REVIEW

Genealogical and Evolutionary Inference with the Human Y Chromosome

Michael P. H. Stumpf[1] and David B. Goldstein[2]*

Population genetics has emerged as a powerful tool for unraveling human history. In addition to the study of mitochondrial and autosomal DNA, attention has recently focused on Y-chromosome variation. Ambiguities and inaccuracies in data analysis, however, pose an important obstacle to further development of the field. Here we review the methods available for genealogical inference using Y-chromosome data. Approaches can be divided into those that do and those that do not use an explicit population model in genealogical inference. We describe the strengths and weaknesses of these model-based and model-free approaches, as well as difficulties associated with the mutation process that affect both methods. In the case of genealogical inference using microsatellite loci, we use coalescent simulations to show that relatively simple generalizations of the mutation process can greatly increase the accuracy of genealogical inference. Because model-free and model-based approaches have different biases and limitations, we conclude that there is considerable benefit in the continued use of both types of approaches.

Genetic data increasingly augment linguistic, archaeological, and paleontological evidence in efforts to reconstruct the history of the human species. Over the past decade, the nonrecombining part of the Y chromosome has become a critical tool in the study of human evolution (*1–3*). The Y chromosome is inherited patrilineally (fathers to sons) and therefore carries information about the evolutionary past of males, complementing information carried by the matrilineal mitochondrial DNA (mtDNA) molecule.

The nonrecombining euchromatic part of the Y chromosome is almost 35 Mb long (*4*). Two randomly chosen Y chromosomes will differ on average at one nucleotide site every 3000 to 4000 bases. The Y chromosome therefore has an essentially unlimited supply of mutations that have been termed unique-event polymorphisms (UEPs) to reflect their low rate of occurrence. UEPs include both single-nucleotide polymorphisms and indels, and because they tend to have a unique mutational origin in samples of realistic size, they unambiguously define related groups of chromosomes, termed haplogroups (*5*). Together with rapidly evolving microsatellites (defining haplotypes), this makes the Y chromosome a uniquely powerful tool in the study of human evolution (*2, 3*). In addition to fine genealogical resolution, the Y chromosome also appears to show greater geographic structure than mtDNA and autosomal systems (*1, 2, 6–8*), although the data are not yet sufficient to fully assess the differences at different spatial scales. This pattern may result from greater female than male migration rates, as has been reported in some traditional societies (*7, 8*), but other explanations have been suggested. Whatever the cause, considerable structure in human Y-chromosome data has been observed in many parts of the world (*2, 6–25*).

The differences in structuring of different marker systems imply that certain aspects of demographic history may be detectable only in the paternal (or in some cases maternal) genetic record. For example, the Y chromosomes of the Basques, thought to be a relic Paleolithic population, show remarkable similarity to those of present-day Celtic-speaking populations (*24, 25*). However, mtDNA and X-chromosome variation suggest that the Celtic populations cluster with other North European populations, but with the Basques again distinct. One explanation for this difference is that the Paleolithic connection between the Celtic speakers and the Basques has been eliminated by female-mediated gene flow, and is now only observable in the paternal record.

There are two overlapping but distinguishable uses of Y-chromosome variation in the study of human evolution. In some cases the Y chromosome, or any other locus, may be used to infer population parameters, such as the growth rate of the population. It has been argued, however, that single-locus systems are not well suited to this purpose because (i) they represent only a single realization of the evolutionary process and therefore inherently lack statistical power, and (ii) selection may interfere with the expected correlation between the genealogy at a single locus and the demography of the population. Population parameters, therefore, are best estimated using data from multiple genomic regions (*26*). In other situations, however, the genealogy (Fig. 1A) may be of interest in its own right. This might occur, for example, in evaluating hypotheses concerning when specific lineages first spread geographically, or in assessing which populations have the oldest lineages at a given locus. For example, the consistent observation that the deepest branches in both the Y chromosome and mtDNA gene genealogies are found in Africa, together with the relative shallowness of these genealogies, has been taken as strong evidence in support of a recent African replacement model for the origin of anatomically modern humans. Genealogical depth may also be of interest in genetic evaluations of oral traditions. For example, Y-chromosome genealogical depth was used to assess the date of origin of patrilineal inheritance of priestly status among the Jewish priesthood (cohanim) (*5*).

Our primary focus here is inferring aspects of Y-chromosome genealogies, as opposed to the use of Y-chromosome genealogies to infer population parameters. However, except in unusual circumstances—for example, very rapid population expansion from small size—it is extremely difficult to draw direct connections between inferred genealogies and population processes (*2, 6*). In particular, casual equations between historical migrations and Y-chromosome lineages should be treated with considerable caution. Instead, genealogical data are better suited to test very specific hypotheses. For example, a test of whether genealogical structure corresponds to the geographic origin of sampled chromosomes can be used to assess whether lineages were or were not recently distributed through their geographic range (*15, 20*). Finally, we note that many applications require translating genealogical depth from generations into years, which introduces a significant source of uncertainty because of our ignorance of the life-styles of past human societies.

Methods used in genealogical inference can be conveniently separated into two groups: those based on explicit population models, and those that do not rely on assumptions about population history. Both approaches must assume a model for the muta-

[1]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. [2]Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK.

*To whom correspondence should be addressed. E-mail: d.goldstein@ucl.ac.uk

tion process, but only the former explicitly specifies the demography.

In the absence of an appropriate population model, genealogical parameters can still be estimated using approaches that do not explicitly specify a demography. For convenience we refer to these approaches as model-free, but this refers only to the absence of a population model. In this case, a sample of alleles is assumed to result from some genealogy (Fig. 1A), aspects of which are inferred from summary statistics of the observed data such as pairwise distance measures. Model-free approaches make only limited use of the data (*27, 28*) and do not use knowledge of population genetics. For illustration, we focus on three genealogical characteristics (Fig. 1A): (i) the time to the most recent common ancestor (TMRCA) of the whole sample, (ii) the ages of UEP mutations, and (iii) the TMRCA of the lineages within each (UEP-defined) haplogroup. All properties can be inferred using population models, but model-free approaches provide reasonable estimates only for a subset of properties; for example, in Fig. 1A, model-free approaches may be used to estimate TMRCA of the whole sample and of the alleles belonging to a particular haplogroup, but cannot give meaningful estimates of property (ii) above.

Model-Based Approaches

In model-based approaches, in addition to the observed data, a specific population model (or class of models) is assumed to be responsible for the ancestral history of the sample. Model-based approaches may be used in a Bayesian or likelihood (frequentist) framework. As the former may be less familiar, we briefly review its structure. The ancestral relationship is understood to be specified by a set of random variables, Ω, for which a prior (i.e., a prior probability distribution function) can be set independently of the sample—for example, by coalescent theory for aspects of the genealogy, or often by assumption in the case of the mutation rate. Population genetics theory is then used to estimate the likelihood of the observed data as a function of the values of Ω. In realistic settings, these likelihoods must be estimated through numerical simulations. For example, importance sampling or Markov chain Monte Carlo procedures may be used. With the likelihoods determined, Bayes theorem can then be used to combine the prior and the likelihoods into a probability distribution for Ω known as the posterior (*29*). Point estimates and confidence intervals (CIs) for all relevant quantities follow from this posterior distribution.

Determination of priors and likelihood surfaces, however, require assumptions about the population that is being modeled. For example, if the underlying model assumes size constancy, then a growing population is not represented adequately. The assumption of constancy would create a strong bias in favor of long branches toward the root and short branches at the tips, whereas rapidly growing populations have the opposite tendency. Unlike population growth (*30*), the effects of population structure are less predictable, and geographically structured models are difficult to justify. In the long term, model-based methods may permit evaluation of the most appropriate representation of human population structure by comparing the support for alternative models, and steps in this direction are already being taken. For example, programs such as FLUCTUATE, GENETREE, and BATWING (*31–33*) all allow for evaluation of the growth model, whereas MIGRATE permits estimation of the effective migration rate between populations (*31*). Given our current ignorance of the most appropriate models to represent human demography, however, continued use of the available inferential methods depends critically on thorough evaluation of the consequences of model misspecification. In particular, methods assuming available models (e.g., migration among N populations at mutation-drift equilibrium) should be applied to simulated data generated under a range of more complicated models in a direct evaluation of the consequences of model specification.

DNA sequence divergence. Focusing on sequence variation, Thomson *et al.* (*34*) used GENETREE (*32, 35, 36*) for Y-chromosome genealogical interference. The GENETREE program runs coalescent simulations, assuming either a stationary or growing population, and estimates a likelihood distribution for genealogical parameters by averaging over a large number of runs. Estimates of absolute genealogical time are obtained using an external estimate of the point mutation rate, in this case inferred using the sequence divergence of chimpanzees and humans and assuming a separation time of 5 million years. It is worth noting that this procedure ignores the effect of lineage divergence within the ancestral population giving rise to humans and chimpanzees, and that the statistical properties of the estimated mutation rate have not been properly evaluated.

Microsatellite data. A similar approach can be taken using microsatellite rather than sequence data (*37*). Here the mutation model presents a particular challenge: Whereas coalescent simulations for DNA sequences generally assume an infinite-sites or infinite-allele model (*38*) where back-mutations can be

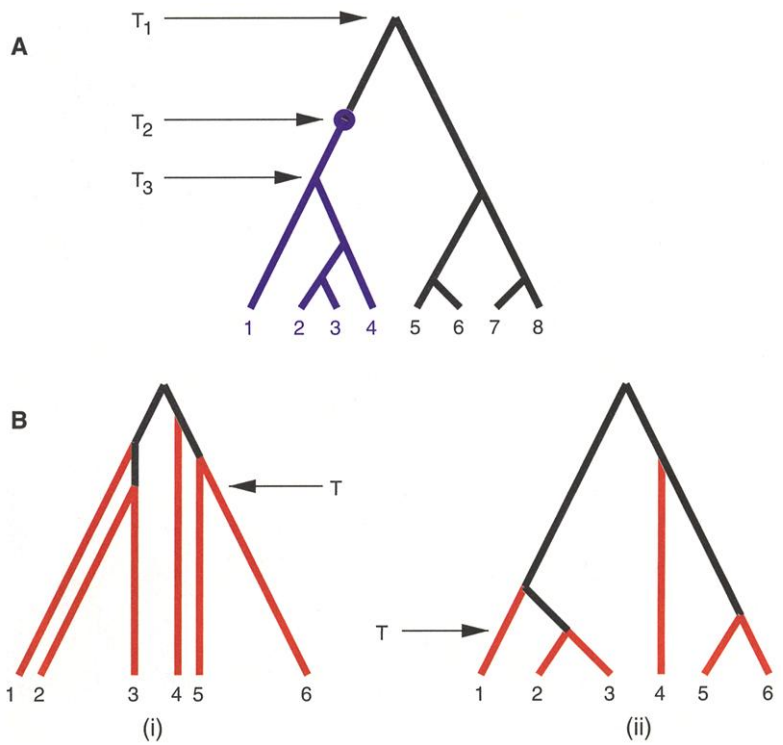

Fig. 1. (**A**) Genealogy of a sample of eight chromosomes with total depth T_1. Chromosomes 1 to 4 belong to one haplogroup defined by the UEP that occurred at time T_2; their most recent common ancestor occurred at time $T_3 < T_2$. (**B**) Genealogies for growing and constant populations. Uncorrelated lineages are shown in red; black branches trace the evolution of two or more present-day chromosomes. T indicates the time during which all lineages in the genealogy evolve independently. Estimates derived for the uncorrelated tree (i) will have narrower CIs than those for the correlated tree (ii).

ignored, the high mutation rate and the stepwise mutation process rule out such a representation for microsatellites. The microsatellite allele lengths at each node in the genealogy are now also members of the random variable Ω, in addition to N_e and μ. Thus, the generalized microsatellite mutation parameter is $\mu\sigma^2$ (*39*), where σ^2 is the variance of the distribution of step sizes, and the height and shape of the genealogy can be investigated simultaneously in this framework. Using this approach, Wilson and Balding (*37*) evaluated the genealogical depth of human Y-chromosome genealogies. Under conditions where the assumptions of the population model are met, the quality of the estimate increases with the number of loci included in the study. When large sample sizes are used, the computational cost is a serious constraint.

Model-Free Approaches

Because of our present ignorance of human demography and how to represent it, it is necessary to compare model-based approaches with those that do not make assumptions about demography in genealogical inference. We discuss approaches that only assume the existence of a genealogy and that seek to estimate features of the genealogy from observed diversity.

Microsatellite repeat variance and genealogical depth. The squared difference between the lengths of two sampled alleles (l_i and l_j) under a stepwise mutation model (SMM) has the expected value

$$E[(l_i - l_j)^2] = 2\mu\sigma^2\tau \qquad (1)$$

(*39, 40*), where τ is the coalescence time of alleles *i* and *j*. Equation 1, evaluated for growing populations, is a reasonable estimator for TMRCA because the average pairwise coalescence time is close to the time back to the most recent common ancestor (*10*) (Fig. 1B).

A variation of this approach, assuming a highly idealized demography, was developed (*19*) to study the situation in which a selective sweep brings a variant quickly to fixation, followed by a return to mutation-drift equilibrium in a now constant population. With a sufficient marker density, this approach would also provide a framework for detecting selective sweeps anywhere in the genome. However, because of the very specific assumptions of this approach and that of Eq. 1, neither is suitable as a general method for dating Y-chromosome genealogies, as they have sometimes been applied.

Ancestral haplotypes and present variation. Irrespective of past population demography, TMRCA can be estimated if the haplotype of the most recent common ancestor can be inferred (*5, 41*). For example, if we know the allele lengths of the haplotype at node 2 in Fig. 1A, then we can calculate the estimated time from nodes 1 through 4 back to the node at time T_2.

For a single locus, the squared distance is given by

$$\Delta_i = (l_i - l_a)^2 \qquad (2)$$

where l_a denotes the ancestral allele length and *i* refers to a present chromosome. Summing Eq. 2 over the *N* present chromosomes yields the averaged squared distance (ASD),

$$\Delta = 1/N\Sigma_i (l_i - l_a)^2 \qquad (3)$$

The squared difference in allele size, Δ, has the expected value

$$\Delta_A = \mu\sigma^2\tau \qquad (4)$$

where Δ_A is the ASD to the ancestral (A) allele and τ is the genealogical branch length separating allele *i* from the most recent common ancestor of the sample. The average over all alleles, Δ_A, is thus an unbiased estimator for TMRCA. In practice, Eq. 4 would be evaluated for each of many loci and averaged. We note that the ASD was originally defined as an estimator of separation time between populations, but the related formulation (Eq. 3) can be used to estimate coalescent times between pairs of sampled alleles, which underlies both the interpopulation formulation (*42*) and Eqs. 3 and 4 (*5, 19, 43*).

Performance of model-free approaches. Model-free approaches have been widely used (*5, 10, 15, 16, 19, 41*), and here we evaluate the performance of the method described above for estimating genealogical depth. To evaluate model-free approaches, we use coalescent simulations with parameters designed to mimic the human Y-chromosome genealogy. We distribute both UEPs and microsatellite mutations throughout the genealogy in order to simulate what is typical or achievable in real studies. In Fig. 2A, we show that Eqs. 3 and 4 yield very good date estimates for coalescence events for the simple SMM (for the strict, single-step mutation model with $\sigma^2 = 1$ and constant μ). Increasing the number of loci from 20 to 30 results in a modest decrease in the mean square error, but using fewer loci would normally lead to

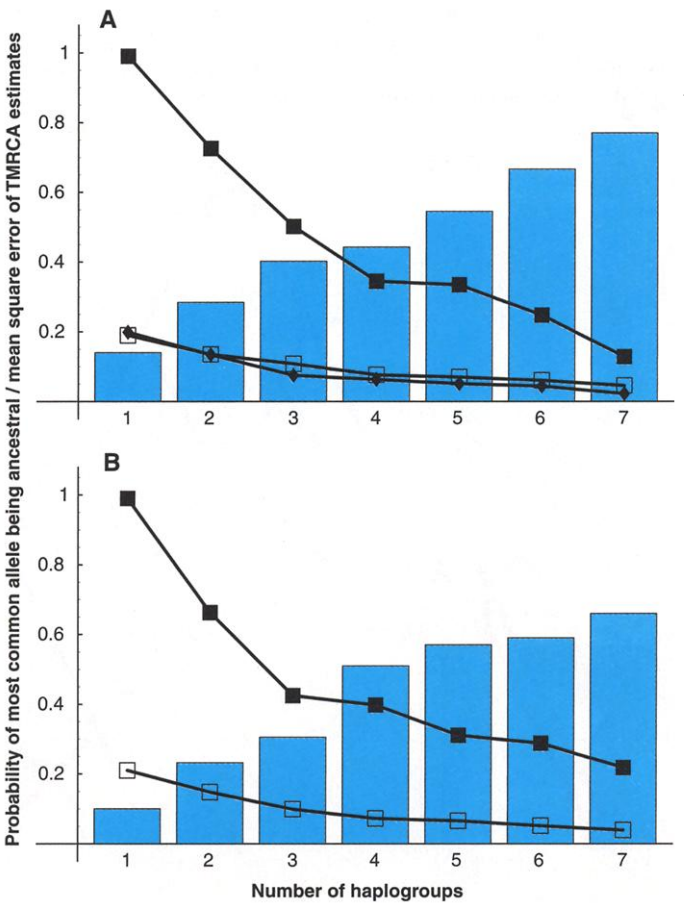

Fig. 2. (**A**) Probability of the most frequent alleles in a haplogroup being ancestral for increasing numbers of haplogroups defined by UEPs. Increasing the number of haplogroups leads to significant improvement in the identification of ancestral alleles. Also shown are the mean square errors of the haplogroup age estimates using 20 microsatellite loci and the true (open squares) and the inferred ancestral allele (closed squares), respectively; diamonds represent estimates obtained for 30 loci and with the true ancestral allele. The error decreases with increasing number of haplogroups and microsatellite loci. Genealogies of a sample of 200 chromosomes were simulated and mutations distributed following the microsatellite mutation process with constant $\mu = 0.0028$ across 20 loci; numbers were averaged over 500 independent runs. (**B**) As in (A), except that for repeat length *l* we now assume a length-dependent mutation rate $\mu = (-6.62 + 0.62l)\,k$, where *k* is an arbitrary constant (*46, 48*). Identification of ancestral haplotype from most frequent alleles is now slightly less reliable, as larger alleles have higher mutation rates; the mean square error of the haplogroup age estimates shows a similar dependence on haplogroup age (increased number of haplogroups results in shallower haplogroup genealogies), as in the constant μ case.

significant increases in the mean square errors (*44*).

In addition to the mean square errors of the estimates for TMRCA using the real ancestral alleles in Eq. 3, we also show the errors that result if the most common allele is assumed to be ancestral. In the analysis of real data, the ancestral haplotype is not known from the outset, and inferring ancestral haplotypes from the most common alleles can be problematic—especially for stationary populations, where most mutations accumulate along the long ancient branches of the genealogy. Expanding the number of UEPs, and hence the number of haplogroups, greatly increases the reliability of inferring ancestral haplotypes (Fig. 2A) from the most common alleles.

Model-free approaches and confidence. Although likelihood approaches include a natural framework for assessing confidence, the difficulty of estimating CIs is a serious limitation for model-free approaches. Moreover, point estimates (e.g., for TMRCA) may be unbiased regardless of the details of the genealogy under consideration, whereas the corresponding CIs depend strongly on the shape of the genealogy. Heuristic arguments for obtaining CIs in a model-free setting are given in (*45*). In the case of star genealogies with n sampled chromosomes, estimation of CIs is straightforward; there are roughly n independent random variables (Fig. 1B) and, irrespective of the process generating differences among lineages, the problem can be simulated or calculated for n independent lineages. For example, for the SMM, for each lineage there will be a Poisson-distributed number of mutation events, whereas the number of repeat size-increasing (and size-decreasing) mutations shows a binomial distribution; it is thus possible to simulate the mutation process on n independent lineages and calculate approximate CIs from the distribution of outcomes. In a correlated genealogy, however, CIs will be much wider (Fig. 1B). One approach for estimating how much wider (*45*) is to assess the effective number of lineages—that is, the number of lineages in an uncorrelated tree that would have properties similar to the correlated genealogy. It should be noted that, ignoring an extreme bottleneck, the true CIs will generally be bounded by the two limiting cases of constant and growing populations.

Estimating Genealogical Depths with Model-Free Methods

Uncertainty concerning the mutation rate and process presents a serious limitation for model-free approaches. In the case of microsatellites, deviations from the SMM (*39, 40*) could result in substantial biases that are hard to detect in model-free analyses. Important possible deviations from the SMM include variable step size ($\sigma^2 > 1$) that has been observed at low frequency (*46*), directional bias in the mutation process (*47*), length dependence in the mutation rate (*48*) and step size, and a dependence on the size of the repeated motif. Finally, it is clear that microsatellite allele length is constrained, in part as a result of the mutation process (*49, 50*), and this will influence the dynamic of distance measures.

Length-dependent mutation rates. A general mutation parameter may thus be written in the form

$$\mu = \mu(l)\sigma^2(l) \quad (5)$$

(*39*), where $\mu(l)$ and $\sigma^2(l)$ are the mutation rate and variance of step size, respectively. In the simplest instance of a length-dependent mutation process, the functional form will be linear; quite generally this form also describes more complicated functional dependencies to first order:

$$\mu(l) = \mu_0 + \mu_1 l \quad (6)$$

To describe length dependence, we use the results of (*48*) for a set of 10 microsatellite loci in a large sample of Y chromosomes. This mutation rate model was implemented in the coalescent simulations as follows. In the generation of the simulated data sets we use $\mu = 0.0028$ but assume that this corresponds to the average allele length (i.e., $l = 16.96$) in (*46*), and we adjust the mutation rate after each mutation event according to the new repeat length. When we estimate the genealogical depth of a haplogroup from Eq. 4, we calculate a length-adjusted mutation rate for each locus from Eq. 6 by using the average repeat size at that locus in the haplogroup. This procedure yields reliable estimates for TMRCA (Fig. 2B). If, however, the mutation rate is assumed to be constant, estimates for TMRCA may significantly deviate from their true values. As expected, accuracy again increases with the number of loci.

Y-chromosome genealogies. If the mutation rate depends on repeat count, this can have quite profound consequences for estimates of genealogical depth (Table 1). This is also a problem for model-based approaches where the initial choice of an underlying population and mutation model may bias the results. For example, deviations from constant μ could be misinterpreted as resulting from a different genealogical depth. In Table 1, we compare estimates of TMRCA for chromosomes sampled from several continents and belonging to four haplogroups (*48*); we find that a length-dependent mutation rate can have important consequences for estimates of genealogical depths. For example, we find that for hg10, including length dependence yields estimates that are more than twice as old as the estimate with constant μ. Differences are even greater for individual loci. The rate for locus DYS388 in hg9 is predicted to be more than 3.5 times the mutation rate at the same locus in hg4. Because Kayser *et al.* (*46*) observed only a single mutation event for simple (i.e., perfect repeats) microsatellite loci, our calculations should be understood as illustrative only. To account for length dependence properly, it is furthermore important (i) to have more conclusive data to estimate μ for simple loci, and (ii) to use locus-specific length dependencies. It is therefore important to further elucidate the microsatellite mutation process and parameterize the SMM such that Y-chromosome data can be interpreted more reliably. As the differences can be expected to be significant if few loci are included, length dependence presents another reason, in addition to inherent stochasticity, to be extremely cautious regarding time estimates based on only a few loci. For these reasons, we consider estimates made on fewer than 10 to 20 loci to be unreliable; as far as we know, all estimates of Y-chromosome genealogical depth published to date fit into this category.

Many questions of interest in human evolution require secure genealogical data for the Y chromosome and other genetic regions. For example, appropriate comparison of male and female patterns of movement requires comparison of Y chromosome and mtDNA lineages of comparable time depths, whereas many questions in anthropology require assessment of the spatial distribution of lineages of known origin (*1, 2, 6, 24*). For this reason, most research studies in human evolution include estimates of genealogical depth. Currently even the more elaborate population models, however, are clearly insufficient descriptions of human demography; this limits the range of questions that can be asked and creates unknown biases. The relative merits of model-based and model-free methods will obviously depend on the quality

Table 1. Ages (in generations) of the most recent common ancestors of alleles belonging to haplogroups hg3 [Var(l) = 0.256], hg4 [Var(l) = 0.316], hg9 [Var(l) = 0.467], and hg10 [Var(l) = 0.277]. Ages (in generations) were determined from Eqs. 3 and 4, assuming constant (column 1) and length-dependent (column 2) mutation rates. Also shown are estimates obtained excluding the six simple loci DYS19, DYS388, DYS392, DYS393, DYS425, and DYS426 without (column 3) and with (column 4) consideration of length dependence. Including length dependence can change estimates for TMRCA by more than a factor of 2, although for individual loci estimates can differ by much more.

Haplogroup	1	2	3	4
hg3	115	214	71	225
hg4	136	215	45	179
hg9	211	319	141	345
hg10	116	265	66	296

of the match between the model used and the population under study, and it is difficult to provide generic guidelines. In some cases—for example, populations distributed across island systems and clearly not at equilibrium—the mismatch between model assumptions and reality is so great that it is hard to see the advantage of using the currently available model-based methods. Even in cases of less obvious violation of model assumptions, we would advocate the continued use of model-free (*51*) methods as a complement and test of the model-based (*31–33*) approaches.

References and Notes

1. M. Seielstad, E. Bekele, M. Ibrahim, A. Touré, M. Traoré, *Genome Res.* **9**, 558 (1999).
2. P. A. Underhill *et al.*, *Nature Genet.* **26**, 358 (2000).
3. P. de Knijff, *Am. J. Hum. Genet.* **67**, 1055 (2000).
4. See www.ncbi.nlm.nih.gov/genome/guide/HsChrY.shtml
5. M. G. Thomas, K. Skorecki, H. Ben-Ami, N. Bradman, D. B. Goldstein, *Nature* **394**, 138 (1998).
6. O. Semino *et al.*, *Science* **290**, 1155 (2000).
7. M. Seielstad, *Am. J. Hum. Genet.* **67**, 1062 (2000).
8. L. B. Jorde *et al.*, *Am. J. Hum. Genet.* **66**, 979 (2000).
9. M. E. Hurles *et al.*, *Am. J. Hum. Genet.* **65**, 1437 (1999).
10. R. A. Kittles *et al.*, *Am. J. Hum. Genet.* **62**, 1171 (1998).
11. A. Helgason *et al.*, *Am. J. Hum. Genet.* **67**, 697 (2000).
12. G. Passarino *et al.*, *Am. J. Hum. Genet.* **62**, 420 (1998).
13. A. Pérez-Lezaun *et al.*, *Am. J. Hum. Genet.* **65**, 208 (1999).
14. B. Su *et al.*, *Am. J. Hum. Genet.* **65**, 1718 (1999).
15. C. Capelli *et al.*, *Am. J. Hum. Genet.* **68**, 432 (2001).
16. A. Ruiz-Linares *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6312 (1999).
17. E. Bosch *et al.*, *Am. J. Hum. Genet.* **65**, 1623 (1999).
18. D. B. Goldstein *et al.*, *Mol. Biol. Evol.* **13**, 1213 (1996).
19. E. S. Poloni *et al.*, *Am. J. Hum. Genet.* **61**, 1015 (1997).
20. M. F. Hammer *et al.*, *Mol. Biol. Evol.* **15**, 427 (1998).
21. P. Malaspina *et al.*, *Am. J. Hum. Genet.* **63**, 847 (1998).
22. J. K. Pritchard, M. T. Seielstad, A. Pérez-Lezaun, M. W. Feldman, *Mol. Biol. Evol.* **16**, 1791 (1999).
23. P. Shen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7354 (2000).
24. J. Wilson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
25. E. W. Hill, M. A. Jobling, D. G. Bradley, *Nature* **404**, 351 (2000).
26. D. B. Goldstein, P. H. Harvey, *Bioessays* **26**, 148 (1999).
27. J. Felsenstein, *Genet. Res.* **59**, 139 (1992).
28. P. Donnelly, *CIBA Found. Symp.* **197**, 25 (1996).
29. D. R. Cox, D. V. Hinkley, *Theoretical Statistics* (Chapman and Hall, Boca Raton, FL, 1974).
30. D. E. Reich, M. W. Feldman, D. B. Goldstein, *Mol. Biol. Evol.* **16**, 453 (1999).
31. See http://evolution.genetics.washington.edu/lamarc.html
32. See www.maths.monash.edu.au/~mbahlo/mpg/gtree.html
33. See www.maths.abdn.ac.uk/~ijw
34. R. Thomson, J. K. Pritchard, P. Shen, P. J. Oefner, M. W. Feldman, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7360 (2000).
35. R. C. Griffiths, S. Tavaré, *Theor. Popul. Biol.* **46**, 131 (1994).
36. S. Tavaré, D. J. Balding, R. C. Griffiths, P. Donnelly, *Genetics* **145**, 505 (1997).
37. I. J. Wilson, D. J. Balding, *Genetics* **150**, 499 (1998).
38. M. Kimura, *Proc. Natl. Acad. Sci. U.S.A.* **63**, 1181 (1969).
39. M. Slatkin, *Genetics* **139**, 457 (1995).
40. P. A. Moran, *Theor. Popul. Biol.* **8**, 318 (1975).
41. M. G. Thomas *et al.*, *Am. J. Hum. Genet.* **66**, 674 (2000).
42. D. B. Goldstein, D. D. Pollock, *J. Hered.* **88**, 335 (1997).
43. D. B. Goldstein, A. Ruiz-Linares, L. L. Cavalli-Sforza, M. W. Feldman, *Genetics* **139**, 463 (1995).
44. M. P. H. Stumpf, D. B. Goldstein, data not shown.
45. D. B. Goldstein *et al.*, *Am. J. Hum. Genet.* **64**, 1071 (1999).
46. M. Kayser *et al.*, *Am. J. Hum. Genet.* **66**, 1580 (2000).
47. B. Harr, C. Schlötterer, *Genetics* **155**, 1213 (2000).
48. D. B. Goldstein *et al.*, *Genetics*, in press.
49. J. C. Garza, M. Slatkin, N. B. Freimer, *Mol. Biol. Evol.* **12**, 594 (1995).
50. H. Ellegren, *Nature Genet.* **24**, 400 (2000).
51. See www.ucl.ac.uk/biology/goldstein/Gold.htm for software to calculate geneological depth using Δ_A and a length-adjusted mutation rate.

52. We thank M. Feldman, H. Harpending, J. Pritchard, M. Slatkin, D. Reich, J. Wilson, and two anonymous referees for helpful comments on earlier versions of this manuscript. Supported by the Wellcome Trust through a fellowship (M.P.H.S.).

REVIEW

Genetic Clues to Dispersal in Human Populations: Retracing the Past from the Present

Rebecca L. Cann

Ongoing debate about proper interpretation of DNA sequence polymorphisms and their ability to reconstruct human population history illustrates a important change in perspective that we have achieved in the past 20 years of population genetics. To what extent does the history of a locus represent the history of a population? Tools originally developed for molecular systematics, where genetic lineages have been separated by speciation events, are routinely applied to the analysis of variation within our species, with conflicting results. Because of automated technologies and linkage analysis, we are poised to harvest a wealth of information about our past, if we are successful in moving beyond a current polarization regarding models of human evolution. Rather than just suggesting that true resolution will only come by considering fossil or archaeological evidence, the realistic and appropriate application of genetic models for analysis of population structure is also necessary. Three examples from different dispersal events are highlighted here.

Studies of single-nucleotide polymorphisms (SNPs), as molecular genetic markers for mapping common disease genes (*1, 2*), have reconfirmed the importance of human population structure. It was originally estimated that about one difference per every 1 kb would exist in the human genome, and two broad surveys (*3, 4*) now suggest that for protein coding regions, this number is likely to be one difference per every 1200 base pairs. Some of these differences will be common around the world, but others will only be associated with local populations. Are there general predictions or principles to assist our interpretation of what is likely to be mere noise and what is genetically important? We know that the history of a gene is easier to determine than an accurate history of a population, because a particular pattern of variation can have multiple evolutionary causes.

In the SNP studies, Africans as a group show greater diversity of alleles and more unique alleles once ascertainment biases had been controlled for, consistent with the antiquity of this gene pool among humans. Comparisons with nonhuman primates also demonstrate that it is possible to use the same DNA chip technology to identify the ancestral state of many common alleles (*5*), so that frequency of a particular polymorphism can be used to infer the time that the allele arose, as predicted by theory (*6*). Yet, a recent study of DNA from Australian skeletal populations has again questioned the African origin of our species and suggested that we are still confused about population dynamics, bottlenecks, and migrations (*7*).

Allele frequencies, first generated from classical markers and more recently with microsatellites, had been the common currency used to compare population isolates because they generated data to estimate gene flow and population subdivision. In order to escape the biases of natural selection that might drive

Department of Cell and Molecular Biology, John A. Burns School of Medicine, University of Hawaii at Manoa, 1960 East-West Road, Honolulu, HI 96822, USA. E-mail: rcann@hawaii.edu