Evolutionary History of Populations Employing a Bifurcating Tree with Horizontal Links

Yuri Kartavtsev*

Professor Far Eastern Branch of Russian Academy of Sciences, Russia

Abstract

Over the past 20 years numerous new trees of recent human populations are published extensively but little attention has been paid to formal phylogenetic synthesis. We utilized the "matrix representation with parsimony" (MRP) method to infer a composite phylogeny (supertree) of recent human populations, supported 257 genetic/genomic, also as linguistic, phylogenetic trees and 44 admixture plots from 200 published studies (1990–2014). The resulting supertree topology includes the foremost basal position of S African Khoisan followed by C African Pygmies and therefore the paraphyletic section of all other sub-Saharan peoples. The sub-Saharan African section is basal to the monophyletic clade consisting of the N African–W Eurasian assemblage and therefore the consistently monophyletic Eastern superclade (Sahul–Oceanian, E Asian and Beringian–American peoples). This topology, dominated by genetic data, is well-resolved and robust to parameter set changes, with a couple of unstable areas (e.g., West Eurasia, Sahul–Melanesia) reflecting the prevailing phylogenetic controversies. a couple of populations were identified as highly unstable "wildcard taxa" (e.g. Andamanese, Malagasy). The linguistic classification fits rather poorly on the supertreetopology, supporting a view that direct coevolution between genes and languages is way from universal.

Introduction

Evolutionary history of recent human populations is an extensively studied topic of great complexity. Human population history is never purely phylogenetic, or tree-like1, as genetic admixture, mediated by processes like migrations, expansions, intermarriage, trade, or slavery, have played a crucial role in shaping human history2. There is, however, a robust hierarchical signal which will be hypothesized as phylogeny in both genetic3,4 and cultural (especially linguistic) data. it's worth noting that even using such terms as "genetic admixture" and "horizontal gene flow" implies an assumption of an underlying tree-like model7. Recently developed phylogenetic methods applied to both genetic and linguistic data10 allow us to see evolutionary history of populations employing a bifurcating tree with horizontal links ("admixture edges"), accounting for both population splits and mixtures.

Today, no unified picture of recent human evolution supported genetic data is out there, as studies that infer human population history have used differing types of genetic markers, from "classical polymorphisms" (such as AB0 blood groups and protein allomorphisms) and uniparental markers (the mitochondrial DNA and therefore the non-recombining portion of the Y chromosome) to genome-wide allele frequency data and data supported whole-genome sequencing. Moreover, individual studies only partially overlap taxonomically. Even the most important published tree (267 populations) supported genome-wide data lacks several population groups important for a comprehensive description of human population history on a worldwide scale (e.g., populations of N Africa, Anatolia, Balkans, E Europe, Indonesia, N Asia, Beringia and N America). A recent meta-analysis of human genomic diversity projects13 has also pointed to the shortage of several key population groups (e.g., Hadza, Sandawe, Fulani, Chadic speakers, Australian Aboriginals, populations of Indonesia, Polynesia and Northern America).

The language phylogenies published so far include up to 542 language varieties but usually cover only one language family each (mostly Bantu, Indo-European, or Austronesian). Formal attempts to reconstruct genealogical relationships between languages beyond the extent of the families are rare so far and nearly all of the proposed linguistic macrofamilies like Eurasiatic/ Nostratic Indo-Pacific and Amerind are considered controversial.

Although an outsized body of comparative data currently exists for a phylogenetic synthesis, integration of all types of data employing a "supermatrix approach" (or "total evidence approach"22) remains unfeasible for the human population, particularly thanks to the distance-based (instead of character-based) nature of some source data and lack of widely overlapping datasets. In light of those problems, a possible strategy is to specialise in published ("source") trees, adopting the "supertree approach" (or "taxonomic congruence approach"23). the first application of supertrees is to summarize existing phylogenetic hypotheses during a sort of an artificial consensus which may be wont to identify and evaluate topological conflicts caused by incongruent or missing data. within the "matrix representation with parsimony" (MRP) method25,26, each source tree is converted into a matrix of additive binary characters; the individual matrices are eventually merged into one character matrix which is then analyzed by the utmost parsimony (MP) method to get a composite phylogeny. The resulting supertree is analogous to a consensus tree when the source trees have different sets of taxa.