. In our case, we used the consensus of the two chimpanzee and one bonobo sequences in Supplementary Table STI and limited the analysis to the 435 robust mtSNPs used in labeling the tree (this list of mtSNPs is given in Supplementary Table STIV). For illustration, we consider three internal nodes, R1, R2, and R3. In Fig. 2, R1 is the split between the L0/L1 superclade and rest of the tree, R2 is the node defining the split between the L2 superclade and the rest of the tree, and R3 is the node separating the L0/L1/L2 subtree from the rest of the tree. We find that the number of mtSNPs that are different between these internal nodes and the chimp-bonobo consensus sequence is: $D_{R1} = 111$, $D_{R2} = 118$, and $D_{R3} = 121$. This identifies R1 as the root of the tree.

Fig. 8 (**a**) The consensus MP tree obtained using 869 mtSNPs from ▶ the union of all polymorphisms identified by PCA and clustering for the clades and M, N, L haplogroups. The tree shown is the consensus over trees from 100 datasets, each of which was created by selecting one sample randomly from each of the 55 haplogroups shown. The branch labels on the consensus MP tree are a measure of the reliability of the branch. This is estimated as the fraction of cases when the branch splits the downstream haplogroups into the sets shown in the tree over the sampled datasets. Ancient branches, corresponding to the clade tree in Fig. 2, are reliably reproduced, as are some recent branches. The middle branches have a lower reliability, with branch accuracies of <10% in many cases. This makes the overall reliability of the tree very low. (**b**) The consensus ML tree obtained using 869 mtSNPs from the union of all polymorphisms identified by PCA and clustering for the clades and M, N, L haplogroups. The tree shown is the consensus over trees from 100 datasets, each of which was created by selecting one sample randomly from each of the 55 haplogroups shown. The branch labels on the consensus ML tree are a measure of the reliability of the branch. This is estimated as the fraction of cases when the branch splits the downstream haplogroups into the sets shown in the tree over the sampled datasets. Ancient branches, corresponding to the clade tree in Fig. 2, are reliably reproduced, as are some recent branches. The middle branches have a lower reliability, with branch accuracies of <10% in many cases. This makes the overall reliability of the tree very low. (**c**) The consensus NJ tree obtained using 869 mtSNPs from the union of all polymorphisms identified by PCA and clustering for the clades and M, N, L haplogroups. The tree shown is the consensus over trees from 100 datasets, each of which was created by selecting one sample randomly from each of the 55 haplogroups shown. The branch labels on the consensus NJ tree are a measure of the reliability of the branch. This is estimated as the fraction of cases when the branch splits the downstream haplogroups into the sets shown in the tree over the sampled datasets. Ancient branches, corresponding to the clade tree in Fig. 2, are reliably reproduced, as are some recent branches. The middle branches have a lower reliability, with branch accuracies of <10% in many cases. This makes the overall reliability of the tree very low. (**d**) The consensus UPGMA tree obtained using 869 mtSNPs from the union of all polymorphisms identified by PCA and clustering for the clades and M, N, L haplogroups. The tree shown is the consensus over trees from 100 datasets, each of which was created by selecting one sample randomly from each of the 55 haplogroups shown. The branch labels on the consensus UPGMA tree are a measure of the reliability of the branch. This is estimated as the fraction of cases when the branch splits the downstream haplogroups into the sets shown in the tree over the sampled datasets. Ancient branches, corresponding to the clade tree in Fig. 2, are reliably reproduced, as are some recent branches. The middle branches have a lower reliability, with branch accuracies of <10% in many cases. This makes the overall reliability of the tree very low


a
L01
L1b
100%
L1c
L2a
90%
100%
L2c
94%
L2b
L3b
83%
L3d
L3f
30%
90%
L3eb
78%
L3ea
W
56%
37%
I
K
53%
J1a
100%
J1b
75%
T2
24%
97%
40%
T1
N9b
9%
B5
25%
19%
X
29%
B4b
48%
B4b
39%
8%
B4b
R5
15%
43%
F2
85%
F1
3%
A
N9a
15%
64%
U5
21%
H4
32%
V
M10
45%
Q
M8
100%
63%
Z
90%
C
12%
M7a
68%
M7b
G2
6%
100%
G1
M9
53%
D5a
100%
D5b
38%
D4b2a
100%
D4b2b
100%
97%
D4b1a
100%
D4b1a
100%
D4h
6%
D4e2
56%
D4e1
4%
D4d1a
69%
D4d1b
8%
D4g
14%
D4a

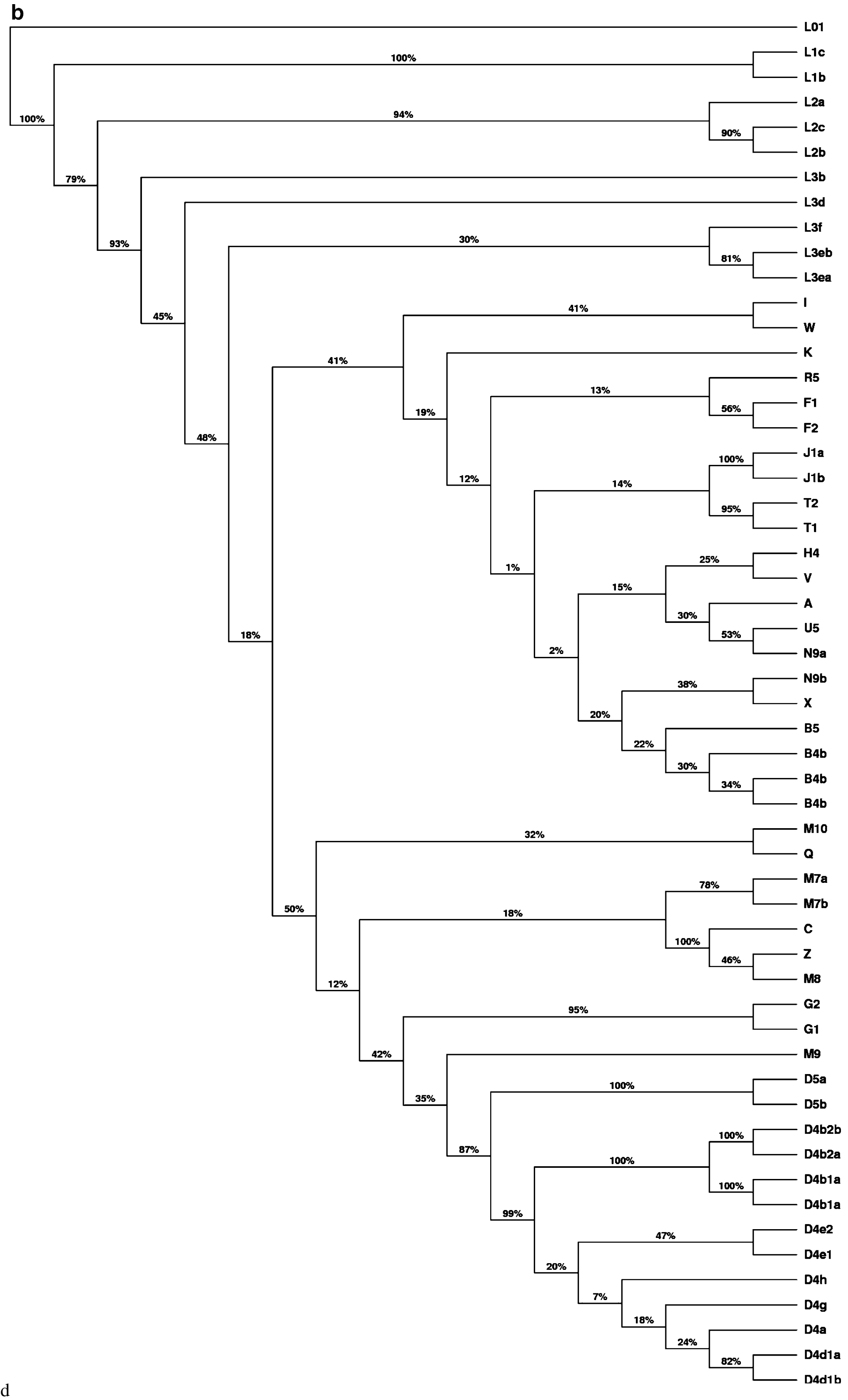


Fig. 8 continued

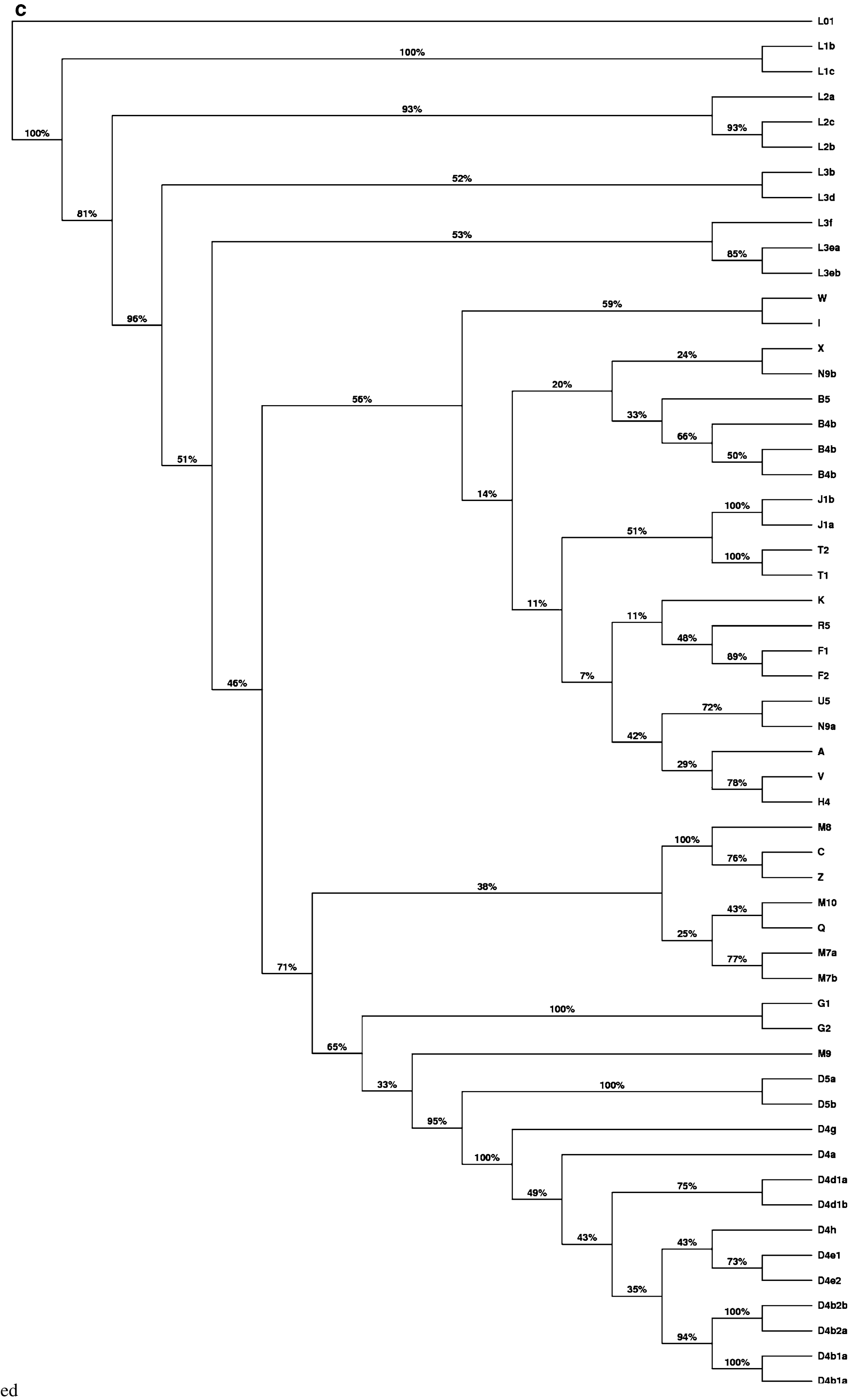


Fig. 8 continued

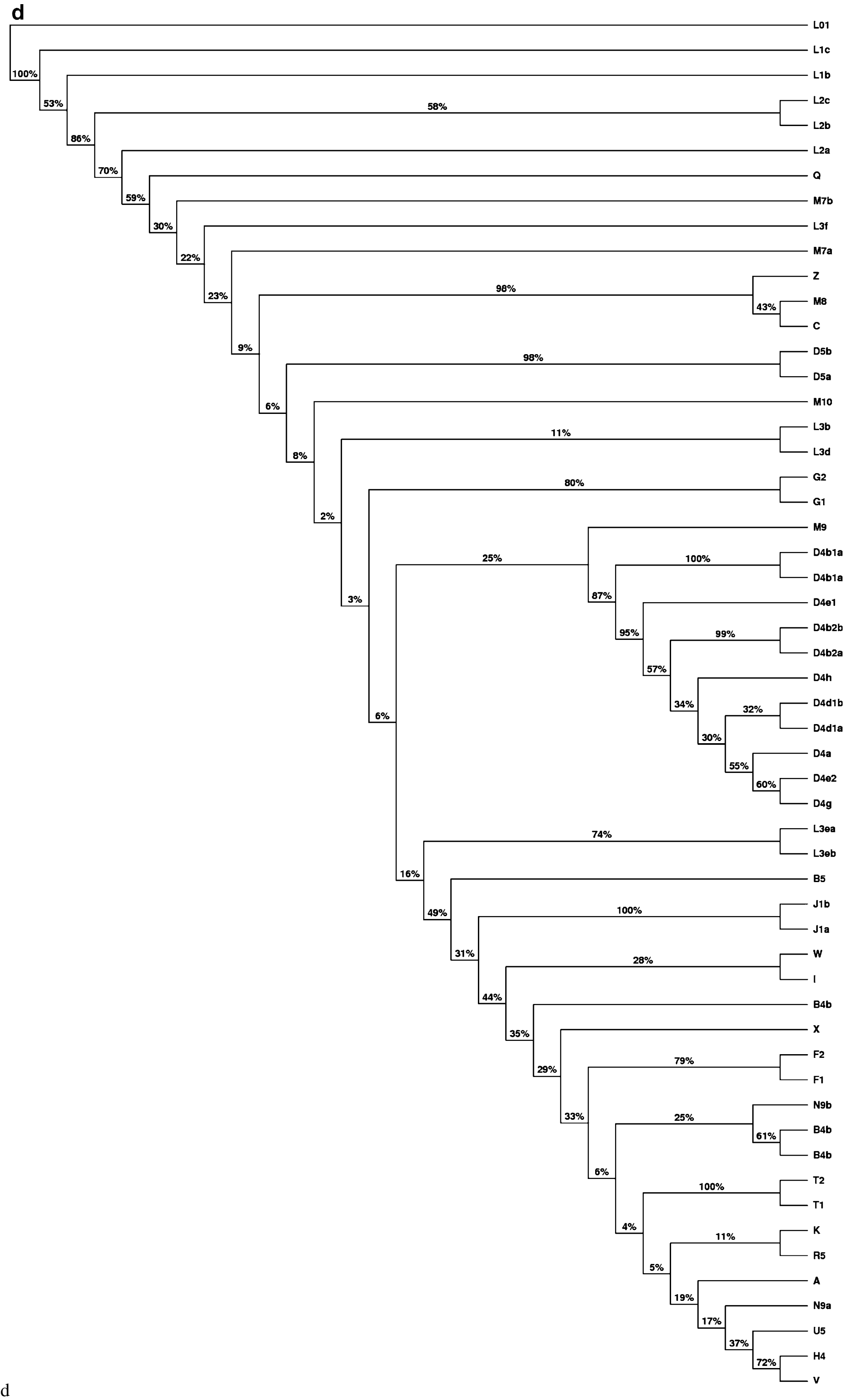


Fig. 8 continued

Method II: Rooting Using Poisson Statistics for the Number of Mutations on Edges

The second method is novel and is based on the fundamental observation that the same amount of time has elapsed from the root to each leaf. Hence, if the mutation rate is assumed to be fixed (as it is here), then the number of mtSNPs from the root to all leaves must be (approximately) the same. We also impose the constraint that the root be the most probable choice to satisfy this criterion (equal distance from leaves) across all possible evolutionary scenarios on the internal branches, as described in greater detail below.

If all loci are equally likely to mutate and the mutation rate is low, the number of mtSNPs on internal branches is a Poisson variable with Poisson parameter proportional to the time corresponding to the branch. The number of actual mtSNPs on the branch is an unbiased estimator of this Poisson parameter. To find the root, we created a number of equivalent evolutionary networks by simulating the Poisson variables (number of mtSNPs on the internal branches), using the observed number of mtSNPs on the edges as the Poisson parameter. The distance $D(R \rightarrow L_i)$ of a leaf L_i from a node R is the sum of the number of mtSNPs labeling the edges in the path from R to L_i. We simulated 1000 evolutionary scenarios, and for each scenario we computed the distances $D(R \rightarrow L_i)$ for all the paths from each possible internal root to the leaves. For each scenario, we then computed the mean and standard deviation (SD) of D over these paths, and from these, the distribution of SDs over the 1000 evolutionary scenarios for each internal node. Figure 9 shows the distribution of SD for the three internal nodes R1, R2, and R3. If we make the reasonable assumption that the best root is the one with the lowest possible SD averaged over all possible evolutionary scenarios, then it is clear from Fig. 9 Fig. AIII.1 that, once again, R1 is the preferred root.

Although both methods give the same result for the root of the mtDNA tree, we prefer the second method, because the averaging over equally probable evolutionary scenarios provides an additional measure of confidence in the identification.

Appendix D: Software to Reproduce Figs. 1 and 2 and Supplementary Table STII

The software is described below and is available for download at: https://biomaps.rutgers.edu/wiki/upload/9/93/MtDNA_utility.tar.gz.

The code uses Python (http://www.python.org) with C extensions. The README file gives specific instructions on how to install the libraries and compile the code. The code takes as input 1737 aligned mtDNA sequences on 3177 polymorphic loci and from them creates a binary "mutation matrix" B, whose rows represent samples and columns represent mtDNA loci. Each column element is assigned the value 1 if the nucleotide matches rCRS and the value 0 if it does not. The process begins by calling migration, which uses parsers.ParseSTI to read Table STI.txt. This parsing class creates a SampleData object containing each of the 1737 samples, their aligned sequences of 3177 polymorphic loci, and their haplogroups. This is then compared to the rCRS sequence pairwise to create the binary matrix B. Next, this matrix is centered so that each column has 0 mean using a call to pca.py. This class also performs a singular value decomposition of the matrix, and the resulting first two eigenvectors, which correspond to the highest eigenvalues, are used to generate a plot (saved to disk using display.py) of each sample in principal component space. This plot corresponds to Fig. 1 in the text. From this plot, using predefined cut-points on the coordinates, clade membership is assigned.

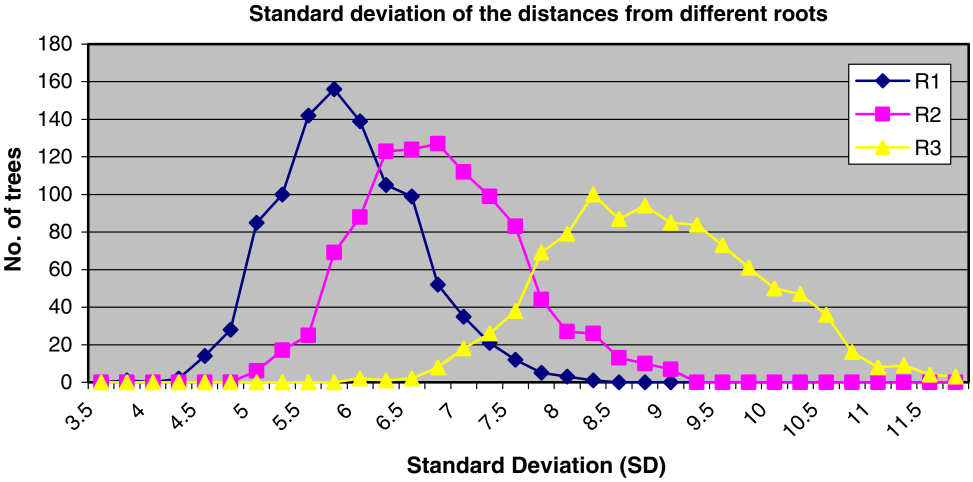


Fig. 9 Standard deviation (SD) of the distances from different roots. Distribution of the SDs of the distances from three possible roots to the leaves over 1000 possible evolutionary scenarios, generated by considering the number of mutations on the internal edges as a Poisson variable with Poisson parameter equal to the observed number of mutations on the edge

Once clade membership is assigned, frequency analysis of the mtSNPs in each clade is performed in migration.py. A mtSNP is selected if it appears in 95% of the samples in one clade and <5% of the samples in any other clade. This identifies the 34 mtSNPs given in Supplementary Table STII.

The call to migration.py also identifies the 410 mtSNPs which occur with high weights (top 25% by absolute value) as coefficients in 166 eigenvectors corresponding to the highest eigenvalues, which represent 85% of the variance in the data. These mtSNPs are listed by clade in Supplementary Table STIX. Migration.py then randomly samples 20 individuals from each of the five clades and reduces their data vectors to just these reliable mtSNPs. These 100 samples are then used as input for cluster.py, which repeatedly samples from within the set and uses *k*-means to divide them into $k = 2, 3, 4, 5$, and 6 clusters. To create a dataset of these 100 samples, cluster.py either randomly selects 80% of the samples, or randomly selects 80% of the mtSNPs for each sample, or randomly selects 80% of the samples and 80% of the mtSNPs for each sample, or leaves the samples and data vectors unmodified. This is repeated 300 times and the data for each *k* are combined into an agreement matrix $\mathbf{M}$, whose entries M_{ij} correspond to the fraction of times samples *i* and *j* were clustered together across the 300 samplings. $(1 - M_{ij})$ may be considered the "distance" between sample *i* and sample *j*. Using this definition of distance, we cluster the samples again using hierarchical clustering with average linkage to assign the final clusters for each *k*. This agreement matrix is then reordered using Simulated Annealing to maximize the similarity between adjacent samples within a cluster and maximize the dissimilarity between samples farther apart. The reordered matrix is recorded to disk using display.py as a heatmap, where bright spots represent samples which were heavily clustered together, and dark spots represent samples which were rarely clustered together.

At this point all five clades are split and the sequence of splits represents (unrooted) Fig. 2, which can then be rooted using the procedure described in Appendix C.

References

Bandelt H-J, Richards M, Macaulay V (2006) Human mitochondrial DNA and the evolution of Homo sapiens (Nucleic acids and molecular biology), 1st edn. Springer, New York

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325(6099):31–36

Cerny V (1985) A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. J Optim Theory Appl 45:41–51

Densmore LD 3rd (2001) Phylogenetic inference and parsimony analysis. Methods Mol Biol 176:23–36

Drummond A, Rodrigo AG (2000) Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol Biol Evol 17(12):1807–1815

Felsenstein J (1996) Inferrring phylogenies. Sinauer Associates, Sunderland, MA

Harpending H, Eswaran V, Macaulay V et al (2005) Tracing modern human origins. Science 309(5743):1995b–1997

Hasegawa M, Kishino H, Saitou N (1991) On the maximum likelihood method in molecular phylogenetics. J Mol Evol 32(5):443–445

Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408(6813):708–713

Jin G, Nakhleh L, Snir S, Tuller T (2006) Maximum likelihood of phylogenetic networks. Bioinformatics 22(21):2604–2611

Jobling MA, Hurles ME, Tyler-Smith C (2004) Human evolutionary genetics: origins, peoples, and disease. Garland Science, New York

Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York

Kaufmann L, Rousserw PJ (1990) Finding groups in data: an introduction to cluster analysis, 1st edn. John Wiley & Sons, New York

Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. Science 220(4598):671–680

Kong QP, Bandelt HJ, Sun C et al (2006) Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15(13):2076–2086

Kumar S, Gadagkar SR (2000) Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies. J Mol Evol 51(6):544–553

Minh BQ, Vinh le S, von Haeseler A, Schmidt HA (2005) pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. Bioinformatics 21(19):3794–3796

Monti S, Tamayo P, Mesirov PJ, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learn J 52(1–2):91–118

Myers E, Miller W (1998) Optimal alignments in linear space. CABIOS 4(1):11–17

Ota S, Li WH (2000) NJML: a hybrid algorithm for the neighbor-joining and maximum-likelihood methods. Mol Biol Evol 17(9):1401–1409

Parsons BL, Heflich RH (1998) Detection of basepair substitution mutation at a frequency of $1 \times 10^{(-7)}$ by combining two genotypic selection methods, MutEx enrichment and allele-specific competitive blocker PCR. Environ Mol Mutagen 32(3):200–211

Pearson WR, Robins G, Zhang T (1999) Generalized neighbor-joining: more reliable phylogenetic tree reconstruction. Mol Biol Evol 16(6):806–816

Saitou N (1990) Maximum likelihood methods. Methods Enzymol 183:584–598

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sanderson MJ (1994) Reconstructing the history of evolutionary processes using maximum likelihood. Soc Gen Physiol Ser 49:13–26

Shinoda K-I (2005) Ancient DNA analysis of skeletal samples recovered from the Kuma-Nishioda Yayoi site. Bull Natl Sci Mus Ser D (Anthropol) 30:1–8

Stewart CB (1993) The powers and pitfalls of parsimony. Nature 361(6413):603–607

Strehl A, Ghosh J (2002) Cluster ensembles: a knowledge reuse framework for combining partitionings. In: Eighteenth National

Conference on Artificial Intelligence, July 28–August 01, 2002, Edmonton, Alberta, Canada, pp 93–98

Stringer C (2001) Modern human origins—distinguishing the models. J Afr Archaeol Rev 18(2):67–75

Studier JA, Keppler KJ (1988) A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol 5(6):729–731

Sullivan J (2005) Maximum-likelihood methods for phylogeny estimation. Methods Enzymol 395:757–779

Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci USA 101(30):11030–11035

Tanaka M, Ozawa T (1994) Strand and symmetry in human mitochondria. Genomics 22:327–335

Tanaka M, Cabrera VM, Gonzalez AM et al (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14(10A):1832–1850

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a dataset via the gap statistic. J Roy Stat Soc Ser B 63:411–423

Yang Z (1996) Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol 42(2):294–307

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13(5):555–556