Insights into recombination from population genetic variation

Garrett Hellenthal and Matthew Stephens

Patterns of genetic variation in natural populations are shaped by, and hence carry valuable information about, the underlying recombination process. In the past five years, the increasing availability of large-scale population genetic data on dense sets of markers, coupled with advances in statistical methods for extracting information from these data, have led to several important advances in our understanding of the recombination process in humans. These advances include the identification of large numbers of 'hotspots', where recombination appears to take place considerably more frequently than in the surrounding sequence, and the identification of DNA sequence motifs that are associated with the locations of these hotspots.

Addresses
Department of Statistics, University of Washington, Seattle, WA 98195, USA

Corresponding author: Stephens, Matthew (stephens@stat.washington.edu)

Current Opinion in Genetics & Development 2006, **16**:565–572

This review comes from a themed issue on
Genomes and evolution
Edited by Chris Tyler-Smith and Molly Pizeworski

Available online 16th October 2006

0959-437X/$ – see front matter

DOI 10.1016/j.gde.2006.10.001

Introduction

Recombination tends to reduce population-level association among alleles at neighboring markers, commonly known as 'linkage disequilibrium' (LD). As a result, the expected amount of LD between markers depends on the recombination rate between them [1], and by measuring LD in natural populations one can attempt to learn about the underlying recombination process. This potential for genetic variation in natural populations to yield insights into recombination has been recognized for some time [2,3]. However, until recently, its usefulness in practice has been limited by two factors: the expense of collecting appropriate data on fine-scale population genetic variation, and the lack of efficient statistical methods for performing analyses. The past five years have seen significant advances on both fronts. The development of cheap, reliable, high-throughput genotyping technologies has facilitated the collection of dense genome-wide data on human genetic variation [4,5]. And advances in statistical methodology (e.g. [6–8]) have made it easier to extract the desired information from these data. Together, these advances have led to several exciting new insights into the recombination process, particularly recombination in humans. Here, we review both the basic ideas behind the use of population data to learn about recombination and some of the new knowledge this approach has recently produced.

Recombination and population genetic variation

Patterns of genetic variation in a sample of unrelated individuals from a population are the product of many mutation and recombination events that have occurred over many generations in the ancestors of that sample. One consequence of this is that population genetic data provide estimates of the *average* recombination rate over many individuals, both males and females, over a long period of time. This is worth remembering when comparing estimates from population data with estimates from other approaches, such as sperm-typing.

Another consequence of genetic variation being shaped over many generations is that it has the potential to inform about recombination rates on a very fine scale: although few recombinations occur in any single meiosis, the ancestral history of a sample of humans typically spans many thousands of meioses, and so even a single kilobase of sequence might have undergone several recombination events. More specifically, the number of crossover events experienced in the ancestry of a given region depends on a quantity known as the 'population recombination parameter', usually denoted by ρ, and defined to be $\rho = 2N_e c$, where c is the probability of crossover occurring in the region in a single meiosis, and N_e is the 'effective population size', which can be thought of as a measure of the average relatedness of individuals in the population. Because patterns of variation in population samples reflect the number of crossover events experienced in their ancestry, population genetic data actually provide estimates of ρ rather than direct estimates of c. In other words, they provide estimates of the recombination rate only up to some (unknown) population-specific scaling constant.

The fact that estimates of recombination rates from population data are scaled by a population-specific constant is important to bear in mind when comparing estimates obtained from samples from different populations. However, it is not really a severe impediment to learning about recombination from population data. For example, provided one is willing to assume that the average relatedness, N_e, is constant across the genome — this is probably reasonable in the absence of strong

selection — then one can learn about *relative* recombination rates in different genomic regions by comparing the respective estimates for ρ. Furthermore, it is also possible to estimate N_e itself, for example by comparing estimates of c from genetic maps with estimates of ρ from population data.

Extracting information from population data

The ability to use population data to learn about fine-scale recombination rates across the genome is an enticing prospect, particularly given that genetic data on pedigrees of closely related individuals, which form the basis of traditional genetic maps [9,10], provide estimates of the recombination rate with a much lower resolution (e.g. available maps do not provide accurate estimates of the recombination rate over regions smaller than about 1 Mb in humans). However, unlike data on pedigrees, the relationships among samples from a population are unknown, which makes the task of inferring recombination rates from population data substantially more difficult, at least conceptually. Although many approaches have been developed for this problem, the most practical and powerful approaches available today all grew out of work based on the coalescent [11,12].

The basic idea behind the coalescent is to attempt to model the ancestry of genetic material sampled in the current population, following it back in time through the mutation and recombination events that have occurred since the sampled individuals shared a common ancestor. This is a natural approach to the problem because, as noted above, it is exactly these mutation and recombination events that have produced the patterns of genetic variation observed in a sample. If one could somehow deduce this sequence of events from the data and identify where each recombination event occurred, then this would provide an estimate of the population recombination rate, ρ, and a powerful means to deduce properties of the underlying recombination process. However, for moderate-sized genetic regions (e.g. a few kb in humans), there are typically huge numbers of possible event sequences that could have created the observed data. As a result, despite several attempts (e.g. [7,13,14]), approaches based on attempting to directly reconstruct the sequence of recombination events are generally computationally intractable for realistic datasets [7].

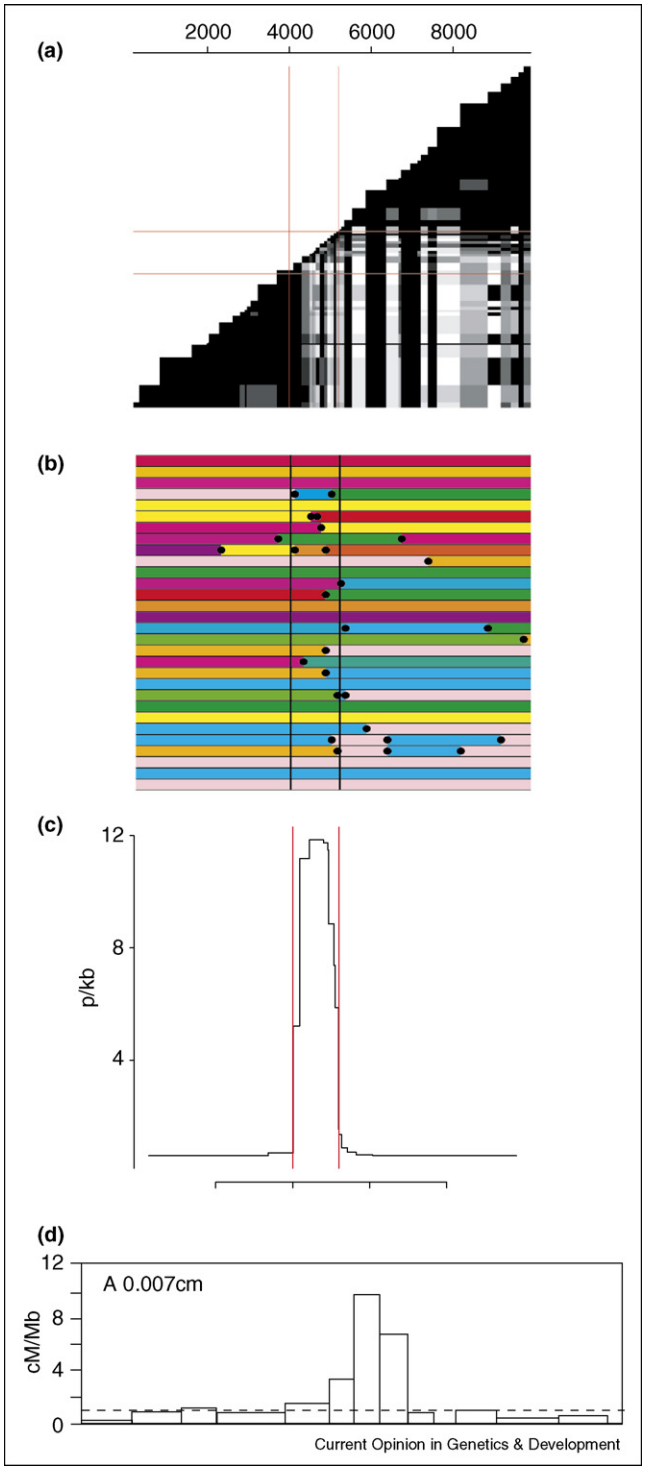
Instead, researchers have turned to approximate approaches that nevertheless capture much of the information in the data. These approaches include the 'pairwise composite likelihood' approach of Hudson [6] (see also [15,16,17•,18••]), the 'region-wise composite likelihood' approach of Fearnhead *et al.* [19,20•], and the 'product of approximate conditionals' (PAC) model of Li and Stephens [8]. The composite likelihood approaches compute likelihoods under well-established (if simplistic) population-genetics models and reduce computation to manageable proportions by separately considering small subsets of the full data (i.e. every pair of single nucleotide polymorphisms [SNPs] for the pairwise approach; non-overlapping sets of ~5–10 contiguous SNPs in the region-wise approach). This information is then combined across subsets by treating them as independent, which is blatantly untrue for the pair-wise approach, but appears to work surprisingly well in practice. By contrast, the PAC model treats all SNPs simultaneously, using a new model that attempts to capture the important qualitative features of established population genetics models while being more computationally tractable. Comparison of these methods on simulated data suggests that they have generally similar average accuracy [21•]. Simulations studies also suggest that inferences regarding relative recombination rates are somewhat robust to deviations from the overly simplistic models for population demography on which all these methods are based [8,21•,22••]. Figure 1 contains an illustration of how patterns of LD in a 10 kb region in *TAP2* (*Transporter 2*) [23] inform about the location and intensity of a recombination hotspot. Further information can be found in the original studies cited above and in the review article by Stumpf and McVean [24].

Insights into crossover

One of the most striking observations to come out of early large-scale data on human genetic variation was that patterns of LD appeared to be somewhat 'blocky'. That is, the genome could be thought of as consisting of ''blocks of variable length over which only a few common haplotypes are observed, punctuated by sites at which recombination could be inferred'' [25]. In one 216 kb segment of the major histocompatibility complex (MHC), Jeffreys *et al.* [26] showed by sperm-typing of eight UK males that crossover events were clustered into narrow, 1–2kb, regions known as 'hotspots'. They also showed, by graphical displays, that the locations of (at least some of) these hotspots coincided with a breakdown of LD, suggesting that patterns of LD might enable the identification of similar such hotspots in other genetic regions. This was an attractive idea because genome-wide data on human genetic variation were just becoming available, and so an analysis of patterns of LD could be performed on a genome-wide scale, whereas such large-scale sperm analyses are currently infeasible.

However, graphical displays of LD, although useful summaries, are limited in that they do not provide quantitative measures of either the intensity of any hotspots identified or, indeed, the confidence one should place in their existence. Using two different statistical approaches that overcome these limitations, and two different datasets, both McVean *et al.* [18••] and Crawford *et al.* [22••] (see also Fearnhead and Smith [20•]) found evidence for substantial variation in recombination rates across the genome. Similar analyses have since been used

Figure 1



Explanation and illustration of how patterns of LD in a 10 kb region in *TAP2* inform about the location and intensity of a recombination hotspot characterized by sperm-typing [23]. **(a)** Graphical display of a measure of LD, D' (see [49]), computed for every pair of SNPs in the population sample of haplotypes obtained by Jeffrey *et al*. [23]. Dark shading corresponds to stronger LD (higher D'). The preponderance of lighter shading in the lower right-hand corner, and of darker shading in the two triangular regions either side, is suggestive of a recombination hotspot near the center of the region. Red vertical and horizontal lines delineate the approximate location of a recombination hotspot found by sperm-typing (see **(d)**). The composite likelihood method of Hudson [6] can be thought of as computing a likelihood (for the population recombination rate, ρ) for each value of D' represented in this plot, and then combining information across these pairwise measures by multiplying together these likelihoods. **(b)** Graphical illustration of how the PAC model from Li and Stephens [8] identifies a recombination hotspot in *TAP2*. Unlike the composite likelihood method, the PAC model considers all SNPs simultaneously. Intuitively, the PAC model can be thought of as considering each observed haplotype in turn and attempting to represent them as a 'mosaic' of the previously-considered haplotypes. The figure illustrates such a set of mosaics obtained for 30 (of the 60) haplotypes from Jeffreys *et al*. [23]. Each row in the figure represents a haplotype, with colors showing the mosaic pieces used to create each haplotype in turn (starting from the bottom haplotype, up). Changes in the colors, emphasized by black circles, can be thought of as being caused by recombination events in the ancestry of the sample, and the frequency of black dots in any region therefore gives a quantitative indication of the recombination rate in that region. The preponderance of black dots within the vertical black lines suggests the existence of a recombination hotspot in the same position as suggested by sperm typing. **(c)** Estimates of the relative recombination rate in the region obtained using the PAC model (implemented in the software PHASE [22••,50]). The axis is scaled to have a minimum value of 1. The estimated crossover rate is elevated in the region identified to have an increased recombination rate in sperm (again delineated by vertical red lines). **(d)** Histogram showing distribution of crossover events identified by sperm-typing in a single male, taken with permission from Jeffreys *et al*. [23].

to identify over 25 000 hotspots in three populations [27••]. All the above studies, and others, have also made some attempt to quantify the extent of the variation in crossover rate across the genome (e.g. in terms of the frequency of recombination hotspots, typically estimated as 1 hotspot per 30–50 kb). These studies have also estimated the proportion of crossover events occurring in a given fraction of the sequence; for example, Myers *et al*. [27••] estimated that 80% of recombination occurs in 10–20% of the sequence. However, although such estimates can be helpful for simulating realistic population data, it would be unwise to place undue emphasis on them in terms of the actual underlying recombination process, both because population data provide estimates based on averages across individuals (and therefore will tend to underestimate the severity of the clustering if there are differences among individuals) and because the inference is likely to be sensitive to underlying assumptions, such as what constitutes a hotspot, or what prior distributions or smoothing penalty should be used for changes in recombination rate. It is worth noting that earlier studies, including those by McVean *et al*. [18••] and Crawford *et al*. [22••], were conducted at a time when the ubiquity of human hotspots was less clear and, therefore, they deliberately made assumptions that were designed to seldom detect hotspots where they did not exist. As a result, these studies are likely to have underestimated the amount of fine-scale variation.

Taken together, the sperm-typing data and population genetic variation suggest strongly that, throughout the

human genome, the locations of crossover events tend to be highly clustered. Further, the identification of a large set of hotspots, even if inevitably somewhat arbitrarily defined, enabled Myers *et al.* [27••] to identify several interesting connections between particular sequence contexts and hotspots. In particular, the sequence motifs CCTCCCT and CCCCACCCC and the two retrovirus-like retrotransposons THE1A and THE1B are all substantially enriched in hotspots. Some types of repeat (e.g. CT-rich and GA-rich repeats) were over-represented in hotspots, whereas others (e.g. GC-rich repeats) were under-represented. Although in most cases these under- and over-represented features might reflect correlations rather than causal relationships, Myers *et al.* [27••] noted that variation across males in the intensity of the *DNA2* hotspot [28] coincides with the presence or absence of a particular CCTCCCT motif within the hotspot — owing to a SNP in this region, some haplotypes carry the motif, whereas others carry CCCCCCT. This strongly suggests a direct role for this motif in regulating recombination. Subsequently, the motif CCCCACCCCC was found to affect the variability of crossover activity across males in the *NID1* hotspot [29•], providing further evidence for at least some regulation of recombination by certain motifs [30].

Aside from such sequence features that appear to be specifically associated with narrow hotspots, other sequence features have also been found to be more generally associated with rates of crossover. For example, just as pedigree studies have found sequence G + C content to be positively correlated with estimated crossover rates on the Megabase scale [10], so population data suggest that this correlation is present at considerably finer scales. The study by Fearnhead and Smith [20•] found G + C content to be slightly elevated in estimated hotspots (46.6% inside versus 44.2% outside). Crawford *et al.* [22••] also found SNP density, or diversity, to be positively correlated with crossover rates, although not specifically with narrow 'hotspots'. By contrast, Fearnhead and Smith [20•] found SNP density to be elevated by a factor of ~1.5 in hotspots compared with outside hotspots, and certain hotspots, including *TAP2*, appear to contain a higher density of SNPs than do surrounding regions [23]. These observations are consistent with recombination being mutagenic, evidence for which has been observed in experiments with yeast [31].

Studies based on pedigrees [10] have found a positive correlation between gene density and recombination rate. By contrast, McVean *et al.* [18••] found that average estimated values for ρ were smaller inside genes than outside. However, it is difficult to know how much of this decrease in estimates might be due to selection acting preferentially near genes, because certain types of selection would be expected to reduce LD and hence decrease estimates of ρ. By contrast, it is harder to imagine how selection might account for the more complex patterns found by Myers *et al.* [27••], who showed that estimated recombination rates are, on average, lower within genes, but that as one moves away from a gene the average estimates increase symmetrically in either direction for about 30 kb before decreasing again. It has also been suggested [32,33] that certain types of selection could give rise to patterns of LD that look like 'hotspots' for crossover, although given the generally good correspondence between sperm-typing and LD-based results it seems unlikely that selection accounts for the majority of hotspots thus-far identified from population data.

Hotspot evolution

Several studies have now been performed comparing patterns of LD in humans and other primates — most notably comparisons of humans with chimpanzees. All have found that the different species appear to share few crossover hotspots [34,35•,36••,37••]. Thus, population data suggest that hotspots can come and go on relatively short evolutionary time-scales. This raises the question of whether there might even be substantial differences in average fine-scale crossover rates among different human ethnic groups, and indeed there does seem to be some evidence for this in population data from different continents [20•,22••]. However, population data are not ideally placed to answer this question, because the fact that human ethnic groups share a substantial amount of their genetic history means that differences — should they exist — among the average recombination rates of current continental groups might not be fully reflected in patterns of LD. Thus, perhaps a stronger indication that average rates are likely to differ across groups is the fact that sperm-typing data suggest sequence polymorphisms can affect hotspot activity across individuals [28,29•], and so one would expect average recombination rates in some regions to differ across ethnic groups owing to frequency differences in such polymorphisms.

The speed of evolution of hotspots can also be studied by comparing estimates of crossover from sperm-typing experiments with estimates based on LD data. Because LD data provide estimates of the *average* crossover rate over many thousands of generations, whereas sperm data provide estimates based on extant individuals, rapid evolution of hotspot locations would be expected to create disparities between the two estimates. Thus, although in some regions there appear to be large differences between estimates based on the two types of data [38•,39•], the generally good concordance between sperm-typing and LD data perhaps suggests that hotspots evolve at an intermediate rate over tens or hundreds of thousands of years.

Insights into gene conversion

Although the vast majority of studies relating recombination to patterns of LD have focussed on crossover (i.e. recombination that results in exchange of flanking

markers), a few have also studied gene conversion (i.e. without exchange of flanking markers). Here, we focus on the simplest case of allelic gene conversion between homologous regions.

The main effect of gene conversion on population data is to decrease LD at small scales (e.g. over a few hundred base pairs), leaving LD at larger scales relatively unaffected. Given that human population data generally appear to have less LD at short scales than expected on the basis of observed patterns of crossover alone, gene conversion might play a non-negligible role in shaping LD [1]. However, the effect of gene conversion on LD is substantially more subtle than that of crossover, and so it might be difficult to use LD to learn about variation in rates of gene conversion on very fine scales. For example, one might like to know whether hotspots for crossover are also hotspots for gene conversion. However, whereas a hotspot for crossover affects patterns of LD among all markers either side of the hotspot, a hotspot for gene conversion affects only markers *within* the hotspot, a hotspot for gene conversion affects only markers within the hotspot, and because many hotspots can contain few markers it seems that detecting hotspots for gene conversion from LD data will be difficult. Characterizing rates of gene conversion by sperm-typing is also difficult but has been done for a few regions known to harbor crossover hotspots [29•,40••,41•], and most appear to be hotspots for gene conversion as well as crossover, consistent with the hypothesis that crossover hotspots are due to increased rates of double-strand break formation.

Therefore, in contrast to studies of crossover, most studies of gene conversion from population data have focussed on estimating *average* rates across the genome, rather than on estimating fine-scale variation. Specifically, these studies have tended to focus on estimating the average ratio of gene conversion to crossover, $f = \gamma/\rho$, where γ, the analogue of ρ for gene conversion, is equal to $2N_e g$, and g is the probability of gene conversion occurring in a single meiosis. Among other things, f has the happy property that it does not depend on the unknown scaling constant N_e, which appears in both γ and ρ. An early study [16] estimated a genome-wide average value for f of ~4–25. Under the double strand-break model for recombination [42], in which gene conversions and crossovers are alternative outcomes of Holliday junction resolution, this would translate into Holliday junctions being resolved as gene conversions some 4–25 times more often than as crossovers. Another study estimated a similar value of f in the range of 3–10 [43]. A later study [17•], based on more data, had substantially smaller estimates: $f \approx 0.3$ for Europeans and $f \approx 1.0$ for African Americans. These values were similar to those of another study that estimated $f \approx 1.6$ in a worldwide sample [44•].

The large differences between estimates from some studies are partly a reflection of the difficulty of the statistical inference problem, and particularly in the relative lack of 'informativeness' of the data for estimating rates of gene conversion. It is also worth noting that there might actually be variation in f among the different sets of regions studied. In support of this possibility, sperm-typing experiments suggest that f can vary across regions; f was estimated to be ~4–15 in two hotspots of the MHC region (hotspots *DNA3* and *DMB2* [40••]), ~0.3 in the hotspots *SHOX* and *NID1* [29•,40••], and <0.1 in the β-globin gene region [41•].

When comparing estimates of f from sperm-typing and LD data, it is also worth noting that all the LD studies mentioned above estimated f assuming that the average length of sequence affected by each gene conversion event (i.e. the average *tract length*, usually denoted t) is 500 bp. Although very limited data are available for estimating t, the most recent data from sperm [29•,40••] suggest that most gene conversion tracts in humans are short (<1 kb), and that average tract lengths could be considerably shorter than 500 bp, perhaps closer to 100 bp. Estimated values for f from LD data scale, approximately, with the inverse of t. Therefore, had the above studies assumed $t = 100$, then their estimates of f would be expected to be roughly five times bigger (e.g. $f = 25$–125 in the study of Frisse *et al.* [16] and $f = 1.5$ (Europeans) and 5.0 (Americans) in the study of Ptak *et al.* [17•]).

Population data have also been used to study gene conversion in organisms other than humans. Indeed, some of the first insights into gene conversion from population data came from Langley *et al.* [45], who explored two regions of the X chromosome in *Drosophila melanogaster* that had low estimated rates of crossover but similar levels of LD breakdown compared with areas that had normal rates of crossover. The authors suggest that this might be due to these regions having a strong bias in favour of resolving Holliday junctions as gene conversion events without crossover [46]. More recently, Plagnol *et al.* [47•] examined rates of crossing over and gene conversion in *Arabidopsis thaliana* using a genome-wide survey consisting of 1347 fragments (500–600 bp each) sequenced in 96 accessions. Using LD techniques, they estimated genome-wide average f of ~1. However, they also found evidence for variation in f across the genome, and, in particular, they found that estimated crossover and gene conversion rates were not highly correlated across regions. In summary, LD analyses have provided substantially less information about gene conversion in humans than crossover. Although one might hope that improvements to existing statistical methods might help this situation, this may be over-optimistic [48].

Conclusions and perspectives

In conclusion, recent analyses of patterns of genetic variation in population samples have provided substantial

insights into recombination in humans: throughout the human genome, recombination events tend to cluster into a relatively small proportion of the total sequence; certain sequence motifs appear to be directly related to recombination in some regions; and, over fine scales, recombination rates in humans appear to differ from those in chimpanzee. In the near future, we can expect to see further progress, particularly with regard to more comprehensive comparisons of recombination rates in humans and chimps, and perhaps also other primates, which should help further elucidate the factors that affect recombination on both fine and broad scales.

Acknowledgements

We thank A Jeffreys for providing the sequencing data for the analyses in Figure 1. The authors were supported by Genome Training Grant HG00035-09/10/11 for GH, and National Institutes of Health Grant 1RO1HG/LM02585-01 for MS.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data**. *Am J Hum Genet* 2001, **69**:1-14.
2. Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD, Kazazian HH: **Nonuniform recombination within the human β-globin gene cluster**. *Am J Hum Genet* 1984, **36**:1239-1258.
3. Hudson RR: **Estimating the recombination parameter of a finite population model without selection**. *Genet Res* 1987, **50**:245-250.
4. The International HapMap Consortium: **A haplotype map of the human genome**. *Nature* 2005, **437**:1299-1320.
5. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: **Whole-genome patterns of common DNA variation in three human populations**. *Science* 2005, **307**:1072-1079.
6. Hudson RR: **Two-locus sampling distributions and their application**. *Genetics* 2001, **159**:1805-1817.
7. Fearnhead P, Donnelly P: **Estimating recombination rates from population genetic data**. *Genetics* 2001, **159**:1299-1318.
8. Li N, Stephens M: **Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data**. *Genetics* 2003, **165**:2213-2233.
9. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL: **Comprehensive human genetic map: individual and sex-specific variation in recombination**. *Am J Hum Genet* 1998, **63**:861-869.
10. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G *et al.*: **A high-resolution recombination map of the human genome**. *Nat Genet* 2002, **31**:241-247.
11. Kingman JFC: **The coalescent**. *Stochastic Process Appl* 1982, **13**:235-248.
12. Hudson RR: **Properties of a neutral allele model with intragenic recombination**. *Theor Popul Biol* 1983, **23**:183-201.
13. Griffiths RC, Marjoram P: **Ancestral inference from samples of DNA sequences with recombination**. *J Comput Biol* 1996, **3**:479-502.
14. Kuhner MK, Yamato J, Felsenstein J: **Maximum likelihood estimation of recombination rates from population data**. *Genetics* 2000, **156**:1393-1401.
15. McVean G, Awadalla P, Fearnhead P: **A coalescent-based method for detecting and estimating recombination from gene sequences**. *Genetics* 2002, **160**:1231-1241.
16. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A: **Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels**. *Am J Hum Genet* 2001, **69**:831-843.
17. • Ptak SE, Voelpel K, Przeworski M: **Insights into recombination from patterns of linkage disequilibrium in humans**. *Genetics* 2004, **167**:387-397.

The authors adapt Hudson's [6] composite-likelihood approach to jointly estimate crossover and gene conversion in the SeattleSNPs dataset, assuming crossover rates to be constant within genes, and f to be constant across the genome. Using a profile-likelihood approach that simultaneously maximizes ρ per gene, and f across genes, they estimate genome-wide average f to be $\sim$1.0 in the African Americans (using 84 genes) and $\sim$0.3 in the Centre d'Etude du Polymorphisme Humain (CEPH) population (77 genes), assuming a tract length of 500 bp.

18. •• McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome**. *Science* 2004, **304**:581-584.

The authors update Hudson's [6] composite-likelihood method to estimate variable crossover rates along a sequence — in this case of kilobase scale. They examine rates in several different datasets, including sampled chromosomes from European and African American individuals (average marker spacing 2.3 kb), and sampled chromosomes of CEPH and Yoruban descent (average marker spacing 5.8 kb). Also, finding hotspots to be narrow, they estimate that about 50% of all crossover events take place in $<$10% of the sequence. They find crossover rates to be diminished in genic regions and note that although G + C content and kilobase-scale crossover rates appear to be correlated, this correlation disappears when looking at finer scales such as hotspots.

19. Fearnhead P, Harding RM, Schneider JA, Myers S, Donnelly P: **Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots**. *Genetics* 2004, **167**:2067-2081.
20. • Fearnhead P, Smith NG: **A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes**. *Am J Hum Genet* 2005, **77**:781-794.

The authors extend the importance-sampling, simulation approach of Fearnhead *et al*. [19] to locate probable crossover hotspots using LD data, considering overlapping sub-regions (assuming them independent) of approximately five SNPs each. Analyzing 89 genes from SeattleSNPs in African Americans and Europeans (both jointly and separately), they find hotspot frequencies of about 1 per 30-40 kb, as well as some evidence that several hotspots (perhaps as many as 30%) do not overlap between the two populations. They also find evidence for increased SNP density and G + C content in hotspots relative to outside them.

21. • Smith NG, Fearnhead P: **A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness**. *Genetics* 2005, **171**:2051-2062.

The authors use coalescent-based simulations with 50 haplotypes, $\theta = 1.0$/kb, and sequence lengths of 2, 10, 25 and 100 kb, to test the crossover estimation methods of Hudson [6], Fearnhead and Donnelly [7], and Li and Stephens [8]. They find that all methods behave similarly, and that crossover estimates in all three are biased upwards in the presence of gene conversion (simulated in rates of 1, 5 or 10/kb). Some extreme SNP ascertainment schemes, when left unaccounted for, can also be a problem.

22. •• Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M: **Evidence for substantial fine-scale variation in recombination rates across the human genome**. *Nat Genet* 2004, **36**:700-706.

The authors find evidence for crossover hotspots in 35 of 74 genes from the SeattleSNPs dataset. Crossover rates on a kilobase scale appear to be correlated with sequence features such as G + C content and nucleotide diversity, although this correlation does not appear to extend to hotspots.

23. Jeffreys AJ, Ritchie A, Neumann R: **High resolution analysis of haplotype diversity and meiotic crossover in the human *TAP2* recombination hotspot**. *Hum Mol Genet* 2000, **9**:725-733.

24. Stumpf MPH, McVean GAT: **Estimating recombination rates from population genetic data**. *Nat Rev Genet* 2003, **4**:959-968.

25. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M *et al.*: **The structure of haplotype blocks in the human genome**. *Science* 2002, **296**:2225-2229.

26. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex**. *Nat Genet* 2001, **29**:217-222.

27. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P:
•• **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**:321-324.
Using the methods outlined by McVean *et al*. [18••], the authors examine crossover hotspots across the genome. They find evidence of >25 000 hotspots, using data from 24 CEPH, 23 African Americans and 24 Han Chinese from Los Angeles, genotyped across 1.5 million common SNPs. They find that hotspots correlate with particular sequence motifs but find an overall lack of correlation between crossover rates at fine (i.e. kb) levels and sequence features, compared with correlations between crossover rates at large (i.e. Mb) levels and sequence features. Chimpanzee and human hotspot locations and intensities appear to be uncorrelated [36••], whereas the authors' population-based crossover rates (on Mb scales) and pedigree-based estimates [10] appear to be strongly correlated. This leads the authors to speculate that crossover rates are constrained over large (i.e. Mb) regions but are perhaps rapidly evolving at finer scales.

28. Jeffreys AJ, Neumann R: **Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot**. *Nat Genet* 2002, **31**:267-271.

29. Jeffreys AJ, Neumann R: **Factors influencing recombination**
• **frequency and distribution in a human meiotic crossover hotspot**. *Hum Mol Genet* 2005, **14**:2277-2287.
The authors examine crossover rates in the hotspot *NID1* of chromosome 1 (near the minisatellite *MS32* hotspot) in seven men (with a mean peak of crossover rate = 70 cM/Mb). They find evidence of crossover asymmetry (the 'crossover-suppressing' allele perhaps reduces crossover by ≥2.8-fold) and transmission distortion; at an allele near the center of the hotspot, the crossover-suppressing allele was over-transmitted by about 3:1. They also looked at rates of gene conversion for one man; rates were only a quarter of the observed crossovers for this man, and transmission distortion rates were similar to those of crossovers.

30. Myers S, Spencer CC, Auton A, Bottolo L, Freeman C, Donnelly P, McVean G: **The distribution and causes of meiotic recombination in the human genome**. *Biochem Soc Trans* 2006, **34**:526-530.

31. Rattray AJ, Shafer BK, McGill CB, Strathern JN: **The roles of *rev3* and *rad57* in double-strand-break-repair-induced mutagenesis of *Saccharomyces cerevisiae***. *Genetics* 2002, **162**:1063-1077.

32. Vander Molen J, Frisse LM, Fullerton SM, Qian Y, del Bosque-Plata L, Hudson RR, Di Rienzo A: **Population genetics of *capn10* and *gpr35*: implications for the evolution of type 2 diabetes variants**. *Am J Hum Genet* 2005, **76**:548-560.

33. Reed FA, Tishkoff SA: **Positive selection can create false hotspots of recombination**. *Genetics* 2006, **172**:2011-2014.

34. Wall JD, Frisse LA, Hudson RR, Di Rienzo A: **Comparative linkage-disequilibrium analysis of the β-globin hotspot in primates**. *Am J Hum Genet* 2003, **73**:1330-1340.

35. Ptak SE, Roeder AD, Stephens M, Gilad Y, Paabo S, Przeworski M:
• **Absence of the TAP2 human recombination hotspot in chimpanzees**. *PLoS Biol* 2004, **2**:849-855.
The authors compare LD patterns between 30 UK humans and 24 resequenced western chimpanzees in the *TAP2* region. In particular, they use the methods of Li and Stephens [8] to estimate ρ across the fine-scale region. They find substantial evidence for a hotspot in the human data, with crossover activity at least 10 times greater than rates outside the hotspot, whereas there is evidence against a hotspot in the chimpanzee data. They also highlight an example of positive association between rates of crossover and mutation, noting that diversity levels are significantly higher than average for the human data in the hotspot region, although this does not appear to be the case in chimpanzees.

36. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG,
•• Przeworski M, Frazer KA, Paabo S: **Fine-scale recombination patterns differ between chimpanzees and humans**. *Nat Genet* 2005, **37**:429-434.
The authors compare LD patterns between the DNA of 23 humans and 8 central chimpanzees, in 14 Mb of sequence from two disjoint regions of chromosome 21 (and its chimp homologue), to determine what proportion of hotspots appear to be shared between the two species. Using the methods of Li and Stephens [8], they find that only 3 of the 39 hotspots inferred in the chimpanzees appeared in humans; lack of power to find chimpanzee hotpots makes it more difficult to infer how many characterized human hotspots appear in the chimps. Through simulation, they find that this observation is consistent with the hypothesis that hotspots are independently distributed along the genome of the two species. This suggests relatively rapid hotspot evolution. The authors also find that background crossover rates — rates outside of hotspots — for 50-kb windows are moderately correlated between the species, suggesting that perhaps recombination rates are conserved at large scales but not at fine scales.

37. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ,
•• Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P, Altshuler D: **Comparison of fine-scale recombination rates in humans and chimpanzees**. *Science* 2005, **308**:107-111.
The authors inferred rates of crossover from resequencing polymorphism data at orthologous loci in western chimpanzees and in CEPH and Yoruban HapMap ENCODE regions. They found statistical support for 18 hotspots in humans and 3 in chimps ($\rho < 0.01$), but with a striking lack of overlap: a crossover hotspot in one species was typically found to have crossover rate 10 to 60 times larger than that of the same region of the other species, which was typically not 'hot' at all relative to its background crossover rates.

38. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P:
• **Human recombination hot spots hidden in regions of strong marker association**. *Nat Genet* 2005, **37**:601-606.
The authors explore eight hotspots in a 206 region of chromosome 1, which contains the gene *NID* and the microsatellite MS32. They carry out sperm analysis on 2-8 individuals per hotspot, and employ the LD methods of Fearnhead *et al*. [19], Li and Stephens [8], and McVean *et al*. [18••], using 80 unrelated UK individuals genotyped at 200 SNPs. The hotspots were found using the D′ association statistic in the 80 UK individuals. Using the method of McVean *et al*. [18••], the authors estimate the historical crossover rate per meiosis, *r*, in the hotspots (assuming $N_e = 10000$), which they compare against that found by sperm analysis. In three of the hotspots, the estimated values from LD analysis were significantly different (e.g. *r* for *NID2a* was estimated as 55 cM/Mb from LD analysis, compared with 10 cM/Mb from sperm analysis).

39. Tiemann-Boege I, Calabrese P, Cochran DM, Sokol R, Arnheim N:
• **High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing**. *PLoS Genet* 2006, **2**:e70.
The authors examine crossover activity across a 103 kb segment of chromosome 21, using the sperm of 240 individuals. They find two hotspots in this region, each of width of 1-2 kb, and estimate that 71% of the total crossover activity occurs in ~12% of the sequence. They find that the patterns of estimated crossover activity from sperm analysis correspond well to those of various LD methods. Furthermore, studying one of the hotspots in more fine-scale detail in three men showed that the majority of the crossover activity for two of the men mapped to one ~2 kb interval, whereas in the third man the majority of activity mapped to an apparently distinct region ~2 kb away. The authors speculate that when one hotspot dies away, a new hotspot arises relatively nearby.

40. Jeffreys AJ, May CA: **Intense and highly localized gene**
•• **conversion activity in human meiotic crossover hot spots**. *Nat Genet* 2004, **36**:151-156.
The authors examine the *DNA3* crossover hotspot of the human MHC by sperm analysis of two men, finding *DNA3* to be a hotspot for gene conversion as well; they speculate that *DNA3* is a hotspot for double-strand-breaks overall. They find that gene conversions occur 4-15 times more often than crossovers in the two men studied. They estimate tract length to be, on average, 55-290 bp per gene conversion event, and note that tracts for conversion accompanying crossover appear to be somewhat larger (~460 bp).

41. Holloway K, Lawson VE, Jeffreys AJ: **Allelic recombination and**
• **de novo deletions in sperm in the human β-globin gene region**. *Hum Mol Genet* 2006, **17**:1099-1111.
The authors use sperm analysis to characterize the crossover activity of 13.5 kb of sequence in the β-globin gene region. Using two donors, they find an intense, narrow crossover hotspot, ~1.2 kb wide, with crossover activity of ~150–260 cM/Mb. Examining the hotspot of one of these men for gene conversion alone, however, detected no gene conversions. This finding does not appear to be attributable to lack of power or marker density. The authors conclude that the relative rate of gene conversion to crossover in

this region must be $<1/12$. Furthermore, they do not find any evidence of association between deletion breakpoints and high crossover activity.

42. Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW: **The double-strand-break repair model for recombination**. *Cell* 1983, **33**:25-35.

43. Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, Winchester E, Lander ES, Kruglyak L: **Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion**. *Am J Hum Genet* 2001, **69**:582-589.

44. • Padhukasahasram B, Marjoram P, Nordborg M: **Estimating the rate of gene conversion on human chromosome 21**. *Am J Hum Genet* 2004, **75**:386-397.

Using simple multilocus summary statistics, the authors estimate rates of gene conversion and crossover in humans. The rates are based on patterns of LD, using the DNA from a worldwide sample of 20 chromosomes representing a 28 Mb region of chromosome 21. They estimate *f* to be $\sim$1.6, assuming a tract length of 500 bp. Estimating rates over overlapping 2 Mb windows, they find rates of crossover and gene conversion to be uncorrelated.

45. Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM: **Linkage disequilibria and the site frequency spectra in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster* X chromosome**. *Genetics* 2000, **156**:1837-1852.

46. Andolfatto P, Wall JD: **Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster***. *Genetics* 2003, **165**:1289-1305.

47. • Plagnol V, Padhukasahasram B, Wall JD, Marjoram P, Nordborg M: **Relative influences of crossing over and gene conversion on the pattern of linkage disequilibrium in *Arabidopsis thaliana***. *Genetics* 2006, **172**:2441-2448.

The authors examine rates of gene conversion in *Arabidopsis thaliana*, using a genome-wide survey consisting of 1347 fragments (500–600 bp each) sequenced in 96 accessions. To do so, they perform coalescent simulations with growth parameters that match the observed allele frequency spectrum, and a crossover rate estimated using the method of Hudson [6]. They then find gene conversion parameters that match the observed LD patterns in their data. They find, on average, that $f = 1$, or that the crossover rate and gene conversion rate are both $\sim$0.3/kb, assuming a tract length of 100 bp. They also estimate rates of crossover and gene conversion across 14 genome-wide 2-Mb windows — each window having, on average, 18 fragments — and find that variable rates of *f*, crossover, and gene conversion across these windows fit the data significantly better than either fixing *f* across regions or taking an identical crossover and gene conversion rate in all regions. Furthermore, they find estimated rates of crossover and gene conversion across 43 such 2-Mb windows to be uncorrelated.

48. Wall JD: **Estimating recombination rates using three-site likelihoods**. *Genetics* 2004, **167**:1461-1473.

49. Hudson RR: **Linkage disequilibrium and recombination**. In *Handbook of Statistical Genetics.* Edited by Balding DJ, Bishop M, Cannings C. Wiley; 2001:309-323.

50. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data**. *Am J Hum Genet* 2001, **68**:978-989.