www.sciencemag.org/cgi/content/full/310/5746/321/DC1

Supporting Online Material for

A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome

Simon Myers, Leonardo Bottolo, Colin Freeman, Gil McVean, Peter Donnelly*

*To whom correspondence should be addressed. E-mail: donnelly@stats.ox.ac.uk

Published 14 October 2005, *Science* **310**, 321 (2005)
DOI: 10.1126/science.1117196

This PDF file includes:

Supplementary Online Material

A fine-scale map of recombination rates and hotspots across the human genome

Simon Myers, Leonardo Bottolo, Colin Freeman, Gil McVean, Peter Donnelly

Contents:

A. Estimation of the recombination rate

Recombination rates were estimated separately for each chromosome in each population, with the data broken up into segments of 2000 SNPs, with an overlap of 500 SNPs. The previously described Markov-Chain Monte Carlo method, LDhat[1] was used, with a block penalty of 5 and 10 million iterations. The first third of the samples were removed as burn-in. In overlapping segments, rate estimates from the last 250 SNPs of the 5' region and 250 SNPs from the 3' region were removed (giving contiguous rate estimates). To convert estimates of $4N_e r$ to centimorgans, we estimated the effective population size from a comparison of total maps lengths of overlapping sections of the deCODE genetic map[2] and the genetic map estimated here (note that the SNPs used in the current study typically extend further than the deCODE markers). Over the scales assessed in this paper, recombination rates and genetic distances are effectively equivalent and will be used interchangeably. Estimated effective population sizes are CEPH: 9,600; African Americans: 15,700 and American Chinese: 10,300. These figures are similar to estimates from levels of genetic variation[3], one consequence of which is that it is unlikely that genome-wide recombination rates (i.e. the total amount of recombination within each chromosome) have changed significantly during recent human history (see also Fig S1).

To create a single map of recombination rate variation across the genome we took a simple arithmetic average of the recombination maps for each population. This approach has the advantage of leading to an unbiased estimate of recombination rate. Alternative approaches were tried (for example, weighting by the variance of local recombination rate estimates). However, these approaches typically led to large decreases in the total estimated genetic map length, and gave poor results in cases where rates had been estimated independently from sperm data.

The coalescent method for estimating recombination rates assumes a simple demographic history for each population (constant population size with no population structure, admixture or natural selection). While this model does not reflect known historical influences on genetic variation (for example ancient bottlenecks in non-African populations or admixture in African American populations), previous simulations[1] have shown that moderate deviations from the assumed model (to the degree likely to be important in human history) have little impact on the accuracy of estimates of recombination rate. Furthermore, where recombination rates have been estimated by independent means, estimates from the coalescent-based approach are in close agreement (see McVean et al[1] and this paper).

B. Comparison of recombination rates estimated from sperm-typing and genetic variation in a 3.3Mb region of the human MHC.

Previous work[4] estimated recombination rates across 3.3Mb of the human MHC region using single-sperm typing in 12 individuals (over 20,000 sperm analysed). The distribution of detected cross-over events in the sperm correlates closely with recombination rates estimated from genetic variation (Fig S2).

C. Detection of recombination hotspots

To identify the location of recombination hotspots with statistical support, we used the previously published composite-likelihood method, LDhot[1;5]. Briefly, we carry out a likelihood ratio test over 200kb windows on each population sample, comparing models with a constant recombination rate against one in which there is a central 2kb recombination hotspot (the analysis window is moved 1kb between analyses). Hotspots are defined as positions in the genome where there is evidence (defined below) of a local recombination rate that is elevated relative to that of the surrounding DNA. The observed difference in log composite likelihood is compared against the null distribution, which is obtained by simulations. Simulations are matched for sample size, SNP density, background recombination rate, the fact we are using genotype data and an approximation to the ascertainment scheme (single-hit SNP ascertainment from a panel of 20 chromosomes with an average depth of 18 for autosomes and a panel of 15 with an average depth of 13.5 for the non-pseudoautosomal region of the X chromosome).

To combine results across populations, we defined hotspots with statistical support as those for which there is at least one population in which the one-tailed *P*-value is less than 0.01 and a second population in which the one-tailed *P*-value is less than 0.05. To identify the centre of each hotspot, we identified the position within each successive series of ‘significant’ 2kb windows where the estimated recombination rate peaks. The width of the hotspot is defined as the region where the estimated recombination is within a factor of 2 of the maximum estimated rate. Occasionally, this approach can result in the identification of a narrow, hot hotspot located within a broader, cooler one. In such cases we take only the narrow hotspot. Of the hotspots detected in other studies (several ENCODE regions[5], the HLA type II region[1] and the region around MS32[6]) where resequencing and very dense genotyping was used, we identify approximately 60% from the data used in this study[7]. Importantly, comparison to data from sperm-typing [1;6] has so far revealed no false positive recombination hotspots for the LDhot approach.

D. Recombination hotspots and sequence features

We have conducted an extensive search for sequence features and motifs that are associated with the presence of recombination hotspots. Briefly, we considered every major repetitive-element class and individual characterised by RepeatMasker[8], all simple tandem repeats as characterised by Tandem Repeats Finder[9] and every possible DNA sequence motif of between 5 and 9 nucleotides.

Among the detected recombination hotspots we selected those with a width less than 5kb (36% of the 25,657 hotspots). For each, we identified a coldspot (a region where there is no support for a hotspot in any population - one-tailed *P*-value=1.0 - matched for SNP density and whether they were in a gene or not) within the surrounding region (median distance between hotspot and coldspot = 56 kb). For each repeat feature we calculated its frequency (and length distribution where appropriate), in hotspots and coldspots, and calculated a *P*-value for the null hypothesis of random distribution using a Fisher's exact test. Repeat features considered, and their associated *P*-values are shown in Tables S1, S2 and S3.

We exhaustively tested for enrichment within hotspots of various sequence motifs, separately within the THE1B/A LTR elements significantly enriched in hotspots (exact matches for all motifs of lengths 5-8), and within non-repeat DNA (after masking all DNA contained within repeats identified using RepeatMasker; exact matches for all motifs of lengths 5-9).

Table S1. Numbers of recombination hotspots and recombination coldspots[1] in which particular repeats are found.

	Number of hotspots	Number of coldspots	Relative Risk Ratio[2]	P-value[3]
L1	4306	5237	0.82	6.6×10^{-41}
MIR	5445	4823	1.13	1.8×10^{-18}
L2	4343	3751	1.16	7.86×10^{-17}
Other	4	29	0.14	3.76×10^{-4}
snRNA	33	64	0.52	7.33×10^{-2}
MER1_type	2404	2226	1.08	9.39×10^{-2}
Alu	5897	5701	1.03	0.11
CR1	723	628	1.15	0.28
Low_complexity	3732	3555	1.05	0.29
MER2_type	694	780	0.89	0.74
MaLR	3323	3183	1.04	1
ERV1	1172	1252	0.94	1
DNA	234	198	1.18	1
scRNA	13	5	2.60	1
PiggyBac	14	24	0.58	1
Satellite	8	3	2.67	1
Simple_repeat	3847	3766	1.02	1
MER1_type?	73	60	1.22	1
rRNA	19	13	1.46	1
AcHobo	359	337	1.07	1
tRNA	22	16	1.38	1
ERV	8	5	1.60	1
ERVL	1336	1310	1.02	1
Mariner	224	235	0.95	1
RNA	11	8	1.38	1
Unknown	15	12	1.25	1
Tip100	269	260	1.03	1
srpRNA	5	7	0.71	1
Tc2	79	76	1.04	1
ERVK	80	79	1.01	1

[1] We considered only those repeat classes for which a total of 10 or more were observed.

[2] Ratio of frequency in hotspots to frequency in non-hotspots

[3] P-values corrected for multiple testing by Bonferroni correction

Table S2. The 25 individual repeat elements showing the greatest difference in abundance between recombination hotspots and recombination coldspots.

	Number of hotspots	Number of coldspots	Relative Risk Ratio	P-value
L2	4343	3751	1.16	1.13×10^{-15}
THE1B	457	261	1.75	4.14×10^{-11}
L1PA4	83	201	0.41	6.35×10^{-10}
MIRb	3508	3062	1.15	4.30×10^{-9}
L1PA7	88	188	0.47	6.48×10^{-7}
L1PA3	53	133	0.40	1.70×10^{-6}
L1P	52	125	0.42	1.76×10^{-5}
L1PA5	90	177	0.51	4.50×10^{-5}
MIR	2715	2396	1.13	8.76×10^{-5}
L1M4	294	415	0.71	2.07×10^{-3}
SVA	4	29	0.14	5.40×10^{-3}
CT-rich	407	296	1.38	1.14×10^{-2}
(TA)n	517	654	0.79	1.99×10^{-2}
L1MB7	113	183	0.62	2.46×10^{-2}
LTR5B	15	0	NAN	3.05×10^{-2}
GA-rich	447	338	1.32	4.02×10^{-2}
THE1D-int	9	35	0.26	5.23×10^{-2}
L1MEc	101	163	0.62	7.35×10^{-2}
THE1A	81	39	2.08	7.56×10^{-2}
GC_rich	20	50	0.40	0.22
L1M2	21	50	0.42	0.38
L1M3	15	40	0.38	0.5
L1ME3B	174	241	0.72	0.52
L1MEb	19	45	0.42	0.77
L1ME1	170	233	0.73	0.88
L1ME2	120	174	0.69	0.9

Table S3. Simple Tandem Repeats (STRs) with significant differences (*P*<0.05) between recombination hotspots and coldspots

STR	Number of hotspots	Number of coldspots
(AT)n	86	186
(ACC)n	18	6
(AAAT)n	234	191
(AAGG)n	151	117
(AATATAT)n	7	18

For the first case, we compared the frequency of each motif within THE1B/A elements overlapping narrow (width<=5kb) hotspots, to its frequency within THE1B/A elements not overlapping any known hotspot (so conditional on the frequencies of the THE1A/B elements themselves in each group). This identified a collection of motifs, all giving highly significant p-values ($<10^{-10}$), all of which remain strongly significant even after Bonferroni correction for the number of motifs tested. Closer examination of the motifs with these low p-values, and alignment to the consensus sequence for THE1B, demonstrated that these motifs strongly clustered into the region between 256bp and 280bp. Two distinct significance peaks aligned to this region, corresponding to distinct motifs, shown below, beneath the consensus sequence to which they are aligned. Each motif represents a 1-bp change from the consensus sequence.

THE1B Consensus, 256bp-280bp:

TGAGGCCTCCCCAGCCATGTGGAAC

Motifs: **CCTCCCT** **CCACGTGG**

P-value: 1.2×10^{-34} 1.6×10^{-20}

(note that the motifs above are imprecise, in the sense that it is difficult to be sure whether to include e.g. either of the bases flanking the motif CCTCCCT; this inclusion does not strongly alter the relative risk ratio observed).

The p-values above refer to tests where we only counted occurrences of the two motifs whose location within their THE1B element aligned to positions 261bp-267bp, and 270bp-277bp, respectively in the THE1B consensus. All occurrences of

CCACGTGG were assumed to align at this location, and occurrences of CCTCCCT matching at least one of the two flanking consensus bases on either side were assumed to align to bases 261-267 within the consensus. This increased the enhancement of the motif CCTCCCT within the hotspot regions (Table S4), suggesting strongly that within this element, the motif context determines whether a recombination hotspot results.

Table S4. Motifs strongly signalled within THE1B hotspots

THE1B	# Non-hot	# Hotspot (≤5kb)	RR[1]	P-value (FET)
Total occurrences	17588	505	-	-
CCTCCCT (261bp-267bp)	425	80	5.85	1.2×10^{-34}
CCACGTGG	1636	119	2.53	1.6×10^{-20}
CCTCCCT (all)	873	92	3.67	6.1×10^{-22}
TACTGTTC	2994	153	1.78	6.9×10^{-13}

[1] The 'relative risk' for each motif, within hotspots vs. outside, obtained by dividing the frequency of the motif in hotspot elements by the frequency in non-hotspot elements.

Only one motif mapping outside the region 256bp-280bp of the consensus shows $p<10^{-10}$; the motif TACTGTTC (p=6.9×10^{-13}), which maps to the THE1B consensus 122bp-129bp, and again represents a 1-bp change from the consensus, TGCTGTTC. However, a number of motifs in other parts of the element also show comparatively low p-values ($<10^{-6}$), worthy of further investigation.

We conducted motif testing for THE1A (where the consensus sequence within the region 256bp-280bp is identical) in the same manner. The motif CCTCCCT at 261bp-267bp again showed strong enhancement within hotspot THE1A's (RR=5.1, $p=1.4 \times 10^{-6}$), with the less significant p-value reflecting the ~5-fold fewer occurrences of THE1A in the genome relative to THE1B. The other two motifs described above showed no significant enhancement within THE1A elements ($p>0.05$). The LTRs of the other THE elements THE1C and THE1D showed no enhancement within hotspot regions, so were not tested for motifs.

Apart from the above testing, we also examined the frequencies of different motifs in hotspots and matched coldspots, after masking repeat sequence. We ordered motifs according to the difference in the number of hotspots vs. coldspots containing the motif (in unmasked sequence), in order to identify those motifs which might explain the most recombination hotspots (Table S5).

Of the five top scoring 8-mers, all bar CCTCCTCT either contain the top scoring 7-mer or are subsequences of the top scoring 9-mer. The top scoring 7-mer, CCTCCCT, exactly matches the top scoring 7-mer found within THE1B elements. A point mutation found to disrupt crossover activity[10] at the hotspot DNA2 in the MHC class II region destroys an occurrence of this motif (see below).

The top-scoring 9-mer is CCCCACCCC. A recent paper[11] identified a second point mutation that appears to disrupt crossover and gene conversion at the NID1 human recombination hotspot, located on chromosome 1. The "non-hotspot" allele removes an occurrence of this motif (present in the “hotspot” allele), suggesting this distinct motif might also promote human recombination hotspots.

Table S5. Motifs showing largest difference in occurrence between hotspots and coldspots in unmasked sequence. The table shows the top 5 motifs, for each motif length between 5 and 9 inclusive (top motif for each length highlighted in bold)

Length	Ranking	Element	# of hotspots[1]	# of coldspots[2]	Difference[3]
9	**1**	**CCCCACCCC**	**987**	**656**	**331**
	2	CCCACCCCC	730	432	298
	3	CCCCCACCC	810	518	292
	4	GAAAAAAAA	3257	2974	283
	5	AAAAAAAAA	4042	3765	277
8	**1**	**CCTCCCTG**	**1868**	**1269**	**599**
	2	CCCCACCC	1844	1280	564
	3	CCCACCCC	1750	1222	528
	4	CCTCCTCT	1950	1431	519
	5	TCCTCCCT	1943	1429	514
7	**1**	**CCTCCCT**	**4366**	**3380**	**986**
	2	CCTTCCC	4272	3351	921
	3	CTCCTCC	4130	3216	914
	4	TCCCCAG	4008	3118	890
	5	CCCCACC	3475	2587	888
6	**1**	**GGGGGT**	**4773**	**3579**	**1194**
	2	CCCCCT	5311	4162	1149

	3	CCCCCA	6316	5178	1138
	4	CCCTGC	6364	5229	1135
	5	CCCCAG	6617	5522	1095
5	**1**	**CCCCC**	**7695**	**6584**	**1111**
	2	CCCCG	4457	3407	1050
	3	GGCCC	7410	6374	1036
	4	GCCCC	7816	6828	988
	5	CCCGC	3825	2865	960

[1] The number of hotspots containing the motif

[2] The number of matched coldspots containing the motif

[3] The difference between the number of hotspots, and the number of coldspots, containing the motif. Motifs were ranked according to this measure (which estimates the number of hotspots explained by each possible motif).

E. Examination of flanking sequence around hotspots DNA2 and NID1.

We obtained the flanking sequence around the mutations disrupting hotspots DNA2 [10] and NID1 [11] from the web pages of Prof. Alec Jeffreys' research group (see http://www.le.ac.uk/genetics/ajj/HLA/index.html, and http://www.le.ac.uk/ge/ajj/MS32/NID.html). To aid the reader, we here reproduce the 20bp on either side of the disrupting mutation in each case, with the hot-spot and disrupting alleles labelled in each case, the location of the disrupting mutation underlined, and occurrences of the motifs CCTCCCT and CCCCACCCC within the sequences shown in blue and red, respectively. Note that for the case of NID1 there are two occurrences of CCCCACCCC within 20bp of the disrupting mutation, only one of which is removed by the mutation.

DNA2

Hotspot allele:

TGCAGGGGGCAGCAACAGGGAGGCTGTCTTTTCTGAGA A/T GG

Disrupting allele:

TGCAGGGGGCAGCAACAGGGGGGCTGTCTTTTCTGAGA A/T GG

NID1

Hotspot allele:

CTTTTTAATTTTAAATCACCCCCCACCCCACCCCAACATAC

Disrupting allele:

CTTTTTAATTTTAAATCACCTCCCACCCCACCCCAACATAC

F. Estimation of the probability a motif causes a hotspot

We estimated the probability that the motif CCTCCCT causes hotspots on different backgrounds, and the proportion of hotspots caused by this motif, as follows. These estimates are necessarily approximate, since they rely on assumptions about our power to detect hotspots, and on empirical counts of motifs on different backgrounds.

For the THE1B (long terminal repeat) background, suppose THE1B's containing the motif CCTCCCT (labelled THE1B+ elements) occur at rate w per base in the genome, and each such occurrence causes a hotspot with probability p_{hot}. We seek to estimate this probability.

Further suppose that with probability p_{call}, a hotspot is identified by our method, and placed within the set of narrow hotspots (of length 5kb or below). Similarly, with probability p_{ncall}, the THE1B element is instead placed in the set of those in our "cold" regions, not overlapping any identified hotspot. Let l_{genome} be the (known) length of the whole genome, l_{hot} be the (known) combined total length of our hotspot regions, and l_{cold} the (known) total length of the non-hotspot regions.

Then the expected number of THE1B occurrences in the hotspot regions is:

$N_{THE1B+hot}=w\times(l_{genome}\times p_{hot}\times p_{call} + l_{hot}\times(1-p_{hot}))$ (1)

and the expected number in the non-hotspot regions is:

$N_{THE1B+nhot}=w\times (l_{genome}\times p_{hot}\times p_{ncall}+l_{cold}\times (1-p_{hot}))$ (2)

Substituting in the observed values in for $N_{THE1B+hot}$ and $N_{THE1B+nhot}$, then dividing (1) by (2), we may solve the resulting equation to produce a moment estimator of p_{hot}, the quantity of interest, provided we make assumptions about p_{call} and p_{ncall}. To obtain the numbers given in the paper, we used the fact that 9292 of the 25637 hotspots identified were 5kb wide or less, and estimated our power to detect a hotspot at 50%, to give $p_{call}=0.5\times 9292/25637=0.18$ and $p_{ncall}=0.5$ (we excluded all hotspots from the cold set; p_{call} is correct assuming a negligible false-positive rate for the hotspots, which would reduce p_{call}). Using l_{genome}=2,669,549kb (for the sequenced human genome), l_{cold}= 2,359,803kb, and l_{hot}=37,833kb, $N_{THE1B+hot}=80$, and $N_{THE1B+nhot}= 476$, and solving gives

$p_{hot}=0.58$

Proceeding in the same way for THE1A's, using $N_{THE1B+hot}=14$, and $N_{THE1B+nhot}=87$ gives

$p_{hot}=0.56$,

suggesting a very similar "hotspot penetrance" in each case of 55-60% for the motif. If we have overestimated our power to detect hotspots (for example, if a lack of SNPs in some regions of the genome means we fail to detect hotspots in these regions), this estimated penetrance would increase (to around 75% with power of 40%), and in our view it is possible that the true figure could even be close to 1 (or, indeed, somewhat below 0.5).

We proceeded very similarly to estimate the probability of causing a hotspot for the CCTCCCT motif outside the THE1B regions, and in non-repeat DNA. To calculate the probability we used the observations of 4366 hotspot and 3380 non-hotspot occurrences. We assumed the same value for as before for p_{call}, but took p_{ncall}=0.0 (since we attempted to compare coldspots with hotspots for this analysis; the results are very similar if we assume as above that our "coldspots" are not in fact perfectly cold). Here l_{genome} is unchanged, but l_{cold}=l_{hot}=37,833kb. Calculating as before gives p_{hot}=0.022 for 50% power to detect hotspots (p_{hot}=0.028 for 40% power). For any assumed value for the power, the estimated probability of causing a hotspot would be much lower for CCTCCCT occurrences in non-repeat DNA than in the THE1A/B elements.

Finally, we (conservatively) estimated the number of hotspots caused by CCTCCCT's within THE1B elements and outside them, by simply subtracting the observed counts in hotspots from those expected based on the non-hotspot regions, to give:

(80+14)-(476+87) ×37833/2359803=85 hotspots of 9292 narrow hotspots (0.9% of hotspots) caused by CCTCCCT's within THE1A/B LTR's.

4366-3380=986 of 9292 (11% of hotspots) caused by CCTCCCT's outside THE1B.

Although the estimated penetrance of the CCTCCCT motif appears much lower outside THE1A/B elements, the much larger number of occurrences outside these elements results in the greater contribution of this element to the hotspot count outside such elements. These estimates are of course very approximate and are only intended as a rough guide to the magnitude of the likely contribution of the motif in the two different settings.

G. Comparison of hotspot signals across populations

We did not directly compare recombination rates across different populations, since assessment of uncertainty in these rates is difficult. Instead, we compared the evidence for hotspots in the different population groups, since this evidence implicitly takes into account uncertainty in recombination rate estimation in each group.

To compare hotspot evidence across populations, we called hotspots exactly as before (see above), but now required only that there was evidence ($p<0.05$) of a hotspot in at least one of the three populations, in order to assess how often evidence was found in other populations. This resulted in 40,115 unique hotspots being called under this criterion across the genome. For each hotspot region identified this way, we recorded the minimal p-value within each population inside the hotspot. Table S6 summarises how often there were differing amounts of evidence ($p<0.05$) in each of the three populations for this new hotspot set. 74% of hotspots had evidence in at least one other population, whilst 45% show evidence in all populations.

Using the criteria of p<0.01 in at least one population, rather than p<0.05, the agreement among populations is improved (table S7), with 83% of 32,541 identified hotspots now showing evidence (p<0.05) in at least one other population, and 54% showing evidence in all three populations. We conclude that at most potential hotspot positions, hotspot evidence is present across multiple populations, and that whilst we cannot discount the possibility of there existing hotspots specific to single populations, lack of power to detect hotspots could equally be responsible for the incomplete correspondence observed at other sites.

Table S6. Evidence in the three populations for potential hotspot regions showing p<0.05 in at least one population.

	AA only	EA only	HC only	AA+EA	AA+HC	EA+HC	All populations
Counts	3782	2908	3927	3300	3773	4387	18038
Proportion	0.09	0.07	0.10	0.08	0.09	0.11	0.45

Table S7. Evidence in the three populations for potential hotspot regions showing p<0.01 in at least one population.

	AA only	EA only	HC only	AA+EA	AA+HC	EA+HC	All populations
Counts	1884	1509	2238	2603	3038	3731	17538
Proportion	0.06	0.05	0.07	0.08	0.09	0.11	0.54

H. Predicting recombination rates at different scales

Using linear models, we assessed the ability of various sequence features and genome annotations to predict recombination rates over 5Mb, 500kb, 50kb and 5kb scales (tables S8-11). Following previous studies[2;12] we considered the predictors:

Factor
GC content
CpG fraction
polyA/polyT fraction
Gene count
Gene fraction
Exon count
Exon fraction
Chromosome
Chromosome arm
Distance from telomere

Factors such as GC content are all calculated for windows of the corresponding size (i.e. 5Mb, 500kb, 50kb, and 5kb). At the 5Mb scale, our results are similar to those of previous studies[2;12] that have compared pedigree-based estimates of recombination rate to various genomic features, although in some cases we have used somewhat different predictors; chromosome size, distance from telomere, GC content, CpG fraction and polyA/polyT fraction are all strongly associated with recombination rate, and the full set of features explain a total of 42% of the variance in recombination rate at this scale. At finer scales, the same factors are important, but their explanatory power is strongly diminished. At the 500kb scale we can explain 34% of the variance in recombination rate, dropping to 15% at 50kb and only 4% at 5kb.

Table S8. Linear model analysis of recombination rates at the 5Mb scale

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)	
Chrom1	0.11	0.13	0.87	0.39	
Chrom2	-3.08×10^{-2}	0.13	-0.25	0.81	
Chrom3	6.05×10^{-2}	0.14	0.44	0.66	
Chrom4	0.28	0.15	1.91	5.74×10^{-2}	•
Chrom5	9.99×10^{-2}	0.15	0.68	0.50	
Chrom6	0.14	0.15	0.94	0.35	
Chrom7	0.11	0.16	0.71	0.48	
Chrom8	-3.60×10^{-2}	0.17	-0.22	0.83	
Chrom9	0.24	0.18	1.31	0.19	
Chrom10	0.17	0.18	0.94	0.35	
Chrom11	-0.11	0.17	-0.62	0.53	
Chrom12	0.29	0.18	1.60	0.11	
Chrom13	0.47	0.21	2.24	2.56×10^{-2}	*
Chrom14	0.46	0.21	2.19	2.90×10^{-2}	*
Chrom15	0.45	0.22	2.06	3.99×10^{-2}	*
Chrom16	0.25	0.23	1.07	0.29	
Chrom17	0.44	0.24	1.80	7.34×10^{-2}	•
Chrom18	0.45	0.24	1.89	5.92×10^{-2}	•
Chrom19	0.40	0.35	1.14	0.25	
Chrom20	-0.29	0.27	-1.08	0.28	
Chrom21	0.93	0.36	2.59	1.00×10^{-2}	**
Chrom22	0.22	0.37	0.58	0.56	
Chromosome arm1	-0.24	8.57×10^{-2}	-2.83	4.91×10^{-3}	**
Distance from telomere	-0.22	4.38×10^{-2}	-4.98	9.15×10^{-7}	***
polyA/polyT fraction	-0.61	0.10	-5.79	1.35×10^{-8}	***
CpG fraction	0.79	0.13	6.10	2.31×10^{-9}	***
GC content	-0.62	0.18	-3.42	6.84×10^{-4}	***
Gene count	-0.47	0.15	-3.09	2.17×10^{-3}	**
Gene fraction	-0.50	0.46	-1.09	0.28	
Exon count	-5.22×10^{-2}	0.15	-0.35	0.73	
Exon fraction	0.59	0.46	1.28	0.20	

Signif. codes: 0-0.001: ***, 0.001-0.01: **, 0.01-0.05: *

Residual standard error: 0.7608 on 438 degrees of freedom
Multiple R-Squared: 0.4583, Adjusted R-squared: 0.42
F-statistic: 11.95 on 31 and 438 DF, p-value: $< 2.2\times10^{-16}$

Table S9. Linear model analysis of recombination rates at the 500kb scale

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)	
Chrom1	-1.20×10^{-4}	4.07×10^{-2}	-3.00×10^{-3}	1.00	
Chrom2	-1.87×10^{-2}	4.08×10^{-2}	-0.46	0.65	
Chrom3	1.94×10^{-2}	4.40×10^{-2}	0.44	0.66	
Chrom4	0.15	4.67×10^{-2}	3.14	1.71×10^{-3}	**
Chrom5	7.16×10^{-4}	4.75×10^{-2}	1.50×10^{-2}	0.99	
Chrom6	2.29×10^{-2}	4.81×10^{-2}	0.48	0.63	
Chrom7	-1.79×10^{-2}	4.98×10^{-2}	-0.36	0.72	
Chrom8	-0.11	5.22×10^{-2}	-2.07	3.82×10^{-2}	*
Chrom9	4.86×10^{-2}	5.68×10^{-2}	0.86	0.39	
Chrom10	-1.36×10^{-2}	5.41×10^{-2}	-0.25	0.80	
Chrom11	-0.20	5.39×10^{-2}	-3.72	1.99×10^{-4}	***
Chrom12	9.49×10^{-2}	5.48×10^{-2}	1.73	8.34×10^{-2}	•
Chrom13	0.25	6.52×10^{-2}	3.87	1.10×10^{-4}	***
Chrom14	0.14	6.77×10^{-2}	2.03	4.28×10^{-2}	*
Chrom15	0.16	6.96×10^{-2}	2.26	2.40×10^{-2}	*
Chrom16	-5.47×10^{-3}	6.88×10^{-2}	-8.00×10^{-2}	0.94	
Chrom17	8.47×10^{-2}	7.09×10^{-2}	1.20	0.23	
Chrom18	0.24	7.20×10^{-2}	3.36	7.85×10^{-4}	***
Chrom19	0.28	8.65×10^{-2}	3.24	1.19×10^{-3}	**
Chrom20	-0.11	7.91×10^{-2}	-1.33	0.18	
Chrom21	0.45	0.11	4.25	2.16×10^{-5}	***
Chrom22	-3.73×10^{-3}	0.11	-3.50×10^{-2}	0.97	
Chromosome arm1	-5.42×10^{-2}	2.66×10^{-2}	-2.04	4.17×10^{-2}	*
Distance from telomere	-0.20	1.37×10^{-2}	-14.95	$<2\times10^{-16}$	***
polyA/polyT fraction	-0.46	1.99×10^{-2}	-23.30	$<2\times10^{-16}$	***
CpG fraction	0.17	2.80×10^{-2}	5.94	3.09×10^{-9}	***
GC content	-6.34×10^{-2}	3.25×10^{-2}	-1.95	5.10×10^{-2}	•
Gene count	-0.18	3.10×10^{-2}	-5.94	3.10×10^{-9}	***
Gene fraction	0.13	4.11×10^{-2}	3.28	1.04×10^{-3}	**
Exon count	-0.11	2.85×10^{-2}	-3.88	1.06×10^{-4}	***
Exon fraction	-0.11	4.31×10^{-2}	-2.45	1.42×10^{-2}	*

Signif. codes: 0-0.001: ***, 0.001-0.01: **, 0.01-0.05: *

Residual standard error: 0.8149 on 5266 degrees of freedom
Multiple R-Squared: 0.3397, Adjusted R-squared: 0.3358
F-statistic: 87.4 on 31 and 5266 DF, p-value: $< 2.2\times10^{-16}$

Table S10. Linear model analysis of recombination rates at the 50kb scale

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)	
Chrom1	-2.38×10^{-3}	1.46×10^{-2}	-0.16	0.87	
Chrom2	4.17×10^{-3}	1.46×10^{-2}	0.29	0.78	
Chrom3	1.92×10^{-2}	1.57×10^{-2}	1.23	0.22	
Chrom4	0.11	1.67×10^{-2}	6.58	4.68×10^{-11}	***
Chrom5	1.27×10^{-2}	1.69×10^{-2}	0.75	0.45	
Chrom6	1.13×10^{-2}	1.71×10^{-2}	0.66	0.51	
Chrom7	-1.88×10^{-3}	1.77×10^{-2}	-0.11	0.92	
Chrom8	-4.19×10^{-2}	1.86×10^{-2}	-2.25	2.43×10^{-2}	*
Chrom9	3.35×10^{-2}	2.04×10^{-2}	1.65	0.10	
Chrom10	-4.46×10^{-3}	1.93×10^{-2}	-0.23	0.82	
Chrom11	-0.12	1.92×10^{-2}	-6.05	1.48×10^{-9}	***
Chrom12	3.41×10^{-2}	1.95×10^{-2}	1.75	8.02×10^{-2}	•
Chrom13	0.16	2.32×10^{-2}	6.97	3.14×10^{-12}	***
Chrom14	7.61×10^{-2}	2.41×10^{-2}	3.16	1.58×10^{-3}	**
Chrom15	9.48×10^{-2}	2.49×10^{-2}	3.81	1.40×10^{-4}	***
Chrom16	-2.66×10^{-2}	2.43×10^{-2}	-1.10	0.27	
Chrom17	-1.82×10^{-3}	2.51×10^{-2}	-7.30×10^{-2}	0.94	
Chrom18	0.15	2.55×10^{-2}	5.72	1.05×10^{-8}	***
Chrom19	5.63×10^{-2}	2.98×10^{-2}	1.89	5.92×10^{-2}	•
Chrom20	-5.49×10^{-2}	2.80×10^{-2}	-1.96	4.97×10^{-2}	*
Chrom21	0.23	3.73×10^{-2}	6.22	5.16×10^{-10}	***
Chrom22	-1.53×10^{-2}	3.74×10^{-2}	-0.41	0.68	
Chromosome arm1	-3.33×10^{-2}	9.49×10^{-3}	-3.51	4.49×10^{-4}	***
Distance from telomere	-0.12	4.79×10^{-3}	-26.02	$<2.00\times10^{-16}$	***
polyA/polyT fraction	-0.26	7.82×10^{-3}	-33.08	$<2.00\times10^{-16}$	***
CpG fraction	3.06×10^{-2}	9.64×10^{-3}	3.17	1.52×10^{-3}	**
GC content	7.21×10^{-2}	1.31×10^{-2}	5.49	3.96×10^{-8}	***
Gene count	-3.05×10^{-2}	6.25×10^{-3}	-4.88	1.09×10^{-6}	***
Gene fraction	-1.68×10^{-2}	5.67×10^{-3}	-2.97	2.96×10^{-3}	**
Exon count	-7.14×10^{-2}	6.33×10^{-3}	-11.27	$<2.00\times10^{-16}$	***
Exon fraction	-6.63×10^{-2}	6.54×10^{-3}	-10.14	$<2.00\times10^{-16}$	***

Signif. codes: 0-0.001: ***, 0.001-0.01: **, 0.01-0.05: *

Residual standard error: 0.9226 on 53170 degrees of freedom
Multiple R-Squared: 0.1494, Adjusted R-squared: 0.1489
F-statistic: 301.2 on 31 and 53170 DF, p-value: $< 2.2\times10^{-16}$

Table S11. Linear model analysis of recombination rates at the 5kb scale

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)	
Chrom1	4.74×10^{-3}	4.87×10^{-3}	0.97	0.33	
Chrom2	6.44×10^{-3}	4.88×10^{-3}	1.32	0.19	
Chrom3	8.32×10^{-3}	5.24×10^{-3}	1.59	0.11	
Chrom4	6.09×10^{-2}	5.57×10^{-3}	10.93	$<2.00\times10^{-16}$	***
Chrom5	9.01×10^{-3}	5.67×10^{-3}	1.59	0.11	
Chrom6	-2.73×10^{-3}	5.73×10^{-3}	-0.48	0.63	
Chrom7	-9.15×10^{-3}	5.93×10^{-3}	-1.54	0.12	
Chrom8	-1.81×10^{-2}	6.23×10^{-3}	-2.91	3.61×10^{-3}	**
Chrom9	1.10×10^{-2}	6.82×10^{-3}	1.61	0.11	
Chrom10	-1.43×10^{-3}	6.48×10^{-3}	-0.22	0.83	
Chrom11	-5.59×10^{-2}	6.41×10^{-3}	-8.72	$<2.00\times10^{-16}$	***
Chrom12	8.05×10^{-3}	6.53×10^{-3}	1.23	0.22	
Chrom13	8.76×10^{-2}	7.75×10^{-3}	11.30	$<2.00\times10^{-16}$	***
Chrom14	4.06×10^{-2}	8.07×10^{-3}	5.03	4.85×10^{-7}	***
Chrom15	4.44×10^{-2}	8.33×10^{-3}	5.33	9.72×10^{-8}	***
Chrom16	-1.07×10^{-2}	8.13×10^{-3}	-1.32	0.19	
Chrom17	-2.17×10^{-2}	8.38×10^{-3}	-2.59	9.51×10^{-3}	**
Chrom18	7.97×10^{-2}	8.54×10^{-3}	9.33	$<2.00\times10^{-16}$	***
Chrom19	-1.12×10^{-2}	9.87×10^{-3}	-1.14	0.26	
Chrom20	-2.21×10^{-2}	9.35×10^{-3}	-2.36	1.81×10^{-2}	*
Chrom21	0.12	1.25×10^{-2}	9.90	$<2.00\times10^{-16}$	***
Chrom22	1.08×10^{-2}	1.25×10^{-2}	0.87	0.39	
Chromosome arm1	-1.60×10^{-2}	3.18×10^{-3}	-5.02	5.09×10^{-7}	***
Distance from telomere	-8.56×10^{-2}	1.58×10^{-3}	-54.32	$<2.00\times10^{-16}$	***
polyA/polyT fraction	-5.64×10^{-2}	2.14×10^{-3}	-26.35	$<2.00\times10^{-16}$	***
CpG fraction	-2.47×10^{-2}	2.31×10^{-3}	-10.68	$<2.00\times10^{-16}$	***
GC content	0.15	3.14×10^{-3}	46.29	$<2.00\times10^{-16}$	***
Gene count	-1.21×10^{-2}	1.73×10^{-3}	-7.01	2.40×10^{-12}	***
Gene fraction	-4.06×10^{-2}	1.68×10^{-3}	-24.14	$<2.00\times10^{-16}$	***
Exon count	-2.97×10^{-2}	1.95×10^{-3}	-15.20	$<2.00\times10^{-16}$	***
Exon fraction	-2.24×10^{-2}	1.91×10^{-3}	-11.72	$<2.00\times10^{-16}$	***

Signif. codes: 0-0.001: ***, 0.001-0.01: **, 0.01-0.05: *

Residual standard error: 0.9775 on 532415 degrees of freedom
Multiple R-Squared: 0.04455, Adjusted R-squared: 0.0445
F-statistic: 800.9 on 31 and 532415 DF, p-value: $< 2.2\times10^{-16}$

I. Comparison of male and female recombination rate

To compare recombination rates on the X chromosome with those on autosomes, we first created a novel autosomal data set that matched the X chromosome for sample size, SNP density and allele frequency spectrum. To achieve this, SNPs on the X chromosome were binned into 10 categories with respect to minor allele frequency. For each bin, we calculated the average SNP density. For each autosome we first selected a random sample of individuals to match the total number of X chromosomes. We then categorised SNPs into the same frequency bins, and sub-sampled sufficient SNPs from each bin to match the SNP density of the same bin on the X chromosome. In addition, X chromosome haplotypes from males were paired to create pseudo-genotype data (comparable to the autosomes). Recombination rates and recombination hotspots were assessed as above.

Because the X chromosome differs systematically in terms of average recombination rate from autosomes, to compare the distribution of recombination hotspots we divided the X chromosome and autosomes into 2 Mb regions, and for each obtained its male and female recombination rates from the deCODE genetic map[2]. Regions were divided into categories by recombination rate (windows of 0.5cM/Mb: male plus female rate for autosomes and female rate for X chromosomes) and for each of the 2Mb region on the X chromosome we sampled a region from the same rate bin on the autosomes. The sampled autosomal regions were concatenated to create a pseudo X-chromosome, and recombination rates were compared using a QQ-plot that compares the smallest proportion of sequence that can explain x% of recombination for the two chromosomes (Fig S4 (B)).

A similar procedure was used to compare recombination rates in regions where the male recombination rate is relatively high with regions where the female recombination rate is relatively high. First, we divided the autosomes into 2Mb regions. Among those regions where we had both fine-scale estimates and estimates of male and female recombination rates from the deCODE[2] genetic map, we identified the 100 with the highest ratio of male to female recombination rates, and 100 with the highest ratio of female to male recombination rate. To match regions for average recombination rate, we categorised each set of 100 regions by sex-averaged rate into bins of 0.25cM/Mb, and for each bin then sub-sampled pairs of regions up to the minimum of the male-biased count and the female-biased count. For the 66 resulting paired regions, we compared the distribution of recombination rates using the concatenation and QQ-plot approach described above (Fig S4 (A)).

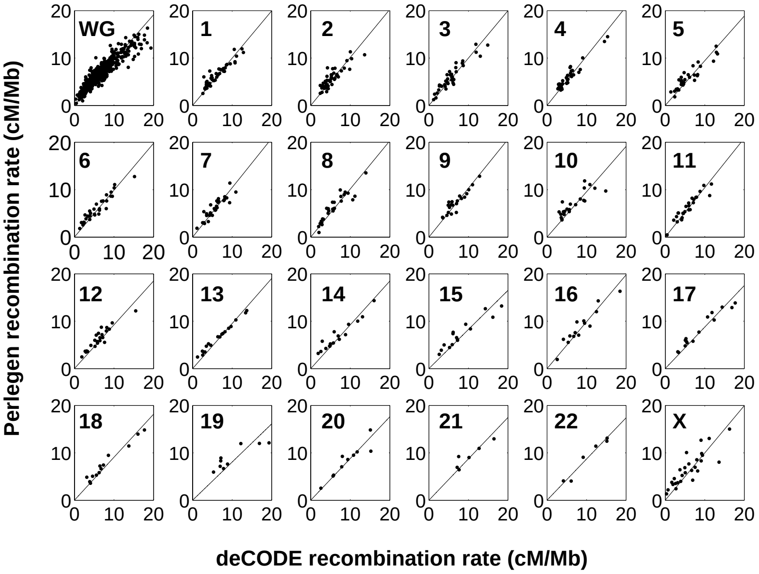
Supplementary Figure Legends

Figure S1. Comparison of recombination rates estimated from population data with those from the existing human genetic map. Scatter plots of the coalescent estimates of recombination rate (Y-axis) with the deCODE pedigree based estimates[2] (X-axis) for 5Mb regions across the genome in both maps, shown for the whole genome (WG) and separately for each chromosome (numbered). The square of the correlation coefficient relating to a linear regression forced to go through the origin (r^2, or the proportion of variance explained) is 0.966 for the whole genome. For autosomes, r^2 ranges from 0.946 to 0.993 and for the X chromosome it is 0.923. 5Mb is towards the lower limit of the resolution of the pedigree-based maps.

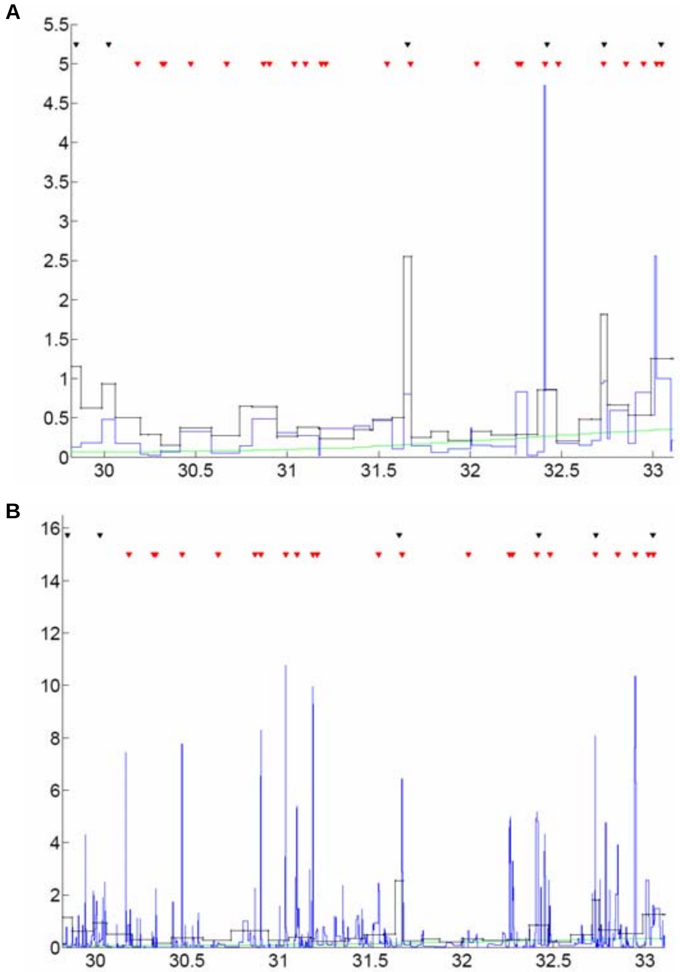
Figure S2. Recombination rates in the 3.3Mb of the human MHC region. Estimates from the deCODE[2] map (green), the sperm-typing study[4] (black)and the present study (blue). The location of identified recombination hotspots from the sperm-typing study are represented as black triangles and from the present study as red triangles. (A) Fine-scale recombination rates averaged over the markers intervals corresponding to the sperm-typing study[4]. (B) Fine-scale recombination rate estimates (not averaged). Note that where a recombination hotspot has been identified in the sperm-typing study at approximately 30Mb corresponds to a region with multiple, low peaks in recombination rate, none of which are identified as statistically-significant hotspots in the current study. However, the combination of these low peaks is sufficient to create a peak in estimate rate that corresponds closely to the rates estimated from sperm-typing.

Figure S3. Recombination rate estimates for the 22 human autosomes and X chromosome. For the X chromosome we have treated the two pseudo-autosomal regions separately from the non-pseudoautosomal region. In each plot the hotspots are shown as triangles above the recombination rates, with colour representing the intensity (total amount of recombination) in the hotspot (from low, blue, to high, red).

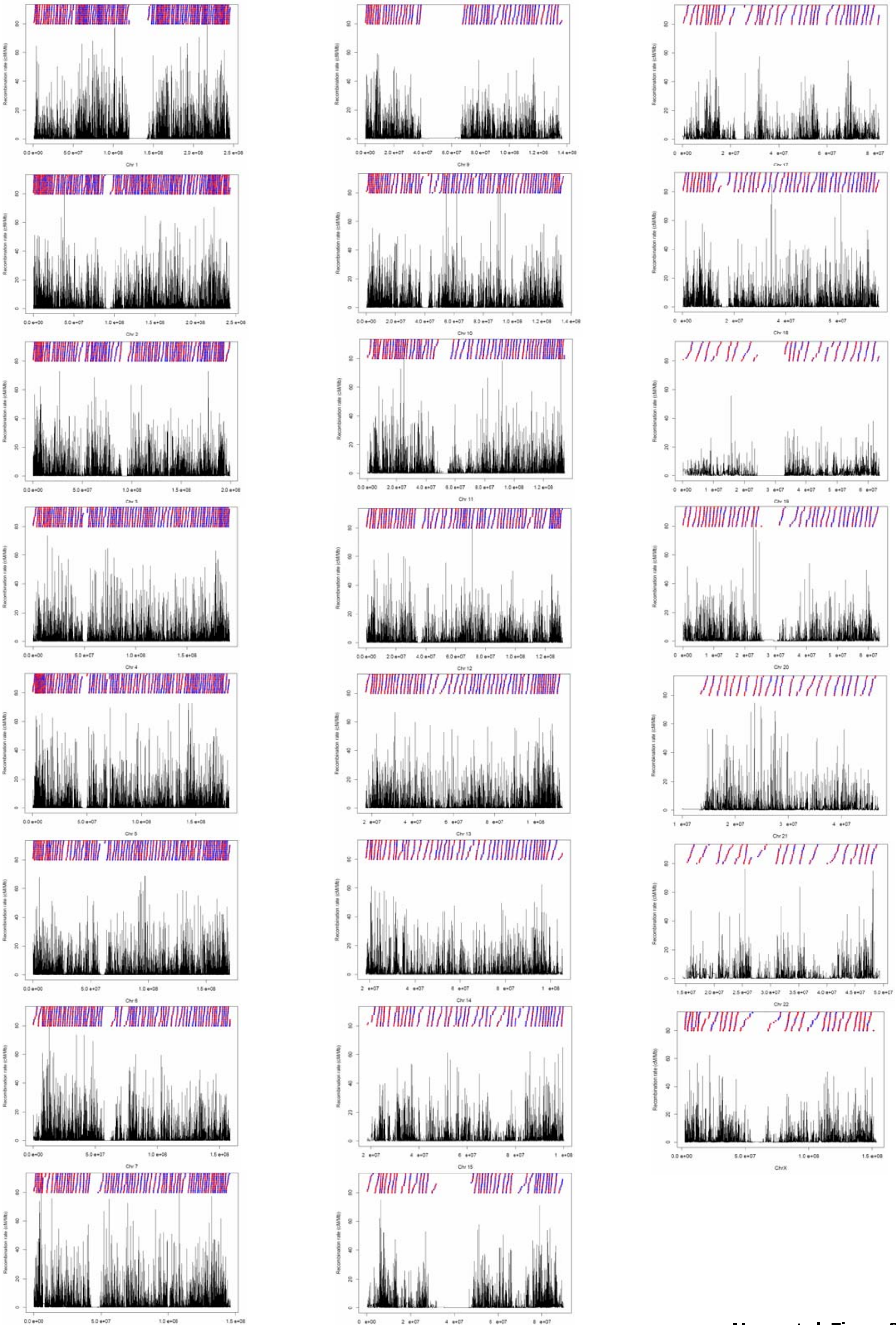
Figure S4. Differences in recombination between males and females. (A) QQ-plot comparing the distribution of recombination rates in regions where males have a relatively high recombination rate (in these regions 68% of all recombination occurs in males, compared to a genome average of 33%) to regions where females have a relatively high recombination rate (in these regions 96% of recombination occurs in females). (B) QQ plot comparing the distribution of recombination rates on the X chromosome with rates on the autosomes in regions matched for sample size, SNP density, allele frequency spectrum and total recombination rate.



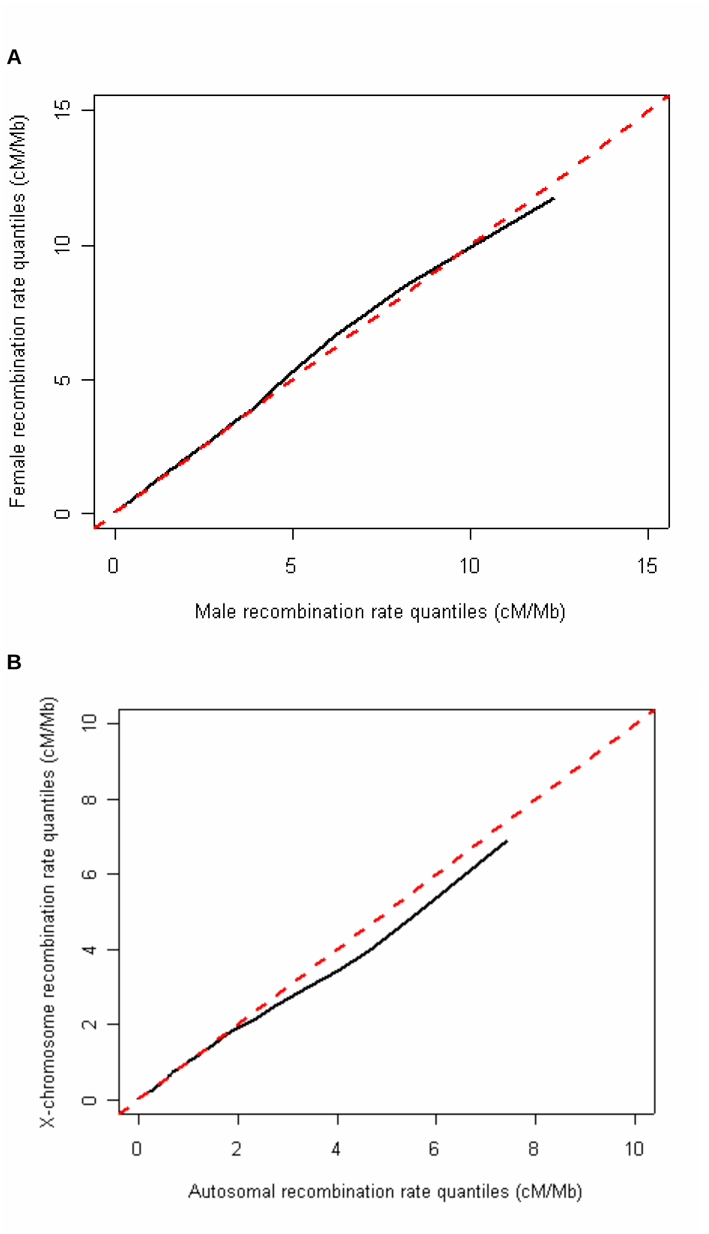
Myers et al. Figure S1



Myers et al. Figure S2



Myers et al. Figure S3



Myers et al. Figure S4

References and notes

1. G. A. McVean et al., *Science* **304**, 581-584 (2004).

2. A. Kong et al., *Nat.Genet.* **31**, 241-247 (2002).

3. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296-302 (2000).

4. M. Cullen, S. P. Perfetto, W. Klitz, G. Nelson, M. Carrington, *Am.J.Hum.Genet.* **71**, 759-776 (2002).

5. W. Winckler et al., *Science* **308**, 107-111 (2005); published online 10 February 2005 (10.1126/science.1105322).

6. A. J. Jeffreys, R. Neumann, M. Panayi, S. Myers, P. Donnelly, *Nat.Genet.* **37**, 601-606 (2005).

7. D. A. Hinds et al., *Science* **307**, 1072-1079 (2005).

8. Smit, A. F. A., Hubley, R., and Green, P. RepeatMasker Open-3.0. 1996-2004, available at http://www.repeatmasker.org

9. G. Benson, *Nucleic Acids Res.* **27**, 573-580 (1999).

10. A. J. Jeffreys and R. Neumann, *Nat.Genet.* **31**, 267-271 (2002).

11. A. J. Jeffreys and R. Neumann, *Hum.Mol.Genet.* **14**, 2277-2287 (2005).

12. M. I. Jensen-Seaman et al., *Genome Res.* **14**, 528-538 (2004).