Linkage disequilibrium: what history has to tell us

Magnus Nordborg and Simon Tavaré

Linkage disequilibrium has become important in the context of gene mapping. We argue that to understand the pattern of association between alleles at different loci, and of DNA sequence polymorphism in general, it is useful first to consider the underlying genealogy of the chromosomes. The stochastic process known as the coalescent is a convenient way to model such genealogies, and in this paper we set out the theory behind the coalescent and its implications for understanding linkage disequilibrium.

Magnus Nordborg*
Simon Tavaré†
Molecular and Computational Biology, University of Southern California, 835 W 37th St, SHS172, Los Angeles, CA 90089-1340, USA.
*e-mail: magnus@usc.edu
†e-mail: stavare@usc.edu

Linkage disequilibrium (LD), or the nonrandom pattern of association between alleles at different loci within a population, has recently received much attention. This is primarily because population associations are potentially useful in the fine-scale mapping of human disease loci [1,2], but also because of the increasing use of haplotype data (e.g. DNA sequence polymorphism) as the basis for historical or evolutionary inference. The statistical aspects of LD mapping are reviewed elsewhere [2,3]. In this paper, we aim to introduce the population genetics theory that is necessary for an understanding of LD and haplotype data.

A genealogical view of LD

Linkage disequilibrium refers to the nonrandom association of alleles in haplotypes. Such associations underlie all forms of genetic mapping. However, whereas linkage analysis is based upon associations in well-characterized pedigrees, LD refers to associations within populations of 'unrelated' individuals. There is nonetheless a close relationship between the two approaches, because the 'unrelated' individuals in a population are unrelated only in a relative and approximate sense. In general, chromosomes sampled from 'unrelated' individuals in a population will be much more distantly related than those sampled from members of traditional pedigrees. This is precisely what makes LD mapping suitable for fine-scale mapping: there will have been more opportunities for recombination to take place. Whereas pedigree studies work with recombination events that exchange megabase chunks of chromosomes, LD studies deal with segments measured in kilobases.

Figure 1 shows a genealogy of three copies of a short chromosomal segment. The segments are traced backward within the pedigree of the individuals. Recombination events allow the resulting pieces to have different genealogical trees. Thus, the genealogy of a sample of chromosomal segments is usually a graph rather than a tree. In linkage mapping the pedigree is known, making it possible to model the genealogy of the chromosomes using basic genetics [4]. When dealing with unrelated individuals, however, the pedigree is almost completely unknown, and different models are needed to take into account this additional level of uncertainty.

So far, we have discussed only the genealogy of chromosomes. There are no alleles in Fig. 1. How does the genealogy relate to LD? Polymorphism in a sample is the result of mutations along the branches of the genealogy that relates the sampled sequences. The genealogies of closely linked loci will tend to be highly correlated, whereas those of distantly linked loci will be effectively independent of one another. Hence, the allelic states of closely linked loci will be correlated (i.e. in LD), whereas those of distantly linked loci will be more-or-less independent.

Figure 2 shows a simple example of particular relevance to LD mapping. All existing copies of a unique mutation can be traced back to the most-recent common ancestor (MRCA) of that mutation. Each haplotype that contains the mutation must also have inherited from the MRCA a small piece of chromosome surrounding the mutation. Different haplotypes might carry different pieces, depending on the history of recombination. Alleles at polymorphic marker loci within the region covered by these pieces of chromosome will be in LD with the mutation. The length of the region depends on the age of the MRCA, and the strength of the association depends on the age of the marker mutation (and the recombination rate and many other factors). The age of the MRCA and that of the marker mutation are reflected in the respective allele frequencies. We shall return later to the frequency dependence of LD.

For the purpose of LD mapping, the pattern of association between the polymorphic markers is of interest only because it contains information about the pattern of recombination. If the recombination history could be inferred directly, there would be no need to rely on genetic markers [5]. LD mapping can therefore be viewed as a missing data problem, where the underlying, unobserved genealogy contains all the information. This suggests that our thinking about LD should focus on descent (i.e. genealogy) rather than on state (i.e. marker mutations). In particular, the traditional pairwise measures of association (Box 1) are not likely to be informative.

What does history tell us about LD?

To understand the properties expected of a sample of haplotypes, we must model their history. The coalescent with recombination provides a flexible and powerful method for doing this (Box 2). Although the coalescent can be adapted to include almost any

http://tig.trends.com 0168-9525/02/$ – see front matter PII: S0168-9525(01)02557-4

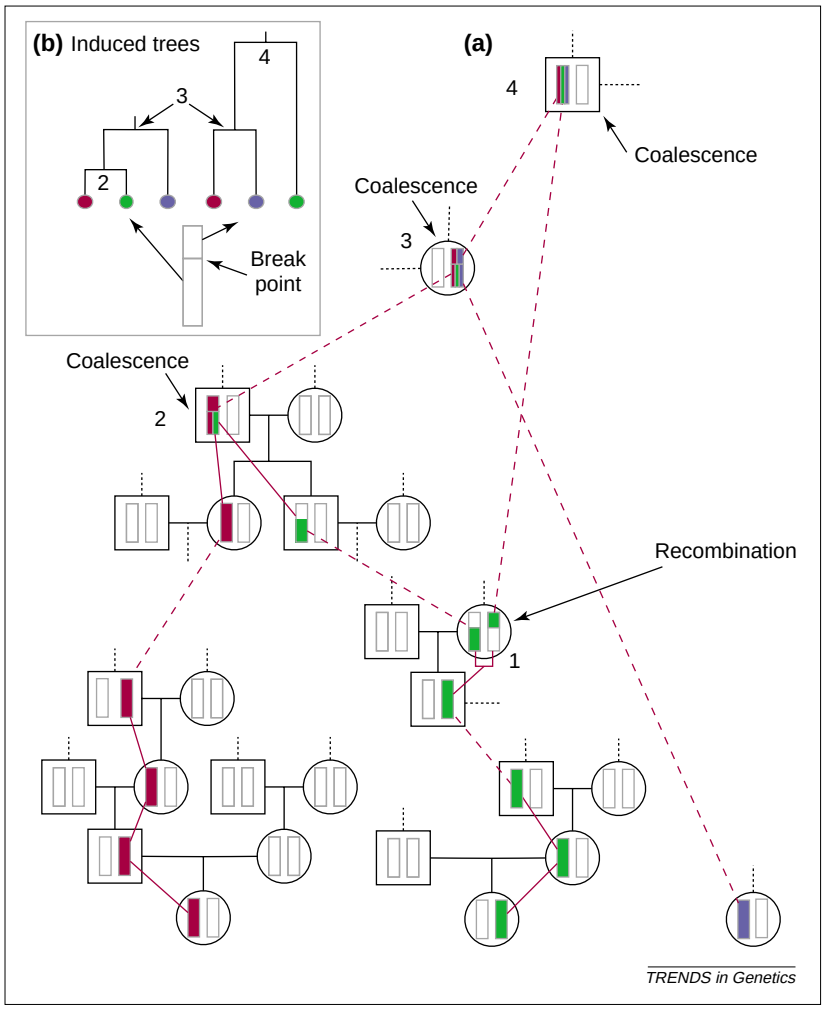

Fig. 1. A genealogy for three copies of a short chromosomal segment. (a) Tracing the segmental lineages back in time, we observe the following events: (1) the 'green' lineage undergoes recombination and splits into two lineages, which are then traced separately; (2) one of the resulting green lineages coalesces with the 'red' lineage, creating a segment, part of which is ancestral to both green and red, part of which is ancestral to red only; (3) the 'blue' lineage coalesces with the lineage created by event 2, creating a segment that is partially ancestral to blue and red, partially ancestral to all three colors; (4) the 'other' part of the green lineage coalesces with the lineage created by event 3, creating a segment that is ancestral to all three colors in its entirety. (b) The recombination event induces different genealogical trees on either side of the break.

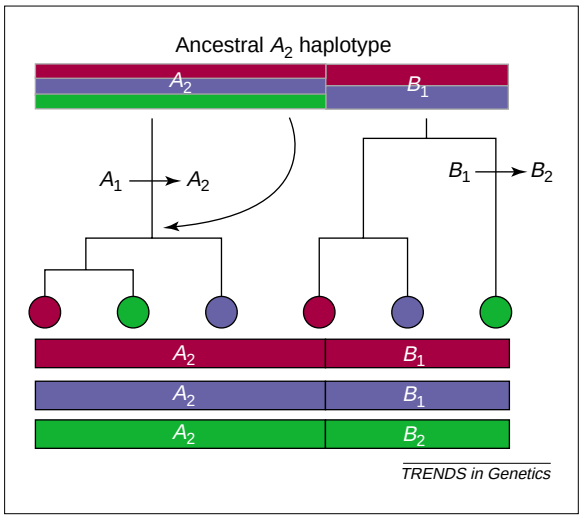

Fig. 2. Genetic polymorphism is the result of mutations along the branches of genealogical trees. The genealogical trees for linked chromosomal positions will not, in general, be statistically independent. Therefore, neither will the allelic states of linked loci be statistically independent, that is, there will be linkage disequilibrium between the loci. This is illustrated here by adding two unique mutations to the trees from Fig. 1. The mutation at locus *B* occurred earlier than the mutation at locus *A*. The A_2 allele therefore arose in a population that was polymorphic with respect to locus *B*. The most recent common ancestor of the A_2 alleles carried a B_1 allele. Two of the three haplotypes shown here still do; the third recombined with a haplotype carrying a B_2 allele (as depicted in Fig. 1).

biological scenario, here we use the basic version. Most of the following conclusions stem simply from the existence of an underlying genealogy, and do not depend on details of the model.

Variability of pairwise LD

Pairwise LD is expected to be extremely variable. This variation is attributable both to the history of recombination and to the history of mutations. To understand the difference between the two, consider the case where there is no recombination, so that the haplotypes are related by a single genealogical tree. As illustrated in Fig. 3, it is nonetheless possible for pairwise measures of LD to vary considerably, simply because of the history of mutation and coalescence. Note that, as there is no recombination, this variability does not reflect physical distance and is therefore uninformative for mapping purposes. Figure 4 shows the result of a coalescent simulation that illustrates the same phenomenon. Association studies have often found markers closely linked to the focal (disease) mutation that show less LD than do more distantly linked ones. Given that much of the variation in LD might not reflect recombination at all, this should not be surprising.

For mapping purposes, it is the recombination history that is important. The mutations are of interest only to the extent that they reveal something about this unobservable history. However, the recombination history is itself variable, and the pattern of LD reflects this too.

The resulting variability is evident in the simulated dataset shown in Figs 5 and 6. The top row of panels in Fig. 5 shows the decay of haplotype sharing with respect to particular focal mutations (disease mutations, say), and the remaining panels illustrate how this underlying genealogical pattern is reflected by markers through two widely used pairwise measures of LD. The relationship between LD and distance is far from smooth, in agreement with what is typically observed. Figure 6 shows the behavior of LD when all pairwise comparisons between sets of loci are made. This illustrates that LD is expected to vary between chromosomal regions as well as between pairs of loci.

It is clear that, although there is a statistical relationship between the LD measure and distance, the variability is too great for reliable estimates to be made from any particular pair of loci. It will rarely be possible in practice to estimate the genetic distance between two markers using LD [6,7]. Note that this is different from the use of multiple markers to map a particular locus [8–12]. It is evident from Fig. 5 that

Box 1. Pairwise measures of linkage disequilibrium

Llinkage disequilibrium (LD) is often quantified using statistics of association between the allelic states at pairs of loci. Consider two loci, *A* and *B*, with alleles A_1/A_2 and B_1/B_2, respectively. Let p_{A_i} stand for the frequency of allele A_i, where $i = 1, 2$, at locus *A*, and similarly for locus *B*. Let $p_{A_iB_j}$ stand for the frequency of the A_iB_j haplotype. Examples of measures of LD used in the present review are:

- $|D'|$, the absolute value of $D = p_{A_1B_1} - p_{A_1}p_{B_1}$, normalized to take values between 0 and 1 regardless of the allele frequencies;
- $r^2 = D^2/(p_{A_1}p_{A_2}p_{B_1}p_{B_2})$, the squared correlation in allelic state between the two loci as they occur in haplotypes.

Both these measures are symmetric, in the sense that it does not matter which allele is associated with which, or, in the context of mapping, which locus is the disease locus and which is the marker locus. A measure for which the latter is not true is:

- $d^2 = (p_{A_2B_1}/p_{A_2} - p_{A_1B_1}/p_{A_1})^2$, which measures the association between the alleles at (marker) locus B and the alleles at (disease) locus *A*.

All these measures are closely related to each other and to the standard χ^2-statistic for a 2×2 contingency table [a,b]. They nonetheless have different properties, as will become clear later.

When 'significant LD' is discussed, it is usually in the sense of a simple contingency-table test of association [c]. However, significant LD can be found even between unlinked loci – because of population structure, for example. In general, the view taken in this paper is that pairwise measures of association make poor use of modern, multilocus data, and their usefulness for mapping purposes is questionable.

References

a Guo, S-W. (1997) Linkage disequilibrium measures for fine-scale mapping: A comparison. *Hum. Hered.* 47, 301–314

b Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* 69, 1–14

c Weir, B.S. (1996) *Genetic Data Analysis*, Sinauer

Box 2. Modeling haplotypes using the coalescent

What is the coalescent?

The coalescent is a stochastic process that describes the history of recombination and coalescence in a sample of *n* homologous sequences [a,b]. For a detailed description, see Refs [c,d]. Of particular relevance here is a version of the coalescent known as the ancestral recombination graph, which allows the modeling of any number of loci, and arbitrary recombination mechanisms [e]. The ancestral recombination graph generates random genealogical graphs of the type depicted in Fig. 1. The behavior of the process depends on the recombination parameter ρ, which determines the rate at which ancestral lineages undergo recombination.

What about mutations?

Random neutral mutations are superimposed on the graph as in Fig. 2. The rate at which mutation events occur along the edges of the graph is determined by the mutation parameter μ. The mutation mechanism itself is arbitrary, and can accommodate microsatellites, single nucleotide polymorphisms (SNPs), etc.

Why is it useful?

The coalescent is useful as a tool for simulating data [c,d]. Each replicate of an ancestral recombination graph with mutations added provides a sample from the evolutionary model, just as in classical forward simulations. However, the coalescent has several advantages, of which the primary one is that only information about those ancestors who leave genetic traces in the sample need be recorded. We do not need to simulate the entire population from which we sample. Coalescent simulations typically result in enormous increases in speed and efficiency. Furthermore, unlike classical methods, the coalescent provides a natural framework for calculating likelihoods for samples, and therefore for statistical inference [f].

Which parameters?

Time in the ancestral recombination graph is measured in units of the effective number of chromosomes in the present generation, and the parameters ρ and μ are scaled accordingly [d]. A direct estimate of μ can be obtained from sequence data. Typical estimates for humans are ~10^{-3} per bp [g]. It is much harder to estimate the recombination parameter ρ from polymorphism data. An alternative is to use the fact that the ratio ρ/μ is equal to the ratio of the per generation probabilities of recombination and mutation, which can be estimated directly. A reasonable guess for humans is that $\rho = \mu = 100$ for a 100-kb region.

What is assumed?

The coalescent can be adapted to incorporate a broad range of biologic scenarios. Founder effects can be modeled by appropriate assignment of types to ancestors or by simulating for a fixed period. Alternative recombination models (e.g. hot-spots and gene conversion), different mutation mechanism (e.g. variable rates), variable population size, selfing, age structure, diploidy, variable reproductive success, and population subdivision have all been studied [c,d]. Modeling the effects of selection is more difficult, and this is an active area of research [d].

References

a Kingman, J.F.C. (1982) On the genealogy of large populations. *J. Appl. Prob.* 19A, 27–43

b Hudson, R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23, 183–201

c Hudson, R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* 7, 1–43

d Nordborg, M. (2001) Coalescent theory. In *Handbook of Statistical Genetics* (Balding, D.J. *et al.*, eds), pp. 179–212, Wiley

e Griffiths, R.C. and Marjoram, P. (1997) An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution* (Donnelly, P. and Tavaré, S., eds), pp. 257–270, Springer

f Stephens, M. (2001) Inference under the coalescent. In *Handbook of Statistical Genetics* (Balding, D.J. *et al.*, eds), pp. 213–238, Wiley

g Przeworski, M. *et al.* (2000) Adjusting the focus on human variation. *Trends Genet.* 16, 296–302

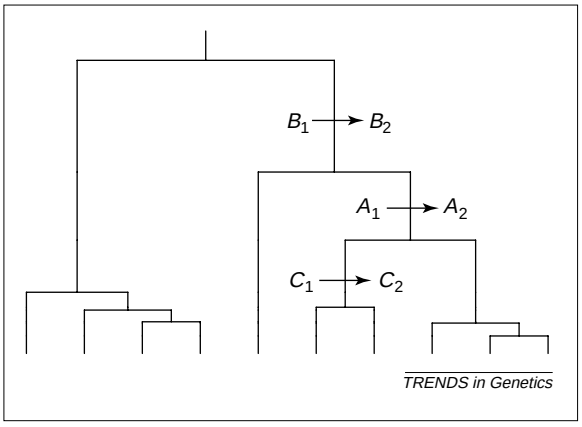

Fig. 3. Most pairwise measures of linkage disequilibrium are highly variable even in the absence of recombination. Three mutations at three loci are shown here superimposed on a genealogical tree for a sample of size ten. Imagine that we are trying to map locus *A* using the polymorphisms at the other loci as markers. Consider, for example, d^2 (Box 1): for locus *B*, we have $d^2 = (0 - 4/5)^2 = 0.64$, but for locus *C*, we have $d^2 = (3/5 - 1)^2 = 0.16$.

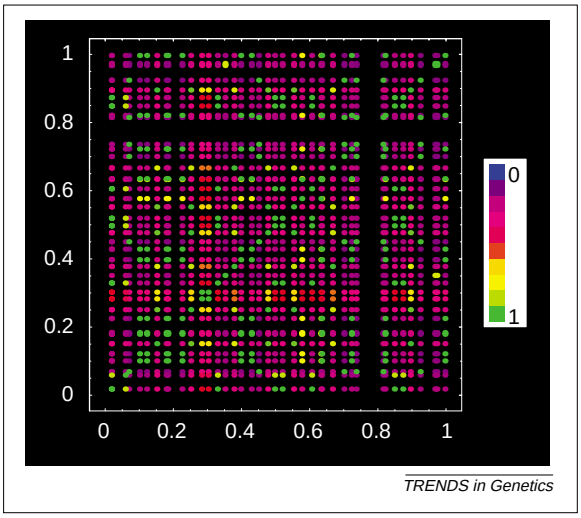

Fig. 4. A further illustration of the behavior of linkage disequilibrium (LD) in the absence of recombination (cf. Fig. 3). Each dot in the figure represents a comparison between a pair of dimorphic marker loci, the chromosomal position of which (in arbitrary units) can be read off the axes. Hence, dots along the diagonal represent comparisons of loci with themselves. The color of each dot represents the strength of LD, measured here as $|r|$ (Box 1). Because this measure is symmetric with respect to the two loci, the figure is symmetric about the diagonal. The figure was calculated from a sample of size $n = 50$, simulated using a standard coalescent with $\mu = 100$. Markers at less than 10% frequency were excluded to allow comparison with Fig. 6. For clarity, only a randomly selected subset of the remaining sites is shown.

the joint pattern of LD exhibited by all the markers might contain considerable information about the position of the locus being mapped.

Ages and frequencies

The behavior of LD depends strongly on the coalescence time of alleles, which under most models will be reflected in their frequencies. From the point of view of LD mapping, Fig. 5 illustrates the key part played by the frequency of the allele being mapped in determining how quickly LD with linked markers decays. As is shown in the top panels, this decay reflects the altered size distribution of the ancestral segment shared by the haplotypes that carry the disease allele. Disease alleles with a younger MRCA will usually be surrounded by a much larger ancestral haplotype. Most successful applications of LD mapping to date have been for rare diseases. If alleles underlying common diseases are more common, they will probably also be much older, and in these cases a very dense map may be needed.

Figure 5 also shows that the extent of LD depends on the ages/frequencies of the markers in addition to those of the disease. The sensitivity to allele frequencies varies between measures. As is further illustrated in Fig. 6, $|D'|$ appears to be much more sensitive than d^2. The reason for this is that $|D'|$ has been normalized to lie between 0 and 1 regardless of allele frequencies. This has some desirable consequences: for example, $|D'| = 1$ between all loci in the absence of recombination (as in Fig. 4). However, it also means that $|D'|$ will tend to be high if there is a young allele at either of the two loci. Therefore, the high values of $|D'|$ in the upper right panel of Fig. 5 do not reflect conservation of the ancestral disease haplotype, but conservation of ancestral marker haplotypes. As we stated earlier, LD depends on the history of mutation and coalescence, and not only on the history of recombination. A consequence is that all measures of LD are frequency dependent [13]. We believe that this frequency dependence is best viewed as age dependence, an integral aspect of LD.

The extent of LD in human populations is controversial [14–30]. We discuss demography and population structure in the next section, but it should be noted that it is necessary to take into account allele frequencies when comparing LD (however measured or defined). Therefore, the statement that LD around a variant at 50% frequency, with an MRCA that is almost certainly older than 5000 generations, is unlikely to extend beyond a few kb [14] is not contradicted by the finding that in pairwise comparisons between random markers, LD with respect to the youngest allele (with an estimated MRCA several hundred generations ago) extends much further [20]. Similarly, studies cannot be compared unless the same measure of LD is used [15]. As can be seen in Fig. 5, LD measured as d^2 might extend only a few kilobases whereas LD measured as $|D'|$ might extend almost 100 kb.

For disease mapping purposes, is it useful to know the typical extent of LD between random markers? After all, the extent of LD around a disease allele is largely determined by the history of that allele. This history might bear little resemblance to the history of random markers, especially if the disease allele is present at a very different frequency or has been subject to selection.

Demography and population structure

The role of population subdivision, bottlenecks and expansions in human evolution has been much discussed. The coalescent can be used to draw general conclusions about the probable effects of each. It is important to understand that, loosely speaking, the timescale of the coalescent reflects the speed at which

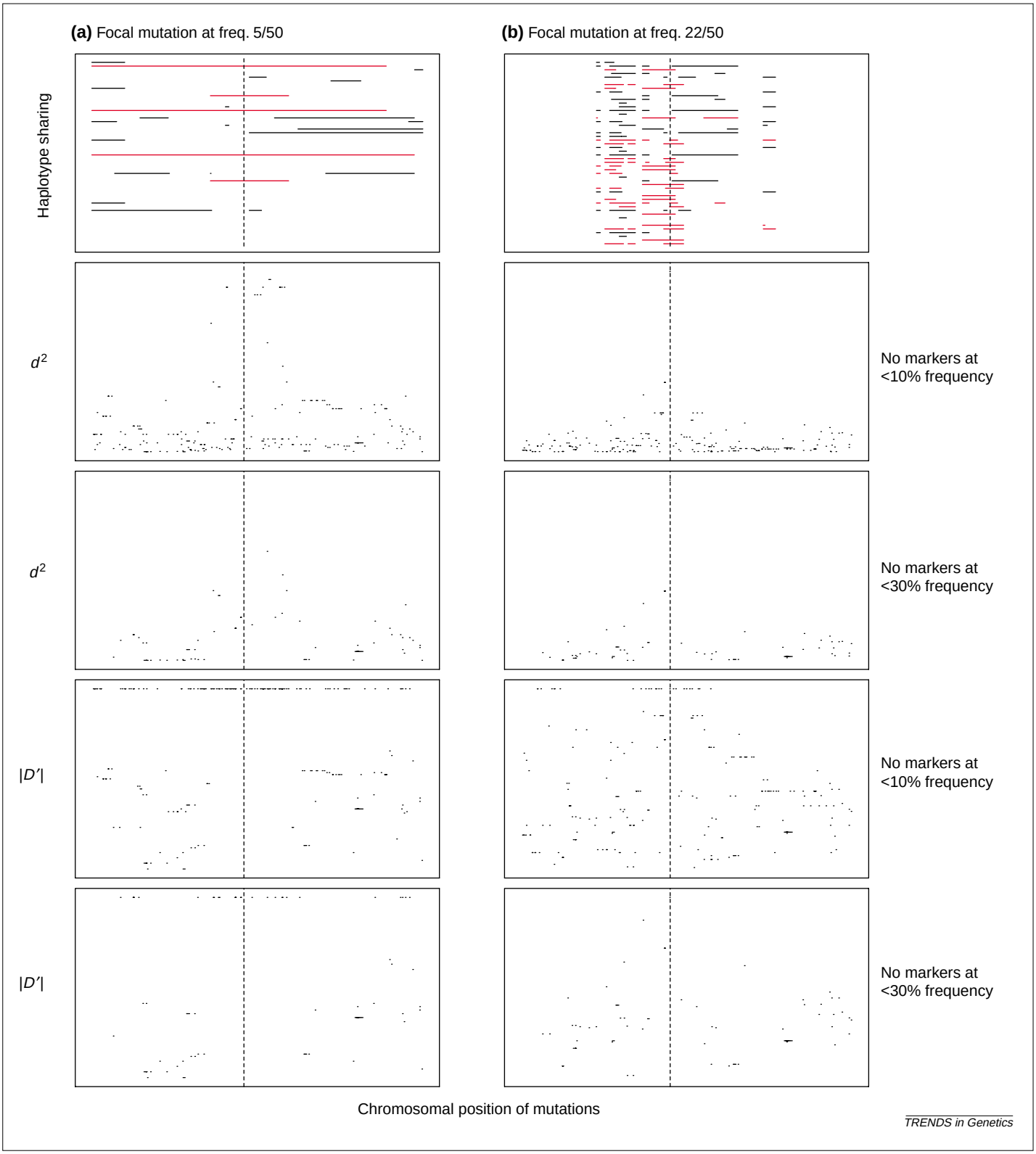

Fig. 5. Examples of haplotype sharing and linkage disequilibrium (LD) in a single simulated sample of size $n = 50$. The standard ancestral recombination graph with infinite-sites mutations and $\mu = \rho = 100$ was used. As explained in Box 2, the horizontal axis, representing chromosomal position, might thus correspond to ~100 kb. The plots illustrate the haplotype sharing and LD with respect to particular focal mutations: (a) the focus is a relatively low-frequency mutation (5/50 = 10%); (b) the focus is a relatively high-frequency mutation (22/50 = 44%). The chromosomal positions of these mutations are indicated by the vertical lines. The top row of plots shows the extent of haplotype sharing with respect to the most recent common ancestor (MRCA) of the focal mutation among the 50 haplotypes. The horizontal lines indicate segments that descend from the MRCA of the focal mutation. Red indicates that the current haplotype also carries the focal mutation; black that it does not. Note that the red segments necessarily overlap the position of the focal mutation. For clarity, segments that do not descend from the MRCA of the focal mutation are excluded, and haplotypes that do not carry segments descended from the MRCA of the focal mutation are therefore invisible. The remaining four rows of plots show the behavior of two commonly used pairwise measures of LD, d^2 and $|D'|$ (Box 1), for different choices of markers. In each plot, the horizontal position of a dot represents the chromosomal position of the marker, and the vertical position the value of the measure (on a zero-to-one scale).

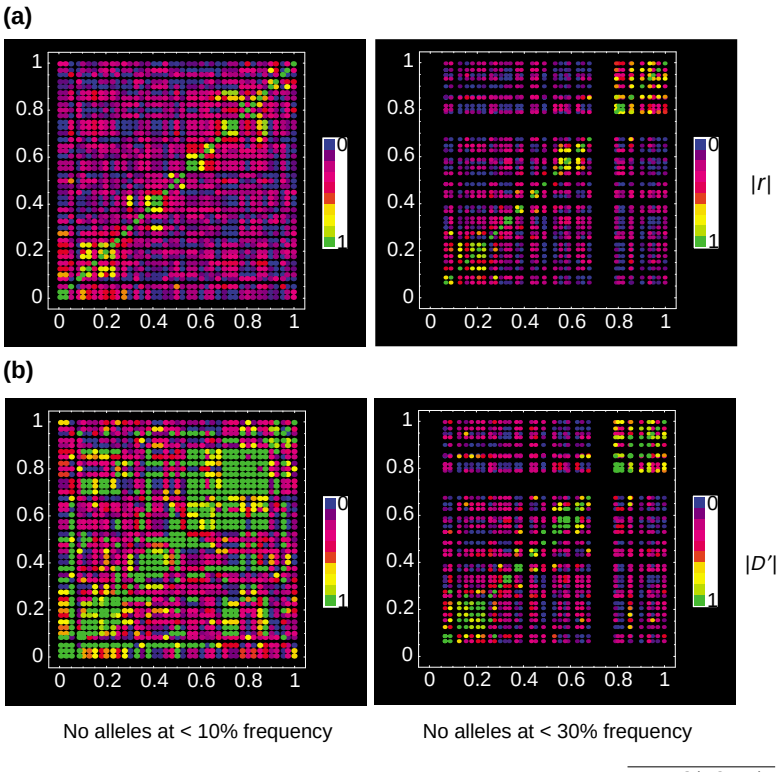

Fig. 6. All pairwise comparisons are shown for the data from Fig. 5. For clarity, only a subset of the sites matching the frequency criteria were included: (a) plots of $|r|$; (b) plots of $|D'|$. The figure is analogous to Fig. 4, except that the data were generated with recombination. The 'cross' of missing points visible in the lower panels demonstrates the variability of this type of data. It appears because the underlying genealogy for that part of the chromosome is such that all minority alleles are present at a frequency of less than 30%.

genetic drift operates (Box 2). Lineages coalesce faster in a small population, and more slowly in a large one. In a population that has experienced recent rapid growth, almost no coalescence events will have occurred during the growth phase. This explains why the recent exponential expansion of the human population might not have had much effect on LD between high-frequency markers [14]. These markers are likely to have ancient MRCAs (perhaps several hundred thousand years old), and the growth phase was too short to have affected the pattern of LD present before growth started.

More generally, any change in the population size does not have much effect on the coalescent unless the change lasts for a number of generations that is of the same order of magnitude as the (new) population size. For example, a bottleneck that instantly reduced the population size from 10 000 to 50 would not be noticeable unless it lasted for around 50 generations; its effects would be similar to a bottleneck of 500 individuals that lasted for around 500 generations. As a consequence, bottlenecks such as those caused by the major European plague epidemics cannot possibly have affected the pattern of variability (except of course at any loci involved in resistance to the disease).

When coupled with the fact that the MRCA of a frequent marker is likely to be very old, these observations explain why LD between high-frequency markers might be little affected by recent demography [14] – most of the recombination events occurred earlier. In a similar vein, the observation [18,19,23–28] that LD between high-frequency markers shows a similar pattern across different populations can be attributed to the common early history of those populations. Of course, this does not mean that all populations must show similar levels of LD; they clearly do not [29–32]. The effects of demography on LD have often been discussed in the context of choosing the right population for LD mapping [14,16–20,22–29,33,34]. However, as noted above, it could be more relevant to focus on the probable effects of demography on the history of the disease alleles, rather than on the pattern of LD between random markers. One undisputed advantage of isolated populations is that allelic heterogeneity for the disease is less likely [35–37]. Such effects of demography might well be more important than the effects on LD. Nonetheless, it is clear that LD can be much more extensive in sufficiently small populations, especially when inbreeding takes place. This might be of relevance not only in humans [31,32], but also in other species [38–40].

Finally, it is well known that population subdivision followed by admixture can increase levels of LD. This can be both a boon and a bane for LD mapping. Essentially, it increases the extent of LD, but it also introduces a significant risk of detecting spurious linkage [41–45].

Extensive LD as a sign of past selection

When the extent of LD surrounding an allele seems unusually large, given the frequency of the allele, it is tempting to conclude that the allele is very young and must have been driven to its present frequency by natural selection. There are at least three common problems with such arguments.

First, it is important to distinguish between the age of the MRCA and the age of the mutation [46,47]. The difference between the two can be substantial [48,49].

Second, the history of an allele that is known to have a certain frequency today is not the same as the history of a random allele [46,48,49]. This makes it hard to evaluate the significance of claims of past selection [46], especially when coupled with scenarios of past population subdivision.

Third, it is worth recalling a classic paradox from population genetics theory: when viewed from the present, the history of a positively selected allele is the same as the history of a negatively selected allele [50]. Thus, extensive LD surrounding a relatively frequent allele that causes disease today might simply reflect past selection against the allele, rather than past selection for the allele. Which is more probable depends on the relative rates at which selectively deleterious and favorable alleles occurred. Given our lack of knowledge of these parameters, this is almost a philosophical question.

Whereas past selection on specific loci can be difficult to demonstrate, it might be easier to infer the

action of selection on a genomic scale, perhaps by looking at the variance of LD [21], or by comparing synonymous and nonsynonymous polymorphisms [23].

Haplotypes and sample size

Population genetic data are correlated because of the single underlying genealogy. This has important statistical consequences. Coalescent theory shows that increasing the sample size n is often surprisingly unhelpful, essentially because one samples more of the same events. For example, when estimating the scaled mutation rate μ, the variance of the estimator decreases at best at a rate of $1/\log n$ rather than $1/n$. Precisely the same problems should apply to LD mapping, albeit that the fine details depend on the assumed demographics. It is clear from Fig. 5 that most of the recombination events are shared between segments. Therefore, increasing the sample size should often yield haplotypes that have already been observed. Of course, large sample sizes might nonetheless be needed to decrease the variability attributable to factors such as incomplete penetrance [51].

Number of recombination events

We noted earlier that the recombination and mutation rates will, on average, be of a similar order of magnitude in many organisms, including humans. It then follows from coalescent theory that the genealogy of the genome contains approximately as many recombination events as it does segregating sites. It is therefore not surprising that some datasets show little tree structure [52]. The ratio of segregating sites to recombination events is likely to be much higher in inbred organisms [53], as well as in regions of low recombination [30]. However, in the latter context it is important to distinguish between actual variation in recombination rates, and apparent variation caused by recombination history.

Prospects

Our main purpose in writing this paper has been to emphasize the value of thinking about LD – or indeed any population genetic data – in terms of the underlying genealogy. Although it is far from clear that explicit modeling of this genealogy will be directly useful for mapping purposes, there can be little doubt that such modeling can provide considerable insight.

There would seem to be two main obstacles to using LD for mapping purposes. The first is the variability of LD. As noted above, LD is the outcome of a complex historical process. Much of the variation in LD is attributable to this process, and cannot be eliminated (unless we re-run evolution). However, LD mapping has been applied successfully on several occasions, often using simple methods, and much more powerful methods based on multiple markers are being developed [8–12]. Some genes might be difficult or impossible to map using LD because of the history of recombination and mutation around them, but this will not be true for all genes.

A potentially much more serious problem is the genetic architecture of the traits being mapped. If a common disease is caused by thousands of rare mutations at hundreds of loci, it will not be amenable to mapping [37]. Population genetics models can be used to predict the likelihood of this scenario [54,55], but given the uncertainties about the relevant parameters, this issue will have to be resolved empirically.

Mapping is not the only motivation for modeling LD. For example, it has been noted that the pattern of LD across the human genome appears not to fit the standard models [15]. More realistic demographic models will probably fit the data better, but it also possible that we shall learn more about fundamental genetic mechanisms such as recombination.

Acknowledgements

We thank David Balding, Warren Ewens, Malia Fullerton, Rosalind Harding, Fengzhu Sun, Kenneth Weiss and the reviewers for helpful comments on the manuscript. This review was started while MN was supported by grants from the Swedish Natural Sciences Research Council (NFR B-AA/BU 12026) and the Erik Philip-Sörensen Foundation. ST was supported in part by NSF grant DBI95-04393 and NIH grant GM 58897.

References

1 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517
2 Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99
3 Clayton, D. (2000) Linkage disequilibrium mapping of disease susceptibility genes in human populations. *Int. Stat. Rev.* 68, 23–43
4 Thompson, E.A. (2000) *Statistical Inferences from Genetic Data on Pedigrees*. IMS
5 Cheung, V.G. *et al.* (1998) Linkage-disequilibrium mapping without genotyping. *Nat. Genet.* 18, 225–230
6 Weir, B.S. and Hill, W.G. (1986) Nonuniform recombination within the human β-globin gene cluster. *Am. J. Hum. Genet.* 38, 776–778
7 Hill, W.C. and Weir, B.S. (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.* 54, 705–714
8 Lazzeroni, L.C. (1998) Linkage disequilibrium and gene mapping: An empirical least-squares approach. *Am. J. Hum. Genet.* 62, 159–170
9 McPeek, M.S. and Strahs, A. (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* 65, 858– 875
10 Morris, A.P. *et al.* (2000) Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am. J. Hum. Genet.* 67, 155–169
11 Rannala, B. and Reeve, J.P. (2001) High-resolution multipoint linkage-disequilibrium mapping in the context of the human genome sequence. *Am. J. Hum. Genet.* 69, 159–178
12 Liu, J.S. *et al.* (2001) Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* 11, 1716–1724
13 Lewontin, R.C. (1988) On measures of gametic disequilibrium. *Genetics* 120, 849–852
14 Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139–144
15 Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* 69, 1–14
16 Laan, M. and Pääbo, S. (1997) Demographic history and linkage disequilibrium in human populations. *Nat. Genet.* 17, 435–438
17 Freimer, N.B. *et al.* (1997) Expanding on population studies. *Nat. Genet.* 17, 371–373
18 Lonjou, C. *et al.* (1999) Allelic association between marker loci. *Proc. Natl. Acad. Sci. U. S. A.* 96, 1621–1626
19 Kruglyak, L. (1999) Genetic isolates: Separate but equal? *Am. J. Hum. Genet.* 96, 1170–1172
20 Collins, A. *et al.* (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl. Acad. Sci. U. S. A.* 96, 15173–15177
21 Huttley, G.A. *et al.* (1999) A scan for linkage disequilibrium across the human genome. *Genetics* 152, 1711–1722
22 Ott, J. (2000) Predicting the range of linkage disequilibrium. *Proc. Natl. Acad. Sci. U. S. A.* 97, 2–3
23 Goddard, K.A. B. *et al.* (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* 66, 216–234
24 Boehnke, M. (2000) A look at linkage disequilibrium. *Nat. Genet.* 25, 246–247
25 Eaves, I.A. *et al.* (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 25, 320–323
26 Taillon-Miller, P. *et al.* (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.* 25, 324–328
27 Dunning, A. *et al.* (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* 67, 1544–1554

28 Abecasis, G.R. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* 68, 191–197
29 Reich, D.E. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* 411, 199–204
30 Daly, M.J. *et al.* (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229–232
31 Kerem, B-S. *et al.* (1989) Identification of the cystic fibrosis gene: Genetic analysis. *Science* 245, 1073–1080
32 Houwen, R.H.J. *et al.* (1994) Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. *Nat. Genet.* 8, 380–386
33 Laan, M. and Pääbo, S. (1998) Mapping genes by drift-generated linkage disequilibrium. *Am. J. Hum. Genet.* 63, 654–656
34 Terwilliger, J.D. *et al.* (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: "drift mapping" in small populations with no demographic expansion. *Hum. Hered.* 48, 138–154
35 Hästbacka, J. *et al.* (1992) Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplàsia in Finland. *Nat. Genet.* 2, 204–211
36 Wright, A.F. *et al.* (1999) Population choice in mapping genes for complex diseases. *Nat. Genet.* 23, 397–404
37 Weiss, K.M. and Terwilliger, J.D. (2000) How many diseases does it take to map a gene with SNPs? *Nat. Genet.* 26, 151–157
38 Riquet, J. *et al.* (1999) Fine-mapping of quantitative-trait loci by identity by descent in outbred populations: Application to milk production in dairy cattle. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9252–9257
39 Lin, L. *et al.* (1999) The sleep disorder canine narcolepsy is caused by a mutation in the *Hypocretin (Orexin) Receptor 2* gene. *Cell* 98, 365–376
40 Schibler, L. *et al.* (2000) Fine mapping suggests that the goat *Polled Intersex Syndrome* and the human *Blepharophimosis Ptosis Epicanthus Syndrome* map to a 100-kb homologous region. *Genome Res.* 10, 311–318
41 Stephens, J.C. *et al.* (1994) Mapping by admixture linkage disequilibrium in human populations: Limits and guidelines. *Am. J. Hum. Genet.* 55, 809–824
42 Ewens, W.J. and Spielman, R.S. (1995) The transmission/disequilibrium test: History, subdivision, and admixture. *Am. J. Hum. Genet.* 57, 455–464
43 Peterson, R.J. *et al.* (1999) Effects of worldwide population subdivision on *aldh2* linkage disequilibrium. *Genome Res.* 9, 844–852
44 Lautenberger, J.A. *et al.* (2000) Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. *Am. J. Hum. Genet.* 66, 969–978
45 Pritchard, J.K. *et al.* (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181
46 Thompson, E.A. and Neel, J.V. (1997) Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am. J. Hum. Genet.* 60, 197–204
47 Rannala, B. and Slatkin, M. (1998) Likelihood analysis of disequilibrium mapping, and related problems. *Am. J. Hum. Genet.* 62, 459–473
48 Griffiths, R.C. and Tavaré, S. (1998) The age of a mutant in a general coalescent tree. *Stoch. Mod.* 14, 273–295
49 Wiuf, C. and Donnelly, P. (1999) Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.* 56, 183–201
50 Maruyama, T. (1974) The age of an allele in a finite population. *Genet. Res. Camb.* 23, 137–143
51 Long, A.D. and Langley, C.H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* 9, 720–731
52 Clark, A.G. *et al.* (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* 63, 595–612
53 Nordborg, M. (2000) Linkage disequilibrium, gene trees, and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics* 154, 923–929
54 Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69, 124–137
55 Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.* 17, 502–510

On the shoulders of giants: p63, p73 and the rise of p53

Annie Yang, Mourad Kaghad, Daniel Caput and Frank McKeon

The discoveries of the p53 homologs, p63 and p73, have both fueled new insights and exposed enigmas in our understanding of the iconic p53 tumor suppressor. Although the pivotal role of p53 in cancer pathways remains unchallenged, because p63 and p73 are now implicated in stem cell identity, neurogenesis, natural immunity and homeostatic control. Despite their seemingly separate tasks, there are hints that the p53 family members both collaborate and interfere with one another. The question remains, therefore, as to whether these genes evolved to function independently or whether their familial ties still bind them in pathways of cell proliferation, death and tumorigenesis.

The realization that p53 [1], the archetypal tumor suppressor in higher mammals, in fact belonged to a family of related genes came in 1997, almost 20 years after the discovery of p53 [2]. The report of the first homolog, p73 [3], and the fact it was located in a long-suspected tumor suppressor locus, was met with great excitement and anticipation. Was this another tumor suppressor? And would decades of work on p53 enable us to understand this close relative readily? The situation was quickly complicated by the appearance of yet another homolog, p63 (also named KET, p51, p40, p73L) [4–8], and the myriad gene products it encoded. Far from fitting into classic p53 roles in tumor suppression, the homologs are claiming their own turf in stem cell biology, neurogenesis and a host of other physiological processes. The past four years of work on p63 and p73 have also added layers of complexity to the p53 family as a whole. Here, we review the individual functions of the p53-related genes and explore evolutionary origins that could offer an intriguing perspective on the p53 family.

The burning question: tumor suppressors or not?

When they were first discovered, it seemed entirely reasonable to imagine that p63 and p73 would follow in the footsteps of p53 and be involved in tumor suppression and cell cycle control. The sequence similarity and conservation of functional domains among the p53 family members are indeed striking [3,4]. p63 and p73 both share the hallmark features that identify p53 across all species – an acidic,

http://tig.trends.com 0168-9525/02/$ – see front matter PII: S0168-9525(01)02595-1