Using Mitochondrial DNA and Phylogenetic Trees to Trace Early Human Expansions

Ian O'Dwyer

December 10, 2001

In the first part of this essay, I review two recent papers which have sought to use mitochondrial DNA (mtDNA) in order to trace human expansion around the globe from times as far back as ∼170,000 years ago. In the second part, I review my own attempts to generate phylogenetic trees to compare with those in the two papers, and discuss the results and problems associated with this.

Mitochondrial DNA (mtDNA) is particularly suitable for the study of early human (female humans in this case since mtDNA is maternally inherited!) expansions into the world from what appears to be a common original location in Africa. mtDNA is a non-recombining (recombinant DNA is a new DNA sequence formed by the combination of two nonhomologous DNA molecules), haploid (contains only one chromosome) molecule, meaning that any differences between two mtDNA sequences are due to mutations. Hence, it is possible to compare mtDNA from various sources in the modern world and try to match the mutations in order to decipher the origins of the DNA. By using modern DNA sequencing and comparison techniques, such as those which everyone on this course has had a chance to experiment with, phylogenetic trees can be produced which allow the roots of modern man to be traced. Once the origin of a DNA sequence is identified and its place in the phylogenetic tree calculated, it is possible to follow the expansions of man from his original location or locations on the planet to the current geographic population of the world. If this information is tied in with archeological data, a strong test of a particular geographic expansion theory or model is possible.

In order to be able to associate historical epochs with the expansion of early man and the mtDNA mutations, a standard clock is required. Enter the *molecular clock hypothesis*, which states that DNA sequences evolve approximately constantly over time in all evolutionary lineages. In addition to this, there is palaeontological and generic evidence that there exists a divergence time between humans and chimapanzees of ∼5 million years. This divergence time is quite controversial and may be much larger, a problem which is exacerbated when dealing with timescales as small as ∼10,000 years. For example, is it more reasonable to claim that humans and chimpanzees suddenly diverged at a specific moment in time, or that the process occured gradually? Assuming that the value above is reasonable, coupling these two theories and considering the mean genetic distance between all humans and one chimpanzee sequence, the estimated mutation rate for mtDNA is 1.70×10^{-8} substitutions per year. For more detailed analysis see [1]. Hence, we now have a clock with which we can compare variations in the mtDNA sequences and which can be used with the phylogenetic tree to match historical periods with the expansions of modern man into the world at large.

In a recent paper by Maca-Meyer et al[2], a study of 42 mtDNA sequences from different hu-

man lineages was conducted. The 42 sequences come from various haplotypes. A haplotype is created as time passes and mutations to the mtDNA accumulate sequentially along the molecules. The initial mutations, that is, those which occur earlier in the mtDNA sequence and are thus shared by more lineages represent haplogroups. The major haplogroups covered in the paper are sub-Saharan African lineages, a group containing European, North African and Western Asian Caucasions and finally a group which describes all of the lineages for Oceania, Asia and native Americans. As time progresses, the mutations which occur within the haplogroups cause the mtDNA to be less and less related until finally we reach the level of the individual who may have unique mtDNA. The phylogenetic tree which was produced by the group is shown in Figure 1., and figure 2. gives the correspondence between the letter representations of the haplogroups in the figure and the individual group or sub-group which it represents.

The basic result is that all of the human lineages tested coalesce into a single unit between 156,000 to 169,000 years ago(YA). Further analysis is also possible based on the phylogenetic tree, giving expansion timescales and geographical paths for the expansion. There appears to have been a split within the African group around 122,000-132,000 YA and another 85,000-95,000 YA forming new clades there, although with no expansion at this time. The next split was 59,000-69,000 YA and corresponds to the first expansion out of Africa. At this point there is some debate as to the route of the expansion, but it was either via Central Asia (haplogroup M) or Southern Asia. One clade which left Africa around 30,000-58,000 YA has subsequently been found in India and Eastern Asia. The second clade leaving Africa around this time produced mainly caucasian lineages in at least three groups , suggesting a Northern route for this expansion into modern Western and Eastern Europe. One group (X) is mainly restricted to Europe, whilst another group (A) is spead across Asia. An interesting point in the paper is that representatives of both of these clusters are found in native Americans, suggesting a possible European ancestry. Also, the X haplotype has been found in Central Asia lacking a particular mutation found in the European X haplotype. It is this same mutation which is lacking in the American X haplotype, giving rise to the possibility that the founders of the New World may have been from Central Asia. The final expansion from Africa was 39,000-52,000 YA and has produced four ancestral clusters, one of which populated Japan and the South-Eastern Pacific. There is also evidence for a return to Africa and further expansions from the population centers established above, the details of which are largely unimportant to an overall understanding of the techinique, except to note that even with the subsequent expansions and more modern movements and changes to the gene pool, it is still possible to use this technique to infer the earliest human routes from them. It should be noted that since the model above is based on comparison with some incomplete genomic sequences, it is tentative, but in line with other similar work[1] at a statistically significant level. Also, there may be some selection bias in the way the DNA samples were chosen. Whilst it is desirable to select sequences from the major haplogroups and clades, using modern demographics to do this is questionable and an alternative method might have been to use linguistic phyla, as was done in [1].

There appeared to be some scope to discover more about creating phylogenetic trees and to confirm the results above, which seemed fascinating, even if a little unbelievable - to be able to infer so much history from something as small as mtDNA was too good to be true! 11 of the sequences used in [1] were retrieved from GenBank. The 11 groups selected from Genbank were: Aborigine, Uzbek, San, Evenki, Effik, Mbuti, Evenki, Buriat, Mbenzele, Hausa and finally the Chimpanzee as a control. These groups were selected as they gave a fairly good cross section of the different branches represented in the Ingman et al paper and it was hoped they would allow for some interesting comparisons with the tree in that paper

(see Figure 3.). Initially, I had anticipated using these sequences in the Biology Workbench, as with the primate sequences which were tested previously, but since the mtDNA sequences were complete, containing ∼16,560 bases and the limit for the biology workbench is 10,000 bases, I was forced to seek alternative software. I decided on a package called fastDNAml [3] which is a super-charged version of the standard PHYLIP (Phylogeny Inference Package [4]) software, widely used in one form or another to perform DNA sequence comparison and to infer phylogenetic tree structure. fastDNAml uses a maximum likelihood method to find the best tree structure for a particular set of DNA sequences. With just four sequences, the program was finished in a second or two and I began to wonder why a 'fast' algorithm was needed. Once I had inserted 11 full mtDNA sequences into the program, the run time was up to around 60 minutes and I started to get the point. Since the two papers discussed earlier compared 40-50 sequences and then generated phylogenetic trees, the need for an efficient algorithm was clear. Whilst my lowly 700MHz AMD Athlon with 640MB of RAM is no supercomputer, a significant amount of CPU cycles are clearly required when the number of sequences and hence the tree complexity, begins to increase. In fact a number of algorithms for parallel processor machines have been developed to speed things up further.

The final tree is shown in Figure 4. Some elements of the tree match well with those in the Ingman paper, and others are quite different. Given my level of expertise as a molecular biologist (i.e. zero) and the differences in the preparation and analysis of the data, this came as no surprise. It was disappointing to see that the chimpanzee branch did not come from the root, but that four groups (Uzbek, Mbuti, San and Evenki) shared a common ancestor with the Chimpanzee! Whilst this indicates serious errors in the method used to construct the tree, there were some more positive aspects. For example, the San and Mbuti branching relationship was preserved and the whole branch of the tree (labelled A in Fig. 3) bore a close relationship to that in the Ingman paper, although the appearance of the Uzbek lineage in this group was anomalous. Since there are clearly more serious problems with the tree, based on the Chimpanzee branch structure, further comparison and analysis does not seem worthwhile, except to note that the production of a tree of any kind was no small task!

In conclusion then, professional biologists are able to apply the techniques of phylogeny to the analysis of mtDNA in order to study the movements and expansions of early woman around the globe. The data suggest that all of the lineages tested came from a single group in Africa some 170,000 YA, which subsequently diversified and spread. Amateur efforts along the same front have demonstrated that there is a wide range of free phylogeny software available on the internet for a diversity of operating systems. The software is relatively easy to install (on a unix system at least), although there is a steep learning curve associated with available parameters, input file formats and so forth, which requires some investment of time up front. Genbank contains a vast supply of complete and partial mtDNA sequences which can be almost directly transferred to an input file for the code. The potential for this method of analysis is limited only by the ability of the operator to understand the underlying assumptions of the various models, and the nuances of the mtDNA structure which may cause unexpected results (think chimpanzee). The computational time required to generate an 11 node tree on a moderate desktop PC is ∼1 hr, indicating that simple analysis is possible without vast computational resources, although a 40-50 node tree would be a considerable challenge for a lone graduate student. The tree bore some structural resemblance to that generated in [1], but a more thorough understanding of the intricacies of phylogenetic computation is clearly required if plausible results are desired.

References

[1] Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. Mitochondrial Genome Variation and the Origin of Modern Humans Nature 2000, 408:708-713

[2] Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM. Major genomic mitochondrial lineages delineate early human expansions. BMC Genetics 2001, 2:13

[3] Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. 1994. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10: 41-48. Also see http://geta.life.uiuc.edu/ gary/programs/fastDNAml.html for details of the fastDNAml algorithm and software package.

[4] The PHYLIP homepage is located at http://evolution.genetics.washington.edu/phylip.html

Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. 1994. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10: 41-48

[5] Tree visualization software used was njplot which can be found at http://pbil.univ-lyon1.fr/software/njplot.html. This takes input in both standard tree formats which are produceed by PHYLIP type programs.

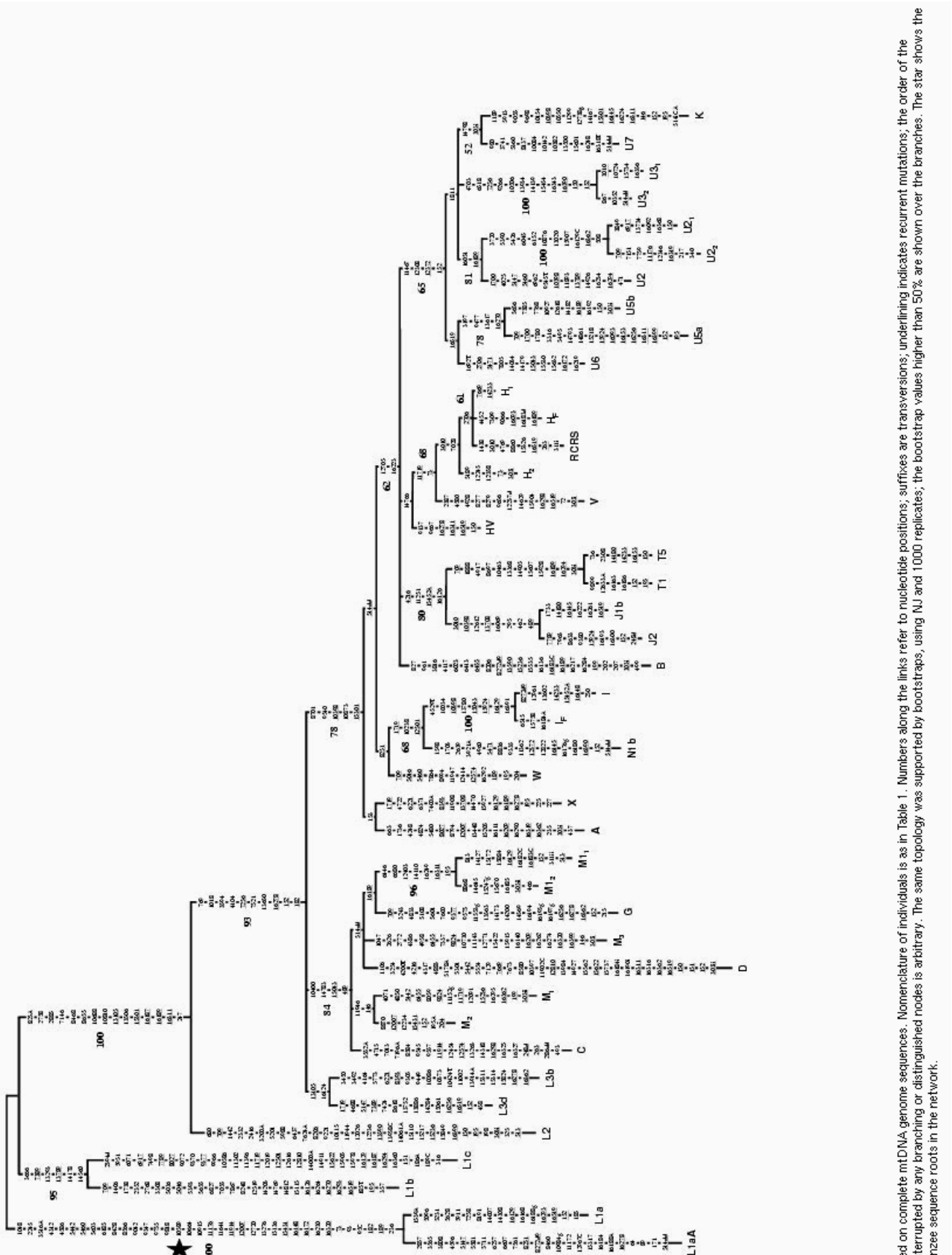

Figure 1: Reproduced from Maca-Mayer et al [2]

Table 1: HVS I motifs

Sample	HVS I motif	Haplogroup	Origin	Ref.[a]
K	145 224 311	K	Iberian	1
U7	248 318T	U7	Iberian	1
$U3_1$	343 356 390	U3	Canarian	1
$U3_2$	343 390	U3	Moroccan	1
$U2_1$	051 092 129C 189 362 368	U2	Jordanian	1
$U2_2$	051 129C 189 319 362	U2	Iberian	1
U2	051 189 234 294	U2	Jordanian	1
U5b	189 192 270	U5b	Berber	1
U5a	093 153 256 270 311 399	U5a1a	Swede	2
U6	172 219	U6	Moroccan	1
H_1		H	Mauritanian	1
H_F	093 183d 189	H		3
RCRS		H	European	4
H_2		H	Iberian	1
V	298	V	Berber	1
HV	278 311	HV	Jordanian	1
T5	126 153 189 294	T5	Moroccan	1
T1	126 163 186 189 294	T1	Iberian	1
J1b	069 126 145 222 261	J1b	Moroccan	1
J2	069 126 193 300	J2	Iberian	1
B	136 183C 189 217 284	B	Japanese	5
I	129 148 223 391	I	Iberian	1
I_F	129 184A 223 391	I		3
N1b	145 176G 180 223 390	N1b	Jordanian	1
W	223 292	W	Iberian	1
X	129 189 223 278	X	Moroccan	1
A	111 209 223 290 319 362	A	Canarian	1
$M1_1$	129 182C 183C 189 223 249 311	M1	Moroccan	1
M12	185 189 223 249 311	M1	Jordanian	1
G	189 194 195G 197G 223 256 278 362	G	Japanese	6
M_3	140 209 223 262 274 320 399	M	Japanese	7
D	184IC 190IC 223 311 316 362	D	Japanese	6
M_1	223 295 362	M	Filipino	1
M_2	223	M	Indian	1
C	223 298 325 327	C	Canarian	1
L3b	124 223 278 362	L3b	Mauritanian	1
L3d	124 223 256	L3d	Jordanian	1
L2	223 278 390	L2	Mauritanian	1
L1c	129 189 223 278 294 311 360	L1c	Mauritanian	1
L1b	126 187 189 223 264 270 278 293 311	L1b	Mauritanian	1
L1a	129 148 168 172 187 188G 189 223 230 278 293 311 320	L1a	Moroccan	1
L1aA	148 172 184 187 188A 189 223 230 311 320	L1a	African	8

[a] 1, This work; 2, GenBank accession number X93334; 3, H and I references [34], we have added for the comparisons the 263, 311i and 16519 mutations in both sequences and 00073 in the I sequence; 4, revised Cambridge reference, GenBank accession number NC 001807; 5, Positive control [35], for comparisons we added 1438; 6, MELAS, P-1 (G) and FICM (D) [36]; 7, (ref [37]); 8, GenBank accession number D38112, for comparisons we added 311i.

Figure 2: Reproduced from Maca-Mayer et al [2]

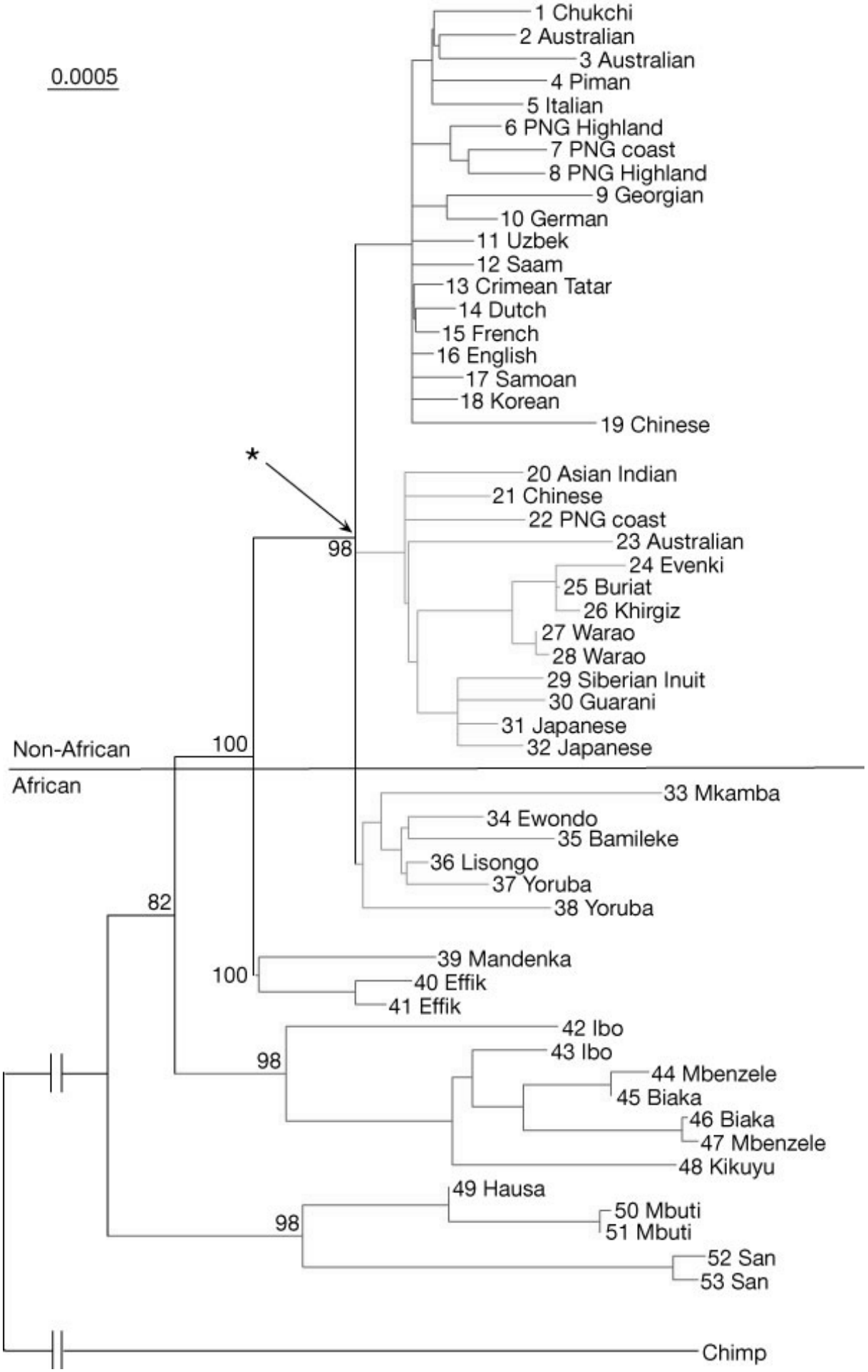

Figure 3: Reproduced from Ingman et al [1]

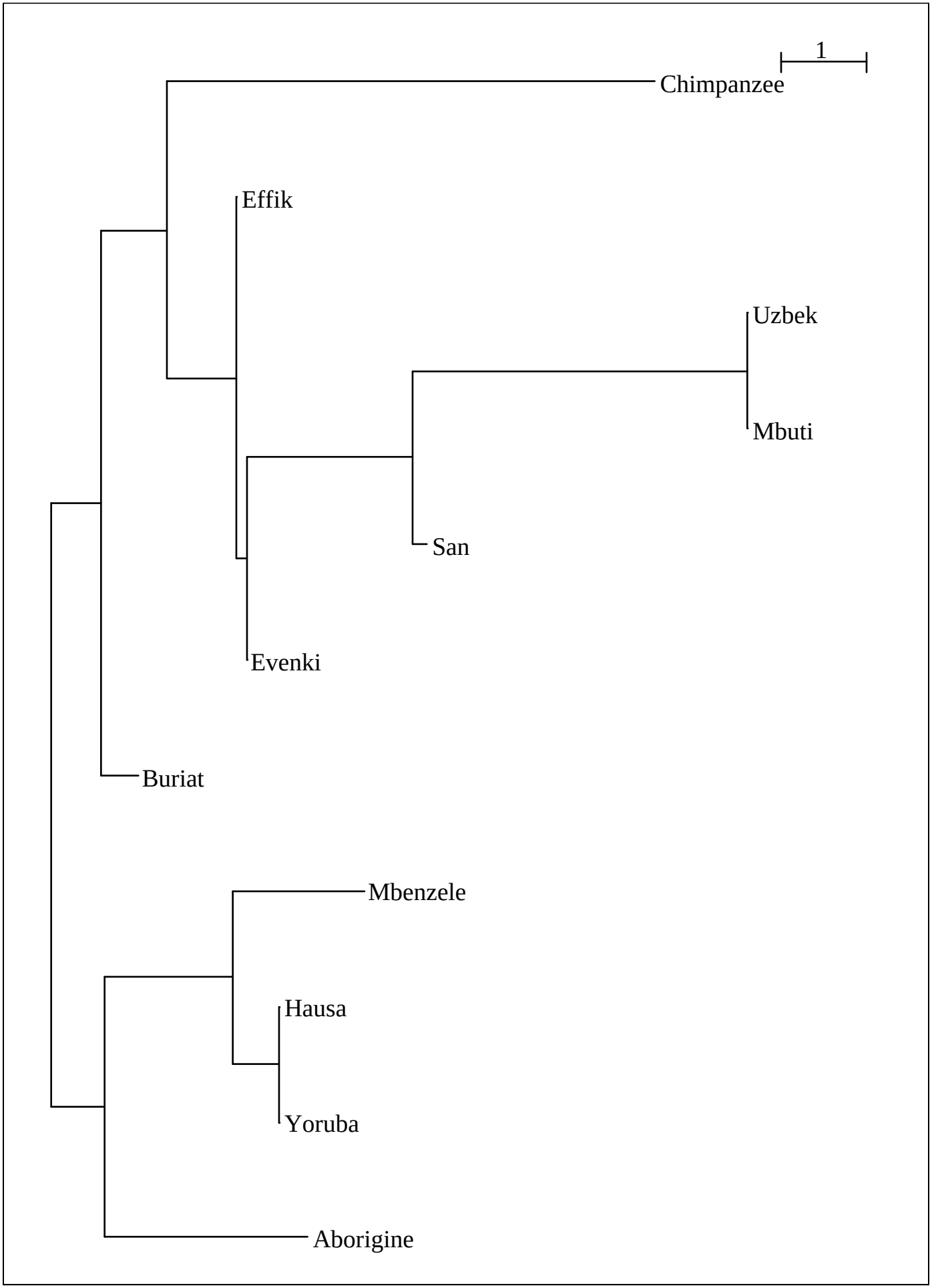

Figure 4: Amateur phylogenetic tree