Prospects for whole-genome linkage disequilibrium mapping of common disease genes

Leonid Kruglyak

Recently, attention has focused on the use of whole-genome linkage disequilibrium (LD) studies to map common disease genes. Such studies would employ a dense map of single nucleotide polymorphisms (SNPs) to detect association between a marker and disease. Construction of SNP maps is currently underway. An essential issue yet to be settled is the required marker density of such maps. Here, I use population simulations to estimate the extent of LD surrounding common gene variants in the general human population as well as in isolated populations. Two main conclusions emerge from these investigations. First, a useful level of LD is unlikely to extend beyond an average distance of roughly 3 kb in the general population, which implies that approximately 500,000 SNPs will be required for whole-genome studies. Second, the extent of LD is similar in isolated populations unless the founding bottleneck is very narrow or the frequency of the variant is low (<5%).

Introduction

Whole-genome association studies have recently been proposed as a powerful approach for detecting the many subtle genetic effects that underlie susceptibility to common diseases[1–3]. In contrast to linkage studies, which look for co-inheritance of chromosomal regions with disease in families, association studies look for differences in frequency of genetic variants between unrelated affected individuals and controls. Such studies have long been used to test the involvement of candidate genes in diseases and to refine the location of disease genes in regions identified by linkage[4]. Improved techniques for high-throughput identification and genotyping of polymorphisms offer the possibility of extending this approach to the entire genome in the near future[5,6].

Association studies can be carried out using one of two general strategies: direct or indirect[7]. Both rely on the hypothesis that common genetic variants underlie susceptibility to common diseases. The direct strategy is to catalogue all common variants in coding and regulatory regions of genes, in the hope that this collection will contain the changes that influence disease susceptibility. Frequencies of these variants would then be compared in patients and controls, with the expectation that a risk-conferring variant will be more common in patients. Although straightforward in principle, the direct strategy faces practical hurdles. A systematic whole-genome application requires the identification of all 50,000–100,000 human genes, as well as their common variants. Currently, complete sequence is available for only a fraction of all human genes. Furthermore, it is difficult to identify variants that affect gene function by means other than changing the coding sequence (for example, regulatory and intronic polymorphisms). For these reasons, near-term applications of the direct strategy will probably be limited to studies of coding variation in selected sets of candidate genes.

The indirect strategy avoids the need for cataloguing potential susceptibility variants by relying instead on association between disease and neutral polymorphisms located near a risk-conferring variant. Such associations may arise as a result of linkage disequilibrium (LD) between the risk locus and nearby polymorphisms. The indirect strategy thus employs a dense map of polymorphic markers to scan the genome systematically for regions associated with disease. Biallelic single nucleotide polymorphisms (SNPs) are the markers of choice because of their high frequency, low mutation rates and amenability to automation[5,6].

A key question facing the indirect strategy is: how many markers are needed to adequately cover the genome? The answer depends on the chromosomal extent of LD in human populations. The factors that govern LD are complex, and the ultimate answer must await comprehensive experimental studies of many genomic regions in many populations. Nonetheless, it is useful to know what to expect based on our current knowledge. Here, I provide estimates of the extent of LD between a marker and a variant in the general human population as well as in isolated populations. These estimates are based on coalescent simulations that use simplified scenarios of population history, mutation and recombination. These simulations suggest that useful levels of LD are unlikely to extend beyond an average distance of approximately 3 kb in the general population. This result implies that roughly 500,000 SNPs will be required for whole-genome association studies in samples drawn from large outbred populations. Furthermore, these simulations indicate that the extent of LD is similar in isolated populations unless the founding bottleneck is very narrow (effective size of 10–100 unrelated individuals) or the frequency of the variant is low (<5%).

Assumptions

Populations were simulated based on the coalescent process[8]. This approach takes into account genetic drift, accommodates different models of population history, allows estimation of LD conditional on variant and marker allele frequencies, does not

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. e-mail: leonid@fhcrc.org

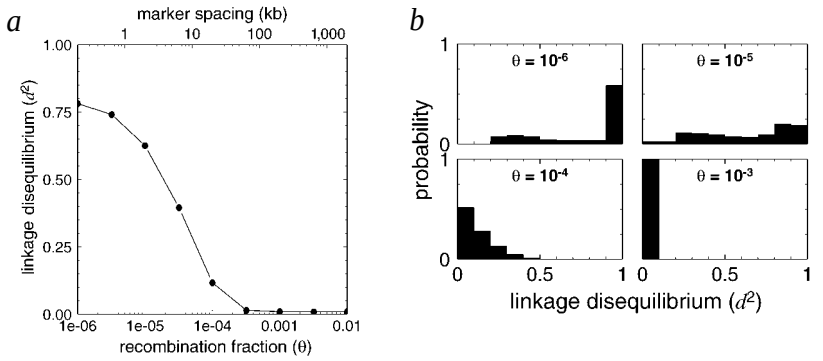

Fig. 1 LD around a variant of 50% frequency in the general human population. ***a***, Average LD between the variant and a marker, plotted as a function of the recombination fraction (bottom *x* axis) and the corresponding average marker spacing (top *x* axis). Grey bars show the standard deviation. The population model is constant size N=10,000 until G=5,000 generations before the present, followed by exponential growth to a present size of 5 billion. ***b***, Probability distributions of LD for selected values of the recombination fraction.

require assumptions about variant age and is computationally rapid. The general human population was assumed to have constant effective size N until G generations before the present, and to then expand exponentially to its present-day size of 5 billion. The standard assumption was N=10,000 and G=5,000 (corresponding to 100,000 years given a 20-year generation time). Although this scenario reflects the best current evidence about human population history[9], the effects of varying N and G were also explored. Isolated populations were modelled as a founding event in which the general population passes through a bottleneck of size F, followed by a period of expansion. The size of the founding population (F) represents the effective number of unrelated founders. The actual census size of the founding population would be larger by a factor that depends on the relatedness of the founders and the details of the founding event. Different founder numbers, founding times and expansion scenarios were examined.

Disease and marker loci were assumed to be biallelic, with variation at each locus arising as a result of a single unique mutational event. The frequency of the risk variant was fixed at a desired value, and the frequency of marker alleles was constrained to lie in the range 0.25–0.75 (other ranges produce similar results).

Several measures of LD are in use for different purposes[10]. An appropriate choice for association studies is d^2, the squared difference between the frequency of the associated marker allele on variant and normal chromosomes[11]. (d^2 is simply related to other common measures such as D^2 and r^2 by $d^2=D^2/[f(1–f)]^2$ and $d^2=r^2m(1–m)/f(1–f)$, where f is the variant frequency and m is the marker allele frequency.) This measure can be interpreted as follows: if the sample size needed to detect association between a variant and disease is S (which depends on the frequency of the variant and the increase in susceptibility that it confers), and if the magnitude of LD between the variant and a nearby marker is d^2, then the sample size needed to detect the association with the marker is approximately S/d^2. If a variant increases the risk of disease twofold or less, the sample size needed to detect association is very large[1,12]. The use of a nearby marker with d^2 of 0.1 would entail a further 10-fold increase in sample size—an increase that is likely to be unacceptable in practice for many studies. For the purposes of discussion, I will consider d^2 values significantly above 0.1 to be useful for association studies, although readers may employ their own criteria.

I assume that all loci are selectively neutral, that there is no gene conversion or recurrent mutation, and that mutation and recombination are independent. In converting results from genetic to physical distance, I assume that recombination occurs at a rate of 1% per Mb. This rate is the genome-wide average, and is thus appropriate for computing the average extent of LD and the average marker spacing needed to detect association. Recombination is widely believed to be highly nonuniform, however, especially over distances below 100 kb, which implies that the physical extent of LD is likely to be much greater than average in some regions of the genome and much less in others.

Extent of LD and SNP density in the human population

Consider a common variant (50% frequency) and a nearby marker separated by recombination fraction θ. The expected mean and standard deviation (s.d.) of LD as a function of θ in the general human population are shown (Fig. 1*a*). Essentially no LD is observed at θ=0.0003 or greater. This genetic distance on average corresponds to a physical distance of 30 kb. A low level of LD (d^2=0.1, s.d.=0.1) is observed at θ=0.0001 (10 kb). The level of LD increases to d^2=0.4 (s.d.=0.2) at θ=0.00003 (3 kb), and approaches a maximal level of d^2=0.8 (s.d.=0.3) at approximately θ=0.000003 (300 bp). Recall that the sample size needed to detect an association is roughly proportional to $1/d^2$. These results then suggest that a marker within θ=0.00003 of every variant is required to avoid a large increase in sample size. This marker density can be achieved with a map of 500,000 SNPs with an average spacing of 6 kb across the genome. Use of sparser maps would entail considerable loss of power. Ideally, marker spacing in different regions of the genome would reflect the recombination rate—and therefore the typical extent of LD—of each region.

The distributions of d^2 for several values of θ are shown (Fig. 1*b*). It is apparent from the distributions that high mean levels of LD are accompanied by high levels of variability. Although the mean level of LD at θ=0.00001 (1 kb) is 0.6, the distribution is very broad, with all values of d^2 between 0.2 and 1 having comparable probabilities. d^2 values below 0.2 were observed in 5% of cases. These results reflect evolutionary vari-

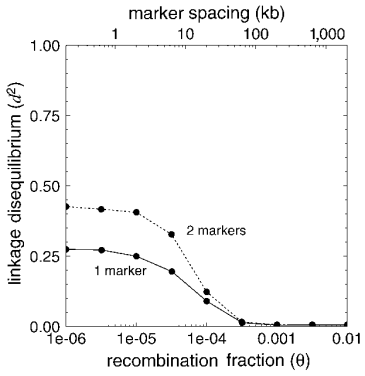

Fig. 2 LD around a variant of 10% frequency. Filled line, LD with a single marker; dashed line, LD with a two-marker haplotype, with the two markers flanking the variant, each at recombination fraction θ.

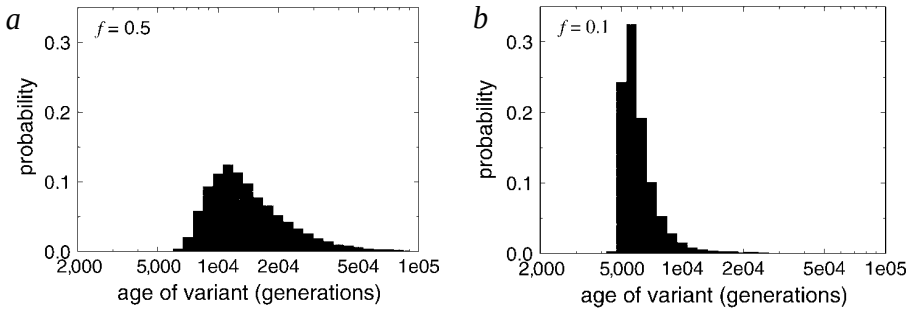

Fig. 3 Distributions of variant age. ***a***, Variant with 50% frequency in today's population. ***b***, Variant with 10% frequency in today's population.

ability of populations and illustrate that pairs of loci separated by the same (small) distance can have very different levels of LD, as is often observed[13,14]. On the other hand, low mean levels of LD are accompanied by low levels of variability—for θ=0.001, d^2=0.01 (s.d.=0.01), with virtually all values of d^2 falling below 0.1.

For variants of lower frequency, the decay of LD with distance is similar, but the maximal level of LD at zero recombination is lower due to the difference in frequency between the variant and the associated SNP allele. An example is shown for a 10% variant (Fig. 2). In principle, it would be optimal to choose the frequency of the associated SNP allele to be the same as the frequency of the variant[15], but this cannot be achieved in practice because the frequencies of risk variants are not known *a priori*. A practical choice is to use a map of SNPs with a range of allele frequencies (here 0.25–0.75, but other choices behave similarly). With such a map, high-frequency variants can be detected with single markers, whereas lower-frequency variants require the use of haplotypes to raise the level of LD. Greater LD with a 10% variant provided by a two-marker haplotype is shown (Fig. 2). LD can be increased further by use of haplotypes of several markers, at the cost of increasing the overall marker density of the map (data not shown). The higher informativeness of a haplotype is a consequence of its lower frequency on non-variant chromosomes.

The rapid decay of LD with distance is a consequence of the relatively ancient origin of most common variants. Simulations show that essentially all neutral variants observed at 50% frequency in today's population would have arisen before the expansion of modern humans 5,000 generations (100,000 years) ago, with most dating back to over 10,000 generations ago (Fig. 3*a*). Variants observed at 10% frequency in today's population tend to be of a more recent origin, but also date almost exclusively to the time of expansion or earlier (Fig. 3*b*). These ancient origins reflect the time required for a neutral variant that initially arose as a mutation on a single chromosome to drift up to high frequency in today's population.

The decay of LD with distance depends on human population history. Neither the date of the expansion nor the ancestral effective population size is known precisely. The former is thought to have occurred between 50,000 and 200,000 years ago[9], and estimates of the latter range from a few thousand to 100,000 (refs 9,16–18). The effect of expansion date on the extent of LD is modest (Fig. 4), whereas the influence of effective ancestral population size is more pronounced (Fig. 5). If the size is as low as 1,000, substantial LD (d^2=0.25) is observed at θ=0.0001 (10 kb). In contrast, if it is as high as 100,000, little LD is observed even at θ=0.00003 (3 kb). None of these scenarios exhibit LD at distances of θ=0.0003 (30 kb) or greater. LD levels are not sensitive to changes of at least two orders of magnitude in current population size (data not shown).

Extent of LD and SNP density in isolated populations

It is frequently argued that LD studies can be carried out with fewer markers in recently founded genetic isolates, because in such populations recombination has had less time to whittle down regions of LD around risk-conferring variants. A distinction needs to be drawn, however, between the age of the population and the age of the variant. If a variant is rare, either arising within the isolate or introduced by a single founder, then its recombinational history is confined to the isolate. In this case, the age of the population is relevant, and the extent of LD is greater in younger isolates. If a variant is common, it is likely to have been introduced by multiple founders. In this case, the variant's recombinational history extends to its origin in the general population, and the extent of LD is not increased. Whether a variant of a given frequency is considered rare or

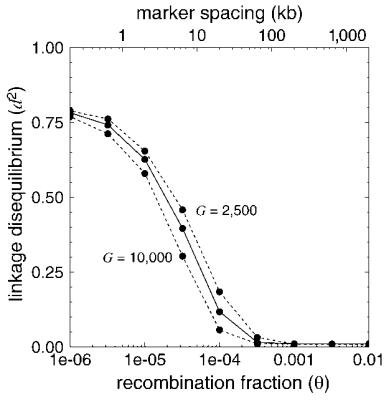

Fig. 4 LD around a variant of 50% frequency for different expansion times. Filled line, G=5,000; dashed lines, G=2,500 or G=10,000, as marked.

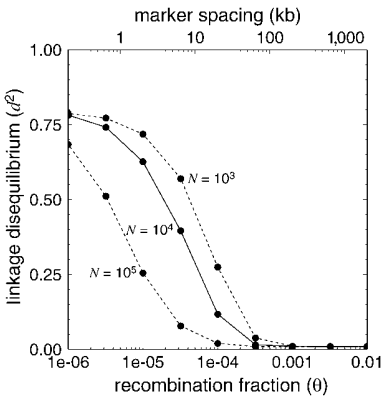

Fig. 5 LD around a variant of 50% frequency for different ancestral population sizes. Filled line, N=10^4; dashed lines, N=10^3 or N=10^5, as marked.

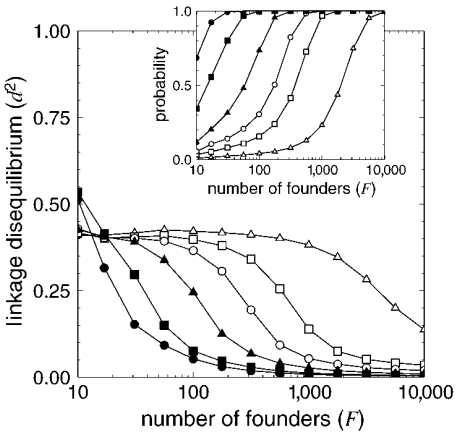

Fig. 6 LD in an isolated population. LD around variants of different frequencies is plotted as a function of founding size *F*. In this population model, the general population passes through a bottleneck of size *F* 100 generations before the present, followed by exponential growth to a present size of 5 million. LD is measured between a variant and 2-marker haplotype, with the markers flanking the variant, each at a recombination fraction of 0.001. Variant frequencies are 50% (●), 25% (■), 10% (▲), 5% (○), 3% (□) and 1% (Δ). Inset, the probability that in a sample of 1,000 chromosomes from today's population, the chromosomes that carry the variant had multiple ancestors at the time of the bottleneck. Symbols designate the same variant frequencies as in the main figure.

common depends on the history of the isolate—in particular, its founding size and rate of growth.

Consider an isolated population that was founded by an effective number *F* of unrelated individuals and then expanded for 100 generations to a current size of 5 million. (This is a model of the population history of Finland[19].) It is useful to focus on LD at θ=0.001 (100 kb)—a distance at which LD is absent in the general population, but is expected to show little decay during 100 generations. The expected LD as a function of *F* for a range of variant frequencies is shown (Fig. 6). In a population with *F*=10, LD is high for variants of all frequencies between 1% and 50%. When *F*=100, little LD is observed for variants with frequencies of 25–50%, but substantial LD persists for variants with frequencies at or below 10%, which were present on relatively few founder chromosomes. When *F*≥1,000, strong LD is observed only for variants with frequency of 1%. These results show that the extent of LD around common variants is increased only in isolates with a very small number of founders. They also illustrate that little LD around common variants may be observed in the same populations that exhibit substantial LD around rare disease mutations. The results are consistent with the idea that LD is low in an isolate when the variant is introduced by multiple founders (Fig. 6, inset).

In addition to the number of founders, isolated populations vary in their rates of growth. A slower growth rate increases the loss of genetic diversity during the early history of the isolate and thereby effectively results in a narrower bottleneck. The effect of growth rate on LD was examined by two approaches. The first was to fix the number of founders at 100 and the founding time at 20 generations before the present, and to vary the growth rate per generation. As expected, lower levels of LD are observed for faster rates of growth (Fig. 7*a*). The second approach was to fix the number of founders at 100, the founding time at 20 generations before the present and present population size at 10,000, then to introduce a period of constant population size either early or late in the history of the isolate. Populations with a longer constant phase early, followed by late rapid expansion, showed increased LD, whereas those with early rapid expansion followed by a longer constant phase late showed decreased LD (Fig. 7*b*). Thus slow early growth, as well as a small number of founders, are key factors for increasing the extent of LD in an isolated population.

Discussion

The results presented here predict that the average extent of useful levels of LD in the general outbred human population is approximately 3 kb. This prediction is consistent with classic population genetics theory for a population with an effective size of 10,000 (ref. 20), and implies that roughly 500,000 appropriately spaced SNPs will be needed for systematic whole-genome LD studies. This number of SNPs is comparable to the expected total number of common gene variants in the human genome[1,2]. The number of SNPs may need to be even larger to accommodate random variations in level of LD, difficulties in matching SNP map density to local recombination rates and the need to construct multi-marker haplotypes. The results also

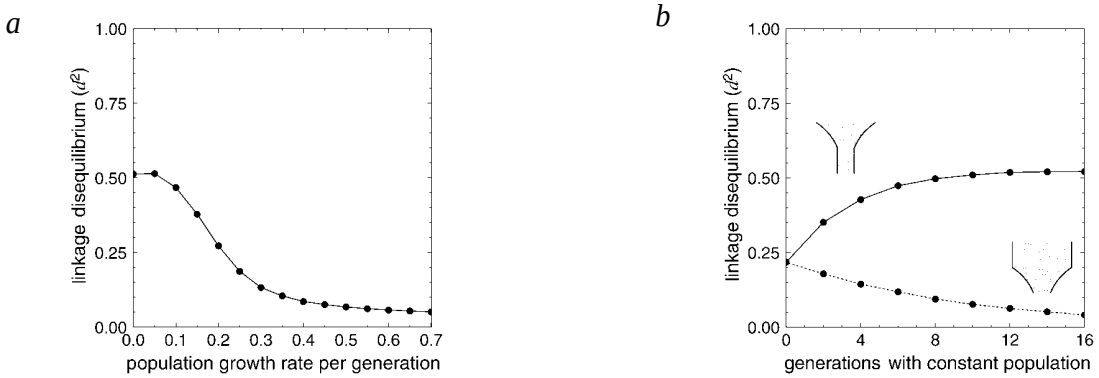

Fig. 7 LD under different scenarios of population growth. LD is measured between a variant of 10% frequency and a 2-marker haplotype as in Fig. 6. ***a***, Exponential growth with different rates for population with founding size *F*=100 and founding time 20 generations before the present. LD is plotted as a function of population growth rate per generation. ***b***, Period of constant size either before (filled line) or after (dashed line) exponential growth for population with founding size *F*=100, current size 10,000 and founding time 20 generations before the present. Schematic diagram above each curve illustrates population growth scenario (present time is at the top). LD is plotted as a function of the number of generations with constant population size.

show that the extent of LD around common variants is similar in isolated populations, unless they have experienced a very narrow bottleneck.

For the short term, these results imply that whole-genome LD studies are likely to be restricted to special populations. The best candidate populations for detecting association with common variants are isolates with a small effective number of unrelated founders (10–100) that experienced slow growth during the early generations following the initial bottleneck. Some populations currently regarded as isolates may not meet these criteria. Because accurate genealogical records are not available for most populations, it would be very useful to have a bioassay for estimating a population's effective number of founders. Another potential advantage of isolates with a small number of founders is that the number of different variants underlying susceptibility to a disease may in some cases be reduced compared with a large outbred population, with the effect that the risk conferred by a given variant would be higher (thereby making association easier to detect).

As noted above, the situation for common variants is in contrast to that of rare disease mutations, which may be detected most easily in young, rapidly growing populations[21,22]. Recently, Terwilliger *et al*.[23] proposed the use of isolates of small constant size for LD studies of common variants. Although an increased level of LD in such populations is consistent with the results presented here, the requirement of small population size—on the order of 1,000 individuals—may make it difficult to collect adequate samples of affected individuals for association studies.

For the longer term, the results presented here establish the required density for the next generation of genetic maps that will be useful for whole-genome LD studies in large outbred populations. There are no fundamental reasons why a map of 500,000–1,000,000 SNPs cannot be constructed. Approximately 3,000,000 common SNPs, or one every 1 kb, are expected to exist in the human genome[6]. It is notable that this total density of SNPs is not much higher than the density required for detecting blocks of ancestral LD in the general human population. Although only a few thousand SNPs are available today[6], efforts to map at least 100,000 SNPs in the next 5 years have been proposed[7,24]. For these maps to be useful for association studies, efficient technologies are needed for genotyping hundreds of thousands of SNPs in thousands of individuals.

The results presented here have several limitations. First, human population history is more complex than assumed, and LD may extend further in some populations because of long-term isolation or recent admixture. In particular, the settlement of some geographic regions after the 'out of Africa' migration may have involved small founding populations, followed by relatively recent expansion. A population whose effective long-term size was as low as 1,000 individuals as recently as 1,000 generations (20,000 years) ago would maintain useful levels of LD up to 30 kb from a common variant (data not shown). Second, higher levels of LD can be created by selection, either for a variant (hitchhiking) or against deleterious mutations in the region (background selection)[25]. In both cases, the effective population size is reduced for the region experiencing selection. Third, LD is expected to be enhanced in regions of lower recombination, or if recombination events are clustered in hotspots that separate regions of little recombination. Finally, LD would be disrupted by gene conversion or recurrent mutation at the variant or the marker, in regions of higher recombination, or if recombination events are correlated with the mutations that give rise to SNPs. The role of all these factors in shaping variation across the human genome is an important question for further study.

Despite these limitations, recent experimental evidence supports the prediction of limited extent of LD in the general population. In the gene *LPL*, many marker pairs within a 9.7-kb region show little or no LD in African-Americans, Europeans and Finns[14]. The extent of LD is limited to approximately 10 kb for a set of SNPs within a 25-kb region on chromosome 21 in CEPH families[26]. As additional data are reported, it is important to ensure that comparable measures of LD are used, or (preferably) that the underlying haplotype counts are given so that any measure can be computed. In particular, it is essential that levels of LD be reported in addition to *P* values, because even low levels of LD may reach statistical significance in large samples.

In conclusion, a systematic application of whole-genome LD mapping in general populations is likely to require very large numbers of SNPs. Studies employing maps of lower density may produce some successes in special populations or in regions of the genome where linkage disequilibrium extends beyond distances expected from neutral evolution and average rates of recombination. A rigorous empirical evaluation of LD mapping—including how frequently such favourable cases occur—requires new data on the extent of linkage disequilibrium across the genome in different populations and around variants of different frequencies. Such data should soon become available as dense SNP maps are developed for selected regions of the genome. Ultimately, dozens of regions, each composed of several hundred kilobases, should be saturated with SNPs spaced at intervals of 1 kb, and the extent of LD in these regions should be examined in many different populations. The regions should include those surrounding known disease loci such as *APOE* (ref. 27), as well as those selected at random. The results emerging from such studies will provide insight into the evolutionary forces that have shaped human genetic variation, as well as into the optimal strategies for identifying the genetic basis of common diseases.

Methods

Population simulations were carried out using described methods for generating multilocus genealogies with recombination[8,28–30]. Briefly, the first step is to generate the history of a sample of present-day gametes back to their common ancestor. This is accomplished by running the simulation backward in time and keeping track of common ancestor and recombination events. This history allows construction of trees that describe the genealogy of any given locus. For a given locus, a tree specifies all the common ancestor events as well as their times. Trees for any two loci are identical in the absence of recombination between the loci, but become increasingly less correlated as the recombination fraction between the loci increases. To generate two-locus trees in populations of variable size, the times t_{CA} for the next common ancestor event and t_{RE} for the next recombination event were simulated at each step. If $t_{CA}<t_{RE}$, a common ancestor event was generated at time t_{CA}; otherwise, a recombination event was generated at time t_{RE}. The time to next recombination event was distributed exponentially with parameter $4N\,a(t)$, where N is the present-day population size and $a(t)$ is the total genetic length of ancestral DNA present at time t. Time to the next common ancestor event in a population of variable size was simulated as described[31].

Next, mutations were placed on the genealogy as follows. At each locus, the number of descendants in today's population was counted for every node. At the disease locus, the mutation giving rise to the variant was placed on a branch leading to a node with fn descendants, where f is the variant frequency and n is the sample size. Mutations were successively placed on each such branch, and the contribution of each placement to the overall statistics was weighted by the length of the branch. This weighting takes into account the likelihoods of different trees conditional on the variant frequency. Including any node with the number of descendants within a small range around fn did not change the results. For each variant placement, mutations at each marker locus were assigned to a node with the

number of descendants in the range of 0.25n to 0.75n. For each marker locus, a single branch leading to a node with an acceptable number of descendants was selected randomly from the possible choices, with the probability of selecting a branch proportional to its length.

For each tree and mutation assignment, d^2 was computed as follows. I denoted the variant allele 'v' and the normal allele '+'. For a single marker, I denoted the two alleles '1' and '2', and tabulated the gamete counts of n_{v1}, n_{v2}, n_{+1} and n_{+2}. Then $d^2=[(n_{v1}n_{+2}–n_{v2}n_{+1})/((n_{v1}+n_{v2})(n_{+1}+n_{+2}))]^2$ (ref. 10). This formula is the same as the intuitive definition, $d^2=(n_{v1}/n_v–n_{+1}/n_+)^2$, where n_v and n_+ are the total number of variant and normal gametes, respectively. For k markers, 2^k haplotypes are possible. For each haplotype H, I tabulated its counts n_{vH} and n_{+H} on variant and normal gametes. Then $d^2=(n_{vH}/n_v–n_{+H}/n_+)^2$ for that haplotype. I chose the maximum of d^2 over all haplotypes as the overall value of d^2. No correction was made for multiple testing that this procedure introduces.

I generated 100,000 trees for each simulation condition, with the mean value of d^2 reported. A sample of n gametes from today's population was considered, with n between 100 and 500 depending on variant frequency (larger sample sizes were used with smaller variant frequencies). Simulations with larger values of n were used to ensure that the choice was adequate to keep sampling effects negligible.

Acknowledgements

This work grew out of conversations with E. Lander, and I am grateful for his advice and encouragement. I thank D. Altshuler, K. Ardlie, M. Cargill, H. Coller, G. Daley, M. Daly, M. Eberle, J. Felsenstein, S. Kruglyak, S.-N. Liu, K. Markianos, D. Slonim, D. Wang and E. Wijsman for helpful discussions and comments on the manuscript, and M. Eberle for invaluable assistance with simulations and figure preparation. The paper benefited from discussion during a meeting on SNPs held at the Banbury Center and from comments of anonymous referees. This work was supported in part by grants from NHGRI and NIMH. L.K. is a James S. McDonnell Centennial Fellow.

Received 23 December 1998; accepted 15 April 1999.

1. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
3. Collins, F.S. Positional cloning moves from perditional to traditional. *Nature Genet.* **9**, 347–350 (1995).
4. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
5. Landegren, U., Nilsson, M. & Kwok, P.-Y. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Res.* **8**, 769–776 (1998).
6. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
7. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
8. Hudson, R.R. Gene genealogies and the coalescent process. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford University Press, Oxford, 1991).
9. Harpending, H.C. *et al.* Genetic traces of ancient demography. *Proc. Natl Acad. Sci. USA* **95**, 1961–1967 (1998).
10. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
11. Kaplan, N. & Weir, B.S. Expected behavior of conditional linkage disequilibrium. *Am. J. Hum. Genet.* **51**, 333–343 (1992).
12. Camp, N.J. Genomewide transmission/disequilibrium testing—consideration of the genotype relative risks at disease loci. *Am. J. Hum. Genet.* **61**, 1424–1430 (1998).
13. Jorde, L.B. Linkage disequilibrium as a gene-mapping tool. *Am. J. Hum. Genet.* **56**, 11–14 (1995).
14. Clark, A.G. *et al.* Haplotype structure and population-genetics inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
15. Muller-Myhsok, B. & Abel, L. Genetic analysis of complex diseases. *Science* **275**, 1328–1329 (1998).
16. Reich, D.E. & Goldstein, D.B. Genetic evidence for a Paleolithic human population explosion in Africa. *Proc. Natl Acad. Sci. USA* **95**, 8119–8123 (1998).
17. Takahata, N. Allelic genealogy and human evolution. *Mol. Biol. Evol.* **10**, 2–22 (1993).
18. Ayala, F.J. The myth of Eve: molecular biology and human origins. *Science* **270**, 1930–1936 (1995).
19. Hastbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2**, 204–211 (1992).
20. Robertson, A. & Hill, W.G. Population and quantitative genetics of many linked loci in finite populations. *Proc. R. Soc. Lond. B Biol. Sci.* **219**, 253–264 (1983).
21. Chapman, N.H. & Wijsman, E.M. Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am. J. Hum. Genet.* **63**, 1872–1885 (1998).
22. Thompson, E.A. & Neel, V.J. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am. J. Hum. Genet.* **60**, 197–204 (1997).
23. Terwilliger, J.D., Zollner, S., Laan, M. & Paabo, S. 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 (1998).
24. Collins, F.S. *et al.* New goals for the U.S. human genome project: 1998–2003. *Science* **282**, 682–689 (1998).
25. Li, W.-H. *Molecular Evolution* (Sinauer, Sunderland, 1997).
26. Carlson, C.S. & Cox, D.R. Linkage disequilibrium of SNPs on human chromosome 21. *Am. J. Hum. Genet.* **63**, A284 (1998).
27. Lai, E., Riley, J., Purvis, I. & Roses, A. A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54**, 31–38 (1998).
28. Hudson, R.R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
29. Hudson, R.R. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631 (1985).
30. Hudson, R.R. The how and why of generating gene genealogies. in *Mechanisms of Molecular Evolution* (eds Takahata, N. & Clark, A.G.) 23–36 (Sinauer, Sunderland, 1992).
31. Donnelly, P. & Tavare, S. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**, 40–421 (1995).