Title: A high-resolution map of non-crossover events in mice reveals impacts of genetic diversity on meiotic recombination

Authors: Ran Li[1,2,†,‡], Emmanuelle Bitoun[1,2,†], Nicolas Altemose[1,2,†,#], Robert W. Davies[1,2,¶], Benjamin Davies[1], Simon R. Myers[1,2,*].

Affiliations:

[1]The Wellcome Centre for Human Genetics, Roosevelt Drive, University of Oxford, Oxford OX3 7BN, UK.

[2]Department of Statistics, University of Oxford, Oxford OX1 3LB, UK.

[‡]Current address: Target Discovery Institute, University of Oxford, Oxford OX3 7FZ, UK.

[#]Current address: Department of Bioengineering, University of California, Berkeley, California, USA.

[¶]Current address: The Centre for Applied Genomics, Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Canada.

[†]These authors contributed equally to this work.

*Corresponding author. Email: myers@stats.ox.ac.uk

Abstract:

In mice and humans, meiotic recombination begins with programmed DNA double-strand breaks at PRDM9-bound sites. These mainly resolve as difficult-to-detect non-crossovers, rather than crossovers. Here, we intercrossed two mouse subspecies over five generations and deep-sequenced 119 offspring, whose high heterozygosity allowed detection of 2,500 crossover and 1,575 non-crossover events with unprecedented power and spatial resolution. These events were strongly depleted at "asymmetric" sites where PRDM9 mainly binds one homologue, implying they instead repair from the sister chromatid. This proves that symmetric PRDM9 binding promotes inter-homologue interactions, illuminating the mechanism of PRDM9-related hybrid infertility. Non-crossovers were surprisingly short (mean 30-41 bp), and complex non-crossovers, seen commonly in humans, were extremely rare. Unexpectedly, GC-biased gene conversion disappeared at non-crossovers containing multiple mismatches. These results demonstrate that local genetic diversity can alter meiotic repair pathway decisions in mammals by changing PRDM9 binding symmetry and non-crossover resolution, which influence genome evolution, fertility, and speciation.

Main Text:

During meiosis, genetic information is exchanged between homologous chromosomes via the process of recombination. In mammals and other species, recombination plays essential roles in ensuring the proper pairing of chromosomes (synapsis) and segregation of chromosomes into gametes, and together with mutation generates all genetic variation[1,2]. In many species, most recombination clusters into small 1-2 kb regions of the genome, called recombination hotspots. In mice and humans, these hotspots are positioned mainly by PRDM9, a zinc-finger protein that binds specific sequence motifs and deposits at least two histone modifications, H3K4me3 and H3K36me3[3,4], on the surrounding nucleosomes[5–10]. Double-Strand Breaks (DSBs) subsequently form near PRDM9 binding sites, and DSB processing results in single-stranded DNA decorated with the strand exchange proteins RAD51 and DMC1[9].

Each DSB can ultimately repair in several ways (Fig. 1a). Because meiotic DSBs occur following replication of DNA, some DSBs – including on the X chromosome in males, which has no homologue – repair invisibly, using the sister chromatid as a repair template, but many DSBs are repaired using the homologous chromosome. A minority of these form crossovers (COs), involving reciprocal exchanges between homologs, while many more DSBs become non-crossovers (NCOs), in which a section of genetic material is copied from the homologue, without the donating chromosome being altered[11,12]. Because NCO tracts are short and often fail to contain any polymorphic markers, these events are difficult to detect and have been less well studied in mammals. Two recent studies[12,13] have shed new light on genome-wide patterns of NCOs in humans, using SNP array data and some sequencing data; no genome-wide study has yet been conducted in any other mammal. Both studies reported that multiple disjoint NCO tracts

cluster in close proximity, and that NCOs show strong allelic bias at heterozygous AT/GC SNPs, with 68% transmitting GC alleles.

Here, we study both CO and NCO event outcomes in mice, including mice humanized at *Prdm9*[14]. The higher genetic diversity present between mouse subspecies compared to humans greatly improves the power and resolution with which we can detect NCOs and provides us with a unique ability to examine how genetic diversity might influence DSB repair. These data also enable us to resolve unanswered basic questions about meiotic recombination, including the total number of homologous recombination events per meiosis, the length of underlying NCO tracts, and the fraction of COs and NCOs occurring within recombination hotspots.

Previous work suggests there might be complex relationships between genetic diversity, PRDM9 binding, DSB formation, and DSB repair. At DSB sites, we previously published evidence that the degree to which PRDM9 binding is "symmetric" – that is, whether PRDM9 binds both homologues equally at each site – can influence properties of recombination and hybrid fertility[14]. Across a range of hybrids, mice with higher levels of symmetric PRDM9 binding genome-wide consistently show improved fertility measures, suggesting PRDM9 binding symmetry aids proper synapsis of homologous chromosomes. Moreover, individual asymmetric hotspots show elevated DMC1 ChIP-seq signals relative to H3K4me3, most consistent with the possibility that these DSBs take longer to repair[14]. In addition, one recent study reported that hotspots with high polymorphism rates tend to overlap fewer crossovers than expected from DMC1 enrichment[15]. We therefore also gathered complementary H3K4me3 and DMC1 ChIP-seq data (DMC1 data generated elsewhere[16]) in the male parental, or closely related, animals[12,14–

[18]. Together, these data provide unprecedented power to study the processes of non-crossover and crossover recombination together.

Results

We crossed two inbred strains of mice to produce F1 hybrids: C57BL/6J, humanized at *Prdm9* (hereafter B6^{Hum}) and CAST/EiJ (hereafter CAST). B6^{Hum} is of predominantly *Mus musculus domesticus* origin and is identical to C57BL/6J except that the portion of the B6 *Prdm9* exon 10 encoding the DNA-binding zinc finger array has been replaced with the orthologous sequence from the human *PRDM9* B allele to produce a new allele we label *Prdm9*^{Hum} [14]. CAST is of mainly *Mus musculus castaneus* origin and possesses a distinct *Prdm9* allele, *Prdm9*^{Cast}. We chose these subspecies due to their high (0.7%) sequence divergence (Supplementary Information), improving power to detect NCO events in offspring. Moreover, the different *Prdm9* alleles allow us to distinguish the properties of *Prdm9*^{Cast} and *Prdm9*^{Hum} controlled hotspots, with the latter allele being of interest because it has not co-evolved with either mouse subspecies' genome. To identify NCO events, we intercrossed these (B6XCAST)F1 mice to generate the F2 generation and deeply sequenced 11 F2 offspring (16-25x sequencing coverage, Fig. 1a) (Supplementary Information), whose genomes reflect recombination events that occurred in the 22 meioses in their F1 parents. We also gathered ChIP-seq data for both DMC1[16] and H3K4me3 in testes from a male (B6XCAST)F1-*Prdm9*$^{Hum/Cast}$ mouse, allowing us to compare these to NCO/CO event outcomes.

We selected 52 (B6xCAST)F2 mice homozygous for the *Prdm9*^{Hum} allele and intercrossed these for three additional generations (Supplementary Information) to generate 108 mice that we

sequenced: 72 (B6XCAST)F5-*Prdm9*$^{Hum/Hum}$ mice (25-30x) and their 36 F4 parents (12-17x; Fig. 1a). The additional generations produce an accumulation of many NCO and CO events, allowing us to study their properties in detail. In aggregate, we were able to map signatures of PRDM9 binding, DSB formation, and NCO/CO events, controlled by two different *Prdm9* alleles, separately in both sexes.

To find CO and NCO events, we developed and applied an HMM-based algorithm (Supplementary Information) to infer ancestral states (B6/B6, B6/CAST and CAST/CAST) across the genome in each mouse (Fig. 1b), and we smoothed to produce a "background" state to test potential gene conversions against. CO events correspond to background changes. SNPs with genotypes not matching their local background represent possible NCO events, but sequencing error also mimics NCO events. Indeed, from an initial set of 863,082 SNPs potentially within NCO events from 11 F2 animals, the vast majority are explained by sequencing errors, and following careful filtering (Extended Data Table 1) using properties of sequencing reads (Supplementary Information), we identified a final collection of 183 NCOs and 295 CO events on autosomes from the 11 F2 animals (Fig. 1c) and 1,392 NCOs and 2,205 CO events in the (B6XCAST)F5-*Prdm9*$^{Hum/Hum}$ mice (Fig. 1d). We used additional sequencing (Supplementary Information) to validate a targeted subset of events occurring in (B6XCAST)F1-*Prdm9*$^{Hum/Cast}$ parents, estimating that 91% of the events that we identified represent true NCOs. To estimate our overall power to identify NCOs (which must include at least one SNP to be observable), we performed simulations (Supplementary Information). These revealed our power to be 63-100%, in various settings (Extended Data Fig. 1a, b).

Overall event properties

NCO and CO events, as well as DMC1 and H3K4me3, show enrichment nearer to telomeres (Fig. 1e and Extended Data Fig. 1c); broadly similar to patterns observed in other mice[19–21] and humans[22]. At least 99.4% of observed NCO events were "simple" and comprised contiguous tracts of converted SNPs, with no non-converted SNPs amongst them. Similarly, 99.4% of crossovers were simple background switches. This implies that complex NCOs are extremely unusual in mice. This pattern contrasts strongly with recent human results, where a large number of complex NCO events, often extending over a kilobase, are seen[12,13]. Complex human events are not strongly enriched in hotspots, occur mainly in females, and show an association with maternal age[12].

In the F5 mice, we were able to identify both *de novo* and parentally inherited NCO and CO events; and we were able to assign a subset to the maternal or paternal meiosis (Supplementary Information). From the H3K4me3 and DMC1 ChIP-seq data, we identified 23,748 DMC1 peaks corresponding to DSB hotspots, and 63,050 PRDM9-dependent H3K4me3 peaks marking PRDM9 binding sites in male (B6xCAST)F1-*Prdm9*$^{Hum/Cast}$ mice (Fig. 2a)[14,23] (Supplementary Information). For most peaks, we could determine which *Prdm9* allele controls them (Supplementary Information). We defined NCO and CO events as occurring within hotspots if they were less than one kilobase away from either peak type (covering 4% of the genome). NCO events can be associated with a genetic background, determining whether they result from a DSB on the B6 or CAST background.

In the F2 mice with parents of the same F1 background used for the ChIP-seq data, 96% of CO events and 92% of NCO events (adjusted for false-positive NCO events and chance overlap; 84% unadjusted) overlap either DMC1 or H3K4me3 ChIP-seq peaks. Thus, recombination hotspots identified by ChIP-seq account for essentially all recombination in mice, with little recombination in the remainder of the genome. NCO and CO events both occur in individual hotspots with probability approximately proportional to their estimated heat using either DMC1 or H3K4me3 signal strength (Fig. 2b, c): over 50% of all hotspot-associated F2 NCO or CO events occur in only the 4,000 hottest hotspots. Strong dominance for *Prdm9*^{Cast}–controlled over *Prdm9*^{Hum}–controlled hotspots is seen in CO and NCO events (Fig. 2d, e) and also in DMC1 and H3K4me3 ChIP-seq data (Extended Data Fig. 2a, b). Thus, dominance is not simply a consequence of evolutionary hotspot erosion[17]. Instead, it could be due to a greater number of strong PRDM9^{Cast} binding targets genome-wide or possibly due to higher expression of *Prdm9*^{Cast}. In F5 mice, where only *Prdm9*^{Hum}–controlled recombination hotspots are active, ChIP-seq peak overlap was only modestly reduced to 88.7% (COs) and 78.8% (NCOs) (Extended Data Table 2), indicating that hotspots and hotspot heats identified by ChIP-seq in the heterozygous F1 mouse are still informative for meioses occurring in homozygous F4 mice.

The extraordinarily high ChIP-seq hotspot overlap, with almost all recombination being explained by only ~24,000 hotspots, substantially exceeds estimates based on human population-averaged recombination maps[22]. This might be explained by diversity in *PRDM9* alleles within the human population. Because F2 events represent both maternal and paternal recombination, but ChIP-seq hotspots are assayed only in males, our results show that few or no truly female-specific recombination sites can exist in mice, although differences in *activity* have been

observed for particular hotspots between the sexes[24]. We applied an approach accounting for Poisson variation in observed event counts (Extended Data Fig. 2c, d) (Supplementary information) to estimate correlations among underlying recombination rates at different scales. We estimate an overall correlation of ~70% between male and female rates (NCO and COs combined) in F5 mice at fine scales (<1 Mb), with a possible decrease at broader scales (Fig. 2f). Moreover, we observe strong correlations (>70%) between (sex-averaged) NCO and CO rates across different scales (Fig. 2g), although we also find very strong evidence that these events do differ in their positioning along the chromosome at broader scales, and the NCO rate is much higher than the CO rate at all scales.

Length, number, and positioning of NCO tracts

We leveraged the high SNP density in our system to estimate properties of the underlying NCO event tract lengths (accounting for the fact that if a NCO event does not contain a SNP, it is not observed) (Supplementary information), separately for hotspots controlled by *Prdm9*^{Cast} and *Prdm9*^{Hum}. The data show relatively good fits to an exponential tract length (Fig. 2h), but with significant differences in estimated mean NCO tract length (p=0.0018): 30 bp (95% CI 25-35 bp) for *Prdm9*^{Cast}, and 41 bp (95% CI 35-48 bp) for *Prdm9*^{Hum}. This is unexpected and implies that new *Prdm9* alleles can change basic properties of how recombination events resolve. These tract length estimates are much more precise than previous studies, and they imply tract lengths at the lowest end of estimates for humans and mice[12,13,23,25].

Using these tract lengths and accounting for our incomplete power to identify NCOs, we estimate that there are 273.7 NCOs (95% CI 231.2-342.7) and 26.8 COs (95% CI 24.4-30.9) per

meiosis in F1 parents (Supplementary information), a ratio similar to previous estimates at individual hotspots or using other approaches[23,24,26], with a similar estimate of 235.2 NCOs in F4 parents. This yields a sex-averaged total of 300.5 DSBs (95% CI 258.5-370.5) per meiosis repairing using the homologue. Previous studies using microscopy have estimated that there are a maximal number of 200-400 visible DMC1 foci per meiosis in mice[11,23,27]. This suggests that the majority of DSBs might be repaired via homologous chromosomes, rather than the sister chromatid[28,29].

We also identified distinct sequence motifs, and their locations, within 97% of hotspots controlled by *Prdm9*^{Cast} and 74% of hotspots controlled by *Prdm9*^{Hum} (Extended Data Fig. 3a)[12,14,17,30]. Both NCO and CO event centres distribute symmetrically around these motifs (Fig. 3a-d and Extended Data Fig. 3b-d). NCO events cluster very near to the PRDM9 binding motifs (potentially overlapping it in 70% of cases; Fig. 3a, d), slightly less strongly than clustering of SPO11-mapped DSBs[31], but with a far tighter range than the DMC1 and H3K4me3 ChIP-seq signals, which identify single-stranded resection tracts around DSBs and histone methylation resulting from PRDM9 binding, respectively (Fig. 3e-f). Thus, NCO gene conversion appears restricted to sites very close to initiating DSBs themselves. A broader positional distribution for observed CO events (Fig. 3b-c) is consistent with previous studies[25,31].

GC-biased gene conversion is controlled by SNP density and explains complex NCO and CO events

Both indirect[32–34] and direct[12,13] evidence for NCO events in humans has revealed a strong (68%) bias from AT towards GC bases, occurring via an unknown mechanism. This phenomenon is

thought to have influenced variability in the GC-content of many species genome-wide[35,36]. Although the mechanisms for this phenomenon remain unknown, the possible causes include either subtle event initiation biases[36,37], or heteroduplex DNA repair pathways[29]. However, simple models of heteroduplex DNA repair favoring G/C bases at mismatching Watson-Crick base pairings are difficult to reconcile with the fact that most NCO events convert a contiguous set of SNPs, with no evidence of repair template switching. Moreover, not all SNPs in individually studied hotspots show GC bias[38].

Our NCO events show strong evidence of AT-to-GC bias, though initially weaker than seen in humans[13], for both *Prdm9*^{Cast}–controlled (64%, binomial test $p=1.3\times10^{-9}$) and *Prdm9*^{Hum}-controlled (60%, binomial test $p=6.2\times10^{-10}$) hotspots (Extended Data Table 3). We next focused on NCO events within *Prdm9*^{Hum}–controlled hotspots for further investigation, because the genomic GC-content has not evolved alongside this allele in such hotspots. We tested for a difference in NCO tracts containing a single SNP with those containing multiple SNPs (Fig. 4a). Surprisingly, this revealed GC-bias to occur exclusively in single-SNP NCO tracts, with no bias ($p=0.92$) for all multiple-SNP tracts combined. Single-SNP NCO events show near-identical GC-bias (68%) in both males and females (Extended Data Table 4), with GC-bias strength unaltered even if DSBs happen only on one homologue (Extended Data Table 4), implying a mechanism driven by heteroduplex repair rather than DSB formation.

A restriction of GC-bias to single-SNP tracts might reflect either some GC-biased process preventing longer events occurring, or a direct impact of the number of SNPs within heteroduplex DNA on whether GC-bias occurs. To distinguish these possibilities, we stratified

SNPs by distance to their nearest SNP and measured their GC-bias if they fell within NCO events (Fig. 4b). Strikingly, SNPs near to other SNPs, and therefore almost always co-converted with them, show no GC bias evidence. Conversely SNPs further than typical NCO tract lengths, >100 bp from the nearest SNP, show the ~68% bias observed in humans, in whom SNP density is much lower[12,13]. This implies that local genetic diversity itself influences GC-biased gene conversion at NCOs, and therefore there must be at least two distinct processes operating to repair heteroduplex stretches formed at DSBs, one which is strongly GC-biased, and another which dominates when multiple mismatches exist and shows no GC bias.

To further characterise GC-bias, we estimated conversion rates of different types of SNP in the donor and recipient chromosomes at single-SNP NCO sites (Fig. 4c) (Supplementary information). We normalised these relative to their conversion rates in multi-SNP events (Extended Data Fig. 3b), or to flanking SNP composition (Fig. 4c), both of which show no GC-bias and gave near-identical results. The simplest model which can explain the data is if there are two distinct conversion rates, with observed NCO rates lower if the *recipient* chromosome (i.e. the homologue in which the DSB occurs) carries a G or a C, and higher if the recipient carries an A or a T. For example, G/C transversions appear to convert at the lower rate. This could be explained by a model where a GC-biased process can resolve heteroduplex DNA in favor of the recipient chromosome, if it carries a G and/or C base – effectively "blocking" conversion of that base. If so, higher local heterozygosity, which disrupts this process, would be expected to actually *increase* local NCO rates.

Interestingly, we do not observe a consistent GC-bias for CO events, which are accompanied by long conversion tracts of ~500 bp in size[25]. However, we did observe a very small number of "complex" events, incorporating non-converted markers surrounded by converted markers, and resulting from the same meiosis. We hypothesised that these might result from occasional operation of the GC-biased process. If so, the above results suggest that complex events might result from "blocking" of conversion of particular markers where the recipient chromosome carries a G or C base. This motivates examining the non-converted markers surrounded by converted markers within complex CO and NCO events. We observed a total of 12 such markers within NCO events and 7 within CO events (Supplementary Information). Remarkably, for 18 of these 19 cases the recipient chromosome carries a G or C base ($p= 7.6x10^{-5}$ by 2-sided binomial test).

Therefore, in our mice essentially all of the complex NCO and CO events we observe can be explained in terms of the action of a GC-biased process which normally only operates within single-SNP conversion tracts. A recent study of one human hotspot[39] found a similar GC-bias of 87-100% for complex CO events, so it seems likely this process operates across species. Moreover, the bias of nearly 100% towards the recipient carrying a G/C, compared to the ~68% bias of all single-SNP NCO events, might suggest that non-biased heteroduplex repair occurs even in some tracts containing only a single heteroduplex site. For example, the bias might only impact either G or C recipient bases, but not both, which would cap the bias at NCO sites to (at most) 67%, very close to the observed fraction.

Hotspots where the homologue is not bound by PRDM9 show increased DMC1 occupancy, yet reduced homologous recombination

For NCO events, we can identify on which homologue the underlying DSB occurred. In the F2 mice, we observed a bias: 60% of our observed NCOs were initiated on the B6 background ($p<10^{-3}$). This B6 vs. CAST background bias is due to the dominance of the *Prdm9*^{Cast} allele, which accounts for 80% of observed NCO events (Fig. 2e), and which shows a strong preference for binding and initiating recombination events on the B6 background (Extended Data Fig. 5a; 66% of NCOs, $p<10^{-3}$), explained by evolutionary hotspot erosion of CAST-controlled hotspots on the CAST genetic background[15,17]. In contrast, the *Prdm9*^{Hum} allele appears to bind and initiate recombination events equally on both backgrounds (Extended Data Fig. 5a, p=0.63). We found that the fraction of NCOs initiating on the B6 background correlates highly with the fraction of DMC1 and H3K4me3 ChIP-seq signal originating from that background (correlations of 0.88 and 0.98, respectively; Extended Data Fig. 5b, e). Because the ChIP-seq data only reflect male meiosis but observed NCO events originate from both males and females, this high correlation implies similar hotspot behavior in both sexes. These increases in PRDM9 binding, DSB formation, and NCO formation on the B6 chromosome imply that no strong compensation mechanism acts to equalise the number of DSBs or recombination events on different homologues, although weaker compensation that we lack power to detect might occur.

Recombination hotspots can be separated into “asymmetric” cases where DSBs occur mainly on one homologous chromosome, and “symmetric” cases where DSBs occur equally on both homologues. Using H3K4me3 and DMC1 ChIP-seq data, we estimated the fraction of PRDM9 binding and DSB formation on the B6 vs. CAST chromosome in each hotspot[14] (Supplementary

Information). Among the most asymmetric hotspots (>95% of signal from one chromosome), we observed SNPs or indel polymorphisms within 96% of identified motifs overall (Extended Data Fig. 5f) (Supplementary Information), implying that asymmetry is mainly driven by sequence changes disrupting PRDM9 binding on one homologue. Therefore, asymmetric binding is expected to be conserved between the sexes and between F2 and F5 animals.

We previously found that the ratio of DMC1 to H3K4me3 ChIP-seq signal increases roughly twofold at asymmetric hotspots compared to symmetric hotspots[14]. This excess of DMC1 signal indicates that DSBs either form at a higher rate or take longer to repair at asymmetric hotspots when compared to symmetric hotspots matched for the same level of PRDM9 binding (as measured by H3K4me3). In this study, we observed a similar excess of DMC1 signal in both *Prdm9*^{Cast} and *Prdm9*^{Hum} controlled asymmetric hotspots, as expected (Extended Data Fig. 5g). However, for the first time we were also able to measure the numbers of NCO and CO events actually occurring in asymmetric vs. symmetric hotspots, and we compared them to their expected counts according to DMC1 and H3K4me3 signal (Supplementary Information).

Given that DMC1 marks DSB sites, then if all DSBs were equally likely to repair by homologous recombination (resolving as COs or NCOs), we would necessarily expect any two groups of hotspots that are matched to have the same total DMC1 signal to also have similar numbers of CO and NCO events. However, when we grouped *Prdm9*^{Hum}-controlled hotspots according to symmetry (defined for each hotspot using the proportion of DMC1 signal from each homologue), we instead observed a twofold depletion of NCO and CO events in the most asymmetric hotspots ($p=10^{-27}$ and $p=10^{-23}$, respectively, after controlling for factors influencing

power; Fig. 5a) (Supplementary Information). Results were similar for *Prdm9*^{Cast}; for both males and females; and for *de novo* and inherited events in F5 mice, as well as events in F2 mice (Extended Data Fig. 6a-d). Because the *Prdm9*^{Hum} allele in particular did not co-evolve alongside the mouse genome, asymmetric hotspots controlled by this allele reflect chance genetic variation disrupting PRDM9 binding sites on one homologue or the other, implying a mechanistic impact of asymmetry on recombination independent of hotspot erosion or other evolutionary forces. Importantly, we found that this homologous recombination deficiency is driven by asymmetry alone rather than SNP diversity elsewhere within hotspots (Supplementary Information). Furthermore, for DSBs occurring on the less-bound chromosome of asymmetric hotspots, we found that NCO events occur at the expected rate for symmetric hotspots (Supplementary Information). This implies that when DSBs occur at asymmetric hotspots on the more frequently bound chromosome, the resulting lack of recombination must stem from a lack of PRDM9 binding to its homologue.

Surprisingly, we also saw a twofold deficit of both NCO and CO events in *Prdm9*^{Hum}-controlled asymmetric hotspots relative to expectations from H3K4me3 ChIP-seq signal (Fig. 5b). This was significant for both males and females ($p<0.04$) (Supplementary Information). Because H3K4me3 reflects the level of PRDM9 binding, the lack of homologous recombination at asymmetric hotspots might be explained if DSBs occur less often at these sites. However, this seems impossible to reconcile with the strong *excess* DMC1 signal observed at asymmetric hotspots for a given level of H3K4me3 signal. Instead, the lack of COs and NCOs at asymmetric hotspots can be explained if their DSBs frequently repair via the sister chromatid rather than the homologous chromosome. If so, this must occur in both sexes. Interestingly, sister chromatid

repair is thought to operate on meiotic DSBs on the X chromosome in males[40], which, similar to autosomal DSBs in asymmetric hotspots, exhibit a very strong increase in DMC1 signal relative to H3K4me3 signal, probably owing to the late timing of their repair[14].

Discussion

It is interesting to compare our results to those of two recent studies of human NCO events, one of which analysed a similar number of events[12,13]. A striking difference is that although both human studies found complex NCO events to be common, particularly in maternal meiosis and with an incidence increasing with maternal age, such events are near absent in mice. We suggest that this difference may reflect differences in the timespan of dictyate arrest, which occurs before the completion of recombination, and lasts decades in humans vs. months in mice. These findings support the idea that complex NCO events in humans might reflect the repair of non-programmed DNA damage, consistent with the fact that they mainly occur outside PRDM9 hotspots[12]. Interestingly, such events are often long and show GC-bias, which could potentially be explained by our model of GC-bias resulting from failure to convert SNPs against a G or C background base.

A second difference is that our results indicate a sex-averaged NCO rate in mice carrying humanized *Prdm9* of around 10^{-6} per base. This is strikingly below human estimates[12], of around 4.1×10^{-6} and 7.7×10^{-6} in males and females respectively, meaning humans show even greater increases in the NCO:CO ratio relative to mice – for unknown reasons. Nonetheless, our minimum estimates of total DSB counts in mice are consistent with previous microscopy studies,

suggesting that most DSBs in mice may repair using the homologous chromosome, at least those not found in asymmetric hotspots or impacted by GC-biased repair processes.

Using DMC1[16] and H3K4me3 ChIP-seq data, we can infer that DSBs form at a particular site on average proportionally to the rate of PRDM9 binding to that site[14]. However, the processing, repair and eventual recombination outcomes at each PRDM9 binding site all depend strongly on the sequence of the homologue. Firstly, we confirmed our previous finding that when DSBs form at "asymmetric" sites where the homologue is not strongly bound by PRDM9 (mostly due to mutations in the PRDM9 binding motif on the homologue), we observe an excess of DMC1 signal relative to H3K4me3 signal, which is consistent with either more DSBs at these sites, or delayed DSB repair[14]. In this study, we additionally found that CO and NCO events occur at only around half the expected rate at asymmetric hotspots (Fig. 5). We also showed that this reduction in homologous recombination at asymmetric sites cannot be explained by genetic diversity alone, as has been suggested[15]; only nearby SNPs that abolish PRDM9 binding symmetry have any effect on the CO or NCO rate. This implies that asymmetric hotspots are in fact compromised in their ability to effectively interact with their homologues and exchange material, which is consistent with the wider asynapsis seen in animals where asymmetric hotspots predominate[14,41]. The reduction of recombination at asymmetric hotspots may also mitigate the effect of hotspot erosion due to the overtransmission of alleles that disrupt PRDM9 binding[14,17].

Taking into account their elevated DMC1 signals, asymmetric hotspots behave oddly, by showing some combination of potentially slower DSB repair, and an inability to engage with their homologues for this repair, via either CO or NCO. Results for NCO events show that for

breaks occurring on the less-bound chromosome of asymmetric hotspots, homologous interactions occur at the expected rate for symmetric hotspots. This suggests that whether the homologue is bound or not is what truly determines the behavior at a PRDM9 binding site. In fact, all our data could be explained by a model whereby DSBs at hotspots where the homologue is not bound sometimes fail to interact with that homologue and are repaired (slowly) from the sister chromatid instead, lowering the observed number of CO/NCO events relative to PRDM9 binding, and increasing the DMC1 signal due to the repair delay. That is, asymmetric hotspots may behave like DSB hotspots on the X chromosome in males, which repair late and from the sister chromatid, and show excess DMC1 signal[14,40]. Other models would involve either more or fewer DSBs occurring at asymmetric hotspots and seem unlikely, because they require strong pairing of homologues prior to DSB formation in order to distinguish symmetric and asymmetric binding sites.

A second impact of genetic differences between homologous chromosomes comes in mediating GC-biased gene conversion (gcBGC). We confirm gcBGC operates downstream of DSB formation, and this implies it must act on repair of heteroduplex DNA with mismatching bases, formed during DSB repair towards recombination. We find gcBGC acts almost exclusively on potential conversion tracts containing only a single SNP (i.e. mismatch), with a strength essentially identical to that observed in humans (68% of NCOs convert A/T to G/C)[12,13]. This single-SNP preference can explain why most multi-SNP NCO tracts are simple stretches of markers without “cherry-picking” of markers converted towards GC. Because heteroduplex DNA is expected to form for all possible NCO and CO tracts containing SNPs, not just those containing single SNPs, our results imply more than one pathway for heteroduplex repair (Fig.

6). In the first, gcBGC acts to favor the strand on which the DSB occurs: if this strand carries a G or C at the SNP, conversion is prevented from occurring. We calculated that such a process would need to block gene conversion at G/C recipient sites 53% of the time to account for the observed 68% overall GC-bias in observed events (Fig. 6) (Supplementary Information). Almost all observed complex NCOs and COs, though very rare, appear to be explained by this GC-bias preventing conversion of individual markers, with the background on which the DSB occurred carrying a G or C base in 95% of such non-converted markers we observed. Similar behavior was observed in a study of COs within a single human hotspot[42]. Among suggested drivers of mammalian gcBGC[32], these properties of strong (almost 100%) base-specific and strand-specific biases, and action on very fine scales (single SNPs), appear most consistent with the action of base excision repair (BER) rather than mismatch repair (MMR) proteins.

The alternative, non-GC-biased repair pathway of heteroduplex DNA instead can act on multiple-SNP stretches, and must show a strand bias – this time favoring the *incoming* strand, copied from the homologous chromosome on which the DSB did *not* occur (Fig. 6). Otherwise, if heteroduplex mismatch repair had no strand bias, then half of potential NCOs would repair invisibly, and so to account for our estimate of the NCO rate (~274 NCOs per meiosis) there would need to be twice as many DSBs per meiosis (~600), which is far outside the range of previous estimates[23,24,26]. Moreover, although resolution of heteroduplex DNA towards the broken chromosome within potential NCO events would be invisible, at 19 observed CO events within highly (>95%) asymmetric hotspots containing a mutation within their PRDM9 motif, we observe transmission of the "cold" PRDM9 allele to offspring in 95% of cases. Therefore, for those longer conversion tracts within CO events at least, heteroduplex repair appears

overwhelmingly biased towards the unbroken homologue. If this mechanism for resolving heteroduplex is also used for single-SNP stretches in some cases, this would explain why gcBGC appears weaker for NCO events in general, compared to complex NCO and CO events. We suggest that this non-GC-biased process, impacting longer stretches including multiple SNPs, is consistent with properties of MMR proteins, several of which are known to be essential for meiosis in mice[43,44]. If these hypotheses are correct, it is interesting that BER and MMR appear able to favor different strands.

GC-bias is near-absent for SNPs adjacent to other SNPs, explained because these SNPs are rarely in single-marker conversion tracts. Notably, these results imply that SNP density within hotspots influences recombination events downstream of DSB formation. For example, the same SNP will show different conversion rates and biases in different individuals, depending on nearby heterozygosity patterns in those individuals. Interestingly, this predicts a slightly higher NCO rate in more diverse than less diverse regions. Another unexpected influence on NCO events, again acting downstream of DSB formation, is *Prdm9* allele, with CAST-controlled NCOs having an average length 11 bp (27%) shorter than human-controlled NCOs. It is unclear whether this reflects PRDM9 binding directly, or some indirect impact, e.g. how PRDM9 binds relative to nucleosome positions.

In mice, male animals with *Prdm9* but with predominantly asymmetric recombination hotspots show high rates of asynapsis and infertility. Our results indicate that in such hotspots in both sexes, DSBs are impaired in their ability to mediate inter-homologue recombination interactions, either NCO or CO events, and DSBs at asymmetric hotspots show repair delay, at least in males.

This suggests that at DSB sites, PRDM9 binding to the homologue and/or the accompanying histone modifications may accelerate homology search[14]. Mouse crosses with asymmetric hotspots show less reduced fertility in females[45], suggesting greater robustness to repair delay in oogenesis. *Prdm9*-null B6 mice show partial synapsis and mainly symmetric DSB hotspot sites at promoter-associated H3K4me3 peaks, yet are infertile[46]. However *PRDM9* is absent in dogs, and although only one human carrying a homozygous null mutation at *PRDM9* has been observed, this female has apparently normal fertility[47]. Our results imply a several-fold lower rate of NCO events, which mark DSB sites, in mice relative to humans. Speculatively, perhaps a greater number of DSBs may aid synapsis, and so fertility, by providing more potential inter-homologue interaction sites during human meiosis. However, an elevation in DSB numbers increases the potential for mispairing at some sites, with consequences including diseases caused by non-allelic homologous recombination.

Data availability

The datasets generated and analysed during the current study will be made available in public repositories (GEO and SRA) prior to publication. The H3K4me3 ChIP-seq data are currently available with GEO accession GSE119727.

Code availability

The computer code developed for the analysis of the datasets in the current study will be made available in Github prior to publication.

Ethical compliance

All experiments involving research animals received local ethical review approval from the University of Oxford Animal Welfare and Ethical Review Body (Clinical Medicine board) and were carried out in accordance with the UK Home Office Animals (Scientific Procedures) Act 1986.

References and Notes:

1. Paigen, K. & Petkov, P. Mammalian recombination hot spots: properties, control and evolution. *Nat. Rev. Genet.* **11,** 221–233 (2010).
2. Lichten, M. Meiotic recombination: Breaking the genome to save it. *Curr. Biol.* **11,** 253–256 (2001).
3. Powers, N. R. *et al.* The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. *PLoS Genet.* **12,** 1–24 (2016).
4. Eram, M. S. *et al.* Trimethylation of histone h3 lysine 36 by human methyltransferase prdm9 protein. *J. Biol. Chem.* **289,** 12177–12188 (2014).
5. Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Genome Res.* **12,** 360–366 (2010).
6. Berg, I. L. *et al.* PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat. Genet.* **42,** 859–863 (2010).
7. Parpanov, E. D., Petkov, P. M. & Paigen, K. Esrrg-1 , and Psmb9 (fig. S1). We located both domain providing a histone methyl transferase ac- 1. *Science (80-.).* **327,** 2010 (2010).
8. Myers, S. *et al.* Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science (80-.).* **327,** 876–879 (2010).
9. Baudat, F., Imai, Y. & de Massy, B. Meiotic recombination in mammals: localization and regulation. *Nat. Rev. Genet.* **14,** 794–806 (2013).
10. Berg, I. L. *et al.* Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 12378–12383 (2011).
11. Cole, F. *et al.* Homeostatic control of recombination is implemented progressively in mouse meiosis. *Nat. Cell Biol.* **14,** 424–430 (2012).
12. Halldorsson, B. V *et al.* The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* 1–11 (2016). doi:10.1038/ng.3669
13. Williams, A. L. *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife* **4,** 1–21 (2015).
14. Davies, A. B. *et al.* Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* **530,**

171–176 (2016).

15. Smagulova, F., Brick, K., Pu, Y., Camerini-otero, R. D. & Petukhova, G. V. The evolutionary turnover of recombination hot spots contributes to speciation in mice. *GENES Dev.* 266–280 (2016). doi:10.1101/gad.270009.115.4

16. Hinch, A. G. *et al.* Novel factors influencing meiotic recombination revealed by whole genome sequencing of single sperm cells (in press).

17. Baker, C. L. *et al.* PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-Specific Initiation of Meiotic Recombination. *PLoS Genet.* **11,** e1004916 (2015).

18. Smagulova, F. *et al.* Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* **472,** 375–378 (2011).

19. Cox, A. *et al.* A new standard genetic map for the laboratory mouse. *Genetics* **182,** 1335–44 (2009).

20. Shifman, S. *et al.* A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol.* **4,** 2227–2237 (2006).

21. Brunschwig, H. *et al.* Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics* **191,** 757–764 (2012).

22. Pratto, F. *et al.* Recombination initiation maps of individual human genomes. *Science (80-.).* (2014). doi:10.1126/science.1256442

23. Baudat, F. & De Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosom. Res.* **15,** 565–577 (2007).

24. de Boer, E., Jasin, M. & Keeney, S. Local and sex-specific biases in crossover vs. noncrossover outcomes at meiotic recombination hot spots in mice. *Genes Dev.* **29,** 1721–1733 (2015).

25. Cole, F. *et al.* Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat. Genet.* **46,** 1072–1080 (2014).

26. Cole, F., Keeney, S. & Jasin, M. Preaching about the converted: how meiotic gene conversion influences genomic diversity. *Ann. N. Y. Acad. Sci.* **1267,** 95–102 (2012).

27. Paigen, K. & Petkov, P. Meiotic DSBs and the control of mammalian recombination. *Cell Res.* **22,** 1624–1626 (2012).

28. Zickler, D. & Kleckner, N. Recombination, Pairing, and Synapsis of Homologs during Meiosis. *Cold Spring*

Harb. Lab. Press 1,2 (2015). doi:10.1101/cshperspect.a016626

29. Borde, V. & de Massy, B. Meiosis: Early DNA double-strand breaks pave the way for inter-homolog repair. *Dev. Cell* **32,** 663–664 (2015).

30. Guénet, J. L. & Bonhomme, F. Wild mice: An ever-increasing contribution to a popular mammalian model. *Trends Genet.* **19,** 24–31 (2003).

31. Lange, J. *et al.* The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. *Cell* **167,** 695–708.e16 (2016).

32. Duret, L. & Galtier, N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285–311 (2009).

33. Lassalle, F. *et al.* GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. *PLOS Genet.* **11,** e1004941 (2015).

34. Pessia, E. *et al.* Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol. Evol.* **4,** 675–682 (2012).

35. Meunier, J. & Duret, L. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21,** 984–990 (2004).

36. Duret, L. & Arndt, P. F. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **4,** (2008).

37. Webb, A. J., Berg, I. L. & Jeffreys, A. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc. Natl. Acad. Sci.* **105,** 10471–10476 (2008).

38. Odenthal-Hesse, L., Berg, I. L., Veselis, A., Jeffreys, A. J. & May, C. A. Transmission Distortion Affecting Human Noncrossover but Not Crossover Recombination: A Hidden Source of Meiotic Drive. *PLoS Genet.* **10,** (2014).

39. Tiemann-Boege, I., Schwarz, T., Striedner, Y. & Heissl, A. The consequences of sequence erosion in the evolution of recombination hotspots. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372,** 20160462 (2017).

40. Lu, L. Y. & Yu, X. Double-strand break repair on sex chromosomes: Challenges during male meiotic prophase. *Cell Cycle* **14,** 516–525 (2015).

41. Gregorova, S. *et al.* Modulation of prdm9-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice. *Elife* **7,** 1–21 (2018).

42. Arbeithuber, B., Betancourt, A. J., Ebner, T. & Tiemann-Boege, I. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci.* **112,** 201416622 (2015).

43. De Vries, S. S. *et al.* Mouse MutS-like protein Msh5 is required for proper chromosome synapsis in male and female meiosis. *Genes Dev.* **13,** 523–531 (1999).

44. Kneitz, B. *et al.* MutS homolog 4 localization to meiotic chromosomes is required for chromosome pairing during meiosis in male and female mice. *Genes Dev.* **14,** 1085–1097 (2000).

45. Bhattacharyya, T. *et al.* X Chromosome Control of Meiotic Chromosome Synapsis in Mouse Inter-Subspecific Hybrids. *PLoS Genet.* **10,** (2014).

46. Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D. & Petukhova, G. V. Genetic recombination is directed away from functional genomic elements in mice. *Nature* **485,** 642–645 (2012).

47. Narasimhan, V. M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science (80-.).* **352,** 474–477 (2016).

48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

49. Lunter, G. & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21,** 936–939 (2011).

50. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477,** 289–94 (2011).

51. Yang, H. *et al.* A customized and versatile high-density genotyping array for the mouse. *Nat. Methods* **6,** 663–666 (2009).

52. Jensen-Seaman, M. & Furey, T. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 528–538 (2004). doi:10.1101/gr.1970304.1

53. Altemose, N. *et al.* A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *Elife* 1–46 (2017).

54. Khil, P. P., Smagulova, F., Brick, K. M., Camerini-Otero, R. D. & Petukhova, G. V. Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. *Genome Res.* **22,** 957–965 (2012).

Acknowledgments: We thank the High-Throughput Genomics Group at the Wellcome Centre for Human Genetics (funded by Wellcome Trust grant reference 203141/Z/16/Z) for the generation of sequencing data. We also thank Gang Zhang and Anjali Gupta Hinch for generating and providing the DMC1 ChIP-seq data for the (B6xCAST)F1-*Prdm9*$^{Hum/Cast}$ mouse. Funding: This work was supported by a Wellcome Trust Investigator Award to S.R.M. (098387/Z/12/Z); N.A. is a Howard Hughes Medical Institute Gilliam Fellow.

Author contributions: S.R.M. designed the study; B.D. generated the humanized B6 mouse; E.B., N.A., and B.D. bred the mice (F0-F2: N.A. and B.D.; F3-F5: E.B.); E.B. performed NCO event validation; N.A. performed H3K4me3 ChIP-seq and data processing; R.L. analysed the data; R.W.D. contributed SNP data; R.L., E.B., N.A. and S.R.M prepared the manuscript; B.D. critically reviewed the manuscript.

Competing interests: Authors declare no competing interests;

Materials & Correspondence: requests should be addressed to S.R.M.

Methods

Mouse breeding and library preparation

CAST/Eij (CAST) mice were sourced from MRC Harwell (UK). The C57BL/6J (B6) line humanized at the Prdm9 zinc-finger array ($B6^{Hum}$) was generated previously[14]. Breeding of CAST and $B6^{Hum}$ mice (F0) was carried out in both directions (using females and males of each type) to generate (B6xCAST)F1 hybrid, heterozygous offspring. F1 mice were genotyped at the *Prdm9* locus as previously described[14], and males and females with the $Prdm9^{Hum/Cast}$ genotype were bred to produce F2 offspring. We selected 26 F2 males and 26 F2 females that were homozygous for humanized *Prdm9*, and we further bred them for two generations to produce F3 and F4 mice. Then we chose 18 F4 males and 18 F4 females and bred them to generate F5 mice. We randomly selected four F5 offspring for sequencing (2 males and 2 females) from each of the 18 pairs of F4 parents. In total, one $B6^{Hum}$ mouse, one CAST mouse, 11 F2 mice and all 18 F4/F5 families (36 F4 parents and 72 F5 offspring) were subjected to whole genome sequencing. Genomic DNA was extracted from spleen using the DNAeasy Blood and Tissue Kit (Qiagen), according to the manufacturer's instructions. Libraries were prepared by the Oxford Genomics Centre at the Wellcome Centre for Human Genetics (Oxford, UK) using established Illumina protocols (with a Nextera DNA Library Prep Kit). Where possible, we preserved spleen, liver, testis, and ear punch samples from each mouse in the final pedigree.

Data processing

We sequenced more than 120 mice, aiming to get 10x coverage for the 2 F0 mice and 36 F4 mice, and 20x coverage for the 11 F2 mice and 72 F5 mice. Sequencing was carried out on the Illumina Hiseq2500 platform for the 2 F0 mice and for 4 of the F2 mice, and on the Illumina

Hiseq4000 platform for the remaining 7 F2 mice and all of the F4 and F5 mice. Genomic DNA was fragmented to an average size of 500 bp and subjected to DNA library creation using established Illumina paired-end protocols (Nextera DNA Library Prep). Sequencing reads were aligned to mm10 using BWA[48] (v. 0.7.0) followed by Stampy[49] (v. 1.0.23, option bamkeepgoodreads). We then used Picard tools (v. 1.115) (http://broadinstitute.github.io/picard) to merge bam files from different lanes for the same sample and mark the duplicated reads. Then we used GenomeAnalysisTK-3.3-0 (GATK) to do local Indel realignment using known Indel targets between B6 and CAST from the 4th version of the Mouse Genome Project (MGPv4) data[50], followed by base quality score recalibration using known sites from SNPs between B6 and CAST in MGPv4, and then we called the variants using UnitedGenotyper in GATK. Next we used the Variant Quality Score Recalibrator (VQSR) from the GATK for variant filtration, where we used the set of variants present on the Affymetrix Mouse Diversity Genotyping Array as a set of true positive variation[51]. We used the annotations "HRun", "HaplotypeScore", "DP", "QD", "FS", "MQ", "MQRankSum", and "ReadPosRankSum" to train the VQSR, and we used a sensitivity threshold of 90% for the true positive set to define the set of newly genotyped sites that passed VQSR filtration. After filtration, about 16 million variants remained. To remove potential hidden heterozygous sites from the F0 individuals and to get a more stringent set of SNPs to start with, we intersected our SNPs with variants that have the homologous reference allele genotype for B6 and homologous alternative allele genotype for CAST from MGPv4[50]. Only SNPs with a PASS quality score were used. After filtering, we obtained 13,946,562 and 13,940,079 reliable autosomal SNPs from F2 samples and F5 samples, respectively, as informative markers to detect recombination events, or roughly one SNP for about every 170 bp.

HMM algorithm to identify events

Using the information from the filtered strain-informative SNPs, we developed a Hidden Markov Model (HMM) to infer the strain origin of each broad segment of the genome. In our HMM, the three possible emitted genotype states B6/B6, B6/CAST and CAST/CAST are represented by 0, 1 and 2, respectively (i.e. the number of CAST allele copies at each strain-informative SNP site). Similarly, the hidden states representing background strain origin are encoded as 0, 1 and 2 copies of a CAST haplotype. Emitted states may be different from hidden states due to sequencing errors or real converted events (e.g. observing a homozygous CAST genotype on an otherwise heterozygous CAST/B6 background). A natural initial stationary distribution is (0.25, 0.5, 0.25) corresponding to state triple (0, 1, 2). The state transition between two sites is driven by recombination events, with the distance between two different states following an exponential distribution with a rate parameter equal to twice the recombination rate. Here we adopted a genome-wide average constant recombination rate of $r=0.625*10^{-8}$ per base pair per generation[21,52]. Thus, the probability of recombination from site i to site j can be written as follows:

$$P_{ij}=1-\exp(-2rD_{ij}),$$

where P_{ij} and D_{ij} stand for the recombination probability and distance between site i and j, respectively. The transition probability matrix from site i to site j is as follows:

$$\mathbf{P}_{ij}=(1-P_{ij})\mathbf{I_3}+P_{ij}\mathbf{Q}, \qquad (2)$$

where I_3 is the 3×3 identity matrix and Q stands for the conditional transition matrix with the entry q_{mn} (m=0,1,2; n=0,1,2) describing the transition probability from state m to state n:

$$\mathbf{Q}=\begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix}. \qquad (3)$$

There is no transition from state 0 to state 2, or vice versa, because it's unlikely that two independent recombination events would happen at exactly the same position with a small sample size. Conditional on there being a recombination event, state 0 or state 2 transitions to state 1 with probability 1, and state 1 transitions to either state 0 or state 2 with equal probability.

Here we defined the emission probabilities from each hidden state by using the quality metrics from GATK for states 0, 1 and 2. Given state *g* in each site *t*, GATK provides a quality score S for three states as follows:

$$s_g^t = -10\log_{10}\frac{p(D|G_t=g)}{\max_{k=0,1,2} p(D|G_t=k)}, \quad (4)$$

where $p(D|G_t = g)$ is the probability that we observe the data D, conditional on the hidden state G_t being *g*. Since for each site *t*, the maximum score is constant, we can inversely infer the probability of observing different states with a constant scale factor:

$$p(D|G_t = g) \propto 10^{-\frac{s_g}{10}}. \quad (5)$$

In our analysis, the scaling parameter was arbitrarily set to 1.

We applied the forward-backward algorithm to infer the posterior distribution of hidden states. Starting with prior state probabilities (0.25, 0.5, 0.25) at the first site, the forward probability of state *j* after seeing the first t sites is

$$A_t(j) = \sum_{i=0}^{2} \alpha_{t-1}(i) p_{ij}\,(t-1) e_j(t), \quad (6)$$

where p_{ij}(t-1) is the (*i*,*j*)th element of transition matrix **P** at site t-1, and $e_j(t) = p(D|G_t = j)$ is the emission probability conditioned on state *j* at site t given by equation 5. At the same time, we define a backward chain with an initialized probability (1, 1, 1) at the end of the site using the following:

$$\beta_t(j) = \sum_{k=0}^{2} \beta_{t+1}(k) p_{jk}\,(t) e_j(t+1), \quad (7)$$

and the probability of hidden state j, given the observed data (j=0,1,2) at site t is

$$p_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=0}^{2} \alpha_t(i)\beta_t(i)}. \quad (8)$$

Finally, we can calculate the stationary distribution of states 0, 1 and 2 for each strain-informative SNP site given the sequencing data, and for each site we choose the hidden state with maximum probability as the real strain background state at that site.

NCO validation by direct Sanger sequencing

To validate a subset of NCO events detected by sequencing in F2 mice, we PCR amplified short regions (around 200 bp) overlapping the identified NCO sites using genomic DNA from the 2 F0 mice, the F2 mouse carrying the NCO, and up to 3 other related and/or unrelated F2 mice, using standard conditions (cycling conditions and primer sequences available upon request). PCR products were purified using the QIAquick PCR Purification Kit (Qiagen) and analysed by direct Sanger sequencing (Source Bioscience, UK). Sequence data comparison and analysis was carried out using Chromas LITE (version 2.1.1). Identification of SNPs allowed assignment of background and genotype at the tested locations, enabling identification of true NCOs or false positives. Of the 79 NCOs identified in F2 mice overlapping a hotspot, we randomly selected 19 NCOs overlapping a hotspot for validation, along with 11 NCOs not overlapping a hotspot (because we suspected the latter might include more false positives). Genotyping results confirmed these sites as genuine NCO events in all 19 cases. These results imply that the vast majority of NCOs identified that overlap hotspots are real events. Of the 11 NCOs identified in F2 mice that do not overlap a hotpot, 2 were ambiguous and manually removed from the final count post-filtering: they showed contradictory signals between the background and the genotype

of the converted base. In the first case, the NCO occurred in an apparently homozygous stretch surrounded by multiple SNPs. In the second case, while the heterozygous state of the NCO appeared correctly assigned, the background was inconsistent. Of the 9 NCOs taken forward for validation, 4 were confirmed by genotyping, a validation rate of 44%. These results suggest that very few real NCOs events occur outside PRDM9-bound hotspots. Given that 84.2% of our F2 NCO events overlap hotspots (Extended Data Table 2), we estimate an overall fraction of validated detected NCO events as 0.842 + 0.44 x 0.158, i.e. 91.1%.

Power to identify NCOs

To estimate the power of our method to detect NCO events of varying tract lengths, we simulated NCOs with different mean tract lengths and ran our pipeline for discovering NCO events, including our filters. Because F2 events are controlled by both $Prdm9^{Hum}$ and $Prdm9^{Cast}$ and F5 *de novo* events are controlled by $Prdm9^{Hum}$ alone, we performed two sets of simulations by using data from 11 F2 samples and 72 F5 samples. Because most recombination events overlap hotspots, we simulated NCOs in hotspot regions. For each mean tract length, we sampled 2000 hotspots with probabilities proportional to their H3K4me3 enrichment. Within each sampled hotspot, we sampled the centre of the NCO tract according to the distribution of NCOs around PRDM9 motifs after correcting for SNP density, and we sampled its tract length from an exponential distribution with a pre-defined mean tract length (which we varied from 10 to 100 bp with step size 10 and from 150 to 300 bp with step size 50). Sampled NCO tracts containing 0 SNPs were not counted as potentially detectable. Across these 2000 tracts, different animals possess different ancestral backgrounds. For each tract in each animal, we checked if any of the other animals has a different ancestral background consistent with a gene conversion event in the

first animal. If so, we sampled such a "donor" mouse (other events were ignored). We copied the sequencing information corresponding to the converted sites from the donor mouse, such as the allele depth, and we copied the sequencing information for the background from the recipient, such as mate-pair information. Then, we applied the same filters to this simulated sequencing data at each sampled tract. We calculated our power by dividing the total number of simulated tracts left after filtering by the total number of simulated tracts overlapping at least one SNP (Extended Data Figure 1a, b).

H3K4me3 ChIP-seq

We performed ChIP-seq against H3K4me3 in testes from an 8-week-old male (B6xCAST)F1-*Prdm9*$^{Hum/Cast}$ mouse C57BL/6J-*Prdm9*$^{Hum/Hum}$ mother, CAST/Eij father) as previously described[14] with several important modifications that increased ChIP stringency (noted here). Lysis was performed in 1% SDS lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris pH 8.0, 2x protease inhibitors). Sonication was performed in a Bioruptor Twin sonication bath at 4°C for three 5-minute periods of 30s on, 30s off at high power. Sonicated lysates were diluted 1:10 in IP wash buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1% NP-40, 1% Na deoxycholate, 2x Protease Inhibitor, filtered) instead of dilution buffer for antibody incubation. This yielded roughly 1 ng of ChIP DNA per testis. ChIP and total chromatin DNA samples were sequenced in multiplexed paired-end Illumina HiSeq2500 libraries (rapid run), yielding 63-71 million 51-bp read pairs per replicate after filtering (one ChIP replicate per testis plus one input sample). Sequencing reads were processed and peaks were called as described in our previous work[14,53]. Haplotype assignment of ChIP signal and removal of PRDM9-independent H3K4me3 peaks were performed as described[14]. The percentage of ChIP-seq read pairs originating from signal (as

opposed to background) was estimated to be 87.4%, a significant improvement over our prior, less stringent, experimental method (which yielded 62-71% of read pairs from signal) [14].

DMC1 ChIP-seq

DMC1 ChIP-seq data were generated elsewhere [16] and provided to us prior to final publication (separate manuscript currently under review). Briefly, single-stranded DNA sequencing (SSDS) DMC1 ChIP-seq was performed as previously described in Khil *et al*. 2012[54], using testes from one of the male F1 mice (B6xCAST)F1$^{Prdm9\ Hum/Cast}$ (C57BL/6J-*Prdm9*$^{Hum/Hum}$ mother, CAST/Eij father). ChIP and total chromatin DNA samples were sequenced in multiplexed paired-end Illumina HiSeq2500 libraries (rapid run), yielding 252 million 51-bp read pairs. We then processed the data for this study by following the algorithm provided by Khil *et al*. 2012 to map the reads to mm10 and obtain type I reads. We then used the same pipeline to call DMC1 peaks as described in Davies *et al*. 2016[14].

Estimation of NCO tract length for human-controlled and CAST-controlled events

To estimate NCO tract length, we assume the converted tract follows an exponential distribution with rate parameter λ, where 1/λ is the mean tract length. (If exponential tract lengths are not a fully accurate model, we can view this as a summary of tract properties, estimating the probability of co-conversion of pairs of markers as the distance between them increases.) We computed a composite likelihood function for our NCOs and estimated λ via maximal likelihood. Specifically, for each converted site, viewing this site as a "focal" site, we examine the SNPs nearby and record for each SNP its distance from the focal SNP, and whether that SNP is also converted. If the SNP is also converted, then it is still in the gene conversion tract, otherwise it is

not. Using this approach allows our approach to be independent of SNP density, because we are conditioning on SNP positions in our analysis. The probability that a SNP nearby a converted site is also converted is

$$\Pr(\text{SNP nearby converted}) = \Pr(\text{in}) = \mathrm{e}^{-\lambda d}$$

d is the distance from the nearby SNP to the converted site. The probability that a SNP nearby a converted site is not in the tract is 1-Pr(in). All the NCOs are independent so we can multiple these probabilities for each SNP in the windows to get the (composite) likelihood of the data:

$$\Pr(\mathrm{D}) = \prod\nolimits_{\text{all\_pairs}} \Pr(\text{in})^{x} \left(1 - \Pr\left(\text{in}\right)\right)^{1-x}$$

Here x=1 if the SNP nearby is also converted and x=0 otherwise. By maximise the likelihood using grid search for $1/\lambda$ from 1 to 1000 with step 0.1, we gained an estimate of tract length. Because pairs of SNPs are not in fact independent, this is not a true likelihood (though the resulting estimator is statistically consistent as the number of independent conversion events increases), and so to estimate uncertainty in the resulting estimates, we utilised bootstrapping of NCO events.

To perform bootstraps, we separated autosomal genomes into 258 non-overlapping 10 Mb blocks (the last block in each chromosome is shorter than 10 Mb). We resample 258 blocks with replacement, where the probability of sampling each block is proportional to the length of that block, and from the resulting bootstrapped set of NCO's, re-estimate tract length via the same procedure. Confidence intervals are calculated from a total of 10,000 bootstraps. We implemented this procedure for two sets of NCO events; those overlapping human-controlled, and those overlapping CAST-controlled, hotspots respectively.

Calculation of number of recombination events in one meiosis

We assume that the average number of DSBs per meiosis resolving as NCO events is K. Because each NCO affects only one of four chromatids, only one quarter of them will be seen in a single offspring.

We take F2 animals as an example. 22 meiosis occur, and generate 11 F2 animals. If D is hotspot SNP density, L is average NCO tract length, and "Power" represents the power to detect a SNP within a NCO event, then if N is the number of converted sites observed, we have:

$$E(N) = \frac{\mathrm{K}}{4} * 22 * \mathrm{Power} * \mathrm{L} * \mathrm{D}$$

Values for N, L, "Power" and D together allow estimation of K. We observe 0.0072 SNPs per bp within hotspots, and N=240 distinct converted sites in total; moreover, we estimate tract length L=30, and a power of 74.3% for these animals. This yields an estimate of $\widehat{K} = 274$ DSBs resolving as NCO events, per meiosis.

For CO events, we have near 100% power to observe these, and half of all recombination CO events are transmitted to a particular offspring. Therefore, based on 295 observed CO events in these mice, the (sex-averaged) estimated number of CO events is 295x2/22=26.8 per meiosis.

The sum of these numbers is the total number of autosomal events repairing using the homologous chromosome, per meiosis (we neglect the X-chromosome in this calculation). To obtain confidence intervals for the number of NCOs, COs and the total number of recombination events per meiosis and for the NCO to CO ratio, we performed bootstrapping as to estimate the

tract length of NCOs. For each bootstrapped sample (of 10,000), we obtained the number of NCOs and number of COs, and used these to re-estimate the total number of recombination events and the NCO/CO ratio.

Hotspot symmetry calculation

Sequence differences between the CAST and B6 genomes allowed us to quantify the fraction of ChIP-seq signal (either DMC1 or H3K4me3), coming from the B6 and CAST chromosomes. This also allows us to determine whether individual hotspots in these hybrids were 'symmetric', with DSBs occurring equally on both chromosomes, or 'asymmetric', with a preference towards either the CAST or B6 chromosome.

Using SNPs distinguishing the B6 and CAST genomes, each type I read pair from a hybrid DSB library (DMC1 ChIP-seq) is assigned to one of the categories 'B6', 'CAST', 'unclassified' or 'uninformative' as in[14], except we replace PWD with CAST. For each DSB hotspot, the B6 cutting ratio was then computed as the fraction of 'B6' reads mapped within 1 kb of the hotspot centre, over the sum of 'B6' and 'CAST' reads in that region.

We followed a similar approach for H3K4me3 ChIP-seq, but we further corrected for background signal, as described in[14]. For both DMC1 and H3K4me3, we only defined the B6 cutting ratio provided we had at least 10 informative reads.

To order hotspots based on their symmetry, if the fraction of cuts estimated on B6 and CAST chromosome respectively were x, and $1-x$, we defined the overall hotspot "symmetry" as $4x(1-x)$.

We obtained additional results for events initiating on a known homologue by using "homologous heat", defined as xh, where h is the estimated total heat of the hotspot, for events initiating on the CAST chromosome, and $(1-x)h$ for events initiating on the CAST chromosome. Note that separate estimates of hotspot symmetry and homologous heat may be obtained from both H3K4me3 and DMC1 ChIP-Seq data, for the same collection of hotspots. Because e.g. the H3K4me3 homologous heat captures how well the homologous chromosome is bound by PRDM9, it may be of stronger direct interest; however, this is only directly available for NCO events, whose initiating homologue is known. For CO events, to be conservative (even though we could attempt to make assumptions regarding conversion tracts to estimate homologous heat), we mainly used hotspot symmetry, which is strand-symmetric and ranges from 0 to 1 for hotspots with events completely on one chromosome, versus equally on both chromosomes[14]. For one set of plots (Extended Data Fig. 6), we used average homologous heat, defined as $2hx(1-x)$ (this averages homologous heat over the strand an event occurs on).

Calculating the fraction of asymmetric/symmetric hotspots containing a disrupting variant in the motif

To estimate the proportion of hotspots of different levels of initiation on B6/CAST chromosomes containing SNPs within their PRDM9 binding motifs, we first filtered to include only hotspots containing a clear motif (posterior probability >0.99). Secondly, we required at least 20 informative reads in our DMC1 data in order to accurately estimate the proportion of reads from B6, and 5 reads from each homologue covering the motif region, to ensure there are enough reads to identify variants if present. In Figure S5f, we then plot the fraction of hotspots in each

binned level of initiation on the B6 chromosome containing a SNP or Indel (as called by GATK prior to VQSR, or Platypus). We found 96% of identified highly asymmetric hotspots where this fraction was <5% or >95% contained such a SNP, after additionally requiring the P-value (binomial test) of asymmetry is $<10^{-10}$, to examine those hotspots most highly asymmetric.

Asymmetry rather than SNP density affects the generation of recombination events

We fitted a generalised linear regression model to discern whether hotspot asymmetry or local SNP density better predicts low CO and NCO rates. For each hotspot containing an identified PRDM9 binding motif, we indicate if there is a CO event overlapping this hotspot. We use this to produce a binary response vector, and fit a binomial generalised linear model. As predictors, we used:

(i) The symmetry of the hotspot

(ii) The log-transformed 'heat' of the hotspot measured by H3K4me3 (the H3K4me3 heat is incremented by a small value 0.0001 as there are a few hotspots with zero heat)

(iii) SNP densities around the PRDM9 binding motif at different scales (±100 bp, ± 500bp, ±800 bp)

We then tested various coefficients for significance, conditional on the others. We did the analysis for *Prdm9*^{Cast}-controlled COs (all of them were generated in the meiosis from F1 where there are two different *Prdm9* alleles) and *de novo Prdm9*^{Hum}-controlled COs (all of them were generated in the meiosis from F4 where there is only one type of *Prdm9* allele) separately to avoid the effect of competition between the two alleles. Results show that conditioned on the heat of H3K4me3 and symmetry of hotspots, SNP density has no significant effect on where

COs happen (p-values from all three scales >0.08) while both heat and symmetry of hotspots have significant positive effects on CO events ($p<0.05$).

For NCO events, we performed a similar analysis, except that we resampled the above hotspots according to the weight generated as described in the section "Rejection sampling for COs and NCOs, construction of Fig. 5 and Extended Data Fig. 6, and testing for impacts of asymmetry on event resolution" to account for higher power to detect NCOs when there is greater local SNP density**.** Some hotspots appeared several times after rejection sampling. The number of these hotspots that are indicated as overlapping a NCO depending on how many NCOs overlap this hotspot. Then we applied the same GLM analysis used for COs. All results show that SNP density has no significant effect on where NCOs happen conditional on the heat of H3K4me3 and symmetry of hotspots ($p>0.2$). For all the *Prdm9*^{Hum}-controlled NCOs, results show that the heat of H3K4me3 and symmetry of hotspots have significant positive effects on NCOs ($p<0.003$). Results from *Prdm9*^{Cast}-controlled NCOs also suggest positive effects on prediction of NCOs, but p-values are not significant (<0.2). We explained the weaker effect of symmetry for *Prdm9*^{Cast}-controlled NCOs in the last section of the supplementary material.

Figure Legends

Fig. 1| Study design and properties of crossover (CO) and non-crossover (NCO) events. a, Study design. Arrows indicate locations of *de novo* CO and NCO events. **b**, Detection of NCOs by comparing observed genotypes and background. **c**, **d**, Distribution of identified COs and NCOs across autosomes from F2 (**c**) and F5 (**d**) animals. **e**, Binning events by their distance to the telomere (x-axis), both NCOs and COs cluster at the telomeric region of chromosomes, more strongly for COs.

Fig. 2| DMC1, H3K4me3, and *Prdm9* allele predict CO and NCO properties. a, DMC1 and H3K4me3 peaks in a 50 kb region on Chromosome 10, with single NCO and CO events overlapping these peaks. **b**, **c**, DMC1 (**b**) and H3K4me3 (**c**) predict well where events occur. **d**, **e**, Dominance of $Prdm9^{Cast}$ over $Prdm9^{Hum}$. After splitting COs and NCOs within hotspots in F2 animals into those controlled by the $Prdm9^{Cast}$ or $Prdm9^{Hum}$ alleles, overlapping hotspots in the *Prdm9* knockout mouse, or non-identifiable (Unknown), $Prdm9^{Cast}$ dominates $Prdm9^{Hum}$ for both CO (**d**) and NCO (**e**) events, although occasionally knockout mouse hotspots are used. **f**, Correlation of underlying recombination rates between females and males (Supplementary Information), for rates binned at different scales (x-axis); dotted lines show 95% confidence intervals for true correlations. **g**, As **f**, but showing correlations between (sex-averaged) NCO and CO rates at different scales. **h**, Decay in probability that nearby SNPs are co-converted, with inter-SNP distance, conditional on a SNP being converted.

Fig. 3| NCOs, COs, DMC1 peaks and H3K4me3 peak positions relative to PRDM9 binding motifs. a, NCOs occurring within hotspots possessing robustly identified PRDM9 binding

motifs. Coloured dots are converted SNPs and grey lines represent upper bound of converted tracts. Yellow shading indicates the identified PRDM9 binding target. **b**, COs around PRDM9 binding motifs. Green dots are SNPs defining CO boundaries within grey delineating regions. COs that have large intervals (>2 kb) between the two defining SNPs are not shown in this plot. **c**, Density of COs occurring around motifs. Bar height at each position is proportional to the probability that break point happens at this position and density in each bin is averaged across the positions. **d**, Density of NCOs occurring around motifs. The distance between a NCO and motif is defined as the mid-point of minimal converted tract to the centre of the nearest identified hotspot motif. Distribution was normalised by SNP density in each bin to more power correct for increased power to see a NCO event where SNP density is high. **e**, **f**, Mean DMC1 and H3K4me3 ChIP-seq read coverage around motifs, for the hotspots shown in **a** and **b**. For DMC1, we separated plus strand (SSDS+) and minus strand (SSDS-) reads. Note x-axis scale differs from **c** and **d**.

Fig. 4| GC-biased gene conversion is absent in multi-SNP NCO tracts. a, Single-SNP tracts show a bias towards conversion of G/C bases (GC-bias), while there is no GC-bias in multiple-SNP tracts. **b**, GC-bias in groups of converted SNPs, binned according to their distance to the nearest SNP. SNPs nearby other SNPs show no detectable GC-bias. **c**, For each of the 12 possible combinations of NCO donor/recipient alleles (x-axis; e.g. A<-C converts recipient C to donor A), we plot the proportion of observed single-SNP NCOs of that type, relative to the corresponding proportion for the nearest non-converted markers, which lack GC-bias. Vertical lines: 95% CI's after pooling strand-equivalent pairs. Horizontal dotted lines: mean relative proportions for NCO events whose recipient types are G/C or A/T respectively, showing under-representation of events whose recipients are G/C could explain observed patterns.

Fig. 5| COs, NCOs are depleted in asymmetric hotspots. a, Human-controlled DMC1 hotspots were separated into 3 bins (asymmetric, intermediate, symmetric) according to symmetry, so that each bin contains the same number of predicted events according to DMC1 heat. Grey bars show the DMC1-predicted expected fraction of events in each bin. The four coloured bars (vertical lines: 95% CIs) show the observed fraction of (sampled) F5 *de novo* events: COs, NCOs, paternal recombination events and maternal recombination events. All are depleted in the asymmetric hotspots. **b**, As **a**, except predicted events were defined using H3K4me3.

Fig. 6| Model explaining influence of local genetic diversity on mismatch repair pathway choice. Three possible gene conversion tracts are depicted, differing in the number and type of heteroduplex mismatch sites on the recipient chromosome (blue). In the first case (left), a single A/T site on the recipient chromosome is converted in a strand-biased manner (perhaps by MMR) to the allele of the donor chromosome, regardless of donor base type (red). When the recipient chromosome contains a G/C at a single mismatch site (middle), a different repair mechanism (perhaps BER) operates 53% of the time and blocks gene conversion (Supplementary Information). The sum of these two effects can explain why 68% of observed gene conversions are converted to G/C. When a second mismatch is present nearby (right), repair reverts to a strand-biased mechanism, and no GC-bias is observed, except in rare complex NCO events.

Figure 1

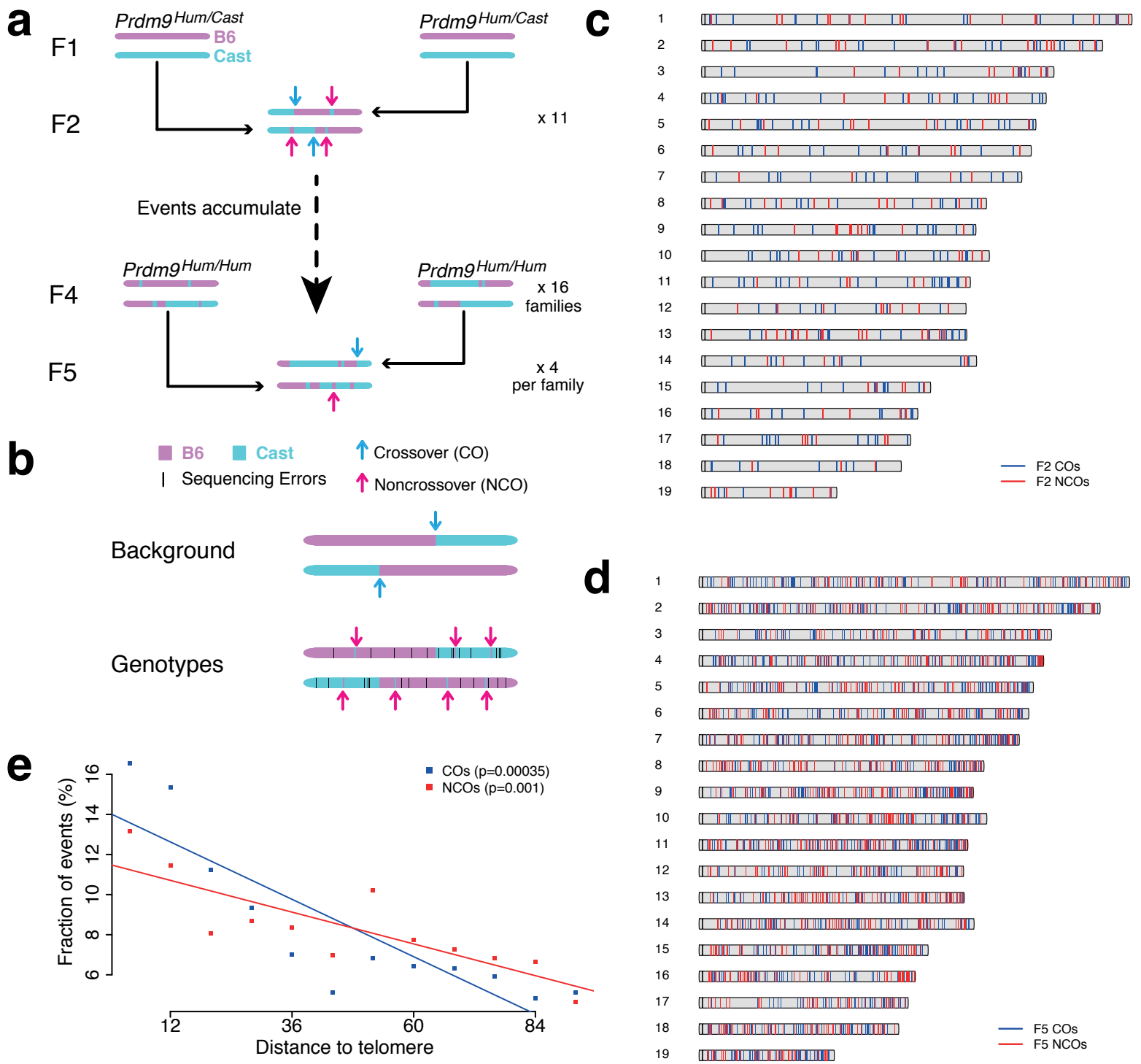

Figure 2

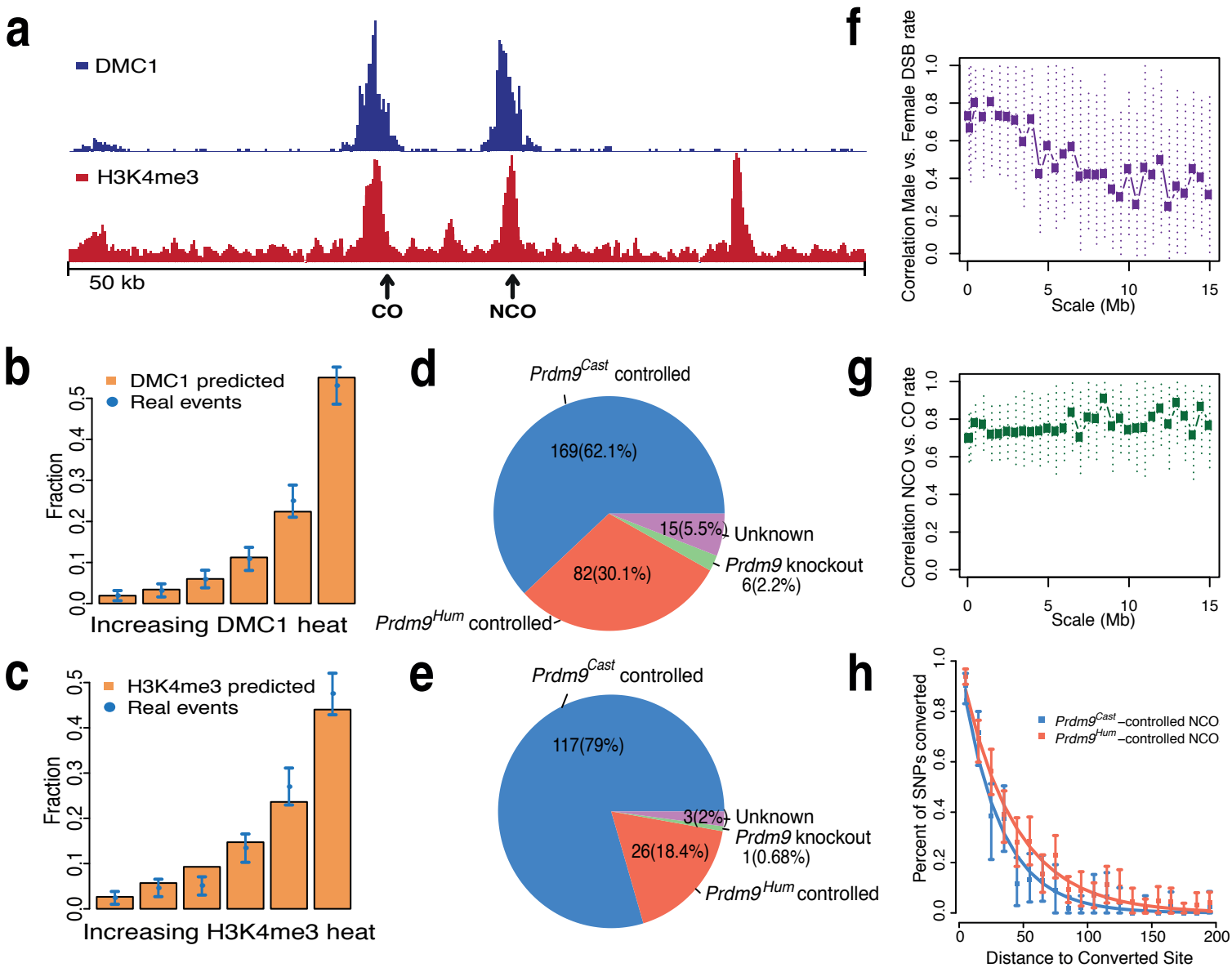

Figure 3

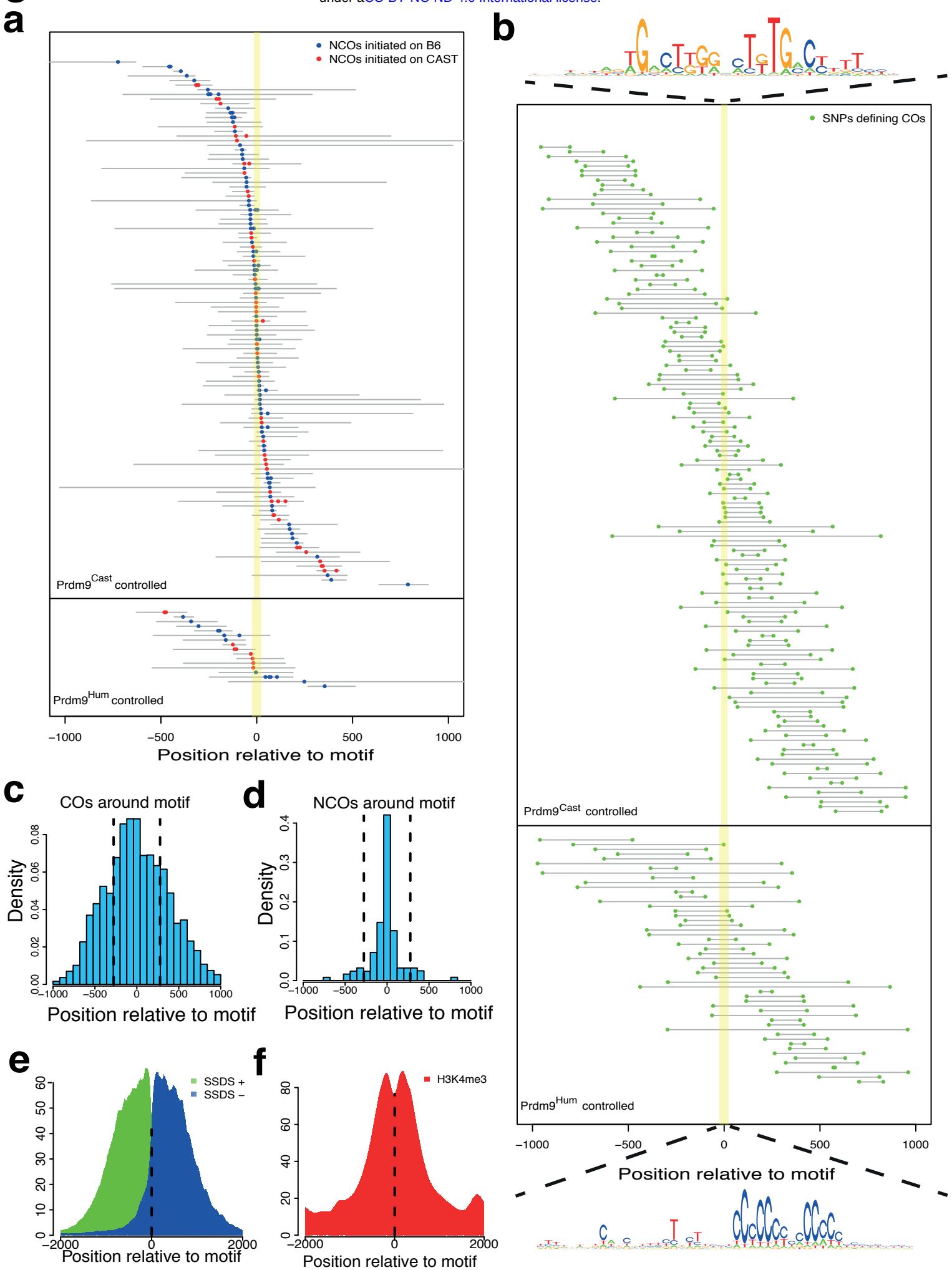

Figure 4

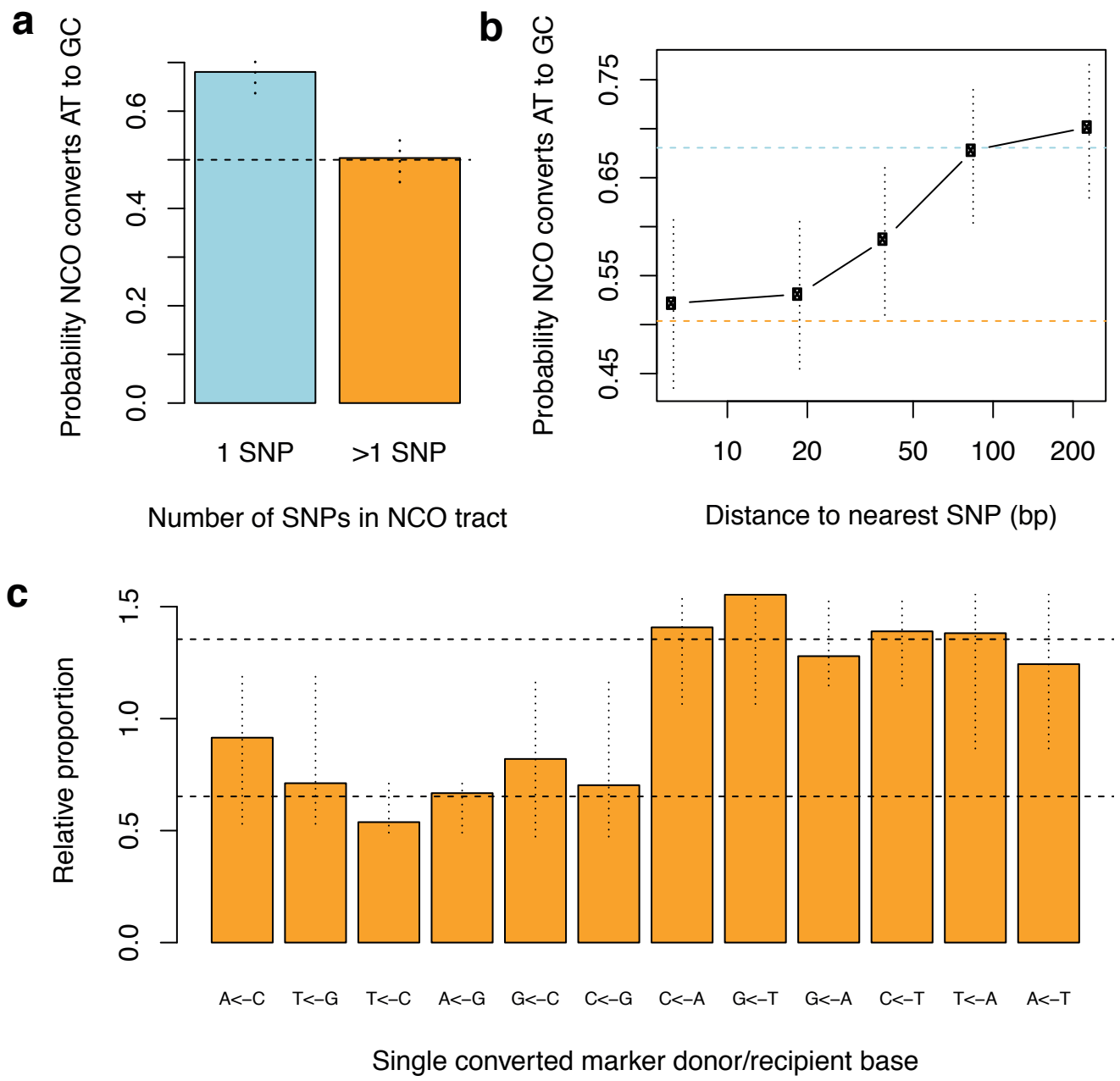

Figure 5

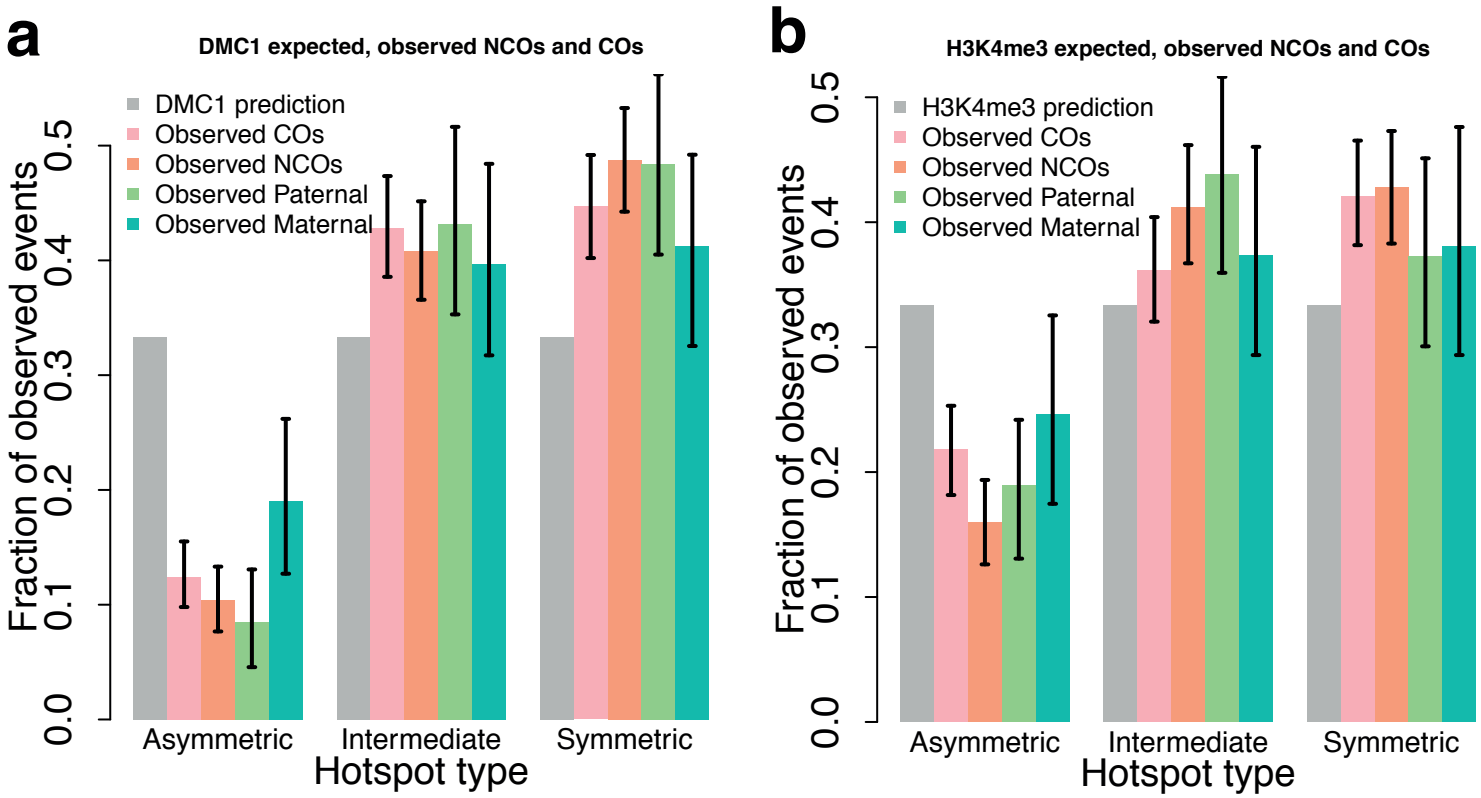

Figure 6