Inferring Common Origins from mtDNA

Ajay K. Royyuru[1], Gabriela Alexe[1], Daniel Platt[1], Ravi Vijaya-Satya[2], Laxmi Parida[1], Saharon Rosset[1], and Gyan Bhanot[1,3]

[1] Computational Biology Center, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598
[2] Department of Computer Science, University of Central Florida, Orlando, FL 32816
[3] Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540
ajayr@us.ibm.com, galexe@us.ibm.com, watplatt@us.ibm.com, rvijaya@cs.ucf.edu, parida@us.ibm.com, srosset@us.ibm.com, gyan@us.ibm.com

Abstract. The history of human migratory events can be inferred from observed variations in DNA sequences. Such studies on non-recombinant mtDNA and Y-chromosome show that present day humans outside Africa originated from one or more migrations of small groups of individuals between 30K-70K YBP. Coalescence theory reveals that, any collection of non-recombinant DNA sequences can be traced back to a common ancestor. Mutations fixed by genetic drift act as markers on the timeline from the common ancestor to the present and can be used to infer migration and founder events that occurred in ancestral populations. However, most mutations seen in the data today are relatively recent and do not carry useful information about deep ancestry. The only ones that can be used reliably are those that can be shown to robustly distinguish large clusters of individuals and thus qualify as true representatives of population events in the past.

In this talk, we present results from the analysis of 1737 complete mtDNA sequences from public databases to infer such a robust set of mutations that reveal the haplogroup phylogeny. Using principal component analysis we identify the samples in L, M and N clades and with unsupervised consensus ensemble clustering we infer the substructure in these clades. Traditional methods are inadequate to handle data of this size and complexity.

The substructure is inferred using a new algorithm that mitigates the usual problems of sample size bias within haplogroups as well as the sampling bias across haplogroups. First, we cluster the data in each of the M, N, L clades separately into $k = 2, 3, 4, \ldots k_{max}$ groups using an agreement matrix derived from multiple clustering techniques and bootstrap sampling. Repeated training/test splits of the samples identify robust clusters and patterns of SNPs which can assign haplogroup labels with a reliability greater than 90%. Even though the clustering at each k is done independently, the clusters split in a way that suggests that the data is revealing population events; a cluster at level k has $k-2$ clusters which are identical with those at level $k-1$ plus two more that

A. Apostolico et al. (Eds.): RECOMB 2006, LNBI 3909, pp. 246–247, 2006.

obtain from a split of one of the clusters at level $k-1$. The clustering is repeated with equal number of samples from the first level clusters. The sequence in which the clusters now split defines a binary network which reveals population events unbiased by sample size. We root the network using an out-group and, assuming a molecular clock, identify an internal node in the bifurcation process which is equidistant from the leaves. This rooting removes the bias across haplogroups which would otherwise influence the order in which the clusters emerge.

Our analysis shows that the African clades L0/L1, L2 and L3 have the greatest heterogeneity of SNPs, in agreement with their ancient ancestry. It also suggests that the M, N clades originated from a common ancestor of L3 in two separate migrations. The first migration gave rise to the M haplogroup, whose descendents currently populate South-East Asia and Australia. The second migration resulted in the N haplogroup, accounting for the current populations in China, Japan, Europe, Central Asia and North and South America. We reveal and robustly label many branches of the mtDNA tree, improving current results significantly. We find that for our choice of robust SNPs, the genetic distances between the NA and NRB haplogroups is smaller compared to that between B and J/T/H/V/U. The detailed N migratory sub-tree is rooted so that the T, J and U haplogroups are on one side of the root and the F, V/H, I, X, R5, B, N9, A and W are on the other. We also find a detailed structure for the M tree consistent with prior literature and we infer additional branches for the MD haplogroup. Finally we provide detailed SNP patterns for each haplogroup identified by our clustering. Our patterns can be used to infer a haplogroup assignment with reliability greater than 90%.