SUPPLEMENTARY TEXT.

Rationale for using *Hop2*$^{-/-}$ mice for DSB mapping.

Due to technical issues associated with the biology of female meiosis in mammals, only males were used in this study. The fraction of cells at the DSB stage represents less than 2% of the male germ cell population [31]. To enrich for DSB-positive cells we used *Hop2* knockout mice (*Hop2*$^{-/-}$). The HOP2 protein is directly involved in the D-loop formation during homologous recombination[32,33], and meiotic DSBs are not repaired in *Hop2*$^{-/-}$ mice[13]. This leads to the arrest of meiotic progression[13]. As a result, cells of later stages of spermatogenesis do not form in the *Hop2* knockout, and the fraction of germ cells at the DSB stage is enriched more than 10-fold (not shown). In wild type mice the sensitivity of hotspot detection is relatively low due to the transient nature of DSBs and to asynchrony in the timing of their formation and repair. This effect is eliminated in the *Hop2*$^{-/-}$ mice where unrepaired breaks persist. Previous studies indicated that lack of the HOP2 protein does not change the number of introduced DSBs or the loading of RAD51 and DMC1 [13,34,35], making this mutant a valuable tool for meiotic DSB mapping.

To map DSB hotspots in the wild type mice we employed juvenile animals undergoing the first wave of spermatogenesis. Mice at the age of 11-12 days are devoid of germ cells beyond the zygotene stage of meiosis and are relatively enriched for DSB-stage cells [36]. Although the sensitivity of this approach is lower than in the case of *Hop2*$^{-/-}$ mice, we were able to detect more than 2500 hotspots (P value = 10^{-4}, FDR=6.7%). These showed a good correlation with *Hop2*$^{-/-}$ hotspots: 94% of the top 2000 hotspots

found in the wild type mice are also present in the *Hop2* knockout mutant (Supplementary Fig. 1).

Correlation of the DSB map with genetic maps.

A distinctive feature of mouse recombination hotspots is their dependence on the genetic background; i.e. the same DNA sequence in different mouse strains can be either recombinationally hot or cold [1,37]. Nevertheless, though the position of individual hotspots may differ, the correlation of recombination efficiencies within longer chromosomal segments is quite significant[15]. According to a two-stage model [11], recombination rates over large domains are tightly controlled, but the strength and position of individual hotspots within these regions change rapidly. This prompted us to compare our DSB map to two available genetic maps generated on different genetic backgrounds [15,16]. Although at higher resolution (250 Kb; the approximate upper limit of the resolution of genetic maps) the maps show low correlations, at low resolution (5 Mb) correlations between the maps are quite significant (R=0.46-0.64) (Fig. 1d). The R values for map comparisons increase linearly to around 7 Mb resolution, then level off and plateau at approximately 10 Mb resolution (Supplementary Fig. 6). Whether this implies the presence of 7-10 Mb chromosomal domains with strain-independent recombination rates will require further investigation including the analysis of DSB maps from multiple mouse strains.

Estimation of the recombination activity of the hotspot cluster in the Pseudoautosomal Region (PAR).

We can estimate the frequency of DSB formation in the PAR using the autosomal *H2Eα* hotspot[14,18,38] as a reference. On average, only one out of 15 meiotic DSBs in mice leads

to the formation of CO[37]. The average estimate of recombination efficiency of the *H2Eα* hotspot is ~1 cM [18] (i.e. 2% of spermatocytes will have a CO in this region). Hence, the fraction of cells with a DSB in this area could be as high as 30%. The combined strength of the DSB hotspot cluster in the PAR is 3.5 fold higher than that of the *H2Eα* hotspot (Fig. 2a), suggesting that about one DSB occurs in the PAR region in every meiosis. If these are the only DSBs formed in the PAR, this would indicate that DSBs in this region form COs with nearly a 1:1 ratio. At the same time, the fact that only ~ 35 Kb of the PAR has been sequenced in the mouse precluded us from characterizing the DSB landscape in 95% of this region. While it seems unlikely that the DSB frequency remains uniformly high throughout the entire PAR, it is reasonable to expect that more DSBs may be generated in unmapped regions, further ensuring that a CO will be made.

Enrichment of Long Terminal Repeats (LTR) of the MaLR superfamily in the hotspots.

It has been previously reported that THE-1 LTRs are particularly enriched in human hotspots[11]. Interestingly, we find that MTC and MTD LTRs belonging to the same MaLR retrotransposon superfamily as THE-1, also show the strongest enrichment among repeated sequences at mouse hotspots (MTC: 72% enrichment, $P = 10^{-28}$; MTD: 63% enrichment, $P = 10^{-42}$, one sided binomial tests). Interestingly, neither THE-1 repeats are found in mouse, nor are MTC/D retrotransposons found in humans. Thus, despite 70 million years of evolutionary time separating human and mouse and substantial sequence divergence between THE-1 and MTC/D LTRs, some functional elements responsible for increased recombination rate in the MaLR LTRs appear to be conserved.

The possible origins of the nucleotide skew.

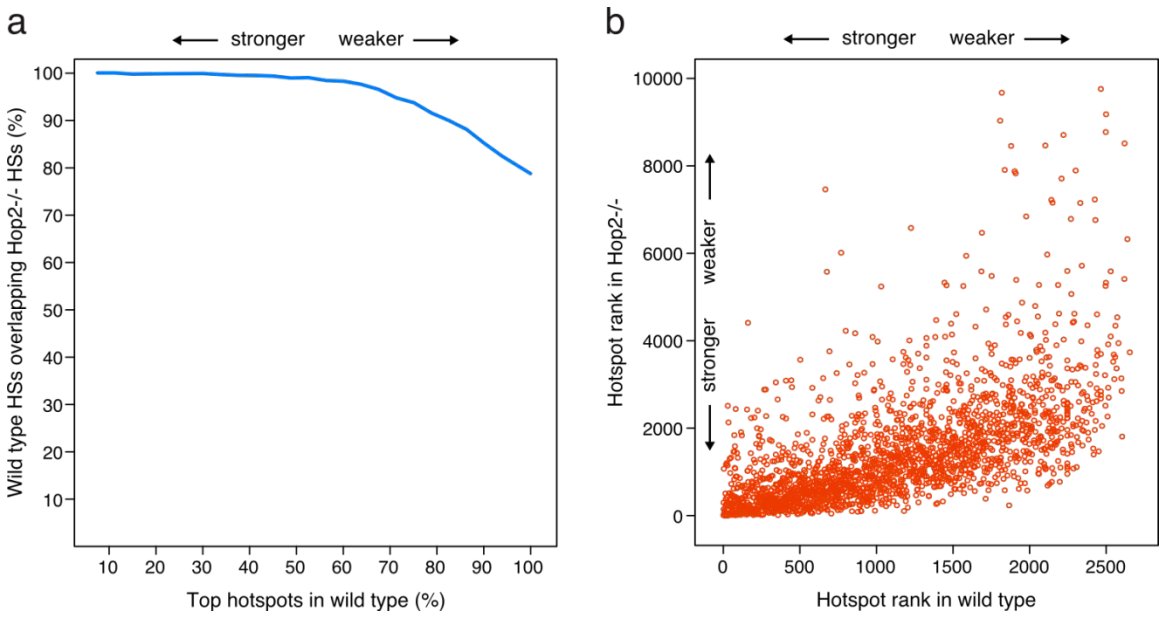
During recombination, DSB ends are resected to produce long single-stranded 3' overhangs. It is therefore conceivable that mutational asymmetry may exist between resected and non-resected DNA strands (Supplementary Fig. 9c). Given that 3' overhangs will form on the opposite strands on the two sides of the DSB site, this may explain the observed flip in the skew at the DSB centre. It is also possible that the skew represents some unknown functional elements of the genome that favour DSB formation.

Interestingly, the skew is more prominent in the repetitive DNA fraction of hotspot DNA, although the unique DNA fraction also exhibits this skew pattern (Supplementary Fig. 9).

H3K4me3 at hotspots and transcription start sites.

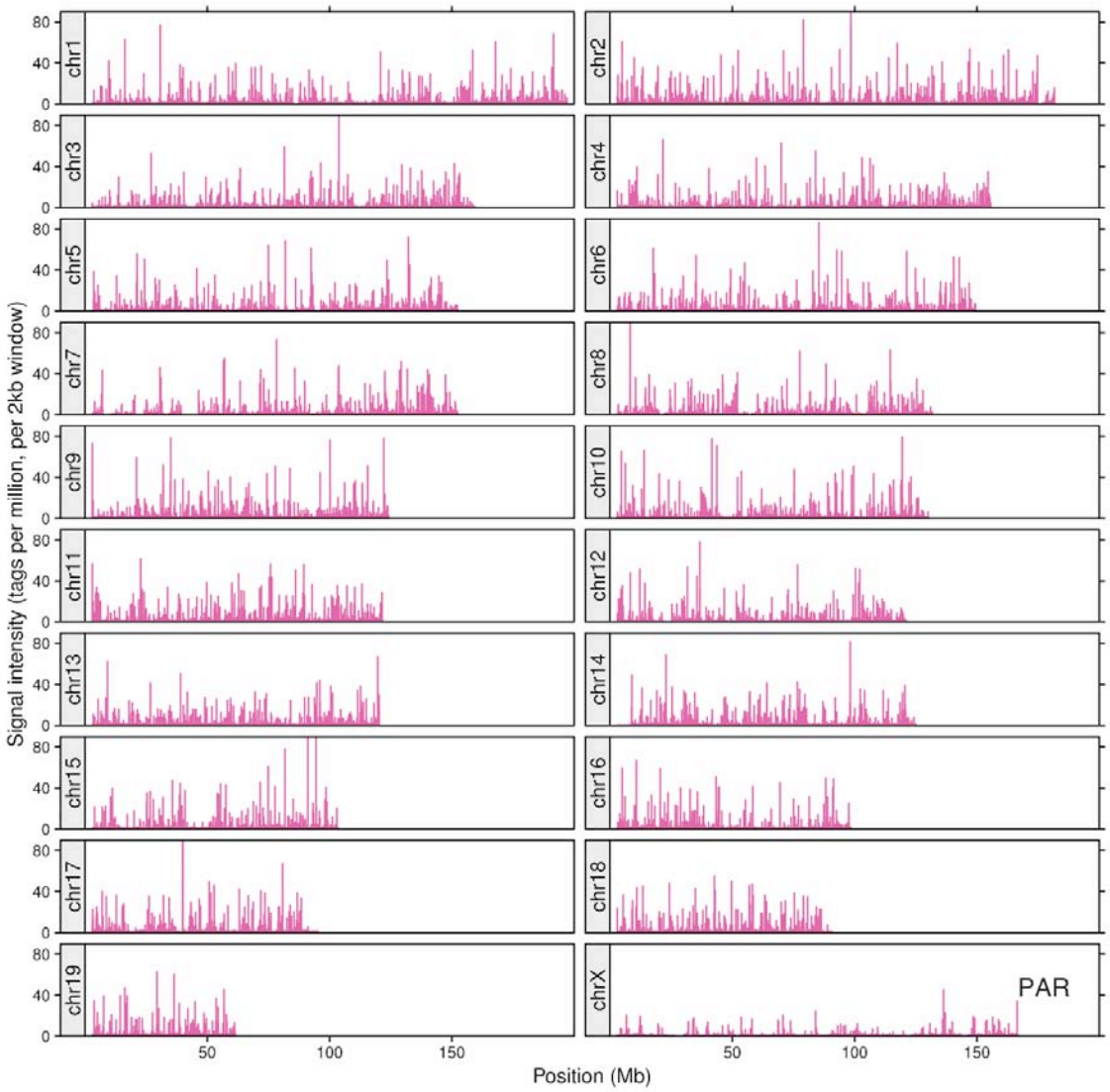
Correlation of H3K4me3 with hotspots was shown genome-wide in *S.cerevisiae* [23], but the meaning of this correlation is difficult to access because hotspots in this organism overlap transcription start sites that themselves are marked by H3K4me3. We found that only 8% of the mouse recombination hotspots overlap transcription start sites, and that the majority of mouse hotspots carry a hotspot-specific H3K4me3 mark (Fig. 3b). Hotspot-associated H3K4me3 marks are significantly weaker than the ones located at the transcription start sites of (presumably) active genes (Fig. 3c and Supplementary Fig. 14), but within this lower range there is a positive correlation of the hotspot strength and the extent of H3K4 trimethylation (Supplementary Fig. 14).

SUPPLEMENTARY FIGURES

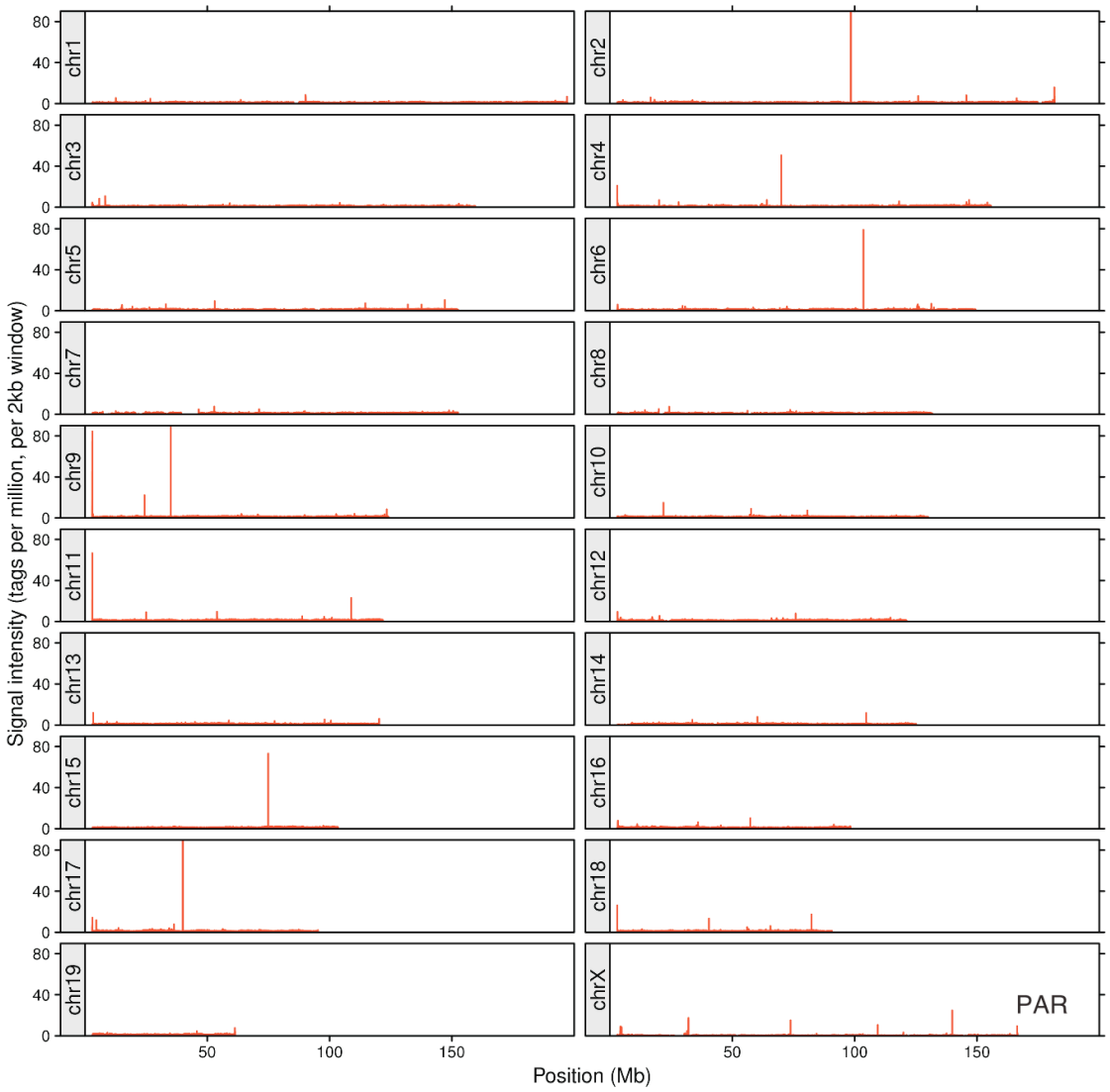


Supplementary Figure 1. The vast majority of the hotspots found in wild type mice are present in the *Hop2* knockout. **a.** Subsets of wild-type hotspots are given in decreasing strength order on the X axis (strongest 10%, strongest 20%, strongest 30%, etc.). The overlap of these hotspots (HSs) with the hotspots found in *Hop2*$^{-/-}$ mice is shown on the Y axis. The wild type set is made up of replicates J and K, with 20 million tags total (Materials and Methods). The *Hop2*$^{-/-}$ set is made up of replicates A, B, C and D, 107 million tags total (Materials and Methods and Fig. 1a). **b.** The strength ranks of hotspots found in wild type mice were compared to the ranks of corresponding hotspots in the *Hop2*$^{-/-}$ mice. It is evident that mostly the strongest hotspots have been identified in the wild type background.

a


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chrX
PAR
80
40
0
50
100
150
Signal intensity (tags per million, per 2kb window)
Position (Mb)

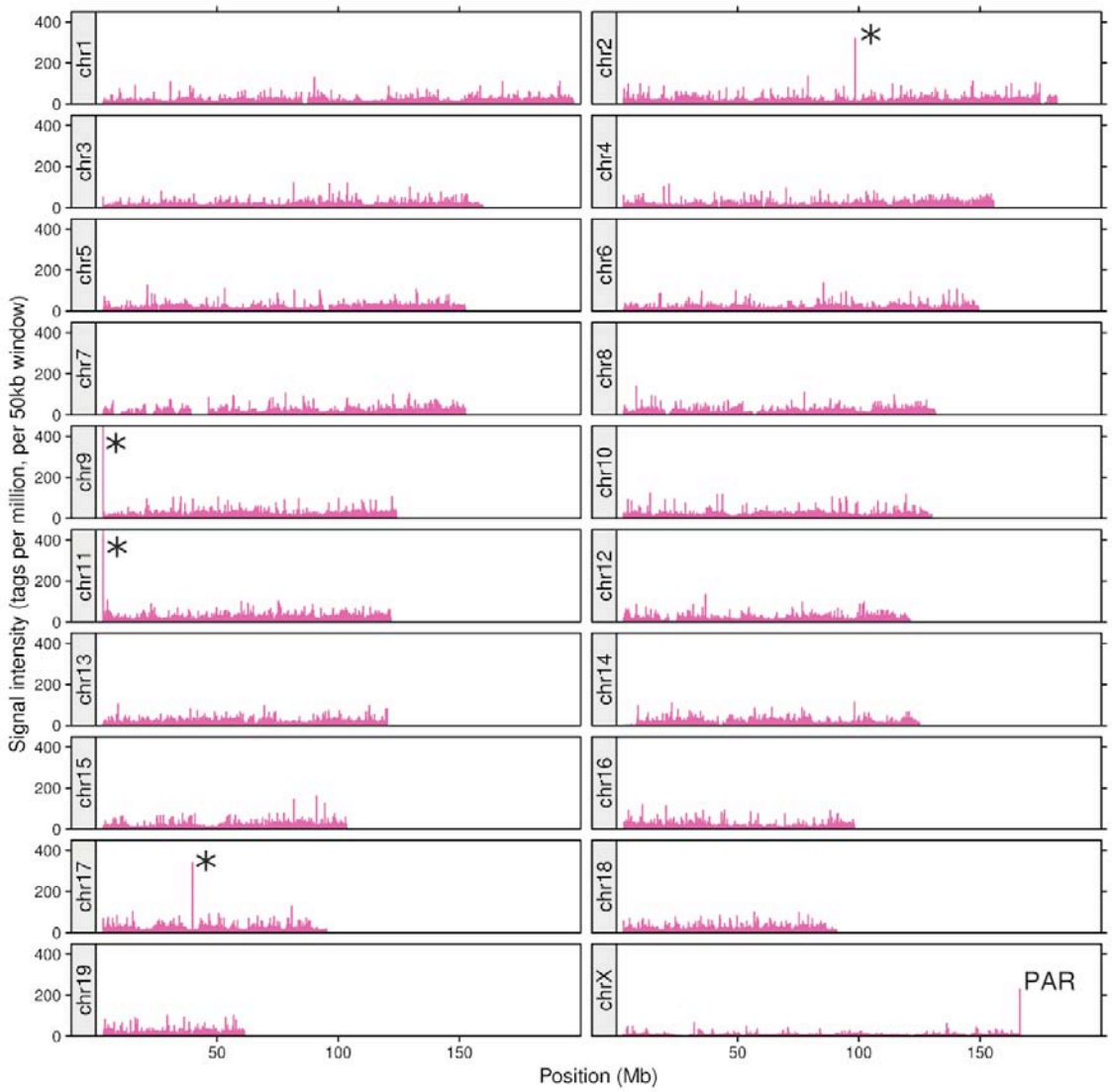
b



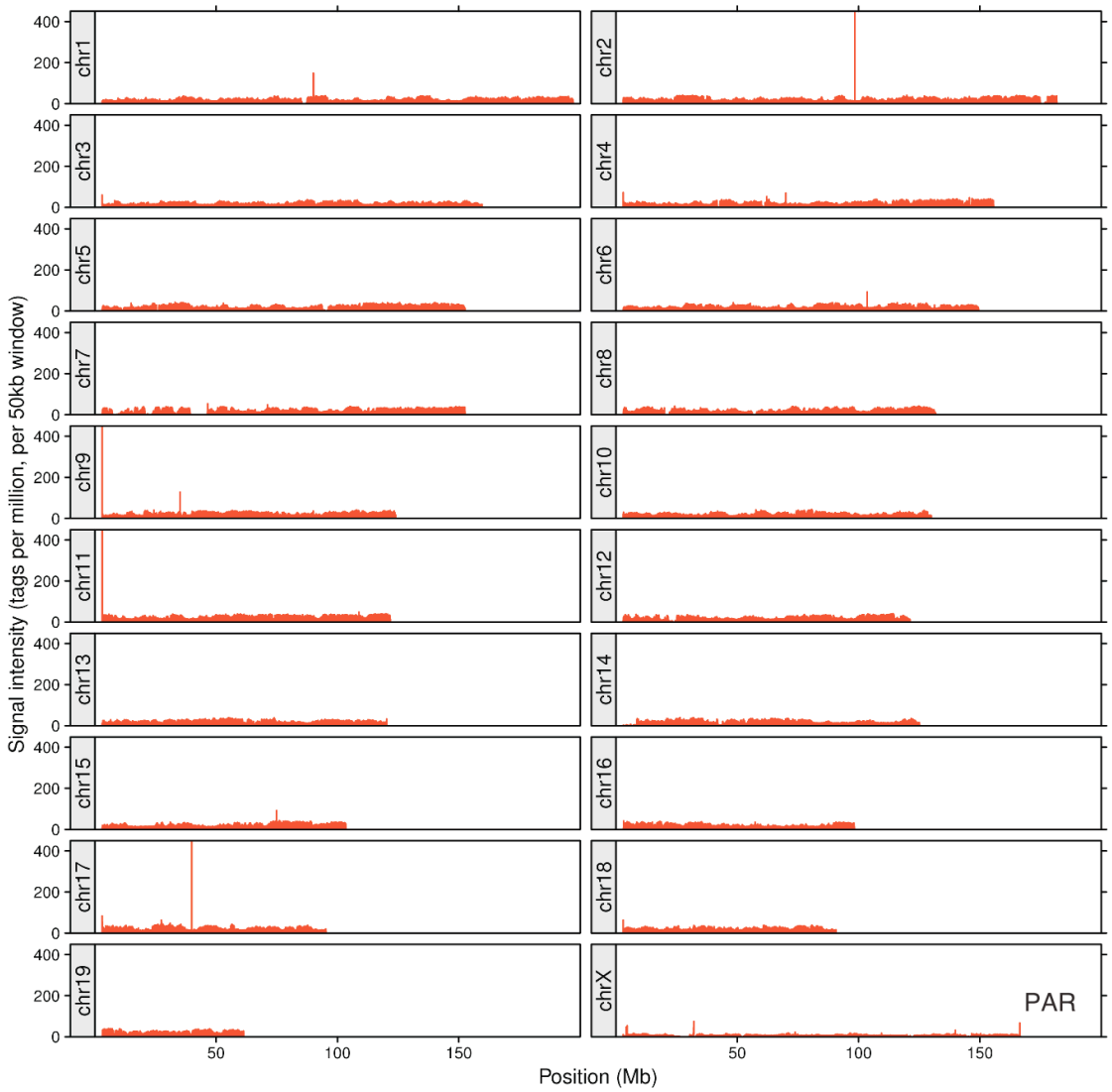
Supplementary Figure 2. Genome-wide ChIP-seq tag distributions. The signal intensity was calculated as the tag coverage in each non-overlapping 2 kb window along the genome using **a.** anti-DMC1 ChIP tags and **b.** control tags (input DNA and anti-IgG ChIP). Coverage was normalized by the total tag count for each sample. Note that the lower hotspot density on chromosome X could be the result of having only a single copy

of X in males that in turn, reduces the statistical power for peak identification.

a

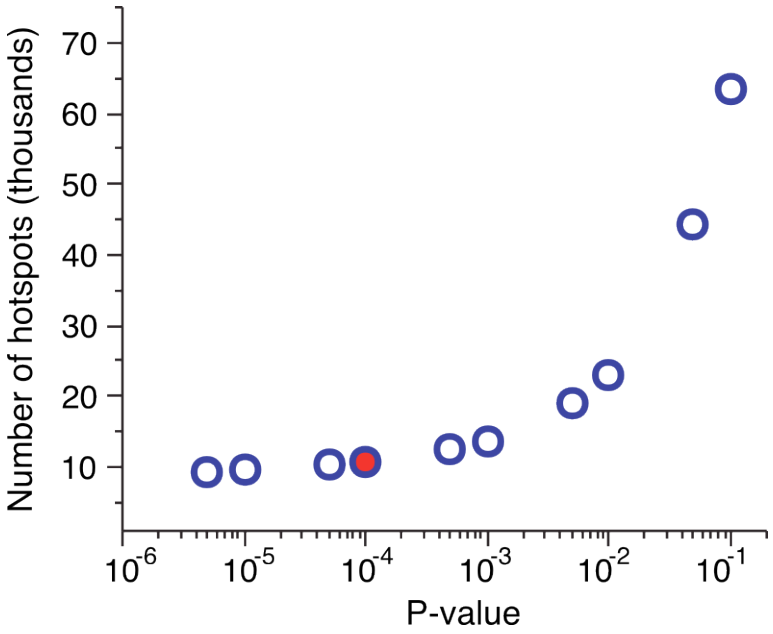

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chrX
PAR
400
200
0
50
100
150
Signal intensity (tags per million, per 50kb window)
Position (Mb)

b

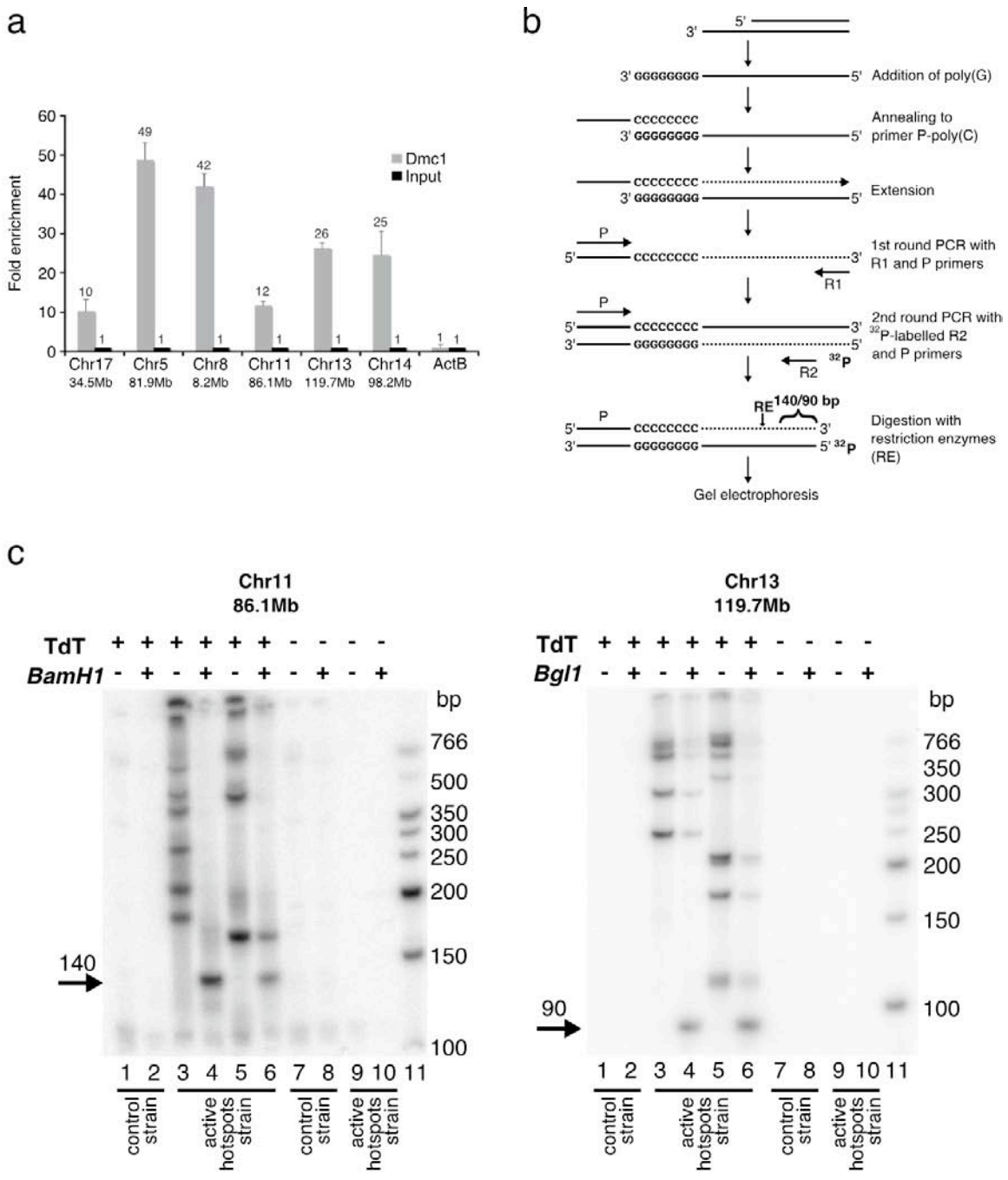


Supplementary Figure 3. Genome-wide ChIP-seq tag distributions at low resolution. The signal intensity was calculated as the tag coverage in each non-overlapping 50 kb window along the genome using **a.** anti-DMC1 ChIP tags and **b.** control tags. Coverage was normalized by the total tag count for each sample. Strong peaks (*) on chromosomes 2, 9, 11 and 17 result from discrepancies between the mm9 / NCBI37 genome assembly

and the genome of our mouse strain. At this resolution, the strongest signal in the genome is at the PAR. Note that the lower hotspot density on chromosome X could be the result of having only a single copy of X in males that in turn, reduces the statistical power for peak identification.

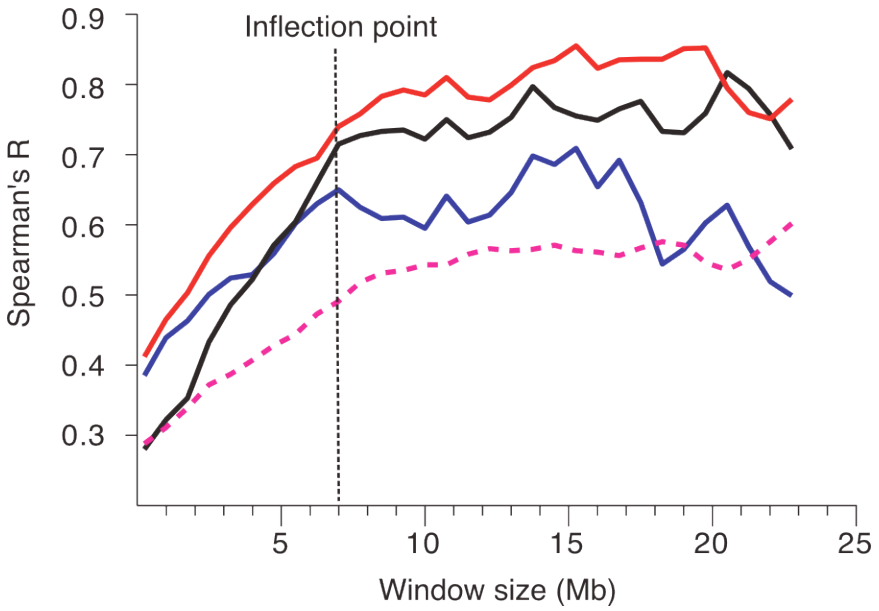


Supplementary Figure 4. Number of hotspots as a function of statistical significance. 10,686 hotspots are identified at a P-value of 10^{-4} (filled circle). FDR correction reduces this to 9,874 (see Supplementary Methods). Although the number of hotspots could be higher when less strict parameters are used we restricted our analysis to the 9,874 hotspots to reduce the effect of false positives.

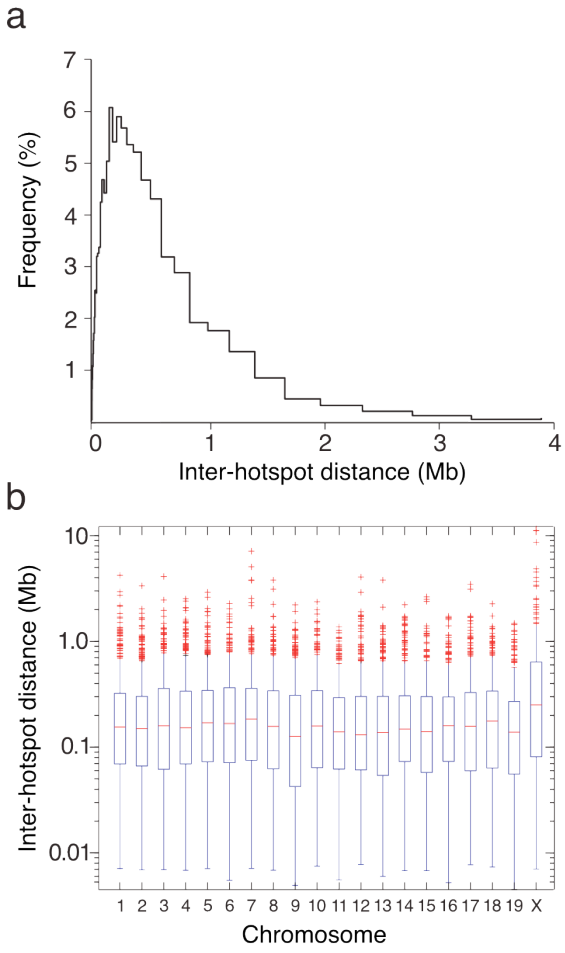


Supplementary Figure 5. Confirmation of newly identified DSB hotspots. **a.** Enrichment of the DNA corresponding to several new strong hotspots by anti-DMC1 ChIP estimated by qPCR. Quantitative data are expressed as the ratio of the enriched (ChIP) to the input DNA. All data were averages of at least three independent

experiments. **b.** Schematic of the physical detection of individual meiotic DSBs[14]. A poly (dG) tail is added to the 3' DNA end of DSB by terminal deoxynucleotide transferase. An annealing and fill-in reaction are used to introduce the P sequence. Hotspot-specific primers (R1 and radio labeled R2) and primer P are used for amplification of the DNA adjacent to the end of the DSB. Since alternative DSB sites within the same hotspot can be utilized multiple radio labeled fragments will be detected after PCR. Treatment of any of these fragments with a hotspot-specific restriction endonuclease will result in a labeled fragment of the same length, corresponding to the distance between the restriction enzyme site and the 5' end of the R2 primer (140 and 90 bp in this study). **c.** Direct detection of individual meiotic DSBs in two new strong hotspots. Testicular DNA was prepared from the mouse strain used for DSB mapping (active hotspot strain) and C57Bl/6J strain that does not have the hotspots in the same positions (control strain). Poly(G) tailing of the ends of the DSBs, and two rounds of PCR were performed to specifically amplify the fragments adjacent to a break site as previously described[14] – see panel b. Restriction enzymes (BamHI, BglI) were added as indicated. Each panel represents a separate hotspot (Chr11, Chr13), and the sample order is kept the same in each gel. Terminal deoxytransferase (TdT) that adds a poly(G) tail was omitted in lanes 7-10 to show the specificity of PCR amplification. Since DSBs are formed at multiple sites within a hotspot, multiple labelled fragments could be seen in lanes 3 and 5 on each panel. After digestion with restriction nuclease these fragments are collapsed into a single fragment with the expected length (arrows). Lanes 3/4 and 5/6 correspond to different aliquots of genomic DNA and therefore show different pattern of DSB sites.



Supplementary Figure 6. Low resolution correlations between DSB hotspot map and published genetic maps. In-hotspot ChIP-seq tag coverage on chromosome I was compared to the published genetic maps of Paigen et al.[15] (blue line) and Cox et al. [16] (black line). The comparisons were performed using non-overlapping window sizes from 0.25 to 25 Mb in 0.25 Mb increments. The Spearman correlation coefficient (R) was calculated between pairs of maps at each resolution and trend lines were generated using a sliding window of 2.5 Mb with a step of 0.25 Mb. The correlation between the two published genetic maps is also shown (red line). The dashed magenta line represents the comparison of the DSB map and genetic map B across the whole genome. In all cases, the correlation between maps increases up to a window size of between 7 and 10 Mb. The R values then plateau. The dip in R values when using the largest window sizes is not apparent in the whole genome comparison, demonstrating that this dip is probably an artefact of having few data points at such low resolution.

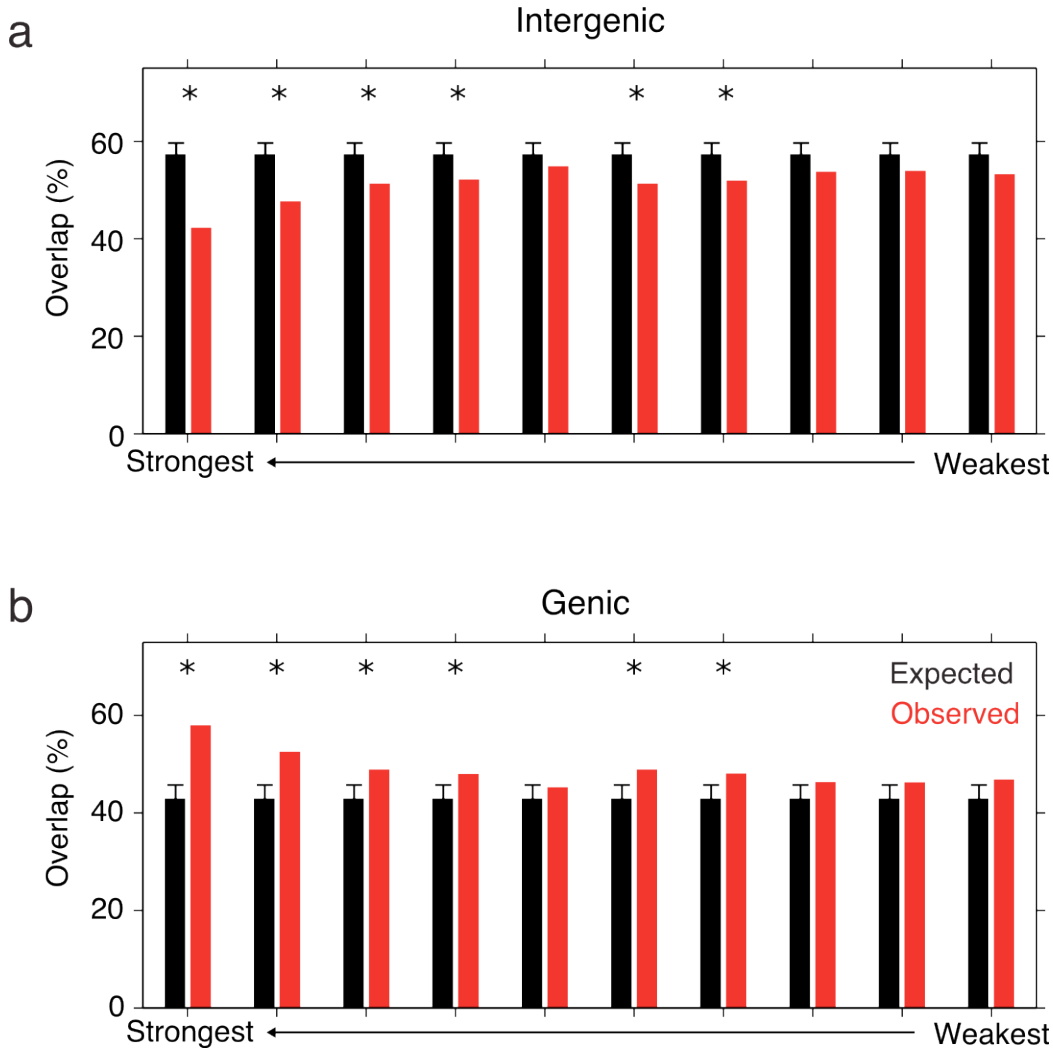


c

Recombination deserts

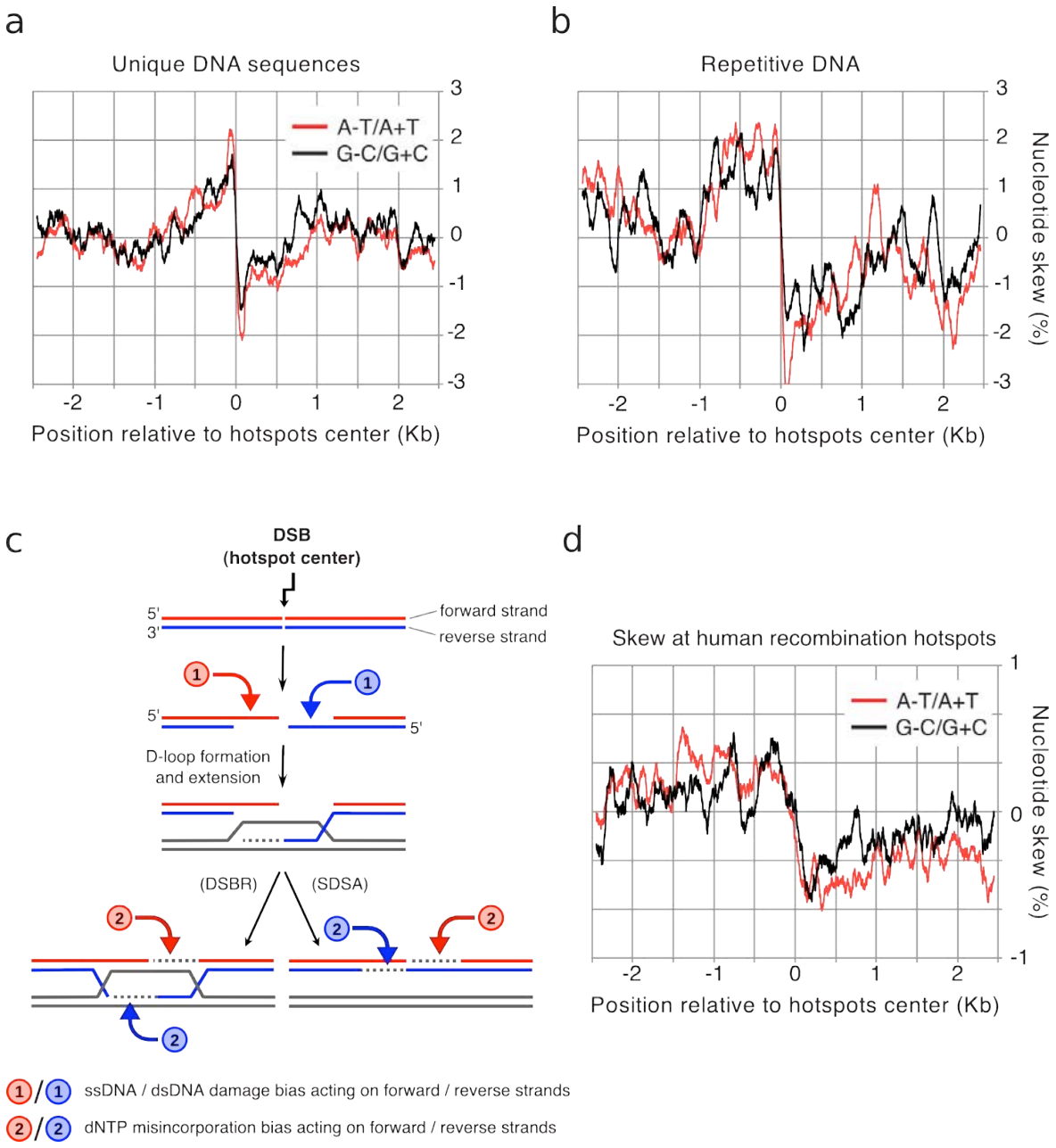
Chromosome	Start	End	Size (Mb)	Gaps (%)
chr17	17,580,258	20,695,178	3.12	0
chr8	55,125,619	58,241,425	3.12	54
chrX	36,087,347	39,409,131	3.33	0
chr2	174,710,150	178,055,666	3.35	54
chrX	62,981,786	66,380,091	3.40	0
chr17	37,158,896	40,622,530	3.47	0
chrX	75,522,490	79,199,381	3.68	0
chr7	20,483,515	24,250,453	3.77	87
chr13	84,053,101	87,847,380	3.80	0
chr8	19,204,190	23,001,081	3.80	54
chr7	31,495,052	35,297,688	3.81	11
chrX	58,107,894	62,127,774	4.02	0
chr12	21,233,942	25,290,393	4.06	50
chr3	41,473,489	45,575,717	4.11	0
chr1	26,174,232	30,385,149	4.21	0
chrX	142,958,957	147,482,219	4.53	43
chrX	84,899,119	89,756,272	4.86	0
chr7	7,254,460	12,318,602	5.07	54
chr7	39,289,425	46,407,254	7.12	98
chrX	115,772,528	124,419,047	8.65	14
chrX	21,813,231	33,417,463	11.61	53

Supplementary Figure 7. Inter-hotspot distances. **a**. Distribution of inter-hotspot distances below 4Mb. Steps are unevenly sized as data were binned by $\log_{10}$(distance). **b**. Inter-hotspot distance distribution by chromosome. The blue box delimits the 25th to 75th percentiles. The red line inside this area represents the median of each distribution. Outliers are represented as red + symbols. **c**. Recombination deserts larger than 3Mb. Some recombination deserts may contain hotspots, which we were unable to detect due to sequencing gaps or highly repetitive DNA. The fifth column (Gaps (%)) represents the proportion of such regions within each recombination desert.



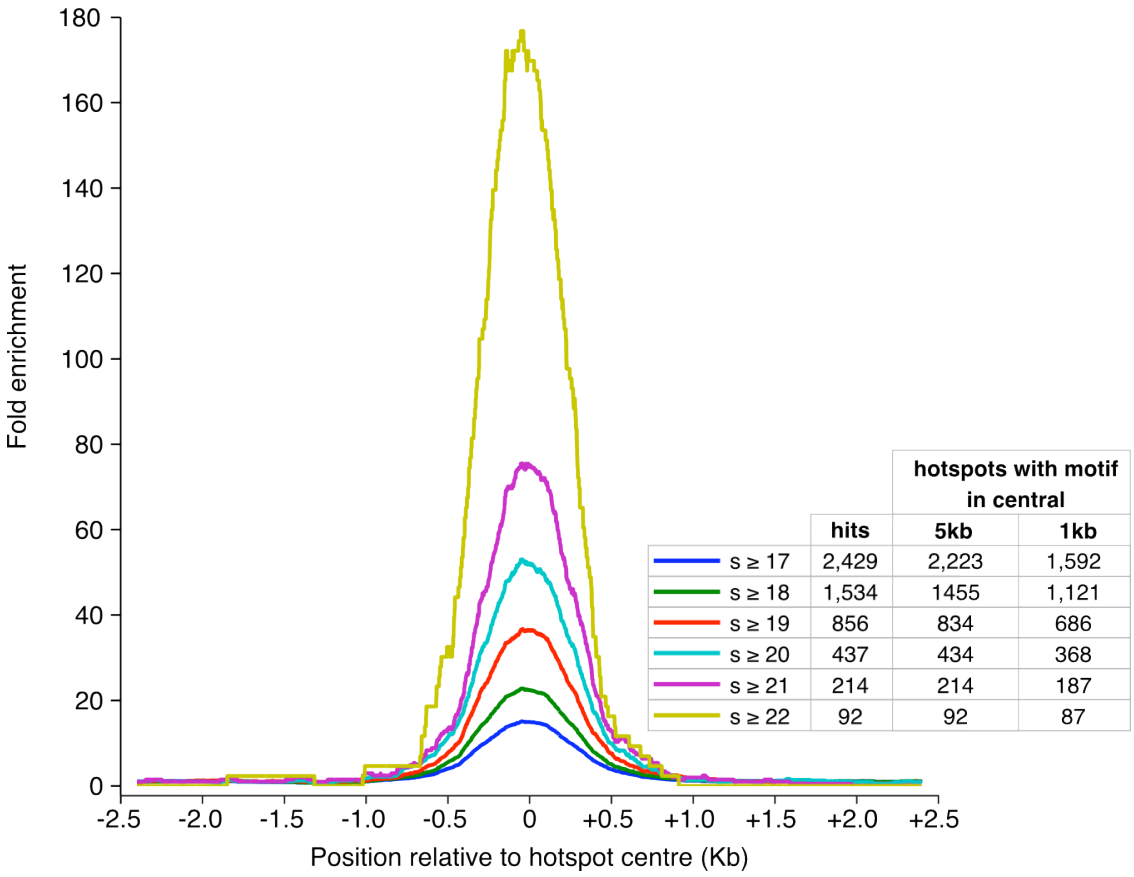
Supplementary Figure 8. The preferred genomic context of mouse recombination initiation hotspots is strength dependent. Mouse recombination initiation hotspots were stratified by strength into ten bins. The strength was calculated as the background-subtracted number of ChIP-seq tags within a hotspot. The frequency of hotspots in (a) intergenic and (b) genic regions was then calculated for hotspots in each bin. Observed frequencies are shown as red bars, median expected frequencies (from 10,000-fold bootstrapping – see Supplementary Methods) as black bars. Error bars represent the 5th to 95th percentiles of the expected value distributions.

Bins with a significant deviation from expectation are highlighted with asterisks. All P-values are $< 10^{-4}$ and derived using a one-sided binomial test.

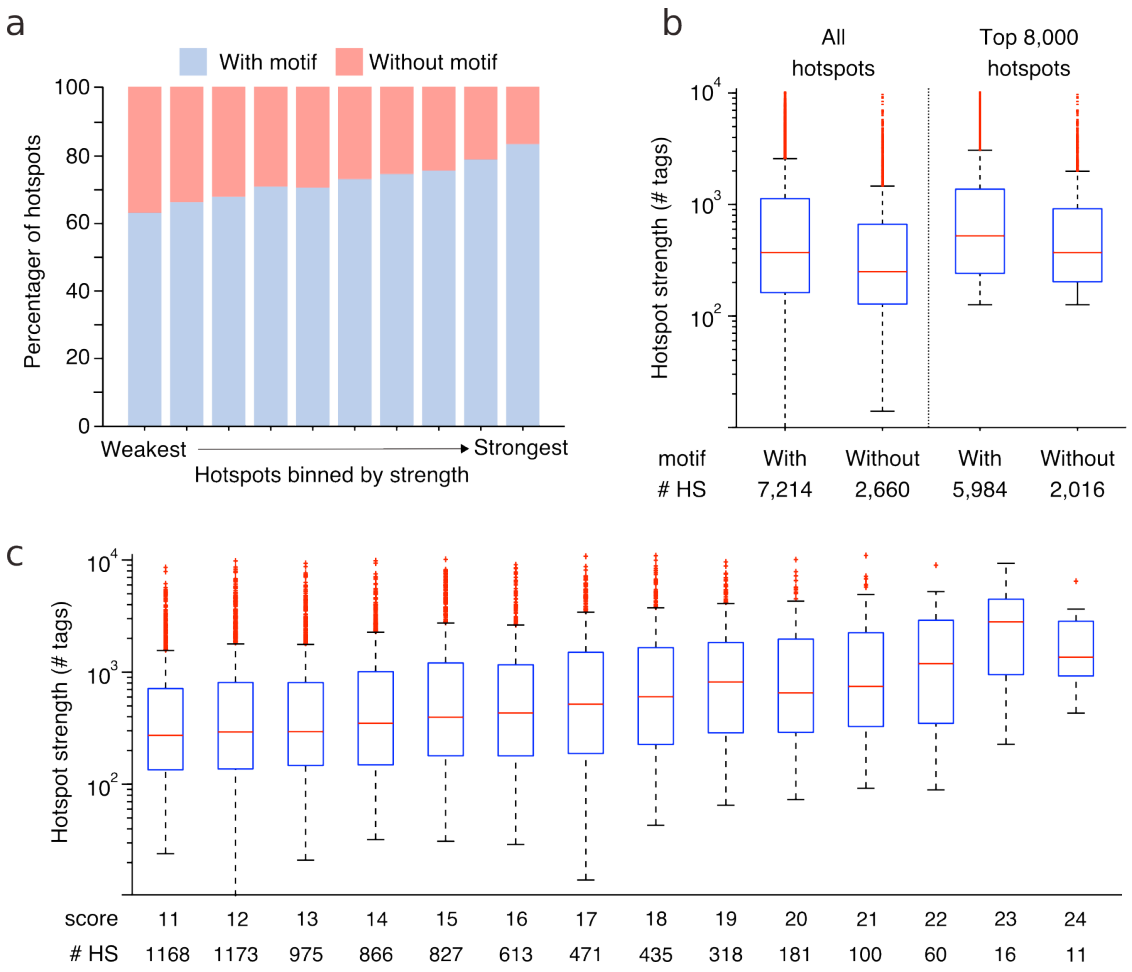


Supplementary Figure 9. Purine/pyrimidine skew in the DSB hotspots. The mean nucleotide skew in unique (**a**) and repetitive (**b**) DNA sequences is plotted for the forward DNA strand for all hotspots aligned at the centre. Mouse chromosomes are acrocentric, the forward DNA strand corresponds to the strand in 5' to 3' direction from the centromeric end. Repeat sequences were defined as all sequences in the RepeatMasker track[39] for the mm9 (NCBI37) version of the mouse genome assembly and

were downloaded from the UCSC genome browser. **c.** Schematic of *hypothetical* events that may give rise to the observed nucleotide skew during meiotic recombination. According to Chargaff's second parity rule[40] the frequency of complementary bases in a long stretch of ssDNA is equal (A=T and G=C). Deviations from this rule (nucleotide skews, also called strand asymmetries) are attributed to DNA mutation/repair biases acting asymmetrically on different (forward and reverse) strands (reviewed in Ref.[19]). (1) The ends of DSBs are nucleolitically processed to have long 3' ssDNA[41] tails. Since single-stranded DNA is more susceptible to certain types of DNA damage compared to double-stranded DNA, ssDNA-biased mutations will affect the immediate area around a DSB site. The same bias would act on the forward strand 5' to the centre of hotspots and the reverse strand 3' to the centre of the hotspot. Therefore, when a DNA sequence is analyzed along either of the strands the strand asymmetry will change polarity at the centre of hotspots. (2) Deoxynucleotide misincorporation bias might occur during D-loop extension and/or gap repair synthesis and will act on opposite (forward or reverse) strands on different sides of a DSB. The same bias acting on opposite strands will lead to a flip in the strand asymmetry at the centre of hotspot when either of the strands is examined. **d.** The purine/pyrimidine skew is also present in human hotspots. Mean nucleotide skew for ~32,000 human recombination hotspots defined in[8]. The skew magnitude is less pronounced than in the mouse, likely due to the lower resolution of the human map. Figures a, b and d are plotted with a sliding window of 100 bp in 1 bp steps.



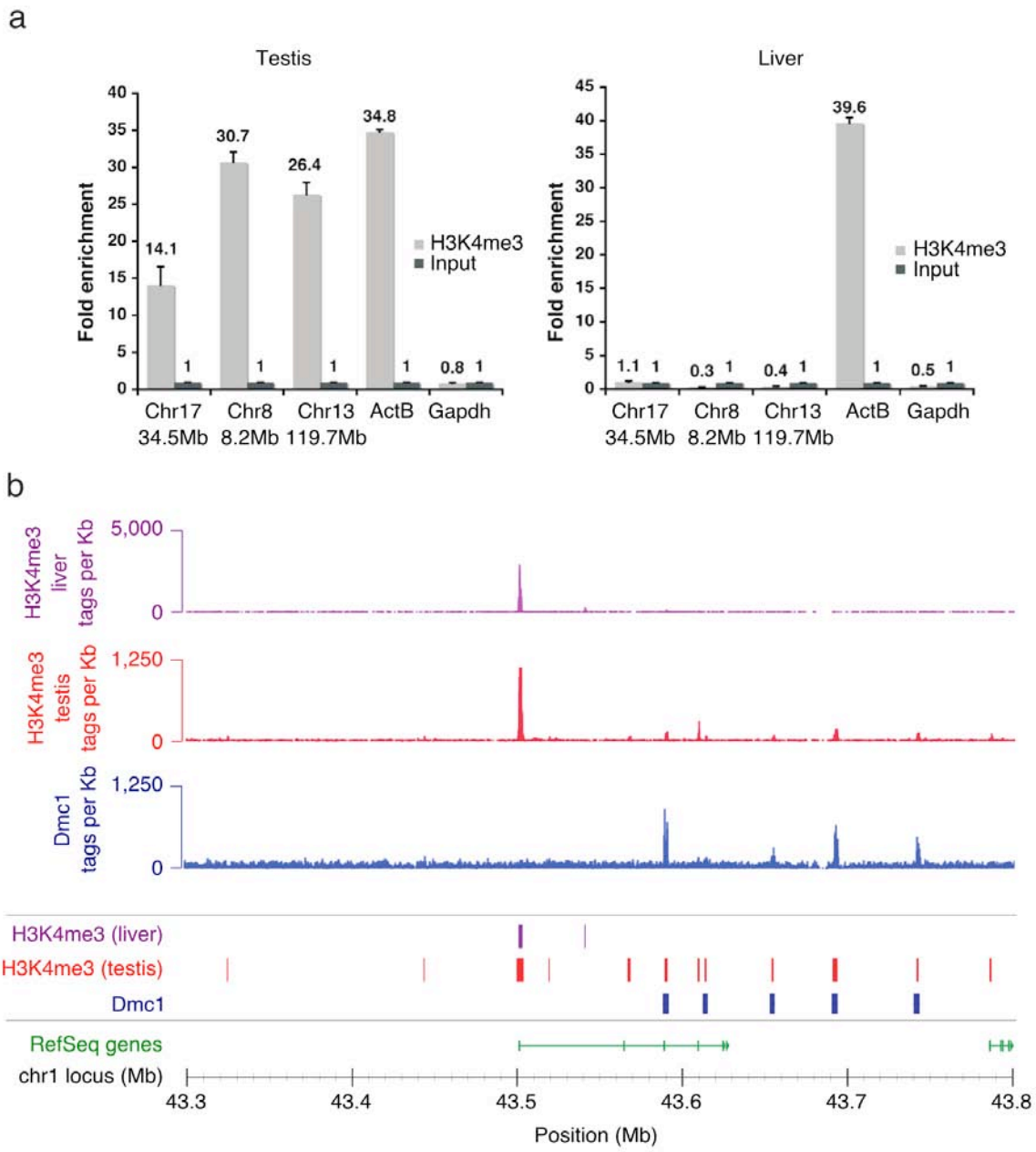
Supplementary Figure 10. Best scoring motifs are over-represented at hotspots almost 180-fold. The spatial distribution around hotspot centres was generated for hits to the position specific scoring matrix (PSSM) of the motif at alignment score thresholds from 17 to 22 (scores were rounded down to the closest integer). The number of hits within hotspot regions at or above each score is shown ("hits"). The number of distinct hotspots containing a motif in the central 5kb or central 1kb is also indicated. The fold enrichment in hotspots increases dramatically at higher alignment scores. Distributions are generated with a 500 bp sliding window in single nucleotide steps. Enrichment is given relative to the average genomic motif frequency.



Supplementary Figure 11. Hotspot strength correlates with motif conservation. Strength is calculated as described in the Supplementary Methods. Only hotspots with the motif in the central 1Kb are considered to contain a copy of the motif. **a.** The proportion of hotspots containing the consensus motif increases with hotspot strength. Hotspots were stratified by strength into ten bins. Blue represents the proportion of hotspots with a motif in each bin, red represents those without. **b.** Distribution of hotspot strength for hotspots with and without a hit to the motif PSSM. Hotspots containing the motif are significantly stronger, both when we

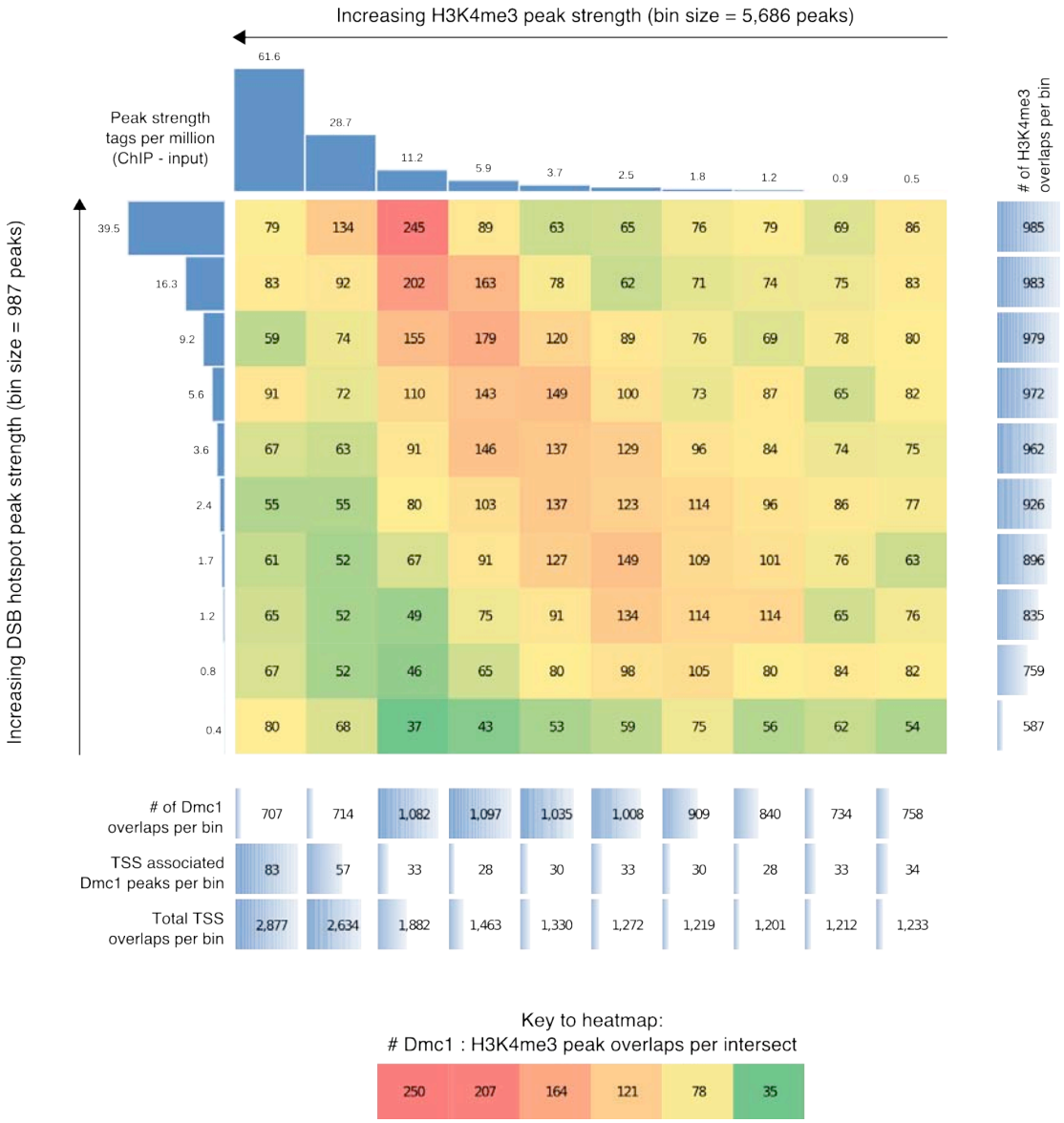
consider (i) all hotspots ($P = 2.4 \times 10^{-36}$, Wilcoxon test) and (ii) when we consider only the top 8000 hotspots in order to minimize the contribution of false positives ($P = 1.2 \times 10^{-22}$, Wilcoxon test). **c.** Hotspot strength is positively correlated with the quality of the motif alignment within hotspots. The quality is given as the best motif alignment score in the central 1 Kb.

Supplementary Figure 12. Alignment of the *Prdm9* alleles in C57BL10.F and C57BL10.S mice used in this study. Each zinc finger in the C-terminal zinc-finger array is shown as a magenta bar. A further zinc finger domain, N-terminal to the array is not shown. The three DNA contact residues used for binding site prediction are shown in red for each zinc finger domain. Sequence conservation between the two alleles is indicated by the yellow bar. There are a total of 13 amino acid differences between these alleles.



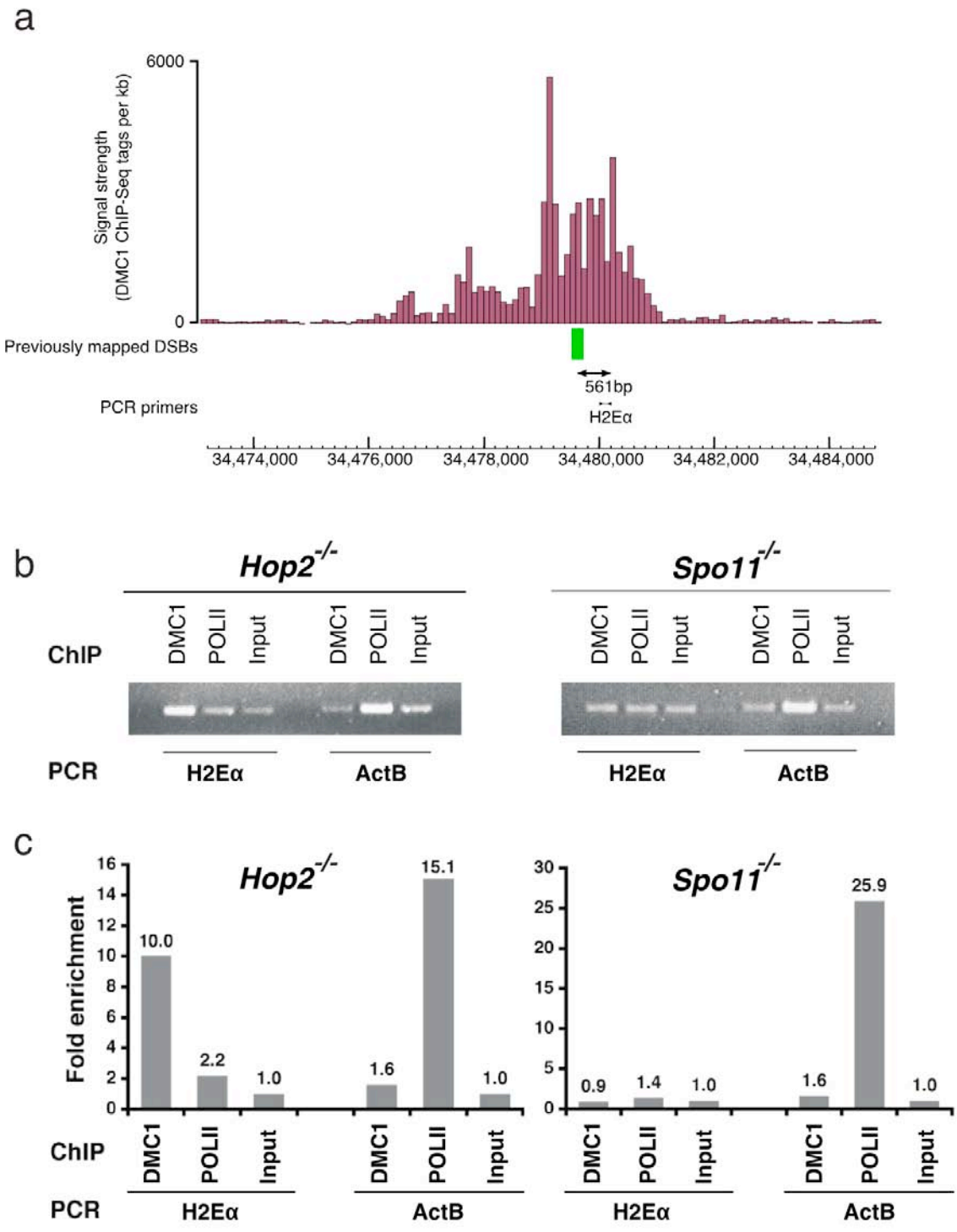
Supplementary Figure 13. DSB hotspots overlap with H3K4me3 marks. **a**. Enrichment of the hotspot DNA by anti-H3K4me3 ChIP. ChIP was performed from testis or liver extracts and the enrichment of the DNA corresponding to several hotspots (Chr17, Chr8, Chr13), the beta actin promoter, and the Gapdh coding region was estimated by qPCR. Quantitative data are expressed as the ratio of the enriched (ChIP) to the input DNA. **b**.

DSB hotspots are associated with a subset of H3K4me3 marks present in testis but not in liver samples. Vertical bars in the middle panel represent peaks for each dataset that were identified as described in Supplementary Methods. The region shown is a representative 0.5 Mb region of chromosome 1.



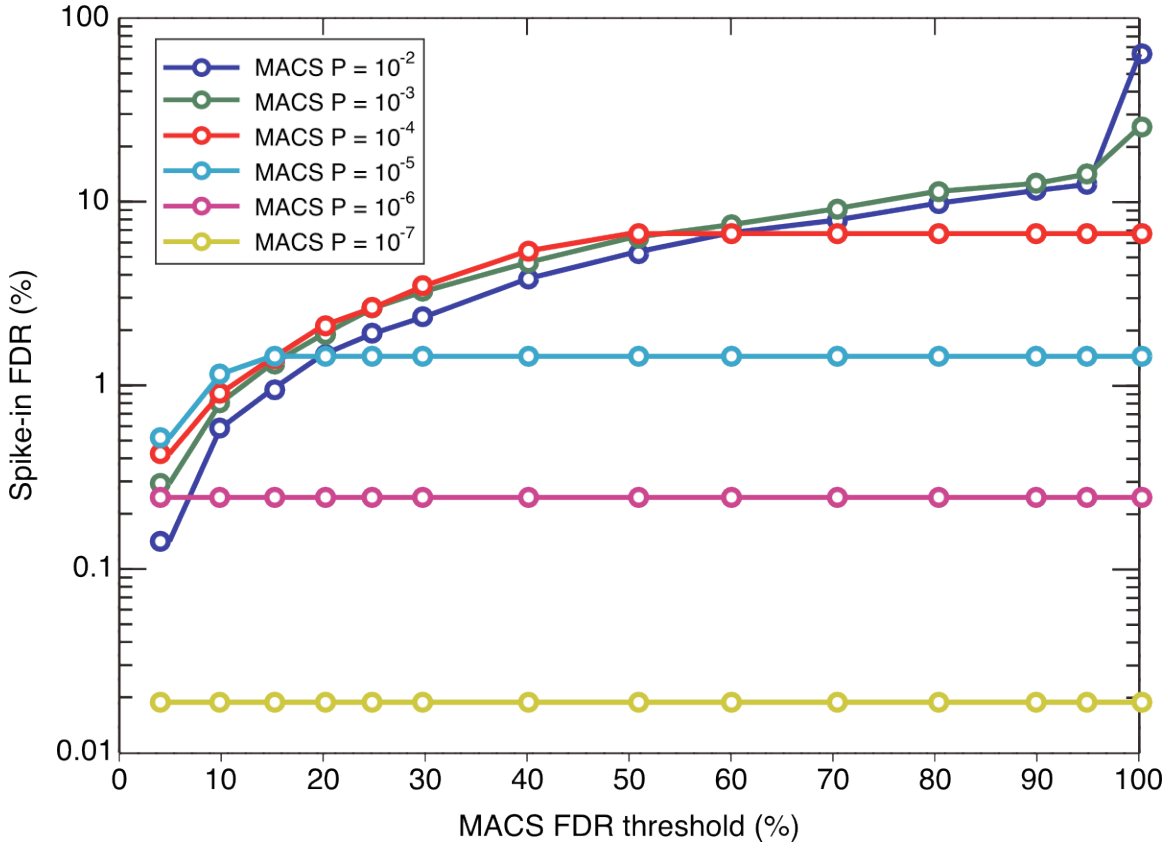
Supplementary Figure 14. DSB associated H3K4me3 marks are weaker than and distinct from those associated with TSS. Both DSB hotspots and H3K4me3 marks were divided into 10 bins by strength. The heat map values represent the number of overlaps between hotspot centers and H3K4me3 mark. DSB strength is correlated with H3K4me3 strength, though not with the strongest H3K4me3 marks. The strongest 20% of H3K4me3 are clearly associated with TSS. It should be noted that while 389 hotspots overlap TSS, they are frequently not co-centered and that these numbers are artificially

inflated because we cannot resolve adjacent H3K4me3 peaks. The right-most column is a sum of the values across each row, and shows that there is a correlation between hotspot strength and the proportion of hotspots that overlap an H3K4me3 mark. This again suggests that the 5.9% of hotspots that do not overlap an H3K4me3 mark (Fig. 3a) may be a consequence of undetected, weak H3K4me3 peaks.



Supplementary Figure 15. Previously known H2Eα hotspot[14,18,38] was used to monitor enrichment of hotspot DNA during ChIP. **a.** The position of the H2Eα hotspot and the primers used for quantitative PCR are shown. 561 bp indicates the distance from the

centre of the previously mapped hotspot region[14] to the qPCR fragment. DMC1 ChIP-Seq tag density obtained in our study is shown in 100 bp non-overlapping windows. **b.** Evaluation of the H2Eα DNA enrichment by regular PCR. Anti-DMC1 or anti-RNA polymerase II (Pol II) antibodies were used for ChIP from *Hop2*$^{-/-}$ or *Spo11*$^{-/-}$ testes as indicated. The recovered DNA was subjected to PCR amplification by H2Eα-specific or beta actin promoter-specific primers as indicated (DNA concentration was normalized). Pol II is expected to bind the promoter region of the actin B and therefore is used as a positive control for the ChIP procedure. *Spo11*$^{-/-}$ mice are defective in DSB formation and therefore are used as a negative control. **c.** ChIP samples described in b were analysed by qPCR. Quantitative data are expressed as the ratio of the enriched (ChIP) to the input DNA.



Supplementary Figure 16. Peak finding false discovery rates as determined by spike-in analysis. Analysis of peaks identified in a simulated peak set reveals that MACS FDR cutoffs are very conservative. See Materials and Methods for spike-in FDR assessment.

Supplementary Table 1: The association between repeats and known human hotspot-associated features with mouse recombination initiation hotspots. The expected overlaps were derived from 10,000 randomized sets of hotspots. P-values calculated using a one-sided binomial test.

Repeat	Repeat parent class	Hotspots with repeat	Deviation from random	# in genome	P-value
Alu	SINE	50.4 %	19.7 %	559,629	10^{-15}
MaLR	LTR	42.9%	20.0%	419,682	10^{-45}
B4	SINE	42.6 %	20.6 %	393,655	10^{-47}
B2	SINE	39.0 %	20.6 %	366,142	10^{-41}
MIR	SINE	16.6 %	21.1 %	121,596	10^{-13}
MER1	DNA	14.3 %	24.3 %	101,357	10^{-14}
ERVL	LTR	13.6 %	17.1 %	109,459	10^{-6}
ID	SINE	10.7 %	34.3 %	63,564	10^{-18}
L2	LINE	9.4 %	26.2 %	65,276	10^{-9}
$(T)_n$	Simple repeat	4.4 %	23.6 %	13,977	10^{-4}
L1	LINE	47.7 %	-18.2%	861,599	10^{-96}
ERVK	LTR	15.0 %	-25.2 %	235,172	10^{-35}

Feature	Details	Hotspots with feature	Deviation from random	# in genome	P-value
CCTCCCT	Human hotspot motif [11]	59.7 %	9.4 %	797,200	10^{-21}
CCNCCNTNNCCNC	Human hotspot motif [28]	38.1 %	3.0 %	413,496	ns
GC content	100Kb regions with > 45% GC	25.6 %	21.1 %	5,259	10^{-26}
CCCCACCCC	Human hotspot motif [11]	20.6 %	11.0 %	176,492	10^{-4}

MATERIALS AND METHODS

Mouse strains

9R (alternative name C57Bl/10.S, Jackson Labs stock number 001650) and 13R (C57Bl/10.F, Jackson Labs stock number 001818) have been received from Dr. N. Arnheim, University of Southern California. *Hop2* heterozygous mice [13] have been backcrossed for at least 10 generations to produce *Hop2* heterozygotes on 9R or 13R genetic backgrounds. 9R and 13R *Hop2* heterozygotes were bred to produce *Hop2*$^{-/-}$ mice on a 9R/13R F1 hybrid background. 9R/13R F1 hybrids were chosen because of the presence of the known recombination hotspot on this genetic background (*H2Eα*, [38]) This was necessary for use as a positive control. Unless stated otherwise, all experiments were performed using adult (2-6 months old) *Hop2*$^{-/-}$ mice on a 9R/13R F1 background. To produce wild type mice 9R and 13R mice were bred and F1 pups were collected at 12 dpp. *Spo11*$^{-/-}$ mice have been previously described [42]. All animal procedures have been approved by the USUHS Animal Care and Use Committee.

Antibodies

The following antibodies were used: anti-DMC1, Santa Cruz (C-20, sc 8973), anti-RAD51, Santa Cruz (H92, # sc 8349), anti-RNA polymerase II, Millipore (clone CTD4H8, # 05-623), anti-H3K4me3, Millipore (#07-473), normal goat or rabbit IgGs, Santa Cruz (#sc-2028 and #sc-2028, respectively).

Chromatin Immunoprecipitation (ChIP)

Testes from three adult or 10 juvenile mice were de-capsulated and fixed for 10 min in 1% formaldehyde (only two adult mice were required for the H3K4me3 ChIP). After quenching the tissue was homogenised and cell suspension was prepared by filtering

through 40 μm cell strainer, and washing in the following buffers: 1) PBS (twice); 2) 0.25% Triton X100, 10mM EDTA, 0.5mM EGTA, 10mM Tris pH8; 3) 0.2M NaCl, 1mM EDTA, 0.5mM EGTA, 10mM Tris pH8. Cells were lysed in 1.5 ml of the lysis buffer (1% SDS, 10mM EDTA, 50 mM TrisCl pH8 plus complete protein inhibitor cocktail from Roche) and the chromatin was sheared to ~1000 bp (~500 bp in case of H3K4me3 ChIP) by sonication. The sample was diluted twice and dialyzed against ChIP buffer (0.01% SDS, 1.1% Triton X100, 1.2mM EDTA, 16.7mM TrisCl , 167 mM NaCl). Chromatin was pre-cleared with Protein G beads (Sigma) and 1/2000 of pre-cleared material was saved and referred to as Input. The rest of the chromatin was incubated with appropriate antibodies overnight at 4C followed by a 2h incubation with Protein G beads. Beads were washed 5 times in the following buffers: 1) 0.1% SDS 1% Triton X100, 2mM EDTA, 20 mM TrisCl, 150 mM NaCl; 2) 0.1% SDS 1% Triton X-100, 2mM EDTA, 20mM TrisCl pH8, 500 mM NaCl; 3) 0.25M LiCl, 1% Igepal, 1mM EDTA, 10 mM TrisCl, pH8, 1% Deoxycholic acid; 4) TE (twice). The chromatin was eluted by 1% SDS, 0.1M $NaHCO_3$ pH9 at 65C and crosslinking was reversed at 65C overnight. DNA was deproteinized for 2h at 45C and DNA was purified with a MinElute Reaction Clean up kit (QIAGEN). The DNA concentration was measured with a Quant-IT Picogreen kit (Invitrogen). ChIP efficiency was evaluated by enrichment of the DNA corresponding to the known hotspot H2Eα[14,18,38] (Supplementary Fig. 15).

Quantitative PCR (qPCR)

Real Time PCR was performed using a Maxima PCR SYBR Green/Rox Kit (Fermentas). Reactions were performed in a 7500 Real-Time PCR machine (Applied Biosystems) according to the manufacturer's instructions. Gene copy number was calculated with ABI

SDS Software Version 1.3.1. Enrichment was calculated as the ratio of gene copy number in ChIP DNA compared to Input. Primers for DSB confirmation are located within 500 bp from the centre of hotspots. Primers with chromosome coordinates are summarized in table below.

	Forward 5'→ 3'	**5' end coordinates**	**Reverse 5'→ 3'**	**5' end coordinates**
H2ea	GCAAGAGCAGTCAGTGTTCTTAACGAA	34480167	GAGGAGATGTCAGTTGATGTTGTAGGA	34480054
Chr5	ATAGATTCATAGCAAGGGTTCCTAC	81929144	CAAATGCATGCCTGAAGTAAG	81929242
Chr8	GAGAGACAGTGTGAAAGAGCCCTAT	8221339	TGGGTTGTTGATTCTCCGTTAC	8221438
Chr11	TGCTGTGTCCACACTACCGTATTCT	86114735	GCAGACCTCCACAAAATGTACCC	86114842
Chr13	GGTAAGTGGGGAATCTGTTGCAT	119699617	GGCCAGCCTAGAATTTTCTACACAT	119699729
Chr14	ATATATAGTCCAGGATTTGAACTGTCG	98235132	CATACAATTTTGGGCAAGACTGA	98235241
ActB/ chr5	CACCCATCGCCAAAACTCTTCATCCT	143665170	CGCACAGTGCAGCATTTTTTTACC	143665301
Gapdh/ Chr6	GCTCACTGGCATGGCCTTCCGTG	125113572	TGGAAGAGTGGGAGTTGCTGTTGA	125113289

Confirmation of Individual Double Stranded Breaks

Testes of adult (3 months old) wild type mice were de-capsulated and incubated in 10 ml of RPMI media with 800 u of type II collagenase and 25mM $CaCl_2$ for 25 min at 32C with gentle rotation. 200µl of 0.5M EDTA was added followed by 25 min incubation at the same conditions. The cells were disaggregated by pipetting and filtered through a 40μm cell strainer. The rest of the procedure was as previously described [14]. Briefly, cells were mixed with equal volume of the low melting agarose (Lonza) to a final concentration of 10^4 cells/µl and pipetted into plug molds (Bio-Rad). Plugs were incubated at 50°C for 40 to 48 h in a lysis solution (0.5 M EDTA pH8, 1.0% Sarkosyl,

100 µg/ml of proteinase K). Plugs were washed 15 times with TE buffer, incubated in TdT buffer at 4°C for 2h, then transferred to 300 µl of fresh TdT buffer containing 300 U of terminal deoxynucleotidyl transferase (NEB). Plugs were incubated on ice for 15 hours and then at 37°C for 2 hours. The plugs were washed with TE buffer, and DNA was extracted by DNA Gel extraction kit (Qiagen). The DNA was eluted in 75 µl of elution buffer and stored at 4°C. Concentration of DNA was estimated with Picogreen kit (Invitrogen). 10 ng of DNA (equal to $1.7x10^3$cells) was used for PCR. The primers used are summarized below. The restriction enzymes are indicated in Supplementary Fig. 5.

Primer C+P: GTTAACCGCAACGTACCGTTGTTTGAGCAGGCCCCCCCCCC

Primer P: GTTAACCGCAACGTACCGTTGTTTGAGCAGG

	Primer R1		**Primer R2**	
Hot spot	Sequence	5' Co-ordinate	Sequence	5' Co-ordinate
Chr11	GTTAATCCCAAGACTTACCAATGC	86115248	ACATATAAGCGAATAATACTGTGTGATCA	86115122
Chr13	CAGTGAGGAGCTCAAGAAAACTATG	119700482	GGTTATCCTGAATAGTGTGTGCTTAG	119700382

Micrococcal nuclease digestion and sample preparation

Cells were prepared from *Hop2*$^{-/-}$ mice (9R/13R F1 background) as in ChIP protocol except for EDTA and EGTA were omitted from the buffers. The pellet was re-suspended in 600 µl of the buffer containing 50 mM Tris-HCl and 5 mM CaCl2 and kept at room temperature for 10 min. Cells were sonicated for 5 seconds at level 1 and treated with 600 gel units of micrococcal nuclease (NEB) for 10 min at 25C. The reaction was stopped by the addition of 24 µl of 0.5 M EDTA and insoluble material was removed by centrifugation. DNA was prepared as in ChIP protocol and 100 ng was used for the library construction. Randomly fragmented chromatin was prepared by sonication to produce ~500 bp fragments.

ChIP-seq Library Construction and Sequencing.

Library construction was done with the ChIP DNA sequencing kit (Illumina) according to the manufacturer's protocol with the exception that DNA size selection was done after the PCR step. Sequencing was done with single-read or paired-end cluster generation kits and 36-cycle sequencing kits (Illumina) according to the instructions provided.

Sequencing and sample description

The Illumina GAII Genome Analyzer was used to perform massively parallel sequencing in either Single Read or Pair End mode. When Paired End mode was used, only the first read was used. Tags were sequenced to a total length of 36 bases and were aligned to the mouse mm9/NCBI37 genome using ELAND versions 1.4, 1.5 or 1.6 [43]. Only tags that passed the quality filter and mapped uniquely to the genome were used. We pooled all tags for replicates of each ChIP sample or control, to generate working datasets. The total number of tags derived from anti-DMC1 ChIP (*Hop2*$^{-/-}$ mice) and controls is approximately 107 and 120 million (M), respectively. The anti-DMC1 ChIP pool comprises four biological replicates consisting of 83 M, 5 M, 9 M and 10 M mapped sequence tags (Fig. 1c, replicates A through D). A fifth biological replicate was also prepared (Fig. 1c, replicate E) but it was removed from further analysis because it had the lowest signal to noise ratio. The control comprises tags from both anti-IgG ChIP and input DNA without ChIP. The two biological replicates of the anti-RAD51 ChIP (*Hop2*$^{-/-}$ mice) (Fig. 1c, samples F and G) are made up of 15 M and 9 M reads. The anti-DMC1 ChIP from *Spo11*$^{-/-}$ mice consists of 5 M reads (Fig. 1c, sample H). In addition, two anti-DMC1 ChIP samples were prepared from wild type mice producing 16 M and 4 M tags (replicates J and K). Tag datasets are fully described and available for download from the

National Centre for Biotechnology Information (US) GEO, accession number GSE24438.

A summary of the tag sets is provided below.

ID	Sample	Strain	# tags	Accession
A	anti-DMC1	*Hop2*$^{-/-}$	83M	GSM602190
B	anti-DMC1	*Hop2*$^{-/-}$	5M	GSM602191
C	anti-DMC1	*Hop2*$^{-/-}$	9M	GSM602192
D	anti-DMC1	*Hop2*$^{-/-}$	10M	GSM602193
E	anti-DMC1	*Hop2*$^{-/-}$	55M	GSM602195
F	anti-RAD51	*Hop2*$^{-/-}$	15M	GSM602196
G	anti-RAD51	*Hop2*$^{-/-}$	9M	GSM602197
H	anti-DMC1	*Spo11*$^{-/-}$	5M	GSM602198
J	anti-DMC1	wild type	16M	GSM602200
K	anti-DMC1	wild type	4M	GSM602201
Control	input DNA& IgG	*Hop2*$^{-/-}$	120M	GSM602199

Tag datasets for anti-H3K4me3 experiments are fully described and available for download from the US National Centre for Biotechnology Information GEO, accession number GSE24438. A summary of the tag sets is provided below.

Tissue	Sample	Strain	# tags	Accession
testis	anti-H3K4me3	*Hop2*$^{-/-}$	36M	GSM602204
testis control	input DNA	*Hop2*$^{-/-}$	39M	GSM602205
liver	anti-H3K4me3	wild type	40M	GSM602206
liver control	input DNA	wild type	32M	GSM602207

Micrococcal nuclease digestion experiments were all run as paired-end. Tag datasets are fully described and available for download from the US National Centre for Biotechnology Information GEO, accession number GSE24438. A summary of the tag sets is provided below.

Sample	Strain	# tags	Accession
MNase digestion	*Hop2*$^{-/-}$	68M	GSM603056
Sonication	*Hop2*$^{-/-}$	69M	GSM603057

Genomic loci represented by an artificially high number of stacked tags where removed from analysis. To determine these over-represented loci, we examined the pooled control that is the largest and most evenly distributed tag set. The nucleotides in this dataset at which 100 or more tags originated on the same strand were identified as offending, and tags derived from these loci were not considered in subsequent analyses (2,040 such positions were detected). This pruning process does not (and it is not intended to) correct for copy number variations between the genomes of the mice used in this study and the current mm9/NCBI37 genome assembly. R/Bioconductor tools and BEDTools were used for locus identification and pruning, respectively[44,45].

Peak calling

Identification of the DSB hotspots and H3K4me3 marks was performed using MACS v1.3.7 [46] with a matching number of tags in treatment and control. Random subsetting of the larger data set was performed to match the number of tags. Peak calling for DMC1 ChIP samples was performed with a P-value cut-off of 10^{-4} and a bandwidth of 800 bases. From the output of the peak finder, only the peaks with a reported FDR of 50% or less were kept. Peak calling of H3K4me3 was also performed using MACS (P-value cut-off =

10^{-5}, bandwidth = 300nt, MACS FDR threshold = 10%). Unlike H3K4me3 ChIP samples, DMC1 and RAD51 ChIP samples were much noisier. This affects the accuracy of the peak finder in determining the boundaries and thus, the centre of the hotspot. To circumvent this problem the peaks were re-centered. We make the assumption that resection will occur symmetrically around DSB sites. We identified each peak centre as the median of the centre of the sequenced fragment locations within the MACS defined peaks. Fragments were assumed to be 120 bp long. We then defined the peak boundaries as the centre ± 1.7Kb (the average width of DMC1 peaks, as calculated by MACS, was 3.4Kb). The list of hotspots reported in this paper is the product of such peak re-centering (Supplementary File F1). Peaks that overlapped as a result of the 1.7Kb extension were merged. The False Discovery Rate (FDR) associated with calling multiple peaks was addressed in two stages. First we simulated a ChIP-seq data by spike-in analysis allowing us to correlate the MACS-estimated FDR with the true FDR (see below). We then filtered peaks using a MACS-estimated FDR threshold of 50% (corresponds to an actual FDR of 6.7%).

Spike-in FDR assessment

It has been shown in unpublished work that the false discovery rate (FDR) reported by MACS is overestimated up to 10 fold. We wished to assess the true FDR rate for our data. To do this, we generated spike-in data using the Simulator algorithm from the USeq package [47]. A logarithmic fit was performed on the strength distribution of DSB hotspots and spike-in tags were generated to reflect points on this strength distribution. This yielded 2,583,657 spike-in tags derived from 10,650 simulated peaks. Tags from our control data pool were added to the spike-in tags in order to reflect the treatment : control

ratio in the real data. We then used MACS to call peaks in this dataset compared to a tag-matched control. We ran MACS using a range of P-value parameters and assessed the true FDR at different MACS FDR thresholds (Supplementary Fig. 16). Using a P-value of 10^{-4} (as we do for true peak calling) we can see that MACS is indeed, very conservative. These analyses prompted us to use a MACS FDR of 50% for peak calling, as the true FDR for these peaks is 6.7%.

Modelling a random DSB hotspot distribution

In order to assess the enrichment of hotspot DNA in functional categories, we generated a representative background upon which to model a random hotspot distribution. We generated a mappability score for each nucleotide in the genome using the gem-mappability algorithm[48]. This allowed us to mask regions of the genome into which randomized hotspot centres could not be placed. The mappability index for each nucleotide is equal to the inverse of the number of occurrences in the genome of the 32 bp oligo beginning at that locus (with up to 2 mismatches). Only nucleotides with a score of one are mappable by our method, which excludes non-uniquely mapping reads. We therefore re-scored each nucleotide as 1 or 0 by this criterion. The average score in each non-overlapping 1 Kb bin across the genome was then calculated. Bins with a score < 0.5 were masked as were all sequencing gaps. All masked regions were then expanded by 1 Kb at either side. Single nucleotide hotspot centres were subsequently randomly distributed in non-masked regions of the genome.

Estimating the precision of hotspot centre identification

In order to define hotspot centres, we assume a model where strand resection is symmetrical around the DSB site. Thus, the median of the tag distribution within a peak

is representative of the likely DSB site. We randomly assigned each DMC1 ChIP-seq tag to one of 20 subsets. Using the same median re-centering method as for the whole dataset, we identified the centres of MACS defined hotspots using each of these 20 tag subsets. This gave us an empirical standard error from the defined centre position for each peak. The median of these values was 179 nt. This measure is referred to as the precision of DSB hotspot mapping.

Peak strength estimation

We sub-set the larger of the control pool and ChIP-Seq pool tags to match the number of tags in the smaller pool. Subsequently peak strength was calculated by subtracting the in-peak control tag count from the in-peak ChIP-Seq tag count.

Motif finding

The frequency of all 9-mers in DSB hotspots was calculated. The expected frequency of each 9-mer was calculated as the mean number of occurrences in 10,000 randomized sets of DSB hotspots (uniformly distributed in non-gap DNA). A Bonferroni corrected p-value for each 9-mer was calculated using a one-sided binomial test.

We next generated consensus motifs from statistically over-represented 9-mers ($p < 10^{-3}$) with more than 10 occurrences in DSB hotspots. We calculated the distribution of mean distance to hotspot centre (d) for each over-represented 9-mer. This distribution showed an elongated tail (cutoff: $d \leq$ mean – 1 standard deviation) containing 172 9-mers found preferentially close to hotspot centers. We considered only these 9-mers for downstream analysis. ClustalW v2.0.12 [49,50] with 1,000 fold bootstrapping was used to align the 172 9-mers. From the resulting tree, we pruned all branches with bootstrapping support of at least 30% and containing at least 5 sequences. Bootstrapping in ClustalW works by

changing residues in the input sequences and then re-running the alignment. Each time a branch remains unchanged it gains bootstrapping support. While 30% bootstrapping support seems low, it is a reasonable choice here given the short length of the sequences. Each chosen sub-branch was converted into a position specific scoring matrix (PSSM), and the resultant PSSMs were clustered using MatLign [51]. The MatLign algorithm's silhouette-based criterion converged to identify two distinct motifs (M1 and M2). M1 was derived from 39 individual 9-mers, while M2 was derived from 16. We also performed this analysis with a range of bootstrapping values and using a manual inspection of the multiple alignments to identify sub-groups. Neither method resulted in noticeable changes to the ultimate motifs. We next identified all in-hotspot occurrences for each of the 39 9-mers for M1 (n = 4,201) and 16 9-mers for M2 (n = 7,201). In order to identify flanking nucleotides of potential importance, we performed an un-gapped multiple alignment for each consensus (M1, M2), using the regions from 15 bp downstream to 15 bp upstream of each in-hotspot motif occurrence.

MAST [52] was used to identify matches to the consensus motifs, M1 and M2, in hotspot sequences. The MAST score (s_m) is derived by sliding a log-odds matrix representing the motif across the genomic sequence. s_m at each position is the sum of the log-odds scores for each nucleotide at that point. We examined the distribution of MAST scores with MAST $p < 0.01$, both in hotspots and in 10,000 randomized hotspots. Positive scoring hits to both motifs were found more frequently than expected at hotspots (Bonferroni corrected, $p < 10^{-150}$, one sided binomial test). In both cases, the enrichment level of motif hits increases dramatically if only hits in the central 1kb of hotspots are considered.

The 3' end of M1 showed remarkable similarity to the 5' end of M2 alluding to the possibility that these two motifs were, in fact part of a single binding site. To examine this possibility, we checked to see if positive scoring M1 hits overlapped with M2 hits at the correct positions. Here, we considered only hits in the central 1kb of hotspots. These M1/ M2 overlaps were significantly more common than expected ($P < 10^{-250}$, binomial test) and we thus aligned the genetic sequences at all occurrences of M1/ M2 overlapping hits in hotspots to derive a single consensus motif (Fig. 2d). We generated a PSSM, and limited the flanks to the most distant positions at which a positive log-odds score was possible. We then used MAST to find hits to this PSSM within hotspots. Hits ≥ 3 were highly significant ($p < 0.001$, 10,000 bootstrapping). 7,214 hotspots contained a hit with an alignment score ≥ 11 in the central 1kb. We use this as an arbitrary score threshold as at the centre of hotspots, such motif hits are more than two-fold enriched. These motif hits are also highly enriched with respect to the genomic expectation ($P < 10^{-300}$, one sided binomial with 10,000 fold bootstrapping). The distribution of hits above this threshold and within hotspots (± 2.5kb) is shown in Fig. 2e.

Prediction of the PRDM9 binding specificity

The sequence of the PRDM9 protein from the 13Rx9R hybrid mice was obtained by sequencing the PRDM9 cDNA from the parental strains. The *Prdm9* allele of the 13R strain (GenBank accession # HQ704390) is different from the *Prdm9* allele in the 9R strain (GenBank accession # HQ704391) (Supplementary Fig. 12), which is identical to the published *Prdm9* allele in the C57Bl/6 strain. Since our preliminary data indicate that the distribution of the hotspots in 9R/13R hybrids is different from 9R mice, but identical to 13R mice we concluded that the *Prdm9* allele of the 13R strain has a dominant effect

on the hotspot distribution. Therefore, the 13R *Prdm9* allele was used for the prediction of the PRDM9 DNA binding specificity.

The amino acids at positions -1, 3 and 6 of the alpha helix of a zinc finger domain may be used to predict nucleotide binding specificity. We used the ZIFIBI database [53] to generate the consensus binding site for each zinc finger in the PRDM9 protein. These tri-nucleotide predictions were joined to give the putative PRDM9 binding site. Each nucleotide was assigned an IUPAC letter representing all possible nucleotides at that position. The consensus motif PSSM was scored against every position in the putative PRDM9 binding site, and the maximum scoring position was identified. To assess significance of the alignment score in the context of this zinc-finger arrangement, we generated randomized binding sites of the same length and containing the same zinc fingers as the putative binding site. This showed the alignment to be significant ($P = 0.024$, 10,000 fold bootstrapping). To assess the importance of the specific sequence of this motif, we maintained the frequency ratios at each position, but randomized the nucleotides. We aligned 100,000 such randomized motifs to the PRDM consensus. This again showed the consensus motif alignment score to be highly significant ($p < 10^{-5}$, from alignment score distribution)

SUPPLEMENTARY REFERENCES

31 Meistrich, M. L. Separation of spermatogenic cells and nuclei from rodent testes. *Methods Cell Biol* **15**, 15-54 (1977).

32 Chen, Y. K. *et al*. Heterodimeric complexes of Hop2 and Mnd1 function with Dmc1 to promote meiotic homolog juxtaposition and strand assimilation. *Proc Natl Acad Sci U S A* **101**, 10572-10577, doi:10.1073/pnas.0404195101 0404195101 [pii] (2004).

33 Petukhova, G. V. *et al*. The Hop2 and Mnd1 proteins act in concert with Rad51 and Dmc1 in meiotic recombination. *Nat Struct Mol Biol* **12**, 449-453, doi:nsmb923 [pii] 10.1038/nsmb923 (2005).

34 Leu, J. Y., Chua, P. R. & Roeder, G. S. The meiosis-specific Hop2 protein of S. cerevisiae ensures synapsis between homologous chromosomes. *Cell* **94**, 375-386, doi:S0092-8674(00)81480-4 [pii] (1998).

35 Tsubouchi, H. & Roeder, G. S. The Mnd1 protein forms a complex with hop2 to promote homologous chromosome pairing and meiotic double-strand break repair. *Mol Cell Biol* **22**, 3078-3088 (2002).

36 Goetz, P., Chandley, A. C. & Speed, R. M. Morphological and temporal sequence of meiotic prophase development at puberty in the male mouse. *J Cell Sci* **65**, 249-263 (1984).

37 Baudat, F. & de Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Res* **15**, 565-577, doi:10.1007/s10577-007-1140-3 (2007).

38 Lafuse, W. P., Berg, N., Savarirayan, S. & David, C. S. Mapping of a second recombination hot spot within the I-E region of the mouse H-2 gene complex. *J Exp Med* **163**, 1518-1528 (1986).

39 Smith, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-3.0. http://www.repeatmasker.org (1996-2010).

40 Rudner, R., Karkas, J. D. & Chargaff, E. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proc Natl Acad Sci U S A* **60**, 921-922 (1968).

41 Keeney, S. & Neale, M. J. Initiation of meiotic recombination by formation of DNA double-strand breaks: mechanism and regulation. *Biochem Soc Trans* **34**, 523-525, doi:BST0340523 [pii] 10.1042/BST0340523 (2006).

42 Romanienko, P. J. & Camerini-Otero, R. D. The mouse Spo11 gene is required for meiotic chromosome synapsis. *Mol Cell* **6**, 975-987, doi:S1097-2765(00)00097-6 [pii] (2000).

43 Bentley, D. R. *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:nature07517 [pii] 10.1038/nature07517 (2008).

44 Gentleman, R. C. *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, doi:gb-2004-5-10-r80 [pii] 10.1186/gb-2004-5-10-r80 (2004).

45 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:btq033 [pii] 10.1093/bioinformatics/btq033 (2010).
46 Zhang, Y. *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:gb-2008-9-9-r137 [pii] 10.1186/gb-2008-9-9-r137 (2008).
47 Nix, D. A., Courdy, S. J. & Boucher, K. M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523, doi:1471-2105-9-523 [pii] 10.1186/1471-2105-9-523 (2008).
48 http://sourceforge.net/apps/mediawiki/gemlibrary/index.php?title=The_GEM_library
49 Larkin, M. A. *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:btm404 [pii] 10.1093/bioinformatics/btm404 (2007).
50 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680 (1994).
51 Kankainen, M. & Loytynoja, A. MATLIGN: a motif clustering, comparison and matching tool. *BMC Bioinformatics* **8**, 189, doi:1471-2105-8-189 [pii] 10.1186/1471-2105-8-189 (2007).
52 Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48-54, doi:btb036 [pii] (1998).
53 Cho, S. Y., Chung, M., Park, M., Park, S. & Lee, Y. S. ZIFIBI: Prediction of DNA binding sites for zinc finger proteins. *Biochem Biophys Res Commun* **369**, 845-848, doi:S0006-291X(08)00382-3 [pii] 10.1016/j.bbrc.2008.02.106 (2008).