Interpreting genetic variability: the effects of shared evolutionary history

Peter Donnelly

Departments of Statistics, and Ecology and Evolution, University of Chicago, 5734 University Avenue, Chicago, IL 60637, USA

Abstract. Data from different individuals at a single locus are positively correlated because of the shared genealogy of the sampled genes. This paper illustrates the qualitative effects on genealogical trees of assumptions about population demography, and it considers the consequences for genetic variability. An understanding of these effects is invaluable in the interpretation of data and for inferences about population history. In contrast, traditional genetic measures of diversity and approximation methods do not seem well suited for addressing the problem.

1996 Variation in the human genome. Wiley, Chichester (Ciba Foundation Symposium 197) p 25–50

Patterns in genetic data result from the superposition of two random mechanisms. The first is the underlying genealogical tree, which describes the ancestral history of the sampled genes. The second is the effect of mutation in changing genetic types. In non-neutral models these two forces interact: the action of mutation on genetic type in turn affects the structure of the genealogical tree as a consequence of natural selection. In contrast, for neutral models the two processes can be treated separately: one can first study properties of genealogies, and then superimpose the effects of mutation. Little is known about genealogy for selective models; however, genealogy is now reasonably well understood under neutrality.

This separation of the effects of genealogy and mutation, via so-called 'coalescent methods', has provided powerful tools for studying neutral models, both analytically and by simulation. It also provides valuable insights into the patterns and correlations within population genetic data.

From the point of view of using neutral genetic data to make inferences about population history, a genealogical approach is extremely natural. Different scenarios for population demography effect only the first of the underlying random mechanisms: they change the probability distribution of the underlying genealogical trees. Thus, if one could observe the trees

themselves, inference about demographic history would be a direct, though rather complicated, statistical problem. In fact, what one observes is the result of a noisy filter, namely mutation, acting on the genealogical tree.

Our perspective throughout this paper is that to understand the consequences of demographic assumptions, and hence to make inferences about them, one should first study their effects on genealogical structure. One can then ask, separately, how certain tree structures will be reflected in patterns in genetic data.

The inference problem is far from simple. Data on different individuals at a single locus are highly correlated, and for most models there is substantial variability in realized genealogical trees. The variability means that even if one were to observe the entire gene tree associated with a sample, precise inferences about population history would be impossible. In any case, because of the dimensionality of the underlying trees, statistical analysis is technically challenging. The problem is compounded by the substantial additional noise imposed by the mutation process. On the other hand, data from unlinked loci have independent genealogies, so that some replication is available.

Traditional summary measures based on pairwise comparisons, such as heterozygosity and pairwise sequence divergence, ignore most of the information in genetic data. They are not well suited to the problem of understanding human population history. Instead, more sophisticated methods, which make better use of information in the data, may be desirable. Further, traditional approximations of more realistic population models, via effective population sizes, can be seriously misleading in this context. Instead, one is now able to study, in some cases analytically and always by simulation, properties of reasonably complicated evolutionary models. Such an approach would appear essential for understanding the consequences of the many simplifying assumptions made in current analyses.

The next section describes and illustrates both the coalescent and the qualitative effects on genealogical structure of various demographic assumptions. (For a more detailed and more quantitative treatment, see Tavaré 1984, Ewens 1990, Hudson 1990, 1992, Donnelly & Tavaré 1995, and references therein.) The consequences of genealogical structure on genetic variability will also be examined, and the implications for inference about population history will be discussed. Neutral loci will be considered throughout the paper, and recombination within a locus will be ignored. (For details of the effects of recombination, see Hudson 1990, 1992, Marjoram & Griffiths 1995.) Much less is known of genealogy in the presence of selection (see, for example, Kaplan et al 1988, 1989, Hudson & Kaplan 1994).

Genealogical trees

The coalescent (Kingman 1982a,b,c) can be thought of as a random tree. Its distribution is a close approximation to that of the genealogical tree associated

with a sample of genes taken from a large panmictic population which has been of the same size throughout its evolution. The approximation is valid for a variety of demographic models.

Throughout this paper, trees are drawn vertically from the number of sampled genes (n) at the tips of the tree up (i.e. backwards in time) to their most recent common ancestor (MRCA). At any time between the present and the MRCA of the sample, the coalescent tree will have one branch for each gene in the population that is an ancestor of at least one of the genes in the sample. Two branches in the tree coalesce each time the corresponding ancestral genes themselves share a common ancestor, until the entire sample is traced back to the single ancestral gene from which all are descended.

The coalescent has a particularly simple structure. Time is measured in units of N/σ^2 generations, where N is the number of haploid genes in the population and σ^2 is the variance of the number of direct copies in the next generation of a gene in the current generation. Throughout this paper I will assume, as is common, that the value of σ^2 is 1. Note, however, that this need not be true for early human populations, particularly for genes carried in males.

In the coalescent, the times $T_j, j = n, n-1, \ldots, 2$ for which the tree has exactly j branches are independent, exponentially distributed random variables with

$$E(T_j) = \frac{2}{j(j-1)}.$$

Thus,

$$\mathrm{Var}(T_j) = \frac{4}{j^2(j-1)^2}.$$

It follows that the total depth of the tree, i.e. the time until the MRCA of the sample, has mean $2(1 - 1/n)$ and variance between 1 and about 1.16, the latter being a good approximation for n larger than 4. Note that most of the depth of the tree, and almost all of the variability in this depth, is due to the times for which there are only a small number of ancestors. In particular, the expected time for which there are exactly two ancestors is over half the expected depth of the tree, and the variance of this time is 1.

Each time the number of branches in the tree decreases, the coalescence is equally likely to involve any of the possible pairs of branches in the tree at that time. When there are j ancestors of the sample, the joint distribution of their numbers of descendants in the sample is uniform. In particular, the distribution of the number of descendants in the sample of a particular one of the final two ancestral genes before the MRCA is uniformly distributed on $\{1, 2, \ldots, n-1\}$.

Figure 1 shows six realizations of coalescent trees. The timescaling is that for an autosomal locus in a population of 5000 diploid individuals. For

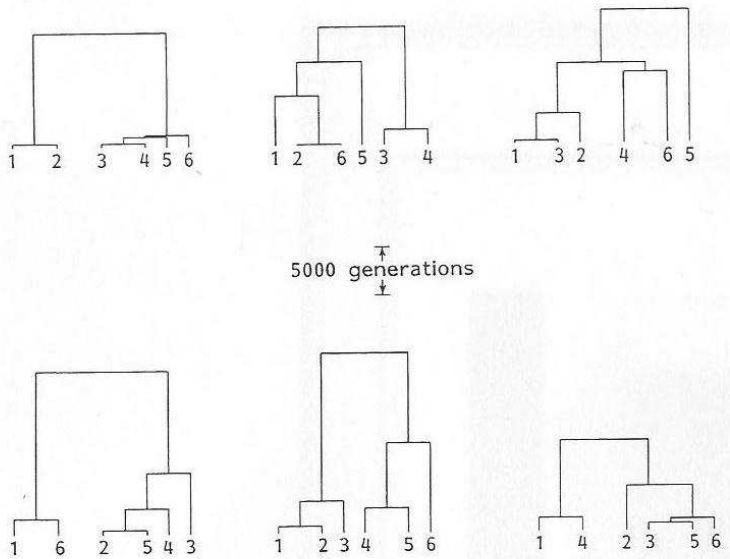

FIG. 1. Six realizations of the gene genealogy for a sample of size $n = 6$ genes from a panmictic population of a constant size of 5000 individuals.

populations of different sizes, all that is necessary is to change the vertical time units. Note the variability, and the fact that it is common to see trees dominated by the final two branches before the MRCA.

All of the figures in this paper relate to small sample sizes. The reason for this is that under most assumptions, 'early' coalescences, i.e. those near the tips of the trees, occur quickly. For larger sample sizes, or indeed for the whole population, the structure of the trees would be similar to those presented here except for dense branching near the tips of the tree. An exception to this is the scenario of continual rapid population growth underlying Fig. 2.

For panmictic populations that have not maintained constant (or approximately constant) sizes, genealogical trees have a different distribution. In general, these trees arise as non-linear time changes of coalescent trees. The nature of this time change depends on aspects of the demography of the population (P. Donnelly & T. G. Kurtz, personal communication 1994). (For early work on the effects on genetic data of variation in population size see Chakraborty 1977.) The simplest and best understood case applies, for example, to populations in which the variation in the population size has been 'exogenous' in a certain sense, or in which it results from independent reproduction by different individuals. This case is, therefore, moderately general, and it appears to have been assumed, at least implicitly, in all published work that uses genealogical techniques to study human evolution.

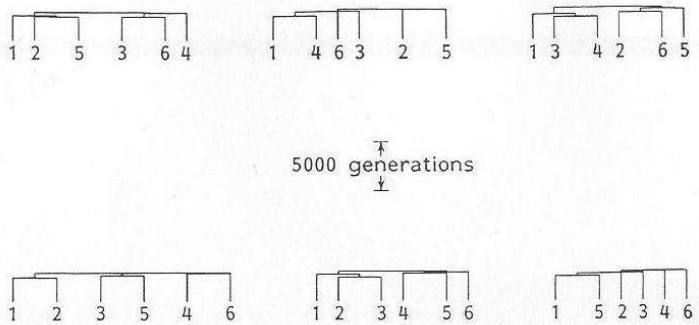

FIG. 2. Six realizations of the gene genealogy for a sample of size $n = 6$ genes from a panmictic population that has grown exponentially, throughout its history, to a current size of 250 000 000 individuals. The value taken by the population size 2500 generations ago is 5000.

However, it may not apply to early human populations; therefore, results described here and elsewhere on the effects of variation in population size should be interpreted with caution.

One way of thinking about the coalescent approximation to constant-sized populations is that each generation of real time corresponds to $1/N$ units of coalescent time. For the class of variable population size models considered here, a generation t for which the haploid size of the population is N_t accounts for $1/N_t$ units of coalescent time.

Consider a population, such as the human population, that has grown in size with time. Recent generations correspond to large population sizes, in which coalescences are less likely than in the smaller-sized generations in the past. The effect of this, which can be quite marked, is to stretch the usual coalescent tree near its tips, and to shrink it near its root. If a population has grown rapidly from a small size, the result is to make the associated genealogical trees resemble a star phylogeny. The intuition behind this is that, going backwards in time, no coalescences occur until the population reaches a relatively small size, at which point all of the coalescences occur in close succession.

Figure 2 shows realizations of six sample genealogies for a panmictic population that has grown continuously in an exponential fashion. Note the similarity in the shapes of the trees for all six realizations, in marked contrast to the variability in the other figures presented here. The parameters, although perhaps not the continual exponential growth, may be plausible for the human population. Any rapid growth, from a small size, will tend to produce the same effect. The exact form of growth is immaterial.

It is important to note that this effect of star-shaped genealogies depends crucially on the fact that the population has grown from a small size. In contrast, for a population that was of approximately constant size before

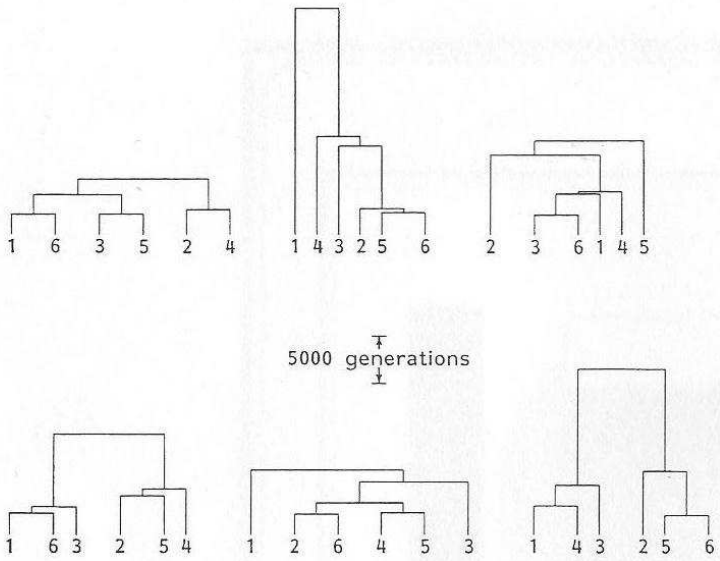

FIG. 3. Six realizations of the gene genealogy for a sample of size $n = 6$ genes from a panmictic population. The population is assumed to have been at a constant size of 5000 individuals until 2500 generations ago, before growing exponentially to a current size of 250 000 000 individuals.

growing exponentially, the shape of the trees will depend on the size of the population before growth. Unless this is small, say fewer than 500 individuals, the trees will again resemble those of the coalescent. Figure 3 illustrates this effect. It shows six trees from a population that had a constant size of 5000 until 2500 generations ago, from which point it grew in size exponentially to a current value of 250 000 000.

The trees in Fig. 3 are much more similar to the standard coalescent trees of Fig. 1 than to those of Fig. 2. The demographic difference between the settings of Figs 2 and 3 is that in the latter case the population is assumed to be of constant, non-trivial size before growth. This may be more plausible for the human population than an assumption that it has grown exponentially from a small value.

Genealogical properties are also well understood for certain models of geographically structured populations. These models posit a population consisting of large, partially isolated colonies. Each such colony is randomly mating, with gene flow resulting from migration of individuals between colonies. The model is then specified by describing the relative sizes of the colonies and the rates and patterns of migration between them.

Attention is restricted here to the qualitative effects on genealogy of such spatial structure. (For a fuller description of the models and the associated genealogical processes, see for example Donnelly & Tavaré 1995 and references therein. For details of genealogy in the presence of this form of spatial structure and variation in population size, see Marjoram & Donnelly 1996.) The models described in the previous paragraph may well not capture important features of early human evolution, so that conclusions drawn from them in this context should be interpreted with caution. It seems likely that many of the qualitative effects on genealogy of geographical population structure may nonetheless hold for other forms of spatial structure, although at this stage little has been established.

If the migration rates are high enough, gene trees from any of the spatial models under consideration will resemble those from panmictic populations. For lower migration rates, the effect of the spatial structure, loosely speaking, is opposite to that of population growth. That is, gene trees tend to be compressed near their tips and stretched near their root. The intuition behind this is that the early coalescences (those near the tips) occur between genes within colonies and these are more rapid because each colony is smaller than the whole population. On the other hand, because coalescences can only occur when the relevant ancestral genes are in the same colony, the final few coalescences in the tree can take substantially longer than in the panmictic case because they must first wait for migration to bring the ancestral genes into the same colony. The extent of this stretching near the root increases as the mutation rate decreases or the extent to which the population is structured increases. A further effect of population structure, in contrast to the setting of population growth, is that there can be enormous variability between realizations in gene trees.

Figure 4 shows six realizations of genealogical trees for a sample of 12 genes from a structured population. The simulations assume a 3×3 stepping-stone model for the population structure. That is, the population is assumed to consist of nine colonies arranged as a square lattice (in fact as a torus to avoid edge effects) with migration allowed only between a colony and its four immediate neighbours: above; below; left; and right. For each gene, per generation, the probability of migration is taken to be 10^{-5}. In the event of a migration the destination is chosen uniformly from the four neighbours. The sample of 12 genes consists of six genes, labelled 1–6 from one colony, and six additional genes, labelled 7–12 from a second, neighbouring colony. The assumptions about population size are as for Fig. 3. Note that the vertical scale in Fig. 4 is different from that in the earlier figures.

The trees in Fig. 4 tend to be much deeper than in the panmictic case (Fig. 3), although the depth of the tree at the bottom right (17 449 generations) is comparable with those for panmictic populations of the same size. Note also the extent to which the trees are dominated by the times for the final one (or for the top left tree, two) coalescence(s). Relative to the panmictic trees, even the

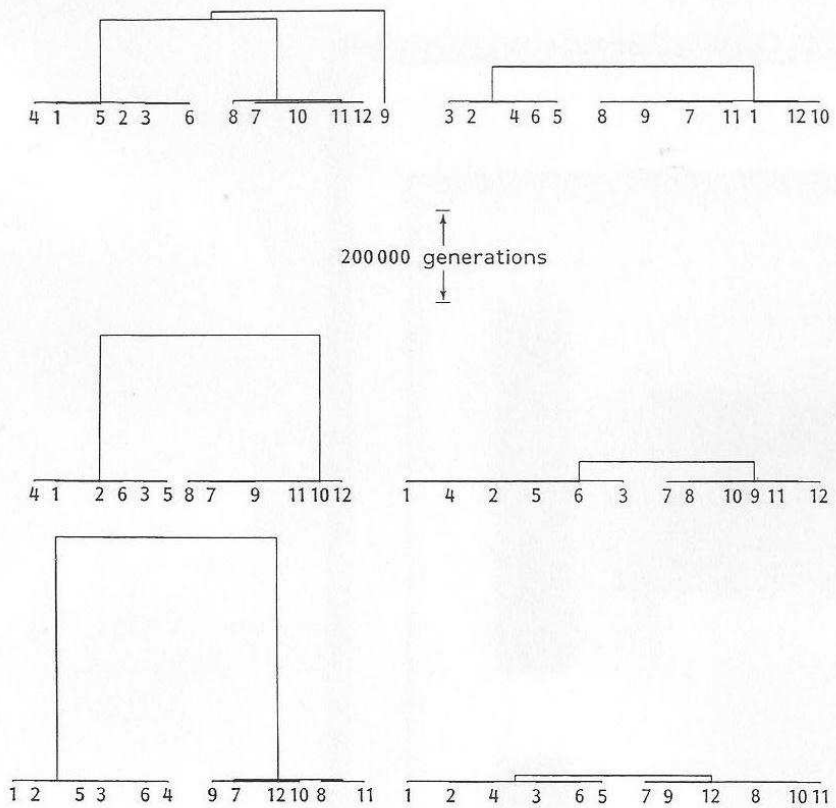

FIG. 4. Six realizations of the gene genealogy of a sample of size $n = 12$ genes from a geographically structured population. The population is structured according to a 3×3 stepping-stone model with migration probability of 10^{-5} per gene per generation. The population is assumed to have been at a constant size of 5000 individuals until 2500 generations ago, before growing exponentially to a current size of 250 000 000 individuals. The sample consists of six genes from each of two neighbouring colonies. Genes labelled 1–6 are from one colony. Those labelled 7–12 are from the second colony.

bottom right tree exhibits this effect: the time for which there are exactly two ancestral genes (about 13 000 generations) is longer than for all but one of the trees in Fig. 3. The effect is extremely marked for the other trees. In the structured population, the early (first nine or 10) coalescences tend to occur more rapidly than the early (first four) coalescences in the panmictic case. In most of the trees, all the genes from within the same colony have a recent common ancestral gene within that colony. The exceptions are the top left tree, in which gene 9 is descended from a recent migrant from one of the

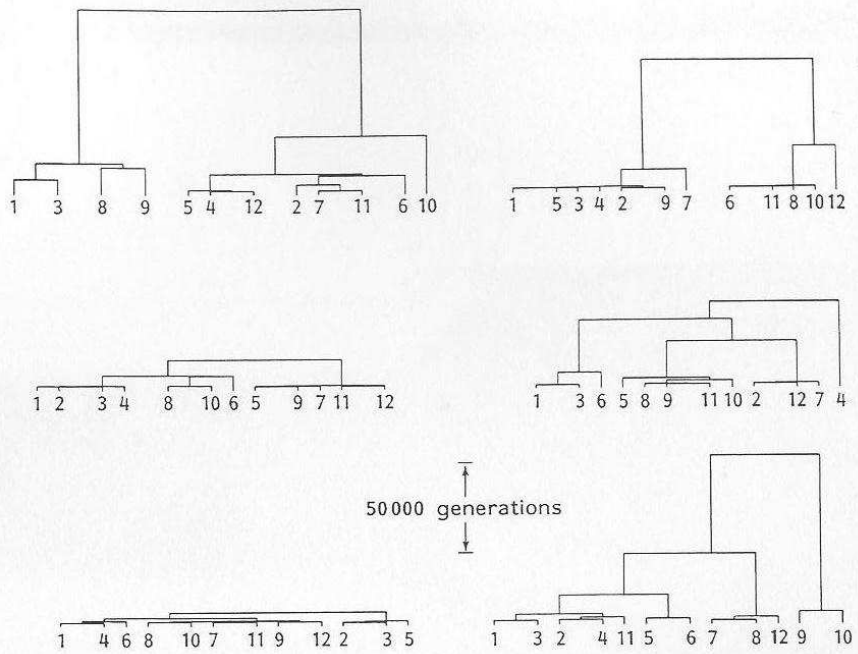

FIG. 5. Six realizations of the gene genealogy of a sample of size $n = 12$ genes from a geographically structured population. The population is structured according to a 3×3 stepping-stone model with migration probability of 10^{-4} per gene per generation. The population is assumed to have been at a constant size of 5000 individuals until 2500 generations ago, before growing exponentially to a current size of 250 000 000 individuals. The sample consists of six genes from each of two neighbouring colonies. Genes labelled 1–6 are from one colony. Those labelled 7–12 are from the second colony.

non-sampled colonies, and the top right tree in which gene 1 is descended from a recent migrant from the other sampled colony. There is also great variability in the depth of the trees in Fig. 4. This results from the variability induced by the migration process in the times for the final coalescences.

Figure 5 shows six genealogical trees for a population identical to that underlying Fig. 4, except that the migration probability has been increased to 10^{-4}. Again, the genes are sampled from two neighbouring colonies. The scale in Fig. 5 is different from that in earlier figures.

The increase in the migration rate has a substantial effect on the structure of the trees. Their shape is more similar to those from panmictic populations: they are dominated less by the final coalescence time and more by the times for the final two or three coalescences. Although genes sampled from the same colony are more likely to share recent common ancestors than those from different

colonies, this is no longer nearly as definitive as in Fig. 4. The increased migration rate means that in a relatively short time in the past the ancestral genes are scattered throughout the population. Final coalescences still rely on migration bringing ancestral genes together, so that the trees in Fig. 5 still tend to be much deeper, and exhibit much more variability in depth, than those for panmictic populations, although these effects are less marked than for lower migration rates.

Genealogical trees were also simulated under the same demographic conditions as in Figs 4 and 5, with the migration probability increased to 10^{-3} (P. Marjoram, personal communication 1995). These trees were similar in structure to those from a panmictic population of the same sizes. They were also more similar in total depth to the panmictic case than were the trees in Figs 4 or 5, but still somewhat longer, with an average depth of about 23 000 generations. The variability in this total depth was comparable with, or perhaps slightly less than, that for trees from panmictic populations.

Consequences for genetic variability

The patterns observed in genetic data from within populations reflect the effects of mutation at the locus in question superimposed on the genealogy of the sampled genes. Although it is not directly observed, the MRCA of the sampled genes will be of a particular allelic type. In the absence of mutation, all the sampled genes would be of this type. The variation actually present in the sample results from mutations along the lineages leading down from the MRCA to the sample.

This paper is confined to a qualitative discussion of genetic variability. In fact, one of the advantages of a genealogical approach is that it often greatly simplifies quantitative analysis. In addition, it provides an efficient and simple method for simulating neutral evolution for quite general mutation mechanisms. (For further details, see, for example, Hudson 1990, 1992, Donnelly & Tavaré 1995 and references therein.)

A mutation that occurs on a particular branch of the genealogical tree will affect all the genes in the sample which are descended from the ancestral gene associated with that branch. For example, under the 'infinite sites' assumption that no back mutation has occurred between the MRCA and the sampled genes, each mutation on the tree will correspond to a segregating site in the sample. For a particular such mutation, all the genes descended from the ancestral gene that underwent mutation will have one base at a particular site, whereas all the other genes will have a different base, that of the MRCA, at that site. Mutations that occur higher up the tree, closer to the MRCA, will tend to be represented in more of the sampled genes.

Patterns in the gene tree will thus tend to produce patterns in the genetic data. Consider, for example, the top left tree in Fig. 1. Because they share a

recent common ancestor, genes 1 and 2 will be similar or identical. The same is true for genes 3, 4, 5 and 6. However, unless the mutation rate is very small, genes 1 and 2 will tend to be different from genes 3, 4, 5 and 6, with the difference reflecting the mutations that occurred on the lineages leading from the MRCA to the common ancestor of genes 1 and 2, and from the MRCA to the common ancestor of genes 3, 4, 5 and 6. Thus, for such a gene tree, the data will tend to consist of two groups of genes, with great similarity within groups and possibly substantial differences between groups. Such an induced pattern is common for coalescent trees from constant-sized populations. For example, it is also likely to apply to each of the trees in the bottom row of Fig. 1.

Trees for populations that were of constant, non-trivial size before recent rapid growth will tend to resemble those from a similar population that has not grown, except that the tips will be lengthened by an amount similar to the number of generations since the onset of growth. If the relative effect of this lengthening is small, as it would be for populations whose size before growth was more than five to 10 times the number of generations since the onset of growth, the induced patterns will be similar to those for constant-sized populations. The effect of this lengthening is not insignificant for the demographic assumptions underlying Fig. 3. Nonetheless, for example, several of the trees may result in two distinct groups of alleles in the sample, for the reasons described in the previous paragraph.

In some sense, the total amount of diversity in the sample will reflect the total length of the tree. One consequence then of the substantial increase in tree depth caused (unless migration rates are large) by geographical subdivision will be a substantial increase in genetic diversity within the sample, compared to the panmictic case, for the same mutation rates. In addition, for small migration rates, subdivision can greatly accentuate the clustering of genetic types within the sample. For example, for the first five trees of Fig. 4, one would expect great similarity within genes sampled from the same colony, relative to the differences between the colonies (with the exception of gene 9 in the top left tree). Because of their increased depth, the trees in Fig. 5 should also result in greater diversity than in a panmictic population. The two trees in the top row will tend to induce a clustering of the sample into two distinct groups of genes, but now the groups will not correspond to the colonies from which the genes are sampled.

Figure 2 illustrates the effect of continual exponential growth, from a small size, in making gene trees star shaped. Such trees will tend to result in quite different patterns in genetic data from those in the other figures. For an appropriate range of mutation rates, the probability of at least one mutation between the root and a particular tip of the tree will be bounded away from 0 and 1. In this case, there may be a single group of identical genes in the sample, those which are identical by descent to the MRCA. The other sampled genes (or all genes in the sample for higher mutation rates) will tend to be 'equally

different' from each other. Of course the randomness inherent in the mutation process will mean that observed samples will not contain faithful reproductions of this 'equally different' property, and chance effects may result in some clustering amongst these genes.

Discussion

In contrast to the setting of classical statistics, single-locus genetic data from different individuals do not consist of independent observations. Rather, population genetic data of this sort is highly dependent because the genes in question share the same underlying genealogy. Recall that all the genes would be identical to the (random) type of their MRCA were it not for the effects of mutation since the MRCA.

Thus, there is limited information about the underlying evolutionary and demographic processes in genetic data of this kind. As a consequence, there is a premium on making maximal use of the information in the data. In any case, inferences based on the complete data, or suitable sufficient statistics, will be more efficient, and more reliable, than those based on other summary measures.

Unfortunately, many traditional genetic measures do not provide efficient summaries of the data. In particular, this is true of measures that are based on pairwise comparisons of the genes within a sample, such as sample heterozygosity or the average pairwise sequence difference. For example, in straightforward problems, such as the estimation of mutation rates under simple assumptions about mutation, estimates based on these measures are not even consistent, in contrast to those which make better use of the data. (For a more detailed discussion see Donnelly & Tavaré 1995.) Fortunately there has been exciting recent progress in the development of full likelihood-based inference procedures for these models (see Griffiths & Tavaré 1994a,b,c, Kuhner et al 1996).

There is substantially more independence between observations, and hence more information about evolutionary parameters, in a star-shaped genealogy than in the other tree structures described here. Provided sensible methods are used, inference in such a setting is, therefore, more reliable than in the other cases. Furthermore, under the demographic assumption of continual rapid growth from a small size, there is also much less variability in the shape and depth of the tree between realizations of evolution, and hence less variability in observed genetic data. This reduced variability also has the effect of substantially increasing the precision of inference procedures.

For growing populations, such star-shaped trees will only arise if the population has grown rapidly from a small size; for example, 500 or fewer individuals. This may be thought a priori to be unlikely for the human population, at least within the last 100 000 or 200 000 years; for example, in

view of the fossil evidence as to the spread of the population (Aiello 1993) and perhaps levels of diversity at certain loci (Takahata 1993). A severe bottleneck effect in the population, which reduces the population size to 100 or fewer individuals, will tend to produce the same effect. This may also be unlikely for a widespread population and it may be inconsistent with some observed levels of diversity. Star-shaped genealogies can also arise for a neutral locus that is closely linked to a selective locus at which a favourable allele sweeps through the population (P. Marjoram, personal communication 1995). The growth in the number of genes linked to the favourable allele will mimic population growth at the linked neutral locus, and the initial frequency of the favourable allele is small. Myself and others (Excoffier 1990, Di Rienzo & Wilson 1991, Marjoram & Donnelly 1994) have argued elsewhere that such a sweep, either at a locus on the mitochondrial genome or on the X chromosome, may be important in interpreting observed patterns in human mitochondrial data.

One traditional approach to population modelling under more realistic assumptions than panmixia and constant population size has been via the concept of effective population size. In this approach, instead of studying the model of interest, one approximates it by a panmictic population with a suitably chosen effective population size. There are several different definitions of effective population size. Loosely speaking, one focuses on some particular one-dimensional summary of the population and defines the effective population size to be the size of a panmictic population for which aspects of the chosen summary behave similarly to those for the more complicated population, over a fixed time horizon.

The above discussion of genealogy allows an assessment of the usefulness of this approach in various situations. For example, for a panmictic population of fixed size in which the variance of the number of descendant genes is different from the value of one in our coalescent approximation, use of an effective population size, defined as the actual number of genes divided by this variance, provides exactly the correct compensation. More generally, however, use of effective population sizes can be quite misleading, except possibly as an informal summary of the extent of genetic diversity in a population.

Gene trees in a constant-sized panmictic population are described by the coalescent. Changes to the value of the population size simply change the vertical scale in these trees. For example, if the population size underlying Fig. 1 was increased by a factor of 10, all that would be necessary would be to increase the vertical scale by a factor of 10. Coalescent trees have a certain intrinsic structure and associated variability. Patterns in genetic data reflect the structure of the underlying gene tree. It follows that if the demographic assumptions about a population are such as to induce gene trees that do not share the structure, or variability, or both, of coalescent trees, then there will be no value of an effective population size for which the patterns in the sample will resemble those from a constant-sized panmictic population.

The trees in Fig. 2 are fundamentally different from those in Fig. 1. There is no way that a linear change in the vertical scale of the coalescent trees in Fig. 1 will result in a close resemblance to those of Fig. 2. In other words, no choice of effective population size will give a good approximation to the patterns observed in populations that have undergone continual rapid growth. Similarly, there is no linear change of scale that will induce coalescent trees to resemble those of the structured population underlying Fig. 4. Again, approximation of the structured model via effective population size could be extremely misleading.

One consequence of recent progress on genealogical methods is that it is now possible to undertake reasonably sophisticated modelling of evolving neutral populations. An understanding of genealogical structure often provides invaluable insights. In some cases, even for relatively complicated models, analytical progress is possible. Most importantly, efficient simulation of such models is always possible via genealogy, so that it is in any case no longer necessary to resort to possibly misleading approximations.

With the exception of Fig. 2, one striking feature of the simulated genealogical trees within each figure is the variability that they exhibit. This is discouraging from the point of view of using genetic data to infer aspects of human population history. Changes in the underlying population demography induce changes in the distribution of the associated gene trees. In general, one is not able to observe the gene trees themselves. Rather, one sees the consequences of an additional random process, namely mutation, superimposed on the underlying gene tree. If one were able to see through the noise added by the mutation process and reconstruct genealogical trees exactly, then, from a single locus, one would have a sample of size one of an object (the tree) whose distribution depends in a complicated way on the underlying demographic process. Note that statistical methods which reconstruct gene trees from data do not have this property. Inference about the demography from this single observation is far from straightforward, and therefore one should be cautious.

This applies to demographic inference from data (regardless of the number of individuals involved) at a single locus, such as the mitochondrial genome. It is difficult to construct plausible demographic scenarios in conventional neutral models for the human population that are consistent with some aspects of observed human mitochondrial DNA (Marjoram & Donnelly 1994, 1996). Some of the inferences as to early human demography on the basis of mitochondrial data should be interpreted with caution until further information from nuclear loci is available.

Data from unlinked loci will have independent genealogies. Thus, although it is impossible to get around the correlations between individuals within a locus, one can gain independent samples from the underlying genealogy-generating mechanism by sampling different loci. Data of this kind for the human population, from a variety of nuclear loci, is now becoming available. The problems associated with the efficient, or even the systematic, use of these data remain open, and one needs to have a better understanding of the consequences for inference of the various simplifying assumptions made in evolutionary models. Nonetheless, the patterns in such data, if not traditional summary measures of them, are potentially extremely informative. They should prove enormously valuable in the understanding of our population's demographic history.

Acknowledgements

Thanks to Paul Marjoram for generating the trees presented in the paper and to Mitzi Nakatsuka for producing the figures. This work was supported in part by a Block Grant from the University of Chicago.

References

Aiello LC 1993 The fossil evidence for modern human origins in Africa: a revised view. Am Anthropol 95:73–96

Chakraborty R 1977 Distribution of nucleotide differences between two randomly chosen cistrons in a population of variable size. Theor Popul Biol 11:11–22

Di Rienzo A, Wilson A 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci USA 88:1597–1601

Donnelly P, Tavaré S 1995 Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421

Ewens WJ 1990 Population genetics theory—the past and the future. In: Lessard S (ed) Mathematical and statistical developments of evolutionary theory. Kluwer Dordrecht, Amsterdam, p 177–227

Excoffier L 1990 Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. J Mol Evol 30:125–139

Griffiths RC, Tavaré S 1994a Sampling theory for neutral alleles in a varying environment. Phil Trans R Soc Lond B Biol Sci 344:403–410

Griffiths RC, Tavaré S 1994b Ancestral inference in population genetics. Stat Sci 9:307–319

Griffths RC, Tavaré S 1994c Simulating probability distributions in the coalescent. Theor Popul Biol 46:131–159

Hudson RR 1990 Gene genealogies and the coalescent process. In: Futuyama D, Antonovics J (eds) Oxford surveys in evolutionary biology, vol 7. Oxford University Press, Oxford, p 1–44

Hudson RR 1992 The how and why of generating gene genealogies. In: Takahata N, Clark AG (eds) Mechanisms of molecular evolution. Sinauer, Sunderland, MA, p 23–36

Hudson RR, Kaplan N 1994 Gene trees with background selection. In: Golding GB (ed) Alternatives to the neutral model. Chapman Hall, London, p 140–153

Kaplan N, Darden T, Hudson RR 1988 The coalescent process with selection. Genetics 120:819–829

Kaplan N, Hudson RR, Langley CH 1989 The 'hitchhiking effect' revisited. Genetics 123:887–899

Kingman JFC 1982a On the genealogy of large populations. J Appl Probab 19:7A–43A
Kingman JFC 1982b The coalescent. Stochastic Processes Appl 13:235–248
Kingman JFC 1982c Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in probability and statistics. North-Holland, New York, p 97–112
Kuhner MK, Yamato J, Felsenstein J 1996 Applications of Metropolis–Hastings genealogy sampling. In: Donnelly P, Tavaré S (eds) Progress in population genetics and human evolution. Springer-Verlag, in press
Marjoram P, Donnelly P 1994 Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics 136:673–683
Marjoram P, Donnelly P 1996 Human demography and the time since mitochondrial Eve. In: Donnelly P, Tavaré S (eds) Progress in population genetics and human evolution. Springer-Verlag, in press
Marjoram P, Griffiths RC 1995 An ancestral recombination graph. In: Donnelly P, Tavaré S (eds) Progress in population genetics and human evolution. Springer-Verlag, in press
Takahata N 1993 Allelic genealogy and human evolution. Mol Biol Evol 10:2–22
Tavaré S 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor Popul Biol 46:119–164

DISCUSSION

Chakraborty: I would like to make two points. First, coalescent theory is mathematically rigorous and can be applied to many complicated gene histories. However, we cannot observe the coalescent tree without the superimposition of mutation events. If a DNA region is hypermutable, then can we say that star-shaped trees reflect the history of this region, or are they just artefacts of the mutation rate?

Second, I would refrain from using the term 'correlation' when talking about sequence similarities because it has a different connotation in the context of the genetic structure of populations. The shared evolutionary history of genes (detected by sequence similarities) should be regarded as different from the correlation of genes between individuals.

Donnelly: Let me take your second point first. This is purely a question of terminology. I did not refer to the term 'correlation' in the same way as human geneticists. Rather, I used the term in the same way as statisticians. The observation of one particular gene in a population is not independent of observations of other genes in the population. Additional information about ancestral history can be gained simply by increasing the sample size; however, each additional observation conveys less information than the previous one. In fact, in the context of coalescent theory, we do not get significantly more information when the sample size is increased from six to 100. Therefore, there's a trade-off between sequencing extra individuals for a given region and sequencing either a different region or a longer region and, in this case, one is better off doing the latter.

Your first point addresses whether star-shaped trees can be an artefact of high mutation rates. I am referring to the underlying ancestral tree, but your point is that this cannot be seen in isolation. The relevant question is that if one takes the tree generated with certain assumptions about demography, and mutations are superimposed on that tree, then what patterns will be observed? Consider the modelling of microsatellite mutation mechanisms. One can still see the effects of the tree. Trees that have a long period when there are only two ancestors will result in 'clumps' of similarly sized alleles in the sample. Highly mutable loci will also have clumps but the differences between the clumps will be larger under a generalized stepwise mutation mechanism, for example. If one draws a histogram of allele lengths in a population, then for microsatellite loci or minisatellite loci, there may be two peaks. This phenomenon is consistent with the clumping involved in coalescents. It does not require there to be two different mutation mechanisms.

W. Bodmer: I'm an old-fashioned population geneticist who hasn't quite come to terms with understanding coalescent theory. Although, intuitively, I would expect similar results from more conventional theories. Please can you clarify the nature of the simulation. How can you simulate gene genealogies without introducing mutations?

Donnelly: Let me explain this by illustrating an alternative simulation. I could just take demography into account in order to simply keep a track of how many descendants a particular gene has in the next generation.

W. Bodmer: But you're not talking about genes, you're talking about individuals.

Donnelly: At a given locus, I can talk about genes because a gene in a particular generation will be descended from one gene in the previous generation, although the genes will be in pairs in the individuals. Therefore, I can simulate the whole population by just keeping track of which genes are descended from which. I can go to the present and look at two particular genes, and then I can look backwards to see how far I have to go before I find that they're descended from a single ancestral gene.

W. Bodmer: You have to assume that each gene is different to start with. A gene has to be identified or somehow labelled because it's in a particular individual and not because of its sequence.

Donnelly: But a DNA sequence in a given region has an ancestral history that is related only to a particular demography on which mutations are being superimposed.

W. Bodmer: Two sequences could be the same. You are actually putting a label on an individual's gene without reference to the sequence of that gene.

Weiss: Wouldn't it have been simpler to say that you're constructing a population pedigree?

Donnelly: Yes, I'm labelling the genes by saying that each one is from a particular individual. They may be the same allele or they may be different. The pictures I drew are realizations of what we would see if we just traced this ancestral history.

W. Bodmer: Are your simulations forwards or backwards?

Donnelly: The simulations are backwards. I could have simulated the whole population and just looked at the relevant bits, but that's inefficient.

W. Bodmer: But you would get the same answers, so it's no different from conventional simulation, except in terms of efficiency, which I'm prepared to accept.

Donnelly: Yes, you're correct. The answers are no different from conventional simulation. However, the approach of focusing first on the gene tree and then asking how will mutations affect that gene tree is different from the usual population genetics approach. One of the advantages of this approach is that the dependencies in population genetics data are there not because of mutation, but because of the shared ancestral history. The patterns are the result of mutation on top of that.

Hartl: This coalescent process is the backwards version of the classical Wright–Fisher model (Hartl & Clark 1989). One of the results from the classical Wright–Fisher model is that mutation rates or migration rates cannot be decoupled from the effective population size because the governing parameter is the product. However, in your answer to Ranajit Chakraborty's question, you seemed to suggest that, by looking at gene genealogies, one could in fact decouple these, and I'm suspicious of that implication.

Donnelly: You have every right to be suspicious. That problem cannot be solved by looking at gene genealogies. Coalescent trees can be converted into real time trees, which requires an estimate of the population size, so that mutations can be superimposed with a real time mutation rate. If one halved the population size, so that the trees were half as deep and the mutation rate was doubled, one would observe the same pattern of variability.

Chakraborty: Irrespective of how inefficient the classic population genetic summary measures are, there are summary measures (e.g. heterozygosity, number of alleles conditioned on the sample size) that allow the decoupling of those parameters by combining data from different loci. Coalescent theory creates problems because different genes have different coalescent histories, which are due to the superimposition of different mutation processes on the same demography.

Donnelly: If one looks at two different, unlinked loci, although the demography of the underlying population of a group of individuals is the same, the gene genealogy is independent because of the time-scale on which it operates.

Dan Hartl also mentioned that coalescent theory is the backward genealogy of the Wright–Fisher model. I would like to give you some examples of insights

that can be obtained by focusing on this. Under the usual Wright–Fisher assumptions of random mating and constant population size, we shouldn't be surprised to observe two 'clumps' of alleles in the genetic data. The similarity within clumps and differences between the clumps are easier to define by looking at the trees than by using the forward equations of the Wright–Fisher diffusion. The patterns are there because of the effects of mutation on the trees, and so we can get some insights into the structure of the data by focusing on the shape of the trees. Also, if one uses the data to estimate the shape of the trees, then what do these shapes tell us about population history? One way to answer that is to ask the question the other way round, i.e. what tree shape would be expected under various sorts of assumptions?

Chakravarti: There are variations in coalescent patterns under any given set of assumptions, and these variations change if the assumptions are altered. It would be much more interesting to determine the coalescent corresponding to a region that has been sequenced. If you consider the sequences of different loci, are you suggesting that there is a single coalescent and the differences between the loci are due to different mutations, or are you suggesting that different loci will have an entirely different coalescent patterns?

Donnelly: At a single locus, the expected coalescent depends on the demographic assumptions. With the same assumptions for two unlinked loci, however, one would expect two independent trees. These independent trees would be star shaped if the population grew exponentially from a small value. Therefore, the loci are independent but the probability distributions change when the demographic assumptions change.

Chakravarti: Are you referring to the demography of the alleles or the populations?

Donnelly: I'm talking about the demography of the populations.

Chakravarti: So for a given set of populations it would be the same?

Donnelly: The probability distribution will be the same, but there would be independent choices from that probability distribution.

Sing: Using data from the French Canadian population we have estimated a star-shaped tree, or cladogram, for the *APOA1/CIII/AIV* gene cluster (Haviland et al 1995). This is consistent with your theoretical expectation. However, when we looked at the apoB gene, we found that it was not star shaped. Therefore, there are two genes within the same French Canadian population, which I believe expanded rapidly within the last hundred years, that have different patterns of allelic variations.

Clark: Is it possible that the founding population had greater genetic diversity at the *APOB* locus?

Sing: Yes. We're talking here in general about the expected differences in the shape of the trees between populations with different demographic histories. Our experience may reflect different allele demographies within a particular population.

Donnelly: If the founder size was two or three individuals, then all trees would be star shaped. If the founder size was a thousand individuals, then it is likely that no trees would be star-shaped. One possible explanation is that there is a range of founder sizes in-between where some trees would be star shaped and some wouldn't.

Scriver: The differences in shapes may also depend on where you did your sampling. Charlie Sing took samples only from French Canadians, so that a small number of people had a disproportionately large effect on the diversity that one sees today (Heyer & Tremblay 1995). However, if French Canadians from eastern and western regions of the province are studied, because they have different demographic histories in those two regions, there is likely to be genetic stratification.

Also, studying gene-related diseases is an important application of this work. It's not strictly a question of anthropology or population history per se.

Harper: The question of what can be inferred from data on a single locus becomes extremely important when dealing with genetic diseases. This was apparent at the European Science Foundation meeting in Strasbourg (November 1993) on genes and genetic diseases in European populations, where there was some confusion between clinical geneticists and population geneticists. A number of people presented valuable data on disease genes in different populations. They traced the spread and evolution of these disease genes with great accuracy but then they proceeded to generalize from the behaviour of these disease genes to the migration and development of entire populations. However, the spread of different disease genes produced different patterns, so that making generalizations about the behaviour of the whole population became impossible. These results also contrasted with more broadly based population studies that involved several loci. Therefore, studies of disease genes are exceptionally valuable in terms of their own particular locus, and they may also form a part of the history of the population, but a multilocus approach is definitely necessary. Also, because of selective forces and other influences that may be involved, the study of disease genes may not give the most accurate answers about the development of the whole population.

Weiss: It's important to look at whether every disease must be studied strictly on its own merits, or whether generalizations can be made that will help in the study of the next disease. How can selection be incorporated into your coalescent approach?

Donnelly: We know a little bit about genealogy for certain sorts of selection. The advantage of the coalescent approach is that in a neutral model, one can focus first on genealogy and then superimpose the genetic types. However, it is not possible to do this with selection because the genetic types are affecting reproductive success and hence demography. The general effect of selection on genealogical structure is really an unsolved problem.

Chakravarti: But wouldn't you expect that alleles which are being strongly selected against would have a short history, so that they probably wouldn't affect the coalescent in a serious way?

Donnelly: Unless the mutation rates are high enough so that, at any given time, some of those alleles are present.

Chakravarti: This suggests that the major effects will be either for deleterious recessive alleles, because they are sheltered within surviving populations for long periods of time, or for common disease-predisposing alleles that have a small effect. The latter alleles may even be the common gene polymorphisms that have survived within the human population for a long period of time. Is it possible that weak selection, either for or against, for a long period of time is likely to have a stronger effect on the coalescent?

Donnelly: It's possible because the time-scale is long and the effect of genetic drift is weak.

Chakravarti: Summary measures, such as heterozygosity and numbers of alleles, have supported the neutral theory. If all alleles are subject to small levels of selection throughout evolution, then are the trees going to look considerably different? In other words, is coalescence going to detect these effects?

Donnelly: If the null hypothesis is a constant-sized population with random mating, then we can do statistical tests with those assumptions. However, problems may arise when tests for neutrality result in the rejection of the null hypothesis not because the neutral assumption has changed but because the other assumptions have changed.

Hartl: I suspect that part of the reason that neutral theory looks good from the summary statistics point of view is that it lacks the power to detect departures from neutrality. In my opinion, if we want to find out the effects of selection on particular genes in the human genome, then we have to look not only at human polymorphisms, but also at the divergence between human and primate genes in comparison with the level of polymorphism in other primates.

Kidd: We have found that, relative to chimpanzees and gorillas, humans are depauperate of genetic variation on a species-wide level. We have exhaustively examined a DNA sequence about 1 kb long, and we have found more common polymorphisms in a few chimpanzees and gorillas than in several-fold larger samples of humans stratified to represent the whole species (Ruano et al 1992, Deinard & Kidd 1995). In a less exhaustive study, we have also found higher rates of polymorphism in a single troop of baboons than in humans (Rogers & Kidd 1993, 1995)

I would like to return to Peter Donnelly's simulations, where he assumed a single demography and had a torus-shaped migration matrix model, and ask whether he has attempted to simulate the actual history of humans. For example, it may be possible to start with the original model, allow the simulation to run for a few generations, and then look at the population in one

corner of the 3 × 3 matrix. This corner could be expanded into one corner of a new 3 × 3 matrix that is gradually filled by expansion of that population. This situation could represent the migration out of Africa, where there is a clear founder effect with some migration across the point where the two matrices join, and also migration (expansion) into the new matrix representing Eurasia. The opposite corner of this second matrix could even be expanded into a third matrix that would represent the New World. Migration would occur within that region, but there would only be a little migration between matrices through the connecting corner. How can the coalescent model explain this in a backward simulation?

Donnelly: The short answer is that I haven't looked at this. The more substantial point you're making, which I'm in entire agreement with, is that the traditional models cannot explain the spread of major human populations. We need to study populations that have smaller groups of migrating individuals. One approach is to think of something realistic and simulate it, but there would be too many variables, so it would be difficult to interpret the conclusions. Another approach is to start with the simplest model, make it slightly more general and look at the quantitative effects, then make it slightly more general and so on. The coalescent story is going through this process, but it is only part way through and not enough is known to study human populations in realistic models.

Weiss: Rogers & Harpending (1992) have performed similar simulations with mismatched distributions. They have evidence for relatively recent (few tens of thousands of years) geographic expansions in every major region of the world. In common with other investigators' models, their model finds that even a small amount of migrations among regions effectively homogenizes different populations.

Donnelly: The approach of picking up signatures of population expansions through pairwise differences has at least three problems. The first is that just looking at pairwise differences can ignore other information in the data. The second is that some population growth models may not be realistic and completely different conclusions may be drawn from more realistic models of human evolution. The third is that the patterns that one sees as a consequence of population growth are broadly the same patterns as one sees as a consequence of selective sweeps. This is particularly applicable to the mitochondrial data, which is what the majority of this work is focused on. One would have to be brave to rule out selection on mitochondria. It's difficult to fit realistic neutral models to the human mitochondrial data.

Weiss: The Rogers–Harpending model is at an early stage but they're also trying to relate it to archaeological evidence for relevant cultural advances or other kinds of ancillary data.

W. Bodmer: It's not difficult to incorporate selection into these models. Luca Cavalli-Sforza made a fundamental point many years ago, which was that

whatever the population structure, if there are families of genes that are behaving differently, then they also have to be influenced differently by selection (Cavalli-Sforza & Bodmer 1972). Results from the analysis of the HLA system support this statement: the pattern of variation at the DNA level in non-synonymous versus synonymous substitutions in particular parts of the genes is so different from other parts of the genes that there has to be selection for one and not for the other (Bodmer et al 1986). At the DNA level, it is possible to define neutral variations in most regions; for example, intron regions, flanking regions, CA repeats and synonymous positions. Therefore, it is possible to define families of differences which are neutral, look for the similarities amongst those and then look for differences in the pattern of variation of sequences that may be subject to selection. The pattern of selection for HLA, for example, is probably a recurrent frequency-dependent selection that selects different variants at different times and can explain the maintenance of combinations of differences that go back through evolutionary time.

Bowcock: Cavalli-Sforza, Ken Kidd and I did some work a few years ago on restriction fragment length polymorphisms (RFLPs) that are mainly derived from non-coding regions, and we found that as many as 30% of the alleles could be subject to selection (Bowcock et al 1991).

W. Bodmer: But if one looks at RFLPs within the HLA region, RFLPs that are defined by polymorphic genes have a different pattern than those defined by genes that are not polymorphic. Therefore, if RFLPs are picked up by functional genes, they are in linkage disequilibrium.

Kidd: I would like to amplify the point that Anne Bowcock made. We found that no more than 30% of the loci are subjected to selection, assuming neutrality and taking the population relationship structure that we inferred from the data. In other words, we found higher values of F_{ST}, the standardized variance, than we expected on our simple model for about 30% of the loci.

W. Bodmer: But you shouldn't base these calculations on any a priori assumptions. If you're looking at F_{ST} values, you should forget about any model and simply ask whether there is any evidence for bimodality or multi-modality. And if there is, then at least one category has to have been subjected to selection.

Kidd: But we found a distribution skewed towards higher F_{ST} values at more loci than we would have expected.

W. Bodmer: What is the source of the sequences that you are using for picking them up? Most RFLPs are based on cDNA sequences. If one were to take a cosmid containing a complete gene that isn't polymorphic, for example the HLA-DRA gene, then one would obtain a different pattern of RFLPs than one would if one took a polymorphic gene, for example HLA-DRB1. In the first case little or no variation is detected, whereas in the second there is extensive variation that correlates with HLA-DR serology because of linkage disequilibrium.

Clark: I agree with Walter's comment that comparing data from different loci can be extremely informative. The Hudson–Kreitman–Aguadé test, in the field of *Drosophila* molecular population genetics is based on this idea (Hudson et al 1987). It examines the levels of polymorphism and interspecific sequence divergence for pairs of loci. Under neutrality, one expects the pattern of polymorphism and divergence to be similar across loci, and divergence from this can be detected by using the χ^2 test. However, Peter Donnelly was incorrect to say that no work on the coalescent with selection has been done. Some work has been done on the coalescent properties of a gene subjected to selection. For example, John Gillespie (1989) showed that his SAS–CFF (stochastic additive scale–concave fitness function) model can produce a neutral coalescent. Also Takahata (1990) showed that under conditions of strict symmetrical over-dominance, one still gets an expected neutral coalescent, i.e. a geometric distribution of the time back to a common ancestor, except that the time-scale is expanded. He also showed that the time-scale can be calculated, based on diffusion. The expected time depth of the coalescent is deeper for a symmetrical overdominance model than under neutrality. This is also true for self-incompatibility loci (Clark 1992), which show a pattern of selection that produces allelic variation having geometric distribution of coalescence times, but with extraordinarily ancient times of coalescence. Coalescence times are so ancient, in fact, that they pre-date speciation, and trans-species polymorphisms are observed (Ioerger et al 1990, Clark & Kao 1991).

A second example is the work of Dick Hudson and Norm Kaplan (Hudson & Kaplan 1995). They proposed a hitch-hiking situation where there is a neutral locus with deleterious mutations hitting the genome at linked sites, and again a neutral coalescent is expected. This represents one situation where the coalescent relates to the population size because the removal of deleterious variation elsewhere in the genome reduces the effect of population size and, therefore, reduces the expected coalescent for that neutral site.

Hartl: I would like to raise to Walter's bait by first conceding a point, and then ask a question.

I don't doubt that a number of methods can be deployed for detecting selection coefficients on the order of 10-fold the effective population size, which is probably what's occurring with sickle-cell anaemia and β-thalassaemia. Selection coefficients that are on the order of the reciprocal of the effective population size may be more common than larger selection coefficients. My question is: are selection coefficients on the order of the reciprocal of the effective population size of any interest or are we only interested in larger values that might be of some immediate clinical significance?

W. Bodmer: We do not fully understand relatively large effects, so we should look at those first. We should also take note of the power that we have to decide what is likely to be neutral a priori at the DNA level. We are in a novel situation. Previously, we only had results from gel electrophoresis and blood

group differences, and we could not be sure that selection was not operating. We should analyse the exact nature of the differences at the DNA level, and use that to determine the neutral situation, whatever model it may fit.

Weiss: There's a difference between the seriousness of a disease and whether it has any selective effects on biochemical communities. Many serious diseases are treated because they reduce your physical 'fitness' at a certain age but this doesn't mean that they have a evolutionary effect, i.e. that they affect your Darwinian fitness.

Harper: Small to moderate selective effects are also important, especially in recessive disorders. For instance, cystic fibrosis and phenylketonuria are two classic examples where people have argued whether heterozygotes have a selective advantage, and if they do, what is it or what might it have been in the past. The fact that these issues are difficult to resolve means that they're worth studying.

Scriver: The cystic fibrosis story has developed sufficiently for us to propose the process and target of selection, and perhaps even the historical period over which the selection process could have taken place. One problem with the phenylketonuria story is that we do not know what the selective agent may have been in the environment, and we do not know the target phenotype in the host. Therefore, we can only begin to guess how old or how current the process might be—if selection has played any role at all.

Beighton: Part of the problem is to define the nature of the selective forces, and whether they are positive or negative. We tend to look only at the present and not in the past. For instance, Gaucher disease in the Ashkenazi Jewish population is found at a high frequency in South Africa because of a founder effect, which reflects the earlier situation in Lithuania (Goldblatt & Beighton 1979). The question then arises as to what has been going on in Lithuania over the past 2000–3000 years.

Also, we regard selection as a gradual generation-to-generation process but is a big-bang, one-off event possible instead? Take Gaucher disease again as an example. It is possible that, in the middle ages, the plague decimated the population of Europe, but did not kill off Gaucher disease heterozygotes selectively. Therefore, within one generation the gene frequency could increase.

W. Bodmer: That idea has been around for a long time in the field of HLA and disease.

References

Bodmer WF, Trowsdale J, Young J, Bodmer J 1986 Gene clusters and the evolution of the major histocompatibility system. Phil Trans R Soc Lond B Biol Sci 312:303P–315P

Bowcock A, Kidd J, Mountain J et al 1991 Drift, admixture and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Cavalli-Sforza LL, Bodmer WF 1972 The genetics of human populations. Ann Hum Genet 36:239–240

Clark AG 1992 Evolutionary inferences from molecular characterization of self-incompatibility alleles. In: Takahata N, Clark AG (eds) Mechanisms of molecular evolution. Sinauer, Sunderland, MA

Clark AG, Kao T-H 1991 Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc Natl Acad Sci USA 88:9823–9827

Deinard AS, Kidd KK 1995 Levels of DNA polymorphism in extant and extinct hominoids. In: Brenner S, Hanihara K (eds) The origin and past of modern humans as viewed from DNA. World Scientific, Teaneck, NJ, p 149–170

Gillespie JH 1989 Molecular evolution and polymorphism: SAS–CFF meets the mutational landscape. Am Natural 134:638–658

Goldblatt J, Beighton P 1979 Gaucher disease in South Africa. J Med Genet 16:302–305

Hartl DL, Clark AG 1989 Principals of population genetics. Sinaeur, Sunderland, MA

Haviland MB, Kessling AM, Davignon J, Sing CF 1995 Cladistic analysis of the apolipoprotein *AI-CIII-AIV* gene cluster using a healthy French Canadian sample. I. Haploid analysis. Ann Hum Genet 59:211–231

Heyer E, Tremblay M 1995 Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. Am J Hum Genet 56:970–978

Hudson RR, Kaplan N 1995 Deleterious background selection with recombination. Genetics 141:1605–1617

Hudson RR, Kreitman M, Aguadé M 1987 A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159

Ioerger TR, Clark AG, Kao T-H 1990 Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc Natl Acad Sci USA 87:9732–9735

Rogers AR, Harpending HC 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552–569

Rogers J, Kidd KK 1993 Nuclear DNA polymorphisms in a wild population of yellow baboons (*Papio hamadryas cynocephalus*) from Mikumi National Park, Tanzania. Am J Phys Anthropol 90:477–486

Rogers J, Kidd KK 1995 Nucleotide polymorphism, effective population size and dispersal distances in the yellow baboons (*Papio hamadryas cynocephalus*) of Mikumi National Park, Tanzania. Am J Primatol, in press

Ruano G, Rogers J, Ferguson AC, Kidd KK 1992 DNA sequence polymorphism within hominoid species exceeds the number of phylogenetically informative characters for a HOX2 locus. Mol Biol Evol 9:575–586

Takahata N 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc Natl Acad Sci USA 87:2419–2423