Adv. Appl. Prob. **31,** 1027–1035 (1999)
Printed in Northern Ireland

DISCUSSION: RECENT COMMON ANCESTORS OF ALL PRESENT-DAY INDIVIDUALS

It is a pleasure to be able to comment on such an interesting and insightful paper. Demographic properties of population genetics models have been of interest for more than 50 years, and suddenly we have a sharp new perspective on properties of a diploid Wright–Fisher model. The results in Chang's paper are all the more interesting for their contrast with the analogue for haploid (that is, single-parent) models. As the author notes, not only is the mean time to a common ancestor, in the sense of the current paper, much smaller in the diploid than the haploid case (lg n rather than order n), but to the appropriate order, the mean captures all the interesting behaviour in the former, but not the latter, case.

The author is certainly correct that the work on mitochondrial Eve captured many imaginations. Indeed, it is almost becoming routine for papers documenting DNA sequence variation within populations to include some kind of estimate of the time since the most recent common ancestor (MRCA) of the sampled sequences. Counterintuitive as it may seem, in studying the ancestry or modelling the diversity in DNA sequences, a single-parent model for reproduction, and as a consequence the coalescent, is appropriate, as the paper notes. The coalescent and its generalizations to more general demographic scenarios thus provide a natural 'pre-data' model, and the inferential question is to calculate the conditional distribution of the time to the MRCA given the molecular genetic data. Evaluation of this conditional distribution is actually a challenging statistical problem, and the subject of considerable current interest. (See for example Tavaré *et al*. (1997); Wilson and Balding (1998); and references therein.)

But what does it all mean? Appreciation of these issues in the human evolution community is somewhat more sophisticated than it was, but some residual confusion is not uncommon. For example, it is apparently a small step from the (logically accurate) quotation from Pääbo given in the paper to the logically inaccurate assertion that the existence of a recent common ancestor for all human mitochondrial DNA implies a recent origin for our *species*. In fact, there is no necessary logical relationship between the time at which MRCAs to particular parts of the genome existed, and the time at which the species originated. Having said that, one scenario for speciation involves it occurring in a small, isolated, population. Under this scenario (but not many others) if the population size were small enough, it would tend to increase the probability that the MRCA occurred close in time to the speciation event.

Notwithstanding its effect on the public imagination, estimates of the dates of MRCAs for particular genomic regions would seem to be of limited interest in their own right. At best they may act as surrogates from which we may learn about the history of the population under study. To do so, however, one needs to build up a picture from many different genomic regions. The realized time for any single region is a sample of size 1 from a distribution which depends, in a complicated way, on the demographic history of the population. It is only when we have a reasonable sample size that we can hope to learn about the distribution, and hence about population history. For example, it is of considerable interest to learn whether the apparent recent common ancestry for mitochondrial DNA is shared by other parts of our genome (as we would expect, for example, if speciation occurred in a small population which subsequently expanded). Even then, it would be better to make inferences about population history directly from the molecular data, rather than indirectly, through one (unobserved) summary measure, the depth of the genealogical tree relating sampled sequences.

The 'correct' tree relating the evolution of particular extant species is naturally an object of curiosity. The author points out that it may not always be well defined, notably when two speciation events occur close in time. The suggestion in the paper for resolving the ambiguity is very interesting. On the other hand, I am inclined to the view that the primary reason for reconstructing evolutionary trees is to structure our thoughts about, and analyses of, genetic diversity. With a single, relatively short, sequence from each species there will be a unique tree, though the topology of the tree may change with the genomic region studied. There may, or may not, be a unique tree for each such region, in the sense that the same topology obtains regardless of which individual from the species is sequenced, and again, these topologies may be locus-specific. In such cases it would seem preferable to be explicit in recognising the distinct patterns of evolutionary history for different genes rather than to construct a single 'species' tree.

Traditional genetics models for geographically structured populations typically posit a number of large, randomly-mating, colonies which interchange genes through migration. This seems unrealistic, for example, for human populations, and there is an urgent need for the development and study of demographic models which maintain a sensible population density in, say, a two-dimensional region. Nonetheless, it is clear that geographical structure of the population can radically alter the shape of genealogical trees, and in particular times to MRCAs. The author notes that the same will be true in his setting, and it would be interesting to understand better the nature of the mechanisms involved. Results on the extent to which common ancestors, in the sense of this paper, are ancestors in the genetic sense, that some of their DNA is inherited by the current generation, would also be of great interest. It is clear that this will depend on the relationship between population size and genome size, possibly in interesting ways.

References

TAVARÉ, S., BALDING, D. J., GRIFFITHS, R. C. AND DONNELLY, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145,** 505–518.

WILSON, I. J., AND BALDING, D. J. (1998). Genealogical inference from microsatellite data. *Genetics* **150,** 499-510.

PETER DONNELLY
Department of Statistics,
University of Oxford,
1 South Parks Road,
Oxford, OX1 3TG, UK.
Email address: donnelly@stats.ox.ac.uk

We have read the paper by Chang with great interest and enjoyed his work very much. A lot of work, within the last decades, has been devoted to understanding the ancestral history of humans and of populations in general. The paper by Chang relates to this work. We found that the questions he addresses are as much motivated from a basic curiosity point of view as from a mathematical point of view. Our attention will be on a few questions that emerge from confronting Chang's work with the 'coalescent'.

In the model by Chang, an individual has two ancestors in each generation with certainty. However, in some non-human populations (for example some bacterial and plant populations) each individual will not necessarily have two ancestors, because each individual has the opportunity to reproduce asexually as well as sexually. Imagine that an individual at random chooses two ancestors in the previous generation with probability $1-p$, and with probability p chooses just one. On average each individual will have $2\cdot(1-p)+1\cdot p = 2-p$ ancestors. We conjecture that the time $\mathcal{T}_p(n)$ until a most recent common ancestor (MRCA) to the entire population will be about $G_p(n) = \log_{2-p}(n)$ generations (where $\log_q$ denotes base-q logarithm) for $0 \leq p < 1$, in line with the result by Chang ($p = 0, G_0(n) = \log_2(n)$). This we have confirmed by simulations. For an entirely asexual population, that is for $p = 1$, we retrieve the standard coalescent result that a MRCA is found $2n$ generations in mean back in time.

The convergence of $\mathcal{T}_p(n)/G_p(n)$, $0 \leq p < 1$, in probability towards 1 cannot happen uniformly in n as a function of p because we have $G_p(n) = \log_{2-p}(n) > 2n$ for large p but $\mathrm{E}\mathcal{T}_p(n) < \mathrm{E}\mathcal{T}_1(n) = 2n$ for all $p < 1$. Moreover, $\mathcal{T}_1(n)/2n$ does not converge in probability towards 1 for n tending to infinity, but towards a stochastic variable with mean 1 and variance 0.29. The latter is well known from standard coalescent theory.

In the proof given by Chang, we can see that Lemma 6 breaks down for $p = 1$. The process $M_t = Y_t\,(2-p)^{-t}$ is a martingale for all p, but the proof needs

$$P\{M_t < b\,(2-p)^{-t} \text{ infinitely often}\} < \text{constant} < 1$$

with b an increasing function of n. The validity of this statement is not guaranteed for $p = 1$. We have checked other lemmas and propositions, and they all seem recoverable for p's less than one.

Furthermore, we could imagine a situation where each individual chooses k, $k \geq 2$, ancestors at random in the previous generation. Such a population could either consist of a hypothetical extra-terrestrial life-form or, for example in the case $k = 4$, be a human population where each individual chooses four grandparents (or two parents and two parents-in-law). One generation in the grandparent setting is almost two generations in Chang's setting. The probabilities that an individual has a, $a = 1, 2, 3$, or 4, different ancestors are in Chang's setting $o(1/n)$, $1/n$, $6/n$, and $1-7/n$, whereas in the grandparent setting these are $o(1/n)$, $o(1/n)$, $6/n$, and $1-6/n$. However, the expected number of descendants to a given number of individuals is asymptotically the same in the two settings, and we will expect that the time $\mathcal{T}_n(k)$ until an MRCA fulfil $\mathcal{T}_n(k)/\log_k(n) \to 1$ in probability.

Concerning the second result by Chang, Theorem 2, we expect to find a version of the statement similar to Chang's for $0 \leq p < 1$ here also. It is interesting to note that for $p = 1$ we have $\mathcal{T}_n(1) = \mathcal{U}_n$ ($\mathcal{U}_n$ as in Chang's notation), that is when the present-day population finds an MRCA in generation $\mathcal{T}_n(1)$ this ancestor is the only individual ancestral to the present-day population.

The way of choosing parents defines the reproductive structure; each individual has a binomially distributed number of offspring with mean $\approx 2-p$ (or k). It is not clear to us how Chang's results are changed if the reproductive structure is different, i.e. if another

exchangeable model is assumed to reproduce the population. Only in the case $p = 1$ is the result known; the time $\mathcal{T}_n(1)$ until an MRCA is scaled with the variance, σ^2, of the number of offspring of an individual, $\mathrm{E}\mathcal{T}_n(1) = 2n/\sigma^2$.

Let us now turn to some genetical aspects of Chang's results. Relative to the time-scale of the Wright–Fisher coalescent process (units of n generations), the time until a common biological ancestor (CBA) of all present-day individuals is just a glimpse of time ($\log_2(n)$ compared to $2n$). If we follow two individuals (or in general a small finite sample) from the present generation and about $\log_2(n)$ generations back in time, the chance that some specific part of the genome of the two individuals has found a common ancestor that relates them genetically is negligible. So, the most recent CBA will in general not be a common genetical ancestor (CGA) to a specific part of the genome of two individuals.

Moreover, a single individual will not necessarily be related genetically to the most recent CBA. This depends on the size of the genome considered and the reproductive model involved. For example, each individual has two copies of each gene, so two of the individual's grandparents are not grandparents in the genetic sense. If we consider one gene at each chromosome, we would have to go 6 generations (assuming all $2^6 = 64$ ancestors are different) back in time to find a biological ancestor who (with certainty) is not a genetical ancestor to any of the genes.

Consider the number of biological ancestors, A_t^i, t generations back in time to a single individual, i. It can easily be seen that $f(x) = E(A_{t+1}^i \mid A_t^i = nx) \approx n(1 - \exp(-2x))$ for n large. Assuming that the process A_t^i at stationarity does not fluctuate very much, we find from solving $f(x) \approx x$ that about 80% of the population is biologically related to individual i. Thus, a biological ancestor will on average be a genetical ancestor to about $g/(0.8n)$ genes (g denotes the number of genes in the genome) of a single individual.

The paper was very elevating for us to read, and we find that it would be interesting to pursue these matters further.

CARSTEN WIUF
Department of Statistics,
University of Oxford,
1 South Parks Road,
Oxford, OX1 3T6 UK.
Email address: wiuf@stats.ox.ac.uk

JOTUN HEIN
Institute of Biological Sciences,
University of Aarhus,
DK-8000 Aarhus, Denmark.

Population genetics theory follows four traditions, three of which rely on probabilistic models to represent the intrinsic unpredictability of reproduction and Mendelian inheritance. Only one tradition, that of models of natural selection and adaptation, employs the deterministic theory of difference and differential equations. A second tradition is concerned with quantitative genetics and is based on linear statistical models that lead to the analysis of variance. A third tradition is based on Markov chains which model genetic drift, possibly in combination with other forces such as selection and mutation. Either the Markov chains themselves or diffusion approximations to them have yielded many results of central importance to the study of molecular evolution. A fourth tradition is concerned with the genealogical history of a sample of genes. Although genealogical methods follow from Sewall Wright's [5] theory of inbreeding coefficients and Malécot's [4] theory of identity by descent, their recent use has blossomed since Kingman's seminal papers on coalescent theory in the early 1980s, for instance [3].

Coalescent or genealogical theory is concerned with gene genealogies that represent the continuity of DNA sequence through successive rounds of replication. Each non-recombining segment of DNA has its own genealogical history that can be modeled by a stochastic process. The development and application of coalescent theory has been driven in part by the increasing abundance of DNA sequences from a variety of organisms. Coalescent theory applied to the new data has made it possible to critically evaluate hypotheses about evolutionary processes and demographic events [2]. One notable example is the theory (still controversial) that all human females today trace their ancestry to a woman that lived in Africa between 100 000 and 200 000 years ago [1]. This theory of human ancestry was based on the maternal inheritance and lack of genetic recombination of the mitochondrial genome and hence the name 'Mitochondrial Eve' for the most recent common female ancestor of living females.

The announcement of mitochondrial Eve caused considerable confusion and consternation among non-geneticists and geneticists alike. One criticism, voiced by many who should have known better, was that such an individual could not have existed, and that instead living females and males too carried genes from many different ancestors. No one could be designated as the most recent common ancestor. That criticism is based on a misunderstanding of the original claim: not all genes are descended from Eve's but only those 37 in the mitochondrial genome, which is inherited as a unit from the mother; fathers make no contribution to the mitochondrial genome of their descendents. As a result of this uniparental inheritance, there must be a mitochondrial Eve, because the number of female ancestors of living females is modeled by a pure death process that has an absorbing state at 1. The only question is when and where Eve lived. The rest of the genome (the 100 000 or so genes found on the 23 pairs of chromosomes in the nucleus) will have different ancestries, different not only from the ancestry of the mitochondrial genome but from one another. Each small segment of nuclear DNA, however, will have an ancestry that can be traced to a most recent common ancestral segment carried in the nucleus of its Eve or Adam at some time in the past. What is currently unknown is how many different ancestral segments there are at any time in the past and the relationship of their different gene genealogies. Simulation methods work for a few linked loci [2] but provide us with no analytic results and little intuition.

Chang provides some important and surprising insights about the ancestry of individuals in a randomly mating population of constant size. His theory is not concerned with genetic ancestry of the kind described by coalescent theory, but what can be thought of as 'practical' or 'potentially genetic' ancestry. Chang assumes that each individual has two immediate ancestors and then examines the properties of the most recent common ancestor (MRCA),

defined to be an individual who is an ancestor of everyone in the population today. Chang shows that the time of the MRCA is concentrated on $\log_2(2n)$ in a diploid species of constant size n and that by roughly $1.77\log_2(2n)$ generations in the past, every individual who has any descendents at all living today is an ancestor to everyone. The logarithmic dependence of this result contrasts sharply with the average time to MRCA of roughly $4n$ generations for a gene genealogy.

As Chang emphasizes, his is a not a genetic result because each chromosomal segment has one not two ancestors. But his result suggests that the ancestries of different chromosomes and different segments of each chromosome have the potential to have very different gene genealogies and that some of them may have most recent common genetic ancestors in the relatively recent past instead of the more distant past suggested by the application of coalescent theory to single genes. New results for multiple independent chromosomes will be needed to join coalescent theory to Chang's theory of biparental ancestry, but his results suggest that such a theory might be accessible and will provide some surprising conclusions.

References

[1] CANN, R. L., STONEKING, M. AND WILSON, A. C. (1987). Mitochondrial DNA and human evolution. *Nature* **325,** 31–36.

[2] HUDSON, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology Vol. 7*, eds. D. Futuyma and J. Antonovics. Oxford University Press, Oxford, pp. 1–44.

[3] KINGMAN, J. F. C. (1982). On the genealogy of large populations. In *Essays in Statistical Science,* eds. J. Gani and E. J. Hannan (J. Appl. Prob. **19A,**). Applied Probability Trust, Sheffield, UK, pp. 27–43.

[4] MALÉCOT, G. (1941). Étude mathématique des populations 'mendéliennes'. *Ann. Univ. Lyon Sec. A* **4,** 45–60.

[5] WRIGHT, S. (1921). Systems of mating. I. The biometric relations between parent and offspring. *Genetics* **6,** 111–178.

MONTGOMERY SLATKIN
Department of Integrative Biology,
University of California,
Berkeley,
CA 94720-3140, USA.
Email address: slatkin@socrates.berkeley.edu

In this paper Chang compares various ancestry properties of a 'two-parent' Wright–Fisher process with parallel properties of Kingman's 'coalescent' process. The Kingman process describes ancestry properties in a variety of genetical models including the 'one gene parent of any gene' Wright–Fisher model. In particular, the question of when 'mitochondrial Eve' lived is discussed in the context of this comparison.

Before taking up the comparisons that Chang makes between predictions of his model and of the coalescent, some issues which are important in themselves, but are not central to these comparisons, should be mentioned. Chang points out clearly that, so far as the human population is concerned, the assumptions made in his model concerning random mating and constancy of population size are unrealistic. The same is true of the parallel assumptions of early coalescent theory (but by not its more recent variants, which attempt to remove these assumptions). Thus, as he notes, his conclusions and comparisons do not apply with any high degree of numerical accuracy when used for the human population. To focus on the main points at issue we consider the simple case where real-world complications are assumed not to arise; specifically, we consider throughout a randomly mating population having the same size n in every generation.

Tracing backwards in time in Chang's 'two-parent' model, it takes on average about $\log n$ generations before some person is found who is an ancestor of all currently living people (all logs are to the base 2). When $n = 10^6$, this is about 20 generations. The coalescent process, describing as it does the progress of genes in a population, is by contrast a 'one-parent' model, since any gene has only one 'parental' gene. Tracing back in time in this one-parent model, one requires on the order of n generations, one million in the above example, before it is likely that all genes in the present-day population are all descended from one single gene in an earlier generation. This is claimed to be in dramatic contrast with the analogous result, namely about 20 generations, for the 'two-parent' model.

But neither result, nor the comparison between them, is disturbing nor new. The comparison is not disturbing because the two calculations concern entirely different questions. Nor is it new. The geometric factor of two increase per generation in the number of ancestors per generation makes it almost obvious that, ignoring multiple ancestry, it requires only about $\log n$ generations to have n ancestors in any generation. Tracing forward in time, since in a population of fixed size each person has on average two children, one expects that it requires on average only about $\log n$ generations before some person is an ancestor to all persons in the later generation. While Chang presents various refinements to these calculations, the logarithmic aspect of his calculation is, as noted, neither new nor disturbing.

Chang introduces his calculations by discussing 'mitochondrial Eve', that is, the most recently living woman from whom we all derived our mitochondrial DNA. It is therefore interesting to address the question: Is the two-parent calculation, or the one-parent coalescent calculation, appropriate for asking when 'mitochondrial Eve' lived?

A woman passes on her mitochondrial DNA to all her children, male and female, that is (in a population of constant size) to two children on the average. It might then be thought, using the arguments sketched above, that 'mitochondrial Eve' lived only about 20 generations ago in a population of size one million. However this conclusion is incorrect. Only women are relevant to the transmission of mitochondrial DNA, so only the female half of the population is relevant to 'mitochondrial Eve' calculations. Each woman passes on her mitochondrial DNA to one daughter, on average, and each woman inherits her mitochondrial DNA from exactly one parent, namely her mother. This implies that the one-parent Wright–Fisher model,

and Kingman's coalescent theory, is the correct vehicle for asking when 'mitochondrial Eve' lived. In the numerical example above, we would estimate this to be on the order of a million generations ago, not 20.

Of course, given the practical problems concerning population stratification, the steadily increasing size of the human population and many other similar considerations, this calculation is not likely to be anywhere near the truth (and would not in any event be made by more modern coalescent calculations). But 'mitochondrial Eve' almost certainly lived many thousands of generations ago, not just 20.

W. J. EWENS
Department of Biology,
University of Pennsylvania,
415 South University Avenue,
Philadelphia,
PA 19104-6018, USA.
Email address: wewens@sas.upenn.edu

This perceptive, and mathematically powerful, analysis shows how careful one must be in carrying over concepts of haploid genealogy to diploid populations. The concept of being a common ancestor of a whole subsequent population is a very weak one, and it is not clear that it has any real significance. The point of genealogy is either to describe inheritance of something (titles, property) or to understand correlations between relatives (inheritance of genes). In neither case is the existence of a single common ancestor relevant in itself.

The author disclaims genetics, but it is worth recalling why the common ancestor concept, in its haploid form, is central to the genetics even of diploid populations. If one asks why there is a diversity of genes at a single chromosome locus in a population, it is natural to start from the observation that each gene comes from a gene at that locus in one of the parents. The parent gene itself has a parent, and so on, tracing the ancestry back from generation to previous generation. Typically, if one carries out this genealogical study for n individual genes, one will find a common ancestor some N generations back, where N is a measure of the population size. Below this common ancestor will be a family tree descending to the n children.

The n genes will be identical unless mutation has affected some or all of the lines of the family tree, and one can make probability statements about n genes by analysing the way mutation is likely to strike the different lines. This leads to the Ewens sampling formula for the partition of n which summarises the diversity of the n genes, and so to the Poisson–Dirichlet formula for the distribution of population frequencies when N is large.

Because of the essential simplicity of the argument, the conclusions are extremely robust to the detailed reproductive mechanism, so long as one crucial assumption is made. Everything turns on the independence of the genealogical structure of the family tree on the one hand, and the process of mutation on the other. If this is true, everything else follows. If not, the true distributions can be radically different (which is of course what the statistician needs).

The independence of the family tree and mutation is violated when mutation implies selective advantage or disadvantage. It is not that mutation is causally influenced by genealogy, but that mutation causes some branches of the tree to develop and bifurcate more strongly than others. The resulting correlations are highly complex, especially in diploid populations with strong dominance or heterosis. And of course different loci affect one another, with hitch hiking and other effects becoming important.

Thus the simple argument that traces back the parents of individual genes (the ‘coalescent’) is of little use when selection is important, but in the neutral models which always need to be considered, if only as null hypotheses, they are powerful and illuminating.

J. F. C. KINGMAN
University of Bristol,
Senate House,
Tyndall Avenue,
Bristol, BS8 1TH, UK.