ESTIMATING RECOMBINATION RATES FROM POPULATION-GENETIC DATA

Michael P. H. Stumpf and Gilean A. T. McVean‡*

Obtaining an accurate measure of how recombination rates vary across the genome has implications for understanding the molecular basis of recombination, its evolutionary significance and the distribution of linkage disequilibrium in natural populations. Although measuring the recombination rate is experimentally challenging, good estimates can be obtained by applying population-genetic methods to DNA sequences taken from natural populations. Statistical methods are now providing insights into the nature and scale of variation in the recombination rate, particularly in humans. Such knowledge will become increasingly important owing to the growing use of population-genetic methods in biomedical research.

LINKAGE DISEQUILIBRIUM (LD). A measure of genetic associations between alleles at different loci, which indicates whether allelic or marker associations on the same chromosome are more common than expected.

**Department of Biological Sciences, Imperial College of Science, Technology and Medicine, London SW7 2AY, UK.*
‡Department of Statistics, University of Oxford, Oxford OX1 3TG, UK.
email: m.stumpf@imperial.ac.uk; mcvean@stats.ox.ac.uk
doi:10.1038/nrg1227

Despite the importance of recombination in genetics many questions remain regarding the details of the recombination process. What determines where recombination occurs along a chromosome? How much recombination occurs in recombination hotspots? Is the recombination process influenced by local polymorphisms? How do rates change over evolutionary time? Answering such questions will help us to understand the molecular basis of recombination, as well as provide important clues to its evolutionary significance. In addition, as our knowledge of how the recombination rate varies within genomes increases, so will our ability to understand and make use of patterns of association between alleles (or LINKAGE DISEQUILIBRIUM[1,2], LD) for mapping the genetic basis of phenotypic variation[3]. Crucially, the ability to identify the genetic components of phenotypic variation depends on our knowledge of how different parts of the genome are correlated, which is in turn determined, to a large extent, by the recombination process.

Unfortunately the direct measurment of recombination rates at high resolution is a difficult and costly process[4–6]. Pedigree studies, because they include only few informative meioses, produce genetic maps that simply do not have the resolution to assess how recombination rates vary at the level of single genes[2,7,8]. Conversely, analyses of sperm samples have provided remarkable insights into how the recombination rate varies at a few locations within the human genome; however, these studies say nothing about recombination rates in females, and are extremely difficult to carry out[4,9]. Without improvements in genotyping efficiency, large-scale crossing experiments in model organisms are also prohibitively expensive. So, it follows that indirect statistical methods for learning about recombination, such as population-genetic methods, can be exceptionally useful; these methods infer recombination rates from patterns of genetic variation among DNA sequences that are sampled randomly from a population[10–12]. With large-scale surveys of genetic variation now becoming an important focus of modern population genetics, researchers need to be aware of the statistical methods that are available for analysing such data and how to interpret the results.

In this review, we discuss how information about recombination can be obtained from population DNA samples, which statistical models can be used to obtain estimates of the recombination rate and how to interpret the application of such methods to empirical data. Finally, we consider some of the challenges that arise from analysing variation in the data using a population-genetic model (so-called model-based inference) and

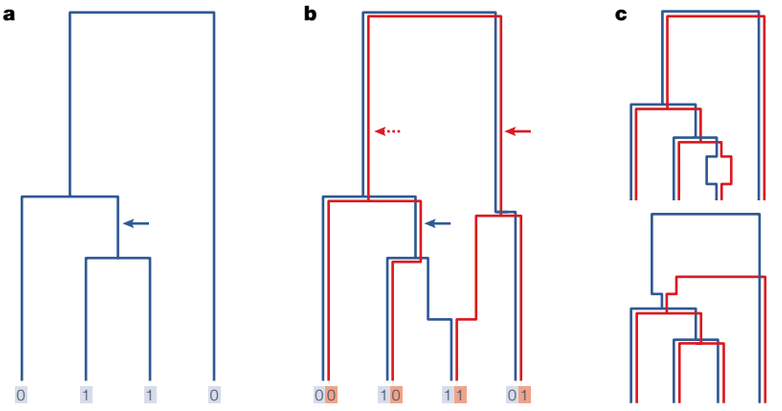


Figure 1 | **Ancestral genealogies and the effects of recombination.** Statistical inferences of evolutionary processes often centre on a description of the genealogy underlying a population or sample. The coalescent is a stochastic process that generates such genealogies. In the legend, 0 and 1 denote ancestral and derived alleles, respectively. **a** | The genealogy of a single hypothetical locus is represented by a single bifurcating tree. A mutation event of 0→1 (arrow) gives rise to a derived allele. **b** | The genealogy of a second locus (red) that is physically close to the locus depicted in part **a** is shown; its genealogy is partially correlated with the original (blue) genealogy (these are known as MARGINAL GENEALOGIES). If mutations occur along the two lineages (indicated by the solid arrows) then the recombination event will be detected in the resulting two-locus gametes, because, as shown here, all four possible gametes (0,0; 0,1; 1,0; 1,1) are observed in the sample. It should be noted that there are two lineages along the red genealogy for which a mutation event can cause the recombination event to be detected (red solid and dashed arrows). **c** | In these two genealogies the recombination event cannot be detected from the resulting data, no matter which lineages mutations occur on. This is because there is no combination of lineages among the two marginal genealogies along which mutations will give rise to all four possible two-locus gametes. For this reason smaller samples are less informative about recombination than larger samples.

how estimates of the recombination rate can aid the application of LD-based strategies for mapping disease-associated loci.

Thinking about recombination

From trees to graphs. The distribution of genetic variation (or polymorphisms) along chromosomes contains a large amount of information about the underlying recombination rate[13]. New mutations arise on a single genetic background in complete association with all of the polymorphisms that are carried by that chromosome. Over time these associations are broken down by the process of recombination, so that, in theory, the degree of association (or LD) between alleles in a sample of chromosomes is simply a function of the age of the mutation and the recombination rate[2,14,15]. However, many other evolutionary forces, such as population history (geographical structure and changes in population size)[14,16], mutation[17,18], natural selection[19,20] and chance events in small populations (genetic drift), also affect patterns of LD[20–24], as can the design of the experiments that are used to determine these patterns, such as MARKER ASCERTAINMENT[25–28]. Consequently, naive, deterministic models of the relationship between recombination and LD fail to capture the enormous stochasticity that underlies the evolutionary process and could generate misleading inferences about patterns of genetic variation[29].

A highly successful way of modelling the impact of evolutionary randomness on genetic variation is to think about the underlying genealogical history of a sample of chromosomes[23,30,31]. Consider a region of the genome that does not recombine, such as the Y chromosome[32], or a single nucleotide position. Looking back in time, we can trace the ancestry of the DNA through its parents, grandparents, and so on. For two DNA samples, the two lineages will meet or 'coalesce' at some point in the past (for example, the Y chromosomes of two brothers coalesce in their father). For a larger sample, we can describe this history of coalescence as an inverted tree[33–35] (FIG. 1a). The differences seen between the sampled DNA sequences are therefore due to mutation events that must have occurred on the tree[36,37]. We are unlikely ever to know the tree in its entirety (including the times at which lineages coalesced and mutations occurred[37]), but we can learn much about the tree from the data[13,38]; for example, we can assess which sequences are most closely related.

Now consider the effect of recombination. At each individual nucleotide we still have a tree, but different parts of the genome will have different trees[34,39–42] (FIG. 1b). Sites that are very close together, which therefore rarely recombine, will probably share the same tree; however, as the recombination distance between sites grows the correlation between the trees decreases[42]. We can therefore describe the ancestry of the sample of recombined chromosomes by using a complex graph[39] that includes a series of coalescence and recombination events (FIG. 1b,c), but which allows us to recover the marginal genealogy at any given position. Again, we can never know the graph in its entirety, but the data provides valuable information about the graph[10].

How does describing data using a graph help us to learn about recombination? First, by thinking about where coalescent, recombination and mutation events have occurred on the tree[42], we can determine what their influence is on patterns of genetic variation[31]. Second, if we can model the process that generates the graph, we can potentially use the data to estimate the parameters of the process (including the recombination rate)[12,39,43].

Counting recombination events

What can we learn about recombination without trying to model the process that generates the underlying recombination graph? The most common statistical approach is to count the number of recombination events that have occurred in the history of a sample: although the family tree of our sample of chromosomes is not known, historical recombination events can leave signature patterns in population-genetic data that can be very informative. However, as we argue below, this method, which does not rely on generating a model of the recombination process, is the least successful. So, although learning about recombination doesn't necessarily mean modelling the process that generates the underlying recombination graph, methods that do (discussed in later sections) are generally the most reliable.

The simplest way of spotting historical recombination events is to look at pairs of single nucleotide polymorphisms (SNPs). For two bi-allelic loci with ancestral and derived alleles *A/B* and *a/b*, respectively, the possible

MARGINAL GENEALOGY
The part of a genealogical graph that corresponds to a single locus or stretch of DNA that is inherited without recombination.

MARKER ASCERTAINMENT
The process by which new genetic markers are obtained — for example, by re-sequencing a subset of chromosomes in a population sample. If those markers are population-specific then inferences that are based on them in other populations might be biased through so-called ascertainment bias.

HAPLOTYPES (or gametes) that can be obtained are: *AB*, *Ab*, *aB* and *ab*. If all of these allelic combinations are observed in a sample then either recurrent mutation or recombination must have occurred somewhere in the history of the sample[44]. Assuming an INFINITE SITES MUTATION MODEL, recombination must be responsible — in this context the FOUR-GAMETE TEST (FGT) scores a recombination event if all four possible two-locus haplotypes occur (FIG. 1b).

Carrying out the FGT on all pairs of sites in a region identifies intervals at which recombination must have occurred. R_m is a conservative estimate of the minimum number of recombination events that have occurred in the history of the entire sample of chromosomes. R_m is obtained by assuming that all overlapping intervals, in which recombination is deemed to have occurred, originate from the same recombination event[23]. However, this assumption is very conservative, and it is often possible to detect that more than one recombination event has occurred in an interval by comparing the number of haplotypes with the number of polymorphic sites[45]. Briefly, if M haplotypes are observed in a region with N segregating sites, then at least $M–N$ recombination events must have occurred (if $M<N$ then the minimum number of recombination events is 0). $M–N$ is therefore a local lower bound and combining these local bounds allows the construction of the global minimum number of recombination events that have occurred in a region, R_h. Crucially, R_h can be used to check whether recombination events are clustered or occur uniformly throughout the genome[45]. The most sophisticated model-free (non-parametric) methods for counting recombination events try to reconstruct the underlying graph that corresponds to the one that would result from the smallest number of recombination events[46]; however, this approach is technically very challenging.

How well do such methods do at counting the true number of recombination events that have occurred in the history of a sample? Extensive simulations[40,47,48] have been used to address this question. Typically, only a small proportion of recombination events in the simulated genealogies can be detected in population-genetic data (FIG. 1c). The reason being that fairly specific conditions need to be met for a recombination event to be detectable: genetic diversity of the region, age of the event[49], sample size and demographic history are all important factors. The issue of sample size is particularly problematic — small samples are unlikely to contain the rarest gamete necessary for the detection of recombination events, but the rate at which adding extra chromosomes improves the estimate is extremely low (of the order of the log of the log of the sample size)[50]. In short, although we can use model-free counting methods to learn about recombination[45,51], we generally greatly underestimate the true level of recombination that has occurred[47,48].

Estimating recombination rates

There are two problems in trying to estimate recombination rates from population-genetic data. First, we cannot simply count the number of recombination events in the underlying graph because, as explained above, most leave no trace in the sample. Second, even if we could, we would have to know how many generations the underlying graph spanned to estimate the per-generation recombination rate. To overcome these difficulties we need to model the underlying process explicitly. The coalescent[34,52,53] is a probabilistic model (or stochastic process) that describes the distribution of the underlying tree or genealogy of a sample of chromosomes in an idealized population. Recombination can easily be incorporated into this model, in which case it is often referred to as the ancestral recombination graph (ARG)[39–42]. As with the standard coalescent[34,35,52], basic ARG models assume a constant population size with random mating, evolutionary neutrality and uniformity of recombination rates across the genome[54]. Generalization to more complicated demographic scenarios, however, is straightforward[33–35].

The population recombination rate. Within the context of the ARG, the key parameter in determining patterns of LD is the product of the PER-GENERATION RECOMBINATION RATE, r, and the EFFECTIVE POPULATION SIZE (N_e)[1]: $\rho = 4N_e r$ (where ρ is the POPULATION RECOMBINATION RATE; in population genetics per-generation rates typically appear in their product with the effective population size). r can depend on genomic factors, such as local sequence or DNA structure, whereas ρ also depends on demographic history (through N_e) and can therefore differ substantially between populations. Although direct measurements, such as those obtained through sperm-typing or pedigree studies, can be used to estimate r directly, population-genetic data can only be used to estimate ρ. Independent estimates of N_e are then required to infer r from ρ.

Model-based approaches. Before we review the various ways of estimating recombination rates within the context of the coalescent, it is worth considering the value of model-based methods of estimating the recombination rate. The idealized model of a population that was developed by Wright and Fisher — on which the coalescent is based — is far from biological reality. Natural populations, including human, do not, of course, conform to any such simple demographic models[55]. So, what can we learn by using a model that we know a priori is incorrect? First, the model might be incorrect in detail, but still offer a good description of the existing patterns of variation. In particular, the effective population size effectively subsumes details of the demographic history of a population into a single number, even though its precise relationship to CENSUS POPULATION SIZE is very difficult to evaluate[56]. Sometimes, however, demographic history cannot be collapsed into a single number[57,58]; moreover, different aspects of the data can result in different estimates for the effective population size. Nevertheless, demographic oversimplification might not be problematic, particularly if the researcher is interested in comparing recombination rates in different parts of the genome. Second, even if a model is incorrect (and all models are, to some degree) it could still allow us to make useful and accurate predictions about the recombination processes and related

HAPLOTYPE
The combination of alleles or genetic markers that is found on a single chromosome of a given individual.

INFINITE SITES MUTATION MODEL
A model that assumes that there are an infinite number of nucleotide sites and consequently that each new mutation occurs at a different locus.

FOUR-GAMETE TEST
(FGT). If all four possible gametes are observed for two bi-allelic loci then this test infers that a recombination event must have occurred between them (under an infinite sites mutation model).

PER-GENERATION RECOMBINATION RATE
(r). The probability of a recombination event occurring during meiosis.

EFFECTIVE POPULATION SIZE
(N_e). The size of the ideal constant-size population, in which the effects of random drift would be the same as those seen in the actual population.

POPULATION RECOMBINATION RATE
(ρ). Population-genetic parameters are generally proportional to the product of a molecular per-generation rate (for example, the per-generation recombination rate, r) and the effective population size (N_e). The population recombination rate has therefore often been defined as $\rho = 4N_e r$.

CENSUS POPULATION SIZE
Actual population size (total number of individuals) as compared to the theoretical effective population size.

Table 1 | **Methods for counting recombination events and estimating recombination rates from population-genetic data**

Brief description of method	Website address	References
Counting recombination events		
R_m	home.uchicago.edu/~rhudson1/source.html events	40
R_h	www.stats.ox.ac.uk/mathgen/software.html	45
Mimimum number of topologies underlying a sample	N/A	46
Moment-based estimators*		
Joint distribution of allele frequencies and LD	home.uchicago.edu/~rhudson1/source.html	23
Estimator based on average characteristics of four sequences	lifesci.rutgers.edu/~heylab	59
Likelihood analysis based on summary statistics of the data	home.uchicago.edu/~rhudson1/source.html	60
Full-likelihood approaches*		
Importance sampling	N/A	39
Lamarc and recombine (MCMC)	evolution.genetics.washington.edu/lamarc.html	12
Bayesian MCMC	N/A	102
IS implemented in two programs: INFS and FINS	www.maths.lancs.ac.uk/~fearnhead/software	10
Approximate-likelihood approaches*		
Pairwise likelihoods	home.uchicago.edu/~rhudson1/source.html	68
Marginal likelihoods/composite likelihoods	www.maths.lancs.ac.uk/~fearnhead/software	11
Pairwise likelihoods	www.stats.ox.ac.uk/mathgen/software.html	69
Approximate multi-locus haplotype diversity	www.stat.washington.edu/stephens/software.html	70

*Estimation of recombination rates. See main text and Box 1 for further details of methods. FINS, Full likelihood for finite site model; INFS, Full likelihood for infinite sites model; IS, Importance sampling; LD, linkage disequilibrium; MCMC, Markov Chain Monte Carlo; N/A, not applicable; R_h, improved estimate of the lower bound on recombination event number; R_m, estimated lower bound on the number of recombination events (see text for more details).

properties, such as LD patterns. Prediction is a great strength, not least because it allows models to be tested experimentally (for example, by comparing the estimated rates to those measured directly in sperm-typing studies).

In the next few sections, we briefly describe various coalescent-based methods for estimating the population recombination rate. These methods are also listed and compared in TABLE 1.

Moment estimators. Several ESTIMATORS have been constructed that use SUMMARY STATISTICS of the data to infer ρ[23,59,60]. For example, the VARIANCE of the pairwise differences between nucleotide sequences can be interpreted as a measure of LD and this or a different summary statistic can be used to estimate ρ[23]. These types of estimators are very easy to calculate and are computationally inexpensive, but they do not include all of the information that is contained in the data[32,37]. Conversely, the estimators that are described below attempt to include all aspects of the data, in particular the fact that the DNA sequences in the sample are correlated through the underlying genealogy.

Full-likelihood approaches. Full-likelihood approaches estimate the probability of observing a given data set under an assumed population-genetic model. They incorporate the genealogical structure that underlies a sample and attempt to use all of the information in the data. However, full-likelihood approaches are computationally extremely expensive. One of the model parameters to be estimated is ρ, but the model might also depend on the mutation rate and demographic parameters. In this approach, extensive simulations are carried out to estimate the LIKELIHOOD SURFACE for the model parameters (BOX 1). The likelihood of a parameter given an observed data set is proportional to the probability of the data given the model parameters[61,62]. The value of ρ at which the probability of observing the data is highest is the maximum-likelihood estimator of the recombination rate[61]. Calculating likelihoods in population genetics is notoriously complicated and the difficulties increase rapidly with the size of the data sets[38,63]. The best full-likelihood methods use MARKOV CHAIN MONTE CARLO (MCMC) and importance sampling[10,39] (IS) strategies to infer population-genetic parameters[13,37,38] (BOX 1).

Approximate-likelihood approaches. Computational cost is a great concern for full-likelihood approaches, which rely heavily on numerically intensive statistical procedures even for moderate data sets (for example, see REFS 64–67). Several approaches have therefore been developed to try to approximate the likelihood surface to scale estimation up to the large population-genetic data sets that are being investigated at present[11]. These approaches either ignore low frequency markers, which hold relatively little information about recombination, or they consider only a small number of markers at a time. Separate likelihoods are calculated for these subsets of the data and then combined to obtain the approximate likelihood estimator.

In the most extreme approximation only the allele distributions of two-locus systems are considered[68], and for each pair of sites the likelihood surface for the recombination (and possibly mutation) parameter is constructed independently. A 'composite likelihood' is then obtained by multiplying all pairwise likelihoods[68,69]. Although such approaches clearly ignore much of the information that is contained in the data, they are very

ESTIMATOR
A statistical method that is used to obtain a numerical estimate for a quantity of interest, such as a model parameter.

SUMMARY STATISTIC
A statistical function that summarizes complex data in terms of simple numbers (examples include the mean and variance).

VARIANCE
A statistic that quantifies the dispersion of data about the mean.

LIKELIHOOD SURFACE
The likelihood of a parameter is proportional to the probability of obtaining the observed data under a parametric model given the model parameter. The likelihood surface is a function/curve that specifies how well the data agrees with the predictions made by a parametric model for different values of the model parameter.

MARKOV CHAIN MONTE CARLO
A computational technique for the efficient numerical calculation of likelihoods.

Box 1 | Calculating coalescent likelihoods

The likelihood of a set of population-genetic parameters — such as the mutation rate and the recombination rate — is proportional to the probability of obtaining the sampled data under a stochastic model that assumes these parameter values[61]. Analytical expressions are impossible to derive under the coalescent model for anything except trivial data sets, so simulation-based techniques that obtain approximate probabilities are used instead[101].

Simulation-based (or Monte Carlo) methods vary considerably in their design, but all use the same central idea, which is to augment the data by its evolutionary history of coalescence, recombination and (usually) mutation[13]. Given a complete history, it is easy to calculate both the probability of the data given the history, and the probability of the history given the coalescent model and its parameters. The central difficulty is that there is an essentially infinite set of histories that could have given rise to the data, and finding those that are highly probable under the assumed model is comparable to the proverbial needle in a haystack dilemma.

Broadly, two approaches have been taken to finding probable genealogical histories. The first approach (known as Markov Chain Monte Carlo; MCMC) is to make an initial guess and then to make subsequent adjustments — changes that are more probable are accepted, whereas steps that are less likely are rejected, with a probability that is proportional to how likely they are to have occurred under the assumed model. The alternative approach is to exploit the fact that the likelihood of the data can be written as a RECURSION over possible ancestral states (that is, the data after a mutation has been removed or a coalescent event has occurred, and so on) and to develop ways of choosing the ancestral states in such a way that the sampler chooses probable histories (known as importance sampling). Both methods have advantages and disadvantages[38,101] and both are computationally intensive.

fast. In terms of bias and variance, simulations show that these so-called point estimates (that is, those with maximum composite likelihood) perform at least as well as any other *ad hoc* approach[68].

In addition to computational efficiency, approximate-likelihood approaches that are based on two- (or more) locus systems, which are known as composite-likelihood approaches, have several other advantages. First, because the number of possible combinations of allelic states in a two-locus system can be easily enumerated, these can be calculated without reference to the data and can be tabulated. The use of the resulting lookup tables can be used to further increase computational speed[68]. Second, composite approaches can deal directly with genotype data without having to infer haplotypes. Third, it is straightforward to consider complicated mutation and population models[68,69].

Several other approximate methods for estimating ρ have been[70], or are being, developed. Such approaches typically aim to capture some aspects of the ARG to estimate the recombination parameter. Increasingly, such methods will be used to consider a wide range of demographic models, different models of the recombination process (for example, those that incorporate GENE CONVERSION), as well as details of the experimental design.

Comparing estimators. How do we compare different estimators? The following factors might be of concern: bias, variance, statistical efficiency, robustness to deviations from the assumed model, computational efficiency and, finally, consistency. Do we expect the estimator to obtain the correct value? How much variability do we expect in estimates for a given value of the underlying recombination rate? How much of the information contained in the data does each estimator use? How wrong will the estimate be if the assumptions are incorrect? How fast can the estimate be computed? Would the estimator be perfect if we had unlimited amounts of data?

Moment methods are typically the fastest to compute estimates, but they also use little information, have considerable variance and can be strongly affected by deviations from model assumptions. Full-likelihood methods, by contrast, use the most information but are computationally expensive (often prohibitively so). Approximate-likelihood methods generally lie somewhere between the two extremes, though some of them lack consistency[71]. Of these methods, two-locus sampling distributions seem to be somewhat less influenced by demography than multi-locus systems and are little biased by SNP ascertainment, which indicates considerable robustness (E. DeSilva and M.P.H.S., unpublished data).

Applications

There are broadly four types of question that can be asked when using estimators that are based on population-genetic data. How much recombination occurs in one region compared to another? Is there variation in the recombination rate within the region that is being analysed? Is the effective population size of one population greater or smaller than that of another population? And, how good is the neutral coalescent with recombination at describing patterns of variation (that is, is there evidence for the action of natural selection or complex demographic processes)? Here we discuss studies that address some of these questions.

BOX 2 shows the results of a comparison between full- and composite-likelihood estimators of the population recombination rate for two human genes, lymphotoxin α (*LTA*) and lymphotoxin β (*LTB*), which are members of the tumour necrosis factor family. The genes show large differences in the amount of recombination — *LTA* shows a considerable amount of recombination, whereas *LTB* shows little or none. Furthermore, the estimate for *LTA* is considerably higher in the African sample compared to the European sample. This pattern, which is consistent across several genes, is probably due to the different demographic histories of the two

RECURSION
A repeated mathematical operation that is often used to aid numerical analysis.

GENE CONVERSION
The non-reciprocal transfer of genetic information between homologous genes as a consequence of mismatch repair after heteroduplex formation.

Box 2 | **Comparing maximum-likelihood estimates.**

The table shows the full- and composite-likelihood estimates of the population recombination rate (ρ) obtained from two human genes, lymphotoxin α (*LTA*) and lymphotoxin β (*LTB*), in African and European populations. The full-likelihood estimates were obtained after PHASING of the genotypes using PHASE[103] (see the online links box). Composite-likelihood estimates were obtained from genotype data (see the online links box). The mutation rate was determined using Watterson's estimator[104]. One possible interpretation of these results is that recombination is suppressed in *LTB* for functional reasons. It is unlikely that association mapping in *LTB* would allow the fine localization of a potential causal functional mutation within the gene. By contrast, it might be possible to do fine-mapping at the gene level in *LTA*, especially in African population samples.

The difference in computational effort between the two approaches is substantial. Calculating the maximum-likelihood estimate in *LTA* in Africans using the full-likelihood approach took just over 6 central processing unit (CPU) hours on a Pentium 4 Xeon 2.4 Ghz, whereas the composite-likelihood estimate was obtained in slightly less than 2 CPU minutes on the same machine.

Gene (gene size)	Full-likelihood estimate of ρ		Composite-likelihood estimate of ρ	
	Africa*	Europe‡	Africa*	Europe‡
LTA (4.9 kb)	7.3	5.5	11	7
LTB (4.4 kb)	0	1	3	0

*n = 24. ‡n = 23. kb, kilobases.

populations (as captured in their respective effective population sizes) and could reflect a population bottleneck during the geographic expansion of modern humans out of Africa. A more consistent picture of the differences in effective population size is obtained by averaging across many genomic regions[14,64,72] (FIG. 2).

Population-genetic methods can also be used to detect local variation in the recombination rate, such as in recombination hotspots. The question of whether recombination events are clustered in hotspots is of enormous interest at present, and being able to answer it unambiguously could have great relevance in the efficient design of ASSOCIATION STUDIES[2,8,29,73,74]. FIGURE 3 shows the estimated recombination rate profile that was calculated from 50 unrelated males in a genomic region that contains a known hotspot[4]. The resulting estimate agrees remarkably with the sperm-typing analysis in terms of the location of the hotspot and the degree to which the rate is elevated. These examples show that even very localized features of the underlying recombination process can be detected given high quality data. Published hotspots are about 1–2 kb wide, so an accurate profile of the recombination process would require an SNP density of >1 per 1kb. Obtaining this density is feasible but expensive.

The great power of model-based estimation methods is their ability to provide testable hypotheses[61]. Testing models has benefits either way. If the model cannot be shown to be wrong (it can never be proved right) it will suffice, and if it is proved wrong we can learn important biological lessons from trying to understand why it is wrong. The idea behind model testing is to estimate parameters within the context of a model, then carry out simulations (or, where possible, derive mathematical expressions) to ask whether particular features of the data are compatible with the assumed model. For example, model-testing approaches to LD have been used to reveal the importance of historical population bottlenecks[64], recombination rate variation[75] gene conversion[76] and adaptive evolution[77] as factors that influence human diversity. An alternative route to model testing is to compare population-genetic based estimates with independent estimates of the same quantities. For example, the ratio of population-genetic estimates of ρ and θ (the

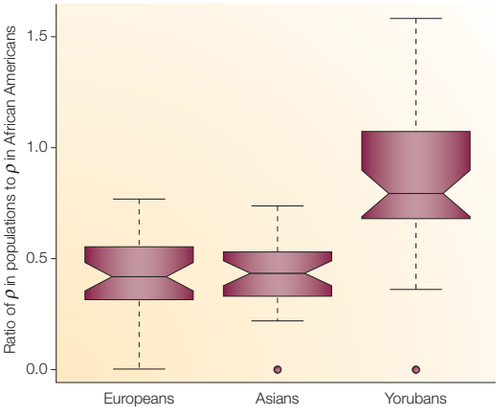


Figure 2 | **The behaviour of estimators is largely independent of genomic region.** The graph shows the ratios of the average inferred population recombination rates (ρ) for 39 genomic regions. For each region, maximum-likelihood estimates for the average recombination rate obtained from European, Asian and Yoruban population samples were divided by the inferred rate in an African American sample. The population samples were taken from Gabriel *et al.*[65] and recombination rates were inferred using a composite-likelihood estimator directly from the genotypes without inferring haplotypes[69]. The box-plots show the 25–75% regions of the distribution of ratios. The horizontal line inside each box denotes the median and if the notches in two different boxes overlap then their medians are not significantly different. The whiskers of each box extend to the approximate 95th percentiles of the distribution; outliers are indicated by the solid circles. We find that the distributions of ratios are relatively tight. This indicates that the estimator behaves in a very similar way in different genomic regions.

PHASING
Determining the haplotype phase (the arrangement of alleles at two loci on homologous chromosomes) from genotype data using statistical methods.

ASSOCIATION STUDIES
A set of methods that are used to correlate polymorphisms in genotype to polymorphisms in phenotype in populations.

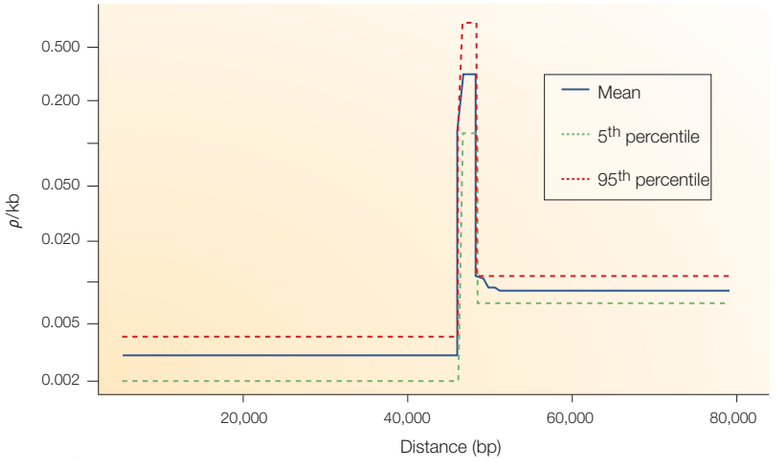


Figure 3 | **Estimating local recombination rate variation in a known recombination hotspot.** We used population-genetic data from 50 unrelated United Kingdom males to estimate the local recombination rate, ρ, in a region with a known hotspot[4]. Both the intensity and location of the hotspot are in very good agreement with the values that are obtained from the sperm-typing analysis that is described in REF 4. Most of the recombination events seem to cluster in a small region. Whereas sperm-typing approaches, by definition, can only estimate male recombination rates, the population-genetic data is a combination of the behaviour of female and male recombination. bp, base pairs; kb, kilobases.

population mutation rate) should be the same as the ratio of experimental estimates of the per-generation recombination and mutation rates. That the ratios do not agree in humans[78] indicates that the assumed model might lack an important element of biological reality.

Further complications of biological reality

Any process of inference is based on explicit or implicit assumptions, and if those assumptions are not correct then they will affect the accuracy of our inferences[32,37,79]. It is therefore important to understand which aspects of biological reality are likely to affect the inferences that are made about recombination rates. Several biological factors might contribute to MODEL MIS-SPECIFICATION. Some of these biological factors are described below, along with some possible ways of addressing the challenges that they might pose.

Mutation. In many species (such as viruses and bacteria) and at certain positions in the human genome (such as CPG ISLANDS), many mutations have occurred at a single nucleotide position in the history of the sample[80]. It is important to detect when this has occurred because recurrent or back mutation can create patterns of variation that resemble those caused by recombination[69] (a phenomenon that is known as homoplasy). Several methods have been developed to try and distinguish between homoplasy and recombination as the genome-wide source of such patterns[81]. The more reliable of these methods considers the fact that recombination generally occurs more frequently between physically distant loci than neighbouring ones. Such methods seem to be robust to the complexities of mutational processes in organisms such as the human immunodeficiency virus (HIV), and coalescent-based methods to estimate recombination rates have also been developed for such genomes.

MODEL MIS-SPECIFICATION
The consequence of using a parametric model in the inference process that is different from the true model under which the data was generated.

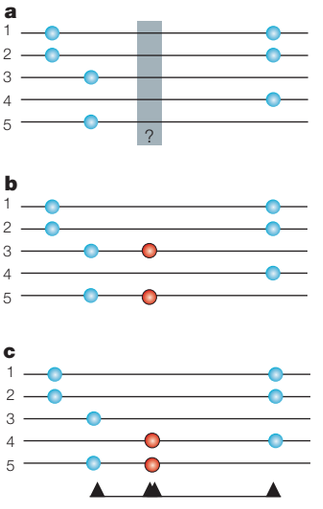
CPG ISLANDS
Genome sequences of >200 base pairs that have high G+C content and CpG frequency.

Box 3 | **Recombination and the hidden SNP problem**

Imagine that a researcher wants to identify a locus that underlies a phenotype of interest by carrying out an association-mapping study using single nucleotide polymorphisms (SNPs). The first step would be to conduct a survey of genetic variation at a given genomic region by typing SNPs collected from several randomly sampled individuals. The sampled haplotypes (1–5) are shown in part **a** of the figure, with the typed SNPs depicted in blue. Before embarking on the association-mapping experiment, it is necessary to ascertain that variation at interspersed sites that have not been examined (the shaded region in part **a** and red dots in part **b**) will be in strong linkage disequilibrium (LD) with variation at the polymorphisms that have been typed. If an interspersed SNP that contributes to phenotypic variation is not in strong LD (part **c**; here an untyped SNP is depicted as no longer being associated with a typed SNP in haplotype 3), subsequent mapping will have low power. Whether this is likely or not will depend on the recombination rate in the region, which can be estimated as described in the main text.

In this example, and assuming the standard neutral model, the conditional probability, P_s, (see part **d**) that an SNP typed in the shaded region is not in LD with the typed SNPs (defined as revealing at least one recombination event) ranges from 0 (if $\rho = 4N_e r = 0$) to 0.24, if there is free recombination. For the observed haplotypes, we can estimate the likelihood of any given recombination rate — the most likely value is 0, but the approximate 95% confidence interval goes up to $\rho = 10$. In this instance, because there is a significant risk that intervening SNPs will not be in strong LD with the typed SNPs, collecting more detailed data (both more SNPs and more chromosomes) would be recommended before proceeding with the association-mapping study.

N_e, effective population size; ρ, population recombination rate; r, per-generation recombination rate; P_s, conditional probability that a SNP is not in LD with the known SNPs.



d

$4N_e r$	P_s
0	0
4	0.054
20	0.119
∞	0.240

TEMPLATE SWITCHING
The process by which RNA templates are switched between viral genomes during reverse transcription.

BOTTLENECK
A temporary marked reduction in population size.

SELECTIVE SWEEP
The process by which positive selection for a mutation eliminates neutral variation at linked sites.

HARDY–WEINBERG EQUILIBRIUM
A state in which the frequency of each diploid genotype at a locus equals that expected from the random union of alleles.

Variation in the recombination process. The process of recombination can also vary between organisms. For example, gene conversion is an integral aspect of recombination in eukaryotes, but it is generally not considered in methods of inference[76]. Similarly, the recombination process in HIV is very different from that of most organisms[82,83], and the rate at which it occurs can depend on the degree of sequence divergence between genomes. However, the evolutionary consequences[84] of gene conversion or TEMPLATE SWITCHING in HIV are easily incorporated into coalescent models and present no major obstacle to methods of inference[82,85,86].

Demographic history. The presence of constant population size with random mating is perhaps the most unreasonable assumption that is made by standard coalescent methods of inference. As explained above, the concept of effective population size goes some way to subsuming many of the details of demographic history, but factors that have a large influence on LD — strong BOTTLENECKS[16,87], population subdivision[21], highly restricted gene flow[14], selfing[88], recent and complete SELECTIVE SWEEPS[78,89], marker ascertainment[27], and so on — also have a considerable impact on estimators and the ability to detect recombination rate variation[54,73,90]. However, such extreme forces should often be readily detectable from other aspects of the data, such as levels of diversity, the frequency distribution of mutations and deviation from the HARDY–WEINBERG EQUILIBRIUM[1,91]. Where such departures from neutrality are not detectable, estimates of the relative recombination rate are likely to be reliable (within the variance of the estimator)[60,92,93]. Furthermore, estimates of the recombination rate to mutation rate ratio can potentially correct for variation in N_e over the genome. Alternatively, estimation methods can attempt to jointly estimate details of the recombination and demographic history. So far there has been little development in these important areas owing to the massive computational burden of full-likelihood calculations under even simple demographic models. However, the advances in the use of approximations to full-likelihood approaches that are described above[11,68,69] are making it possible to make joint inferences about recombination and other evolutionary forces.

Neutrality. The assumption of evolutionary neutrality[1] can also introduce serious bias into the recombination rate estimators if it is invalid. In particular, if, as is certain[94], selection has varied across the genome (for example, through the effects of localized selective sweeps[89]), then local estimates of the recombination rate might be biased. Despite this, recent research indicates that at least some estimators are robust to all but extreme selection events, which can be detected by standard neutrality tests (C. Spencer and G.A.T.M., unpublished data). Alternatively, a joint inference of recombination and natural selection might be possible. For example, Przeworski[95] considered the joint estimation of recombination rate and the parameters of a selective sweep by using the summary statistic likelihood estimator of Wall[92].

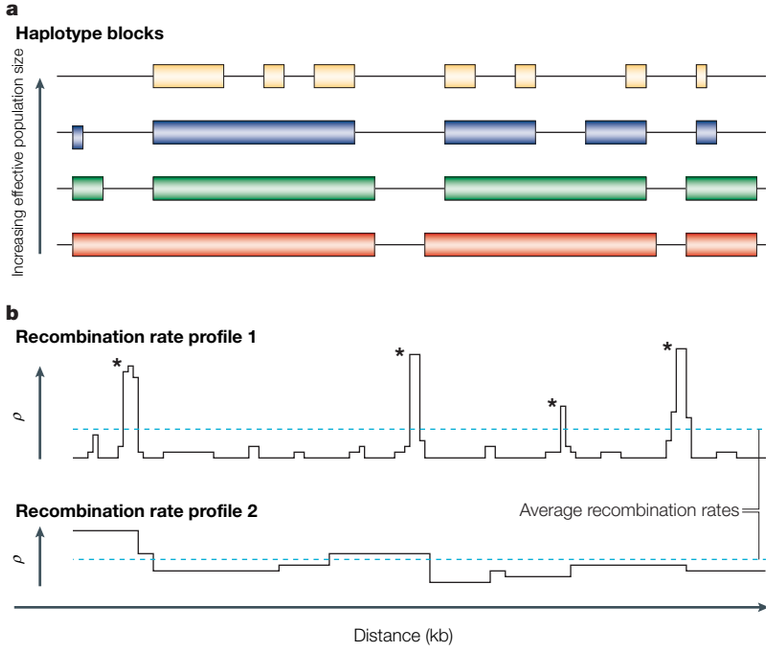


Figure 4 | **Blocks and the interplay of recombination rate and demography. a** | Haplotype and/or linkage disequilibrium (LD) blocks are expected (and seen[65]) to depend on the sample populations. Generally, the larger the effective population size the smaller the blocks will be. **b** | It is well known that haplotype and/or LD blocks will arise by chance even if the recombination rate is uniform[18]. If recombination hotspots (profile 1; denoted by *) are ubiquitous features of the human genome, then some aspects of blocks will be transferable between populations, with details of the block pattern dependent on demography. If, however, recombination shows only mild levels of variation then blocks reflect past recombination events and only very old recombination events can result in block boundaries that are shared between populations (profile 2). So, whether or not blocks offer a convincing description of genetic diversity depends on how the recombination rate varies along a stretch of DNA. kb, kilobases; ρ, population recombination rate.

Interpreting LD data

So far we have largely considered what can be learned about recombination from patterns of LD. However, estimates of the population recombination rate can also be used to inform the design of experiments that use LD to map the genetic basis of human variation. At the simplest level, an estimate of the recombination rate can be thought of as a summary statistic of LD, which can be compared directly between genes and populations. By contrast, patterns of LD from different samples are often very difficult to interpret[2,14,72]. However, far more important is the ability to use estimates of the recombination rate (and the coalescent framework[49]) to model patterns of genetic variation, either in regions of the genome that have not been directly assayed in the experiment (for example, by typing sparse sets of SNPs), or the same region but in a subsequent survey (such as in a different population).

This ability will be of considerable importance in the application of SNP-based LD surveys such as the HapMap project[26]. The HapMap project aims to reduce human genetic variation to a set of representative markers that can then be used in HAPLOTYPE-BASED

HAPLOTYPE-BASED APPROACH
An approach to association studies in which the co-inheritance of phenotypes and haplotypes — as opposed to single markers — is statistically analysed.

TAGGING APPROACH
Identifying sub-sets of markers ('tags') that describe patterns of association or haplotypes among larger marker sets.

MINIMUM-DESCRIPTION LENGTH APPROACHES
A concept from information theory, in which all of the information contained in a system (for example, a sample of DNA sequences) is described in the most compact form possible.

or TAGGING[96] APPROACHES for the study of complex diseases, for example, as markers in genome-wide association studies. The success of the approach requires that typed SNPs adequately capture patterns of variation at untyped loci (through LD). Whether or not this is true depends, to a large degree, on the level of recombination (BOX 3). Similarly, haplotype diversity in a region is determined by the local recombination rate[70]. Estimating fine-scale variation in the recombination rate (for example, the location of hotspots) could therefore have profound implications for marker selection, not least because it can provide us with an idea of how certain we can be that typed SNPs adequately capture variation within the region.

A related issue is that estimates of the local recombination rate can be used to address whether haplotype blocks[97,98] are real (in the sense that they are regions of low recombination that are bounded by recombination hotspots) or stochastic (in the sense that they represent chance historical events) features of the human genome (FIG. 4). If most recombination events fall within small and easily defined regions of the genome, that is, if within-hotspot events account for most recombination events, then blocks might be transferable between populations and offer a useful description of genetic diversity in genetic association studies. In the absence of true hotspots, however, block definitions that are based on summary statistics of haplotype diversity and/or LD can still give rise to blocks, but they are population (and potentially sample) specific[25,26]. By contrast, MINIMUM-DESCRIPTION LENGTH APPROACHES, at least in simulated data, often locate block boundaries at recombination hotspots[99,100]. Estimating local recombination rate variation is therefore crucial for assessing whether or not haplotype blocks reflect genuine recombination rate variation or are just artefacts of the block-detection algorithm[25,99].

Conclusions

It has long been known that knowledge of the recombination rate will improve understanding of patterns of LD in genomes. As population-genetic approaches are becoming increasingly important in biomedical research, through genetic association and/or functional studies, understanding the recombination process is an important challenge. The recent theoretical developments that have been reviewed here make the estimation of reliable recombination rates from population-genetic data possible, even if estimated recombination rates will, of course, be biased by ignoring factors such as demography and selection. In addition, they allow us to extract considerable information about the recombination process. We therefore expect that knowledge of (estimated) recombination rates will augment LD studies and aid in their design and interpretation.

1. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer, Sunderland, 1998).
2. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
This work highlights issues that are related to the application of LD data to association studies.
3. Kaplan, N. & Morris, R. Prospects for association-based fine mapping of a susceptibility gene for a complex disease. *Theor. Popul. Biol.* **60**, 181–191 (2001).
4. Jeffreys, A. J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human *TAP2* recombination hotspot. *Hum. Mol. Genet.* **9**, 725–733 (2000).
5. Badge, R. M., Yardley, J., Jeffreys, A. J. & Armour, J. A. Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Hum. Mol. Genet.* **9**, 1239–1244 (2000).
6. Cullen, M., Erlich, H., Klitz, W. & Carrington, M. Molecular mapping of a recombination hotspot located in the second intron of the human *TAP2* locus. *Am. J. Hum. Genet.* **56**, 1350–1358 (1995).
7. Zhao, H. Family-based association studies. *Stat. Methods Med. Res.* **9**, 563–87 (2000).
8. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).
9. Jeffreys, A. J., Murray, J. & Neumann, R. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol. Cell* **2**, 267–273 (1998).
10. Fearnhead, P. & Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
11. Fearnhead, P. & Donnelly, P. Approximate likelihood methods for estimating local recombination rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 657–680 (2002).
12. Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401 (2000).
13. Stephens, M. & Donnelly, P. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62**, 605–635 (2000).
14. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
A comprehensive review of LD and its dependence on demography; the paper also examines the connection between theoretical models and experimental data.
15. Golding, G. B. The sampling distribution of linkage disequilibrium. *Genetics* **108**, 257–274 (1984).
16. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
17. Calafell, F., Grigorenko, E. L., Chikanian, A. A. & Kidd, K. K. Haplotype evolution and linkage disequilibrium: a simulation study. *Hum. Hered.* **51**, 85–96 (2000).
18. Wang, N., Akey, J. M., Zhang, K., Chakraborty, R. & Jin, L. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71**, 1227–1234 (2002).
19. Barton, N. H. Genetic hitchhiking. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **355**, 1553–1562 (2000).
20. Charlesworth, B., Nordborg, M. & Charlesworth, D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* **70**, 155–174 (1997).
21. Chapman, N. H. & Thompson, E. A. Linkage disequilibrium mapping: the role of population history, size, and structure. *Adv. Genet.* **42**, 413–437 (2001).
22. Freimer, N. B., Service, S. K. & Slatkin, M. Expanding on population studies. *Nature Genet.* **17**, 371–373 (1997).
23. Hudson, R. R. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631 (1985).
24. Garner, C. & Slatkin, M. On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. *Genet. Epidemiol* **24**, 57–67 (2003).
25. Phillips, M. S. *et al.* Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genet.* **33**, 382–387 (2003).
A study of a dense marker map on chromosome 19 that, together with a detailed theoretical analysis, highlights problems in defining haplotype blocks.
26. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19**, 135–140 (2003).
27. Akey, J. M., Zhang, K., Xiong, M. M. & Jin, L. The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* **20**, 232–242 (2003).
28. Nielsen, R. & Signorovitch, J. Correcting for ascertainment bias when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.* **63**, 245–255 (2003).
29. Rannala, B. & Slatkin, M. Likelihood analysis of disequilibrium mapping, and related problems. *Am. J. Hum. Genet.* **62**, 459–473 (1998).
30. Zollner, S. & von Haeseler, A. A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **66**, 615–628 (2000).
31. Nordborg, M. & Tavare, S. Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**, 83–90 (2002).
A careful attempt at discussing the effects of population history on LD in a genealogical framework.
32. Stumpf, M. P. H. & Goldstein, D. B. Genealogical and evolutionary inference with the human Y chromosome. *Science* **291**, 1738–1742 (2001).
33. Donnelly, P. & Tavare, S. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**, 401–421 (1995).
34. Nordborg, M. in *Handbook of Statistical Genetics* (eds Balding, D. J. M. B. & Cannings, C.) 179–212 (Wiley, Chichester, 2000).
A modern exposition of the coalescent and its application in modern population genetics.
35. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* (ed. Futuyama, D. J. A.) 1–43 (Oxford University Press, Oxford, 1990).
36. Tavare, S. A genealogical view of some stochastic-models in population-genetics. *Stochastic Processes and their Applications Abstr.* **19**, 10 (1985).
37. Tavare, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
38. Stephens, M. in *Handbook of Statistical Genetics* (eds Balding, D. J. M. B. & Cannings, C.) 213–238 (Wiley, Chichester, 2001).
A detailed and highly accessible account of statistical inference in population genetics using the coalescent.

39. Griffiths, R. C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502 (1996).
40. Hudson, R. R. & Kaplan, N. L. The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840 (1988).
41. Wiuf, C. & Hein, J. The ancestry of a sample of sequences subject to recombination. *Genetics* **151**, 1217–1228 (1999).
42. Wiuf, C. & Hein, J. Recombination as a point process along sequences. *Theor. Popul. Biol.* **55**, 248–259 (1999).
43. Kuhner, M. K., Beerli, P., Yamato, J. & Felsenstein, J. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156**, 439–447 (2000).
44. Weir, B. S. Inferences about linkage disequilibrium. *Biometrics* **35**, 235–254 (1979).
45. Myers, S. R. & Griffiths, R. C. Bounds on the minimum number of recombination events in a sample history. *Genetics* **163**, 375–394 (2003).
46. Wiuf, C. On the minimum number of topologies explaining a sample of DNA sequences. *Theor. Popul. Biol.* **62**, 357–363 (2002).
47. Posada, D. & Crandall, K. A. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl Acad. Sci. USA* **98**, 13757–13762 (2001).
48. Wiuf, C., Christensen, T. & Hein, J. A simulation study of the reliability of recombination detection methods. *Mol. Biol. Evol.* **18**, 1929–1939 (2001).
49. McVean, G. A. A genealogical interpretation of linkage disequilibrium. *Genetics* **162**, 987–991 (2002).
This paper discusses LD in a genealogical framework and shows how features of the genealogy are connected to LD summary statistics.
50. Myers, S. *The Detection of Recombination Events Using DNA Sequence Data*. Thesis, Univ. Oxford (2003).
51. Wiuf, C. & Hein, J. On the number of ancestors to a DNA sequence. *Genetics* **147**, 1459–1468 (1997).
52. Kingman, J. F. C. The coalescent. *Stochastic Processes and their Applications* **13**, 235–248 (1982).
53. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet.* **3**, 380–390 (2002).
54. Wiuf, C. & Posada, D. A coalescent model of recombination hotspots. *Genetics* **164**, 407–417 (2003).
55. Cavalli-Sforza, L. L., Mennazzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, 1996).
56. Rannala, B. Gene genealogy in a population of variable size. *Heredity* **78**, 417–423 (1997).
57. Wakeley, J. & Lessard, S. Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics* **164**, 1043–1053 (2003).
58. Nordborg, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with selfing. *Genetics* **154**, 923–929 (2000).
59. Hey, J. & Wakeley, J. A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846 (1997).
60. Wall, J. D. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* **17**, 156–163 (2000).
61. Cox, D. R. & Hinkley, D. V. *Theoretical Statistics* (Chapman and Hall, London, 1974).
62. Casella, G. & Berger, R. L. *Statistical Inference* (Duxbury, Pacific Grove, 2002).
63. Steel, M. & Penny, D. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* **17**, 839–850 (2000).
64. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
65. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
An influential experimental study that investigates the presence of haplotype blocks in different populations across 52 genomic regions.
66. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
A beautiful experimental study of recombination hotspots and associated patterns of LD in a human population sample.
67. Clark, A. G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
68. Hudson, R. R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817 (2001).
The first study to estimate recombination rates using pairwise approximation to the likelihood.
69. McVean, G., Awadalla, P. & Fearnhead, P. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241 (2002).
70. Li, N. & Stephens, M. A new multilocus model for linkage disequilibrium, with application to exploring variations in recombination rate. *Genetics* (in the press).
71. Fearnhead, P. Consistency of estimators of the population-scaled recombination rate. *Theor. Popul. Biol.* **64**, 67–79 (2003).
72. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Rev. Genet.* **3**, 299–309 (2002).
73. Stumpf, M. P. & Goldstein, D. B. Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr. Biol.* **13**, 1–8 (2003).
74. Stumpf, M. P. Haplotype diversity and the block structure of linkage disequilibrium. *Trends Genet.* **18**, 226–228 (2002).
75. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
76. Frisse, L. *et al.* Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**, 831–843 (2001).
77. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
78. Przeworski, M. & Wall, J. D. Why is there so little intragenic linkage disequilibrium in humans? *Genet. Res.* **77**, 143–151 (2001).
79. Griffiths, R. C. & Tavare, S. Ancestral inference in population-genetics. *Stat. Sci.* **9**, 307–319 (1994).
80. Smith, J. M., Smith, N. H., O'Rourke, M. & Spratt, B. G. How clonal are bacteria? *Proc. Natl Acad. Sci. USA* **90**, 4384–4388 (1993).
81. Smith, J. M. The detection and measurement of recombination from sequence data. *Genetics* **153**, 1021–1027 (1999).
82. Holmes, E. C. On the origin and evolution of the human immunodeficiency virus (HIV). *Biol. Rev* **76**, 239–254 (2001).
83. Fu, Y. X. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol. Biol. Evol.* **18**, 620–626 (2001).
84. Awadalla, P. The evolutionary genomics of pathogen recombination. *Nature Rev. Genet.* **4**, 50–60 (2003).
85. Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–1320 (2002).
86. Grassly, N. C. & Holmes, E. C. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* **14**, 239–247 (1997).
87. Hey, J. & Harris, E. Population bottlenecks and patterns of human polymorphism. *Mol. Biol. Evol.* **16**, 1423–1426 (1999).
88. Nordborg, M. & Donnelly, P. The coalescent process with selfing. *Genetics* **146**, 1185–1195 (1997).
89. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189 (2002).
90. Posada, D. & Wiuf, C. Simulating haplotype blocks in the human genome. *Bioinformatics* **19**, 289–290 (2003).
91. Gillespie, J. H. *Population Genetics: a Concise Guide* (Johns Hopkins Univ. Press, Baltimore, 1998).
92. Wall, J. D. Recombination and the power of statistical tests of neutrality. *Genet. Res.* **74**, 65–79 (1999).
93. Brown, C. J., Garner, E. C., Dunker, A. K. & Joyce, P. The power to detect recombination using the coalescent. *Mol. Biol. Evol.* **18**, 1421–1424 (2001).
94. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1991).
95. Przeworski, M., Charlesworth, B. & Wall, J. D. Genealogies and weak purifying selection. *Mol. Biol. Evol.* **16**, 246–252 (1999).
96. Johnson, G. C. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
This paper pioneered the concept of haplotype tagging to describe genetic variation.
97. Wall, J. D. & Pritchard, J. K. Assessing the performance of haplotype block models of linkage disequilibrium. *Am. J. Hum. Genet.* **73**, 502–515 (2003).
98. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–597 (2003).
99. Anderson, E. C. & Novembre, J. Finding haplotype block boundaries by using the minimum-description-length principle. *Am. J. Hum. Genet.* **73**, 336–354 (2003).
100. Koivisto, M. *et al.* in *Pac. Symp. Biocomput. 2003* (eds Altman, R. B., Dukner, A. K., Hunter, L., Jung, T. A. & Klein, T. E.) 502–513 (World Scientific, Singapore, 2002).
101. Liu, J. S. *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2003).
102. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942 (2000).
103. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
104. Watterson, G. A. On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).

Acknowledgments

We thank A. Jeffreys and P. Donnelly for useful discussions, and C. Wiuf, M. Slatkin, L. Cardon, G. Coop, C. Spencer and three anonymous referees for their helpful comments on earlier drafts of this manuscript. Generous support through research fellowships from the Wellcome Trust (to M.P.H.S) and the Royal Society (to G.A.T.M.) is gratefully acknowledged.

Conflicting interests statement

The authors declare that they have no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/
LTA | *LTB*

FURTHER INFORMATION

Michael Stumpf's laboratory:
http://www.imperial.ac.uk/biologicalsciences/research/stumpf
Gilean McVean's laboratory:
http://www.stats.ox.ac.uk/people/mcvean/index.htm
***LTA* and *LTB* genotypes:** http://pga.gs.washington.edu/data/
SHOX genotypes: http://www.leicester.ac.uk/ge/ajj/SHOX/
Data of Gabriel *et al.* :
http://www.genome.wi.mit.edu/mpg/hapmap/
PHASE software:
http://www.stat.washington.edu/stephens/software.html
Access to this interactive links box is free online.