The distribution and causes of meiotic recombination in the human genome

S. Myers*†, C.C.A. Spencer*, A. Auton*, L. Bottolo‡, C. Freeman*, P. Donnelly* and G. McVean*†

*Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, U.K., †The Broad Institute of MIT and Harvard, 1 Kendall Square, Building 300, Cambridge, MA 02139-1561, U.S.A., and ‡Department of Epidemiology and Public Health, Imperial College, St. Mary’s Campus, Norfolk Place, London W2 1PG, U.K.

Abstract

Using the statistical analysis of genetic variation, we have developed a high-resolution genetic map of recombination hotspots and recombination rate variation across the human genome. This map, which has a resolution several orders of magnitude greater than previous studies, identifies over 25,000 recombination hotspots and gives new insights into the distribution and determination of recombination. Wavelet-based analysis demonstrates scale-specific influences of base composition, coding context and DNA repeats on recombination rates, though, in contrast with other species, no association with DNase I hypersensitivity. We have also identified specific DNA motifs that are strongly associated with recombination hotspots and whose activity is influenced by local context. Comparative analysis of recombination rates in humans and chimpanzees demonstrates very high rates of evolution of the fine-scale structure of the recombination landscape. In the light of these observations, we suggest possible resolutions of the hotspot paradox.

A high-resolution genetic map of the human genome

Conventionally, genetic maps have been estimated from the co-transmission of alleles in pedigrees [1,2]. In humans, the best such genetic map has resolution on the centimorgan (cM) scale, which means that recombination rate variation below the megabase scale cannot be documented. However, in recent years, it has become clear that most of the recombination in humans is concentrated in short regions of 1–2 kb known as recombination hotspots [3–7]. Such hotspots cannot be detected from pedigree-based studies; however, they can be identified with high accuracy from patterns of genetic variation [8–11]. Using the statistical analysis of genome-wide surveys of genetic variation [12,13], we have generated a high-resolution genetic map detailing over 25,000 recombination hotspots across the human genome [8]. The methods used have been extensively validated by comparison with sperm-based studies over the fine scale and pedigree-based studies over the broad scales [8,11,14]. Figure 1 illustrates the nature of recombination rate variation and its relationship to genomic features over a 2.5 Mb region of chromosome 21 where estimates of cross-over rate from sperm [15] show strong correspondence with estimates based on genetic variation.

A high-resolution genetic map allows us to characterize the nature and extent of recombination rate heterogeneity. After correcting for the power of statistical methods to identify recombination hotspots, we estimate there to be approx. 30–50,000 recombination hotspots in the human genome, corresponding to one for every 50–100 kb on average [14]. Of those hotspots that were identified by the statistical methods (and which are likely to be biased towards more active hotspots), the average amount of recombination is 0.075 cM (or one cross-over event per 1300 meioses), whereas the most extreme hotspot has a map length of 0.9 cM (or one cross-over event per 110 meioses) [16]. However, both the density and intensity of hotspots vary considerably across the genome, leading to the large-scale (megabase and greater) variation in recombination rate documented by pedigree-based genetic maps [1,2,17]; for example, the general increase in recombination rate in the sub-telomeric regions and the centromeric suppression.

The nature of recombination rate variation can be quantified in two ways. First, we can measure how concentrated recombination is by estimating the minimum amount of DNA sequence that can account for a given fraction of cross-over activity. Analysis of 5 Mb of sequence across the genome with very high SNP (single nucleotide polymorphism) densities suggests that 80% of all recombination takes place in approx. 15% of the sequence [13]; see also [8,11]. While demonstrating the dominance of recombination hotspots, this result also shows that while recombination is typically low outside hotspots, it is not entirely suppressed. Indeed, we find no region greater than about 200 kb in which recombination is apparently absent (beyond the Y chromosome) [8]. The second approach to describing variation in recombination rate is to measure how much of the variation in rate is accounted for by variation over different physical scales. Using wavelet analysis (C.C.A. Spencer, S. Myers, P. Deloukas, S. Hunt, J.C. Mullikin, D. Bentley, P. Donnelly and G. McVean, unpublished work), we find that the landscape of recombination rate variation is dominated by...
Figure 1 | Recombination rate variation over a 2.5 Mb region of chromosome 21

The estimated sex-averaged recombination rate is shown alongside the locations of statistically significant recombination hotspots (black bars), the motif most strongly associated with hotspots (grey triangles) [8], known genes on the plus and minus strands, base composition (graded such that darker shades represent higher GC content) and CpG islands (black bars). The upper panel shows cross-over rates estimated in three males [15]; each bar represents a region over which average rates were estimated (darker indicates higher average recombination rate; range 2.1–168 cM · Mb⁻¹). Where short regions of high cross-over activity have been identified, these correspond closely to recombination hotspots identified from genetic variation.

fine-scales of 2–50 kb. However, the distribution is bimodal, with variation over scales of 2–30 Mb generating macrochromosomal variation in recombination rate.

Although our estimates of recombination are, by necessity, sex-averaged, it is possible to learn about differences in the landscape of recombination between males and females in two ways. First, the X chromosome only undergoes recombination in females. Secondly, we can identify regions from the sex-specific genetic maps estimated from pedigree data where there is a much higher average female rate than males (and vice versa). Analysis of such genomic regions indicates that males and females have qualitatively similar fine-scale recombination landscapes, in the sense that both have hotspots at similar densities and intensities [8]. We can, for example, rule out the possibility that recombination hotspots (which have only been characterized experimentally in sperm studies) are restricted to males.

Scale-specific correlates of recombination

What genomic features are associated with heterogeneity in the recombination rate? As shown above, recombination rates vary over multiple scales, from the location of specific hotspots at the nucleotide scale, to variation over the scale of chromosome arms. Therefore it is possible that variation at different scales is associated with different genomic features.

To assess the scale-specific correlates of recombination, we can use wavelet transformations (C.C.A. Spencer, S. Myers, P. Deloukas, S. Hunt, J.C. Mullikin, D. Bentley, P. Donnelly and G. McVean, unpublished work; [19]) to decompose the original signals (recombination rate and other genomic properties) into a series of coefficients that measure heterogeneity at each scale and location. Because wavelet transformations typically result in a series of coefficients that have low autocorrelation and follow a Normal distribution (at least approximately), it is possible to use linear model analysis to identify significant correlations between factors at different scales.

Figure 2 shows an analysis of recombination rates along chromosome 1 that includes a variety of genomic features, namely gene location, base composition, repeat density, CpG islands, the locations of segmental duplications and conservation, as well as epigenetic features such as replication timing and DNase I hypersensitivity and the locations of a motif identified in our recent research that is strongly associated with hotspots [8]; see also below. Only two factors stand out as being strong and consistent correlators of recombination rate: GC content (typically over scales of 8–512 kb) and the location of the motif most strongly associated with hotspots (at scales of 8 kb and less). Both repeat DNA and exonic sequences are negatively associated with recombination over moderate to broad scales; indeed recombination rates seem to reach a peak at distances of 10–50 kb from genes [8]. In contrast with studies in Saccharomyces cerevisiae [20], there is no association between recombination rate and DNase I hypersensitivity; rather this is strongly associated with the location of CpG islands (results not shown). However, it should be noted that DNase I hypersensitivity has been measured in CD⁺ T-cells [21], and the conservation of such markings across cell type is not clear.

Finally, it should be noted that while strong correlations with recombination rate can be identified, overall such factors have low predictive power. Indeed, combined they explain less than 10% of the variance in rate.
Figure 2 | Wavelet-based analysis of genomic features potentially associated with recombination

Tables showing $-\log_{10} P$ values (two-sided t test) for linear model analyses of wavelet coefficients (generated using the Haar function) for two 65536 kb regions on the short and long arms respectively (data were obtained from http://www.genome.ucsc.edu; hg17). Cells are shaded according to significance ($P < 10^{-1}$, $10^{-2}$ etc.), factors with a significant ($P < 0.01$) negative effect on recombination rate are underlined. Weights are used in the linear model analysis to correct for missing data and a zero intercept is fitted at each level. DNase HS, DNase hypersensitivity; Seg. Duplins, segmental duplications.

DNA sequence motifs associated with recombination hotspots

Sequence motifs promoting crossover activity have been previously found in several prokaryotes, for example the Chi sequence of Escherichia coli [22,23] and at a single eukaryotic hotspot, M26 in Schizosaccharomyces pombe [24]. Recently, it was shown that a sequence closely related to the M26 motif stimulates DSBs (double-strand breaks) more generally within the Schizosaccharomyces pombe genome [25]. No strong evidence of such sequence motifs has previously been found in higher eukaryotes, despite the intensive study of hotspots in both humans and mice. The localization of 25000 hotspots across the human genome offers an unprecedented opportunity for understanding their sequence basis [8,13]. Exhaustive testing of transposable elements for enrichment or depletion in these hotspots revealed a surprising, strong influence of local sequence features on local recombination rate. For example, the long terminal repeats (but not the accompanying internal sequences) of THE1A and THE1B retrotransposons are 2-fold enriched in hotspots, whereas certain L1 elements are more than 2-fold depleted.

The enrichment of the THE1A/B elements suggests a causal relationship – we therefore asked whether motifs were enriched in those elements occurring within narrowly defined hotspots (5 kb or less), relative to the genomic THE1A/B population. Such testing automatically controls for local background influences such as GC content, while the small size ($\sim 350$ bases) of these elements potentially improves hotspot localization considerably. We found a striking effect – the seven-base-pair oligonucleotide CCTCCCT was 5 times enriched in the hotspot THE1Bs ($P < 10^{-30}$), an effect matched in the related THE1A elements. A second motif, CCACGTGG, located two base-pairs downstream, showed independent 2-fold enrichment in hotspot THE1Bs.

Strong evidence suggests that this motif directly acts to promote recombination more generally throughout the genome. First, after masking all repeat sequence, CCTCCCT was the most hotspot-enriched of all possible 16384 7-bp oligonucleotides (comparing $\sim 9000$ autosomal hotspots defined to within 5 kb to matched cold regions of the genome). Secondly, compelling evidence comes from sperm studies. Jeffreys and Neumann [26] identified a single nucleotide polymorphism within the DNA2 hotspot where the derived allele, C, is overtransmitted in sperm, and postulated to strongly reduce hotspot activity compared with the ancestral allele, T. Examination of the flanking sequence surrounding this SNP reveals an ancestral occurrence of CCTCCCT, disrupted by the SNP to produce CCCCCCT. Similar evidence strongly suggests a role for other, distinct, motifs. A second hotspot-weakening mutation, identified in the human hotspot NID1, disrupts the most hotspot-enriched 9-mer (CCCCACCCC) [27].

The results therefore strongly suggest that both SNPs shown to directly reduce human crossover activity do so by disrupting specific sequence motifs that somehow stimulate recombination. Much remains mysterious about the action of these motifs, which do not match known transcription factor binding sites, and have not been identified in previous recombination studies. First, it appears that the local sequence context around CCTCCCT, disrupted by the SNP to produce CCCCCCT. Similar evidence strongly suggests a role for other, distinct, motifs. A second hotspot-weakening mutation, identified in the human hotspot NID1, disrupts the most hotspot-enriched 9-mer (CCCCACCCC) [27].

Evolution of human recombination hotspots

Recent surveys comparing human linkage disequilibrium patterns with those of chimpanzees have provided strong evidence that although both humans and chimpanzees possess hotspots, these are rarely, if ever, shared between the species [28–30]. Comparison of sperm-based estimates of
current hotspot intensity in males with coalescent-based rate estimates, which use information from ancient recombination events, at seven hotspots, suggested rates of evolution of hotspot even higher than human genealogical timescales [14]. This poses an intriguing question of why hotspots exist in the first place – if hotspots have short life-spans, so most present-day hotspots have evolved recently, what genetic changes have generated them?

Of potential particular importance to understanding recombination rate evolution is the phenomenon of ‘biased transmission’ at recombination hotspots. Under the DSB model of recombination, mutations that disrupt hotspots by preventing DSBs from occurring locally on their own strand gain an evolutionary advantage. By being preferentially copied to repair breaks on the other strand, they become over-transmitted in heterozygotes, and more likely to become fixed in the population as a result. This is known to occur in humans – it was directly observed at the two hotspot-disrupting SNPs described above [14]. In the NID1 case, simulations suggest that the hotspot-disrupting mutation is almost certain to become fixed in the population in future (removing the more active motif from the population in the process), thus severely weakening the hotspot [14].

Evolutionary models suggest that biased transmission can prevent new hotspots from fixing in the population, and lead to the rapid removal of existing hotspots [31,32]. This phenomenon therefore offers an explanation of why humans and chimpanzees might differ in their hotspots, but not why a large number of hotspots are in fact present in the human genome, and for this reason has been labelled the ‘recombination hotspot paradox’. Random drift, direct selection for recombination and population demographic effects (e.g. bottlenecks) do not appear to be strong enough forces to explain this contradiction, particularly in the case of highly active hotspots in the genome, which are more strongly affected by the bias (G. Coop and S. Myers, unpublished work). The issue is particularly acute when there are specific motifs that promote local hotspot activity, since these motifs offer an obvious target for disrupting mutations.

What is the solution? One possibility is that sequence changes can promote hotspot activity non-locally – though perhaps relatively nearby – in the genome. Such mutations could generate hotspots, without suffering accompanying biases in transmission. One specific suggestion for how this could occur is the idea that removal of one hotspot (perhaps aided by biased transmission) could lead to an increase in intensity at neighbouring hotspots, with genetic constraints acting to maintain the amount of recombination on broader genomic scales. Evidence of such competition between hotspots has been found in S. cerevisiae [33–35] and mice [36]. Our results support this hypothesis, because we observe strong correlation between present-day recombination rates, estimated via linkage [2], and our historical rate estimates, over large distances ($r^2 = 0.97$ over 5 Mb) [8]. Despite the evidence that recombination rates have evolved locally over the timescales influencing our rate estimates, we see no corresponding signal of broad-scale rate evolution.

An alternative proposal is that the recombination machinery itself might evolve so rapidly that it differs between humans and chimpanzees [29,30]. This could result in simultaneous genome-wide shifting of hotspot locations, and human hotspot motifs might even be specific to our species. We are currently testing this hypothesis by analysing the evidence for hotspot activity using genetic variation data in positions in the chimpanzee genome orthologous to human CCTCCCT-associated hotspots where the motif is conserved. If such hotspots are not conserved, this will be strong evidence for a change in the recombination machinery. One piece of evidence that suggests a whole-scale change in machinery might have occurred is that we see no greater rate of molecular evolution at the CCTCCCT motif when it is associated with a human hotspot (82% conservation) than when not associated with a hotspot (81% conservation). Yet if biased transmission is a common feature of hotspot activity, we might expect a higher rate of substitution at hotspot-associated motifs. Clearly, much is yet to be done to provide a coherent picture of how the landscape of recombination evolves. However, comparative studies of recombination in

---

Table 1 | Fraction of motif occurrences associated with hotspots in different genetic backgrounds

<table>
<thead>
<tr>
<th>Sequence background</th>
<th>CCTCCCT$^a$</th>
<th>CCACGTGG 2 bp downstream</th>
<th>Occurrences in genome</th>
<th>Probability of hotspot, localized (&lt; 5 kb)</th>
<th>Probability of hotspot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outside repeat</td>
<td>Yes</td>
<td>No</td>
<td>417 704</td>
<td>0.018</td>
<td>0.166</td>
</tr>
<tr>
<td>DNA</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>0.000</td>
<td>0.330</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>18 831</td>
<td>0.017</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>744</td>
<td>0.095</td>
<td>0.333</td>
</tr>
<tr>
<td>THE1B</td>
<td>Yes</td>
<td>Yes</td>
<td>72</td>
<td>0.139</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>3735</td>
<td>0.022</td>
<td>0.151</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>157</td>
<td>0.089</td>
<td>0.401</td>
</tr>
<tr>
<td>THE1A</td>
<td>Yes</td>
<td>Yes</td>
<td>5</td>
<td>0.000</td>
<td>0.800</td>
</tr>
</tbody>
</table>

$^a$For the THE1A/B background, the motif CCTCCCT must occur in the optimal sequence context of bases 261–267 in the THE1B consensus [8].

©2006 Biochemical Society
closely related species offer a powerful approach to learning about the forces that drive recombination.

References

18 Reference deleted

Received 16 March 2006