Human origins and genetic evolution

Paul Grayson

December 1, 2001

1 Introduction

One area where phylogenetic trees can reveal interesting information is in the history of human evolution. Archaeologists have worked for centuries to determine how people have spread throughout the world and are especially interested in finding out exactly where our species originated. Careful examination of the genetic relationships between people in different areas of the world has started to answer some of the archaeologists questions. In this essay I will discuss two recent studies [3, 2], the later of which was mentioned in a review of the field by Rebecca Cann [1]. The two studies, done by the same group, were very similar and used an overlapping set of samples from the population, but focused on different parts of the genetic material.

Section 2 describes the sampling techniques common to the two studies. Section 3 discusses the differences between the sections of genetic material picked for the two studies, and Section 4 explains what we can conclude from them.

2 Language group–based sampling

The two studies used speakers of languages from almost all linguistic groups, in order to get most complete phylogenetic tree possible. For a phylogenetic tree to be representative of the genetic diversity present in humans today, it needs to have samples of as many genetically different humans as possible. Two nearly-identical samples of DNA are not useful when constructing a global tree of human evolution, since they do not add large branches to the tree. However, it is not possible to know how different two people's genes

are before the genes are sequenced, so it is not obvious how to choose the distribution of samples ahead of time. A good guess would be to try to sample groups of people in proportion to their sizes (*i.e.* weight all people on earth equally in a random drawing). However, this approach is very likely to miss many small, genetically isolated groups. The studies used 69 and 53 samples, respectively — if they were chosen based on population size, genetic groups containing less than 100 million people (almost all genetic groups!) would likely be missed, while the few groups that have grown in size over the last thousand years would be over-sampled.

Luckily, there is an easy way to find genetically isolated groups of people. Languages evolve slowly, and in a similar way to genes: they are almost always passed from parent to child. Therefore, groups that are have very distantly related genes are likely to speak very different languages. Sampling by language group is, as a result, better than sampling by population size at identifying individuals from a large number of genetic groups. The people used in these studies came from 16 out of 17 major linguistic divisions, and beyond that were (probably) picked to maximize linguistic diversity, so we can assume that they form a representative set of the genetic groups of people alive today.

3 Differences between the studies

The two studies built trees based on different sections of DNA, and the differences are typical of what is seen in other genetic studies. The first study [3] used a non-coding segment of the X chromosome that is known to have very low mutation and recombination rates. It is important to use non-coding regions to eliminate variable drift caused by natural selection, and regions with low recombination are known to have a history that is tree-like (rather than web-like), so that construction of a phylogenetic tree can be useful. The low mutation rate sets the scale on which the phylogenetic analysis can be conducted: the 69 people sampled in this study have only 20 distinct versions of the 10^4–base-pair sequence, so the resulting tree only has 20 leaves. Unsurprisingly, all four intermediate nodes of the tree correspond to sequences actually found in the population.

The second study [2] also used a piece of DNA with a low recombination rate: the entire mitochondrial DNA sequence. To the best of current knowledge, this DNA is always inherited from the mother. Since this DNA

does contain genes, it was important to check whether its evolution can be assumed to be constant over time. Other studies have verified that this assumption is valid for human mitochondrial DNA, except within the D-loop. Because of its non-constant mutation rate, the D-loop was excluded from this study. The mitochondrial DNA has a similar size to the X chromosome DNA used in the earlier study, but a much higher mutation rate. Each of the 53 people sampled in this study has a different sequence, and only one intermediate node is found in the sample. As a result, an approximation to the full tree of human evolution can be drawn with this data.

4 Conclusions about human origins and future research

The studies agreed very well with each other and with others. They both clearly showed a much larger diversity among Africans than non-Africans, and the bell-shaped distribution of non-African genetic diversity shown in the later study indicates a recent population boom. These findings contribute to the growing support for the "out-of-Africa" theory of human origins, but cannot locate the place of origin precisely.

The two studies each show interesting possible evolutionary pathways for a large number of different groups of people. However, the accuracy of any phylogenetic tree is questionable, and it would be interesting to see how the trees derived here agree with others. In particular, it should be possible to check how well the trees derived in the two studies here correspond, since they involve the same groups of people. A lack of correspondence would not necessarily indicate inaccuracy, since the two segments of DNA used could have very different evolutionary histories. The authors mention briefly that *individuals* can have very different relationships in the two trees, but it is unclear how incompatible the overall trees are.

References

[1] Rebecca L. Cann. Genetic clues to dispersal in human populations: Retracing the past from the present. *Science*, 291:1742–1748, March 2001. Note: Rebecca Cann was one of the first to demonstrate the utility of mtDNA in analyzing human origins, in 1987.

[2] Max Ingman, Henrik Kaessmann, Svante Pääbo, and Ulf Gyllensten. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408:708–713, December 2000.

[3] Henrik Kaessmann, Florian Heißig, Arndt von Haeseler, and Svante Pääbo. Dna sequence variation in a non-coding region of low recombination on the human x chromosome. *Nature Genetics*, 22:78–81, May 1999.