Handbook of Human Symbolic Evolution

Edited by

ANDREW LOCK

Massy University,
Palmerston North,
New Zealand

CHARLES R. PETERS

University of Georgia,
Athens.
Georgia. USA

CLARENDON PRESS • OXFORD

1996

Oxford University Press, Walton Street, Oxford *0x2 6DP*
Oxford New York
Athens Auckland Bangkok Bombay
Calcutta Cape Town Dares Salaam – Delhi
Florence Hong Kong Istanbul Karachi
Kuala Lumpur Madras Madrid Melbourne
Mexico City Nairobi Paris Singapore
Taipei Tokyo Toronto
and associated companies in
Berlin Ibadan
Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press Inc., New York

A catalogue record for this book is available from the British Library
Library of Congress Cataloging in Publication Data
(Data available)
ISBN 0 19 852153 7
Typeset by Pure Tech Corporation, Pondicherry, India
Printed in Hong Kong

Contents

3

Evolutionary trees of apes and humans from DNA sequences

Peter J. Waddell and David Penny

Abstract

Developments over the past decade have made DNA sequences the primary source of information for inferring relationships between organisms. Originally sequences were used for studying relationships between species, but increasingly they are now used to study relationships between individuals and between populations. In this chapter we show how sequences have changed, and continue to change, our views of human origins and evolution. Techniques used to go from DNA sequences to evolutionary inference are outlined, because they are crucial in evaluating this vast new source of data. In addition to a review we report some of the latest research findings, and where necessary have developed appropriate statistical methods. The main points of this chapter are:

1. There is consistently strong support for the human and chimpanzee lineages' being the closest relatives to each other, and the next closest the gorilla lineage, with the orang-utan being the closest non-African relative of these African hominoids.

2. A calibration of these evolutionary trees is given, with estimated dates of divergence for the living hominoids, together with estimates of the expected errors—an important consideration to those interested in assessing the compatibility or otherwise of fossil (or palaeoanthropological) data with molecular inferences. We estimate that the divergence of human and chimpanzee lineages took place approximately 6.5 million years ago, while the standard error of such dating methods is at present about 1 million years.

3. Our evaluation of the 'Out of Africa hypotheses' (mitochondrial 'Eve') leads to the conclusion that this set of four hypotheses (pertaining to the when, where, who, and how of modern humans' origins) does indeed stand up to scrutiny; a point reinforced by our reanalysis of specific features of the data. No single data-set gives overwhelming support to all four aspects of the Out-of-Africa scenario; but it is consistent with several data-sets, while overall the data contradict the 'multiregion' hypothesis of human origins.

4. A re-evaluation of the molecular evidence confirms that the 'when' was almost certainly less than 200 000 years ago, as inferred from both mitochondrial and nuclear DNA data calibrated using both biological and palaeoanthropological data. Africa is most consistently inferred as the 'where'. The mitochondrial DNA sequences give us a glimpse of 'who' founded populations outside Africa and 'how', as populations appear to have expanded rapidly at some point after their arrival into new lands.

Novel maximum likelihood methods were developed to estimate trees with other statistical techniques to infer the reliability of branching points and species divergence dates.

3.1 Introduction

DNA sequences are now used to study two important aspects of human evolution, relationships between humans and higher primates, and relationships among modern humans *(Homo sapiens sapiens).* This chapter illustrates both aspects. For reconstructing evolutionary relationships (phylogeny), the rationale for each step, from sequence data to statistically-justifiable inferences of evolutionary events, is explained as simply as possible. We illustrate these steps with recent data which allow us to address questions such as whether common (*Pan troglodytes)* and pygmy *(Pan paniscus)* chimpanzees are the closest living relatives of modern humans.

For elucidating relationships within the group of modern humans, we introduce and extend the analyses of the data used to support the 'Eve' hypothesis (Wilson and Cann 1992)—that of a recent, African origin for all modern groups. We conclude that, despite recent controversies, these and other molecular data are consistent with the hypotheses that *Homo sapiens sapiens:*

(1) is a very recent species (less than 200 000 years old);
(2) originated in a localized region of Africa; and
(3) close to 100 000 years ago spread out of Africa to replace all other hominids living in Europe (Neanderthals) and Asia (for example, the Solo specimens).

There is no evidence as yet of any interbreeding between modern humans and these other species, which in some areas (for example. Western Europe) appear to have become extinct shortly (perhaps less than 2000 years) after they came into contact with modern humans (Stringer 1990). This does not necessarily imply any direct interaction between species, such as warfare, but could result from indirect competition (Zubrow 1989).

The use of DNA sequences is now standard for inferring evolutionary relationships between species, though it certainly was controversial two decades ago. The use of sequences for studies of relationships within populations and species is newer, and still controversial in some quarters. Nevertheless, the scope and power of such studies is increasing rapidly, and we expect them to quickly become routine. A common theme we use is to build trees from different data-sets and compare the results to see if there is more agreement than would be expected to come about by chance (Penny *el al.* 1982). We conclude the chapter with a discussion of the advances in our understanding of human evolution that we might expect to achieve over the next ten years through the use of molecular data.

3.2 Reconstructing relationships: from DNA sequences to evolutionary history

We will outline the general biochemical approach to establishing evolutionary relations in the phylogeny of apes and humans (that is, hominoids). Using DNA sequences, the three steps in analysing the data are:

(1) estimating the separation (branching) pattern of all species of living hominoids, thus establishing an evolutionary tree;
(2) calibrating this tree, so that we can infer how many millions of years ago (m.y.a.) different lineages diverged; and
(3) placing statistical confidence limits on the estimates of divergence times.

3.2.1 *Basic steps in obtaining a tree for a selected stretch of DNA*

DNA is made up of ordered sequences of the four nucleotide bases that are abbreviated as a, c, t, and .g. The chromosomes of living hominoids contain approximately 3 billion bases, arranged linearly on 44 to 48 chromosomes (the number varies in different species owing to chromosome fusions and/or splitting). The chromosomes are in the nucleus of each cell, and contain the nuclear DNA. Chromosomes are inherited equally, but randomly, from both parents (excepting the male Y chromosome). In addition, there are just under 17 000 base pairs (b.p.) of DNA in the mitochondria (mtDNA). Mitochondria are organelles in the cytoplasm of cells, and are inherited solely from the mother via the egg cell. In the present context we are only interested in what DNA sequences can tell about evolutionary history. For our purpose a DNA locus (plural loci) is a contiguous stretch of DNA.

Our interest in DNA is not its function, but how we can use the changes (mutations) at particular loci to trace the evolution of the DNA, and hence to gain insight into the evolution of the species. Figure 3.1 shows a short piece of DNA sequence from a human mitochondrion lined up with the equivalent sequences from apes. During each cell division the total DNA of an individual is copied with great precision, although very occasionally (about once per billion nucleotide replications) a mistake is made.' This mistake, if it is either advantageous or effectively neutral (neither helping nor hindering an individual) may persist and spread in later generations. Almost all the substitutions occurring in the DNA are neutral (Nei 1987; Penny 1994).

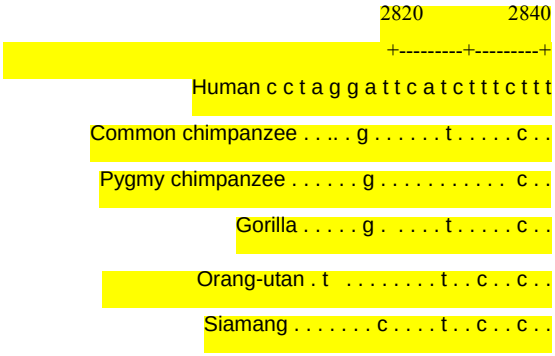

Fig. 3.1 An example of 21 nucleotide sites (base pairs) of aligned DNA sequences from hominoids (sites 2820-2840 from Horai *el at.* 1992). The convention of having a dot when a species has the same nucleotide as the first species allows patterns in the data to be seen more readily. The pattern of changes at site 2832 groups humans and chimpanzees together, while site 2835 groups humans, chimpanzees, and gorillas. There are also sites where only one species differs from the rest (for example, sites 2821, 2829, 2838); these provide evidence for the length of time since the divergence of that species. Site 2826 is an example of a site that must have changed at least twice. Maximum likelihood takes all these patterns into account when working out which tree best fits the data, and also allows a test of which, if any, tree model fits the data adequately.

Since neutral changes are essentially invisible to the processes of natural selection their occurrences can be modeled accurately using mathematical and statistical theory. A neutral mutation will usually disappear by chance (it is initially present in only one individual); but occasionally it may—also by chance—spread throughout a species.[2] Between the different hominoid species there has been sufficient time to ensure that most neutral mutations have by chance either become lost or become predominant within a particular species; when predominant, they are called substitutions.

It can scarcely be emphasized highly enough that treating changes as neutral removes one of the most important difficulties that beset earlier generations of researchers, namely, the difficulty involved in treating differences between humans as evidence for inferiority or superiority. The traditional European view of Natural Theology was that everything, including differences, must have a 'purpose'. This combined with the idea of a Great Chain of Being comprising a hierarchy of living forms led naturally to explanations involving value judgments about human differences. Nothing of which we are aware limits the capacity to coin such explanations to European cultures. Last century Darwinian evolutionary theory introduced probabilistic reasoning as a major concept in science; and this, and other aspects of the theory (Penny 1994) removed the need for assuming everything must have a purpose. The application of probabilistic thinking to evolutionary studies has increased, particularly over the past few decades, with the development of Kimura's theory of neutral evolution. Treating the vast majority of DNA changes as neutral therefore has the double advantage both of allowing more detailed mathematical modelling and of removing value judgments from the study of human variation.

When substitutions in DNA are relatively rare it is likely that sequences sharing a substitution are more closely related to each other than to others with a different base. Thus, for example, site 2832 of mtDNA as shown in Fig. 3.1 suggests that humans and chimpanzees are a group separate from the other apes, i.e. the site has a 'pattern' supporting a chimpanzee-human grouping. Occasionally, however, two or more individuals might share the same innovatory base at a site as a result of independent mutations. Such events can lead to erroneous conclusions about relationships if considered just by themselves, since some evidence can be found for almost any hypothesis! This problem of parallel mutation is much less likely to occur if the total amount of change in the DNA being examined is reasonably low; and, additionally, with longer sequences we get more accurate estimates of the true frequencies.

Evolution is a stochastic (probabilistic) process, and so the same change will occur on different lineages just by chance. Because of these multiple changes it can be difficult to get the correct tree directly from DNA sequences (see Fig. 3.2 for commonly used terms). A useful statistical criterion for deciding which weighted tree best fits the data is 'maximum likelihood' (Swofford and Olsen 1990). This assesses the likelihood of the observed nucleotide patterns given a weighted tree (Fig. 3.2 (II)) and the relative rates of substitutions (such as a —> c, a —> g). The 'weight' of an edge is the average number of substitutions per site expected on that edge. Computer programs are available that search for both the weighted tree and the mechanism of change that gives the best overall fit between the model and the data. The maximum likelihood criterion has the advantage of allowing confidence estimates on all parameters in the model. An important test is whether a tree model even fits the data adequately. A model may fail such a test for a number of reasons (including selection for certain changes); and in such cases we must be careful in placing confidence in the results. These tests may indicate that we need to consider more complicated models that include factors such as hybridization between different lineages.

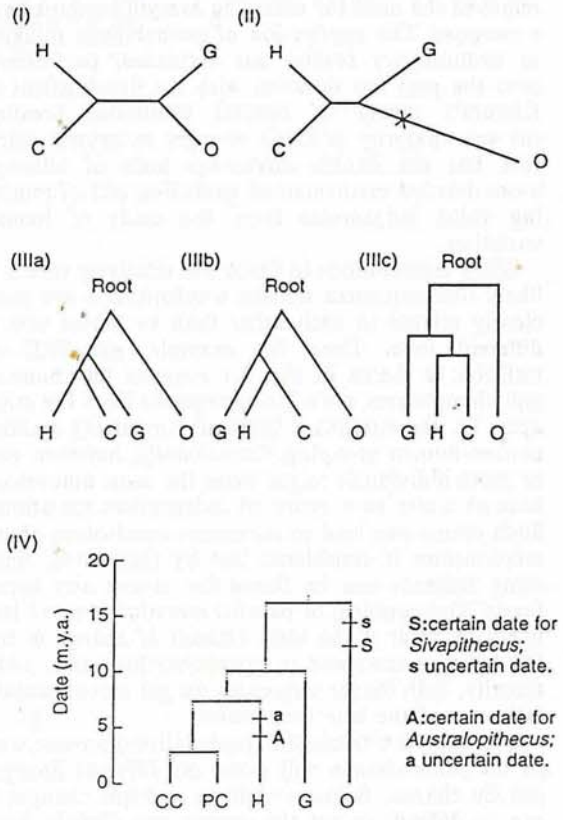

Fig. 3.2 Types of evolutionary tree:
(I) An unrooted tree (H = human, C = common chimpanzee, G = gorilla, O = orang-utan).
(II) An unrooted weighted tree, where edge (= internode) lengths are proportional to the amount of change (here DNA substitutions) on that edge. The X marks the position where the siamang joins this tree.
(IIIa) A rooted weighted tree; the tree in II rooted by the outgroup method (i.e. the point at which the sequence of a more distantly related species joins the tree). Here the siamang is used as the outgroup. **(IIIb)** The same rooted tree as in IIIa. **(IIIc)** The same tree again, but drawn so that the edge lengths are represented by only the vertical component. This is not intended to indicate anything different about the process of evolution, but rather, is useful in comparin edge lengths in a rooted weighted tree.
(IV) The maximum likelihood tree of 5 kilobases of mtDNA sequence, with the pygmy chimpanzee also included, calibrated by dating the origin of the edge leading to the orang-utan at 16 million years ago. CC is the common chimpanzee, while PC is the pygmy chimpanzee *Pan paniscus*. This tree is the optimal tree for the best-fitting maximum likelihood model (Kimura's 3ST mechanism of nucleotide change, with the rate of change across sites falling into two classes), with the constraint that the edge tips all meet at the same level (i.e. follow a molecular clock). The dates of the earliest known fossils attributed to specific lineages are shown, as well as the dates of less certain, more fragmentary, but similar remains (Groves 1989; Martin 1990; Campbell, this volume, Chapter 2).

Figure 3.2 (IV) shows the tree which best fits 5000 base pairs of mtDNA from apes and humans (Horai *et al.* 1992). One of the interesting things about this tree is that it groups humans and chimpanzees together to the exclusion of other apes—a conclusion that is consistent with other data-sets discussed below. The sampling error resulting from a finite sequence length is shown for each edge, and the error on even the shortest edge (human–chimpanzee) is clearly less than the support for that edge.[3] The weighted tree we have illustrated is the maximum likelihood tree, with the constraint of a molecular clock (see below) so that all tips meet at the same level (Fig. 3.2(IV)). This constraint results in a slightly worse fit of model to data (as is to be expected, because it has fewer adjustable parameters); but it is the type of model that is most useful when calibrating the tree. Further details of the maximum likelihood models used to analyse the data are given in a footnote to the caption of Table 3.1.

3.2.2 *Putting dates on a tree*

There are two ways in which we can add a temporal dimension to these trees: with respect to other species and using the 'molecular clock' hypothesis. The orientation of the hominoid tree with respect to time (called 'rooting' the tree, Fig. 3.2(III)) has been inferred using the knowledge that the Great Apes are more closely related to each other than to the gibbons (lesser apes), a view confirmed when trees are built from larger sets of species, including Old and New World monkeys, tarsiers, and lemurs. We will refer to this technique as the 'outgroup' method of tree-rooting.

The conclusion that a sequence of DNA will evolve at the same average rate in different species is called the 'molecular clock' hypothesis. The average rate is characteristic for each DNA sequence, and depends on the mutation rate and the number of constraints on the sequence. The 'clock' is a natural outcome of

Table 3.1 Estimated divergence dates of DNA sequences for humans, chimps, and gorillas, calibrated assuming an orangutan divergence date of 16 m.y.a. Because of polymorphism in the ancestral populations these dates will tend on average to be greater than the actual dates of species divergence (see text). The loci are mtDNA (Horai *el at.* 1992), and ψη) and γ-globin nuclear DNA loci (Bailey *el at.* 1992). The 'difference' column is the estimated time between the gorilla separation and the-subsequent splitting of the lineages to humans and chimpanzees. These sequences were analysed using a variety of maximum likelihood models based around the Kimura 3ST model, allowing the rate of substitution to vary between sites (Steel *et at.* 1993; plus see technical notes at the bottom of this table). The values for the DNA hybridization experiments designated 'DNA hybrid.-1' are maximum and minimum values from trees shown in Sibley *el al.* (1990) and constrained to fit a molecular clock. The values in 'DNA hybrid.-!' are for the clock-constrained tree estimated using a least-squares fit on the DNA hybridization data of Caccone and Powell (1989). Note the discrepancy of the inferred time of the lineages leading to humans and gorillas for these two data-sets, a feature which must make us cautious in over-interpreting the DNA hybridization results, at least until the reasons for it are understood. Values are in millions of years (values in parentheses are relative divergence dates, i.e. proportion of time back to orang-utan divergence). 'Chimp-chimp' values are for the divergence dates found for common vs Pygmy chimpanzees *(Pan troglodytes* vs *Pan paniscus).*

	human-chimp	human-gorilla	difference	chimp-chimp
mtDNA	7.2(0.45)	9.8(0.61)	2.6(0.16)	2.7(0.17)
gamma	6.5(0.41)	7.8(0.49)	1.3(0.08)	—
psi-eta	7.6(0.47)	8.2(0.51)	0.6(0.04)	—
DNA hybrid.-l	5.9(0.37) to 8.1(0.51)	8.2(0.52) to 9.1(0.57)	0.6(0.04) to 2.3(0.14)	3.0(0.19)
DNA hybrid.-2	7.2(0.45)	11.4(0.71)	4.2(0.26)	3.6(0.22)

Technical notes: Our evaluations with maximum likelihood included novel models assuming thai substitution rates of sites followed a gamma distribution (similar in shape to the lognormal). The general model used here is the extended Kimura 3ST model of Steel *el al.* (1993). All free parameters in all models were optimized using a Newion method. The 'logarithm of the likelihood of the data" is the sum over all site patterns of (observed frequency of the *i*-th site pattern multiplied by the natural logarithm of its probability under the model). The fit of model to data gives strong evidence that rates of substitution in the mtDNA sites vary considerably (due, no doubt, to stabilizing selection, i.e. many sites do not accept substitutions because they are functionally constrained). Here, the log likelihood ratio fit statistic (InLR or $G^2/2$) decreased from 391.6 to 151.6 when the gamma distribution was allowed (with the optimized shape parameter equalling 0.351). An 'invariant sites' model gave an even better fit, suggesting a demarcation between those sites which can *vs* cannot change (InLR 139.7 with 59.2% of sites assumed unable to change; the variable sites are mostly 'third' position sites, see Nei 1987). Interestingly, a mixed gamma-invariant sites model did not further improve the fit. We found no evidence to suggest unequal site rates in the nuclear 'non-coding regions'. For the mtDNA alone, the divergence times in this table are the average of 21 different submodels of the generalized Kimura 3ST. This allowed inference of errors due to choice of evolutionary model. These fluctuations in divergence times, relative to orang-utan, had range: human-chimp (0.38 to 0.53), human-gorilla (0.56 to 0.66), and chimp-chimp (0.14 to 0.21). Further details of these analyses are available from PJW (email: farside@ massey.ac.nz).

a probabilistic process where most changes are neutral (Penny 1994). Because changes are stochastic the 'clock' can only be an average rate; but it gives another way of rooting a tree, by estimating the midpoint in the tree (that farthest from the living species) and taking this as the earliest ancestor. The molecular clock hypothesis can be difficult to test without reliable outgroup rooting; but it is useful, as we will see in the later section on relationships within the human species. For the mtDNA data, and two other hominoid nuclear DNA loci that we analyse below, both outgroup and molecular clock methods of rooting are in agreement, thus providing further reassurance that the root is reasonably placed.

There are two reliable methods used to calibrate the branching points on a tree. The first is a well-dated fossil reliably associated with a particular lineage in the tree, preferably close to the origin of that lineage. In the case of hominoids the *Sivapithecus,* fossils known from Asia, with good fossils from Pakistan (Pilbeam1984), seem to fit this role, as may Dryopithecus from Europe (Solà and Kühler 1993). In the last twenty years better examples of these fossil apes, and a revision of systematics, have shown that they share a number of unique skeletal features with orang-utans *(Pongo pygmaeus).* This suggests that *Sivapithecus* was somewhere along the lineage leading to the orang-utan. Since these fossils are known to date back 12 million years, the point where the orang-utan edge joins the rest of the tree must indicate an event at least 12 million years old (Pilbeam 1984; Groves 1989).

Speculation about the time required for *Sivapithecus* to acquire its unique features, and the trend of the past twenty years of finding somewhat older *Sivapithecus* fossils, suggests

that the true date of the origin of the edge leading to the orang-utan was probably 2 to 6 million years older again. Thus fossil evidence suggests that the node where the orang-utan last shared an ancestor with the other Great Apes would be 16 m.y.a. ± myr (million years), consistent with Pilbeam's (1984) expectations. The most crucial assumption here is that *Sivapithecus* is most closely related to orang-utans amongst the living apes. Estimates of the relationships of fossil hominoids have changed much over the past twenty years, and it would be reassuring if the discovery of post-cranial bones of *Sivapithecus* supported its present placement.

Other fossils which can help calibrate the hominoid tree are australopithecine fossils on the edge of the tree leading to humans. At present these are dated back with certainty to 4 million years, and possibly even to 5.5 m.y.a. (Campbell, this volume. Chapter 2; Groves 1989). Figure 3.2(IV) shows these dates plotted on to the tree calibrated with the expected divergence time of the orang-utan. At present there are no other fossils that we can confidently associate with any other edge of the tree. Groves (1989) gives a useful overview of the status of fossils which, with better evidence, may eventually be assigned to particular parts of the hominoid tree.

A second way of estimating the date of a split in a tree is by a dated biogeographic event. An often-cited example is the separation of the lineages of the ratite birds ostrich (Africa) and rhea (South America) due to the continental rifting which caused the Atlantic Ocean to appear approximately 80 m.y.a. The opposite biogeographic effect occurred approximately 18 m.y.a., when the Arabian peninsula collided with Eurasia and allowed the biota of both areas to mix (Thomas 1985; Pilbeam 1984). Since the earliest fossil apes come from Africa it seems plausible that the ancestors of living and fossil Asian apes emigrated from Africa not earlier than 20 m.y.a. The evidence of other fossil groups emigrating from Africa to Asia at this time is consistent with the fossil dating given above for the orang-utan.

There are other methods for dating the branching points of trees, but these are less reliable than those just noted. Rates estimated by the above techniques in one group are extrapolated to those of another group. An example is to estimate an average rate of nucleotide substitution for different mammalian orders, using a diversification time of approximately 60-80 million years ago. This average can then be applied to any group of mammals that has a poor fossil record. Unfortunately, it is suggested that even within the mammals DNA substitutions may be as much as three times higher than the average in some groups, such as certain rodents (Nei 1987). Accordingly this approach is good for a ballpark figure, but has additional uncertainties.

The calibration of hominoid divergences has been attempted using DNA from a range of older primate divergences, such as: the divergence of Old World monkeys and hominoids; the earlier split of the ancestors of these two groups from the New World monkeys; and earlier events, going back to the supposed origin of primates (as estimated from fragmentary fossils). Such an approach has two major drawbacks:

(i) The suggested divergence times for the above events vary, because the earliest known fossils for a group may occur well after the origin of that group (Martin 1990); and

(ii) there is evidence that the rate of DNA substitution in hominoids, especially the larger ones, has slowed down in relation to that in other primates (Bailey *el ai.* 1992).

Both of these effects are expected to cause an underestimation of the divergence dates of hominoids. Together with sampling errors they largely explain why some published dates for the divergence of humans and chimpanzees from DNA sequences are too recent (as little as 3.3 m.y.a. in one case (Hasegawa *et at.* 1985)).

3.2.3 *Results from other molecular data*

In addition to the 5 kilobases of mtDNA sequence used above, there are sequences approximately 10 000 base pairs long for two regions of nuclear DNA, the yr) (psi-eta) and γ (gamma) globin loci (Bailey *el ai.* 1992). These two loci are contiguous, and form part of a region of about 100 kilobases known as the β globin gene cluster. Trees for these loci have been estimated using maximum likelihood methods based upon Kimura's 3ST (Nei 1987) model of nucleotide change, with the option of allowing the relative probability of substitution to vary at different nucleotide sites (Steel *at at.* 1993; Waddell. in preparation).

Statistical tests indicate that in all cases the fit of the tree model to the data is acceptable. Using a likelihood ratio statistic none of the other trees (including any tree grouping chimpanzee and gorilla. *Gorilla gorilla}* provide an adequate fit for the data. Thus three loci clearly favour humans and chimpanzees as closest relatives, with the gorilla being the next closest living relative. Results such as these have overturned the prevailing view of the last hundred years that the Great Apes are all more closely related to each other than any of them is to humans. Even though it is accepted that the African hominoids are our closest relatives, there is still some reluctance among morphologists to separate the knuckle-walking apes (chimpanzees and gorillas). As we show below, the molecular evidence is consistently in favour of the human-chimpanzee grouping.

Another set of data often referred to in studies of hominoids comes from the method of DNA hybridization. This estimates the overall nucleotide change between two species, but without determining the actual sequences. There are four published data-sets of DNA hybridization distances that include at least four living hominoids (Sibley *et at.* 1990). All favour the human-chimpanzee tree over the chimpanzee-gorilla tree, with one data-set (that of Caccone and Powell 1989) giving this tree over 99.7 per cent of the time in a statistical resampling procedure (Marshall 1991). Note, however, that different experimental procedures (hybrid-1 *vs* hybrid-2 in Table 3.1) can yield different results, sounding a note of caution in interpreting this type of data. In general, the DNA hybridization results are reassuringly consistent with results from sequence data, and, experimental errors aside, are expected to be indicative of the tree for the majority of DNA loci in these species.

Since we have reasonable dates on at least one node in our tree we can estimate dates for others. Table 3.1 shows these dates from three DNA loci on the basis of models assuming a molecular clock. In general they agree quite well. A notable exception is the estimated time from the divergence of the gorilla lineage to the separation of humans and chimpanzes (Table 3.1, column 4). This is a point of some interest to researchers; for example, if the combined human-chimpanzee lineage was relatively long humans may still share some ancestral characteristics with chimpanzees (perhaps language abilities or behaviour) that are not shared with gorillas.

3.2.4 *Polymorphisms and population variability*

A probable reason for the differences in these estimates of the divergence time from gorillas, apart from sampling error, is Molecular polymorphism. There are different, but related, sequences (alleles) in a population at any one time.[4] The degree of DNA polymorphism in a population is directly proportional to the long-term size of its population (also known as effective population size: Nei 1987). Consequently, at any time a population will have alleles that originated well back in the past, and these times will vary for different alleles. When a population subdivides, leading eventually to two species, there is a random component as to which of the ancestral alleles become prevalent in each population. Consequently we require many DNA sequences before we can confidently predict the exact separation time of species, and not just the earlier divergence times of alleles.[5]

The trees for the three loci discussed are also consistent with a number of shorter sequences. These loci, such as 28S rRNA and the associated spacer region (Gonzalez *el ai.* 1991), favour the human-chimpanzee tree, though individually they do not statistically reject the alternatives (in the mentioned example a claimed significant result is doubtful after we found four possible alignment errors in the original data). Less decisive data-sets, such as chromosome structure and allozyme frequencies, are also consistent to the limit of their resolution with the human-chimpanzee grouping. Given the molecular results, palaeontologists are reappraising the fossil data, which some now contend (for example. Begun 1992) are really most consistent with the human-chimpanzee grouping.

3.2.5 *Total error on estimated divergence times*

So far we have identified four independent sources of error on divergence times: fossil calibration; sequence length (sampling error); ancestral polymorphism; and the variety in methods and models that can be used to infer trees, edge lengths, and node times. If these errors are independent and additive we may estimate the total error on divergence times, since the overall variance from independent sources is then the sum of the individual variances. In this example all errors are independent, but some of them are multiplicative, and since they are such the total error we derive here will be an underestimate. (Later, in note 10, we show how to calculate some of these multiplicative errors in the context of dating the origin of modern humans.) We will describe the exact statistical error structure of molecular divergence times elsewhere (Waddell in preparation). We now illustrate these calculations with the divergence time of human and chimpanzee mtDNA. plus the additional step of inferring the divergence times of the actual populations. The standard error is the square root of the variance estimated from the sample.

1. The fossil calibration for the origin of the orangutan edge has a standard deviation of about 1 million years (myr), so the variance of this is $1^2 = 1$. But, since the human edge is only about half as long as the orang-utan edge, the relative error becomes ½ myr, giving a variance of $0.5^2 = 0.25$.

2. The standard deviation (due to sampling error) of the ratio of the height of the human-chimpanzee node to the divergence of the orang-utan lineage is approximately 0.05, which translates to ¾ myr (variance = $0.75^2 = 0.56$).

3. Different models of sequence evolution will also give slightly different ratios of edge lengths, and it is hard to be sure which will give the best estimate. In addition there are alternative ways of estimating divergence times without imposing a molecular clock. For these mitochondrial data the observed standard deviation of edge-length ratios due to these two causes together is equivalent to about ½ myr (variance = 0.25) (Waddell, unpublished).

4. While the polymorphism effect is still an unknown quantity, in our samples it may have introduced an average error of about 1 myr if the effective population size of the last human-chimpanzee ancestor was about the same as that of chimpanzees before human impact. If there was a similar amount of polymorphism at the origin of the orang-utan lineage then this effect is reduced by about a half, to about ½ myr (if the effective population size was constant for long enough, then the distribution will be approximately exponential, which implies the mean is equal to the s.d., so the variance = ¼ . This is an approximation.

Adding up all these variances we have an overall variance of 0.25 + 0.56+ 0.25+0.25= 1.31. The overall standard deviation of our estimates is then √1.31 or about 1.15, which is expected to be close to normally distributed, and consequently the 95 per cent confidence interval is ± 2 standard errors. Thus we estimate the time of human-chimpanzee mtDNA divergence at 7.2 m.y.a.. with a 95 per cent confidence interval of ±2(1.14), this is approximately 4.9 to 9.5 m.y.a. There is an expected upward bias of 0.5 to 1.0 myr due to ancestral DNA polymorphism, making about 6.2-6.7 m.y.a.[6] the most likely time for the divergence of the actual populations, but still with a standard deviation of just over 1 myr.

The results shown in Table 3.1 still have a rather large uncertainty in estimating divergence times (even if we could fix the divergence time of the orang-utan lineage exactly). While a refined estimate of divergence times will require the sequencing and analysis of more DNA loci, it appears most likely that the lineages leading to humans and chimpanzees diverged about 6 to 7 m.y.a. The population leading to gorillas probably diverged somewhere between 0.5 to 2.5 myr earlier again. Similarly, the evidence points to the divergence of the two chimpanzee species about 2.0 to 2.5 m.y.a.[7]

In conclusion, almost all the data collected so far are consistent with humans and chimpanzees being the closest relatives, and from the diverse data amassed it would be surprising if this view were overturned. The example illustrates the usefulness of molecular data when resolving what was probably a fairly closely-spaced series of population divergences. Even with molecular data quite long sequences are needed to have confidence in the results. The relative duration of an ancestral population leading eventually to humans and chimpanzees should become clear with additional sequences, and these are becoming available at an increasing rate. This will allow reliable inferences about some of the population dynamics of these ancestral species. Having established a reliable phylogeny we consider next what can be learned from molecular data regarding the origin and expansion of one particular species. *Homo sapiens sapiens.*

3.3 Human genetic data: mtDNA sequences

Here we look at the genetic evidence of the origin and interrelationships of humans. We consider the controversial findings from human mtDNA, and then how these results compare with evidence of genetic relationships from nuclear DNA.

3.3.1 *Out-of-Africa, or mitochondrial Eve*

The 'Out-of-Africa' (or mitochondrial 'Eve') hypothesis was the result of Cann, Stoneking, and Wilson's (1987) study using sequence markers on mtDNA to trace the maternal ancestry of 147 people from widely dispersed indigenous populations (see Wilson and Cann 1992 for a general review). The Eve hypothesis is a set of hypotheses (see below) proposing that all modern humans have a common maternal ancestor who lived in Africa around 200 000 years ago. Similar ideas had been developed earlier for nuclear-coded protein polymorphisms (Nei and Roychoudhury 1982). An opposing hypothesis, that modern human races evolved *in situ* from interbreeding populations of *Homo ereciiis* and its descendant populations (for example Neanderthals), is referred to as the 'multi-regional hypothesis' (Wilson and Cann 1992). Perhaps the most common misunderstanding of the Eve hypothesis is that there was just a single female in the population. However, calculations referred to below suggest a population of 1000 to 10000 females. It is probabilistic processes, referred to earlier under the neutral theory, that eventually lead to all mitochondria being derived from just a single female.

Because sequences are expected to be more informative than sequence markers for building trees, we will use the data of Vigilant *et al.* (1991). They used 630 base pairs from the fastest-evolving region of mitochondrial DNA (the origin of DNA replication, also known as the D-loop), and 135 different sequences were obtained. We will use this data-set both to explore the 'Out-of-Africa hypothesis' and to outline the techniques developed, and which need to be developed, to analyse such data fully. The 'Out-of-Africa' hypothesis is really a set of hypotheses, each one of which predicts a different characteristic of the tree of human maternal relationships. The hypotheses are:

1. All human populations can trace their maternal ancestry back to a common ancestor, who was surprisingly recent (possibly less than 200 000 years ago). Further, because of the effect of molecular polymorphism (described above), this date may be an overestimation of the actual age of the species *H. s. sapiens.*
2. The most probable ancestral region of modern humans was Africa, because the earliest divergences in the rooted tree seemed to have purely African descendants.
3. Most major populations (for example Asians, Europeans) show a number of distinct maternal lineages, suggesting that they were founded by populations of diverse individuals, and not by small, closely-related groups.

A more recently identified feature of human mtDNA trees that adds to the original hypothesis is:

4. Some time after the deepest divergences in the tree there appear many lineage separations, consistent with a rapid expansion in the size of the human population (Di Rienzo and Wilson 1991). This feature is noted especially amongst the non-African sequences: it is a feature predictable from 'Out-of-Africa', but not specifically predicted by the 'multi-regional hypothesis'.

The first two parts, the 'when' and the 'where', are the most critical in deciding which hypothesis (Out-of-Africa or the multiregional) is better. The second two parts, the 'who' and the 'how' of founding populations. add detail. All four of these parts are logically independent (a subset of them could be true and the others false), and each requires a specific type of test.

In conjunction with fossil evidence, the maternal mtDNA data implies some startling features in human evolution. There is good fossil evidence that *Homo erectus* occupied much of the

Old World (Africa and Eurasia) by 1 m.y.a. and after that differentiated into regional forms (such as Neanderthals in Europe). An age of only 200 000 years for the last common ancestor of all modern humans implies that only one of these *Homo erectus* populations has left any maternal descendants. In other words, a single geographically localized lineage of *Homo erectus* must have evolved into modern humans and then spread out to replace other living hominids. Molecular population geneticists are generally comfortable with such a hypothesis, because it is consistent with known processes of mutation and replacement. They do not have to appeal to unknown mechanisms or to mechanisms 'special' to humans.

Africa has been suggested as the ancestral area of humans by two criteria: (1) the location of the root of the human mtDNA tree, with its first lineages leading to large African branches; and (2) the fact that present-day African populations include the most divergent human mtDNA sequences (Vigilant *et al.* 1991). The mere presence of such large blocks of purely African sequences argues strongly against the multiregional hypothesis, which requires a large amount of interbreeding between all descendant populations of *Homo erectus* (Wilson and Cann 1992). Further, the very recent nature of all human mtDNA diversity makes even the whole of Africa look unlikely as the place where *Homo sapiens* first evolved, and suggests that the earliest origin of modern humans was in a part of Africa perhaps supporting a population of the order of 10 000 or even less (from the expected polymorphism of different population sizes under the neutral model). As yet there is no direct genetic evidence for the exact location.

Such major claims regarding the evolution of modern humans have not gone uncriticized (for example; Templeton 1993). There are criticisms directed against the reliability of trees obtained from such data, and also of hypotheses related to the location of the root (for example: Maddison, *et al.* 1992) and the date of the root (for example: Nei 1992) of the human mtDNA tree.

We will proceed as follows. First, we shall note some of the problems that are involved in trying to infer evolutionary hypotheses from the available data. Second, we shall briefly note the approach we are currently pursuing in a re-analysis of Vigilant *et al.'s* 1991 data-set. Third, we shall draw on our earlier discussion of tree-building and dating techniques (developed above with respect to the hominoids) to consider the status of the related hypotheses of when, where, who, and how. And finally, we shall consider other sources of molecular data, independent of the mtDNA sets, and attempt to link them to fossil and other data. This enables us to estimate the accumulated sources of error, and to establish the 'ballpark' within which the evolution of the hominid lineage occurred. We attempt, in the midst of a great number of sources of uncertainty, to indicate what seems to us, at this point, the best interpretation of the data available. For an opposing view, see Templeton 1993; but even here the critics of the Out-of-Africa model do not find support in DNA sequences for a multi-region model.

For this section, **we** shall introduce a new term: *branch.* A branch is a collection of edges in a tree emanating from one node, and, just as with a real tree, when you detach a branch you take with it all the edges further out from the root. As such, the term 'branch' is not interchangeable with the term 'edge' (internode).

3.3.2 *Problems with trees from large numbers of sequences*

The problems of determining the branching order and the rooting of human mtDNA trees are, for a number of reasons, extreme: the tree has many edges; the sequences are relatively short; only $_{1/3}$ of the sites have patterns directly useful in estimating the branching pattern; and the distance to the closest human out-group (chimpanzees) is approximately 20 times as long as the maximum distance between human sequences. Maximum likelihood methods of tree reconstruction slow down[8] as the number of sequences increases, and are as yet impractical for the types of study these data require.

Instead we have chosen to use the method of parsimony (Swofford and Olsen 1990), which searches for the tree requiring the smallest number of nucleotide substitutions (mutations). When the overall rate of change is small parsimony is expected to perform in a similar manner to maximum likelihood. A complication is that, with a similar number of sequence and sites informative to parsimony, there can be many trees with equivalent support. This is especially true with so many sequences (there are over 10^{260} possible trees for 135 sequences—compared with about 10^{70} elementary particles in the universe).

In such cases the fundamental questions the researcher needs answered are: do the optimal trees seem to be converging towards a common answer; and, if so, what are the general features of the best-fitting trees; are these trees consistent with other relevant but independent data

3.3.3 *Results from re-analysing the data*

To study the reliability of the analysis of Vigilant *et at.* (1991) we began searches for optimal trees from many random starting-points, using a recently developed search method (the Great Deluge algorithm. Penny *el at.* 1994) that has been shown to work well on other problems of similar complexity. The results of over 400 separate runs have been studied (Penny *et at.* 1995). By using measures of distances between trees and basic geometry, the locally optimal trees can be viewed as forming a single peak, implying that shorter trees are indeed converging to a relatively small subset of similar trees that we expect to be good estimators of the main features of the underlying tree.

Here we have chosen to infer the form of the true tree by finding the median tree (Fig. 3.3) that is, in a geometrical sense, the middle of all the best trees found. It turns out that this median tree is also one of the shortest trees on this data-set (one step longer than the shortest). Many of the edges in this tree have estimated lengths of zero, and for robustness we only show those edges which are supported by one or more changes. The tree has been rooted by the midpoint method,

which depends on a molecular clock. This tree is generally similar to those found by other tree-building methods (for example, the neighbour-joining method: Swofford and Olsen 1990). It is different from the original tree of Vigilant *et at.* (1991) in that it unites the! Kung.

Other interesting features are found in the tree. Some African sequences tend to form clusters according to ethnic origin, for example the !Kung and the Pygmies. Such features (assuming a random sample) probably indicate moderate population sizes and degree of isolation from other groups, rather than founding by a few closely-related individuals. The depth of these clusters indicates that these populations have existed for a relatively long time with respect to the depth of the root. Sequences from other African groups (the Yoruban, the Herero, and the Hadza) also form distinct clusters within larger assemblages of African sequences, but are also found intermixed with the Asian, European, and New Guinean sequences. Such a pattern is indicative of either a sudden population expansion or else a period of exponential population growth (Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Harpending *et at.* 1993). The relationships between African and non-African sequences is revealing; the Africans clearly form a super-set, with only some of the African clusters being found amongst non-Africans—which is in agreement with an African origin of *H. s. sapiens,* followed by expansion out of Africa by a subset of the total African diversity.

The positions in the tree of the African-American sequences are consistent with their being members of the African groups (excepting the Pygmies and the !Kung) recently displaced by the slave trade. Are the two Asian sequences in the midst of these otherwise African clusters similarly due to a slave trade or a mixed Afro-American/Chinese ancestry of Californian subjects.

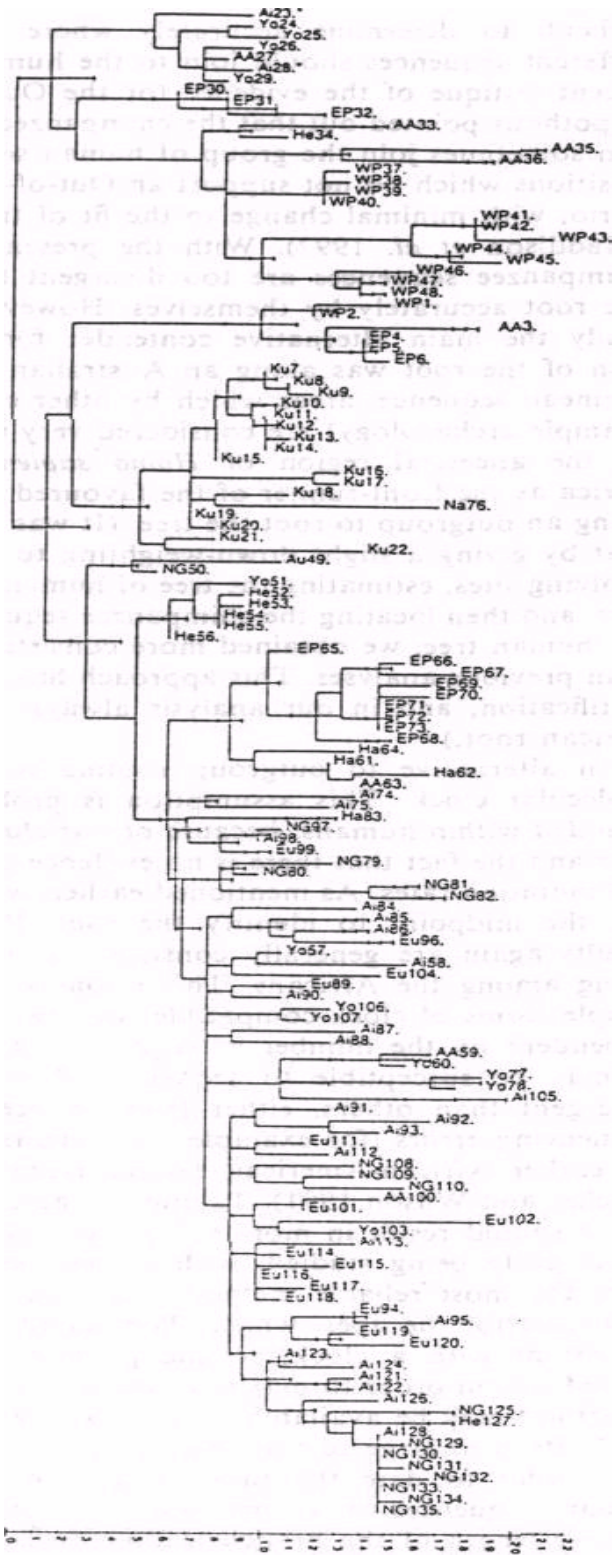

Fig. 3.3 Median tree from the re-analysis of the 135 human mtDNA sequences of Vigilant *et al.* (1991). This is the median tree of over 400 independent runs with the maximum parsimony criterion (as found by

If the time between the origin of *Homo sapiens sapiens* and major expansions in its range were reasonably close (say expansion was less than one-third the time after origin) then this will add to the difficulty in locating the root. In addition, populations such as the !Kung have possibly lost some of their original diversity, and this loss increases the difficulty of finding evidence for an early rapid expansion among African populations prior to the expansion of *Homo sapiens sapiens* out of Africa.

In general, we suggest that the above pattern is most consistent with an African origin, followed relatively soon by migrations out of Africa by people who possessed only a subset of the mtDNA diversity, resulting in the superset-subset relationship between the African sequences and those from the rest of the world. Some of the African groups, such as the !Kung, were possibly separate by this time, and may have had a relatively small effective population size since then. Alternatively, if the tree and its root are substantially correct, and the population size of the !Kung has not been so small as to lose mtDNA diversity over the millennia, then we have the suggestion that the !Kung were an early migration into Southern Africa from elsewhere. That our tree alone, of all those yet published, places the Naron (another group of San people) with the! Kung reinforces this possibility (also relevant to this hypothesis is Deacon (1992), who argues that humans in Southern Africa were quiet isolated from 125 000 to 10 000 years ago). This in turn suggests that Southern Africa was *not* the place of origin of modern humans. If this conjecture is correct then little by little we are whittling down the area in Africa which provides genetic evidence of being the place of origin of modern humans. Interestingly, since this analysis an associate has found very high mtDNA diversity in a small region near the border of Kenya and Ethiopia (E. Watson, in preparation), and this has fired these expectations further.

3.3.4 *When, where, who, and how*

We now look in more detail at these four aspects of the Out-of-Africa hypothesis.

3.3.4.1 *When and where*

It is difficult to date the trees of human sequences using the same methods for the human-ape trees, because the sequences of our closest living relatives, chimpanzees, are over twenty times more divergent than the greatest differences within human mtDNA. The short sequences presently available make it difficult to determine accurately where these very different sequences should join to the human tree. A recent critique of the evidence for the Out-of-Africa hypothesis pointed out that the chimpanzee sequences can sometimes join the group of human sequences at positions which do not support an Out-of-Africa scenario, with minimal change to the fit of tree to data (Maddison *el a!.* 1992). With the present data the chimpanzee sequences are too divergent to indicate the root accurately by themselves. However, in that study the main alternative contender for the location of the root was along an Australian

or a New Guinean sequence, areas which by other criteria (for example archaeology) are considered very unlikely to be the ancestral region of *Homo sapiens,* leaving Africa as the front-runner of the favoured regions by using an outgroup to root the tree. (It was interesting that by giving a slight down weighting to the fastest evolving sites, estimating the tree of human sequences first, and then locating the chimpanzee sequences onto the human tree, we obtained more consistent rooting than previous analyses. This approach has theoretical justification, and in our analysis always implied an African root.)

An alternative to outgroup rooting is using the molecular clock. This assumption is probably well founded within humans, because of our close relatedness and the fact that here is no evidence of differential mutation rates. As mentioned earlier, we can then use the midpoint to identify the root. Preliminary results again are generally consistent with the root being among the Africans. This is one of the most simple forms of clock-compatible rooting, and is not dependent on the number of sequences in a cluster. It may be susceptible to sequences that are more divergent than others, either from chance or from sequencing errors (for example the 5 errors found in an earlier African-American sample resequenced bv Kocher and Wilson 1991). Taking averages of sets edges should result in more robust rooting, and this is currently being studied, with similar results. Perhaps the most reliable method would be similar to that used for the apes, which allocates edge lengths, consistent with a clock, to the predetermined unrooted tree in order to maximize the likelihood. Such a method may be available in the next few years for application to large sets of sequences.

In order to date the time of divergence of the human sequences it is necessary to estimate the relative length of the branch leading from chimpanzees to humans, using the longest D loop sequences.[9] With data for 1000 nucleotide positions under the ideal model (all changes equally likely, all sites equally likely to change) the standard error of the distance measured will be at least 4 per cent. In reality, measuring the true distance from human to chimpanzee D loop sequences is more difficult; and, after taking into account factors such as variation of rates at different sites and different rates of substitutions, the relative error climbs to about 33 per cent! (Tamura and Nei 1993). We have to add to this the uncertainty of the exact time of divergence of chimpanzee from human mtDNA, which we estimated above to be about 5 to 9 million years ago once the different sources of error were taken into account.

The same statistical reasoning used earlier to estimate the total error of the human-chimpanzee divergence date can be applied to K.ocher and Wilson's data[9] in order to estimate the date of the deepest root in the human mtDNA tree by Tamura and Nei's (1993) method of measuring relative distances. We estimate the age of the human mtDNA ancestor from the D loop sequences to be 240 000 years ago, with a standard error of about 220000 years.[10] This is a substantially larger Standard error than that calculated by Tamura and Nei (1993) for the human mtDNA ancestor."

Recently, Hasegawa *et at.* (1993) have used approximately 300 of the fastest-evolving (third position) sites of a protein-coding region of the mtDNA (from Kocher and Wilson 1991), and have estimated the human mtDNA root using a maximum likelihood method to be 100000 years ago, with a standard deviation of 50 000 years. However, they took the divergence of human and chimpanzee mtDNA as 4 m.y.a., which seems too recent given the australopithecine fossils and the expected polymorphisms in the ancestral population. Their analysis of these data calculates the ratio of the root of human mtDNA sequences to the divergence of human and chimpanzee mtDNA as 1/40. with a standard error of 1/80. The expected divergence date, using a figure of 7.2 m.y.a. for human-chimpanzee divergence (with variance = 1.31), then becomes 1/40 x 7.2 m.y.a. = 0.18 m.y.a. with a standard error of 1/80 (so variance = $(1/80)^2$). The variance of this product estimated by the formula given in note 10 is $(1/40x1.31+7.2\ x(1/80)^2 + (1/80)^2x\ (1.31))$ which gives a standard error of 0.185 m.y.a. or 185000 years, which is close to that of the D-loop region.[12]

Because both of (he above estimates are for the origin of the same thing (human mtDNA), we can further improve our estimate of the date of the root of the human mtDNA tree as their weighted average (here for simplicity we ignore the weights since they are nearly equal). This gives (I 80 000 + 240 000) / 2 = 210000 years ago, and, because the standard errors of the two estimaies are about equal, the variance of the average is approximately halved, reducing the standard error by about 30 per cent, in this case about 150000 years. This date is a useful calibration point for other studies of human mitochondrial DNA, such as Harpending *et at.* (1993). Estimates of average mammalian rates of mtDNA evolution also give estimated divergence dates that are close to those obtained by the above two methods (Wilson and Cann 1992), increasing our confidence that we are not wrong in the basic tenet that the genetic evidence indicates that our species comprised a single small population less than 200 000 years ago.

Even though the variance of the estimated date of the deepest root in the human mtDNA tree is large, its answer to 'When?' gives strong evidence that *Homo sapiens* is a recent species derived from the descendants of *Homo erectus* on only one of the continents of Africa, Asia, or Europe. If *Homo sapiens* were derived from a mixture of intercontinental populations then some human mtDNA types should date back to before the time that *Homo erectus* colonized Europe and Asia, a time generally taken to be close to 1 m.y.a. This is a highly unlikely date given the present data, even with their large variance. The evidence from mtDNA so far has tended to rebut the multiregional hypothesis of human evolution as an adequate explanation.[13]

3.3.4.2 *Dating trees with archaeological evidence*

Archaeological evidence can also help give more precision in dating the human mtDNA tree. At least two major events promise to help here. The first is the arrival of people into Australia (which, owing to the lower sea levels of the last Ice Age, was then connected to New Guinea and some of Melanesia). Occupation of this region has been pushed back over 10000 years in the past decade, to approximately 50 000 BP, by archaeological finds (Jones 1989). Given that the archaeological sites almost certainly represent minimum ages. the first colonization of the Australian region could be about 60 000 years ago, with an approximate 95 per cent confidence interval being 50 000 to 70 000 years ago. This period also coincides with a period of minimum sea levels that would have reduced the largest sea crossing between South-East Asia and Australia-New Guinea to about 100 kilometres—still a major feat, however, with no evidence of its having been accomplished by *Homo erectus.*

A concern in using such an 'ingroup' dating technique is sampling enough sequences to be reasonably accurate in estimating the closest relatives among populations. Unfortunately, there are still few mtDNA sequences of Australians (including New Guineans) and South-East Asians. Accuracy would be enhanced if we could identify the oldest exclusively Australian assemblages of types, so as not to risk biasing the results towards a greater degree of Australian divergence due to the effect of polymorphism. While the present sample of Australian-New Guinean sequences is small (about 15 sequences) they tend to branch quite deeply, and often most closely with Asian sequences. A crude estimate of the depth of these branchings relative to the root is 2/5 that of the root. A similar picture emerges from the data of Cann *el at.* (1987), who had more Australians in their study, and estimated sequence divergence across the mitochondria using the observed changes in genetic markers.

A second event is the migration of people into the Ameircas; but there is still uncertainty on this dating, though it is generally accepted to have occurred 15 to 40 thousand years ago. Recently larger samples (72 and 63: Horai *et al.* 1992) of American Indian mtDNA have been sequenced, allowing more confidence in the true depth of indigenous mtDNA groups. These groups appear to be approximately 7 per cent of the depth of the root of human mtDNA, which is consistent with the first colonization of the America's 14 to 20 thousand years ago (though there could still be older immigrants not represented in the study). The results of these two studies, particularly that of the Australians, lend support to the hypothesis of a last common ancestor for human mtDNA 200 000-250 000 years ago.

3.3.4.3 *Who and how*

The diversity and distribution of mtDNA from people outside Africa (as shown, for example, in Fig. 3.3 indicates that the colonization of each of the main continents involved genetically diverse individuals, or possibly more than one wave of colonists. The mixture of Yoruban, Herero, and Hadza mtDNA lineages with those from outside Africa suggests that some of the ancestors of all these groups may have been among those who left Africa (possibly in contrast to more specialized peoples such as the Pygmy and the !Kung, who may already have been ethnically distinct at that time). Questions of 'Who' left Africa should become much clearer with further sampling of other African ethnic groups.

The final feature of the mtDNA tree we shall discuss pertains to the 'How' question of human population expansion. Nearly all the deepest lineages connecting Asians, Europeans, New Guineans, and some African sequences arose in a very short time. This feature has also been studied, using pairwise distances between sequences, by a number of authors (for example Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Harpending *et al.* 1993). While their methods do not yet have formal statistical tests, simulations performed by Rogers and Harpending show that the real data fit well with a rapid population explosion, but fit very poorly with a constant population size. The most rapid branching occurs at approximately half the height to the root, which, with present estimates of the time to the root, makes this diversification about 100000 years ago (and may coincide with modern people settling the Middle East and adjacent lands). Such a feature is expected under the Out-of-Africa model. Archaeological evidence fits this picture quite well, with some interesting punctuations. For example, modern humans did not colonize the bulk of Europe until about 35 000 years ago (a glacial period), although they may have been in the Middle East from 100000 years ago (Stringer 1990). It will be interesting to see if genetic evidence can help explain such mysteries. In addition, their analysis (Rogers and Harpending 1992; Harpending *et al.* 1993) also suggests that *Homo sapiens sapiens* probably evolved from a population with a breeding population of 1000 to 10000 females, suggesting a

total population of at most a few tens of thousands of humans. Such analyses help show the tremendous potential power of DNA sequences.

3.4 Trees of human relationships from nuclear genetic data

3.4.1 *Alleles and polymorphisms*

Nuclear DNA evolves more slowly than mtDNA (which has a higher mutation rate) and consequently nuclear gene sequences would need to be long (perhaps 10 000 base pairs each) to provide a similar amount of resolution of the branching patterns within humans. However, over the past thirty years a large data-base has been built up of the frequencies of different alleles of proteins for many human populations. The majority of differences in these protein alleles are due to the neutral evolution of the protein (that is, the stochastic replacement of one amino acid with another equally suitable amino acid, so that the protein continues to function quite adequately). More recently, additional alleles at different DNA loci have been detected by using enzymes which cut the DNA only at specific short sequences (4 to 8 base pairs) (Bowccock *el al.* 1991).

A general feature of these data is that there are significantly more allelic variants in Africa than in any other region of the world, and that the non-African populations appear as subsets of the diversity in Africa. These data do not rely upon tree-building, but parallel the situation found for human mtDNA: the most divergent forms of mtDNA are all African, with the non-African forms being derived from a subset of the deepest branches. To a non-specialist the concentration on amounts of diversity may not seem significant; but under population genetics models the amount of diversity increases as a function of time after a population has expanded in numbers, and so the amount and type of genetic diversity is a powerful indicator of population history.

After a population divides the relative frequencies of the various alleles change (become less alike), and we can thus measure the genetic distance between two populations. The rate at which the frequencies of neutral alleles diverge is also a function of breeding population size; but if population size remains fairly constant then the degree of divergence in allele frequencies is expected to be proportional to time. To keep the variance of such measures of genetic distance reasonable it is necessary to measure many loci (100 or more if possible).

Data on the frequency of alleles are also useful for estimating the phylogenetic relationships of human populations. If a sufficient number of genetic loci are mapped in large samples in different populations this allows polymorphism in the ancestral populations to be taken into account. Early results with allelic frequencies (including blood-group data, Nei and Roychoudhury 1982) argued for an African origin about 100000 years ago. Figure 3.4 reproduces the results of a later study undertaken by Bowcock *et al.* (1991). Their data-set is large, and the model used to estimate the phylogeny is one of the most detailed yet developed. Initial analysis of these data showed that a tree did not fit the data well (Bowcock *et al.* 1991); but they then modelled the possibility that each population was founded by a mixing of two others. Of the different possibilities only that of Europeans' being founded by a mixing of people with both Asian and African origins allowed the data and model to agree within statistical limits. The data are also consistent with a constant rate of evolution in all lineages in the phylogeny (not a tree, since it has rejoining or reticulate lineages), and these can be rooted by assuming a molecular clock. With trees based on allele frequencies the alternative of rooting or dating trees using chimpanzees as the outgroup is even less certain than in the case of mtDNA sequences, because most allelic variants become fixed (either becoming predominant or else disappearing altogether) during the time of separation of humans and chimpanzees.

3.4.2 *Ingroup dating of the tree*

Bowcock *et al.* (1991) dated the root of the tree in Fig. 3.4 by assuming that human populations first moved out of Africa 100000 years ago, on the basis of fossil evidence (which has been disputed) of apparent *H. s. sapiens* in Israel at about that time. We have recalibrated this tree by using an estimated time of divergence of Melanesians and Asians. This time should be nearly coincidental with the people crossing the sea channels to reach the greater Australian continent, an estimated 60 000 ±10 000 years ago (see above). To this date we will add 5000 years to allow for the separation between South-East Asian people (assuming they are the closest relatives of the Australian-New Guinean-Melanesian peoples) and the more northerly Asians who constitute part of the Chinese sample used here (see the tree figure in Cavalli-Sforza 1991). We will also raise the standard deviation to 7000 years to take into account some uncertainty as to exactly how much difference should be allowed for in using the more northerly Chinese population.

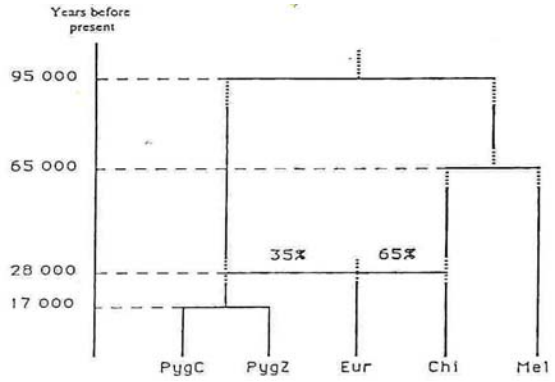

Fig. 3.4 A redrawing of the phylogeny of selected human populations as estimated by Bowcock *et al.*''^1) from the frequencies of DNA variants at 100 loci (PygC and PygZ = Pygmy populations in the Central African Republic and Zaire respectively, Eur = Europeans, Chi = Chinese, Mel = Melanesians). We have calibrated the phylogeny with an assumed date for the divergence between Melanesians and Asians of 65 000 years ago. The dots show the standard error on each edge resulting from sampling a finite number of DNA loci. The standard error for the estimated divergence time between African and non-African peoples (the root of this tree), after taking into account other known sources of error, is close to 20 000 years (see text). The percentages on the edges leading to Europeans (which they treat as a hybrid population) are the maximum likelihood estimates of the amount of genetic material contributed from each ancestral lineage.

When we recalibrate the tree this way two other dates in the tree agree with known fossil evidence of the spread of humans. The first appearance of fossils of modern human form in the Middle East is about 100 000 years ago; and the first appearance of modern humans in Europe (for example the Cro-Magnon specimen) is about 28 000-34 000 years ago (Stringer 1990). Note that this last date excludes the possibility that Neanderthals contributed substantially to the genetics of modern Europeans, as they are a unique lineage that evolved in Europe over at least 120000 years, and are now favoured as being in a line of descent back to the earliest European Homo erectus-like fossils of 500 000 or more years ago (Stringer 1990). That is, if they had contributed even 1/8 of the genetic material of modern Europeans, then modern Europeans should form a noticeably deeper edge in the phylogeny of Fig. 3.4. This is strong evidence that Neanderthals were a species distinct from Homo sapiens sapiens, but of uncertain biological status with respect to other lineages descended from Homo erectus. Thus an independent dating point not relying upon a cententious assignment of fossils to fully modern people, gives a similar result to Bowcock et al. (1991), and clearly supports the recent origin of modern humans.

There is also good evidence that, on the whole, modern Europeans (including Basque people) are closely related to peoples of the Middle East (for example, Iraqi, Iranian: see Cavalli-Sforza 1991) which could well have been a mixing place of the peoples moving between the South and East (Africa and Asia respectively). Such a hypothesis is also testable with mtDNA data. Notably, it does seem that more European sequences of mtDNA associate with either an African or an Asian sequence then African and Asian sequences are inferred as direct relatives (Fig. 3.3). although there are as yet relatively few sequences and it is not certain what effect sampling errors may be having. Many of the published trees derived from the genetic distances of protein alleles are in good agreement with these findings (Cavalli-Sforza 1991) despite not taking into account the mixing of populations (such as is implicated in the origin of Europeans and Polynesians, for example). We can now take a further step, and attempt to estimate the date for the movement of modern humans out of Africa. The date we have assigned for the divergence of Melanesian and Chinese populations is 65 000 years ago, with a standard deviation of 7000 years (variance = 7000^2). The ratio of the root of this tree (Figure 1, Bowcock el al. 1991) to the separation of Melanesians from Chinese is 100/68 = 1.47. Making the approximation that the errors in estimating the relative times of the root and the separation of Chinese and Melanesians are independent (with variances of 10.0^2 and 7.5^2 respectively; see Bowcock et al. 1991 and Fig. 3.4), then the variance of this ratio = $(100/65)^2$ x $(l0^2/l00^2 + 7.5^2/68^2) = 0.05$ (as given by the formula already used in footnote 10).

Thus the ratio of the divergence of African from non-African populations relative to the separation of Chinese and Melanesians is 1.47, with a standard error of ^0.05 = 0.23. The two numbers that make up this ratio are in fact positively correlated, which makes this a slight overestimate of the true variance. Our estimate of the Out-of-Africa event from this data-set is 1.47x65000 year =95500. As was described above, the variance of this last number is given by the formula for the variance of a product of two independent numbers, and equals ($65\ 000^2$ x 0.05 + 1.47^2 x 49 000 000 + 0.05 x 49 000 000 = 319 600 000, giving a standard error of about 18000 years. While we have not taken into account variability due to choice of model, this estimate is probably reasonably accurate, given the overestimate we made of the variance of the ratio we calculated.

This estimate and its variance (which is noticeably smaller than for the mtDNA estimates) clearly reject the idea of Homo sapiens sapiens being anything like one million years old, while the expansion out of Africa was almost certainly less than 140 000 years ago. Most importantly, this estimate comes from a random sample of over 100 of our DNA loci, making it highly improbable that it is atypical, something which is never so certain when studying a few loci (one, in the case of the mtDNA data).

The dates we have produced here from both the mtDNA and the nuclear data suggest that the Out-of-Africa event most probably occurred at 100 000 years ago. Such a date is very informative in the light of known fossil evidence. The first skulls with a distinctly modern aspect appear in Africa about 100000 to 120 000 years ago, and are preceded over the previous 200 000 years by skulls from throughout Africa sharing some unique features with modern humans (Groves 1989). About 100000 years ago quite modern-looking human skeletons are known from caves in the Middle East, where they were apparently contemporaneous with Neanderthal forms (Stringer 1990). While the exact nature of these modern-looking skulls is still in dispute (for instance. whether they are fully modern), the dating of the genetic evidence, so far, is consistent with the notion that they were amongst the first modern humans to have migrated out of Africa. There is no clear fossil evidence as yet that these two lineages interbred—another finding consistent with the genetic evidence.

Finally, there are now sufficient nuclear sequences to begin to make some statements about longer-term aspects of our genetic structure. Applying aspects of genetic drift theory to these sequences, they give a hint of the long-term population size of the lineage of hominids that led to modern humans. Recent analyses such as that of Takahata (1993) suggest that

the long-term effective population size of the hominid lineage leading to modern humans never fell below 10000 for any noticeable period. While this is just the beginning of a most interesting area of research, where more data and theoretical work is eagerly awaited, it highlights just how quickly genetic information is uncovering new sources of knowledge about the pattern and demographics of human evolution.

3.5 Conclusions and prospects

Molecular data have over the last thirty years elucidated many points of the evolutionary history of hominoids, though each problem considered was initially controversial. Sequence data have confirmed the findings of the early immunological explorations of the relationships of apes and humans: that the African apes and humans are the closest relatives; and that their divergence was much more recent than had previously been believed (5 to 10 million years, vs 20 to 30 million years). Such results have forced palaeontologists to reappraise their own assumptions about fossil relationships and to reconsider their methodologies. The most recent data and analyses most strongly support the grouping of humans and chimpanzees as the closest relatives, contradicting the apparent morphological similarities of gorillas and chimpanzees.

As we noted, the exact dating of the divergences of hominoids is still somewhat general, but importantly does not exclude any of the australopithecine fossils as hominid, and frames the origin of human ancestors in a 4.5 to 8.5 million-year period, despite the lack of any decisive fossil evidence. Claims of australopithecine fossils dating back to 5.5 m.y.a. challenge those molecular biologists who confidently estimate 4 m.y.a. as the human-chimpanzee divergence time, and highlight the importance of considering the compound uncertainties in calculating evolutionary dates.

It is only recently that relevant molecular data have been available for a large number of humans, and these have led to an argument for a very recent African origin of modern humans (Cann *et al.* 1987). First critiques of these data and the conclusions drawn from them were often relatively easy: biologists simply had not, and in many ways still have not, developed the techniques with which to analyse such large data-sets adequately. It was easy to criticize the original results for inadequate analyses without considering any alternative hypotheses or trying to integrate all the lines of evidence available. Reanalyses with appropriate techniques are now supporting the original claims when all the evidence and alternatives are considered. Balancing this there is also a need to consider the myriad possible sources of error in making quantitative estimates from molecular data. Molecular data do not stand on their own in the larger field of biological knowledge. Yet it remains true that the most powerful way to test the evolutionary ideas we have discussed here will be with the adequate sampling, sequencing, and analysis of other DNA loci.

Using a statistical framework we have outlined the major sources of error in reconstructing an accurate chronological phylogeny of ape and human evolution. We expect that within ten years the uncertainty from each of these sources will be more than halved. Given the trend of the past ten years we expect new fossil and archaeological finds that will improve the absolute time calibration of human-ape evolutionary trees, although the overall problem remains of assigning fossils accurately to lineages. There will also be a much greater number (20+) of independent DNA loci with which to build trees, while the methods expected to be available will allow more refined statistical estimates. By .that time we should be getting a clear picture of the long-term breeding population size of our distant ancestors, answering questions such as: 'How big was the population of our ancestors before it diverged into the lineages leading to chimpanzees and humans?' or 'Was the long-term effective population size of the hominids leading to humans really as small as 10000?'; 'Was it perhaps even smaller?'

Another area of interest is obtaining DNA sequences from (sub) fossil bones over 30 000 years old. Much effort is being put into this, especially to obtain verified sequences from Neanderthal bones that are not contaminated from handling by modern humans. This would allow a direct test of the hypothesis that the deepest divergence in the mtDNA of modern humans significantly postdated the divergence of non-African descendants of *Homo erectus.* It may eventually be possible to use DNA sequences from ancient bones to determine relationships amongst the populations of *Homo erectus.* Unfortunately there is evidence that the DNA in such old bones (unless frozen or else preserved in exceptionally dry conditions) is much degraded, making sequencing impractical with current techniques. The recovery of DNA from specimens in amber (now of *Jurassic Park* fame) is a different mailer: but humans have yet to turn up in this predicament.

We expect an even more profound understanding of human evolution to be exposed by molecular genetics over the next few decades. The human genome project will supply a huge amount of detailed information on the genetic structure of our own species and those of our closest living relatives. Combining this with advances in our understanding of developmental biology we may finally be able to identify which sets of genes regulate such features as body and brain development. Phylogenetic analysis of such sequences should allow us to estimate when such genes changed their function, and hence when, for example, areas of the brain associated with language evolved. It is no exaggeration to say that we will have previously unimagined insights into how we ended up being, well, human.

Epilogue

We take the opportunity to update the latest developments since the completion of the main manuscript in September 1993. There 'have been some exciting new developments, most of which reinforce our main conclusions. The whole mtDNA genome of all the Great Apes has been sequenced (Horai et al. 1995) and this again verifies both the closer relationship of human and chimpanzee sequences and that the gap back to the gorilla sequence is about 2 myr. The non-synonymous substitutions in this data set show very few multiple substitutions and seem to allow inference of the divergence dates without need of a specific model (although we would like the full out-group sequences of a gibbon to confirm that the orang-utan lineage is not evolving faster than the African hominoids). When we use the methods in this paper to estimate divergence times (plus total standard error) for mtDNA using just the non-synonymous substitutions, we arrive at the following divergence dates: human-gorilla, 7.6 myr (1 s.e. = 0.71), human-chimp 5.6 (s.e. = 0.60), and chimp-pygmy chimp 3.0 (s.e. = 0.44). Taking into account unknown ancestral polymorphism (using the assumptions already made in the text), the species divergence date is expected to be about 1/2 myr more recent in each case (and the overall s.e. will rise by approximately 0.25 myr in each case). The full sequence of a divergent African mtDNA by Horai et al. (1995) strongly supports the hypothesis that the root of the human mt DNA was less than 200000 years ago (although exact dating is still contingent upon reducing the other 3 main sources of error in making this calibration). Recent fossil finds have also been claimed to support a Dryopithicus (Sivapithecus, orang-utan) group (Solà, and Köhler 1993). This hypothesis also looks reasonable on biogeographic grounds, since dryopithecines are found in Europe, Sivapithecus near the Indian sub-continent, and orangutan in S.E. Asia. This reinforces our anticipation that orang-utan divergence was in the period 14 to 18 m.y.a., especially since dryopithecine fossils are known to date back to approximately 14 m.y.a.

There have been more papers showing evidence of the greater genetic diversity in Africa (e.g. alu elements, Batzer et al. 1994; mtDNA; E. E. Watson, in preparation). There are also analyses of nuclear sequence variation which are beginning to rival the lineage resolving power of mtDNA (Bowcock et al. 1994), and these are supporting Out-of-Africa. Recalibration of the age of some Javan fossils at nearly 2 myr old (Swisher et al. 1994), and new fossil finds in the near east (Gabunia and A. Vekua 1995) suggests Homo erecius spread out of Africa even earlier than assumed, making the multi-regional explanation of human origins even harder to defend. Overdue analyses are revealing that each chimp species has much more genetic variation than all humans (Morin el al. 1994), and similar results are coming to light for gorillas. This further bolsters the argument for humans evolving from a relatively small local population in the recent past, fully consistent with Out-of-Africa. On the theoretical front, there are signs that a variety of computationally feasible maximum likelihood models for estimating population histories (e.g. expected ancestral size, evidence of population expansion or migration) will be available in the next few years (Mary Kuhner and Joe Felsenstein, Bob Griffiths and Simon Tavare, pers. comm.). These should greatly help in the quantitative interpretation of past population events.

Substantial finds of fossils 4 to 4.5 myr old from Ethiopia have been assigned to a new species Australopilhecus ramidiis (White et al. 1994). Overall they appear chimp-like, and so far no characters amongst are conclusive in assigning this taxon to either human, chimp, or human-chimp ancestor lineages (although the base of the skull particularly appears to have some special features in common with at least one australopithecine species). More material is required and is rumoured likely to be reported soon. If this does turn out to be an early hominid we strongly resist it being assigned to a new genus. Most mammalian genera are at least 5 million years old. To have 3 or 4 named genera within the human lineage goes against the very truth revealed by the genetic studies of hominoid relationships—that humans are a recent group surprisingly closely related to the African apes.

Taken all together the evidence now seems to be favouring the period 4.5 to 6.5 m.y.a. more than the period 6.5 to 8.5 m.y.a. as the divergence time of human and chimp lineages. A conclusive resolution of the question exactly when, will require more sequences, better fossil calibrations of gene trees, and techniques to reliably infer the genetic variability of ancestral populations. -Overall, results relating to human origins are, as expected, accumulating at an ever increasing rate, with the human genome project yet to make its presence felt. Perhaps it is just as well that we learn more about our own past history, before we answer the next millennium's issues relating to modifying our future evolution.

Notes

1. A good analogy is the form of a surname, which can have related forms (e.g. Davey, Davis, Davies, etc.), each usually the result of a change that is passed on to direct descendants.
2. A good analogy is again given by surnames. If a family has only daughters (who to make the analogy strictly correct must marry to have children who inherit the husband's surname) that branch of the family name will die out. Conversely, a family may have all sons, and the family surname is then more likely to increase in frequency. These effects are often noted in small villages, where, after a period of time, many people end up with the same surname.

3. When a Chi-square goodness-of-fit test was performed on the observed and predicted nucleotide patterns the optimal tree model was not rejected. However the fit of all other possible trees was very poor (including having chimpanzees and gorilla *(Gorilla gorilla)* together), and accordingly we reject all alternative trees.

4. A good analogy for this effect is again the inheritance of family names (surnames). Envisage a situation where all the surnames in a town have evolved from one form; they are slightly different, but clearly related. The town was then divided into two parts, which were isolated from each other (by a dragon, a spell, or nationalistic armies). In each town one of the forms of the ancient name became predominant. A linguist then came along who knew how quickly names change in form, and deduced accurately how long it had taken for the two names to have changed from the ancestral form into their present forms. However she realized it was still not possible to estimate exactly when the two towns were separated, because the initial differences between the names probably predated the division of the old town.

5. In compensation there is the bonus that the distribution of the divergence dales from many different alleles will allow us to estimate the effective population size of all the nodes of the tree, including our distant ancestors. With the human genome project, which is also committed to sequence large stretches of DNA from apes for comparison with humans, data relevant to doing just this should be flowing in at an increased rate over the next decade.

6. These apparent divergence dates need not be minimal population divergence dates, because a large degree of polymorphism existing at the time the organ-utan lineage diverged could bias downwards the estimated times of later divergences.

7. Allowing 1 myr as the mean increase in divergence expected from DNA polymorphism in their ancestral population.

8. That is, the numbers of calculations required to estimate them are non-trivial, and could occupy even modern computers for long periods of time, and searching across the tree space could take years.

9. The sequences used here are from Kocher and Wilson (1991), who sequenced complete Dioop sequences (1135 base pairs) for about 20 individuals. They sequenced the DNA going in both directions, and we expect that the sequencing error rate was about 1 in 2000. This can be expected to inflate by about 5% our estimate (below) of the age of the root of human mtDNA.

10. (A) Estimated distance between the most divergent human lineages is 0.024, with variance $=0.006^2= 0.000\ 036$.

(B) Estimated distance between human and chimp mtDNA = 0.752, with variance = $0.224^2 = 0.0502$.

(C) Estimated divergence time of human and chimpanzee mtDNA is 7.2 m.y.a., with variance $1.15^2= 1.32$.

The ratio of the deepest human distances relative to human-chimpanzee distances is 0.024/0.722=0.033. The variance of this ratio is given by the formula $(\chi_1 / \chi_2)^2(\text{var}\ (\chi_1) / \chi_1^2 + \text{var}\ (\chi_2)/\chi_2^2)$ where *x* is the average and var(.v) its variance (Stuart and Ord 1987, p. 325), p. 325, which in our case is 0.000467, with the standard error being 0.021. Taking the time of divergence of human and chimpanzee mtDNA to be 7.2 myr (standard deviation = 1 m.y.a.), we have an estimated age for the human mtDNA ancestor of 7.2 myr x 0.033 = 0.24 m.y.a. The variance of the product of these two independent numbers $= \chi^2_1 x\ \text{var}\ (\chi_2) + \chi^2_2 x\ \text{var}(\chi|_1) + \text{var}\ (\chi|_1) \times \text{var}\ (\chi|_2)$ $=(7.2 \times 0.000467 + 0.033 \times 1.31 + 1.31 \times 0.000467)^{0.5} = 0.047$ which gives a standard error of 0.22 (Stuart and Ord 1987, p. 325), which equates to 220000 years.

11. This is due to omitting the interaction of errors on the two distances, in taking first the ratio, and then the errors on the product calculated above. We have not included the errors expected from the choice of model used for the distance measure (including the estimate of the variation of rates across sites) nor the difficulty in locating the exact root of the tree **of** human sequences.

12. Notice that when 300 sites were used in the above study the standard error was very similar to the standard error for the 630 sites from the D-loop, which shows the gain that can be made using sites that are evolving in a more predictable way. If the sequencing error rate is approximately 1 in 2000 this last figure is probably biased less than 4% upwards (expected error rate (1/2000) x sequence length (300) divided by average difference between human sequences in this region (4), times 100).

13. The sequencing of 5 kilobase stretches of the mtDNA from different humans could well reduce the standard error (due to finite sequence length) of the age of the last known human mtDNA ancestor to within 25 000 years. While such data may be available in the next couple of years, any further significant reduction in the region of error inherent in the above method of dating will require a significant improvement in the dating of the divergence time of human and chimpanzee mtDNA.

References

Bailey, W. J., Hayasaka, K., Skinner, C. G., Kehoe, S., Sieu, L. C., Slightom. J. L.. and Goodman, M. (1992). Reexamination of the African Hominoid trichotomy with additional sequences from the primate p-globin gene cluster. *Molecular Phytogeny and Evolution,* 1, 97-135.

Batzer, M. A., Stoneking, M., Algeria-Hartman, M., Bazan, H., Kass, D. H., Shaikh, T. H., *el a!.* (1994). African origin of human specific polymorphic Alu insertions. *Proceedings of the National Academy of Sciences (USA).* 91, 12288-92.

Begun, D. R. (1992). Miocene fossil Hominoids and the chimp-human clade. *Science.* 257, 1929-33.

Bowcock. A. M., Kidd. J. R., Mountain, J. L., Hebert, J. M., Caroienuto. L.. Kidd. K. K., and Cavalli-Sforza, L. L. (1991). Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. *Proceedings of the National Academy of Sciences (USA),* 88, 839-43.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd. J. R.. and Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature,* 368, 455-7.

Caccone. A. and Powell, J. R. (1989). DNA divergence among hominoids. *Evolution, 43,* 925-42.

Cann, R. L., Stoneking. M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature,* 325, 31-6.

Cavalli-Sforza, L. L. (1991). Genes, peoples and languages. *Scientific American,* 265, 72-8. Deacon, H. L. (1992). Southern Africa and modern human origins. *Phil. Trans. R. Soc. Land. B.,* 337, 177-83.

Di Rienzo, A. and Wilson, A. C. (1991). Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proceedings of the National Academy of Sciences (USA),* 88, 1597-1601.

Gabunia, L. and Vekua, A. (1995). A Plio-Pleistocene hominid from Dmanisi, East Georgia, Caucasus. *Nature,* 373, 509-12.

Gonzalez, I. L., Sylvester, J. E., Smith, T. F., Stambolian, D., and Schmickel, R. D. (1990). Ribosomal RNA gene sequences and Hominoid phylogeny. *Molecular Biology and Evolution,* 7, 203-19.

Groves, C. P. (1989). *A theory of human and primale evolution.* Clarendon Press, Oxford.

Harpending, H. C., Sherry, S. T., Rogers, A. R., and Stoneking, M. (1993). The genetic structure of ancient human populations. *Current Anthropology,* 34, 483-96.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution,* 22, 160-74.

Hasegawa, M., Di Rienzo, A., Kocher, T. D., and Wilson, A. C. (1993). Toward a more accurate time scale for the human mitochondrial DNA tree. Journal of Molecular Evolution, 37, 347-54.

Horai, S., Sana, Y., Hayasaka, K., Kondo, R., Inoue, T., Ishida, T., *et al.* (1992). Man's place in the Hominoidea revealed by mitochondrial DNA genealogy. *Journal of Molecular Evolution,* 35, 32-43.

Horai, S., Kondo, R., Nakagawa-Hattori, Y., Hayashi, S., Sonoda, S., Tajima, K. (1993). Peopling of the Americas, founded by four major lineages of mitochondrial DNA. *Molecular Biology and Evolution,* 10, 23-17.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K., and Takahata, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proceedings of the National Academy of Sciences (USA),* 92, 532-6.

Jones, R. (1989). East of Wallace's line: issues and problems in the colonisation of the Australian continent. In The human revolution (ed. P. Mellars and C. Stringer), pp. 743-82. Princeton University Press.

Kocher. T. D. and Wilson, A. C. (1991). Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding region. In Evolution of life (ed. S. Osawa and T. Honjo). pp. 391-413. Springer-Verlag, Tokyo.

Maddison. D. R., Ruvolo, M., and Swofford, D. L. (1992). Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. Systematic Biology, 41. 11 1-24.

Marshall, C. R. (1991). Statistical tests and bootstrapping: assessing the reliability of phylogenies based on distance data. Molecular Biology and Evolution. 8, 386-91.

Martin, R. D. (1990). Primate origins and evolution. Chapman and Hall, London.

Morin, P. A., Moore, J. J., Chakraborty, R., Jin, L., Goodall, J., and Woodruff, D. S. (1994). Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science* 265, 1193-201.

Nei, M. (1987). *Molecular evolutionary genetics.* Columbia University Press, New York.

Nei, M. (1992). Age of the common ancestor of human mitochondrial DNA. *Molecular Biology and Evolution,* **9,** 1176-8.

Nei, M., and Roychoudhury, A. K. (1982). Genetic relationship and evolution of human races. *Evolutionary Biology,* 14, 1-59.

Penny, D. (1994). Darwinian evolution; molecular evolution; molecular phylogeny. In *Encyclopedia of molecularbiology* (ed. J. Kendrew). Blackwell, Oxford.

Penny, D., Foulds, L. R., and Hendy, M. D. (1982). Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature,* 297, 197-200.

Penny, D., Steel, M. A., Waddell, P. J., and Hendy M. D. (1995). Improved analyses of human mtDNA sequence support a recent African origin for *Homo sapiens. Molecular Biology and Evolution,* 12(5) 863-82

Pilbeam, D. (1984). The descent of hominoids and hominids. *Scientific American,* 250, 60-9.

Rogers, A. R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution,* **9,** 552-69.

Sibley, C. G., Comstock, J. A. and Ahlquist, J. E. (1990). DNA hybridization evidence of Hominoid phylogeny: a reanalysis of the data. *Journal of Molecular Evolution,* 30, 202-36.

Solà, S. M. and Kohler, M. (1993). Recent discoveries of *Dryopithecus* shed new light on evolution of great apes. *Nature,* 3<i5, 543-5.

Steel, M. A., Szekely, L., Erdos, P. L., and Waddell, P. J. (1993). A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model.

New Zealand Journal of Botany (Conference Issue). 31, 289-96.

Stringer, C. B. (1990). The emergence of modem humans. *Scientific American,* 263, 68-74.

Stuart, A. and Ord, J. K. (1987). *Kendall's advanced theory of statistics.* Charles Griffin and Co., London.

Swisher, C. C., Curtis, G. H., Jacob, T., Getty, A. G., Suprijo, A., and Widiasmoro (1994). Age of the earliest known hominoids in Java, Indonesia. *Science.* 263, 1118-21.

Swofford, D. L. and OIsen, G. J. (1990). Phylogeny reconstruction. In *Molecular systematic* (ed. D. M. Hillis and C. Moritz), pp. 411-501. Sinauer Associates, Sunderland, Mass.

Takahata, N. (1993). Allelic genealogy and human evolution. *Molecular Biology and Evolution,* 10, 2-22.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution,* 10, 513-26.

Templeton, A. R. (1993). The "Eve" hypothesis: a genetic critique and reanalysis. *American Anthropologist.* 95**,** 51-72.

Thomas, H. (1985). The early and middle Miocene land connection of the Afro-Arabian plate and Asia: a major event for Hominoid dispersal? In *Ancestors: the hard evidence* (ed. E. Delson), pp. 42-50. Alan R. Liss, New York.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science,* 253, 1503-7.

Waddell, P. J., Penny, D., Hendy, M. D., and Arnold, G. C. (1994). Variance-covariance matrices for evolutionary trees using Hadamard transforms. *Molecular Biology and Evolution,* 11, 630—42.

White, T. D., Suwa, G., and Asfaw, B. (1994). *Australopithecus ramidus,* a new species of early hominid from Aramis, Ethiopia. *Nature,* 371, 306-12.

Wilson, A. C., and Cann R. L. (1992). The recent African genesis of humans. *Scientific American.* 266, 22-7.

Zubrow, E. (1989). The demographic modelling of Neanderthal extinction. In *The human revolution* (ed. P. Mellars and C. Stringer), pp. 212-31. Princeton University Press.