Computational reconstruction of non-crossover recombination

Marteinn Þór Harðarson

Dissertation submitted to the School of Technology, Department of Engineering, at Reykjavík University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

2026

Thesis committee:

Bjarni V. Halldórsson, Supervisor
Associate professor, Reykjavík University, Iceland

Daníel Guðbjartsson, Committee member
Associate professor, University of Iceland, Iceland

Jón Guðnason, Committee member
Professor, Reykjavík University, Iceland

Sorin Istrail, Examiner
Professor, Brown University, United States of America

Amy L. Williams, Examiner
Associate Professor, Brigham Young University, United States of America

ISBN 978-9935-577-05-4
ORCID 0000-0003-1130-8601

Table of contents

Abstract

As a diploid organism, humans inherit two copies of each autosomal chromosome, one from the father and one from the mother. The copy of a chromosome transmitted by a parent is reshuffle, via meiotic recombination, of the two parental copies through meiotic recombination. Most apparent are large chromosome parts of alternating chromosomal origin separated by crossover recombination. The other type of meiotic recombination is gene conversion, small segment of one of the parental homologous chromosomes on the background of the other. Gene conversions arise through process called non-crossover recombination (NCO). Recombinations can only be detected indirectly in offspring, at heterozygous markers, sites where the two homologous chromosomes differ. Sequencing errors, genotyping errors, phasing errors and structural variations can all mimic the signature of gene conversions making them difficult to be detected reliably. Not all NCOs create gene conversions, due to their short span and limit number of heterozygous markers. Large NCOs have a better chance to generate gene conversion than short NCO, due to higher likelihood of overlapping heterozygous markers. This makes it difficult to infer information about NCOs from gene conversions.

This thesis makes three main contributions. First, three-generation families are used to find and verify gene conversions showing age and sex dependent patterns for these events. Second, a statistical framework modelling the underlying NCO processes generating these gene conversions is developed in order to obtain the underlying length distribution and their quantity. Finally, combining crossovers and non-crossover recombinations to construct a complete human recombination map revealing their contribution to de novo mutagenesis.

Key words: meiosis, recombination, gene conversion, statistical modelling, partial observability inference

Útdráttur

Sem tvílitna lífverur erfir mannfólk tvö eintök af hverjum litningi, eitt frá föður og eitt frá móður. Eintakið sem foreldrið lætur í té er samblanda tveggja eintaka foreldrisins í gegnum meiótískar endurraðanir. Mest áberandi eru stórir litningshlutar upprunir á víxl frá á hvoru litningaeintakinu og eru aðskildir með krossunarendurröðun. Hin gerð meiótískra endurraðana eru genavendingar, þar sem lítill bútur af öðru eintakinu situr á bakgrunni svæðis sem er frá hinu eintakinu. Genavendingar verða til við ferli sem nefnast endurröðun án krossunar. Endurraðanir er einungis hægt að sjá óbeint í afkvæminu á stöðum sem eru arfblendnir, staðir þar sem tvö eintök litninga foreldrisins eru ólíkir. Villur í raðgreiningu, erfðaflokkun og fösun sem og erfðafæðilegir byggingarbreytileikar geta öll skilið eftir sig ummerki sem svipar til ummerkja eftir genevendingu sem verður til þess að erfitt er að greina þær með vissu. Þar sem enduraðanir án krossunar eru stuttar og þar sem arfblendnir staðir eru hlutfallslega fáir búa fár þeirra til genavendingar. Stórar endurraðanir án krossunar hafa því meiri líkur á að búa til genavendingar en smáar, þar sem þær eru líklegri til að ná yfir arfblendnar staðsetningar. Af þessum sökum er erfitt að ákvarða eiginleka enduraðana án uppstokkunar út frá genavendingum.

Þessi ritgerð hefur þrjú megin framlög. Í fyrsta lagi eru þriggja kynslóða fjölskyldur notaðar til að finna og staðfesta genavendingar sem sýnir að þær hegða sér ólíkt með aldri og kyni foreldris. Í öðru lagi er þróað tölfræðilegt líkan af undirliggjandi endurstokkunum án krossunar sem búa til genavendingarnar sem við sjáum og líkanið notað til að ákvarða fjölda og lengdardreifingu slíkar endurraðana. Að síðustu eru endurraðanir með og án krossunar settar saman til búa til heildar genetískt kort fyrir manneskjuna sem undirstrikar framlag þeirra til stökkbreytinga.

Efnisorð: meiósa, endurröðun, genevending, tölfræðileg líkanasmíð, ályktanir um að hluta til sýnileg gögn

List of publications

This thesis consists of three peer-reviewed journal papers and one pending paper submission, found in sections labelled Paper I-III.

Paper I

Halldorsson, B.V., Hardarson, M.T., Kehr, B. *et al.* The rate of meiotic gene conversion varies by sex and age. *Nature genetics* **48**, 1377-1384 (2016). https://doi.org/10.1038/ng.3669

In this paper I was responsible for family phasing and identifying gene converted markers.

Paper II

Hardarson, M.T., Palsson, G., Halldorsson, B.V. NCOurd: modelling length distributions of NCO events and gene conversion tracts. *Bioinformatics*, **39**, 8 (2023). https://doi.org/10.1093/bioinformatics/btad485

Paper III

Palsson, G., Hardarson, M.T., Jonsson, H., *et al*. Complete human recombination maps. Nature **639**, 700-707 (2025). https://doi.org/10.1038/s41586-024-08450-5

In this paper I was responsible for estimating the length distribution of NCO events as mixture of negative binomial distributions using NCOurd and PRDM9 motif analysis.

Other of publications

For all the following papers my contribution mostly entailed data preparation, data analysis and data visual representation.

Jónsson, H., Sulem, P., Kehr, B. *et al.* Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017). https://doi.org/10.1038/nature24018

Halldorsson, B. V., Palsson, G.; Stefansson, O. A., *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map, Science **363**, eaau1043 (2019). https://doi.org/10.1126/science.aau1043

Eggertsson, H. P.; Kristmundsdottir, S., Beyter, D. *et al*., GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs, Nature communications **10**, 5402 (2019). https://doi.org/10.1038/s41467-019-13341-9

Beyter, D., Ingimundardottir, H., Oddsson, A., *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits, Nature genetics **53**, 779-786 (2021). https://doi.org/10.1038/s41588-021-00865-4

Halldorsson, B. V., Eggertsson, H. P., Moore, K., *et al.* The sequences of 150,119 genomes in the UK Biobank, Nature **607**, 732-740 (2022). https://doi.org/10.1038/s41586-022-04965-x

Kristmundsdottir, S., Jonsson, H., Hardarson, M.T. *et al.* Sequence variants affecting the genome-wide rate of germline microsatellite mutations. *Nat Commun* **14**, 3855 (2023). https://doi.org/10.1038/s41467-023-39547-6

Gisladottir, R. S., Helgason, A., Halldorsson, B. V. *et al*. Sequence variants affecting voice pitch in humans, Science advances **9**, eabq2969 (2023). https://doi.org/10.1126/sciadv.abq2969

Stefansson, O. A.; Sigurpalsdottir, B. D.; Rognvaldsson, S. *et al*. The correlation between CpG methylation and gene expression is driven by sequence variants, Nature genetics **56**, 1624-1631 (2024). https://doi.org/10.1038/s41588-024-01851-2

UK Biobank Whole-Genome Sequencing Consortium, Whole-genome sequencing of 490,640 UK Biobank participants, Nature **645**, 692-701 (2025). https://doi.org/10.1038/s41586-025-09272-9

Acknowledgements

I would like to express my deepest gratitude to my family, who has supported me throughout my career and studies. I would also like to thank my friends and colleagues at deCODE genetics. Without their support, enthusiasm and insights this work would not have been possible.

Introduction

An individual inherits two near-identical sets of DNA from each parent. Each parent transmits a copy of their DNA to the offspring that differs slightly from the two copies carried by each parent, resulting in genetic diversity. Genetic diversity in humans is shaped by two fundamental mechanisms: mutations and meiotic recombination. Mutations introduce novel genetic variants into the DNA sequence, while meiotic recombinations reshuffle homologous chromosome pairs, creating new combinations of existing genetic variants. Genetic diversity in turn influences evolution by altering adaptability and contributes to the risk of various diseases.

Meiotic recombinations are of two types: crossover recombination where large chromosomal segments are reciprocally exchanged and non-crossover (NCO) recombination involving short tracts of DNA copied from one homologous chromosome to the other without reciprocal exchange. Crossovers have been widely studied in various species while NCOs are less well understood. However, NCOs' have short span and are only indirectly observable. When the genome of both the parent and the offspring is known NCOs can be partially inferred from studying differences between the parent's two homologous chromosomes and what the parent transmitted to its offspring. as gene conversion, where one or few parent's alleles (genetic differences) of one homologous chromosome appear on the background of alleles from the homologue, making NCOs difficult to study.

In this thesis we address the following research questions:

1. How can gene conversions be reliably detected at population scale using whole-genome data and pedigrees?

2. How can the number of underlying non-crossover events and their length distribution be inferred from gene conversions observed in a population?

3. How do non-crossover recombination patterns vary by the sex and age of the parent and how do they interact with de novo mutagenesis?

To answer these questions integration of large-scale genomic datasets together with development of algorithms and statistical models is required. At its core the problem is computational as the underlying biological process is only partially observable.

Genetic Markers and Inheritance

A **single-nucleotide polymorphism (SNP)** is position in the genome where the residing nucleotide differs between individuals. For example, one chromosome may carry an adenine (A) while another carries a cytosine (C). The **alleles** of SNP are the different nucleotides which can be seen at SNP and the **genotype** for the SNP is the combination of the two alleles carried at the SNP. SNPs are the most common form of genetic variations in humans and act as markers for tracking inheritance.

Humans are diploid organisms, meaning they have two copies of each autosomal chromosome: one inherited from the father and one from the mother. If the two alleles of a SNP differ the individual is said to be **heterozygous** at the position. Heterozygous SNPs are crucial for detecting recombinations because they allow for the two homologous chromosomes to be distinguished.

A **haplotype** is the collection of alleles residing at single copy of a chromosome. Genotype calls typically do not distinguish between alleles on the two homologous chromosomes. **Phasing** is the process of determining the two haplotypes for an individual. Phasing of both parent and offspring is essential for locating recombination events from the parent as recombinations are defined as changes in the haplotype origin along a chromosome.

Pedigrees provide the necessary structure to reconstruct inheritance patterns. A **pedigree** is a representation of familial relationships across generations, typically including parents and multiple offsprings. By comparing the genotypes of parents and offspring, one can determine the haplotype origin of the offspring. In three-generation-families the recombinations can be verified in the grandchildren providing additional confirmation of candidate recombination events.

Biological Basis of Non-Crossover Recombination

Meiotic recombinations are initiated with the formation of programmed double-strand breaks (DSBs) in DNA. These double-strand breaks are introduced by the SPO11 protein at locations influenced by PRDM9 binding specificity. Following the break formation the DNA ends around the breaks are resected creating single-stranded overhangs, which invade the homologous chromosome to repair the DSB.

Repairs of the DSB can follow alternative pathways. In crossovers, a structure called double Holliday junction (dHj) is formed. The dHj entangles the two homologous chromosomes in such a way that the information on haplotype origin is lost across the dHj. This allows for the transmitted chromosome to be a copy of the two different homologous chromosomes on either side of dHj after its resolution. In non-crossovers only one of the overhangs invades the homologous chromosome. The invading strand is then extended via DNA synthesis using the homologous chromosome as a template, after which the invading strand returns and bridges the gap created by the DSB. The newly synthesized DNA will differ from DNA on the chromosome subject to the DSB at heterozygous SNPs, creating heteroduplex DNA. If mismatch repair resolves these heteroduplexes in favor of the donor allele, the event becomes detectable as a gene conversion.

Importantly, many NCOs leave no detectable trace. If an NCO occurs in a region with no heterozygous markers, or if mismatch repair restores the original allele, no gene conversion is observed. Therefore, observed gene conversions only represent a biased subset of the underlying NCO events. The probability of observing an NCO as gene conversion depends on the size of the NCO tract, heterozygous marker density and the repair dynamics.

Understanding the true number of NCO events and their length distribution therefore requires modelling the relationship between the observed gene conversions and the latent NCO events.

Computational Structure of the Problem

At its core, the study of non-crossover recombinations is an inference problem under partial observability. The underlying biological events, double strand breaks and their repair outcomes are not directly measurable in human meiosis. Instead, we observe the offsprings genotypes and from them we attempt to reconstruct the hidden processes that generated them.

The detection of gene conversion events at a population-scale requires algorithms which operate on billions of transmitted alleles over large pedigrees. Gene conversions are short, many tracts consist of only a single SNP, and the algorithm must be able to distinguish them from sequencing errors, genotyping errors, phasing errors or signals generated by de novo mutations and structural variations, which all can mimic gene conversions. Robust identification of gene conversions therefore requires careful modelling of inheritance constraints and using stringent validation, for example in form of allele transmission in three-

generation families. Finally, it must be computationally and memory efficient to deal with the vast amount of data.

Even after identifying the gene conversions many key features of NCO remain latent such as their number and length distribution. The size of the non-crossover event is not the same as the span of the gene conversion it generates due to lack of heterozygous sites. Small NCOs are less likely to be observed as they are less likely to overlap heterozygous SNPs and are therefore underrepresented among the detected events while long events are more likely to be observed. To complicate matters, not all heterozygous markers overlapping an NCO will be gene converted as some may be reverted to their original state. A gene conversion tract is therefore a projection of a hidden event.

Estimating the number of non-crossovers per meiosis and their length distribution entails explicit modelling of detection probabilities depending on the underlying NCO tract length as well as the landscape of heterozygous markers around the observed event. In this thesis we use mixture models and expectation-maximization based algorithms to extract these latent distributions from the observed gene conversions. This modelling framework integrates marker density, gene conversion penetrance and event-length heterogeneity into unified inference procedure.

Finally, combining crossover and non-crossover data to construct complete human recombination map involves correcting for different detection biases across genomic regions, smoothing of discrete observations and quantifying regional sex specific double-strand break resolution patterns. The resulting maps are a computational reconstruction of the human recombination landscape derived from family-based whole-genome sequencing data. Additionally, they give valuable insights into their effect on human de novo mutagenesis.

Scope of the Thesis

The work presented in this thesis progresses from detection of events to statistical inference to population-scale application. First, gene conversions are identified and verified in three-generation families shedding light on age and sex dependent patterns for these events in the process. Second, development of a statistical framework modelling the underlying NCO processes generating these gene conversion events to obtain the underlying length distribution and their quantity. Finally, combining gene conversion with crossovers constructing a complete human recombination map revealing their contribution to de novo mutations.

Together, these studies provide quantitative and mechanistic characterization of non-crossover recombination in humans, from population-scale genomic data facilitated by computational modelling.

Paper I

Title:

The rate of meiotic gene conversion varies by sex and age

Authors:

Bjarni V. Halldorsson[1,2], Marteinn T. Hardarson[1], Birte Kehr[1], Unnur Styrkarsdottir[1], Arnaldur Gylfason1, Gudmar Thorleifsson1, Florian Zink1, Adalbjorg Jonasdottir1, Aslaug Jonasdottir1, Patrick Sulem1, Gisli Masson1, Unnur Thorsteinsdottir1,3, Agnar Helgason1,4, Augustine Kong1, Daniel F. Gudbjartsson1,5, Kari Stefansson1,3

Affiliations:

1 deCODE genetics / Amgen Inc., Sturlugata 8, Reykjavik, Iceland

2 School of Science and Engineering, Reykjavik University, Reykjavík, Iceland

3 Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland

4 Department of Anthropology, University of Iceland, Reykjavik, Iceland

5 School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland

*Correspondence to: Bjarni V. Halldorsson (Bjarni.Halldorsson@decode.is)

Abstract

Meiotic recombination involves a combination of gene conversion and crossover events that along with mutations produce germline genetic diversity. Here, we report the discovery of 3,176 SNP and 61 indel gene conversions. Our estimate of the non-crossover (NCO) gene conversion rate (G) is 7.0 for SNPs and 5.8 for indels per Mb per generation, and the GC bias is 67.6%. For indels we demonstrate a 65.6% preference for the longer allele. NCO gene conversions from mothers are longer than those from fathers and G is 2.17 times greater in mothers. Notably, G increases with the age of mothers, but not fathers. A disproportionate number of NCO gene conversions in older mothers occur outside double strand break (DSB) regions and in regions with relatively low GC content. This points to age-related changes in the mechanisms of meiotic gene conversions in oocytes.

Introduction

New combinations of alleles, generated through the shuffling of material between homologous chromosomes during meiosis, contribute to genetic diversity. Meiotic recombination is mostly thought to stem from programmed DNA double-strand breaks (DSBs)[1] generated by the evolutionary conserved SPO11 protein[2] at hotspot locations determined by the DNA-binding specificity of PRDM9[3]. These DSBs are either resolved through the reciprocal crossing-over of large regions between homologous chromosomes (often referred to as recombination in previous publications) or NCO gene conversion, the non-reciprocal transfer of short DNA segments between homologues. Crossover events, which arise through the resolution of the double Holliday junction (dHj)[4], are also frequently accompanied by gene conversions, hereafter referred to as CO gene conversions. NCO gene conversions most commonly arise via the synthesis-dependent strand annealing (SDSA) pathway[5], generating short regions of unidirectional homologous exchange[6,7], but the resolution of the dHj can also lead to NCO gene conversions[8,9]. A schematic overview of these DSB repair mechanisms is given in Supplementary Figure 1.

Repair through meiotic gene conversion uses sequence copied from a homologous chromosome. Heteroduplex DNA is formed in the recombination intermediates[10] and mismatched nucleotides in the heteroduplex DNA are subsequently corrected. A preference of strong (G/C) over weak (A/T) base-pairs of mismatch repaired nucleotides in converted segments, known as GC bias, has been previously observed[11]. Consequently, gene conversion plays a key role in shaping the genome's GC content[12,13] , has a potential confounding impact on mutation rate inferences and evolutionary divergence time estimates[14,15] and may skew allele frequencies increasing the disease burden of recessive alleles[16].

In previous publications, we quantified the rate of chromosomal crossover in meiosis, its determinants and genomic distribution in humans[17–19]. Here we report on the discovery of allelic gene conversions in meiosis inferred from the germline genotypes of probands and their relatives. We adopted a study design similar to that of Williams et al.[20], based on probands in three-generation families, where genotypes are available for a proband, its spouse, both its parents, at least two of its siblings and at least one child. This family structure was chosen to limit the impact of genotyping errors that can mimic gene conversion, such that haplotypes carried by the parents can be independently verified in the siblings and the gene conversion can be verified in the child. We searched for contiguous tracts of markers

(no longer than 100kb) consistent with non-reciprocal transfer of chromosome fragments in a transmission from parent to proband. This study design is adapted to the detection of NCO gene conversions, which are typically shorter than 100kb and where the converted tract is flanked by sequence from the original recipient chromosome on both sides[4,20]. Our study design does not allow for the detection of the subset of CO gene conversions, where the converted tract is flanked by sequence from the original recipient chromosome on only one side[4,8]. However, we are able to detect crossover recombination events that are accompanied by one or more such gene conversion events – hereafter referred to as complex crossover (CCO) gene conversions[20–22].

Using this approach, we sought gene conversions in two overlapping datasets. The first consists of 7,219 proband-family sets genotyped on Illumina HumanHap and Omni BeadChip arrays (*chip dataset*). The second consists of 101 whole genome sequenced (at > 30x) proband-family sets (*sequencing dataset*), 91 of which are contained in the first dataset (cf. Figure 1).

Results

As gene conversions can only be detected at polymorphic markers, we restricted our analysis to high quality SNPs and indels with a minor allele frequency over 0.5%. Indels were further restricted to be shorter than 10 bp, as longer indels are rare and yield less reliable genotypes. This restriction left 624,955 SNPs in the chip dataset and 8,195,014 SNPs and 465,054 indels in the sequencing dataset. For each proband-family set, we considered only variants where at least one parent was heterozygous, the haplotypes transmitted by both parents could be verified in a sibling and the gene conversion could be verified in an offspring. Using this approach, we assessed 214,241,663 marker proband pairs (*mpps*) in the chip dataset, yielding 2,192 mpps involved in a gene conversion (cf. Table 1,Table 2). In the sequencing dataset, 147,368,280 SNP and 8,715,873 indel mpps were assessed, yielding 1,027 and 61 gene converted mpps, respectively.

As most sequenced individuals also have chip data, it follows that a subset of 5,233,293 chip data mpps can be informative about the sensitivity and specificity of gene conversion calling in the two sets. From this subset 1,060,127 were omitted from the sequencing dataset due to quality control thresholds, leaving 4,173,166 overlapping mpps that were assessed in both datasets. All yielded concordant results, including an overlap of 43 gene conversions (cf.

Supplementary Note 1.1). We further confirmed phased haplotypes using read pairs (Supplementary Note 1.2, Supplementary Table 1) and genotypes using Sanger sequencing (Supplementary Note 1.3) and whole genome sequencing (Supplementary Note 1.4, Supplementary Table 2), which revealed error rates between 0.0% and 1.1%.

NCO gene conversions in DSB regions and recombination hotspots

We first examined the distribution of NCO gene conversions in the genome and their rate, G. As most such events are thought to be due to programmed DSBs, we compared our predicted NCO gene conversions to a map of DSB regions generated from spermatocyte samples of five human males[23], which have a mean size of 1,464 bp, s.d. 586bp. Table 1 shows that NCO gene conversions are highly overrepresented in spermatocyte DSB regions (odds ratio > 10). In paternal transmissions, the overrepresentation was 42.3 and 45.7 fold in the chip and sequencing data, respectively. In maternal transmissions, the overrepresentation was only 5.4 fold and 7.1 fold in chip and sequencing datasets, respectively, albeit highly significant (p-value < 0.001, for both datasets). As the locations of crossover recombination hotspots are known to differ between male and female meioses[18], and hot spots are largely determined by DSB regions[23], it follows that a stronger elevation of maternal G would be expected against a map of DSB regions from oocytes (such a map is currently not available). A previous study[20] also showed an overrepresentation of NCO gene conversions in male DSB regions and indications of a sex difference in their localization. Recent research has shown that PRDM9 alleles carried by the parent strongly influences the locations of DSBs[23] and consequently the locations of crossovers and NCOs in chromosomes transmitted to the offspring. Our results confirm these findings. The PRDM9 allele of the proband strongly affects the distribution of NCO gene conversion, but not their overrepresentation in DSB regions (cf.Supplementary Note 2, Supplementary Table 3).

We next compared NCO gene conversions to a sex specific map of crossover recombination hotspots[24]. Crossover recombinations, when estimated in the sexes separately, are enriched 38.0 and 27.6 fold in male and female crossover recombination hotspots, respectively[17]. Table 1 shows that for males, the enrichment of G in male crossover recombination hotspots is 8.2 and 12.7 fold in chip and sequencing datasets, respectively. For females, the enrichment of G was 4.6 and 8.0 fold in female crossover recombination hotspots.

NCO gene conversion rate

Williams et al.[20], estimated G as 5.9/Mb/generation, using a genomewide approach similar to the one presented here (but restricting to events shorter than 5kb). Comparable estimates have been obtained based on sperm genotyping[6,7] (not genomewide) and coalescent inferences[14,15,25]. Our genomewide estimate of G is 9.5/Mb/generation in the chip dataset (unadjusted for SNP ascertainment) and 5.9/Mb/generation in the sequencing dataset (cf. Table 1). While the sequencing dataset is minimally affected by ascertainment bias, the markers on Illumina BeadChip arrays are preferentially selected in genomic regions of low linkage disequilibrium[26], i.e. regions with a high rate of crossover recombination (and underlying DSBs). After correcting for this ascertainment bias in the chip dataset, we estimate the genomewide G to be 7.0/Mb/generation (95% CI 6.0–8.0). The difference between this estimate and the one obtained from the sequencing data is not significant (p-value: 0.12) and both are consistent with previous estimates.

Genomic distribution and sex differences

An assessment of the genomic distribution of G in the two sexes (cf. Supplementary Note 3) shows that G is elevated near telomeres and therefore decreases with chromosome length (Figure 2b). This pattern is analogous to that observed for crossover recombinations[18,27] and is consistent with a higher stationary GC-content near telomeres[28]. Moreover, as in the case of crossover recombinations, the proportion of events near telomeres is greater in fathers than in mothers. Overall, G is 2.17 (95% CI 1.94–2.45) and 1.91 (95% CI 1.34–2.58) times higher in maternal transmissions than paternal transmissions when assessed in the chip and sequencing datasets, respectively. This sex difference is very similar to that observed for the crossover recombination rate[17], which is 2.03, although the difference in G can largely be attributed to longer events in maternal transmissions.

G increases with maternal age

As previous studies have reported an age-related increase in the crossover recombination rate in females that is not seen in males[19,29], we examined the impact of age on G. In the chip dataset, there is a marked increase of G with maternal age of 0.58/Mb/year (95% CI 0.38–0.78, p-value: $1.4 \cdot 10^{-8}$) that is not observed in fathers (Figure 2a, Supplementary Note 4). Based on a very small number of events, the maternal age effect in the sequencing dataset is 0.33/Mb/year (95% CI –0.18 – 0.83). Although this estimate is not significantly different

from 0 (p-value: 0.21), this effect is also not significantly different from the chip data estimate (p-value: 0.35).

Interestingly, the increase of G with maternal age in the chip is comparable inside and outside of both crossover recombination hotspots and male DSB regions (Figure 2c,d), resulting in a decrease of odds ratios for co-occurence with these genome features (cf. Supplementary Note 5, Supplementary Figure 2). The average female crossover recombination rate of maternally transmitted NCO gene conversions further decreases with age (cf. Supplementary Figure 3). Thus, the NCO gene conversions that accumulate with age appear to be less tied to programmed DSBs than those transmitted by younger mothers.

G increases with local GC content

G increases with local GC content, defined as the GC content of the 100 base pairs surrounding each mpp (cf. Supplementary Note 6), a result consistent with GC-biased NCO gene conversions occurring repeatedly at similar locations in evolutionary history. Local GC content is elevated in DSB regions (cf. Supplementary Figure 4 and Supplementary Table 4), in part due to the GC composition of PRDM9 motif[30]. We consider the correlation with local GC content inside and outside of DSB regions separately as well as distinguishing between the chip and sequencing datasets and paternal and maternal transmissions. Remarkably, even after we have conditioned the data on DSB region status, in seven out of eight cases G is positively correlated with local GC content (Figure 3 and Supplementary Table 5).

In the chip dataset, we observe a decrease (p-value: 0.0068) in local GC content with age in maternally transmitted NCO gene conversions. As this result might be attributed to the high GC content of the PRDM9 binding motif, we restricted to mpps outside of male DSB regions, where we also observed a decrease (p-value: 0.0037) (Figure 2e). The result also remained significant after adjustment of local GC content for the HapMap based recombination rate[31] (p-value: 0.0029). Again, this suggests that the NCO gene conversions that accrue in aging mothers are different in mechanism from those found in younger mothers.

We investigated whether G was dependent on the SNP type, the base pair composition of the SNP‘s two alleles, and found no such effect (cf.Supplementary Table 6).

Complex NCO events

We grouped gene converted mpps into events, based on their proximity. In NCO gene conversions, most events contain only a small number of gene converted mpps, with an average of 1.24 mpps per event for the chip dataset and 1.78 gene per event for the sequencing dataset. The smaller number of mpps per event for the chip dataset than the sequencing dataset can in part be attributed to the lower marker density. Interestingly, in NCO gene conversions, maternally transmitted events are tagged by more mpps than paternally transmitted events; 1.37 vs. 1.04 (p-value < 0.001) and 2.34 vs. 1.23 (p-value < 0.001) mpps per event for chip and sequencing data, respectively. However, we did not observe an age dependence in the number of mpps per event (cf. Supplementary Figure 5).

We partitioned the set of NCO gene conversions into short and long NCO events, based on a distance of 1,000 bp (roughly the size of a DSB region), between the first and last gene converted mpp per event. Supplementary Table 7 shows the length distribution of long NCO events by distance between the first and last marker. Due to the denser marker set, the length of the event can be better estimated in the sequencing dataset, hence some events classified as short NCO events in the chip data would be classified as long NCO events, if the denser sequencing data were available (cf. Supplementary Note 7). Table 1 shows that short NCO events are highly overrepresented in male DSB regions and crossover recombinations hotspots, while long NCO events are not overrepresented in male DSB regions, but are overrepresented in crossover recombination hotspots.

The tracts of long NCO events contain both gene converted mpps and non-gene converted mpps (cf. Supplementary Note 7, Supplementary Table 8). A similar pattern has been previously observed[20] and are referred to as complex NCO gene conversions. Within complex events both the gene converted and non-gene converted mpps show a GC bias.

In the chip and sequencing datasets we estimate that, respectively, at least 46.1% and 65.3% of all long NCO events are complex (cf. Supplementary Table 9). The true rate of complex events is likely to be higher (cf. Supplementary Note 7), leading us to hypothesize that all long NCO events may be complex. These long, and mostly complex, NCO events are more common in maternal transmissions (p-value < 0.001, for chip and sequencing datasets). A significant increase in G with mother‘s age is observed both for short and long events (cf. Supplementary Note 7).

Gene conversion of indels

We estimate G for indels as 5.8/Mb/generation (95% CI 4.1–7.9), comparable to that for SNPs. Our results show a bias of 65.6% (95% CI 53.3–76.6, p-value: 0.018, cf.Table 3) toward the longer allele for indels in allelic gene conversions (cf. Supplementary Note 8). A direct estimate of gene conversions involving indels has to our knowledge not been previously reported. Comparisons between species have yielded conflicting results; a deletion bias has previously been reported for non-allelic gene conversion[32], while a bias towards insertion has previously been reported for allelic gene conversion[33].

Complex crossover (CCO) gene conversions

In crossovers we are only able to detect complex events (cf. Supplementary Note 9, Supplementary Figure 6). The rate reported for CCO gene conversions should be interpreted with caution, as it refers to the fraction of mpps within a distance of 100kb from a crossover recombination that show evidence of gene conversion (cf. Supplementary Note 9) and is not a genomewide rate. We observe a greater CCO gene conversion rate in the sequencing dataset, where more events are detectable (cf.Table 2). Due to our inability to detect all complex events the true CCO gene conversion rate is likely to be higher than the estimates in both datasets.

The CCO gene conversion rate is greater than the NCO gene conversion rate (p-value <0.001 chip dataset, < 0.001 sequencing dataset). This confirms that as a group CCO gene conversions are not independent of crossovers. How CCO gene conversions are related to crossover recombinations remains to be elucidated.

A large majority of the CCO gene conversions we identified are maternal, demonstrating that complex crossovers are more common in maternal transmissions, as is the case for NCO gene conversions. Another similarity is that, we observe an increase in CCO gene conversion rate with maternal age of 14.0/Mb/year (95% CI 0.7– 27.3, p-value: 0.04), in the chip dataset (cf. Figure 4a). Moreover, the fraction of crossovers that are complex increases with maternal age (p-value: 0.02 in the chip dataset) (cf. Supplementary Note 9,Figure 4b). In the sequencing dataset, we observe that 0.31% of all paternally transmitted crossover recombinations are complex. This result is in close agreement with a previous estimate of 0.33% obtained using sperm analysis[22].

GC bias

Like previous studies[7,20], our results reveal a significant GC bias for NCO gene conversions[4], where strong base pairs (G or C) preferentially appear on polymorphic gene converted base pairs (cf.Table 1): the GC bias is 67.6% (95% CI 65.7–69.8) in the chip dataset and 69.3% (95% CI 65.8–72.3) in the sequencing dataset. Short and long NCO events exhibit the same GC bias (cf.Table 1).

Our results (cf. Supplementary Note 10) indicate that the bias is greater in maternal than paternal transmissions (p-value: 0.032 chip, 0.004 sequencing). Further, CpG SNPs show a greater GC-bias than other SNPs (p-value: 0.038 chip, <0.001 sequencing).

The GC bias in CCO gene conversions is 70.2 (95% CI 62.5-77.8) in the chip dataset and 70.1 (95% CI 63.1-78.8) in the sequencing dataset, which is not significantly different from that observed in NCO gene conversions (p-values were: 0.56 and 0.73 for the chip and sequencing data, respectively).

Discussion

In summary, we have used both SNP chip and whole genome sequencing datasets from three-generation families to search for meiotic gene conversions in humans. Overall, we identified 3,237 mpps involved in gene conversions. Based on the sequencing data, we obtained a sex-averaged estimate of G, as 5.9/Mb/generation. Crucially, our results demonstrate that G varies with both age and sex. Thus, the rate for mothers (7.7/Mb/generation) is 1.91 (95% CI 1.34 – 2.58) times greater than for fathers (4.1/Mb/generation), in the sequencing dataset. Given that the fraction of heterozygous loci (where gene conversions can be detected) in Icelanders[34] is $6{\cdot}10^{-4}$, it follows there is an expectation of 7 detectable NCO gene conversions from fathers and 14 from mothers. These numbers are 12 and 23, respectively, based on a worldwide average heterozygosity[35] of $1{\cdot}10^{-3}$.

A surprising result was the magnitude of the age-related increase of G in females, where we estimate a 2.42 (95% CI 1.83–3.10) fold increase from the ages of 20 to 40 years. In the case of crossover recombination, the estimated increase between these age points is only 1.042 fold (4.2%), which is thought to be the result of greater viability of eggs with more crossover events[19,29]. Such selection might account for some of the age-related increase of G in females. However, given the more drastic increase in G, some other age-related factor must be

at work. The increased G in older mothers is less biased towards male DSBs, sex specific crossover recombination hotspots (cf. Supplementary Note 5) or regions of high local GC content (cf. Supplementary Note 6). Although our results are not conclusive, they indicate that a large fraction of the increase in crossover recombinations with maternal age are due to complex crossover recombinations (cf. Supplementary Note 9).

Overall, the large fraction of NCO gene conversions in spermatocyte DSB regions and crossover recombination hotspots is consistent with the view that most of them occurred in response to programmed DSBs prior to the meiotic prophase 1 arrest of oocytes in the fetal ovary[36]. However, an accumulation of programmed DSBs over subsequent decades does not seem a likely source of the age-related increase of NCO gene conversions in females. Other possible sources may be linked to the age-related deterioration of the oocytes across the decades that they are in dictyate arrest, possibly leading to non-disjunction[37,38]. This deterioration may be due to damage-induced DSBs, deficiencies in checkpoint mechanisms[36,39], failure of cohesins to maintain the cohesion of sister-chromatids[36,40] or that cohesive linkages are not restored at the same rate as they are lost[38]. Further research is needed to determine the source of these additional NCO gene conversions in older females and whether they have the same source as the additional CCO gene conversions. It is obviously interesting to note in this context that risk of aneuploidies increases drastically with the age of mothers[41] – although there is no direct evidence to link aneuploidies with the age-related increase of gene conversions.

Our results further demonstrate that the control of gene conversions differs between the sexes. While most paternally transmitted NCO events are short and complex crossovers are rare in paternal transmissions, maternally transmitted NCO events tend to be long and complex. Our results suggest that there is a different biological mechanism underlying short and long NCO events and consequently maternal and paternal transmissions. As the definition of an event is based solely on proximity of gene converted mpps, we cannot discern whether the gene converted mpps within the same long event occurred simultaneously in a single process or in several collocated processes. Crossover interference has recently been shown to decrease with maternal age[42], possibly leading to double crossover recombination events which in our study design could be detected as long NCO events. The complex nature of long NCO events and their GC bias make it unlikely that crossover interference explains a large fraction of long NCO events (cf. Supplementary Note 7). Paternal NCO gene conversions may be enriched for those derived from the SDSA pathway and maternal NCO gene conversions may be

enriched for those resulting from dHj resolution[4]. The long NCO events, which are complex and mostly maternally transmitted may also arise from a more complex set of underlying biological mechanisms[43], including repeated template switching[44]. Analysis of sperm[6,22], oocyte[45] or tetrad analysis[8] are promising approaches for obtaining a more complete picture of meiotic gene conversion and its mechanisms.

Our results and others[12,13,20] show that gene conversions are biased towards GC base-pairs, while mutations are biased towards AT base-pairs and increase with age in both sexes[46,47], but more strongly with father‘s age[48]. Now it is clear that gene conversions increase with mothers' age. On average, the number of gene conversions per generation is comparable to that of mutations. Intriguingly, this means that the nucleotide composition of the human genome represents an equilibrium that is maintained by an unwitting battle between the sexes, where male driven AT-biased mutations[48] are offset by female driven GC-biased gene conversion events.

Acknowledgements:

This work was supported in part by NIH (NIDA) (R01–DA017932).

Author contributions:

BVH, DFG and KS designed the experiments. BVH wrote the first draft of the paper. BVH, MTH, BK, US, PS, AH, AK, DFG and KS reviewed and contributed to subsequent drafts of the paper. BVH, MTH and AG implemented the methodology. BVH, MTH and BK prepared tables and figures. AsJ and AdJ performed the Sanger sequencing. UT oversaw the operations of the genotyping facility. BVH, MTH, FZ, GT, AG and GM processed the data. BVH and MTH analyzed the data. All authors contributed to the final version of the manuscript.

References for main text

1. Sun, H., Treco, D., Schultes, N. P. & Szostak, J. W. Double-strand breaks at an initiation site for meiotic gene conversion. *Nature* **338,** 87–90 (1989).

2. Lam, I. & Keeney, S. Mechanism and regulation of meiotic recombination initiation. *Cold Spring Harb. Perspect. Biol.* **7,** a016634 (2015).

3. Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327,** 836–40 (2010).

4. Haber, J. *Genome stability*. (Garland Science, 2013).

5. McMahill, M. S., Sham, C. W. & Bishop, D. K. Synthesis-dependent strand annealing in meiosis. *PLoS Biol.* **5,** e299 (2007).

6. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* **36,** 151–6 (2004).

7. 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,** e1004106 (2014).

8. Cole, F. *et al.* Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat. Genet.* **46,** 1072–80 (2014).

9. Allers, T. & Lichten, M. Differential Timing and Control of Noncrossover and Crossover Recombination during Meiosis. *Cell* **106,** 47–57 (2001).

10. Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. & Stahl, F. W. The double-strand-break repair model for recombination. *Cell* **33,** 25–35 (1983).

11. Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics* **159,** 907–911 (2001).

12. Glemin, S. *et al.* Quantification of GC-biased gene conversion in the human genome. *Genome Res.* **25,** 1215–1228 (2015).

13. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285–311 (2009).

14. Narasimhan, V. M. et al. A direct multi-generational estimate of the human mutation rate from autozygous segments seen in thousands of parentally related individuals. bioRxiv (Cold Spring Harbor Labs Journals, 2016). doi:10.1101/059436

15. Palamara, P. F. *et al.* Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. *Am. J. Hum. Genet.* **97,** 775–789 (2015).

16. Lachance, J. & Tishkoff, S. A. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am. J. Hum. Genet.* **95,** 408–420 (2014).

17. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–1103 (2010).

18. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31,** 241–247 (2002).

19. Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nat. Genet.* **36,** 1203–1206 (2004).

20. Williams, A. L. *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife* **4,** e04637 (2015).

21. Guillon, H., Baudat, F., Grey, C., Liskay, R. M. & de Massy, B. Crossover and noncrossover pathways in mouse meiosis. *Mol. Cell* **20,** 563–73 (2005).

22. 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. U. S. A.* **105,** 10471–10476 (2008).

23. Pratto, F. *et al.* Recombination initiation maps of individual human genomes. *Science* **346,** 1256442–1–9 (2014).

24. Kong, A. *et al.* Common and low-frequency variants associated with genome-wide recombination rate. *Nat. Genet.* **46,** 11–16 (2014).

25. Padhukasahasram, B. & Rannala, B. Meiotic gene-conversion rate and tract length variation in the human genome. *Eur. J. Hum. Genet.* (2013). doi:10.1038/ejhg.2013.30

26. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38,** 659–62 (2006).

27. Pardo-Manuel de Villena, F. & Sapienza, C. Recombination is proportional to the number of chromosome arms in mammals. *Mamm. Genome* **12,** 318–22 (2001).

28. Duret, L. & Arndt, P. F. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **4,** e1000071 (2008).

29. Martin, H. C. *et al.* Multicohort analysis of the maternal age effect on recombination. *Nat. Commun.* **6,** 7846 (2015).

30. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40,** 1124–1129 (2008).

31. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

32. Assis, R. & Kondrashov, A. S. A strong deletion bias in nonallelic gene conversion. *PLoS Genet.* **8,** e1002508 (2012).

33. Leushkin, E. V & Bazykin, G. A. Short indels are subject to insertion-biased gene conversion. *Evolution* **67,** 2604–2613 (2013).

34. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).

35. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

36. Handel, M. A. & Schimenti, J. C. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat. Rev. Genet.* **11,** 124–136 (2010).

37. Subramanian, V. V & Bickel, S. E. Aging predisposes oocytes to meiotic nondisjunction when the cohesin subunit SMC1 is reduced. *PLoS Genet.* **4,** e1000263 (2008).

38. Weng, K. A., Jeffreys, C. A. & Bickel, S. E. Rejuvenation of meiotic cohesion in oocytes during prophase I is required for chiasma maintenance and accurate chromosome segregation. *PLoS Genet.* **10,** e1004607 (2014).

39. Leland, S. *et al.* Heterozygosity for a Bub1 mutation causes female-specific germ cell aneuploidy in mice. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12776–12781 (2009).

40. Hodges, C. A., Revenkova, E., Jessberger, R., Hassold, T. J. & Hunt, P. A. SMC1beta-deficient female mice provide evidence that cohesins are a missing link in age-related nondisjunction. *Nat. Genet.* **37,** 1351–1355 (2005).

41. Nagaoka, S. I., Hassold, T. J. & Hunt, P. A. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat. Rev. Genet.* **13,** 493–504 (2012).

42. Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D. & Auton, A. Escape from crossover interference increases with maternal age. *Nat. Commun.* **6,** 6260 (2015).

43. Martini, E. *et al.* Genome-wide analysis of heteroduplex DNA in mismatch repair-deficient yeast cells reveals novel properties of meiotic recombination pathways. *PLoS Genet.* **7,** e1002305 (2011).

44. Tsaponina, O. & Haber, J. E. Frequent Interchromosomal Template Switches during Gene Conversion in S. cerevisiae. *Mol. Cell* **55,** 615–25 (2014).

45. de Boer, E., Jasin, M. & Keeney, S. Local and sex-specific biases in crossover vs. noncrossover outcomes at meiotic recombination hotspots in mouse. bioRxiv (Cold Spring Harbor Labs Journals, 2015). doi:10.1101/022830

46. Wong, W. S. W. *et al.* New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7,** 10486 (2016).

47. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48,** 935–939 (2016).

48. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).

Figure legends for main text

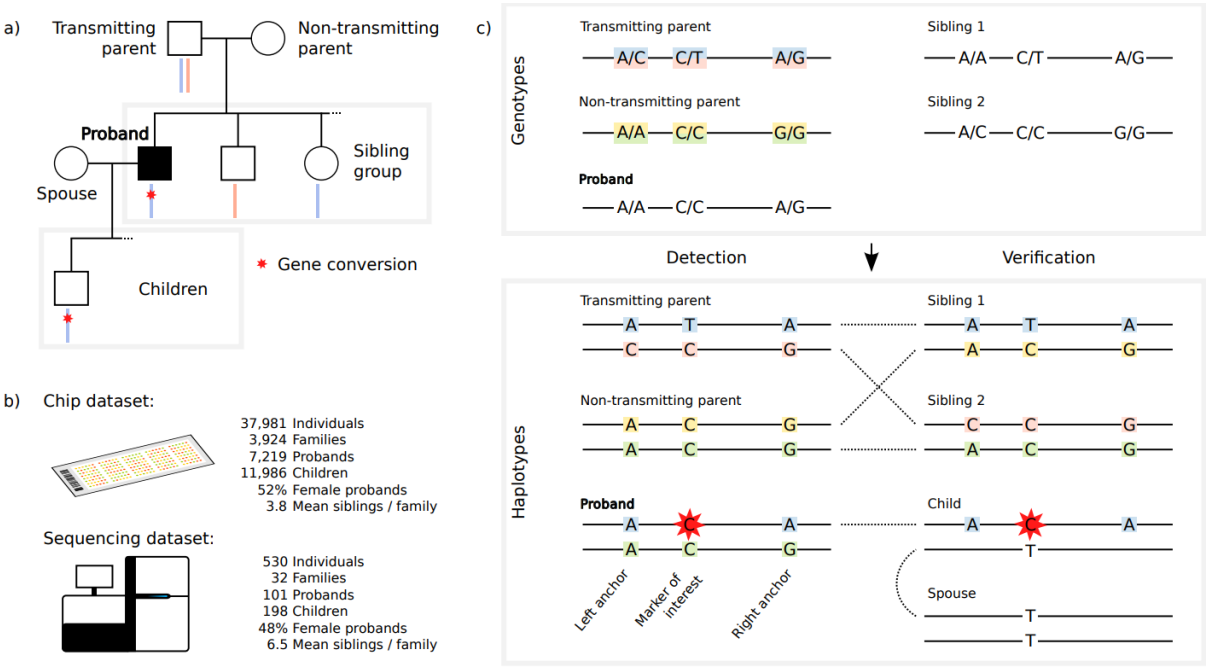


Figure 1 The study design and method used for detecting gene conversions. Including a) the family structure and b) the two sources of genotypes. c) The genotypes of the siblings are necessary to verify the haplotypes of both parents. We require at least one child of the proband to verify the gene conversion. The method is shown for NCO gene conversion.

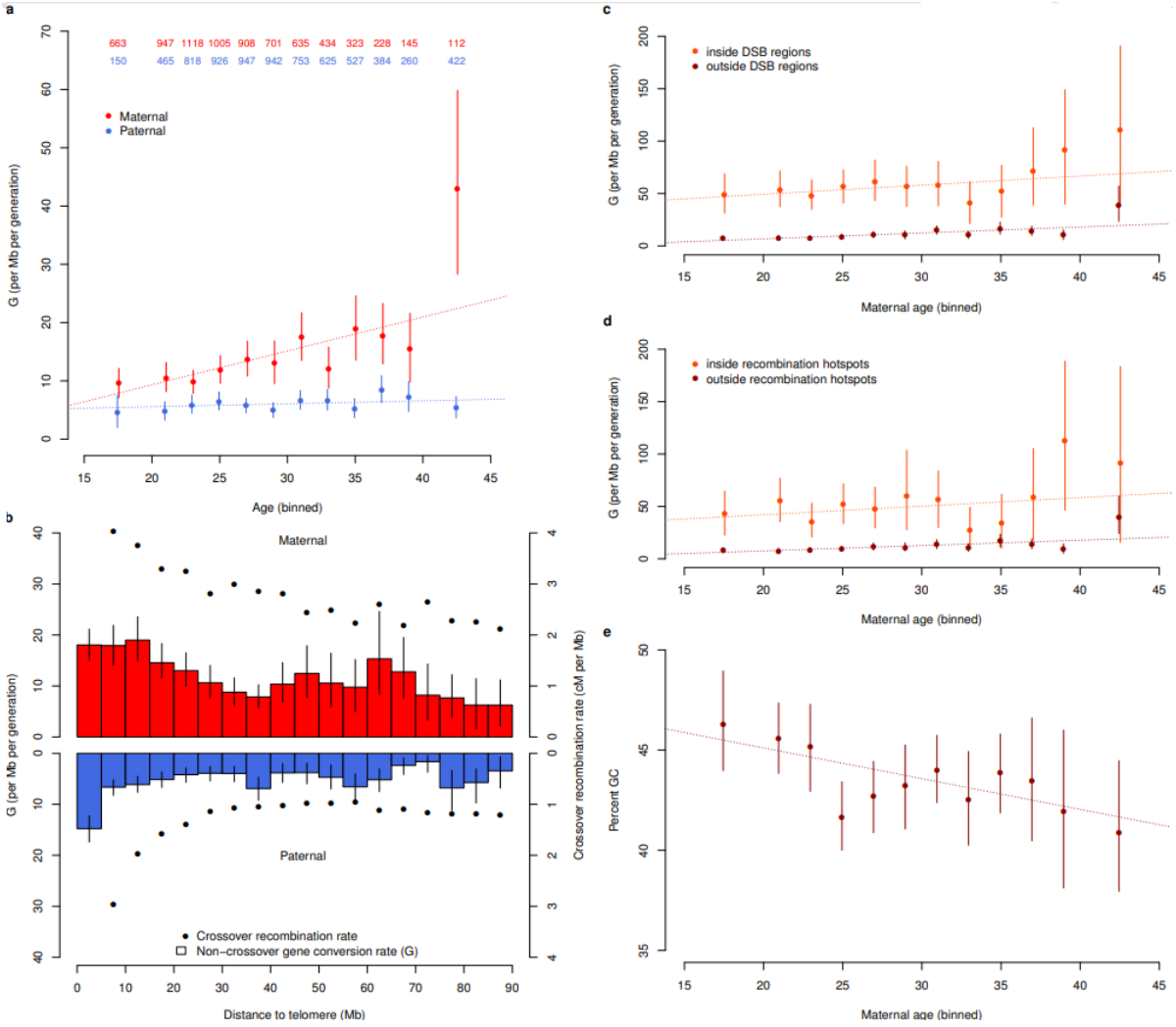


Figure 2 NCO gene conversion rate, G. As a function of **a)** age of parent, **b)** distance from telomere, **c)** age of mother stratified by presence in DSB region and **d)** age of mother stratified by presence in crossover recombination hotspots. **e)** percent GC content of neighboring 100 bases for gene converted markers outside of double strand break regions as a functionof age of mother. Blue represents paternal and red maternal transmission. Error bars represent 95% confidence intervals. **a)** Number of parents in each age bin are presented in top of figure. **a,c,d,e)** Individuals are grouped into 2 year age bins with all individuals younger than 20 grouped together and all individuals 40 years or older grouped together. **b)** mpps are grouped into 5Mb bins. Crossover recombination rate is computed as the average over all marker proband pairs in a bin of average crossover recombination rate in +/− 5kb interval around a marker. cM = centiMorgan, Mb = 1 million base pairs.

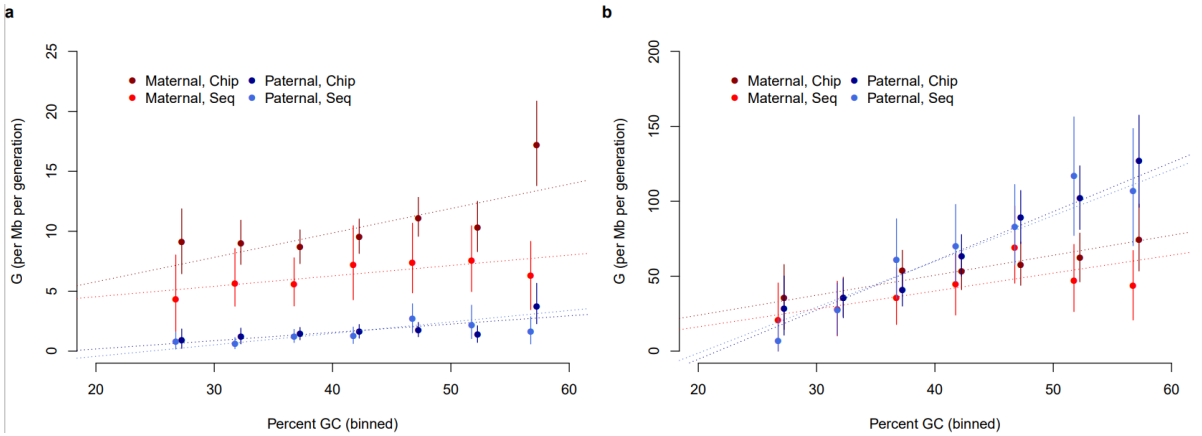


Figure 3 NCO gene conversion rate G, as a function of GC content of neighboring 100bp. a) G outside of male DSB regions. b) G inside of male DSB regions. Error bars represent 95% confidence intervals. The dotted line represents a linear model. NCO = non-crossover, DSB = double strand break

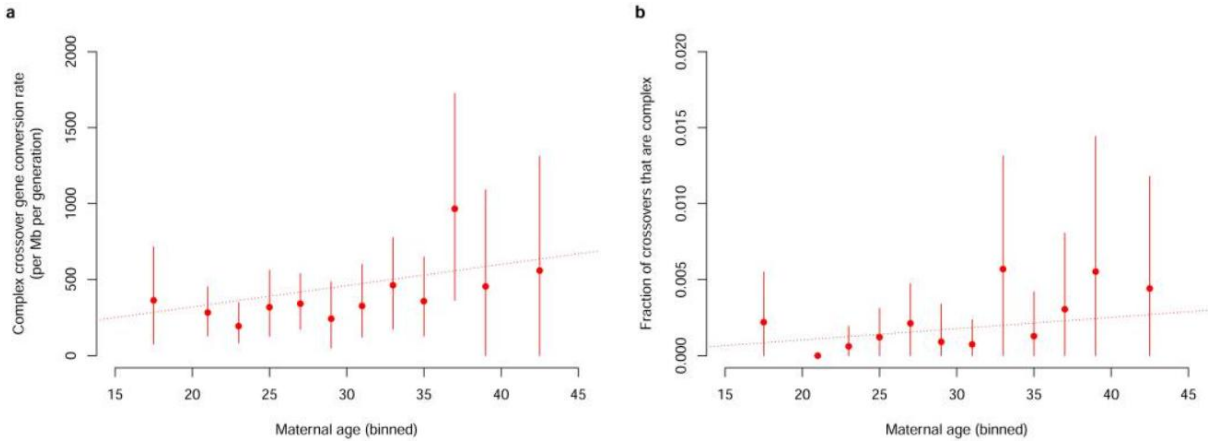


Figure 4 Complex crossovers (CCO) and maternal age. a) Complex crossover (CCO) gene conversion rate by maternal age in the chip dataset. b) Fraction of maternally transmitted crossovers in the chip dataset that are complex, as a function of mother's age. Error bars represent 95% confidence intervals. The dotted line represents a linear model.

Table 1 Estimates of NCO gene conversion of SNPs

	Chip dataset			**Sequencing dataset**		
	Paternal	Maternal	Total	Paternal	Maternal	Total
Converted mpps	641	1,393	2,034	300	565	865
N mpps considered	106,740,626	106,761,366	213,501,992	73,919,056	73,016,289	146,935,345
Observed gene conversion rate per Million bp (95% CI)	6.0 (5.5–6.5)	13.0 (12.0–14.1)	9.5 (9.0–10.1)	4.1 (3.4–4.8)	7.7 (5.7–9.9)	5.9 (4.9–7.0)
Corrected gene conversion rate per Million bp (95% CI)	3.9 (3.5–4.4)	10.0 (8.5–11.6)	7.0 (6.0–8.0)	–	–	–
N events	618	1,018	1,636	244	241	485
Short	606	849	1,455	234	196	430
Long	12	169	181	10	45	55
Gene converted mpps / event (SE)	1.04 (±0.01)	1.37 (± 0.03)	1.24 (±0.02)	1.23 (±0.06)	2.34(±0.28)	1.78(±0.15)
Short	1.01 (±0.00)	1.04 (±0.01)	1.03 (±0.01)	1.09 (± 0.02)	1.17 (±0.04)	1.13 (±0.02)
Long	–	3.01(±0.13)	2.96 (±0.12)	–	7.44 (±1.15)	6.93 (±0.99)
GC bias (95% CI)	64.4 (60.7–68.3)	69.1 (66.6–71.6)	67.6 (65.7–69.8)	63.1 (57.1–69.1)	72.5 (69.1–76.1)	69.3 (65.8–72.3)
Short	64.7 (61.1–68.5)	69.5 (66.4–72.3)	67.5 (65.3–69.7)	65.3 (58.3–72.0)	73.0 (66.5–79.7)	68.9 (64.1–74.1)
Long	–	68.6 (64.5–72.3)	67.9 (64.0–71.7)	–	72.2 (67.2–77.4)	69.8 (64.9–74.5)
OR DSB (95% CI)	42.3 (35.2–50.9)	5.4 (4.7–6.2)	10.7 (9.6–11.9)	45.7 (32.8–67.1)	7.1 (5.2–9.7)	14.9 (11.3–19.8)

Short	49.7 (41.2–61.1)	9.1 (7.9–10.5)	17.5 (15.9–19.4)	66.9 (49.4–92.1)	20.3 (15.7–25.8)	37.0 (30.2–45.7)
Long	–	1.1 (0.7–1.5)	1.1 (0.7–1.5)	–	1.6 (0.9–2.3)	2.0 (1.2–3.0)
OR crossover recombination hotspots (95% CI)	8.2 (6.6–10.1)	4.6 (3.8–5.4)	–	12.7 (8.6–18.3)	8.0 (5.7–11.1)	–
Short	8.8 (7.1–10.6)	6.5 (5.4–7.7)	–	14.8 (10.7–20.5)	14.1 (9.8–19.5)	–
Long	–	2.0 (1.1–3.2)	–	–	4.0 (1.5–6.9)	–

Corrected gene conversion rate is corrected for crossover recombination rate. This correction and the odds ratio for hotspots are computed using sex specific maps. Short events < 1,000 base-pairs, long events >= 1,000 base-pairs. Results for long NCO events are not presented in paternal transmissions due to small sample size. As crossover maps are sex specific no joint analysis is provided. NCO = non-crossover, OR = odds ratio, DSB = double strand break, mpp = marker proband pair, CI = confidence interval, SE = standard error

Table 2 Estimates of CCO gene conversions of SNPs

	Chip dataset			**Sequencing dataset**		
	Paternal	Maternal	Total	Paternal	Maternal	Total
Converted mpps	8	150	158	5	157	162
N mpps considered	289,364	450,307	739,671	170,612	262,323	432,935
Observed gene conversion rate per Million bp (95% CI)	28 (10–52)	333 (265–411)	214 (168–259)	29 (6–58)	598 (289–1,011)	374 (182–624)
N events	7	107	114	5	34	39
Gene converted mpps / event (SE)	1.14	1.40 (±0.08)	1.39 (±0.07)	1	4.76(±1.29)	4.26(±1.16)
GC bias (95% CI)	12.5	73.3 (65.2–80.7)	70.2 (62.5–77.8)	75	70.2 (63.1–78.6)	70.1 (63.1–78.8)
N crossovers considered	45,025	70,079	115,104	1,624	2,490	4,114
Complex crossovers, per thousand	0.16 (0.07–0.28)	1.5 (1.3–1.8)	1.0 (0.8–1.2)	3.1 (0.6–6.0)	13.3 (8.5–18.2)	9.2 (6.2–12.5)

Confidence intervals for CCO gene conversions in paternal transmissions are not reported due to small number of events. CCO = complex crossover, OR = odds ratio, DSB = double strand break, mpp = marker proband pair, CI = confidence interval

Table 3 Gene conversion of indels

	Sequencing dataset		
	Paternal	Maternal	Total
Converted mpps	23	38	61
NCO	22	28	50
CCO	1	10	11
N mpps considered	4,391,571	4,324,302	8,715,873
NCO	4,382,511	4,310,148	8,692,659
CCO	9,060	14,154	23,214
Observed gene conversion rate per Million bp (95% CI)	5.2 (3.0–8.0)	8.8 (5.5–12.1)	7.0 (5.0–9.0)
NCO	5.0 (2.7–7.8)	6.5 (3.9–9.5)	5.8 (4.1–7.9)
CCO	110	707	474
Insertion bias (95% CI)	55.5% (33.3–77.5)	71.1% (53.2–84.5)	65.6% (53.2–77.4)
NCO	54.5 (31.0–75.4)	67.8 (50.0–82.8)	62.0 (48.0–74.7)
CCO	–	80	81.8
OR NCO in DSB (95% CI)	35.6 (14.8–86.0)	4.9 (1.0–11.8)	13.9 (7.0–24.1)
OR NCO in crossover recombination hotspots (95% CI)	10.4 (2.0–29.0)	8.4 (0.0–25.4)	

OR = odds ratio, DSB = double strand break, mpp = marker proband pair, CI = confidence interval, CCO = complex crossover, NCO = non-crossover.

Online methods

In order to detect gene conversions, we use three-generation pedigrees, where a proband as well as both of its parents, at least two of the proband's siblings, a child of the proband and the proband's spouse are all genotyped (cf. Figure 1).

Data descriptions

We use two datasets of Icelandic samples collected as a part of disease association efforts at deCODE genetics[34]: Chip data consisting of 7,219 probands genotyped on Illumina HumanHap and Omni BeadChip arrays and sequencing data consisting of 101 whole genome sequenced probands.

The study was approved by the Icelandic Data Protection Authority (ref. 2004120649) and the National Bioethics Committee, Iceland (ref. VSN 13-028). All participating subjects who donated blood signed informed consent. Personal identities of the participants and biological samples were encrypted by a third party system approved and monitored by the Icelandic Data Protection Authority.

A description of how families were selected and datasets were proprocessed is given in Supplementary Note 11. A list of the chip types used in the chip dataset is given in Supplementary Table 10.

Identifying gene conversion mpps in pedigrees

Gene conversions appear in the proband's haplotype as short tracts of mpps from one of the parent's haplotype on the background of its other haplotype, i.e. the markers in question are inherited from one haplotype while nearby markers on both sides are inherited from the other haplotype. We can detect them if the two haplotypes of a parent and the haplotype transmitted from the parent to the proband are known.

To determine the haplotypes of the parents we phase their nuclear family (the parents, the proband and its siblings). This is done in three steps; In the first step we construct a set of mpps, where we are confident of the inheritance pattern from a given parent of the sibling group. We refer to these mpps as anchors. In the second step, we phase the remaining mpps by minimizing the discrepancy between the genotypes of the mpps compared to the inheritance pattern observed at neighboring anchors. In the third step, we determine the location of crossover recombinations in order to phase the proband.

Finally, we identify gene conversions as tracts of inconsistencies between genotypes of the proband and the phased inheritance pattern at each mpp. If the tracts of inconsistencies indicate a gene conversion we attempt to verify them in the proband's children.

A schematic overview of the algorithm is given in Supplementary figure 7.

Determining anchor mpps

An anchor is an mpp where one parent is heterozygous, the other is homozygous and the genotypes of the proband, its parents and all of its siblings meet the accuracy thresholds defined above. With respect to the heterozygous parent, the phase of the sibling group (the proband and its siblings) is unambiguous at anchors and the sibling group can be partitioned into two sets, determined by which allele was inherited from the heterozygous parent. Two adjacent anchors, with the same heterozygous parent, induce the same partition in the sibling group, unless either anchor is genotyped incorrectly or a gene conversion or crossover recombination occurred between them.

Given an mpp and a parent we define a left anchor as the closest anchor with a lower numerical coordinate where the parent is heterozygous. A right anchor is defined analogously with a higher numerical coordinate.

Unless a genotyping error occurred we can be confident in the inheritance pattern for all anchor mpps. We now remove mpps that appear to be the result of a genotyping error from the set of anchor markers. When removing these we may also remove mpps that are the result of a gene conversion. In a later step we determine whether the mpp is the result of a gene conversion or a genotyping error.

Mpps whose partition does not agree with neighboring anchors are removed. To formally delineate which markers are removed from the set of anchors, we define the discrepancy between two anchors as the minimum number of individuals that need to be moved between the sibling group partitions of the anchors such that the partitions become identical. We compute a local discrepancy score for an anchor, A, as the sum of the discrepancy between A and its two closest anchors to the left and two closest anchors to the right. The anchor A is removed if, when doing so, the sum of the discrepancy scores of all other anchors is reduced.

Phasing the parents

At a given mpp the sibling group can be split into four inheritance groups based on which of the two haplotypes they inherit from each parent. The inheritance groups are not known but when there are no crossover recombinations, gene conversions or genotyping errors, the haplotypes will agree with the haplotypes of both parents inherited at the neighbouring anchors. We define two inheritance groupings, left and right. The left inheritance grouping is determined by the left anchors of both parents and the right inheritance grouping is based on the right anchors of both parents. Both the left and right inheritance groupings should be identical to the inheritance grouping at the given mpp unless there has been a crossover recombination for either parent in the region or the mpp being examined is gene converted in the proband or one of its siblings. A genotyping error in one of the siblings or either parent may also occur, causing the genotypes not to agree with the left and right inheritance groupings even if they are identical to the true inheritance grouping at the mpp.

Given the genotypes of the individuals in the nuclear family and the inheritance groupings we assign alleles to the parents' haplotypes. For binary mpps, there are a total of $2^4 = 16$ possible assignments of the two alleles to the four haplotypes. For each such assignment we infer genotypes of both parents and the siblings according to the left inheritance grouping and compare them to observed genotypes. We define left phasing discrepancies in the nuclear family as the combined number of mismatches between observed and inferred genotypes. Right phasing discrepancies are defined analogously from the right inheritance groups. For each assignment of alleles to haplotypes we define the number of phasing discrepancies as the smaller of the left and right phasing discrepancies. A phasing discrepancy can be explained with a crossover recombination, a gene conversion or genotyping error.

If there exists exactly one assignment of the alleles to the parents' haplotypes with fewer than two phasing discrepancies, the mpp is considered phased by the assignment. All other Mpps are removed from further consideration as candidates when searching for gene conversions. When the assignment has no phasing discrepancy either there is no gene conversion or the mpp is part of a gene conversion tract that includes neighboring anchors. Assignments with a single phasing discrepancy are further candidates for where a gene conversion may have taken place. Not all single phasing discrepancies will represent a gene conversion as they may also represent a genotyping error or a non-gene converted mpp in a long gene conversion tract including both neighboring anchors.

When there are more than one assignments that have fewer than two phasing discrepancies we cannot reliably determine which assignment is the correct one, since all of them can arise from a single genotyping error or gene converted mpp. An example of when an mpp has multiple assignments with fewer than two phasing discrepancies is when one of the parents' haplotypes is not carried by any members of the sibling group. If all individuals are correctly genotyped and there is no gene conversion then the correct assignment of haplotypes leads to zero phasing discrepancies, while switching the assignment of the haplotype not carried by any members of the sibling group leads to one phasing discrepancy.

When all assignments of alleles to the parents' haplotypes yield at least two phasing discrepancies, then the mpps genotypes are not consistent with the left and right inheritance groupings, even when allowing for a single individual to be either carrier of a gene conversion or having a genotyping error. This can occur due to multiple recombinations, a structural variant at the locus, a misplacement of the marker in the assembly or repeated genotyping errors at the marker.

Determining crossover recombinations

For each proband we locate crossover recombinations from each parent separately. We refer to an mpp as informative if the phase of the proband can be determined directly at the mpp without any assumption about the phase inherited from the other parent. Thus, the set of informative mpps includes anchor mpps as well as mpps where the parent of interest is heterozygous and either the proband or the other parent are homozygous. In particular, the set includes mpps where some of the siblings of the proband do not meet genotyping accuracy thresholds.

In order to distinguish candidate gene conversions from crossover recombinations, we look for all inheritance tract changes in the proband. Initially, we assign inheritance tract changes to all regions between two adjacent informative mpps where the proband inherits alleles from different haplotypes of the parent of interest. The region of an inheritance tract change can be further narrowed if the parent of interest is heterozygous for additional mpps between the two informative markers, if the haplotype of the other parent is known, and if the two informative mpps agree on a haplotype for the other parent. In this case, we assume that the proband inherited the same haplotype for the whole region from the other parent and we can determine the proband‘s phase at the all mpps in the region where the parent of interest is heterozygous. If the two informative mpps do not agree on a haplotype for the other parent,

the inheritance tract change region is excluded from the search for gene conversions. Once the inheritance tract change region has been narrowed, we assign an inheritance tract change to the center of the region.

A crossover recombination is assigned to all tract changes where no other tract change occurs within 100 kb. All other tract changes are candidate gene conversions. Additionally, a crossover recombination is assigned if there is an odd number of multiple consecutive tract changes within 100 kb of each other; more precisely, the crossover recombination is assigned to the leftmost or rightmost tract change depending on which induces fewer gene converted mpps (see below). In this case a crossover recombination has occurred along with possible gene conversions.

Determining gene conversion mpps

Having assigned crossover recombinations and phased the haplotypes of the parents, we search in the proband for mismatches in observed genotypes compared to the genotypes defined by the haplotypes inherited from the parents.

We search for gene conversion from each parent separately for all mpps in the phased mpp set after applying the quality filters described in Supplementary Note 12.

Mpps passing these filters are counted towards the denominator in our rate computations. If there is a mismatch between the proband‘s genotype and its phase-determined haplotypes, we examine whether this mismatch can be due to a gene conversion.

If the mismatch can be explained with the other haplotype of the parent of interest, we attempt to verify the gene conversion in the proband’s children. We verify the genotype in one of two ways; If a child is homozygous, we verify the gene converted haplotype without requiring the proband’s spouse to be genotyped. Otherwise, if the proband’s spouse is genotyped and homozygous we use that together with the child’s genotype to verify the gene converted haplotype. We may be unable to verify the haplotype due to inconsistencies such as structural variants, a misplacement in the assembly or repeated genotyping errors.

In order to verify a putative gene conversion in the proband’s child, we first determine whether the child carries the gene converted haplotype. We search for the closest mpps to the left (of lower numerical order) and to the right (of higher numerical order) where the proband is heterozygous and the spouse is homozygous, ignoring mpps at the end of chromosomes where there is no such left or right mpp. At these mpps we can determine

which haplotype the child inherited from the proband. If the child inherited the same haplotype from the proband at both of these mpps and the distance between these mpps is less than 1 Mb we assume that the child is carrying the corresponding haplotype. The gene conversion can be verified in the child if this is the gene conversion haplotype.

If a marker shows evidence of being part of a structural variant, being misplaced in the assembly or having more than one genotyping error, the marker is flagged as problematic in the family. Specifically, a marker is flagged as problematic either if all of the assignments of alleles to the parents' phases produces two or more phasing discrepancies or if we fail to verify a gene conversion in one of the proband's children. Markers that are flagged as problematic in more than one family are removed from the set of quality markers input to the algorithm.

Crossover regions and CCO vs. NCO gene conversions

For each crossover recombination we defined a crossover region as the adjacent 100 kb in each direction. If a gene converted mpp was found within the crossover region the region was iteratively extended so that it contained all mpps within 100 kb of the crossover recombination and the gene converted mpps found.

Mpps determined to be within and outside a crossover region were used in the computation of CCO gene conversion rate and NCO gene conversion rate, respectively.

Determining gene conversion events

Gene conversion events are determined once all gene converted mpps have been determined. While searching for gene converted mpps we restrict our search to contiguous tracts of mpps where the length of the tract is 100kb. Gene conversion events may however contain both gene converted mpps and non-gene converted mpps.

Within non-crossovers, we arranged mpp positions from a parent-proband pair into distinct gene conversion events by traversing the chromosome in numerical order. We considered the first gene converted mpp found on a chromosome to be part of a new event and iteratively extended the event if a gene converted mpp was found within 100 kb of the previous gene converted mpp. Consequently, gene conversion events may be longer than 100kb.

Within crossovers, mpps neighbouring the same crossover recombination were considered a part of the same event.

Crossover recombination map data

We use a dataset of crossover recombinations[24] preprocessed as described in Supplementary Note 13.

Computation of rates, confidence intervals and p-values

We compute the observed gene conversion rate as the number of mpps where a gene conversion occurs divided by the number of mpps that were tested for a gene conversion.

For two events, A and B, we compute the odds ratio as (N11*N22)/(N12*N21), where N11 represents the number of mpps that are part of event A and B, N12 the number of mpps that are part of A and not B, N21 the number of mpps that are part of B and not A and N22 the number of mpps that are part of neither event. Odds ratios for crossover recombination hotspots are computed considering only those markers where the crossover recombination rate has been estimated.

All confidence intervals presented are 95% confidence intervals and all p-values are two-sided. Confidence intervals for G, odds ratios, number of mpps per event, rate of increase in G between age of 20 and 40 and GC bias, are computed using a bootstrap method[49]. For the chip dataset 1,000 sets of 7,219 individuals are sampled with replacement from the set of 7,219 probands. For the sequencing dataset 1,000 sets of 101 individuals are sampled with replacement from the set of 101 probands. The statistic in question is computed within each set, creating a list of 1,000 statistics. Following the sorting of this list, the lower bound of the confidence interval is computed as the mean of entries 25 and 26 and the upper bound is computed as the mean of entries 975 and 976.

Age of parent effects are determined using a weighted linear regression using the function lm in R[50]. To determine age effect of G we first compute G for each proband-parent pair separately. The final model can be expressed as: lm(G ~ ParentAge, weights = sqrt(N)), where N is the number of mpps considered for the proband. To determine age effect on other statistics we compute for each proband-parent pair separately the statistic, S, in question. The final model can be expressed as: lm(S ~ ParentAge, weights = sqrt(N)).

Linear regression, its confidence intervals and p-values, for distance to the telomere and length of chromosome, were computed using the lm function in R, using a matrix containing all mpps where a gene conversion event could be ascertained. All other linear regressions were implemented using Python[51].

All other p-values, not previously discussed, were computed using bootstrapping. 1,000 simulations are used analogously to the description above and a p-value is computed by counting the number of times the single-sided event of interest occurred and dividing by the number of simulations. The single-sided p-value was then multiplied by 2, in order to obtain a double-sided p-value. In cases when the event of interest did not occur in 1,000 simulations the p-value was reported as < 0.001.

In order to compute G corrected for crossover recombination, a linear regression was performed with gene conversion as a response and local sex specific crossover recombination rate as an explanatory variable. All marker proband pairs where an NCO gene conversion could be ascertained and the crossover recombination rate had been determined were used. A corrected G was computed by inserting the genomic average crossover recombination rate[24], of 1.572 cM/Mb for maternal transmissions and 0.772 cM/Mb for paternal transmissions, into the regression formula. Confidence intervals were computed using the predict.lm function in R.

Methods-only references

49. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (CRC Press, 1994).

50. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5,** 299–314 (1996).

51. van Rossum, G. & Drake, F. L. *PYTHON Reference Manual*. (Centrum voor Wiskunde en Informatica, 1995).

Competing financial interests

All authors are employees of deCODE genetics/Amgen.

Paper II

Title:

NCOurd: modelling length distributions of NCO events and gene conversion tracts

Authors:

Marteinn T Hardarson [1,2,✉], Gunnar Palsson [1], Bjarni V Halldorsson [1,2,✉]

Affiliations:

1 deCODE genetics, Reykjavik 102, Iceland

2 School of Technology, Reykjavik University, Reykjavik 102, Iceland

*Correspondence to: Bjarni V. Halldorsson (Bjarni.Halldorsson@decode.is)

Abstract

Motivation

Meiotic recombination is the main driving force of human genetic diversity, along with mutations. Recombinations split into crossovers, separating large chromosomal regions originating from different homologous chromosomes, and non-crossovers (NCOs), where a small segment from one chromosome is embedded in a region originating from the homologous chromosome. NCOs are much less studied than mutations and crossovers as NCOs are short and can only be detected at markers heterozygous in the transmitting parent, leaving most of them undetectable.

Results

The detectable NCOs, known as gene conversions, hide information about NCOs, including their number and length, waiting to be unveiled. We introduce NCOurd, software, and algorithm, based on an expectation–maximization algorithm, to estimate the number of NCOs and their length distribution from gene conversion data.

Availability and implementation

https://github.com/DecodeGenetics/NCOurd

1 Introduction

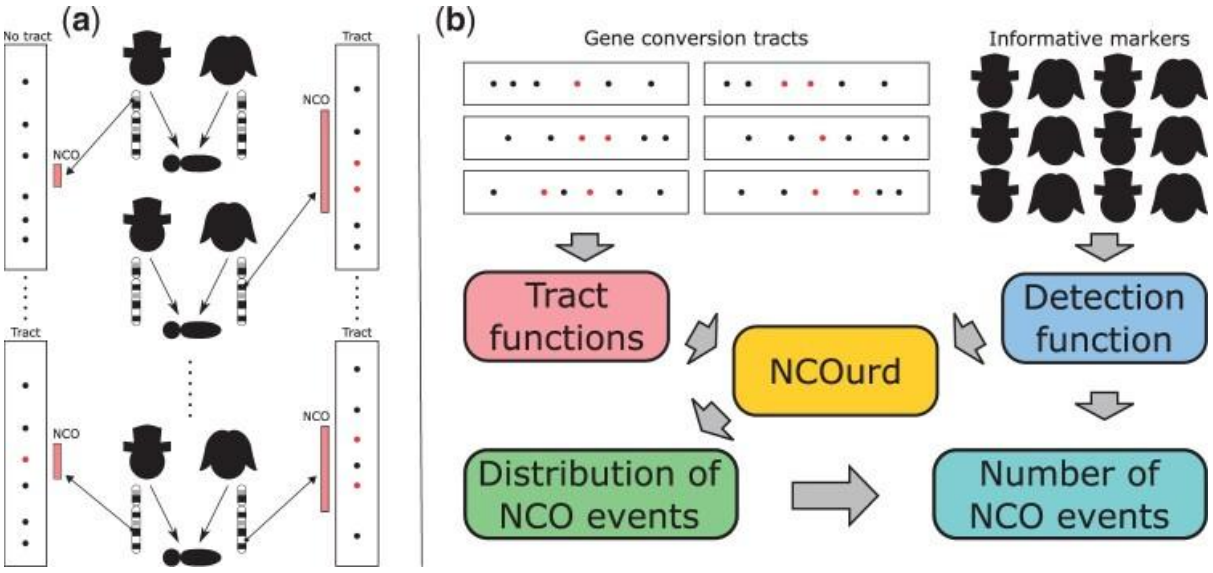
The name NCOurd (pronounced encored) is a combination of NCO and Urd. In Nordic mythology, Urd is one of the three Norns (witches) who decide the fate of men and gods. Of the three witches, Urd represents the past, and we would love to understand NCOs happening in the past, during meiosis, but we can only measure gene conversion in the present, from the DNA of living individuals. With mathematical (mathe-magical) witchery, we are able to estimate the past from the present.

In sexually reproducing organisms, the gametes (reproductive cells) are created in a special cell division called meiosis. Meioses are initiated by diploid cells, but the product, the gametes, are haploid and each of their chromosomes is a combination of the two homologous chromosomes of the parent cell created by recombinations. Recombinations allow offspring to inherit a mosaic of the homologous parental chromosomes rather than a copy of either one of them.

Determining from which homologous parental chromosome a segment originates is only possible at heterozygous markers; sites where the two chromosomes differ. Meiotic recombinations are initiated by a double strand break (DSB) in one parental chromosome and are repaired using the homologue as a template. This can result in either a crossover, which separates large chromosomal segments from different homologues, or non-crossover (NCO). DSBs that are repaired with NCOs can result in gene conversions where alleles from one haplotype are embedded on the background of the homologous haplotype. The gene-converted markers from a single NCO form a gene conversion tract. (See Supplementary Fig. S1 and Lam and Keeney (2014) for more information on the meiotic recombination process.)

Not all heterozygous markers overlapping an NCO event will become gene conversions (Supplementary Fig. S1) so NCO events can extend beyond the heterozygous markers flanking the gene conversion tract and the gene conversion tract does not need to consist of consecutive heterozygous markers. Some NCO events will fail to result in a gene conversion tract either because the NCO event did not overlap any heterozygous markers or because none was gene converted (Fig. 1a).

Figure 1.



(a) NCO events (red rectangles) occur during meiosis. Some NCO events overlap heterozygous markers (points) and create gene conversion (red points), alleles from one haplotype embedded in the background of the homologous haplotype (black points). If at least one marker gets gene converted a gene conversion tract is created. (Note that gene conversion tracts do not need to consist of consecutive heterozygous markers.) The penetrance is probability of a heterozygous marker overlapping the NCO event becoming a gene conversion. (b) For each observed gene conversion tract, a tract function is computed, representing the probability that the gene conversion tract was produced by a NCO event of length x. A detection function, representing the probability that a gene conversion can be detected from a NCO event of length x, is calculated based on the all heterozygous markers in all the parents. NCOurd uses these functions to estimate the length distribution and the number of NCO events.

Consequently, short NCO events are very unlikely to produce any gene conversion, but as events get larger, the probability increases, upwardly skewing the length distribution of NCO events leading to gene conversions compared with the underlying length distribution of all NCO events. Furthermore, looking at individual gene conversions gives limited insights into the length of the original NCO event as flanking heterozygous markers can be far away or may have overlapped the NCO event but not been gene converted. These two factors make it impossible to estimate the length distribution of NCO events without some statistical modelling.

The scope of this article is limited to directly observable allelic gene conversion, which have been studied in various species. For example, Williams *et al.* (2015) and Halldorsson *et al.* (2016) used three-generation families to study gene conversions in humans, Miller *et al.* (2016) studied gene conversions in fruit flies, Li *et al.* (2019) in mice, and Wall *et al.* (2022) in baboons.

Gene conversions have also been studied with population-based methods (Betrán *et al.* 1997, Setter *et al.* 2022). Non-allelic gene conversion, where a different locus is used as a

template for DNA repair, has also been studied (Mansai and Innan 2010). Both settings require statistical models different from the one presented here.

Some studies attempt to estimate length distributions and the number of NCO events (Mansai *et al.* 2011, Miller *et al.* 2012, Li *et al.* 2019). Until recently, the models used have been limited to a single exponential distribution or their discrete analogue, geometric distribution. The authors of Wall *et al.* (2022) found that a single geometric distribution did not fit their data, consisting of gene conversions in baboons, well and proposed a mixture of two geometric distributions. Modelling gene conversions with a mixture distribution is reasonable, as DSBs can be resolved with an NCO via multiple pathways. However, a mixture of geometric distributions has the drawback of forcing NCO events to become increasingly more likely as they get smaller, neglecting the possibility that the strand invasion process might enforce a minimum NCO event size. Further, the models used thus far are oblivious to the fact that some heterozygous markers overlapping NCOs fail to produce gene conversions allowing NCO events the possibility to extend beyond flanking heterozygous markers of the gene conversion tract.

To improve upon this and to utilize ever larger gene conversion datasets, we introduce NCOurd, a software package to infer the length distributions of NCO events (Fig. 1b). We demonstrate the robustness of our method using simulations and apply it to published datasets (Halldorsson *et al.* 2016, Li *et al.* 2019).

Given a set of gene conversions and the length distribution of NCOs, we can estimate the number of NCOs occurring per meiosis and better understand their underlying biology and how they shape human diversity and the human evolutionary process in general. Further, reliable length estimates allow us to estimate which regions of the human genome are most affected by NCOs, whether they affect disease, fertility, or other human phenotypes, and how NCOs are affected by human genetics and environmental conditions such as parental age and sex.

2 Methods

Our goal is to estimate the length distribution of NCO events from a set of gene conversion tracts together with the set of heterozygous markers in the transmitting parent where gene conversions would be detectable. We assume gene conversions have already been identified

and grouped into gene conversion tracts. Examples of how gene conversions can be identified can be found in Halldorsson *et al.* (2016) and Li *et al.* (2019).

We first define terminology to model NCO events and their resulting gene conversion tracts. Then, we show how a likelihood function can be written in terms of the length distribution of NCO events. This allows for a maximum-likelihood estimate of the observed gene conversion tract over the parameter space of the length distribution of NCO events, assuming that the length distribution of NCOs follows a mixture of negative binomial distributions. We then use an expectation–maximization (EM) algorithm (Dempster *et al.* 1977) to solve the maximum-likelihood problem.

2.1 Definitions and model

NCO events can lead to 'gene conversions'; where one or more alleles from one chromosome are embedded in a haplotype originating from the homologous chromosome, the 'background haplotype'. 'Informative markers' are all sites where a gene conversion can be detected and, thus, are a subset of the sites where a gene conversion can occur, i.e. the heterozygous markers in the parent (some heterozygous markers may be filtered for quality control). A 'gene conversion tract' is the set of all the gene-converted markers detected from a single NCO event. Note that a gene conversion tract need not be consecutive informative markers as some informative markers within an NCO event can be from the background haplotype. To account for the fact that not all heterozygous markers result in a gene conversion, we define the 'penetrance' p as the proportion of informative markers that are gene converted within NCO events. (Informative markers overlapping the NCO event have the probability $1-p$ of being reverted back to the background state).

Let (L, M, C, O) be a random NCO event where L is the length of the event, M is the starting point (lowest chromosomal coordinate), C is the chromosome, and $O = (R, S)$ is the observed gene conversion tract, where R denotes a parent–offspring pair in which the event occurred and S is the set of gene-converted markers produced by the event (S may be empty).
Both L and M are discrete random variables taking integer values, C and R are random categorical values, and S is a random set of integers. We also define I_R as the informative markers of the parent in R.

The set S consists of zero or more informative markers on chromosome, C, of the parent in R overlapping the NCO tract, i.e. $S \subseteq [M, M\ S \subseteq [M, M+L-1] \cap I_R$. The random NCO event produces a gene conversion tract if and only if $S \neq \emptyset$.

Let t be a gene conversion tract. Let $o_t = (r_t, s_t)$ be the pair consisting of r_t, the transmitting parent–recipient offspring pair, and s_t, the set of gene-converted markers in the tract t.

As an input for our EM algorithm, we need the following:

- *Detection function*, $D(x)$, for an integer x is the probability of a random NCO event of length x producing a gene conversion tract, i.e. $D(x) = \Pr(S \neq \emptyset | L = x)$.

The detection function measures the probability of an NCO event of length x is detected (as a gene conversion) and depends on the distribution of informative markers across the genome and the penetrance.

- *Tract functions*, $T_t(x)$, for an integer x and a gene conversion tract t are the probability of a random NCO event of length x producing the gene conversion tract o_t, i.e. $T_t(x) = \Pr(O = o_t | L = x)$.

The tract functions measure the probability of an NCO event of length x producing a tract t. The probabilities depend on the placement of informative markers around the gene conversion tract t and the penetrance.

We note that these functions can be calculated without knowledge of the length distribution of NCO events.

We need the penetrance p for these calculations. All informative markers within all NCO events are assumed to have a probability p of becoming gene conversions independently of each other. The probability of an NCO event resulting in a specific gene conversion tract contained within the NCO event is, therefore, the product of p and $(1-p)$ for each informative marker within the NCO event depending on whether it was a gene conversion or not and if the NCO event has n informative markers the probability of any gene conversion being created is $1-(1-p)^n$ (i.e. all cases except when all informative markers inside the NCO event are reverted to the original state).

The value of the detection function at x is the average probability that an NCO event of length x creates any gene conversion tract over all possible placements in the genome.

For each gene conversion tract t, the value of $T_t(x)$ is the sum of the probabilities of an NCO event of length x resulting in t over all possible placements containing t, weighted with the probabilities of the placement.

The penetrance can be estimated by computing the fraction of gene-converted markers in all gene conversion tracts, excluding boundary markers (the first and last marker of each tract). It

is possible that the penetrance depends on the NCO event length (i.e. short tracts could be either more or less likely to produce gene-converted markers). The calculations of detection and tract functions can easily be augmented to accommodate penetrance as a function of NCO event length. However, estimating penetrance as a function of NCO event length is difficult and would likely suffer from lack of power.

A more detailed description of how the detection and the tract functions are calculated and the approximations used together with information on how the penetrance can be estimated can be found in the Supplementary Information.

2.2 Likelihood function

To calculate a likelihood function, we must be able to evaluate the probability of our outcomes. Using our model, we assume random NCO events are generated, but we can only observe them if they create a gene conversion tract ($S \neq \emptyset$). In other words, for gene conversion tract *t*, we need to estimate

$$\Pr(O = o_t) \,|S \neq \emptyset).$$

Using the law of total probability, we get:

$$\Pr(O = o_t|S \neq \emptyset) = \sum_{x=1}^{\infty} \Pr\,(O = o_t|S \neq \emptyset, L = x) \cdot \Pr\,(L = x|S \neq \emptyset). \qquad (1)$$

The two factors inside the sum can be rewritten using our definitions of detection and tract functions. The former, together with the definition of conditional probability, gives: since *t* is a gene conversion tract, $O = o_t$ implies that $S \neq \emptyset$. The latter combined with the Bayes' theorem and the law of total probability giving:

$$\begin{aligned} \Pr(O = o_t|S \neq \emptyset, L = x) &= \frac{\Pr\,(O = o_t, S \neq \emptyset|L = x)}{\Pr\,(S \neq \emptyset|L = x)} \\ &= \frac{\Pr\,(O = o_t|L = x)}{\Pr\,(S \neq \emptyset|L = x)} = \frac{T_t(x)}{D(x)}, \end{aligned} \qquad (2)$$

$$\Pr(L = x|S \neq \emptyset) = \frac{\Pr\,(S \neq \emptyset|L = x) \cdot \Pr\,(L = x)}{\Pr\,(S \neq \emptyset)} \qquad (3)$$

$$= \frac{\Pr(S \neq \emptyset | L = x) \cdot \Pr(L = x)}{\sum_{y=1}^{\infty} \Pr(S \neq \emptyset | L = y) \Pr(L = y)}$$

$$= \frac{D(x) \cdot \Pr(L = x)}{\sum_{y=1}^{\infty} D(y) \Pr(L = y)}.$$

Putting the three equations together (1)–(3), we get:

$$\Pr(O = o_t | S \neq \emptyset)$$

$$= \sum_{x=1}^{\infty} \Pr(O = o_t | S \neq \emptyset, L = x) \cdot \Pr(L = x | S \neq \emptyset)$$

$$= \sum_{x=1}^{\infty} \frac{T_t(x)}{D(x)} \cdot \frac{D(x) \cdot \Pr(L = x)}{\sum_{y=1}^{\infty} D(y) \cdot \Pr(L = y)} \tag{4}$$

$$= \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \Pr(L = x)}{\sum_{x=1}^{\infty} D(x) \cdot \Pr(L = x)}.$$

We are now ready to define our likelihood function. We assume that we have a set of gene conversion tracts, $\mathcal{T}$, and L is a discrete parametric distribution with parameters θ, denoted L_θ.

Then we want find θ which maximizes the likelihood function the second equation is due to Equation (4).

$$\mathcal{L}(\theta | \mathcal{T}) = \prod_{t \in \mathcal{T}} \Pr(O = o_t | S \neq \emptyset, L = x) = \prod_{t \in \mathcal{T}} \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \Pr(L = x)}{\sum_{x=1}^{\infty} D(x) \cdot \Pr(L = x)}. \tag{5}$$

Note that even though the gene conversion tracts are drawn from a subset of all NCO events, namely requiring $S \neq \emptyset$, we have written the likelihood functions in terms of length distribution of all NCO events with $\Pr(L_\theta = x)$together with the detection function and the tract functions assumed to have already been calculated.

2.3 EM algorithm

In the Supplementary Information, we show that if the length distribution of NCO events is a mixture distribution with *n* components, then the length distribution of gene conversion producing NCO events is a mixture distribution with *n* components, and the probability mass function (PMF) of each component in the second mixture distribution can be derived from the PMF of the corresponding component in the first mixture distribution. Also, if the PMFs and

mixture weights for all the components are known for one of the mixture distributions, we can calculate all the weights for the other mixture distribution.

We have assumed that the length distribution of NCO events is a mixture of *n* components and let α_i and $\hat{\alpha}_i$ be the mixture weights for component *i* in the length distributions of NCO events and gene conversion producing NCO events, respectively. Similarly, let f_i and $\hat{f}_i$ be the PMFs for component *i* in the length distributions of NCO events and gene conversion producing NCO events, respectively. We add to our model a latent random integer, *Z*, representing membership in the mixture components. Now we have:

$$\alpha_i = \Pr(Z = i)$$

$$f_i(x) = \Pr(L = x | Z = i)$$

$$f(x) = \Pr(L = x) = \sum_{i=1}^{n} \alpha_i f_i(x)$$

$$\hat{\alpha}_i = \Pr(Z = i | S \neq \emptyset)$$

$$\hat{f}_i(x) = \Pr(L = x | Z = i, S \neq \emptyset)$$

$$\hat{f}(x) = \Pr(L = x) = \sum_{i=1}^{n} \hat{\alpha}_i \hat{f}_i(x).$$

We assume that the mixture components are negative binomial distributions having parameters θ_i for the *i*-th component and write f_{θ_i} instead of f_i. The NCO events associated with our gene conversion tracts are from the length distribution having the PMF: and we want to determine the parameters $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_n)$ and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$ (the parameters of the underlying negative binomial distributions of the NCO length distribution mixture).

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i \hat{f}_{\theta_i}(x)$$

Using Equation (5), our likelihood function for a single tract, *t*, now becomes: and the complete likelihood function is then:

$$\mathcal{L}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta} | \mathrm{t}) = \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \Pr\left(L_{\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta}} = x\right)}{\sum_{x=1}^{\infty} D(x) \cdot \Pr\left(L_{\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta}} = x\right)}$$

$$= \sum_{i=1}^{n} \Pr(Z = i | S \neq \emptyset) \cdot \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \Pr\left(L_{\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta}} = x | Z = i\right)}{\sum_{x=1}^{\infty} D(x) \cdot \Pr\left(L_{\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta}} = x | Z = i\right)}$$

$$= \sum_{i=1}^{n} \hat{\alpha}_i \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \Pr\left(L_{\theta_i} = x\right)}{\sum_{x=1}^{\infty} D(x) \cdot \Pr\left(L_{\theta_i} = x\right)}$$

$$= \sum_{i=1}^{n} \hat{\alpha}_i \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \mathrm{f}_{\theta_i}(x)}{\sum_{x=1}^{\infty} D(x) \cdot \mathrm{f}_{\theta_i}(x)}.$$

$$\mathcal{L}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta} | \mathcal{T}) = \prod_{t \in \mathcal{T}} \sum_{i=1}^{n} \hat{\alpha}_i \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \mathrm{f}_{\theta_i}(x)}{\sum_{x=1}^{\infty} D(x) \cdot \mathrm{f}_{\theta_i}(x)}. \tag{6}$$

We can now use an EM algorithm to estimate the parameters $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\theta}$ that maximize the likelihood function. Assume that at step k in the EM algorithm, the estimated values of $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\theta}$ are $\hat{\boldsymbol{\alpha}}(k) = \left(\hat{\alpha}_1(k), \cdots, \hat{\alpha}_n(k)\right)$ and $\boldsymbol{\theta}(k) = \left(\theta_1(k), \cdots, \theta_n(k)\right)$, respectively. For a gene conversion tract t, we get the ith membership weight (E-step):

$$w_i^t = \frac{\hat{\alpha}_i \Pr\left(O = o_t | \theta_i, S \neq \emptyset\right)}{\sum_{j=1}^{n} \hat{\alpha}_j \Pr\left(O = o_t | \theta_j, S \neq \emptyset\right)} = \frac{\hat{\alpha}_i \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \mathrm{f}_{\theta_i}(x)}{\sum_{x=1}^{\infty} D(x) \cdot \mathrm{f}_{\theta_i}(x)}}{\sum_{j=1}^{n} \hat{\alpha}_j \frac{\sum_{x=1}^{\infty} T_t(x) \cdot \mathrm{f}_{\theta_j}(x)}{\sum_{x=1}^{\infty} D(x) \cdot \mathrm{f}_{\theta_j}(x)}} \tag{7}$$

That is, the relative probabilities that each component would generate the given gene conversion tract.

Then we update the parameters (M-step): and

$$\hat{\alpha}_i^{(k+1)} = \sum_{t \in \mathcal{T}} \frac{w_i^t}{|\mathcal{T}|},$$

$$\begin{aligned} \theta_i^{(k+1)} &= \arg\max_{\theta} \sum_{t \in \mathcal{T}} w_i^t \log(\Pr(O = o_t | \theta, S \neq \emptyset)) \\ &= \arg\max_{\theta} \sum_{t \in \mathcal{T}} w_i^t \log\left(\frac{\sum_{x=1}^{\infty} T_t(x) \cdot \mathrm{f}_{\theta_i}(x)}{\sum_{x=1}^{\infty} D(x) \cdot \mathrm{f}_{\theta_i}(x)}\right) \end{aligned} \tag{8}$$

The log-likelihood will improve by at least the difference of the values of the target function at $\theta_i^{(k+1)}$ and $\theta_i^{(k)}$ for each component i (Dempster *et al.* 1977).

This guarantees a monotonic increase in the total log-likelihood. We repeat this process until the parameters have converged (by default the Euclidean norm of the relative changes in all parameters is less 10^{-7}). We use `scipy.optimize` to find a value of θ that maximizes the sum.

Additional information on how we approximate the infinite sums is provided in the Supplementary Information.

2.4 Simulated data

We used the marker set from Jónsson *et al.* (2017) to construct a human-like set of heterozygous markers to use as an informative marker set. We used allele frequencies to determine the probability of each marker to be included in the heterozygous marker set. The mean and median distance between consecutive markers in the set was 1465 and 754, respectively. We tested our method in two experiments, E1 and E2, using simulated gene conversion tract datasets from NCO events drawn from negative binomial length distributions placed uniformly in the genome. The gene-converted status of each heterozygous marker overlapping the simulated NCO event was determined using a Bernoulli trial with the penetrance as the probability of each marker becoming gene converted. In Experiment E1, we simulated gene conversion tracts for 63 different parameter combinations, having mean 100, 300, and 1000 bp; 7 different values for the variance; and 0.5, 0.75, and 1.0 as the value for the penetrance—the proportion of heterozygous markers overlapping NCO events leading to gene conversions.

Since previous studies modelled the length distribution of NCO events with an exponential distribution, we reimplemented one of them, the method used by Li *et al.* (2019), for comparison. We also augmented that method to include the penetrance to see whether improvements on previous methods are due to the inclusion of penetrance or due to a more flexible model.

In Experiment E2, we simulated a mixture distribution by reusing two groups of datasets from E1, i.e. the ones with means 100 and 1000. For each pair of datasets from those two groups, we created a mixture with an equal probability of producing gene conversion tracts from both distributions. This results in 147 different parameter combinations (three different values of penetrance and seven different values for the variances of each of the two distributions having means 100 and 1000).

Each experiment was repeated 200 times to get a distribution for the inferred mean tract length and inferred number of NCO events.

Using the data from the first 20 repeats of E2 and each parameter combination (a total of 2940 datasets), we show that confidence intervals for the distribution mean and the total number of NCO events can be estimated by running NCOurd on resampled gene conversion tracts.

For each such dataset, we create 200 bootstrap samples of gene conversion tracts and run NCOurd on each sample. Thus, we get 200 estimates of the mean length and the number of NCO events from which we compute 95% confidence intervals.

Finally, we count how often the actual mean falls within the computed confidence intervals. This should happen in approximately 95% of the cases for all parameter combinations.

We calculated the confidence intervals for the published datasets using this bootstrapping method.

2.5 Determining the number of mixture components

Choosing the appropriate number of mixture components for the model is important. If too few components are used, the estimated length distribution of NCO events can lack some features of the true length distribution. More mixture components will always produce higher likelihood estimates, but too many will lead to overfitting the data.

We use a likelihood ratio test (Neyman and Pearson 1933) to determine the appropriate number of mixture components. A model with the length distribution of NCO events being a mixture of n negative binomial distributions is nested in a model with the length distribution of NCO events being a mixture of $n + 1$ negative binomial distributions. The larger model has three additional degrees of freedom (an extra mixture weight and two extra parameters for the negative binomial distribution). Let $\mathcal{L}_{\mathrm{n}}$ and $\mathcal{L}_{\mathrm{n+1}}$ be the maximum likelihood with models having n and $n + 1$ components, respectively. The null hypothesis is that the data obey the distribution of the model with n components. Given the null hypothesis, we have that $-2\log(\mathcal{L}_{\mathrm{n+1}}/\mathcal{L}_{\mathrm{n}})$ approximates a χ^2 random variable with three degrees of freedom (Wilks 1938), which in turn can be used to obtain a P-value.

Similarly, a model with the length distribution of NCO events as a geometric distribution is nested in a model with the length distribution of NCO events as negative binomial distributions. The larger model has one additional degree of freedom (an extra parameter).

Let $\mathcal{L}_G$ and $\mathcal{L}_N$ be maximum-likelihood estimates with geometric and negative binomial distribution models, respectively. Assuming the null hypothesis that geometric distribution is the correct choice of a model, then $-2\log(\mathcal{L}_N/\mathcal{L}_G)$ approximates a χ^2 random variable with one degree of freedom, which can be used to obtain a *P*-value.

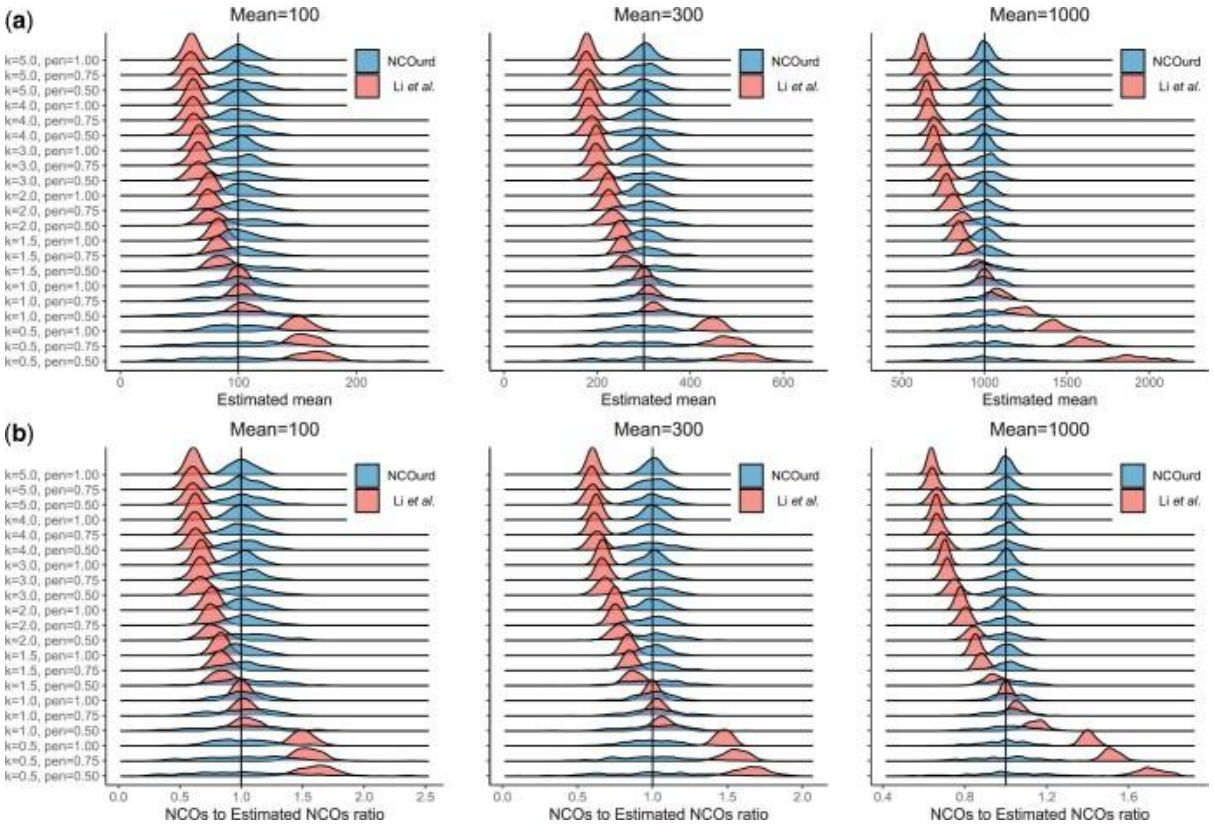
To determine the number of mixture components to use, we start with one component and repeatedly add a component while the likelihood ratio gives a significant *P*-value less than .05.

3 Results

For the E1 experiments, we evaluated how well the two methods inferred the mean length and the number of NCOs for 1000 gene conversion tracts from a single distribution (one mixture component). We did this for each of the simulated datasets, repeating each experiment 200 times to get a distribution for the inferred values.

For NCOurd, the median of the inferred values for the mean length and number of NCOs agrees with their true values in all cases, while the Li *et al.* method is biased away from the mean when the distribution deviates from an exponential distribution (Fig. 2a and b). Supplementary Table S2 shows the downward bias of the mean estimate when the penetrance is omitted in the Li *et al.* method. Supplementary Fig. S2a shows how the variance of the mean estimation decreases with an increased number of tracts; the variance is inversely proportional to the number of tracts (Supplementary Table S3). As the standard deviation is the square root of the variance, a 4-fold number of input gene conversion tracts is needed to halve the standard error of the estimated mean NCO event length.

Figure 2.



Results for simulated data. ***(a)*** *Distribution of estimated mean NCO event length.* ***(b)*** *Distribution of the number of NCOs. NCOurd in blue and the method from Li et al. in red. Each experiment consists of 1000 simulated tracts from the given negative binomial distribution and was repeated 200 times to obtain a distribution. The penetrance for the datasets is 'pen' and the variance is mean²/k.*

Next, for the E2 experiments, we evaluated how well NCOurd inferred mean NCO event length and the number of NCO events for 2000 gene conversion tracts from a mixture of two distributions with means 100 and 1000, using 1000 tracts from each distribution. For a mixture of two simulated datasets, the median values for mean length and the number of NCOs are close to the true values, but the variance is large if the small distribution has small *k* and the penetrance is low (Supplementary Fig. S2b and c).

For the simulated mixture distributions, we bootstrapped the input gene conversion tracts to estimate 95% confidence intervals. The mean NCO event length used for the simulations was within the estimated confidence interval 2784 out of 2940 or 94.7% [95% confidence interval (CI), 93.8%–95.4%], and the number of NCO events in the simulations was within the estimated confidence interval 2786 out of 2940 or 94.8% (95% CI, 94.0%–95.5%). This shows that bootstrapping the input gene conversion tracts accurately estimates the CIs.

Finally, we employed NCOurd on two publicly available gene conversion datasets.

The first dataset is from Li *et al.* (2019) consisting of 1575 gene conversion tracts from cross-breeding two highly divergent mouse strains. The mouse dataset contains a very dense set of markers with mean and median distance between consecutive markers of 179 and 67, respectively. Each strain was almost fully inbred, making the task of identifying gene conversions that accumulated over a few generations relatively straightforward since there were only three possible haplotype pairs: homozygous for either strain and heterozygous. On average, the set of heterozygous markers is half of the genome for each mouse. Penetrance was estimated as 0.9, computed as the fraction of gene-converted markers in all gene conversion tracts, excluding boundary markers (first and last marker of each tract).

For the inbred mouse data, we used a likelihood ratio test to determine the number of components for the mixture. We reject the null hypotheses that a single geometric distribution fits the data as well as a single negative binomial distribution with P-value $1.7 \cdot 10^{-23}$, and that a single negative binomial distribution fits the data as well as a mixture of two negative binomial distributions with P-value $5.1 \cdot 10^{-28}$. We cannot, however, reject the null hypothesis that a mixture of two negative binomial distributions fits the data as well as a mixture of three negative binomial distributions, P-value 0.36 (Supplementary Table S1). Out of the three models, the single geometric distribution is most comparable to the methodology of Li *et al.* (2019).

Using two negative binomial distributions gives an estimate of mean NCO event length of 42 bp (95% CI, 24–48) and an estimate of the number of NCO events repaired with the homologous chromosome per meiosis of 172.4 (95% CI, 152.2–290.1) (Table 1). Combining the estimated number of NCO events with the estimated number of crossovers gives 199.3 (95% CI, 179.0–316.9.1) DSBs repaired with the homologous chromosome using NCOurd. Li *et al.* estimate the average number of NCO events as 274 and the number of DSBs as 300.5. It is worth noting that Li *et al.* (2019) proposed a strand-aware model to obtain this estimate, while under a standard strand-unaware model, their estimate was 465.

Table 1.

	Distributions	Mean	NCOs	*P*-value
Mice	1	18.0 (12.5–26.0)	396 (273–552)	1.7·10−23
Li *et al.* (2019)	2	41.8 (24.3–47.9)	172 (152–290)	5.1·10−28
Human maternal	1	10.6 (7.3–17.5)	24445 (16202–31 603)	4.3·10−46
Halldorsson *et al.* (2016)	2	41.9 (16.4–2925)	6858 (108–16 692)	0.0035
Human paternal	1	2.6 (1.65–11.9)	41994 (8676–51 537)	5.3·10−48
Halldorsson *et al.* (2016)	2	177 (61.0–389)	791 (330–2098)	4.1·10−13

Estimated mean length of NCO events and the number of NCO events per meiosis using a single negative binomial distribution and a mixture of two negative binomial distributions for the datasets from Li et al. (2019) and Halldorsson et al. (2016).

Notes: Estimates are obtained with NCOurd with 95% CIs shown in parentheses. P-values were obtained with likelihood ratio tests, comparing single negative binomial distributions against single geometric distributions and mixtures of two negative binomial distributions against single negative binomial distributions.

The authors estimated the mean NCO event length separately depending on the controlling PRDM9 allele and inferred the mean length of NCO events as 30 and 41 bp for events controlled by the *Cast* and *Hum* PRDM9 alleles, respectively. Applying NCOurd, we estimate the mean NCO event lengths 17 (95% CI, 9.5–47) and 48 (95% CI, 40–57) for *Cast* and *Hum* controlled events, respectively, suggesting we lack power to determine whether their means differ.

A detailed description of the method used to estimate the number of NCO events can be found in the Supplementary Information, together with a comparison with the method used in Li *et al.* (2019).

The second dataset is from Halldorsson *et al.* (2016) consisting of 504 gene conversion tracts found and verified in large sequenced families. The dataset contained 257 paternal and 247 maternal gene conversion tracts. As gene conversions are known to behave differently depending on the sex of the transmitting parent (Halldorsson *et al.* 2016) we ran NCOurd separately on the paternal and maternal gene conversion tracts. Two maternal gene conversion tracts were excluded since they spanned more than 100 000 bp. Penetrance was estimated 0.52 using the method described above for the mouse dataset, and the mean and median distance between consecutive informative markers in the dataset is 2745 and 561, respectively.

Using a likelihood ratio test on paternal and maternal gene conversion tracts, we rejected the null hypothesis that a single negative binomial distribution fits the data as well as a mixture of two negative binomial distributions with *P*-values $4.1 \cdot 10^{-13}$ and 0.0035, respectively. But we could not reject the null hypothesis that a mixture of two negative binomial distributions fits the data as well as a mixture of three negative binomial distributions, *P*-values 0.867 and 0.256, respectively.

We estimate the mean paternal and maternal NCO event length as 177 bp (95% CI, 61.0–389) and 41.9 bp (95% CI, 16.4–2925), respectively, and the number of paternal and maternal NCO events as 791 (95% CI, 330–2098) and 6858 (95% CI, 108–16 692), respectively (Table 1).

For NCOurd, the confidence intervals for the mean length and the number of maternal NCO events span two orders of magnitudes suggesting that a much larger dataset is needed for an accurate estimate. The confidence intervals for the mean length and the number of paternal NCO events are more reasonable. Using the method from Li *et al.* (2019), we get the estimate for mean NCO event length as 2506 (95% CI, 840–5415) and 21014 (95% CI, 9642–35 100) for paternal and maternal gene conversion tracts, respectively. Due to the presence of very long gene conversion tracts, the method from Li *et al.* (2019) fails to produce plausible mean estimates.

4 Discussion

We modelled the length distribution of NCO events and implemented NCOurd—an EM algorithm to infer NCO event lengths as a mixture of negative binomial distributions that best fit a gene conversion tract dataset. We have shown by simulations that the NCOurd accurately

infers the mean length of NCO events and the number of NCO events occurring during meiosis and can be used to estimate confidence intervals for the inferred values. We demonstrated NCOurd on publicly available mouse and human gene conversion datasets to infer the mean lengths and the number of events for the underlying NCOs creating these gene conversions. For both datasets, a mixture of two negative binomial distributions fits the data significantly better than a single distribution. We presume that this is due to NCO events arising via multiple pathways, such as the synthesis dependent strand annealing and double Holliday junction pathways (Supplementary Fig. S1).

The main limitation of NCOurd is that the method requires more extensive data than is generally made available in gene conversion studies and that the method could be biased if the gene conversion process deviates drastically from the model assumptions. However, future researchers can easily create the extra data needed, the informative marker sets, and informative markers flanking each gene conversion tract. Most of the model assumptions are used to calculate the input for the EM algorithm, the detection function, and the tract functions. The strongest modelling assumptions are that the penetrance is fixed (i.e. all informative markers contained in any NCO event have the same probability of becoming gene converted) and NCO events are uniformly distributed in the genome. In the Supplementary Information, we discuss how the model assumptions can be relaxed.

As NCOs are understudied compared with the other contributors to genetic diversity, mutations, and crossovers, we hope this method will help us gain a better insight into NCOs and the meiotic recombination process in general. The true length distributions of NCO events will likely remain unknown for some time, but NCOurd can shed light on some of the features of the distribution, difference between the sexes, and eventually further our understanding of processes in oocytes during meiotic arrest leading to an increased number of crossovers and gene conversions.

Supplementary Material

btad485_Supplementary_Data

Click here for additional data file. (1MB, pdf)

Contributor Information

Marteinn T Hardarson, deCODE genetics, Reykjavik 102, Iceland; School of Technology, Reykjavik University, Reykjavik 102, Iceland.

Gunnar Palsson, deCODE genetics, Reykjavik 102, Iceland.

Bjarni V Halldorsson, deCODE genetics, Reykjavik 102, Iceland; School of Technology, Reykjavik University, Reykjavik 102, Iceland.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

The authors are employees of deCODE genetics/Amgen.

Financial support

None declared.

Data availability

The datasets were derived from sources in the public domain: Li R, Bitoun E, Altemose N *et al.* A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nat Commun 2019;10:3900–15. https://www.nature.com/articles/s41467-019-11675-y#Sec24. Halldorsson BV, Hardarson MT, Kehr B *et al.* The rate of meiotic gene conversion varies by sex and age. Nat Genet 2016;48:1377–84. https://www.nature.com/articles/ng.3669#Sec25. Jónsson H, Sulem P, Kehr B et al. Data descriptor: whole genome characterization of sequence diversity of 15,220 icelanders. Sci Data 2017;4:170115. https://www.ebi.ac.uk/ena/browser/view/PRJEB15197.

References

1. Betrán E, Rozas J, Navarro A. et al. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 1997;146:89–99. [DOI] [PMC free article] [PubMed] [Google Scholar]
2. Dempster AP, Laird NM, Rubin DB. et al. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol) 1977;39:1–22. [Google Scholar]
3. Halldorsson BV, Hardarson MT, Kehr B. et al. The rate of meiotic gene conversion varies by sex and age. Nat Genet 2016;48:1377–84. [DOI] [PMC free article] [PubMed] [Google Scholar]
4. Jónsson H, Sulem P, Kehr B. et al. Data descriptor: whole genome characterization of sequence diversity of 15,220 icelanders. Sci Data 2017;4:170115. [DOI] [PMC free article] [PubMed] [Google Scholar]
5. Lam I, Keeney S.. Mechanism and regulation of meiotic recombination initiation. Cold Spring Harb Perspect Biol 2014;7:a016634. [DOI] [PMC free article] [PubMed] [Google Scholar]
6. Li R, Bitoun E, Altemose N. et al. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nat Commun 2019;10:3900–15. [DOI] [PMC free article] [PubMed] [Google Scholar]
7. Mansai SP, Innan H.. The power of the methods for detecting interlocus gene conversion. Genetics 2010;184:517–27. [DOI] [PMC free article] [PubMed] [Google Scholar]
8. Mansai SP, Kado T, Innan H. et al. The rate and tract length of gene conversion between duplicated genes. Genes (Basel) 2011;2:313–31. [DOI] [PMC free article] [PubMed] [Google Scholar]
9. Miller DE, Takeo S, Nandanan K. et al. A whole-chromosome analysis of meiotic recombination in *Drosophila melanogaster*. G3 (Bethesda) 2012;2:249–60. [DOI] [PMC free article] [PubMed] [Google Scholar]
10. Miller DE, Smith CB, Kazemi NY. et al. Whole-genome analysis of individual meiotic events in Drosophila melanogaster reveals that noncrossover gene conversions are insensitive to interference and the centromere effect. Genetics 2016;203:159–71. [DOI] [PMC free article] [PubMed] [Google Scholar]
11. Neyman J, Pearson ES.. IX. On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc Lond Ser A 1933;ccxxxi:289–337. [Google Scholar]

12. Setter D, Ebdon S, Jackson B.. et al. Estimating the rates of crossover and gene conversion from individual genomes. Genetics 2022;222:iyac100. 10.1093/genetics/iyac100. [DOI] [PMC free article] [PubMed] [Google Scholar]
13. Wall JD, Robinson JA, Cox LA.. High-Resolution estimates of crossover and noncrossover recombination from a captive baboon colony. Genome Biol Evol 2022;14. 10.1093/gbe/evac040. [DOI] [PMC free article] [PubMed] [Google Scholar]
14. Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 1938;9:60–2. 10.1214/aoms/1177732360. [DOI] [Google Scholar]
15. Williams AL, Genovese G, Dyer T.. et al. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. eLife 2015;2015. 10.7554/eLife.04637. [DOI] [PMC free article] [PubMed] [Google Scholar]

Associated Data

This section collects any data citations, data availability statements, or supplementary materials included in this article.

Supplementary Materials

btad485_Supplementary_Data

Click here for additional data file. (1MB, pdf)

Data Availability Statement

The datasets were derived from sources in the public domain: Li R, Bitoun E, Altemose N *et al.* A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nat Commun 2019;10:3900–15. https://www.nature.com/articles/s41467-019-11675-y#Sec24. Halldorsson BV, Hardarson MT, Kehr B *et al.* The rate of meiotic gene conversion varies by sex and age. Nat Genet 2016;48:1377–84. https://www.nature.com/articles/ng.3669#Sec25. Jónsson H, Sulem P, Kehr B et al. Data descriptor: whole genome characterization of sequence diversity of 15,220 icelanders. Sci Data 2017;4:170115. https://www.ebi.ac.uk/ena/browser/view/PRJEB15197.

Paper III

Title:

Complete human recombination maps

Authors:

Marteinn T Hardarson [1,2,✉], Gunnar Gunnar Palsson[1], Marteinn T. Hardarson[1,2], Hakon Jonsson[1], Valgerdur Steinthorsdottir[1], Olafur A. Stefansson[1], Hannes P. Eggertsson[1], Sigurjon A. Gudjonsson[1], Pall I. Olason[1], Arnaldur Gylfason[1], Gisli Masson[1], Unnur Thorsteinsdottir[1,3], Patrick Sulem[1], Agnar Helgason[1,4], Daniel F. Gudbjartsson[1,5], Bjarni V. Halldorsson[1,2,*], Kari Stefansson[1,3,*]

Affiliations:

1 deCODE genetics / Amgen Inc., Sturlugata 8, Reykjavik, Iceland

2 School of Technology, Reykjavik University, Reykjavík, Iceland

3 Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland

4 Department of Anthropology, University of Iceland, Reykjavik, Iceland

5 School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland

* Corresponding authors

*Correspondence to: Bjarni V. Halldorsson (Bjarni.Halldorsson@decode.is)

Abstract

Human recombination maps are a valuable resource for association and linkage studies and crucial for many inferences of population history and natural selection. Existing maps[1–5] are based solely on crossover (CO) recombination, omitting the more common form of recombination[6] – non-crossovers (NCOs) – due to the difficulty in detecting them. Using whole-genome sequence (WGS) data in families, we estimate the number of NCOs transmitted from parent to offspring and derive complete, sex-specific recombination maps including both NCOs and COs. Mothers have fewer but longer NCOs than fathers, and oocytes accumulate NCOs in a non-regulated fashion with maternal age. Recombination, primarily NCO, is responsible for 1.8% (95% CI: 1.3-2.3) and 11.3% (95% CI: 9.0-13.6) of paternal and maternal de novo mutations (DNMs), respectively, and may drive the increase in DNMs with maternal age. NCOs are substantially more prominent than COs in centromeres, possibly to avoid large-scale genomic changes that may cause aneuploidy. Our results demonstrate that NCOs highlight, to a much greater extent than COs, the differences in the meiotic process between the sexes, where maternal NCOs may reflect the safeguarding of oocytes from infancy till ovulation.

Main

Sexually reproducing organisms make gametes with meiosis[7], in which homologous chromosomes recombine, ensuring proper segregation of chromosomes as well as mixing the genetic material that is passed on to the offspring. Human genetic diversity and our ability to evolve, and thereby, adapt, is generated by DNMs and meiotic recombination. DNMs are known to occur, in part, as a result of recombination[1,8–11], but, as NCOs are underreported, the extent to which recombination contributes to DNMs in the offspring remains unknown.

Meiosis is initiated by the duplication of the genetic material and subsequent formation of double-strand breaks (DSBs)[12], occurring mainly in hotspots defined by the histone methyltransferase PRDM9[13]. PRDM9 recruits SPO11, which, together with TOP6BL[14], mediates DNA cleavage, resulting in DSBs[15]. The 5' ends of double stranded DNA at a DSB are resected (Figure 1a), leaving an overhang of 3' single-stranded DNA (ssDNA) branches[16]. One or both 3' branches subsequently invade the homologous chromosome to repair the DSB through DNA resynthesis resulting in a recombination, either a CO or an NCO[6,17].

For DSBs repaired as NCOs, the resected region around the DSB is partially repaired using the homologous chromosome as a template. Thus, NCOs manifest as short transitions between haplotypes of different grandparental origin, distinguished only at markers that are heterozygous in the transmitting parent (Figure 1b). NCOs that have incorporated a donor sequence that includes a heterozygous marker in the parent will contain heteroduplex DNA[16] with mismatched base-pairings that must be resolved. Some of these are resolved to match the original haplotype, leaving no evidence of the NCO in the offspring, but, when resolved to match the donor sequence, a haplotype transition will be detectable, referred to as a gene conversion[18].

Meiotic DSBs may be repaired using the sister chromatid as the template in which case neither COs nor NCOs would be detectable due to the near perfect sequence identity between the sister chromatids. Sister chromatid repair is suppressed in meiosis and therefore rare[6]. We focus on DSBs that were repaired using the homologous chromosome in meiocytes that had resulted in live births.

The meiotic process is dramatically different in males and females[19]. Spermatogonia undergo continual mitosis from puberty throughout life, while meiosis is initiated in spermatocytes only a few weeks before haploid sperm cells are fully formed. In contrast, females are born

with a limited supply of oocytes with the meiotic process suspended in prophase I, only to be completed at ovulation and fertilization decades later[20]. Oocytes are thus exposed much longer to possible exogenous agents that may adversely affect DNA integrity[21] and some of their NCOs may be due to repair of DSBs sustained and accumulated over the years.

We derive complete sex-specific recombination maps, including both NCOs and COs, thereby completing the high-resolution mapping of sex-specific human recombination, a task we started over 20 years ago[2]. We identified NCOs by looking for gene conversions (Supplementary Note 1.1) transmitted to offspring in 5,420 trios in 2,132 Icelandic nuclear families where both parents and at least two children have been WGS[22] (Table S1). We estimated the length distribution of NCOs[23] and their average number per meiosis, and constructed genetic maps of NCOs (Figure 1c). By identifying DNMs near NCOs in the same individuals, we derive sex-specific mutational spectra of DNMs resulting from NCOs and estimate the overall contribution of recombination to mutagenesis.

Gene conversions and observed NCOs

We restricted our analysis to autosomal variants with frequencies above 0.5% (Supplementary Note 0), resulting in 8,893,878 sequence variants, 8,270,254 SNPs and 623,624 indels (Table S2). Gene conversions can only be determined at variants where the transmitting parent is heterozygous, and we require the other parent to be homozygous (Supplementary Note 0), giving a total of 4,229,340,533 and 4,288,524,590 informative marker-proband-pairs (MPPs) for paternal and maternal meiosis, respectively (Extended Data Table 1).

When the genotypes of parent and offspring are phased with parent-of-origin determined, contiguous haplotype segments of a given grandparental origin can be identified in offspring (Supplementary Note 0). Telomeric haplotype segments and haplotype segments longer than 100kb are considered background, consistent with reciprocal recombination, while haplotype segments shorter than 100kb are considered gene conversion candidates[24] (Figure 1b-i). Consecutive gene conversion candidates flanked by two background haplotypes of the same grandparental origin are considered to be an observed NCO (oNCO), and we note that oNCOs may be complex, containing both gene-converted and non-gene-converted MPPs (Figure 1b-ii). When the grandparental origins of the flanking background haplotypes differ, the result is a complex CO[1] (Figure 1b-iii).

We identified 17,109 paternal and 45,653 maternal gene-converted MPPs – over 30 times more than the largest family-based study to date[24] (Extended Data Table 1). Gene conversion rates are consistent with earlier estimates[24,25]. The total number of oNCOs was 12,948 and 15,712, an average of 2.39 (95% CI: 2.33-2.44) and 2.90 (95% CI: 2.83-2.97) oNCOs per offspring for paternal and maternal meioses, respectively, with the mean number of converted MPPs per oNCO being 1.32 (95% CI: 1.26–1.40) and 2.91 (95% CI: 2.79–3.03) for paternal and maternal meiosis, respectively. A larger study[26] of gene conversions identified from identity by descent segments was published concurrently with this research.

We observed a GC bias in gene-converted SNPs (Extended Data Table 1), where strong (C or G) alleles are more likely to be transmitted than weak (A or T), consistent with earlier studies[24,25,27,28]. We note a slightly higher GC bias (P = 0.002, bootstrap test) for maternal transversion: 67.1% (95%CI; 65.8-68.5) vs. 65.1% (95% CI: 64.2-65.9) for transitions, with 50% indicating no bias. For indels, we replicate[11,24] insertion bias (the longer allele is retained more frequently than the shorter one). This bias is restricted to maternal meiosis, where it is 62.5% (95% CI: 60.4-64.4, $P = 1.7 \cdot 10^{-36}$, binomial test).

Mothers transmit fewer and longer NCOs

From oNCOs identified for each proband we estimate, separately for fathers and mothers, the length distribution of all NCOs, including those that are not observable due to the absence of gene converted MPPs. We model the length distributions as mixtures of several components (Supplementary Note 0) using our previously described method[23]. Very short NCOs (<10bp) would not be accurately identified from parent offspring transmissions; however, our data suggest that these are rare (*Extended Data Fig. 1*). The length distributions consist of multiple components (Table S3, Table S4), which we group into short components (<1kb) and extended components. Short NCOs are, on average, 123bp (95% CI:94-135) and 102bp (95% CI:71-125) for paternal and maternal transmissions, respectively (Extended Data Table 1). The size difference between the sexes is not significant (P = 0.272, bootstrap test) and the results are comparable to earlier NCO-length estimates of 55-290bp in humans[29], 30bp in baboons[30], 155bp in rhesus macaques[31], and in mice[32]: 86bp (range 23-148) in oocytes, 68bp (range 15-124) in spermatocytes. We note that all these estimates are shorter than the size of resected regions around DSBs, estimated as approximately 1400bp in human testis[10], 1640bp in yeast[33] and 2200bp in mice[34], all with wide variability.

Extended NCOs are, on average, 7.2kb (95% CI: 1.8-11.8), and 9.1kb (95% CI: 6.8-10.8), for paternal and maternal transmissions, respectively. Extended NCOs are just 1.14% (95% CI: 0.62-5.88) of paternal NCOs but are more common in mothers, 6.89% (95% CI: 5.10-9.93), and appear to be generated from a different process than the paternal ones with stronger allele selection biases (Extended Data Table 1, Table S5).

Based on the inferred length distribution, we estimate that there are on average 105.0 (95% CI, 95.9-125.0) and 81.6 (95% CI, 66.7-103.1) NCOs per offspring for paternal and maternal meiosis respectively (Figure 1c, Supplementary note 4.3), indicating that the vast majority of NCOs are unobserved. We estimate that there are more NCOs in paternal meiosis ($P = 0.042$, bootstrap test) in contrast with the greater number of COs in maternal meiosis[1] (~42 vs. ~26 for paternal meiosis). Estimating DSBs by combining our estimates for NCOs and COs (Supplementary note 0) we find that the average number of DSBs is not significantly different between the sexes ($P = 0.120$, bootstrap test), with the average number of DSBs per meiocyte estimated as 474 (95% CI: 438-554) and 410 (95% CI: 350-496) for paternal and maternal meiosis, respectively. Cytological measurements of early recombination intermediates in human meiocytes give an estimate of approximately 400 DSBs in spermatocytes[35], and 370 DSBs in oocytes[36].

NCOs and COs are positively correlated for both paternal and maternal meiosis ($r = 0.058$, $P = 2.0 \cdot 10^{-5}$ paternal, $r = 0.100$, $P = 1.6 \cdot 10^{-13}$ maternal, Supplementary note 0), consistent with an increase in DSBs leading to an increase in both NCOs and COs, in both fathers and mothers, indicating that an increase in COs, does not come at the cost of fewer NCOs. These results are consistent with a model where chromosomal axis lengths vary between germ cells while the distance between DSB events varies less[37]. The relatively modest correlation may in part be explained by crossover homeostasis[17,38].

Genic DSBs favor NCO resolution

We constructed sex-specific maps of human NCO recombination on a grid of overlapping 3Mb windows at intervals of 1Mb (Supplementary note 0), achieving resolution on par with early CO maps[2,4]. Human recombination maps[1–5] have only accounted for CO recombination but combining our NCO maps with existing CO maps[1] allows us to explore all recombination as well as variations in DSB resolution through differences in the NCO and CO maps. We display the sex-specific NCO maps for chromosome 19 in Figure 2a, and the full set of recombination maps in Fig. S1.

Large scale features of the CO map, such as the elevation of paternal rates near telomeres, are mirrored in the NCO map, and we find that these maps are highly correlated – the genome-wide correlation is 0.68 (95% CI: 0.65-0.71) and 0.36 (95% CI: 0.32-0.40) for paternal and maternal maps, respectively. There is significant positive correlation between the paternal CO and NCO maps for all autosomes, while the maternal maps are significantly positively correlated for all autosomes except for 6, 16, 17, 21, and 22 (Table S6).

Highlighting the shared etiology of NCOs and COs, locations of oNCOs are highly correlated with sex-specific CO recombination hotspots[1] (Extended Data Table 2, Supplementary Note 0), where their frequency is increased 22.4 and 13.7-fold for fathers and mothers, respectively. Underscoring this further, the hotspot usage of oNCOs associates with a *PRDM9* sequence variant (rs2973614, MAF = 3.2%, effect = $-0.6 \cdot SD$, $P = 1.1 \cdot 10^{-21}$, Supplementary Note 0) that also associates strongly with hotspot usage of COs[1,13] (effect = $-1.7 \cdot SD$, $P = 4.3 \cdot 10^{-2382}$).

NCOs are also highly enriched within regions of annotated DSB activity (Extended Data Table 2), with the highest enrichment found in DNA Meiotic Recombinase 1 (DMC1) hotspots[10,11], 42.4 and 19.3-fold, for paternal and maternal NCOs, respectively. NCOs also co-localize with binding sites of PRDM9[39] (Supplementary Note 0), and this binding is biased with respect to the center position of oNCOs; the center of the PRDM9 binding motif is, on average, 36 bp (95% CI: 26.5-41 bp) downstream of the center position of the oNCO (Table S7), consistent with PRDM9 inducing DSBs preferentially upstream of its binding site.

There is an excess of NCOs in exons and transcribed regions[40] (Extended Data Table 2); the enrichment in transcribed regions is 1.52 (1.38, 1.67) and 1.26 (1.17, 1.36) for paternal and maternal meiosis, respectively. These results run counter to the well documented depletion of COs in such regions[1,41] (Extended Data Table 2). Transcription is marked by H3K36 trimethylation[40] – known to be deposited by the DSB-associated PRDM9[42] – which may be implicated in regulating DSB pathway choice[43], possibly explaining why DSB resolution is NCO-enriched in transcribed regions.

Pericentromeric DSBs are resolved as NCOs

The NCO/CO ratio is an indicator of the propensity of DSBs to be resolved as NCOs rather than COs and here we estimate this ratio as 7.84 (95% CI: 7.24-9.49) and 3.91 (95% CI: 3.15-4.93) per meiocyte for paternal and maternal meiosis, respectively. These estimates are

concordant with prior estimates[29,44] and with the Housworth-Stahl modelling of crossover interference (Supplementary Note 0).

We analyze localized variations in the NCO/CO ratio through a normalized NCO/CO difference, Δ_{DSB}, which measures how DSB resolution deviates locally from the genome-wide average of the NCO/CO ratio (Figure 2b). Values of Δ_{DSB} range from −1 (only COs) to 1 (only NCOs), where $\Delta_{DSB} > 0$ indicates that DSBs are being resolved as NCOs at a higher rate than the genome-wide average; referred to as NCO-enriched DSB resolution, whereas $\Delta_{DSB} < 0$ indicates CO-enriched DSB resolution.

Paternal DSB rates are highly elevated in the 10Mb region closest to the telomere, with both NCOs and COs contributing (Figure 2c, *Extended Data Fig. 2*). Paternal DSB resolution is NCO-enriched in the 1Mb interval closest to the telomere, but in the range of 3-10Mb from the telomere the resolution is CO-enriched[45]. Maternal DSB rates, on the other hand, are only mildly elevated in the 10Mb region closest to the telomere, driven by elevation of NCOs and NCO-enriched resolution up to 5Mb from the telomere. Further away, up to 10Mb, maternal CO rates are also elevated, and the DSB resolution is balanced between COs and NCOs (*Extended Data Fig. 2*).

While COs are known to be suppressed near centromeres[46–48], our results (Figure 2e) demonstrate that the processes that govern that suppression do not affect NCOs[49], and consequently, we see NCO-enriched DSB resolution near centromeres (Figure 2f, Fig. S1). This is especially evident for paternal meiosis, where the average Δ_{DSB} is about 0.85 within 1Mb of the centromere, (indicating that almost all DSBs are resolved as NCOs) and remains significantly positive up to 10Mb away from the centromere. For maternal meiosis this suppression is less effective as the average Δ_{DSB} is positive only within 1Mb from the centromere.

Recombination map values and DSB resolution as functions of both GC content and replication timing are displayed in *Extended Data Fig. 3* and *Extended Data Fig. 4*, respectively.

DNMs are enriched near oNCOs

We have previously described genomic regions with high NCO rate and maternal C>G DNM rate (C>G enriched regions; CGER[9]). However, as relatively few NCOs are detected per individual and oNCO data sets have been limited in size, we did not detect co-occurrence of DNMs and NCOs in the same meiosis. Here, we expanded this and explored co-localization

of oNCOs with 382,566 DNMs identified in the probands of our study (Supplementary Note 0). For each DNM, we measured the distance to the center of the nearest oNCO and found that within 1kb of the oNCO centers DNM rates are elevated 142-fold (95% CI: 106-183) and 125-fold (95% CI: 66-197) for paternal and maternal meiosis, respectively. To assess whether this enrichment is due to nucleotide composition or other genomic features of NCOs, we permuted the DNMs across the probands and recalculated the enrichment per permutation. We found no significant enrichment in the permutation, indicating that DNM and oNCO co-occurrence is not due to nucleotide composition at oNCO sites (Table S8). DNM rate elevation drops rapidly for paternal meiosis and is mostly observable within 3kb from the NCO, while significant elevation can be found up to 100kb away from NCOs for maternal meiosis (Extended Data Table 3, *Extended Data Fig. 5*, Table S9, Table S10).

Sex difference in DNM spectra near NCOs

To highlight the differences in the mechanisms leading to DNMs that arise due to NCOs versus other DNMs, we investigated the mutation spectra of the two groups of DNMs. Using our analysis of the extent of rate elevation, we use paternal and maternal DNMs within 3kb and 100 kb, respectively, to represent DNMs that arise due to NCOs. We group together mutations and their reverse complement (mutation class), and consider C>T DNMs inside and outside of CpG context separately and compute the frequency of each mutation class (Mutation spectrum; Figure 3a, Table S11). We then test whether the spectrum of DNMs near NCOs is different from the genomic background ($P = 0.075$, paternal; $P = 1.5 \cdot 10^{-87}$, maternal, χ^2-test). The strong maternal difference is mostly explained by a very large increase in the percentage of DNMs in the C>G mutation class near NCOs, 37.2% (95% CI: 31.6 to 43.2) compared to 7.55% (95% CI: 7.41 to 7.68) genome-wide ($P = 4.65 \cdot 10^{-49}$, Fisher's test). This directly implicates NCOs with the regional enrichment of maternal C>G DNMs in CGER. We next checked whether the mutational processes for NCO are CO are similar by comparing spectra for DNMs near NCOs and COs (Extended Data Fig. 6). We do not find a significant difference between the spectra for NCO-proximal and CO-proximal DNMs in paternal meiosis ($P = 0.10$, χ^2-test), while for maternal meiosis they are different ($P = 1.9 \cdot 10^{-12}$, χ^2-test).

NCOs and COs are the product of DSB repair with single strand DNA (ssDNA) intermediates which are mutation prone[50], and mutational/repair processes operating on specific types of nucleotides in single strand context would create strand asymmetry in the mutation spectra[51].

In line with this expectation, we found strand asymmetry of CpG>TpG DNMs near paternal COs[1] and the same pattern has been observed for C>N DNMs and sequence variants around DSB hotspots[11,52]. Here we detect strand asymmetry in the spectrum for NCO-proximal DNMs (Figure 3b, *Extended Data Fig. 7*). CpG>TpG asymmetry, a signature of spontaneous deamination of methylated cytosines[53], is present in paternal meiosis, OR = ∞ (95% CI: 2.51,∞ , $P = 4.7 \cdot 10^{-3}$, Fisher's test), but not in maternal meiosis. No other asymmetry signature was found in paternal meiosis, but in maternal meiosis we observed asymmetry in all C>N classes (*Extended Data Fig. 7*). These results, the long-range enrichment of C>G DNMs near maternal NCOs, and the hefty accumulation of strand coordinated C>G DNMs in aging mothers[9,54], indicate that the resected region around DSBs is larger than NCOs.

Age-related NCOs are not regulated

For maternal meioses, we find that the number of NCOs per offspring increases ($P = 5.73 \cdot 10^{-7}$, t-test), with maternal age at the birth of proband by 20.3 (95% CI: 15.7-24.9) events per decade (Table S12, Figure 3g), in line with the age-dependence observed for both CO[1] and gene conversion rates[24]. No age effects are observed for paternal meiosis (P = 0.78, t-test), as with paternal COs[1]. Combining the maternal estimates, we find an increase in the number of DSBs per meiocyte of 82.5 (95% CI: 64-01) events per decade (Table S12), with most new DSBs resolved as NCOs (Figure 3h). Thus, pregnancies carried to term by 20-year-old mothers are products of meiocytes with an average of 349 (95% CI: 329-368) DSBs, while those of 40-year-old mothers result from meiocytes with an average of 514 (95% CI: 488-539) DSBs, a 50% increase. We note that for parental age of 20, maternal meiocytes contain significantly fewer DSBs ($P<2 \cdot 10^{-3}$, bootstrap test) than paternal meiocytes, which contain on average 474 (95% CI: 438-554) DSBs, corresponding roughly to the average number of DSBs in 35-year-old mothers.

The increase in NCOs with maternal age occurs entirely outside of DMC1 hotspots with no change in the number of oNCOs inside the DMC1 hotspots (p = 0.88, t-test). Consequently, over a span of two decades from the age of 20, the percentage of oNCOs within DMC1 hotspots decreases from 29.0% (95% CI: 27.6-30.4) to 15.9% (95% CI: 14.0-17.8) (Table S12). Both values are significantly smaller than the corresponding results for paternal oNCO, which are 51.7% (95% CI: 50.7-52.7) and depend minimally on paternal age (Effect = −1.75% per decade, P = 0.032, t-test).

Maternal CO hotspots are another measure of programmed DSB activity in maternal meiosis. Here also, we see a decrease in the percentage of oNCOs that fall within hotspots, going from 32.8% (31.4, 34.3) at the age of 20 years to 22.1% (20.1, 24.1) at the age of 40 years (Table S12). Thus, nearly all the age increase in NCOs takes place outside CO hotspots.

These results suggest that maternal NCOs become less tightly regulated with age, but at the same time we find that maternal NCO rates increase disproportionally in CGER[9], with the proportion of oNCOs in CGER almost doubling for maternal age between 20 and 40, going from 13.8% (95% CI: 12.5-15.0) to 24.6% (95% CI: 22.9-26.3). This increase is much higher than the increase for COs and thus, the meiotic NCO/CO ratio increases dramatically in CGER, going from 3.0 (95% CI: 2.2-3.7), to 8.4 (95% CI: 7.3-9.5), while outside CGER it goes from 3.3 (95% CI: 3.1-3.5) to 4.7 (95% CI: 4.4-5.0) for maternal age of 20 to 40 (Table S12).

Sex specific contribution of DSB to DNMs

Extrapolating the mutation rate increase to all NCOs (Supplementary Note 0) allows us to assess the fraction of DNMs that can be attributed to NCOs, yielding an average of 1.69% (95% CI: 1.22-2.15) and 10.95% (95% CI: 8.74-13.03) for paternal and maternal meiosis, respectively. Mutagenicity was already established for COs[1] (Table S9), leading to estimates of about 0.11% and 0.38% for the CO contribution to the paternal and maternal mutation rates, respectively (Extended Data Table 3). The total contribution from both NCOs and COs, i.e., due to DSBs is therefore, 1.80% (95% CI: 1.29-2.31) and 11.3% (95% CI: 9.0-13.6) for paternal and maternal meiosis, respectively.

The elevation of the mutation rate in the regions from 3-100kb from maternal NCOs is more pronounced ($P<0.002$, bootstrap test) for NCOs inside CGER. Within CGER, we estimate that DSBs contribute 38.8% (95% CI: 31.4-46.7) of maternal DNMs. This percentage is strongly age dependent because of the age-related increase in the number of maternal DSBs. Thus, at the maternal age of 20 years the DSB contribution to the genome-wide DNM rate is 2.0% (95% CI: 0.8-3.7), increasing to 31.0% (95%CI: 18.6-44.8) at the age of 40. Within CGER, the corresponding contribution percentages are about 8.2% (95%CI: 1.9-15.8) and 80.4% (95% CI: 44.0-122.2). Thus, age-related NCOs within CGER are the main contributors to increased DNM rates in older mothers.

Discussion

We present, for both sexes, the first human NCO recombination maps for live offspring, along with complete recombination maps incorporating both COs and NCOs, and maps of DSB resolution. These maps are an important tool for exploring the meiotic process and a major steppingstone towards building a better understanding of the distribution of NCOs, DSB resolution, and the interplay between recombination and mutation, the two key processes underlying the generation of human genetic diversity. A better understanding of the recombination process may also allow us to recognize the conditions under which this process fails and results in aneuploidies and pregnancy loss.

The number of NCOs and COs are positively correlated in both paternal and maternal meiosis, and thus, the age increase observed in maternal COs[1,55,56] is not due to DSBs being preferentially resolved as COs as the mother ages, but to an overall increase in meiotic DSBs in oocytes that result in successful pregnancies.

While the human maps for NCOs and COs are highly correlated, there are, as in other species[6,17,57–59], genomic regions where COs are clearly avoided – not because of a depletion of DSBs, but due to DSB resolution. In transcribed regions COs are likely suppressed because of possible disruptive effects on coding integrity of genes[41], and pericentromeric regions where their presence has been associated with meiotic segregation errors and aneuploidy[49,60].

Chromosomal abnormalities may constitute a major cause of infertility and pregnancy loss as they are found in over 50% of miscarriages and only 0.1% of live births[61]. Chromosomal abnormalities are mostly maternally transmitted, with abnormal placement of COs likely playing a key role. We have recently discovered[62] that a missense variant in *SYCE2* associates with both pregnancy loss as well as several meiosis related phenotypes, among them distance of COs from telomere[1]. SYCE2 is a key protein involved in the assembly of the synaptonemal complex backbone, the protein lattice that affects pairing of homologous chromosomes during meiosis. Proper segregation of chromosomes during meiosis depends on a tightly regulated placement of COs.

Paternal DSBs appear much more tightly regulated, evidenced by the much higher fraction of oNCOs occurring within DMC1[10,11] and CO[1] hotspots. The difference becomes more prominent with maternal age, consistent with nearly all maternal age increase occurring

outside of DMC1 and CO hotspots. We postulate that the age increase in NCOs occurs under a different process than other NCOs; however, our data does not reveal these processes.

DNMs are enriched near NCOs with sex differences in both the range of impact on NCO-induced DNMs and the associated spectrum of DNMs. Paternal DNMs are primarily overrepresented within 3kb from an NCO, consistent with almost all paternal NCOs impacting a limited range around DSBs. Whereas we observe elevation of the maternal DNM rate up to 100kb.

The total mutational contribution of DSBs is about 1.8% and 11.3% in paternal and maternal meiosis respectively. For mothers, the contribution increases with age and is larger in CGER, where our results suggest that the age-related increase in DNMs is largely driven by NCOs.

The generation of new sequence diversity can be seen as an unwitting battle between the sexes. Mothers contribute mostly through COs whereas fathers do it through DNMs. Fathers also contribute more through NCOs than mothers, in that they yield more of them, whereas NCOs from mothers are longer and therefore contain more MPPs. We show that recombination contributes a substantial fraction of the DNMs. The rate of DNMs is increased in the areas flanking DSBs, no matter whether they are resolved through CO or NCO or come from the mother or the father.

References

1. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science (80-.).* **363**, eaau1043 (2019).

2. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).

3. Bhérer, C., Campbell, C. L. & Auton, A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.* **8**, 14994 (2017).

4. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861 (1998).

5. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).

6. Haber, J. *Genome stability*. (Garland Science, 2013).

7. Zickler, D. & Kleckner, N. Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harb. Perspect. Biol.* **7**, 1–28 (2015).

8. Bergman, J. & Schierup, M. H. Evolutionary dynamics of pseudoautosomal region 1 in humans and great apes. *Genome Biol.* **23**, (2022).

9. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).

10. Pratto, F. *et al.* Recombination initiation maps of individual human genomes. *Science* **346**, (2014).

11. Hinch, R., Donnelly, P. & Hinch, A. G. Meiotic DNA breaks drive multifaceted mutagenesis in the human germ line. *Science (80-.).* **382**, (2023).

12. Sun, H., Treco, D., Schultes, N. P. & Szostak, J. W. Double-strand breaks at an initiation site for meiotic gene conversion. *Nature* **338**, 87–90 (1989).

13. Baudat, F. *et al.* PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327**, 836–840 (2010).

14. Robert, T. *et al.* The TopoVIB-Like protein family is required for meiotic DNA double-strand break formation. *Science* **351**, 943–949 (2016).

15. Keeney, S., Giroux, C. N. & Kleckner, N. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88**, 375–384 (1997).

16. Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. & Stahl, F. W. The double-strand-break repair model for recombination. *Cell* **33**, 25–35 (1983).

17. Zickler, D. & Kleckner, N. Meiosis: Dances Between Homologs. *Annu. Rev. Genet.* **57**, 1–63 (2023).

18. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8**, 762–775 (2007).

19. Handel, M. A. & Schimenti, J. C. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat. Rev. Genet.* **11**, 124–136 (2010).

20. Gray, S. & Cohen, P. E. Control of Meiotic Crossovers: From Double-Strand Break Formation to Designation. Annual Review of Genetics **50**, (2016).

21. Collins, J. K. & Jones, K. T. DNA damage responses in mammalian oocytes. *Reproduction* **152**, R15–R22 (2016).

22. Gudbjartsson, D. F. *et al.* Sequence variants from whole genome sequencing a large group of Icelanders. *Sci. Data* **2**, 150011 (2015).

23. Hardarson, M. T., Palsson, G. & Halldorsson, B. V. NCOurd: Modelling length distributions of NCO events and gene conversion tracts. *Bioinformatics* btad485 (2023). doi:10.1093/bioinformatics/btad485

24. Halldorsson, B. V. *et al.* The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* **48**, 1377–1384 (2016).

25. Williams, A. L. *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife* **4**, (2015).

26. Browning, S. R. & Browning, B. L. Biobank-scale inference of multi-individual identity by descent and gene conversion. *Am. J. Hum. Genet.* **111**, 691–700 (2024).

27. 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).

28. Kostka, D., Hubisz, M. J., Siepel, A. & Pollard, K. S. The role of GC-Biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.* **29**, 1047–1057 (2012).

29. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* **36**, 151–156 (2004).

30. Wall, J. D., Robinson, J. A. & Cox, L. A. High-Resolution Estimates of Crossover and Noncrossover Recombination from a Captive Baboon Colony. *Genome Biol. Evol.* **14**, (2022).

31. Versoza, C. J. *et al.* Novel Insights into the Landscape of Crossover and Noncrossover Events in Rhesus Macaques (Macaca mulatta). *Genome Biol. Evol.* **16**, (2024).

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

33. Mimitou, E. P., Yamada, S. & Keeney, S. A global view of meiotic double-strand break end resection. *Science (80-.).* **355**, 40–45 (2017).

34. Paiano, J. *et al.* ATM and PRDM9 regulate SPO11-bound recombination intermediates during meiosis. *Nat. Commun.* **11**, 1–15 (2020).

35. Oliver-Bonet, M., Campillo, M., Turek, P. J., Ko, E. & Martin, R. H. Analysis of replication protein A (RPA) in human spermatogenesis. *Mol. Hum. Reprod.* **13**, 837–844 (2007).

36. Lenzi, M. L. *et al.* Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis i in human oocytes. *Am. J. Hum. Genet.* **76**, 112–127 (2005).

37. Wang, S. *et al.* Per-Nucleus Crossover Covariation and Implications for Evolution. *Cell* **177**, 326-338.e16 (2019).

38. Martini, E., Diaz, R. L., Hunter, N. & Keeney, S. Crossover Homeostasis in Yeast Meiosis. doi:10.1016/j.cell.2006.05.044

39. 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* **6**, e28383 (2017).

40. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).

41. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).

42. Powers, N. R. *et al.* The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. *PLoS Genet.* **12**, e1006146 (2016).

43. Pai, C. C. *et al.* A histone H3K36 chromatin switch coordinates DNA double-strand break repair pathway choice. *Nat. Commun.* **5**, (2014).

44. 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).

45. Hinch, A. G. *et al.* Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science* **363**, eaau8861 (2019).

46. Centola, M. & Carbon, J. Cloning and Characterization of Centromeric DNA from Neurospora crassa. *Mol. Cell. Biol.* **14**, 1510–1519 (1994).

47. Puechberty, J. *et al.* Genetic and physical analyses of the centromeric and pericentromeric regions of human chromosome 5: Recombination across 5cen. *Genomics* **56**, 274–287 (1999).

48. Mahtani, M. M. & Willard, H. F. Physical and genetic mapping of the human X chromosome centromere: Repression of recombination. *Genome Res.* **8**, 100–110 (1998).

49. Vincenten, N. *et al.* The kinetochore prevents centromere-proximal crossover recombination during meiosis. *Elife* **4**, (2015).

50. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).

51. Chan, K. & Gordenin, D. A. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annu. Rev. Genet.* **49**, 243–267 (2015).

52. Pratto, F. *et al.* Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell* (2021). doi:10.1016/J.CELL.2021.06.025

53. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72–77 (2017).

54. Goldmann, J. M. *et al.* Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet. 2018 504* **50**, 487–492 (2018).

55. Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nat. Genet.* **36**, 1203–1206 (2004).

56. Martin, H. C. *et al.* Multicohort analysis of the maternal age effect on recombination. *Nat. Commun.* **6**, (2015).

57. Li, R. *et al.* A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nat. Commun.* **10**, 1–15 (2019).

58. 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**, 1721–33 (2015).

59. de Boer, E., Stam, P., Dietrich, A. J. J., Pastink, A. & Heyting, C. Two levels of interference in mouse meiotic recombination. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9607–12 (2006).

60. Wang, H. & Xu, X. Microhomology-mediated end joining: New players join the team. *Cell and Bioscience* **7**, 6 (2017).

61. Wartosch, L. *et al.* Origins and mechanisms leading to aneuploidy in human eggs. *Prenat. Diagn.* **41**, 620–630 (2021).

62. Steinthorsdottir, V. *et al.* Variant in the synaptonemal complex protein SYCE2 associates with pregnancy loss through effect on recombination. *Nat. Struct. Mol. Biol.* (2024). doi:10.1038/s41594-023-01209-y

63. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).

64. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).

65. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

66. Cheung, V. G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).

67. Furey, T. S. & Haussler, D. Integration of the cytogenetic map with the draft human genome sequence. *Hum. Mol. Genet.* **12**, 1037–44 (2003).

68. R Core Team. R: A Language and Environment for Statistical Computing. (2017).

69. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw.* **82**, (2017).

70. Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with python. in *9th Python in Science Conference* (2010).

71. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).

72. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).

73. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).

74. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

75. McMahill, M. S., Sham, C. W. & Bishop, D. K. Synthesis-dependent strand annealing in meiosis. *PLoS Biol.* **5**, 2589–2601 (2007).

76. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).

77. Sveinbjornsson, G. *et al.* Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).

78. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

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

80. Housworth, E. A. & Stahl, F. W. Crossover interference in humans. *Am. J. Hum. Genet.* **73**, 188–97 (2003).

81. Broman, K. W. & Weber, J. L. Characterization of human crossover interference. *Am. J. Hum. Genet.* **66**, 1911–1926 (2000).

82. Wang, S. *et al.* Inefficient Crossover Maturation Underlies Elevated Aneuploidy in Human Female Meiosis. *Cell* **168**, (2017).

83. Broman, K. W. xoi: Tools for Analyzing Crossover Interference. R package version 0.72. (2012).

84. Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D. & Auton, A. Escape from crossover interference increases with maternal age. *Nat. Commun.* **6**, 1–7 (2015).

85. Döring, A., Weese, D., Rausch, T. & Reinert, K. SeqAn An efficient, generic C++ library for sequence analysis. *BMC Bioinformatics* **9**, 11 (2008).

86. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**, 3–36 (2011).

87. Jonsson, H. *et al.* Differences between germline genomes of monozygotic twins. *Nat. Genet.* **53**, 27–34 (2021).

88. Gelman, A. *et al.* Bayesian Data Analysis. (2013).

89. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol. 2020 3811* **38**, 1347–1355 (2020).

90. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* (2020). doi:10.1038/s41586-020-2287-8

91. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* (2019). doi:10.1016/j.cell.2018.12.019

92. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 1–26 (2016).

93. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).

Figures

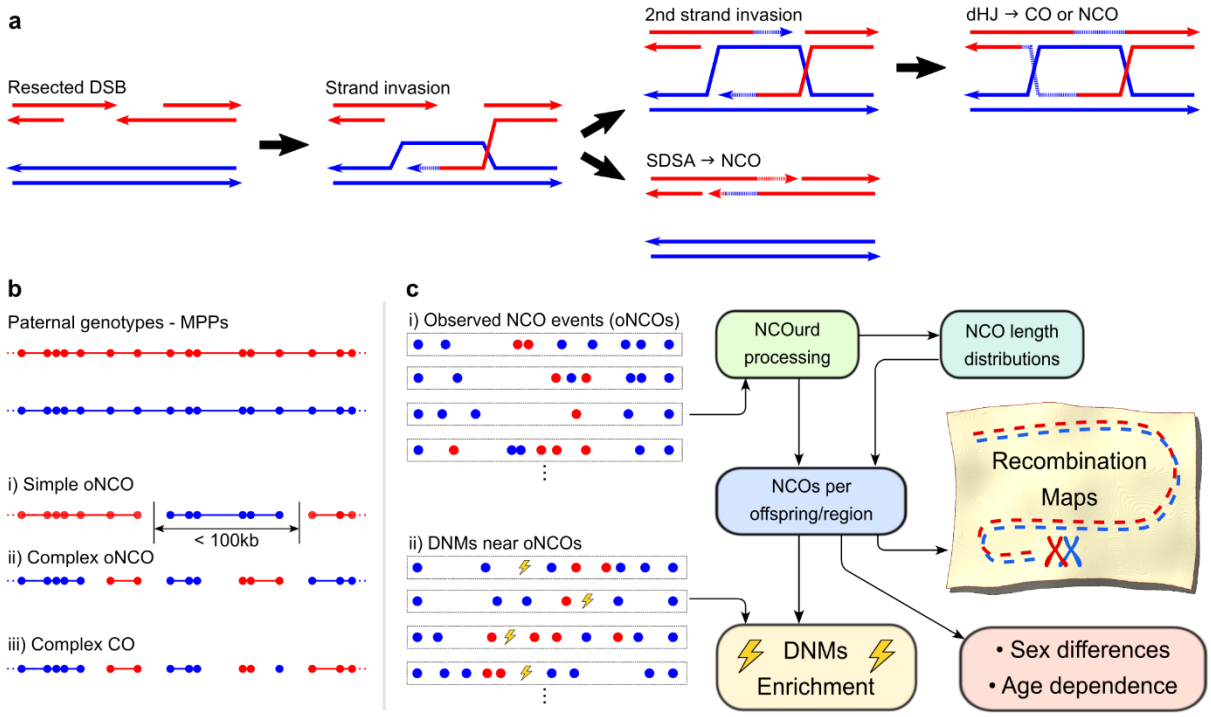


Figure 1: Meiosis, NCOs and data analysis: a| Schematic view of NCO/CO resolution. *A DSB is induced on one chromosome (red) and the 5‘ strands near the DSB are resected. The 3‘ strands invade the homologous chromosome (blue) and DNA is synthesized (dotted lines) to bridge the DSB. When only one strand invades, the synthesis dependent strand annealing (SDSA) pathway is used, leading to NCOs. When both strands invade, a double Holliday junction (dHJ) is generated, which are the primary source of COs.* ***b| Schematic view of recombination events.*** *The points denote MPPs in a meiosis with the color indicating the grandparental origin of each MPP. Short haplotype segments are gene conversion candidates flanked by background haplotypes forming i) a simple oNCO with a single converted segment or ii) a complex oNCO with alternating gene-converted and non-gene-converted segments, if background haplotypes are of same grandparental origin, and otherwise iii) CO with associated gene conversions.* ***c| Schematic overview of the NCOurd process and subsequent analysis.*** *i) oNCOs are specified by a set of gene-converted MPPs (red) and the surrounding background haplotype MPPs (blue). Our previously described method[23] (NCOurd) derives length distributions for NCOs from the oNCOs. These are used to compute the number of NCOs per offspring/region. NCOs per offspring allows us to explore sex differences and age dependence of the meiotic process, as well as interaction with DNMs estimated from ii) DNMs found near oNCOs. NCOs per region are used to compute the number of NCOs throughout the genome to create maps of NCO activity and DSB resolution.*

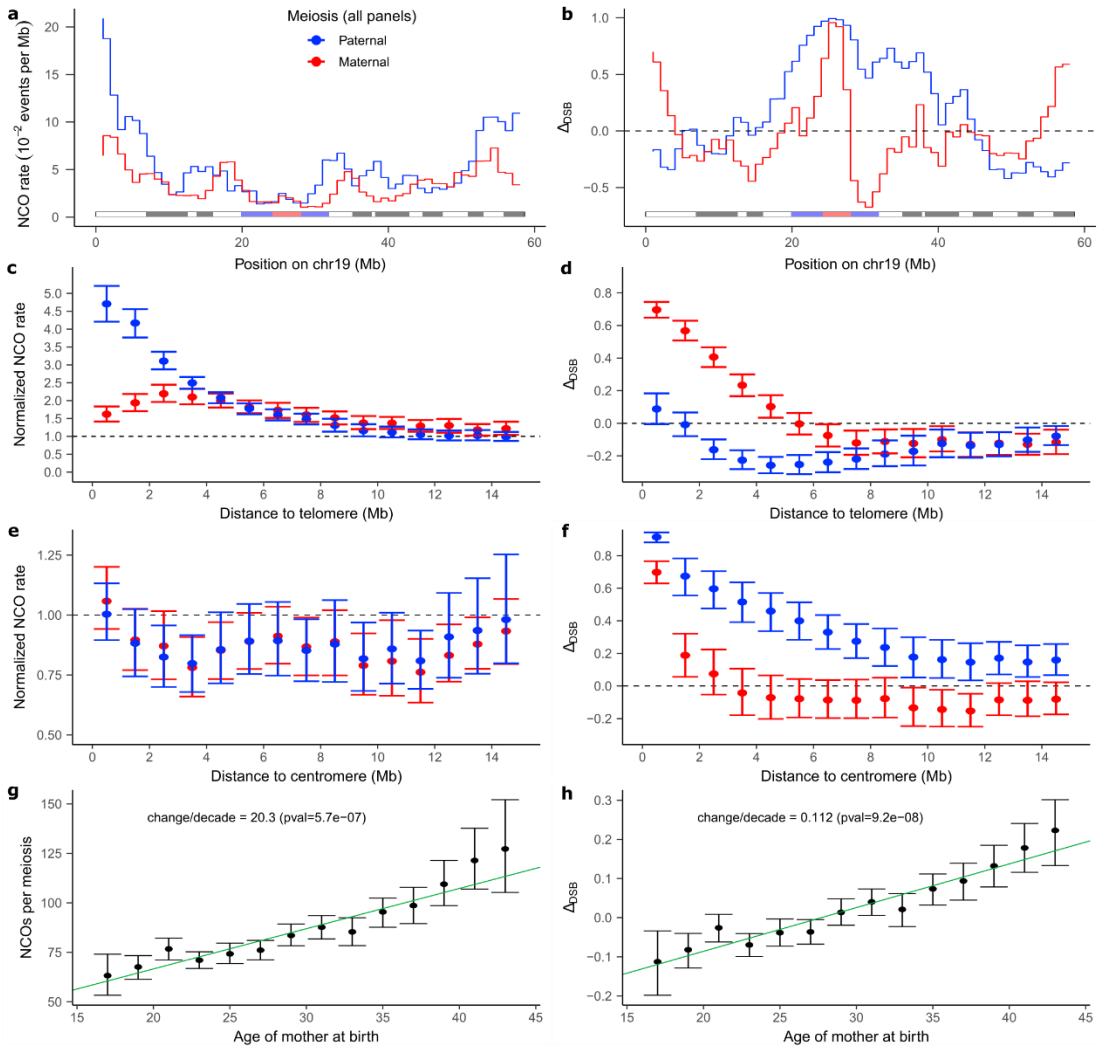


Figure 2: Recombination map and maternal age effects. a| *NCO maps for chr19.* ***b|*** *Δ_{DSB} measure for chr19. Cytobands are shown below the graphs with the centromere indicated in red, gneg bands in white, all gpos bands in gray, and gvar and stalk bands in blue.* ***c|*** *and* ***d|*** *Average values of the NCO and Δ_{DSB} recombination maps near telomeres, with the NCO data normalized to the autosomal average. Error bars show 95% confidence intervals computed by bootstrapping 1000 samples based on map data for the 22 autosomes.* ***e|*** *and* ***f|*** *Same as c/d for map values near centromeres.* ***g|*** *and* ***h|*** *Results for per-offspring NCO count and Δ_{DSB} of maternal meioses vs. maternal age. Offspring are grouped by maternal age in bins of size 2 years; the points show group averages, omitting bins with fewer than 25 offspring. Error bars show 95% confidence intervals computed by bootstrapping 1000 samples from the 5240 probands. The green lines show linear regression results using the inverse of the size of the confidence intervals as weight. P-values for regression results are based on Student's t-distribution.*

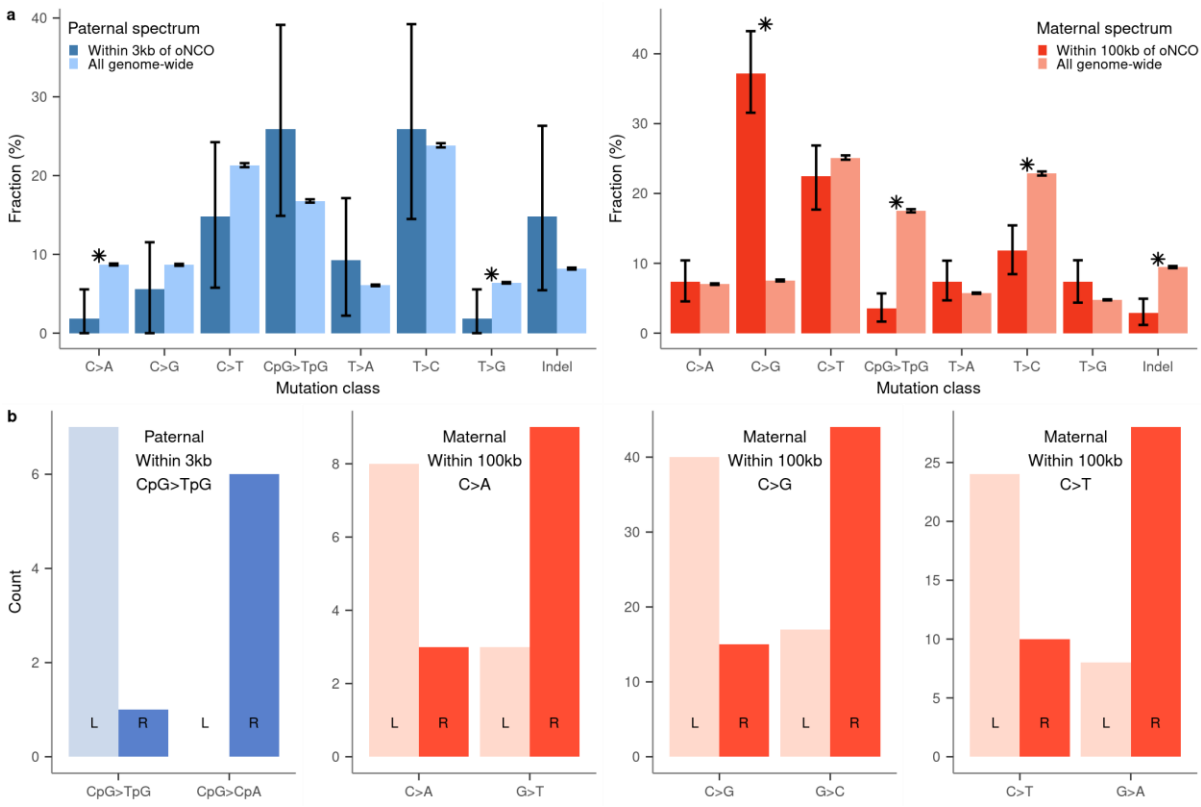


Figure 3: Mutation spectra. a| Mutation spectra for phased DNMs proximal to oNCOs and genome-wide. DNMs are considered proximal to oNCOs if they are within 3kb and 100kb for paternally and maternally phased DNMs, respectively. The length of the bars indicates the mutation class fraction for the complete cohort of the study. Error bars show 95% confidence intervals computed by bootstrapping 1000 samples and asterisks indicate mutation classes where the NCO-proximal and genomewide spectra are significantly different ($P < 0.05$, bootstrap test). **b| Strand asymmetry for phased DNMs around oNCOs.** L and R denote DNMs to the left and right of the oNCO center, respectively.

Methods

Data

To call gene conversions we used whole genome sequence data for parents and children in 2,132 Icelandic families with two or more children, comprising a total of 10,840 meioses. The dataset is made up of Icelandic samples collected as part of disease association efforts at deCODE genetics consisting of 173,025 SNP chip-typed individuals of which a subset of 63,118 are whole genome sequenced[63,64] (Supplementary Note 0). All participants were Icelanders who signed an informed consent form and donated biological samples for genotyping as part of various research projects at deCODE genetics approved by the National Bioethics Committee (NBC) in Iceland. Processing of data was performed in agreement with the approvals issued by the National Bioethics Committee (NBC) and conditions set by the Data Protection Authority (DPA) (ref. PV_2017060950ÞS) on procedures to ensure security in the processing of personal data for the scientific research within the health sector conducted by deCODE genetics and the Act on Scientific Research in the Health Sector No. 44/2014.

NCO calls

Gene conversions are detected in offspring by phasing the genotypes of both children and parents on a curated set of SNPs and indel variants. Phasing can be done on informative variants (MPPs), i.e., those that are heterozygous in one parent and homozygous for the other. The phasing is trivial for the offspring as one of parents is homozygous. The parents are then phased against the set of children. The phasing allows us to assign a grand-parental origin to haplotypes in the offspring and thus detect where the grandparental origin of haplotypes changes, indicating either COs or NCOs. Haplotypes which span less than 100kb between such changes are regarded as gene conversion candidates and may be grouped into an observed NCO (oNCO) if they are consecutive and do not form part of a crossover (Supplementary Note 4).

Quality checks

The phasing of the parent at variants where a candidate gene conversion transmission has been detected in offspring is verified with two approaches. Phasing may be confirmed if variants of alternate grandparental origin in the offspring are close enough to that they may be

observed together on a sequence read for the parent. If the variants are further apart, we use the Icelandic genealogical database to find one or more relatives that share, IBD with the parent, large haplotypes containing the variants to be checked. If those relatives happen to be also homozygous at those variants, we can use the relationship of the parent and the relative to establish whether the variant is paternally or maternally inherited (Supplementary Note 0).

We also benchmark our methodology by comparison with earlier methods[24] that use three generation family structures with three or more siblings (Supplementary Note 0). All summary statistics are comparable (Table S13, Table S14) with no unexplained discrepancies (Table S15).

Finally, to check whether family size impacts our results, we verify that all summary statistics are comparable between families with only two children and for larger families (Table S16).

Length distributions, NCOs and DSBs

The length distributions of the NCOs (observed or not) are modeled as mixtures of negative binomial distributions using the NCOurd approach[23] (Supplementary Note 0). NCOurd models the mismatch repair of ssDNA at the site of NCOs and requires the complete set of oNCO along with the set of MPPs for each proband. The set of MPPs is used to compute a tract function for each oNCO, and a detection probability function, which depends on the length. The mixture components of the resulting length distribution are separated into two groups: *short* and *extended,* having mean length under and over 1kb, respectively.

The NCOurd results and the detection probability function can be used to derive the expected fraction of NCOs detected as oNCOs. Using this detection fraction and the total number of oNCOs we can estimate the average number of NCOs (N_{NCO}) per offspring and then the average number of DSBs (N_{DSB}) per meiocyte is estimated from N_{NCO} and the average number of COs (N_{CO}) per offspring as (Supplementary Note 0):

$$N_{DSB} = 4 \cdot N_{NCO} + 2 \cdot N_{CO}$$

Statistics, confidence intervals, and p-values

Unless otherwise specified, all statistics pertain to the data for the full cohort (possibly in groups) and the length distributions that are computed with the full cohort. Reported averages are computed using the full cohort with confidence intervals computed by bootstrapping from the set of parents (Supplementary Note 0). Thus, quantities are computed per parent based on

the data for the children and then statistics are computed for 1000 bootstrap samples from the set of parents. The lower and upper limits of the confidence interval correspond to the 2.5 and 97.5 percentile of the bootstrap statistics. The same bootstrap approach is used for computing matched pairs for estimating P-values for comparison of two datasets (Supplementary Note 0).

The exception to the approach above is in the estimates of confidence intervals for length distribution parameters, and the number of NCOs and NCO derived quantities. Here we computed 1000 different length distributions by sampling data from the set of parents as described above. Reported values correspond to a computation done with the full cohort, and the confidence intervals are computed based on the bootstrap estimates as outlined above.

All computed p-values are two-sided.

NCO Maps

NCO maps are computed on a grid of overlapping 3Mb windows at intervals of 1Mb (Supplementary Note 0). Self-consistent distributions for the number of paternal and maternal NCOs per window are derived based on the number of oNCOs falling within the window and the expected NCO detection fraction in the window. The expected number of NCOs per window can then be obtained from those distributions.

DSB maps can be created from the NCO and CO maps, with the DSB map value (n_{DSB}) given in terms of the NCO map value (n_{NCO}) and the CO map value (n_{CO}) as:

$$n_{DSB} = 4 \cdot n_{NCO} + 2 \cdot n_{CO}$$

We explore DSB resolution with a normalized NCO/CO difference map, Δ_{DSB}, given by:

$$\Delta_{DSB} = \frac{\frac{n_{NCO}}{N_{NCO}} - \frac{n_{CO}}{N_{CO}}}{\frac{n_{NCO}}{N_{NCO}} + \frac{n_{CO}}{N_{CO}}} .$$

Here N_{NCO} and N_{CO} indicate the average number of NCOs and COs per offspring, respectively. We note that Δ_{DSB} takes the values 1 and -1 if all DSBs within a given map window are resolved as NCOs and COs, respectively, and 0 if the resolution conforms to the genome-wide average.

Analysis of telomere/centromere distances

Distances are measured from the center of each map window. Distance to telomeres are measured to the ends of the chromosomes as given by GRCh38[65]. Distances to centromeres are measured to the edge of centromere[66,67]. If the window center is within the centromere the distance is regarded as zero. Distances are binned in bins of size 1Mb and the x-coordinate of each point in the graph is the center of the corresponding bin.

DNMs

Mutagenicity of NCOs is explored by analyzing the enrichment of DNMs near oNCOs (Supplementary Note 0). I.e., we compare the observed number of DNMs in bins at different distances from oNCOs to the expected number of DNMs based on the normative DNM rate. The normative DNMs rate is computed per proband based on the age of the parent at the birth of the proband. A Bayesian approach using a Beta distribution prior is employed to estimate the paternal-maternal split of unphased DNMs. We analyzed whether the rate enrichment might be due to sequence context by computing the enrichment after permuting the DNMs among probands. No enrichment was found after permutation ($P>0.2$), indicating that the observed enrichment is due to the mutagenicity of NCOs. The enrichment analysis is performed separately for each component of the NCO length distributions, allowing us to compute the final enrichment of DNMs near NCOs, weighted appropriately for NCOs rather than oNCOs. Finally, we can estimate the contribution of NCOs to the overall number of DNMs by multiplying the total number of NCOs and the enrichment of the DNMs rate near NCOs.

Comparison of DNMs spectra is computed with a two-sided χ^2 test. The χ^2 statistic and p-value is computed for the difference between the oNCO-proximal spectrum and the genomewide spectrum for the probands in the study. The null distribution for the χ^2 statistic is simulated by sampling one million times from the complete set of DNMs for the probands in the study. We sample the same number of DNMs as in the oNCO-proximal spectrum, using data from the same number of DNM transmitting parents as in that spectrum. For fathers, we find 74950 events with χ^2 statistic larger than the one computed for the oNCO-proximal spectrum, giving a p-value of 0.075. For mothers, no events are found with larger χ^2 statistic than the one for the oNCO-proximal spectrum. For the maternal distribution we compute an inflation factor as the number of degrees of freedom (7) divided by the average of the χ^2

statistic across the 1M simulations (6.96). As this factor (7/6.96 = 1.006) is greater than 1.0 we quote the unadjusted p-value for the difference between the spectra.

Odds ratios (with p-values and confidence intervals) for the strand asymmetry of oNCO-proximal DNMs are calculated using the fisher.test function of R[68]. For each class we omit DNMs transmitted by parents to two or more distinct children in the class. For the reported classes this only affected the maternal C>T class, where two DNMs were omitted from the calculation and had a minimal effect on the p-value and odds-ratio.

To determine whether the fraction of oNCO-proximal DNMs within a mutation class differs significantly from the genome-wide fraction shown in **Figure 1** we use bootstrapping (Supplementary Note 0) to compute an odds-ratio and 2-sided test to determine if the bootstrap values differ significantly from 1.

Age effects

We perform linear regressions to analyze how parental age at birth of offspring affect various recombination statistics. For statistics that pertain to oNCOs and COs, where we have an observed statistic per proband we compute the linear regression directly on that data. For estimated statistics, such as the number of NCOs and DSBs, we split the cohort into groups based on parental age and compute the average statistic for each group. Confidence intervals are estimated for each group using bootstrapping as described above. The linear regression is then performed on the average, weighted with the inverse of the estimated confidence interval.

Mixed model linear regression is employed for the analysis of the DMC1 and maternal CO hotspot annotations of oNCOs. This is done with the lmerTest[69] package in R[68]. Here we can compute the statistic for each proband and perform the regression on the complete set of data.

Confidence intervals for linear regression results are computed based on Student's t-distribution within the packages used for computation[69,70].

Acknowledgements

We thank the study participants.

Author Contributions

Paper was written by GP and BVH with input from MT, HK, VS, OAS, HP, UT, PS, AH, DFG and KS. Analysis of gene conversions and NCOs was performed by GP and MT. DNM analysis was performed by HJ. Annotation analysis was performed by OAS and SAG. Genotyping was performed by HPE. Imputation was performed by PIO and AG, supervised by GM. Statistical analysis was performed by GP, MTH and BVH, supervised by BVH and DFG. Study was supervised by BVH and KS. All authors agreed to the final version of the manuscript.

Competing interests

All authors are employees of deCODE genetics/Amgen.

Additional information

Address correspondence to: Bjarni V. Halldorsson, deCODE genetics / Amgen Inc., Sturlugata 8, 102 Reykjavik, Iceland. bjarnih@decode.is, Phone: 354-5701808, fax 354-5701901

Kari Stefansson, deCODE genetics / Amgen Inc., Sturlugata 8, 102 Reykjavik, Iceland. kstefans@decode.is, Phone:354-5701900, fax 354-5701901.

Data availability

Recombination maps, proband information, gene conversions, oNCOs and DNMs are available at https://doi.org/10.5281/zenodo.14025565.

GRCh38, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/.

GIAB WGS samples https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/.

Beyter et al. SV data https://github.com/DecodeGenetics/LRS_SV_sets.

Refseq https://www.ncbi.nlm.nih.gov/refseq/.

Eichler SV calls are shared on dbVar (https://www.ncbi.nlm.nih.gov/dbvar) under accession dbVar:nstd162.

DMC1 hotspots
https://www.science.org/doi/suppl/10.1126/science.aau1043/suppl_file/aau1043_datas2.gz

Altemose et. al. PRDM9 data:

https://ftp.ncbi.nlm.nih.gov/geo/series/GSE99nnn/GSE99407/suppl/GSE99407%5FChIPseq%5FPeaks.YFP%5FHumanPRDM9.antiGFP.protocolN.p10e%2D5.sep250.Annotated.txt.gz

Altemose et. al. H3K4me3 data:

https://ftp.ncbi.nlm.nih.gov/geo/series/GSE99nnn/GSE99407/suppl/GSE99407%5FChIPseq%5FPeaks.YFP%5FHumanPRDM9.antiH3K4me3.protocolN.p10e%2D5.sep250.txt.gz

Gnomad SVs http://gnomad-sg.org/downloads.

CO recombination data
https://www.science.org/doi/suppl/10.1126/science.adh2531/suppl_file/science.adh2531_data_s1.zip

https://www.science.org/doi/suppl/10.1126/science.adh2531/suppl_file/science.adh2531_data_s2.zip

The raw sequence data and the Icelandic genealogical database cannot be made publicly available because Icelandic law and the regulations of the Icelandic Data Protection Authority prohibit the release of individual-level and personally identifying data. Data access for raw data can be granted for scientific purposes only at the facilities of deCODE genetics in Iceland, subject to Icelandic law regarding data usage. Anyone wishing to gain access to the data should contact B.V.H. (bjarni.halldorsson@decode.is) or K.S. (kstefans@decode.is). Requests for access are generally considered monthly.

Summary statistics for GWAS studies are available at https://www.decode.com/summarydata/

Code availability

We used publicly available software (URLs are listed below) in conjunction with the above-described algorithms. GraphTyper (v2.7.1), https://github.com/DecodeGenetics/graphtyper. NCOurd, https://github.com/DecodeGenetics/NCOurd. R (v4.2.2 with lm v4.2.2, xoi v0.67-1), https://www.r-project.org/. Python (v3.8.1 with numpy v1.24.2, pandas v1.4.0, scipy v1.10.1, statsmodels v0.13.2), https://www.python.org/downloads/.

Extended Data Table 1 | Cohort and recombination data. The total number of meioses and MPPs in the study, along with the number of oNCOs and gene-converted markers. Also shown are the number of oNCOs and NCOs, split into short and extended events. 95% confidence intervals are shown in parentheses where applicable.

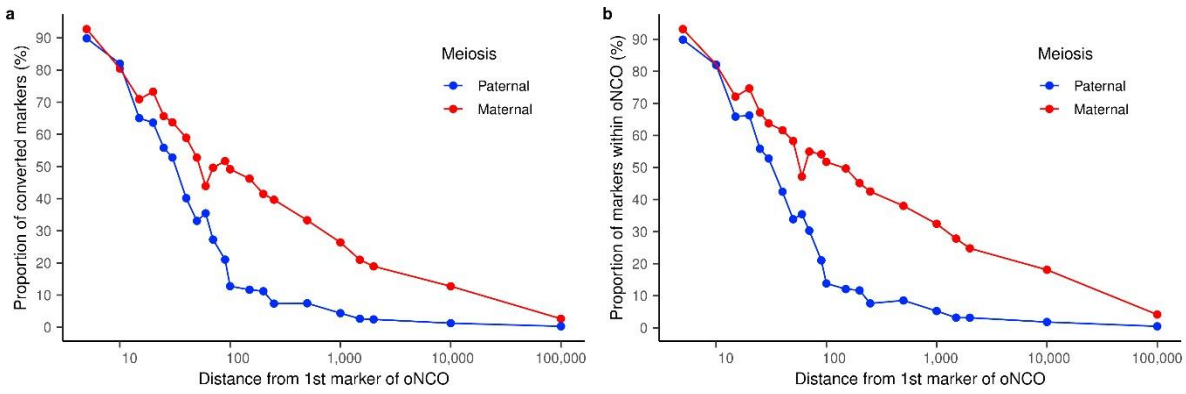
	Paternal	Maternal
Number of offspring	5,420	5,420
Total MPPs	4,229,340,533	4,288,524,590
Observed NCOs (oNCOs)	12,948	15,712
Gene-converted sequence variants	17,109	45,653
SNPs	15,955	43,108
Indels	1,154	2,545
Gene conversion rate (per million MPPs)	4.05 (3.83-4.32)	10.6 (10.2-11.2)
Gene-converted variants per oNCO	1.32 (1.26-1.40)	2.91 (2.79-3.03)
oNCOs per offspring	2.39 (2.33-2.44)	2.90 (2.83-2.97)
Short oNCOs	2.05 (1.69-2.14)	1.21 (1.00-1.31)
Extended oNCOs	0.33 (0.27-0.69)	1.69 (1.60-1.89)
# NCOs per offspring	105.0 (95.9, 125.0)	81.6 (66.7, 103.1)
Short NCOs	103.8 (94.2, 123.7)	76.0 (61.4, 97.6)
Extended NCOs	1.2 (0.7, 6.4)	5.6 (4.7, 7.7)
Proportion of NCOs		
Short NCOs (%)	98.86 (94.12, 99.38)	93.11 (90.07, 94.90)
Extended NCOs (%)	1.14 (0.62, 5.88)	6.89 (5.10, 9.93)
Proportion of oNCOs		
Short oNCOs (%)	86.01 (69.30, 88.85)	41.83 (33.86, 44.98)
Extended oNCOs (%)	13.99 (11.15, 30.70)	58.17 (55.02, 66.14)
Average length of NCOs		
Short NCOs (bp)	123 (94, 135)	102 (71, 125)
Extended NCOs (kb)	7.2 (1.8, 11.8)	9.1 (6.8, 10.8)
# COs per offspring	26.8 (26.7, 26.9)	41.7 (41.5, 42.0)
# DSBs per meiocyte	474 (438, 554)	410 (350, 496)
NCO/CO ratio per meiocyte	7.84 (7.16, 9.33)	3.91 (3.20, 4.94)
Allele selection biases - gene-converted variants		
SNPs: GC bias (%)	61.9 (60.6-63.3)	65.5 (64.7-66.2)
Indels: Insertion bias (%)	44.9 (42.1-47.8)	62.5 (60.4-64.4)
Allele selection biases - non-gene-converted variants		
SNPs: GC bias (%)	47.6 (44.2-51.9)	68.0 (66.6-69.5)
Indels: Insertion bias (%)	48.8 (40.3-58.6)	74.6 (70.9-77.9)

Extended Data Table 2 | Relative rates of oNCOs and COs[1] in annotated regions of the genome. The CO hotspots used are from sex-specific genetic maps[1]. The ChromHMM annotation[40] for paternal and maternal meiosis is a consensus annotation (Supplementary note 0) for several samples measured in testis and ovary, respectively.

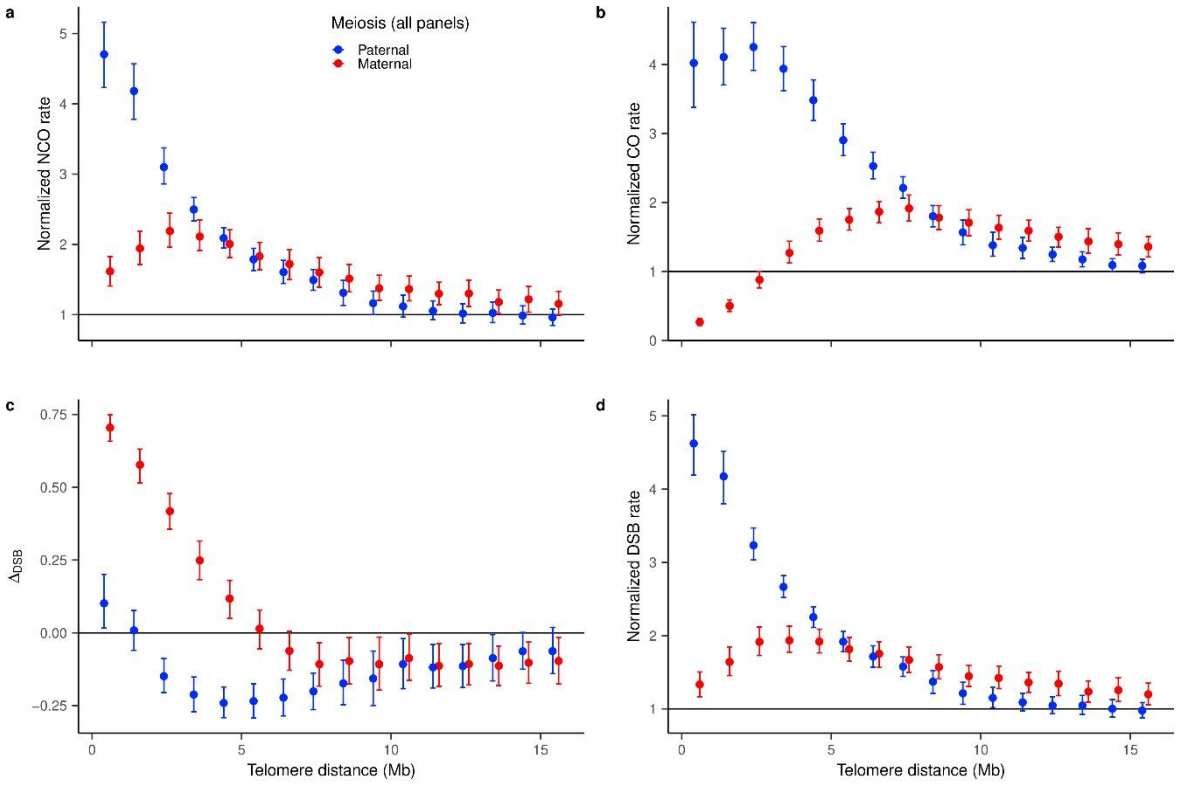
	Paternal		Maternal	
Annotation	**oNCOs**	**COs**	**oNCOs**	**COs**
deCODE CO hotspots[1]	22.4 (21.9,22.8)	50.03 (49.98,50.08)	13.7 (13.4,14)	43.07 (43.03,43.12)
DMC1 hotspots[10,11]	42.4 (41.6,43.2)	54.62 (54.49,54.74)	19.3 (18.8,20)	38.81 (38.73,38.90)
Altemose PRDM9 ±500 bp[39]	7.47 (7.31,7.64)	7.29 (7.27,7.31)	4.95 (4.82,5.09)	7.12 (7.11,7.14)
Altemose H3K4me3 ±500 bp[39]	3.90 (3.83,3.98)	3.93 (3.92,3.94)	2.96 (2.89,3.02)	4.00 (3.99,4.00)
C>G enriched regions[9]	0.96 (0.92,1.01)	1.47 (1.46,1.48)	1.54 (1.48,1.60)	1.16 (1.15,1.16)
Refseq Exons[71]	1.69 (1.43,1.97)	0.91 (0.90,0.92)	1.62 (1.38,1.89)	0.96 (0.95,0.97)
ChromHMM[40] Strongly Transcribed (Tx)	1.52 (1.38,1.67)	0.37 (0.36,0.37)	1.26 (1.17,1.36)	0.51 (0.50,0.51)

Extended Data Table 3 | Enrichment of DNMs near NCO and COs and the estimated contribution of recombination to the total DNMs rate. The contribution to the DNMs rate is computed from the elevation and the estimated numbers of NCOs and COs per offspring. For COs we use the enrichment estimates from a prior publication[1] and the CO count of the current cohort.

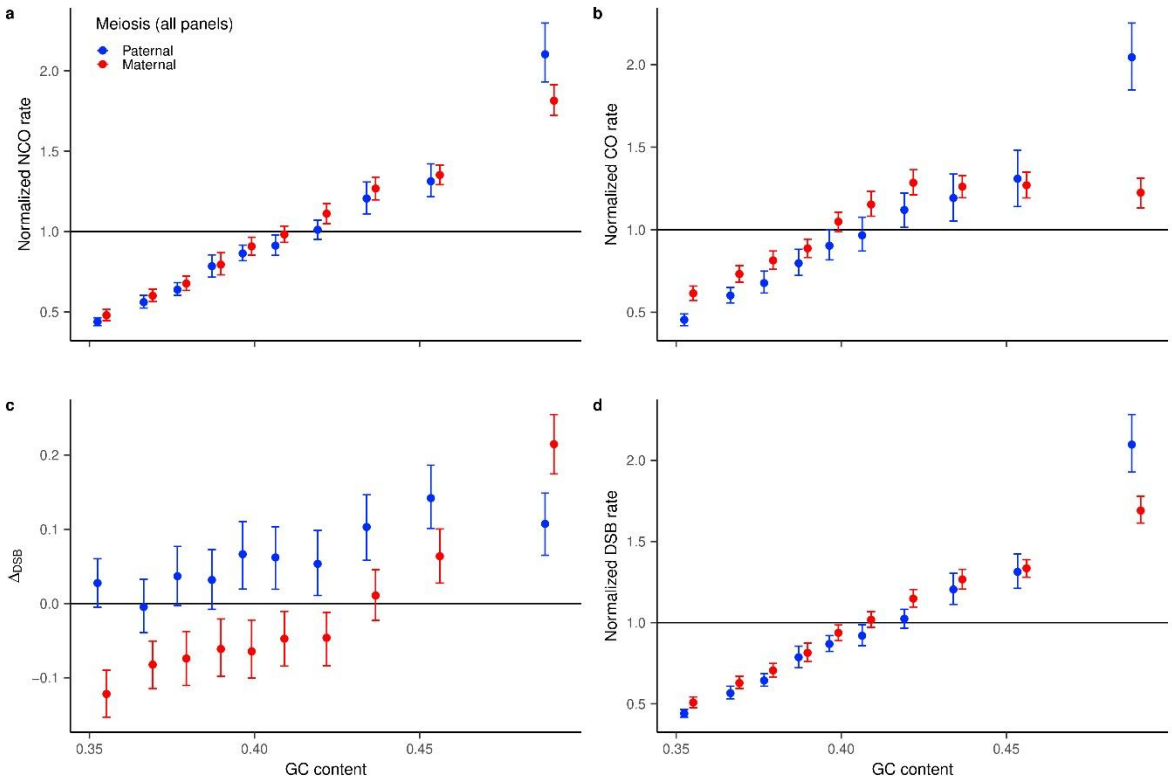
	Paternal	Maternal	
	All genome	**All genome**	**In C>G enriched regions**
DNMs enrichment near NCOs – full cohort			
Enrichment within 0 to 1 kb	142.3 (105.8, 183.2)	125.5 (65.7, 197.0)	119.1 (35.4, 228.7)
Enrichment within 1 to 3 kb	10.4 (3.7, 19.1)	48.2 (21.7, 79.0)	84.8 (31.0, 151.3)
Enrichment within 3 to 40 kb	1.9 (1.2, 2.7)	37.0 (29.3, 44.9)	91.4 (72.7, 111.8)
Enrichment within 40 to 100 kb	1.3 (0.9, 1.7)	4.9 (3.3, 6.8)	12.4 (7.7, 18.0)
DNMs enrichment near NCOs – maternal at age = 20 years (17.5 to 22.5)			
Enrichment within 0 to 1 kb		132.4 (5.6, 339.5)	47.5 (0.0, 142.9)
Enrichment within 1 to 3 kb		14.7 (0.0, 49.6)	47.9 (0.0, 151.0)
Enrichment within 3 to 40 kb		6.2 (1.6, 13.2)	37.2 (6.0, 73.8)
Enrichment within 40 to 100 kb		0.6 (0.2, 0.7)	1.0 (0.3, 1.6)
DNMs enrichment near NCOs – maternal at age = 40 years (37.5 to 42.5)			
Enrichment within 0 to 1 kb		476.8 (128.5, 919.2)	253.1 (0.0, 644.0)
Enrichment within 1 to 3 kb		7.3 (0.0, 18.4)	9.0 (0.0, 25.7)
Enrichment within 3 to 40 kb		72.7 (40.6, 111.9)	105.0 (51.4, 166.8)
Enrichment within 40 to 100 kb		12.8 (4.5, 23.6)	21.2 (5.1, 40.2)
Contribution to DNMs rates			
From NCOs	1.69% (1.22%-2.15%)	10.95% (8.74%-13.03%)	38.39% (31.12%-45.99%)
From COs	0.11% (0.07%-0.16%)	0.38% (0.24%-0.59%)	0.44% (0.28%-0.67%)
Total due to DSBs	1.80% (1.29%-2.31%)	11.3% (9.0%-13.6%)	38.8% (31.4%-46.7%)
Maternal age = 20 years		2.0% (0.6%-3.9%)	8.2% (1.9%-15.8%)
Maternal age = 40 years		31.0% (18.6%-44.8%)	80.4% (45.0%-122.0%)



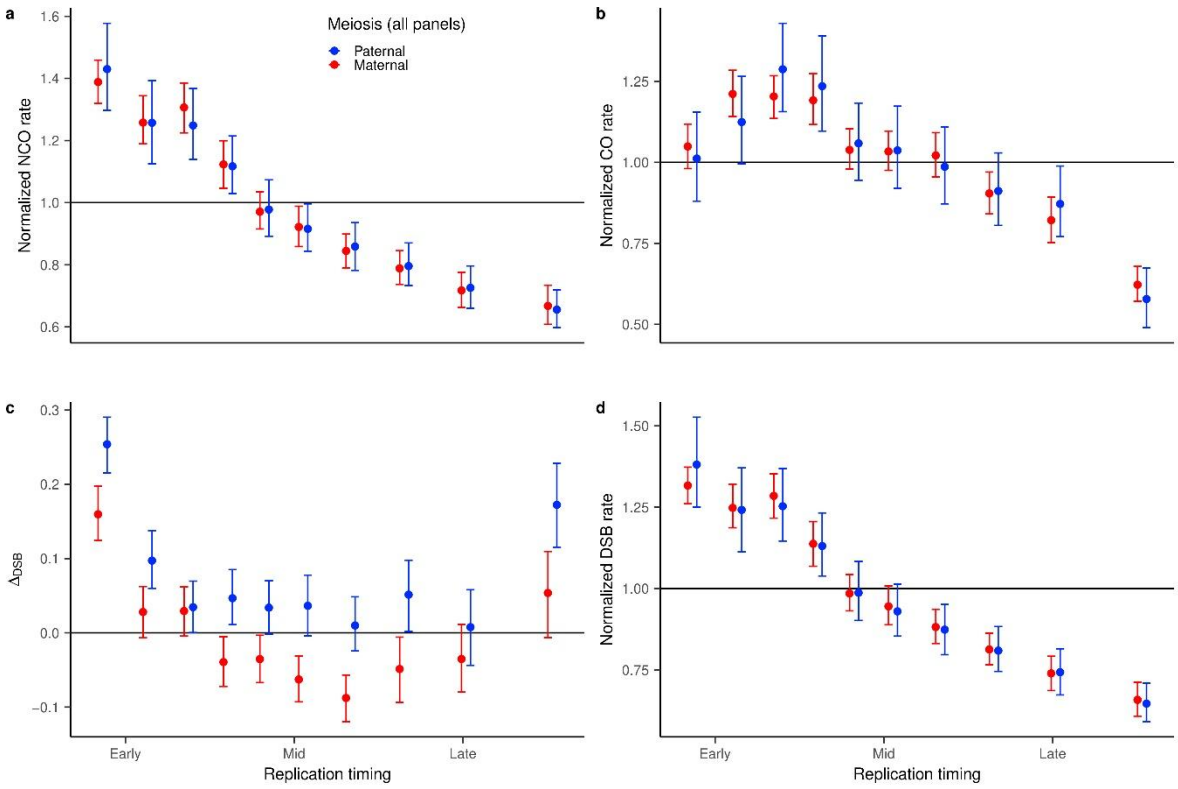
Extended Data Fig. 1 | Co-conversion probability and oNCO extent.
a| Proportion of gene converted markers as function of distance from the first marker of oNCOs. b| Proportion of markers that are within the oNCO (gene converted or not) as a function of distance from the first marker of oNCOs. The fraction of markers within oNCOs for a given distance provides a lower bound on fraction of oNCOs shorter than that distance.



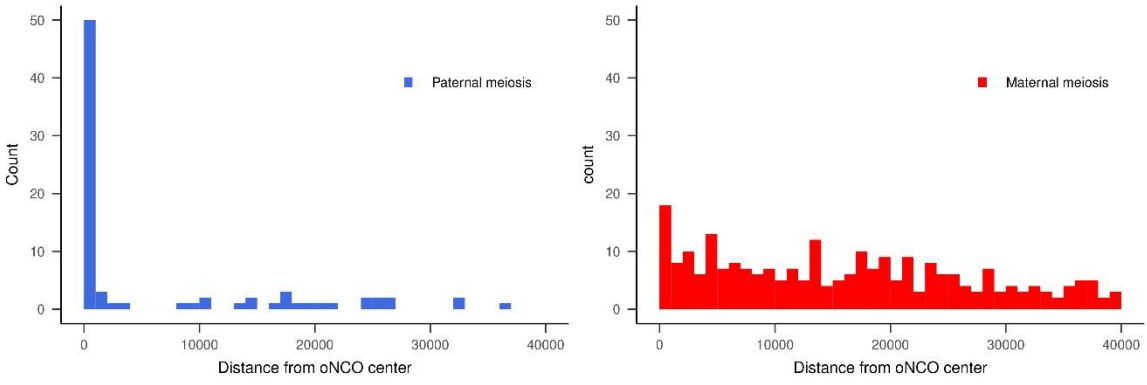
Extended Data Fig. 2 | Variation of recombination parameters with distance to nearest telomere.
The figure shows the average map values vs. distance to the nearest telomere, computed on a grid of 3Mb overlapping windows and normalized individually with respect to their genome-wide mean. The x-coordinates for the telomere distance are shifted slightly in opposite directions for paternal/maternal meiosis so that error bars don't overlap for nearby points. The error bars indicate 95% confidence intervals and are computed by bootstrap sampling 1000 times from the set of windows within each distance bin.



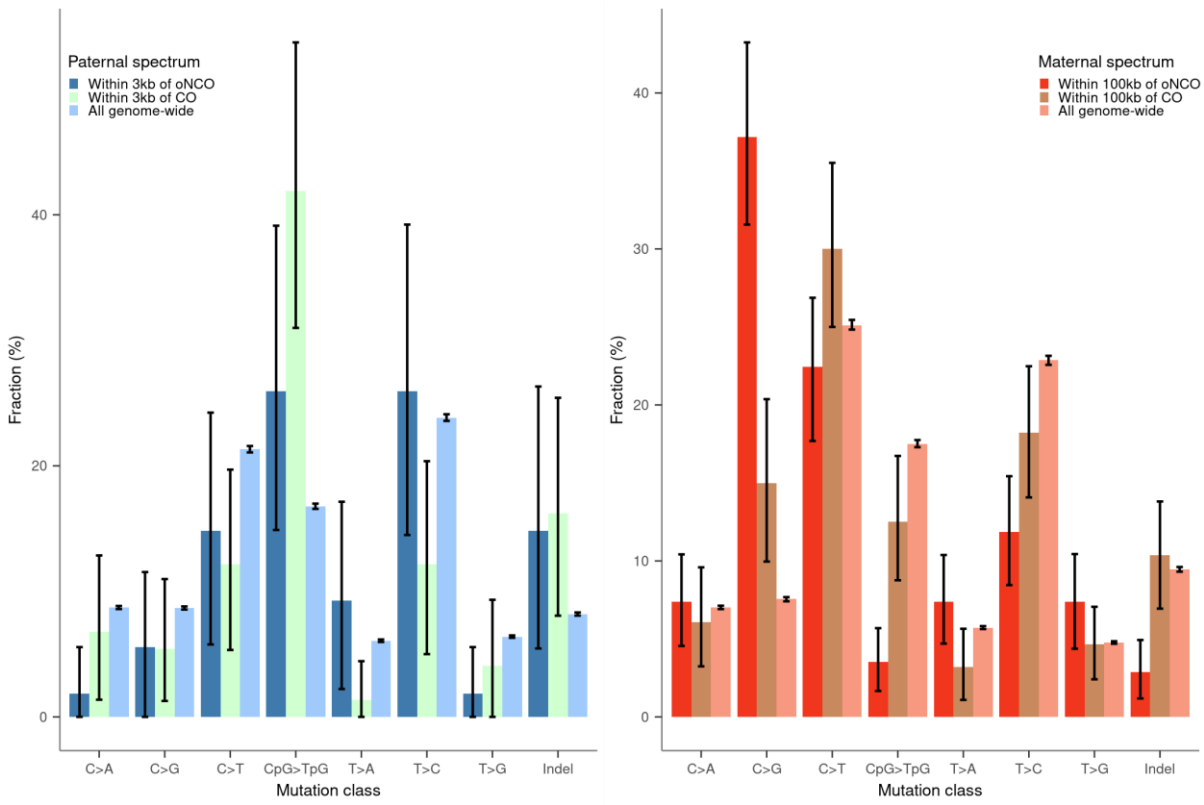
Extended Data Fig. 3 | Variation of recombination maps with GC content.
The GC content is computed on the same overlapping 3Mb windows as the maps, split into deciles and the mean map values computed for the windows that fall into each decile. All maps except Δ_{DSB} are normalized individually with respect to their genome-wide mean. The x-coordinate in the figures shows the median GC content in each decile, shifted slightly in opposite directions for paternal/maternal meioses so that errors bars don't overlap for nearby points. The error bars indicate 95% confidence intervals and are computed by bootstrap sampling the map values 1000 times from the set of windows within each decile.



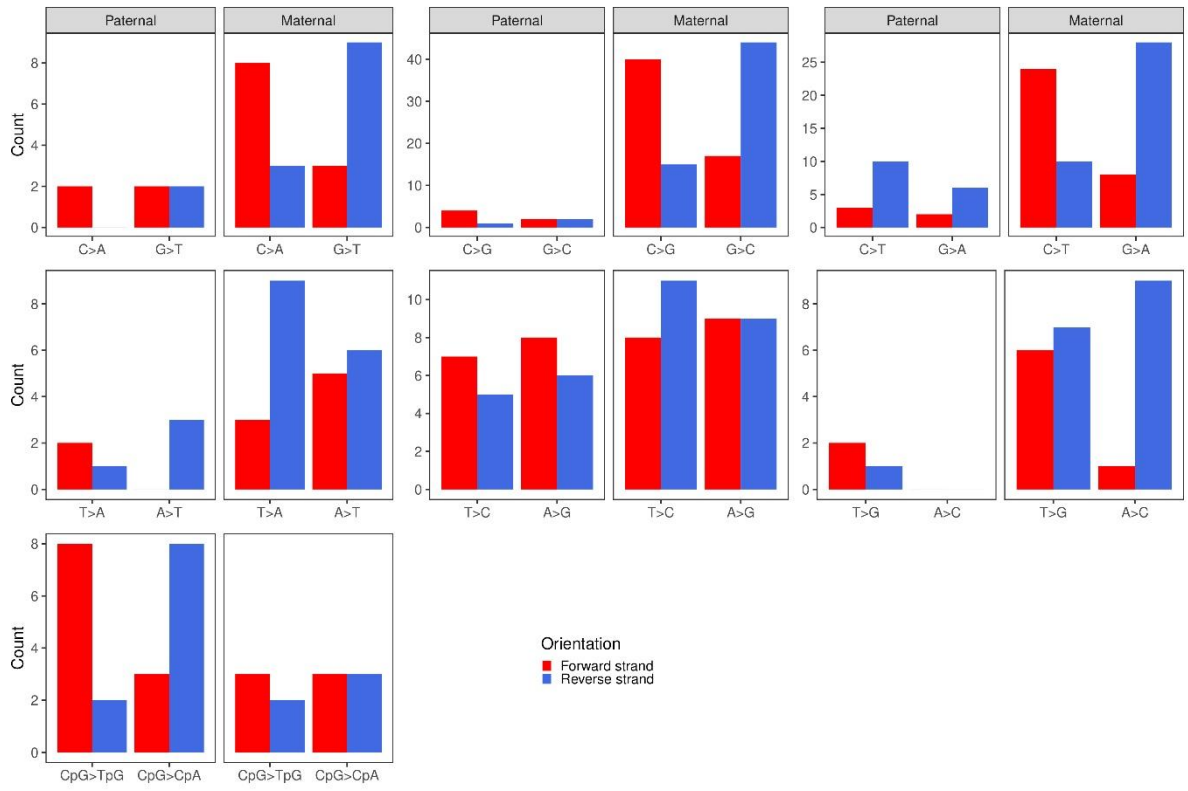
Extended Data Fig. 4 | Variation of recombination maps with replication timing.
The replication timing is computed on the same overlapping 3Mb windows as the maps, split into deciles and the mean map values computed for the windows that fall into each decile. All maps except Δ_{DSB} are normalized individually with respect to their genome-wide mean. The x-coordinate in the figures corresponds to the median replication time in each decile, shifted slightly in opposite directions for paternal/maternal meioses so that errors bars don't overlap for nearby points. ***Early*** *replication time corresponds to replication timing of 1.0,* ***Mid*** *corresponds to 0.0, and* ***Late*** *corresponds to -1.0. The error bars indicate 95% confidence intervals and are computed by bootstrap sampling the map values 1000 times from the set of windows within each decile.*



Extended Data Fig. 5| Distribution of NCO-proximal DNMs.
The count of phased SNP/Indel DNMs near oNCOs vs. distance from the center of the oNCO.



Extended Data Fig. 6 | Comparison of mutation spectra for phased DNMs.
DNMs are considered proximal to oNCOs and COs if they are within 3kb and 100kb for paternally and maternally phased DNMs, respectively. The length of bars indicates the mutation class fraction, computed for 5400 probands. Error bars represent 95% confidence intervals, computed using 1000 bootstrap samples.



Extended Data Fig. 7 | Oriented spectra for SNP DNMs.

The count of phased SNP DNMs within the regions of enriched DNM rate around oNCOs, i.e. within 3kb for paternal DNMs and within 100kb for maternal DNMs. Strand asymmetry of DNM variants and their complement is observed in four mutation classes: maternal C>A (Fisher's test p-value: 0.039), maternal C>G (Fisher's test p-value: $2.3 \cdot 10^{-6}$), and maternal C>T (Fisher's test p-value: $7.0 \cdot 10^{-5}$), and paternal CpG>TpG (Fisher's test p-value: $4.7 \cdot 10^{-3}$).

Supplementary material: Complete human recombination maps

Authors:

Gunnar Palsson[1], Marteinn T. Hardarson[1,2], Hakon Jonsson[1], Valgerdur Steinthorsdottir[1], Olafur A. Stefansson[1], Hannes P. Eggertsson[1], Sigurjon A. Gudjonsson[1], Pall I. Olason[1], Arnaldur Gylfason[1], Gisli Masson[1], Unnur Thorsteinsdottir[1,3], Patrick Sulem[1], Agnar Helgason[1,4], Daniel F. Gudbjartsson[1,5], Bjarni V. Halldorsson[1,2,*], Kari Stefansson[1,3,*]

Affiliations:

1 deCODE genetics / Amgen Inc., Sturlugata 8, Reykjavik, Iceland

2 School of Technology, Reykjavik University, Reykjavík, Iceland

3 Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland

4 Department of Anthropology, University of Iceland, Reykjavik, Iceland

5 School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland

* Corresponding authors.

Setup

Study sample

For this study of gene conversions, genome-wide associations and DNMs we used SNP chip-typed and whole genome sequenced (WGS) Icelandic samples collected as part of disease association efforts at deCODE genetics[22]. The Icelandic genealogical database is used to identify WGS families with two or more children. A single representative child is included from groups of monozygotic siblings. A total of 2,132 families comprising 10,840 meioses are used in the study (Table S1).

Genotype data

A set of 173,025 SNP chip-typed individuals were long-range phased and the parent-of-origin of their haplotypes determined[63,64]. Multiple Illumina SNP chips are used for chip-typing (Table S17) with imputation performed on an imputation marker set of 619,525 markers (611,116 autosomal, 8,409 chrX), which is a subset of the union of the marker sets present on the different chips. These data allow us to compute IBS haplotype sharing between pairs of individuals and confirm parent-child relationships.

Sequence data

A set of 63,118 individuals, a subset of the 173,025 SNP chip-typed individuals, was whole genome sequenced at an average coverage of over 30x using Illumina GAII, HiSeq, HiSeqX and NovaSeq sequencing instruments (Table S17).

The genotypes were determined using GraphTyper v2.7.1[72], using joint genotyping of all 63,118 samples. GraphTyper computes Phred-scaled likelihoods of all the possible genotypes of an individual. The most likely genotype gets a score of 0 and the alternative genotypes get a score x, representing that the likelihood of the alternative genotype relative to the most likely genotype is $10^{-x/10}$. The *Genotype Quality (GQ)* is the Phred score of the second most likely genotype.

The sequence data were subsequently phased, and haplotypes imputed into the 173,025 SNP chip-typed individuals[73].

Sequence variant selection

Gene conversions can only be detected at polymorphic sequence variants and here we restrict our analysis to biallelic SNPs and indels in autosomal sequence data based on the following criteria:

1. Imputation minor allele frequency > 0.5%.
2. Imputation Info > 0.98[73], as measured when the genotypes of the sequenced individuals were imputed into a set of 173,025 chip-typed individuals.
3. Total sequencing depth over all sequenced individuals at marker < 2×(Total genomic average sequencing depth over all individuals).
4. Marker is not inside a homopolymer repeat of length 6 or greater.
5. Sequencing information[73] is in the range of 0.6 – 1.4 for SNPs
6. The genotypes of at least ½ of all genotyped individuals have GQ greater than 40.
7. At least 27.5% and at most 72.5% of reads overlapping variant for individuals genotyped as heterozygous carry the alternative allele.
8. At least 96.5% of reads overlapping variant for individuals genotyped as homozygous carry the called allele.
9. Some individual must have at least 30% of their reads overlapping the variant supporting the alternative allele.
10. Some individual must have at least 4 reads supporting the alternative allele.
11. Root mean square of mapping quality of all reads overlapping a variant is greater than 30.
12. At least 10% of all reads overlapping the marker come from each strand.
13. Hardy-Weinberg equilibrium p-value > 10^{-3} in sequenced samples.
14. Inheritance error < 0.1%, as measured by comparing the genotypes of 35,621 parent offspring pairs, only using those genotypes with GQ greater than 40.

Additionally, to prevent structural variants from introducing false positives in our results we remove genomic regions that are known to harbor structural variants and regions where variant calling with Illumina sequence reads is known to be difficult. See Table S18 for information about the regions removed and Table S2 for details on the variant count per chromosome.

Informative markers – MPPs

Gene conversions transmitted from a father, say, to a proband, can be found at markers where the grandparental origin (GPO) of the paternal alleles inherited by the proband can be determined. Such markers are called informative (for paternal meiosis) and must be phasable for both the father and the proband. Thus, the father, which we refer to as the transmitting parent, must be heterozygous, and in our analysis, we require the mother (the non-transmitting parent) to be homozygous to ensure that the given marker is in the informative marker set irrespective of whether the proband receives a gene conversion or not. The roles are reversed when we search for maternal gene conversion; then the mother is the transmitting parent, and the father is the non-transmitting one. The set of informative markers for a given proband and meiosis is a marker-proband pair, and the markers are referred to as MPPs. As defined, the set of MPPs for paternal and maternal meiosis are disjoint and the MPPs of siblings generally overlap fully apart from markers which may fail quality criteria in some probands.

Quality criteria for MPPs

Phred score

For each MPP we require that its GQ be 80 or higher. This criterion is applied to the genotypes of both the parent and the proband as well as the non-transmitting parent and must also be fulfilled in at least half of the siblings.

Read checks for informative markers

To ensure that MPP variants are correctly characterized as heterozygous in the transmitting parent they must fulfill the following criteria:

a) The variant must be supported by no fewer than 12 qualifying sequence reads (see below).
b) The frequency of each variant allele among qualifying sequence reads is not less than 20%.
c) Each variant allele is supported by at least one qualifying sequence read with mapping quality above 20 and minimum base calling quality above 20 for all bases of the called allele.

d) Each variant allele must be supported by at least one qualifying sequence read in the following categories:
    a. readreverse
    b. not readreverse
    c. firstinpair
    d. secondinpair

Qualifying sequence reads are those that conform to the following criteria:

1. The difference between primary and secondary alignment mapping quality is greater than or equal to 20.
2. Reads must be mapped in a proper pair.
3. Reads must not fail, i.e., SAM flags[74] must not overlap mask 3852.
4. Read and mate must represent the same allele if both overlap the MPP variant.

Other quality considerations

Mendelian errors

We filter out MPPs that display Mendelian inconsistencies, i.e., where the genotype of the MPP is not consistent with the genotype of the homozygous parent. These may be the results of mutations, genotyping errors, structural variants, or aneuploidies. Such MPPs are rejected along with all MPPs within 1kb in both parents. If the total size of rejected regions on a given chromosome is larger than 1Mb we reject the complete chromosome. Note, that this rejection is done for the parent and thus the rejected regions show up in all children even if the errors were only evident in one child.

Abnormal haplotype sharing

Phased chip-genotype data allows us to compute IBS sharing of haplotypes between individuals, and, specifically, we can compute the sharing between parents and children in our study. When such sharing is indicative of an anomaly (such as a large deletion or disomy) larger than 2cM on any given chromosome, the chromosome in question is dropped from the data. In two-children families we remove also the same chromosome for the sibling as we need at least two children to conduct phasing.

Phasing, gene conversions and oNCOs

The genotypes of both the offspring and the transmitting parent must be phased at MPPs. The phasing is straightforward for the offspring as the non-transmitting parent is homozygous at MPPs and thus it is always clear which allele the offspring has inherited from the non-transmitting parent. The genotypes of the transmitting parent must be phased at all MPPs of the children. These data are available from the long-range phasing and imputation processes[63], and thus, we can determine the grandparental origin (GPO) of the allele that each offspring has inherited from the transmitting parent, assigning a GPO-phase value of 1 when the allele has been inherited from the grandfather and 2 when the allele has been inherited from the grandmother.

Considering however, that the imputation of a phase for sequence variants is a probabilistic process, their phasing may have to be adjusted, and thus, we conduct supplementary phasing of the parental data as described in below.

Stable Majority Phasing – offspring-based phasing

For any offspring, a change in the GPO-phase between adjacent MPPs is an indication of either a CO or a NCO. The supplementary phasing process uses these GPO-phase switches and is based on what we refer to as the *Stable Majority Phasing (SMP)* rule, i.e., we assume that no two offspring share a location for changes to their GPO-phase, and thus, the locations for CO and NCOs are assumed to not overlap for two or more siblings in a family. Hence, as we move along the chromosome from one MPP to the next, we would not expect GPO-phase to change for more than one offspring at a time. If more than half of the siblings in a family show GPO-phase changes at the same marker we regard the phasing of the parental marker as inaccurate, and it is adjusted by interchanging the paternal/maternal alleles of the parent as well as the GPO-phases of the probands. Following the adjustment, we interchange the alleles and the originally assigned GPO-phases in all subsequent markers until another adjustment is required, switching us back to the original phasing of the data.

Identification of oNCOs

After the GPO-phase assignment has been completed, we construct contiguous haplotype segments for each offspring based on the GPO-phase of qualifying MPPs. Thus, our data for each parent-offspring pair consists of haplotype segments of a given GPO, starting, and

ending at specific MPPs. The haplotype segments are now classified based on whether their length is larger or shorter than 100kb as either background haplotypes or gene conversion candidates (see Figure 1b). But first and last telomeric segments are never regarded as gene conversion candidates irrespective of size. The largest gene conversion candidate contains 204 MPPs (Table S19). We consider run of consecutive gene conversion candidates an observed NCO (oNCO) if the flanking background haplotypes are of same grandparental origin.

Quality control for oNCOs

After the oNCOs have been identified, we implement quality controls to reduce the likelihood of false positives due to quality issues in sequence data or due to other unidentified genetic variations. This also serves to verify the correct phasing of the transmitting parent, which may need adjustment in families with only two children as the SMP may not assign the detected oNCO to the correct sibling. This can be addressed, either by looking directly at the sequence reads or by applying the surrogate parent concept of the long-range phasing methodology[63] to the sequence data. We outline these processes below.

Shared gene conversions

The phasing process ensures that GPO-phase changes between two adjoining MPPs cannot be shared by more than half of the siblings. Generally, such GPO-phase changes would only show up in one of the siblings as NCOs or COs are unlikely to occur at the same place for two siblings in a family. Recombinations that demonstrate such an overlap are disqualified from our dataset. A total of 620 overlapping oNCOs are omitted, 271 paternal and 349 maternal.

Read-pair-validation

For all oNCOs, we attempt to find read-pairs in the sequence data of the transmitting parent that overlap adjacent MPPs both within and just outside the oNCO. When such information is available, we can often verify that the phasing of the parent is correct and thus confirm the oNCO. We only consider MPP pairs where more than one overlapping read can be found – such pairs are called *links* in what follows. For each link we prepare the following counts:

1. gf_gf : reads that match phasing and correspond to paternal-paternal allele combination

2. gm_gm : reads that match phasing and correspond to maternal-maternal allele combination
3. gf_gm : reads that contradict phasing and correspond to paternal-maternal allele combination
4. gm_gf : reads that contradict phasing and correspond to maternal-paternal allele combination
5. bad : reads that do not match any phasing combination
6. inph = gf_gf + gm_gm : reads that are consistent with the phasing
7. ooph = gf_gm + gm_gf : reads that contradict the phasing
8. total = inph + ooph + bad : total number of reads
9. mainphase = max(inph,ooph) : reads that define the majority phasing configuration
10. maindiff = total – mainphase : number of reads not supporting the majority phasing configuration
11. mainfreq = mainphase/total : the read frequency of the majority phasing configuration

Links are regarded as OK if `maindiff` is less than 3 or `mainfreq` is greater than 90%. oNCOs that have links that are not OK are regarded as demonstrating evidence of three haplotypes and are discarded.

Note, that we don't assume that the phasing is consistent between the markers in the link and thus the *majority phase configuration* is defined as the one supported by the greatest number of reads. If the out-of-phase reads are more numerous, then the original phasing of one of the markers is not supported by the read-pair evidence and a phase adjustment is required. However, such adjustments can only be done on links that connect markers where the grandparental origin of the transmitted alleles is not the same, and where the number of children in the family is two. If adjustments are called for in any other situation, we would be forced to attribute a shared gene conversion or a CO to two or more children and thus the oNCO is rejected.

Hence, read-pair-validation for gene conversions in families with more than two children serves only to identify problematic oNCOs, which are subsequently discarded. For two-children families we have the option of adjusting the phase and thus effectively transferring an oNCO from one sibling to another.

Phasing with surrogate parent

The read-pair-validation only works when the distance between the MPPs is short, within size of the sequencing libraries, which are typically around 500bp. In the majority of cases the distance between converted and unconverted MPPs is longer than the length of typical Illumina read-pairs and thus such validation is not possible.

To overcome this limitation, we use surrogate parents[63] of the transmitting parent. Surrogate parents are locus specific – they can only be surrogates in regions where they share haplotypes with the transmitting parent, but in those regions, they are, just like regular parents, informative of phase at markers where their genotype is homozygous. Surrogate parents are identified with the help of the Icelandic genealogical database[22], and are restricted to individuals less than 11 meiotic steps away.

In rare cases, neither validation method is available. When this happens in two-children families the owner of the oNCO is ambiguous. Those oNCOs are tagged as AO (ambiguous owner) and not used for analysis of allele selection bias and DNMs colocation.

Comparison with earlier methods

In a previous publication[24] a method was developed to search for gene conversions in larger families. Here we refer to this approach as the Large Family Method (LFM). The main difference of the two methods is that LFM requires the genotypes for three generation families as potential gene conversions are verified in the third generation. Also, to verify the parents' haplotypes genotypes LFM requires two siblings of the proband whereas using our current method only one sibling is needed. As shown in Table S16, oNCOs for two-sibling families (proband and a sibling) have almost identical properties as oNCOs in larger families. However, LFM can find gene conversions at markers where both parents are heterozygous. Estimates for our current method (shown in Table S13) are comparable to estimates obtained from LFM (Shown in Table S14).

The LFM first phases the parents using their genotypes as well as the genotypes of their children including the proband. Markers are only considered phased if a potential gene conversion can be distinguished from a single genotyping error otherwise the markers is considered **unphased**.

Then, to verify potential gene conversions in the children of the proband using the genotypes of the proband's children and the children's other parent. This can only be done in regions

where a proband's child carries the haplotype subject to the potential gene conversion, when no such child exists the potential gene conversion is considered **untested**. More detailed information on the LFM is given in the original paper[24].

We used the LFM to search for gene conversion in all probands with the required family structure and compared them with the gene conversions found with our current method on the same set of probands. Overview of this comparison is shown in Table S15.

The exact marker set affects the positions of switches between parent's haplotypes in the child. This can result in a gene conversion in one dataset can become part of a complex crossover in the other, two approximate crossovers in one dataset to become a run of gene converted markers in the other dataset, or a single oNCO in one set can become two oNCOs in the other dataset.

In the set of probands shared between the two cohorts there are 1,459 oNCOs that are common in the two methods. In addition, there are 2,602 oNCOs found only with our current method and 1,381 only found with the family-based method.

Almost all (Table S15) the gene converted markers that were found only with our current method and not with the LFM were either unphased within the LFM or were in a region where no sibling carried one of the parental haplotypes or where gene conversions couldn't be verified. The remaining 29 gene converted markers were missed by the LFM because the placement of switches between haplotypes in the proband resulted in the markers being classified as a part of a complex crossover.

Gene converted markers that are only found with the LFM and not our current method are mostly either markers where both parents were heterozygous or markers where the genotypes of the proband or the parents had too low Phred scores. The remaining gene converted markers only found with the LFM were missed by the current method due to marker quality checks, oNCO quality checks or because the placement of switches between haplotypes in the proband changed the classification of the markers.

Length distribution

We model the distribution of physical NCO lengths as a mixture of negative binomial distributions following the NCOurd methodology[23]. This is done separately for paternal and maternal meioses.

To implement this approach, we prepare two sets of functions: 1) for each offspring, a detection probability functions, $D(x)$, computed from the set of MPPs for the offspring in question, and representing the probability of observing a given NCO of length x. 2) for each oNCO, a tract function $T(x)$, giving the probability that the oNCO represents an NCO of length x, based on the markers within and in the vicinity of the oNCO, where GC bias – indicating mismatch repair – is also observed (Table S20).

We compute the length distributions with varying number of sub-distributions (components), using a likelihood ratio test to determine the number of components required to capture the features of the main distribution while avoiding overfitting. We find that the paternal distribution is well described with four components while the maternal one requires five. The results are displayed in Table S4. In our analysis, the two shortest components are regarded as representing *short* NCOs and the remaining ones as *extended NCOs*. These distributions are shown graphically in Fig. S2.

Model hyperparameters

Two hyperparameters need to be selected; the number of mixture components and the penetrance, the probability of heterozygous markers within an NCO to become gene converted. To choose the number of components we use likelihood ratio tests to evaluate whether increasing mixture components significantly improves the model fit.

To estimate the penetrance, using the fraction of non-boundary markers within oNCOs was suggested, giving an estimate of the value of the penetrance as 0.678 (Table S21A). An alternative method to estimate the penetrance is to compare the likelihood of the observed data given of the model produced by the NCOurd for each value of the penetrance and the choosing the penetrance having the highest likelihood (Table S21B, Fig. S3). This leads to a similar estimate. Both the paternal and the maternal oNCOs have the highest likelihood for the penetrance value of 0.66 and we use this value for the penetrance in this study.

Additionally, we restricted the parameter space such that the standard deviation for each mixture component of the NCO length distribution is no greater than it's mean.

NCO components and short and extended events

After computing the probability mass function of the NCO length distribution, $p(x)$, as a weighted sum of components, $p_c(x)$:

$$p(x) = \sum_i \alpha_c p_c(x)$$

we can for each oNCO compute the likelihood, w_c, of it belonging to subdistribution c by using its tract function, $T(x)$, namely:

$$w_c = \frac{\sum_x \alpha_c p_c(x) T(x)}{\sum_x p(x) T(x)}$$

Thus, each paternal oNCO will be represented by four weights, $w_1 \ldots w_4$, and maternal oNCOs by five weights, $w_1 \ldots w_5$. The short and extended proportions of the oNCO are then given by:

$$w_{short} = w_1 + w_2 \text{ and } w_{extended} = 1 - w_{short}$$

Estimating NCO events – per meiosis and age effects

Assuming the mixture component are fixed but the mixture weights can change between subpopulations we can investigate how the mixture weights and the estimated number of NCOs change with age of the parent.

Using the mixture component, c, with probability mass function, $p_c(x)$, and the detection probability function, $D(x)$, we can estimate the fraction of NCOs in component c expected to be detected as oNCOs. This **detection fraction** is given by:

$$\delta_c = \sum_x p_c(x) D(x)$$

The estimated number of NCOs, N_c, in component c, can then be computed from the number, T_c, of oNCOs in component c as $N_c = \frac{T_c}{\delta_c}$.

Note, that for the whole cohort, we can compute the detection fraction for the total number of NCOs, namely:

$$\delta = \sum_x p(x) D(x) = \sum_{x,c} \alpha_c p_c(x) D(x) = \sum_c \alpha_c \delta_c$$

and get the total number, N, of NCOs from the T oNCOs as:

$$N = \frac{T}{\delta}.$$

However, if we want to look at subsets of the cohort, the relative weights of the component may be different in different subsets. Thus, it is important to compute the number of NCOs separately in each component and sum them up to get the total.

The results for the number of NCOs per offspring are displayed in Table S22. In Table S12 we show the results from a linear regression for maternal age effect on recombination event count. This is computed by splitting the cohort into groups based on the age of the mother at birth of the offspring with each group spanning a range of two years. This analysis did not yield any significant age effect for fathers.

Double-strand breaks

When estimating the total amount of DSBs per meiocyte we note that each offspring represents only one of the four chromatids of the meiocyte. Evidence of DSBs that are resolved as COs will show up in two chromatids, while those that are resolved as NCOs will affect only one chromatid – assuming that NCOs are mainly produced through the SDSA channel of DSB resolution[75]. Thus, the number, N_{DSB}, of DSBs generated in the meiocyte is:

$$N_{DSB} = 4 \cdot N_{NCO} + 2 \cdot N_{CO}$$

where N_{NCO} and N_{CO} denote, respectively, the number of NCOs and the number of COs measured in the proband.

Estimating NCOs – per individual

We saw above how the average number of NCOs can be estimated. The actual number of NCOs per individual is harder to estimate as only a few NCOs are observed per proband. To approach this, we assume here that the number, n_c, of NCOs in component c is a random variable defined on the cohort of probands and we will derive a probability distribution for this random variable, from which we can estimate the number of NCOs per proband.

For this we need to consider the total number of oNCOs that we detect per component per proband, and the fraction, δ, of NCOs that we can expect to detect from each component. For proband a, with detection probability function $D_a(x)$, the detection fraction of NCOs in component c is given by:

$$\delta_{c,a} = \sum_x p_c(x) D_a(x) \, .$$

oNCO count probabilities

For each proband, a, we have a (possibly empty) list of oNCOs, $\{t_{a,m}\}$, where each oNCO, $t_{a,m}$, is represented by 4 (paternal) or 5 (maternal) component weights, $w^c_{a,m}$, with $\sum_c w^c_{a,m} = 1$. The probability that the proband obtained a total of k oNCOs in component c is denoted by $q^c_{a,k}$ and is equal to the coefficient multiplying x^k in the polynomial:

$$q^c_a(x) = \prod_m \left(w^c_{a,m} x + (1 - w^c_{a,m})\right) = \sum_{k=0}^{T_{a,t}} q^c_{a,k}\, x^k$$

Here $T_{a,t}$ is the total number of oNCOs for the proband and if $T_{a,t} = 0$ we get an empty product, which is simply equal to 1 and indicates 100% likelihood of obtaining zero events.

NCO count probabilities

We assume the existence of an underlying probability distribution, $p_{c,nco}$, of the number of transmitted NCOs per proband in component c. This probability distribution is a-priori unknown but will be computed along the way.

For each proband, a, we have the probability, $q^c_{a,k}$, that k oNCOs were observed for component c. The probability that these k oNCOs were due to n NCOs is given by:

$$p_{a,c}\left(n \middle| \delta_{a,c}, k, p_{c,nco}\right) = \frac{1}{Z_{a,c}} p_{c,nco}(n) \binom{n}{k} \delta^k_{a,c} (1 - \delta_{a,c})^{n-k}$$

where $Z_{a,c}$ is a normalization factor such that:

$$\sum_n p_{a,c}\left(n \middle| \delta_{a,c}, k, p_{c,nco}\right) = 1.$$

The distribution $p_{c,nco}$ is computed iteratively, and we denote the results for iteration $j = 1,2, \ldots$ as $p^j_{c,nco}$. We start the iterations with a prior $p^0_{c,nco}$, which is simply taken as the uniform distribution from 0 to L ($L = 3000$ was used). Then given $p^j_{c,nco}$, we compute the updated probability distribution as:

$$p^{j+1}_{c,nco}(n) = \frac{1}{A} \sum_a \sum_{k=0}^{T_{a,t}} q^c_{a,k}\, p_{a,c}\left(n \middle| \delta_{a,c}, k, p^j_{c,nco}\right).$$

Here, A is the number of probands in the cohort. Iterations are continued until the total absolute difference between successive iterations is less than 10^{-5} or till minimum variance is reached, in which case further iterations would lead to overfitting.

After convergence is reached, we can compute the statistics for the number of NCOs per individual a from the probabilities, $p_{a,c}\left(n\middle|\delta_{a,c},k,p_{c,nco}\right)$. The expected number of NCOs of component c is given by:

$$\langle n_{a,c}\rangle = \sum_{n}\sum_{k=0}^{T_{t,a}} n \cdot q_{a,k}^{c}\, p_{a,c}(n\,|\delta_{a,c},k,p_{c,nco}).$$

We use this expectation value to rank the NCO counts for the probands in the cohort and then map those rankings to the $p_{c,nco}$ distribution. The mapped value is regarded as the NCO count for the proband. The estimated distributions of the number of NCOs, COs, and DSBs per individual in the cohort are shown in Fig. S4.

NCO maps

We create genetic NCO maps in a similar fashion to the genetic maps created for CO recombination. Thus, like in the early genetic maps[2], we group our data into overlapping bins of size 3Mb with step size 1Mb. To compute the number of NCOs from the number of oNCOs within each bin we use the same methodology as we used when computing the number of NCOs per proband, i.e., we have an underlying distribution of the number of NCOs per bin and then use the detection probability per bin to estimate the number of NCOs per bin. The detection probability per bin is computed in the same way as the detection probability for the complete genome[23], i.e. by averaging the detection probabilities for each proband, restricted to the bin in question.

Annotations

Correlation of oNCOs with genomic annotations is estimated from their relative rate in the annotated regions. We define the following:

- n_A: Number of oNCOs in annotated regions
- N_A: Number of sequence variants in annotated regions
- n_G: Total number of oNCOs
- N_G: Total number of sequence variants.

The relative rate is now given by:

$$RR = \frac{n_A}{N_A} / \frac{n_G}{N_G}$$

An oNCO is regarded as being in an annotated region if its center is located within it.

Annotations for crossovers

For comparison between COs and NCOs we compute relative rates for COs using data from our earlier publication[1]. This is computed in a similar fashion as above:

- x_A: Number of COs in annotated regions
- R_A: Size of annotated regions in bases
- x_G: Total number of COs
- R_G: Size of autosomal genome excluding centromeres (2,756,098,793)

The relative rate is now given by:

$$RR = \frac{x_A}{R_A} / \frac{x_G}{R_G}$$

A CO is regarded as being annotated if its center is located within the annotated region. The center is computed as the median location of the CO using a published recombination map[1].

Altemose annotations

We explore enrichment of recombination with two sets of motif binding maps by Altemose et. al.[39] (see Data availability section). The maps provide a single location for motif matches. We perform liftover[76] of this data from GRCh37 to GRCh38 and create 1000 bp regions centered on the motif locations, merging overlapping regions.

ChromHMM annotations

To explore enrichment of recombinations within regulatory elements[40] we use tissue specific (testis/ovary) samples and combine several samples for each tissue into a consensus annotation with a reduced number of states. When samples disagree on the annotation of a segment of the genome that is shared between them the consensus annotation for that segment matches the annotation with the higher priority of the two discordant annotations. A list of annotations, consensus states, and priority order is given in Table S23.

GWAS

We performed genome-wide association studies (GWAS) to explore how the genetic makeup of parents affects phenotypes derived from NCO data. This analysis is analogous to the one conducted for CO data in our earlier publication[1]. Thus, we have five phenotypes, namely:

- *Rate* – number of NCOs per child
- *GC content* – average GC content within 50 bases of oNCO center
- *Telomere distance* – the distance from oNCO to nearest telomere.
- *Hotspot usage* – percentage of oNCOs within recombination hotspots
- *Replication timing* – replication timing value at oNCO location.

Association is performed for both paternal and maternal phenotypes as well as joint, giving 15 phenotypes in total. Of these 15 phenotypes, the only ones where genome-wide significant association is found are the ones measuring hotspot usage of oNCOs (Table S24, Fig. S7).

The strongest association is at the *PRDM9* locus on chr5: rs2973614 (MAF = 3.17%) (effect = $-0.6 \cdot$SD, p = $1.1 \cdot 10^{-21}$) for the joint phenotype, and rs2973613 (MAF = 3.17%) (effect = $-0.6 \cdot$SD, p = $2.1 \cdot 10^{-12}$) and (effect = $-0.6 \cdot$SD, p = $1.0 \cdot 10^{-11}$) for the paternal and maternal phenotypes, respectively. rs2973614 associates very strongly with hotspot usage in COs (effect = −1.7, p = $4.3 \cdot 10^{-2382}$).

One other variant associates at Bonferroni corrected genome-wide significance level[77] with maternal hotspot usage (Fig. S7b), but is no longer significant when correcting for the fifteen phenotypes tested. This is a missense variant on chr12, rs765589892, *POU6F1*:p.Asn14Ser (MAF = 0.15%) (effect = 1.7 SD, p = $4.0 \cdot 10^{-8}$).

Motif co-localization

We investigate the co-localization of oNCOs with seven non-degenerate binding motifs for the PRDM9 reference allele (B-allele)[39] and perform independent analysis for each motif. Using Biopython[78,] we calculate a threshold for motif matches such that the false-positive rate in randomly generated sequence of base pairs is one in ten million.

For each oNCO, we define a search region based on the span of the markers defining the oNCO extended by 2000 base pairs in each direction. We then search for motif-matches in the GRCh38[65] reference genome and its reverse complement within each search region.

To estimate co-localization of motif matches and oNCOs it is important to have an accurate placement for both the oNCO and the binding site of the PRDM9 motif. To obtain the best estimate of the chromosomal position of oNCOs, we limited the data set to oNCOs having exactly one gene-converted marker. For motifs we use the chromosomal position of their central basepairs.

To evaluate the reliability of the PRDM9 motif binding site as being informative for DSB co-localization we look specifically for consecutive motif matches passing the threshold anywhere in the GRCh38 reference genome. We observe that motif 4 has an abundance of pairs of motif matches within 50 base pairs of each other. Motifs 3 and 7 also have a considerable fraction of motif matches within 50 base pairs of a previous match (Fig. S5). Thus, these motifs are not as reliable for accurate placement estimates and we only report results for motifs 1,2,5, and 6, which we refer to as primary motifs.

For each oNCO with proximal (<2000 bp) matches to the primary motifs we identified the highest scoring match and measured its distance and direction with respect to the oNCO. Thirty-nine oNCOs were omitted because of multiple highest scoring matches within the search region. Significant bias (Fig. S6) was found for the motif match being downstream of the oNCO center; 60.3% (95% CI, 57.7%-62.9%, p-value $2.6 \cdot 10^{-14}$) and 55.3% (95% CI, 52.3%-58.4%, p-value $5.0 \cdot 10^{-4}$) for paternal and maternal oNCOs, respectively (Table S25).

Computing the median of the primary motif-match densities we observe that their location is 36 bp (95% CI, 26.5-41 bp) downstream of the motif center. Confidence intervals for these estimates were calculated with bootstrapping the oNCOs having a proximal match. Results for individual motifs are show in Table S7.

There are two possible explanations for this downstream bias of the highest scoring motif match relative to the oNCO. The first is that the placement of the double strand breaks is biased towards the region upstream of the PRDM9 motif binding site. An asymmetry of DSB placements relative to PRDM9 motif has been reported for individual DSB hotspots in mice[79]. Such asymmetry can explain this bias if the human DSB hotspots predominantly create DSBs upstream of the PRDM9 motif. The second possible explanation is that there is a strand bias for the invading strand in the SDSA pathway favoring the unbound strand. In this case the unbound strand would invade more frequently upstream relative to the motif, possibly leading to upstream gene conversions.

Crossover interference

We used the crossover data to estimate parameters of the Housworth-Stahl model[80,81]; crossover interference (ν) and escape from crossover interference (p). Crossover formation is a well-regulated process known to be under strong genetic control[82]. The formation of one crossover is known to reduce the probability of a second crossover occurring nearby under a process known as crossover-interference. A subset of crossovers, however, appears to escape crossover interference during female meiosis[82]. Larger crossover interference parameter (ν) means that the crossovers are less clustered and more evenly distributed, while ν=1 represents no crossover interference and random distribution of crossovers across the chromosome. High levels of the crossover escape parameter, (p), similarly represents more random placement of crossovers across each chromosome.

Crossover interference parameters were computed using the function fitStahl in the software package xoi[83], using data described previously[1]. The data consisted of crossovers for 56,291 and 70,035 paternal and maternal meiosis, respectively, for each of the 22 autosomes. We estimated (ν) in as 8.05/6.59 and (p) as 0.050/0.039 and in paternal and maternal meiosis, respectively. These numbers have previously been estimated[84] as (ν) 8.93/7.19 and (p) 0.067/0.078. Our estimates correspond to NCO/CO ratios of (8.05-1)·(1-0.05)+0.05=6.75 paternal and (6.59-1)·(1-0.039)+0.039=5.21 maternal per meiocyte or 6.75/2=3.37 paternal and 5.21/2=2.61 maternal per transmitted haplotype.

De novo mutations (DNMs)

Identifying DNMs in Icelandic trios

We scanned the sequence variants and extracted DNM candidates in a similar manner as before[1,9] for 9,643 trios by comparing the genotypes of the parents and offspring, 5,400 of which were common with the study of NCOs. Briefly, we defined a DNM candidate with permissive cutoffs for the genotype of the proband requiring that allele balance is greater than 0.25 and depth of 12 reads at the position (supporting either the reference or alternative allele). For the genotypes of the parents, we require at least 12 reads, maximum of one read supporting the alternative allele and the allelic balance to be less than 5%. Likely (N_{LIK}) and

possible carriers (N_{POSS}) of the DNM allele outside the descendants of the parent pair were defined as before[9]. We restricted to DNM candidates with less than 50 likely carriers and either of following less than 10 possible carriers or the ratio N_{LIK}/N_{POSS} is greater than 80%. We tuned the DNM candidate filtering by using segregation of DNM candidates in three generation families (2,042 probands) and the following quality covariates in generalized additive model with a logistic link:

- AAScore: Prediction probability from Graphtyper that the variant is a true positive.
- Carrier_regression_beta: Slope from the alternative allele depth regression for the sequence variant. The alternative allele counts were regressed (AD) on the depth (DP) conditioned on the genotypes (GT) reported by GraphTyper[72]. For a well-behaving sequence variant, the mean alternative allele count for a homozygous reference genotype should be 0, for a heterozygous genotype it should be DP/2 and for a homozygous alternative genotype it should be DP. Under the assumption of no sequencing or genotyping error, the expected value of AD should be DP conditioned on the genotype, in other words an identity line (slope 1 and intercept 0). Deviations from the identity line indicate that the sequence variant is spurious or somatic.
- Carrier_regression_alpha: Intercept from the alternative allele depth regression (see description for Carrier_regression_beta) for the sequence variant.
- Proband_het_AB: The allelic balance of the proband.
- MaxAAS: The maximum read support for the sequence variant across all individuals.
- Alignment_Alt_Reads: The number of reads supporting the alternative allele. This covariate and the following covariates were derived by identifying the reads in the BAM files supporting DNM allele.
- Alignment_Alt_Unique_Positions: The unique number of starting positions for the reads supporting the alternative allele.
- Alignment_Alt_Soft_clipped: The number of soft clipped bases (S in CIGAR string).
- Alignment_Alt_Matched_bases: The number of matched bases (M in CIGAR string).
- Alignment_Alt_Score_diff: The difference of the alignment score of the best and the second best hit as reported by BWA mem.
- Alignment_Alt_Pair_sw_nm: The pairwise mismatches between reads supporting the alternative allele using Smith Waterman implementation in SeqAn[85].
- Alignment_Alt_Pair_align: The number of bases in the pairwise alignments.

With following formula for the gam function from the mgcv R package[86]:

```
threegen_Consistent_hs~
        I(cut(alignment_Alt_Unique_Positions,c(-1,2,4,8,10,Inf)))+
        s(I(AAScore))+
        s(Carrier_regression_beta)+
        s(Carrier_regression_alpha)+
        I(ifelse(alignment_Alt_Reads>0,
            (alignment_Alt_Score_diff/alignment_Alt_Reads)>10,
            FALSE))+
        I(ifelse(alignment_Alt_Pair_align>0,
            (alignment_Alt_Pair_sw_nm/alignment_Alt_Pair_align)>0.05,
            TRUE))+
        I(ifelse(alignment_Alt_Matched_bases>0,
            alignment_Alt_Soft_clipped/alignment_Alt_Matched_bases>0.5,
            TRUE))+
        s(Proband_het_AB)+
        s(ifelse(MaxAAS>15,
            16,
            MaxAAS))+
        I(NPOSS == 0)
```

We restricted to instances of the DNM candidates where we see both of the proband's haplotypes at a locus transmitted to the offspring of the proband (See Figure 1C in[9]). Briefly, if the DNM is a true germline variant then the allele of the DNM candidate should be present in one of the proband's offspring. On the other hand, if it is absent from the children then this assay suggests that the DNM candidate is a false positive DNM call (See more detailed description in[9]). Like before we fitted the generalized additive model using the mgcv R package[86].

We used the cutoff of 0.5 to call high quality DNM candidates for the predicted probability of correct segregation in three generation families. To validate the false positive detection rate of the DNMs we also used the genotype consistency between pairs of monozygotic twins, we find that 3.8% of DNMs are unobserved in the monozygotic twin of the proband. Note that this approach overestimates the false positive rate as there are instances of high frequency post-zygotic mutations that differ between pairs of monozygotic twins[87].

Phasing of mutations in Icelandic trios

Like before we used two complementary approaches to phase DNMs[1,9], one by tracking of transmission of the DNMs to the offspring of the proband (142,190 DNMs; three generation phasing) and another one using read tracing of the DNM allele to phase informative alleles (247,287 DNMs; read phasing). We combined the phase of the DNMs from the different approaches resulting in 333,950 phased DNMs. If the phase of DNMs differed between the approaches they were set to unphased DNMs in the downstream analysis.

Enrichment of DNMs near oNCOs

We compute enrichment of DNMs around oNCOs by looking for DNMs within 1Mb from their centers. The center is taken as the midpoint between the two outermost gene-converted markers that define the oNCO.

We split the search region into six bins, namely:

1. 0 – 1kb size: 2kb
2. 1kb – 3kb size: 4kb
3. 3kb – 40kb size: 74kb
4. 40kb – 100kb size: 120kb
5. 100kb – 0.5Mb size: 800kb
6. 0.5Mb – 1Mb size: 1Mb

DNMs found near oNCOs are assigned to a **distance bin** depending on their distance to the center of the oNCO.

Component-wise calculations

The enrichment of DNMs is computed separately for each distance bin, parent type, and each mixture component of the NCO length distribution. When considering enrichment within and outside CGER, we separate the set of oNCOs and DNMs into those that fall inside and outside CGER and calculate the enrichment separately for each group. Similarly, we can compute age dependence for the DNMs enrichment by separating our cohort into groups based on parental age.

For the remainder of this section, we calculate the enrichment for a specific **distance bin, parent type and component** combination and sum over the NCO components in the following section. Each transmitted oNCO has weight in the component equal to the likelihood of it belonging to the component (section 0). Each DNM found close to an oNCO is associated to this component with the same weight as the oNCO.

Let w_i be the weight of oNCO number i. Let $p_i, \overline{p}_i$ and u_i be the number of DNMs co-located with oNCO number i phased to the transmitting parent of the oNCO, phased to non-transmitting parent and unphased, respectively. For most oNCOs all three numbers will be zero.

Using the age specific DNM rates for the parents of each proband (Table S26) we can for each oNCO, i, assign two numbers (α_i, β_i) denoting the expected (fractional) number of mutations from the oNCO-transmitting parent (α_i) and the non-transmitting parent (β_i). These numbers are computed from the mutation rate and the size of the distance bin under consideration.

For the component and bin under consideration we now define:

$p = \sum_i w_i \cdot p_i$	$\overline{p} = \sum_i w_i \cdot \overline{p}_i$	$u = \sum_i w_i \cdot u_i$
$t = p + \bar{p} + u$	$\alpha = \sum_i w_i \cdot \alpha_i$	$\beta = \sum_i w_i \cdot \beta_i$.

Now, to estimate the fraction, f_p, of DNMs within each bin, that originate from the oNCO transmitting parent, we regard the number of DNMs as the results of a binomial process, and thus we can estimate f_p using a Beta distribution prior $Beta(\alpha, \beta)$[88]. Hence, we get:

$$f_p = \frac{p + \alpha}{p + \alpha + \overline{p} + \beta}.$$

And then the total estimate m, of DNMs originating from the oNCO-transmitting parent is given by:

$$m = \max(p, t \cdot f_p)$$

And **the elevation of the mutation rate** within this distance bin is given by:

$$e = m/\alpha$$

NCO contribution to mutation rate

To compute the contribution of NCOs to DNM rate we continue with the computation from above and first sum up the total number of phased DNMs for all distance bins. Thus, we add indices j and c to the quantities (m, α above become $m_{j,c}, \alpha_{j,c}$) we computed in section 0, signifying distance bin j and component c. Now we get:

$$m_c = \sum_j m_{j,c} \text{ and } \alpha_c = \sum_j \alpha_{j,c}.$$

Note, that m_c represents the total number of DNMs near oNCOs and thus to find the contribution from the NCOs we need to subtract the DNMs expected from the background DNM rate, α_c.

Now let n_c denote the sum of oNCOs for component c, and N_c denote the corresponding expected number of NCOs. Here, N_c is computed from the age of the parent, using the parameters presented in Table S12 for the age dependence of NCOs.

The number of DNMs that can be attributed to NCOs of this component is now given by:

$$m_{nco}^c = (m_c - \alpha_c) \cdot \frac{N_c}{n_c}$$

where we have subtracted the DNMs, α_c, attributable to the background mutation rate. Then the total number of DNMs attributed to NCOs is given by:

$$m_{nco} = \sum_c m_{nco}^c.$$

Based on the age-dependent mutation rates for the parents of our cohort we can now compute the total expected number of DNMs, m_{tot}, and then get the fractional contribution of NCOs to the mutation rate as:

$$f_{nco} = \frac{m_{nco}}{m_{tot}}.$$

The elevation of mutation rates as computed with these methods are displayed in Extended Data Table 3 and Table S9. In this table we have also shown the elevation results separately within C>G enriched regions[9].

Confidence intervals and p-values

Unless specified otherwise, confidence intervals for all statistics are computed by bootstrapping from the set of parents. I.e., quantities are computed per parent based on the data for the children and then population statistics are computed for 1000 bootstrap samples from the set of parents. The 1000 sample-sets of parents are prepared independently, and the same sample-sets used wherever such bootstrapping is called for. The reported value of a statistic is then given by the point estimate with the 95% confidence intervals computed with interpolation from the bootstrapped statistics.

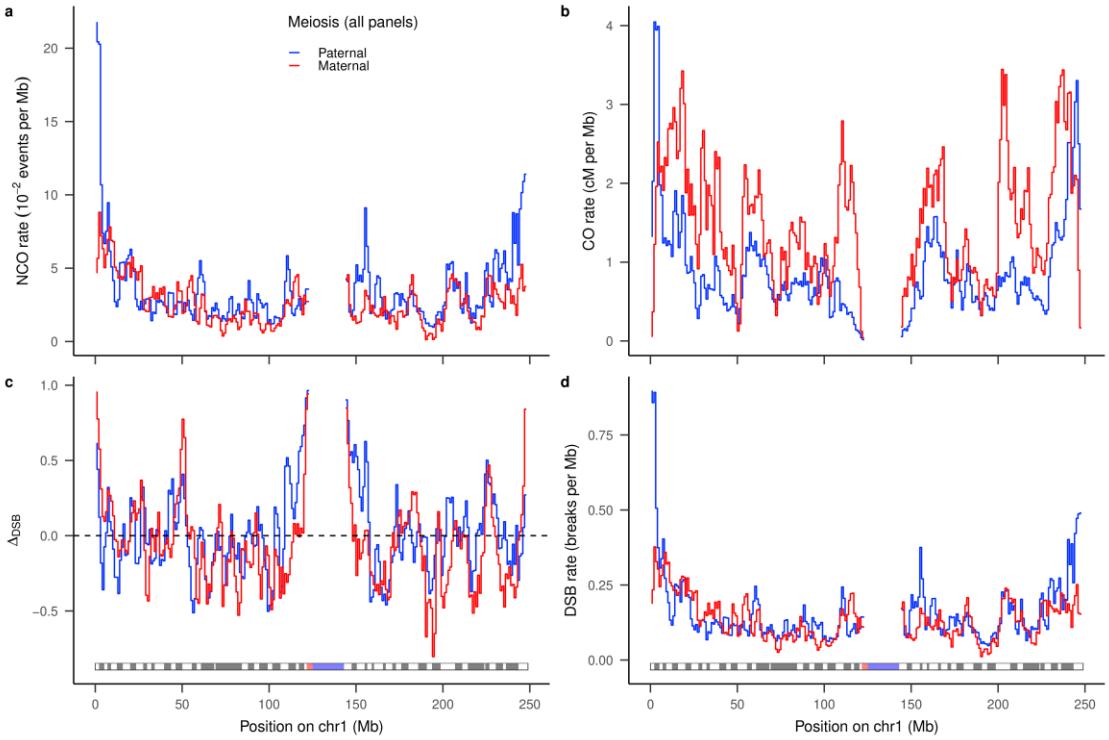
We compute a two-sided p-value for the hypothesis that the (μ) of two datasets, A and B, are the same by merging their bootstrap sampling results by sample ID and counting the, p, proportion of cases where the results for A are larger than the matched results for B. The p-value is equal to 2 times the minimum of p and (1-p). As we are using 1000 samples for our bootstrapping process a p-value shown equal to 0 indicates that $p < 0.002$.

Figures

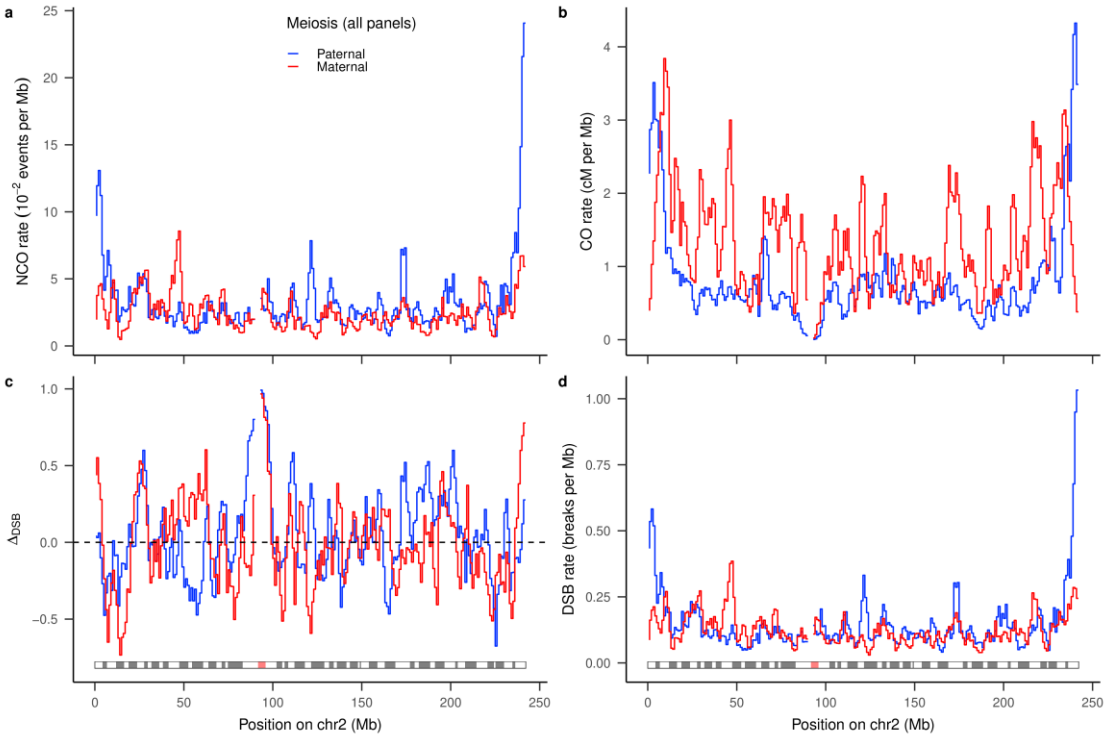
Fig. S1. RECOMBINATION MAPS

Recombination maps for autosomal chromosomes. Panels show: **a|** NCO maps, **b|** CO maps[1], **c|** Δ_{DSB} maps, and **d|** DSB maps. GRCh38 cytobands[66,67] are shown below the graphs in panels **c** and **d**. The centromere (**acen**) is red, **gneg** bands are white, all **gpos** bands are gray, and **gvar** and **stalk** bands are blue. Gaps in the maps indicate regions with no MPPs.

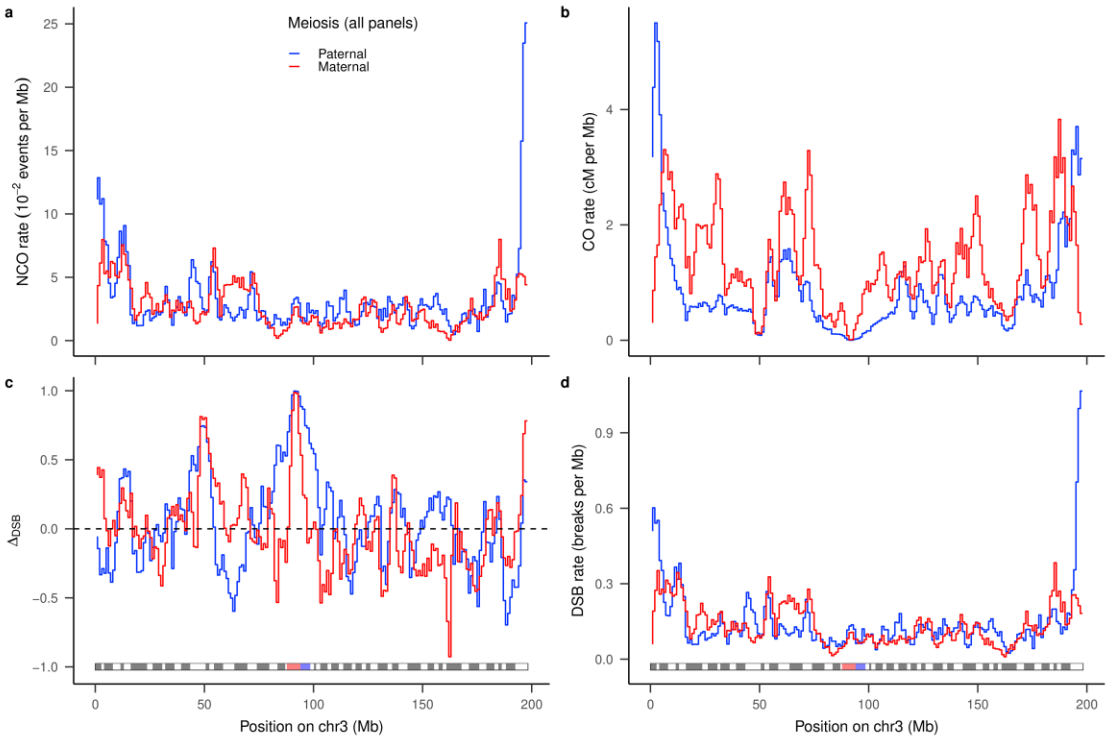
a) Recombination maps – chromosome 1



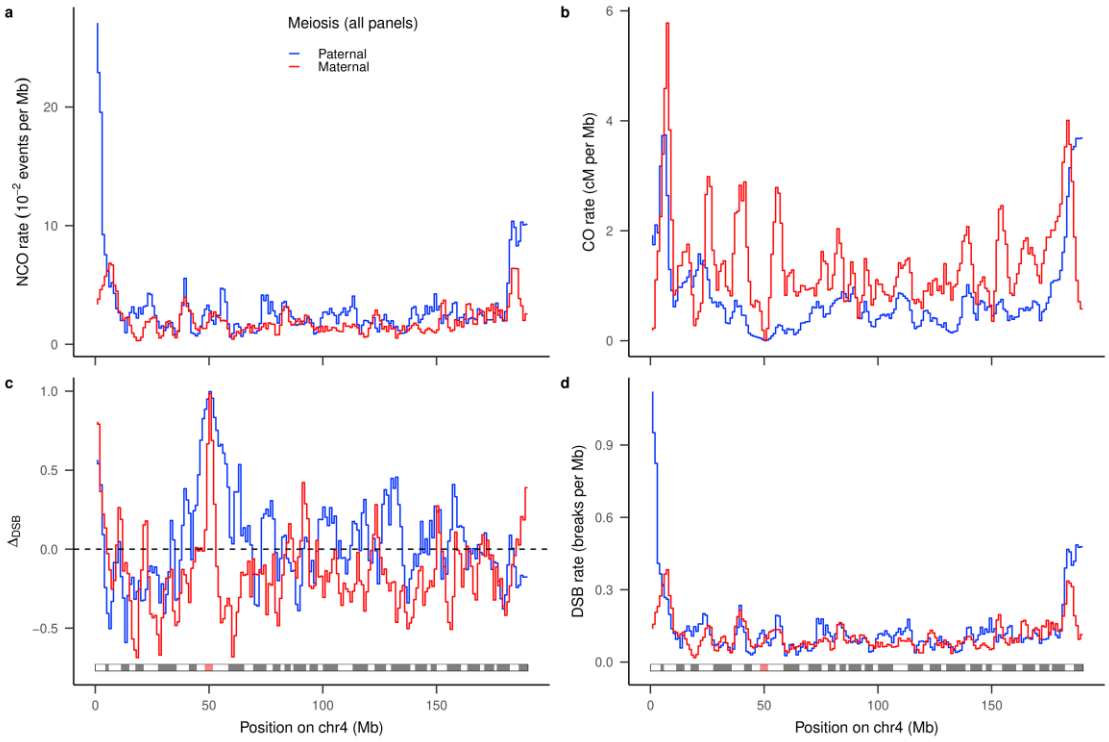
b) Recombination maps – chromosome 2



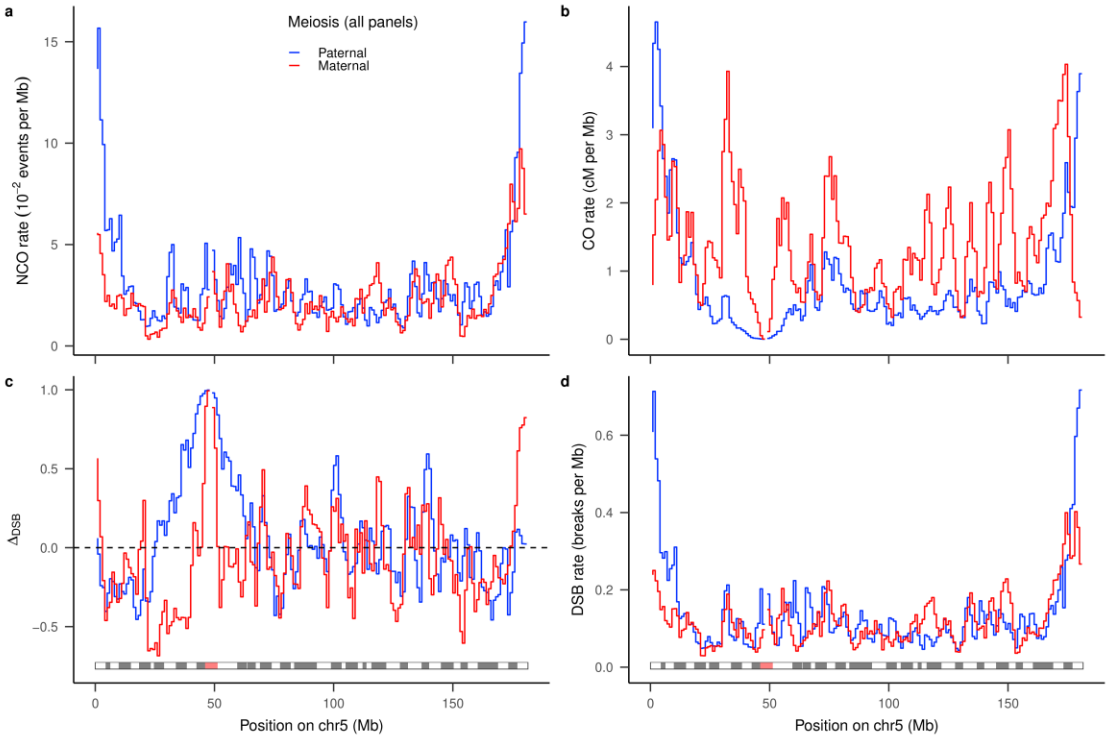
c) Recombination maps – chromosome 3



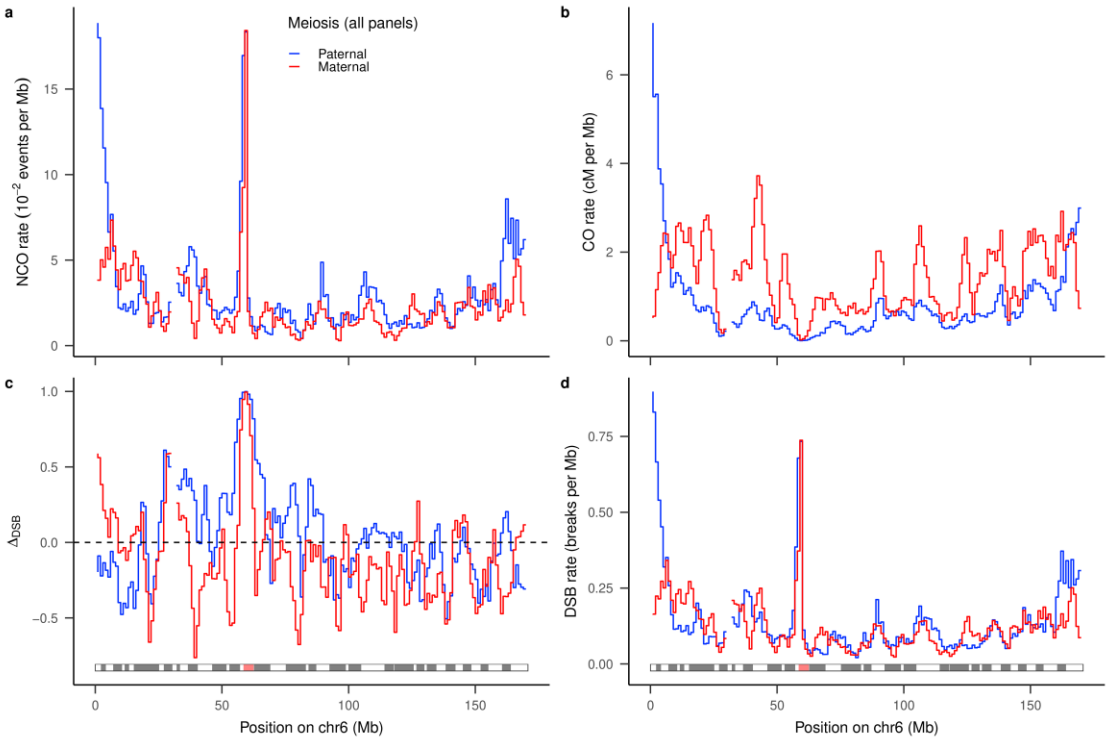
d) Recombination maps – chromosome 4



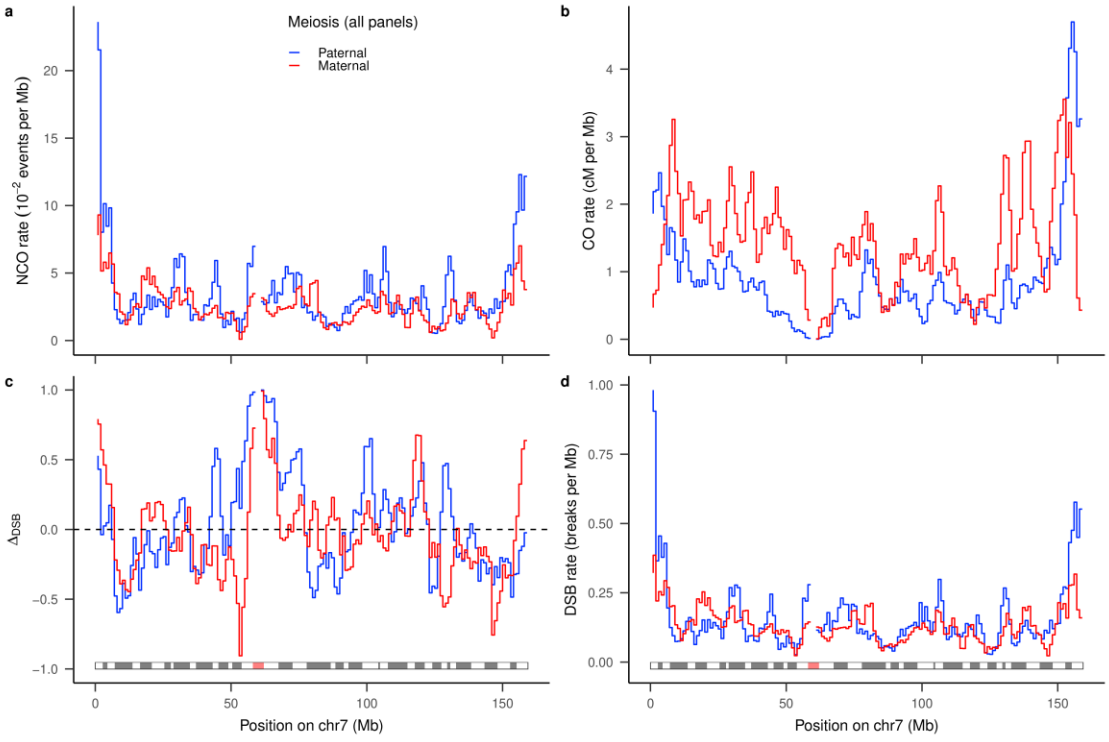
e) Recombination maps – chromosome 5.



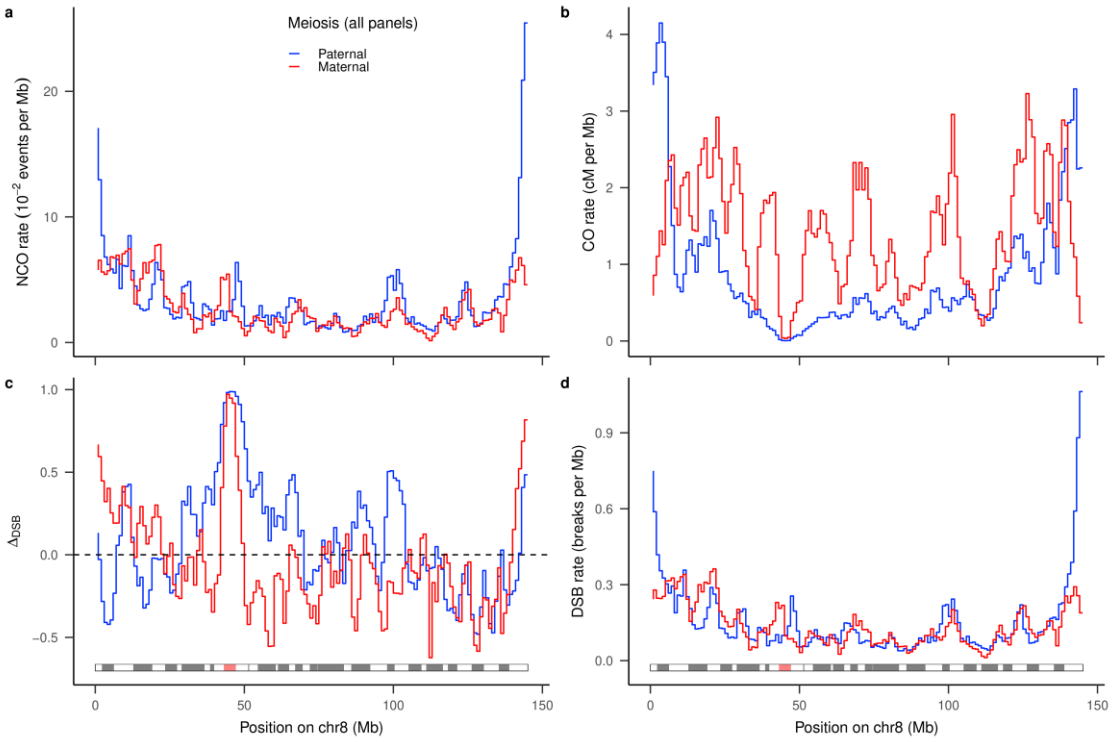
f) Recombination maps – chromosome 6



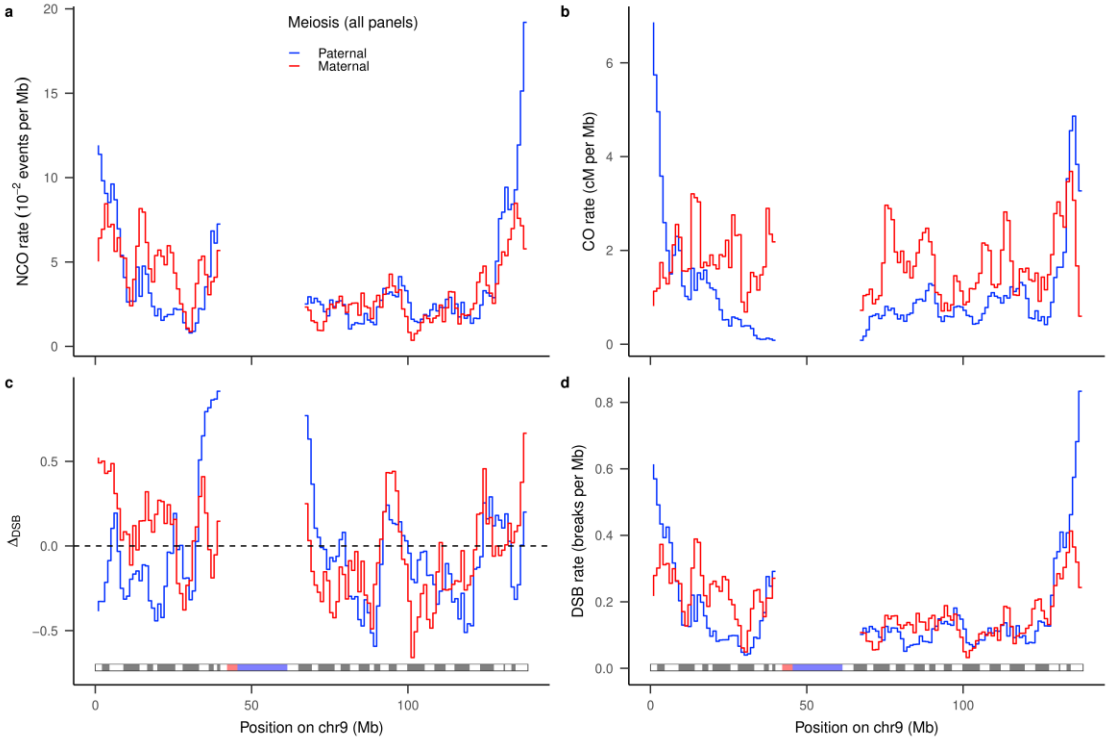
g) Recombination maps – chromosome 7



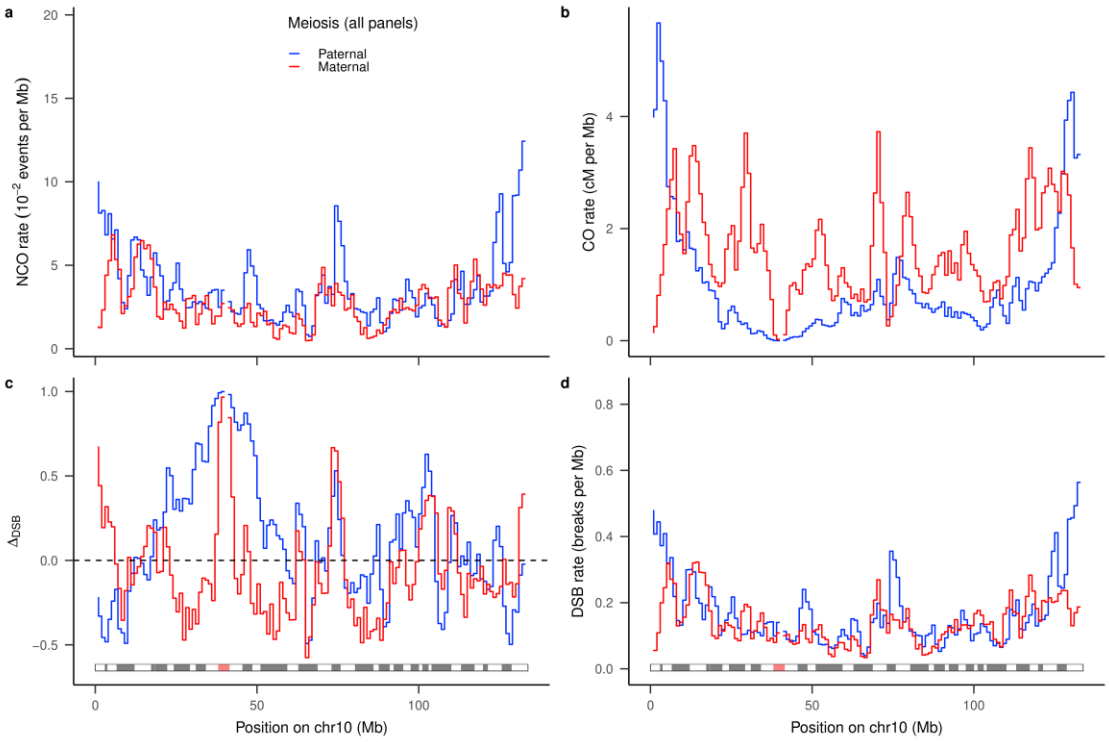
h) Recombination maps – chromosome 8



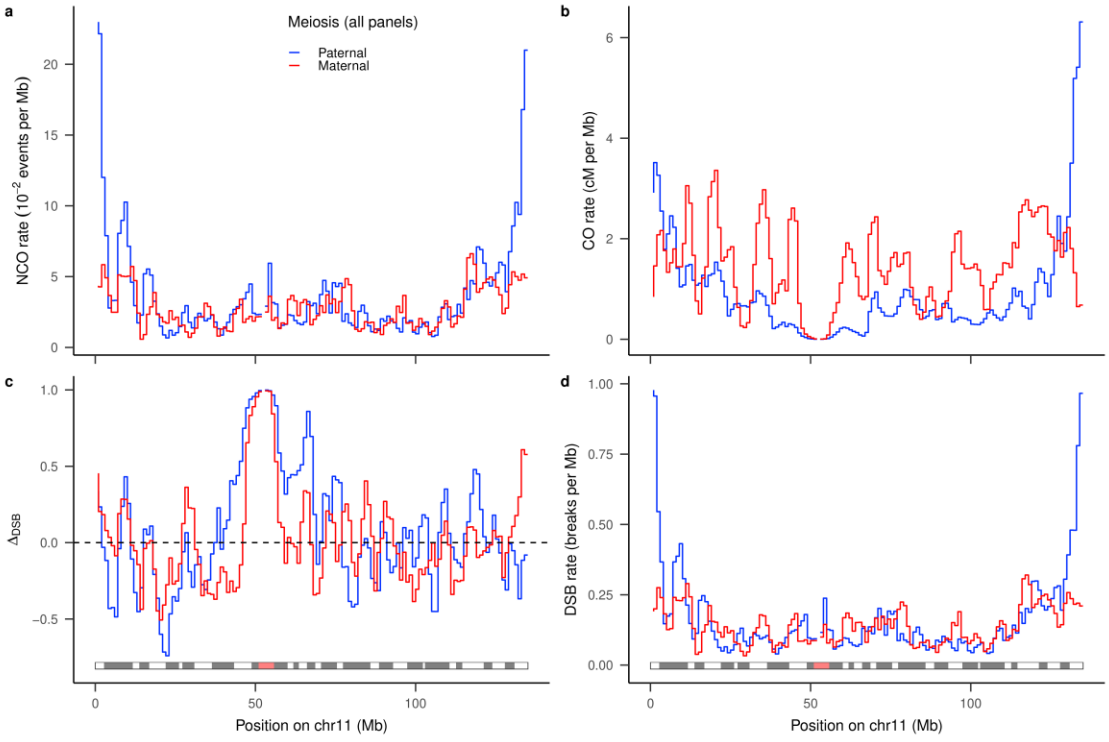
i) Recombination maps – chromosome 9



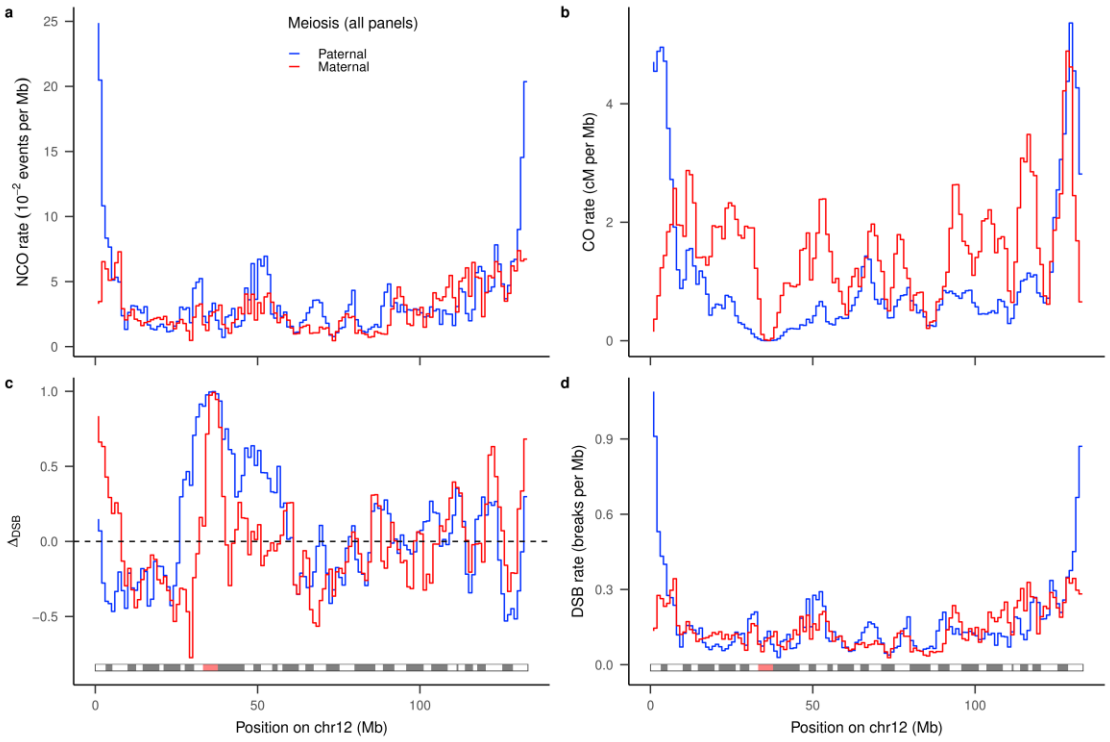
j) Recombination maps – chromosome 10



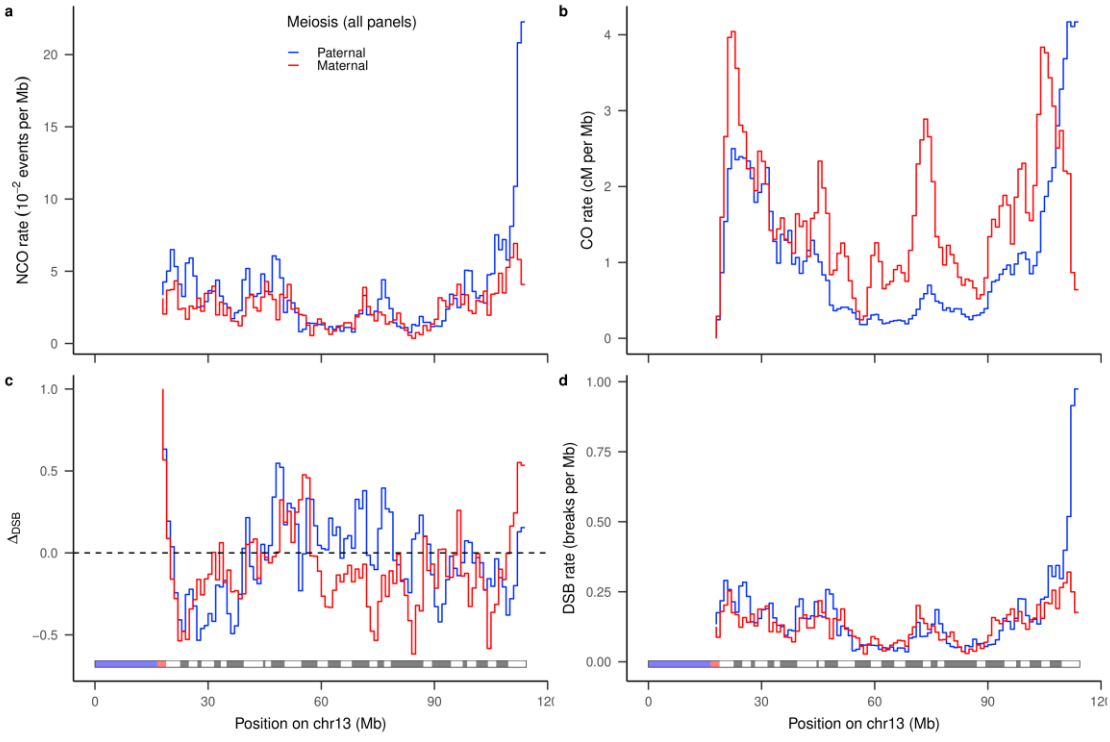
k) Recombination maps – chromosome 11



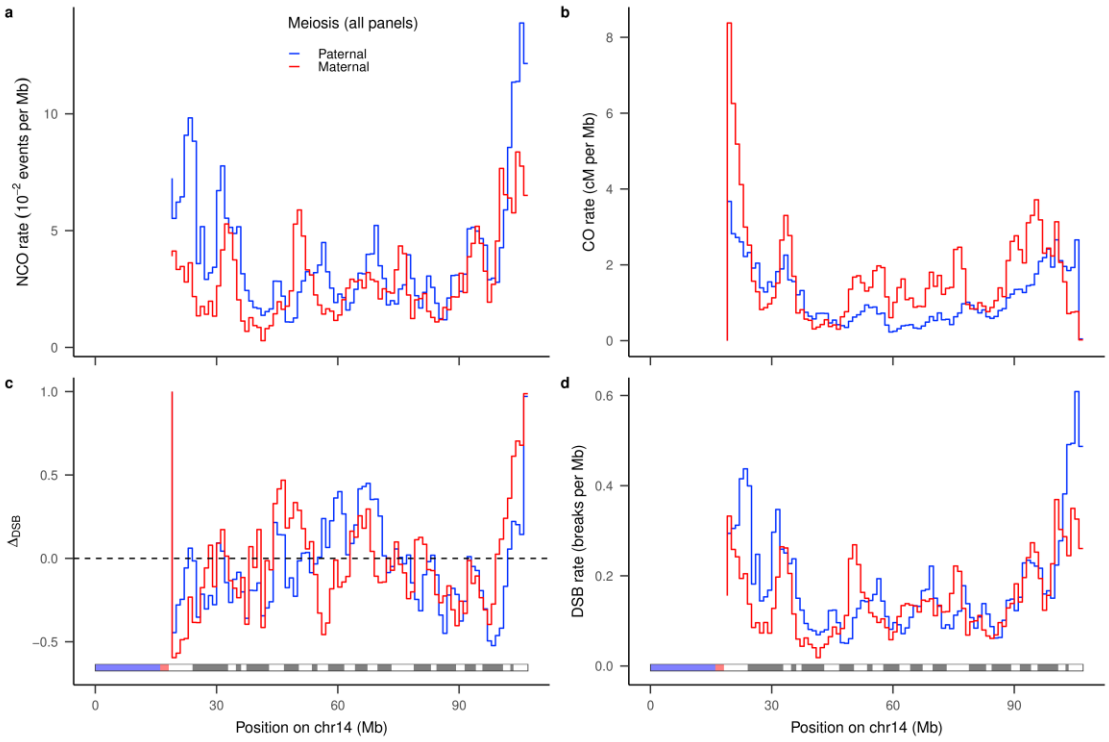
l) Recombination maps – chromosome 12



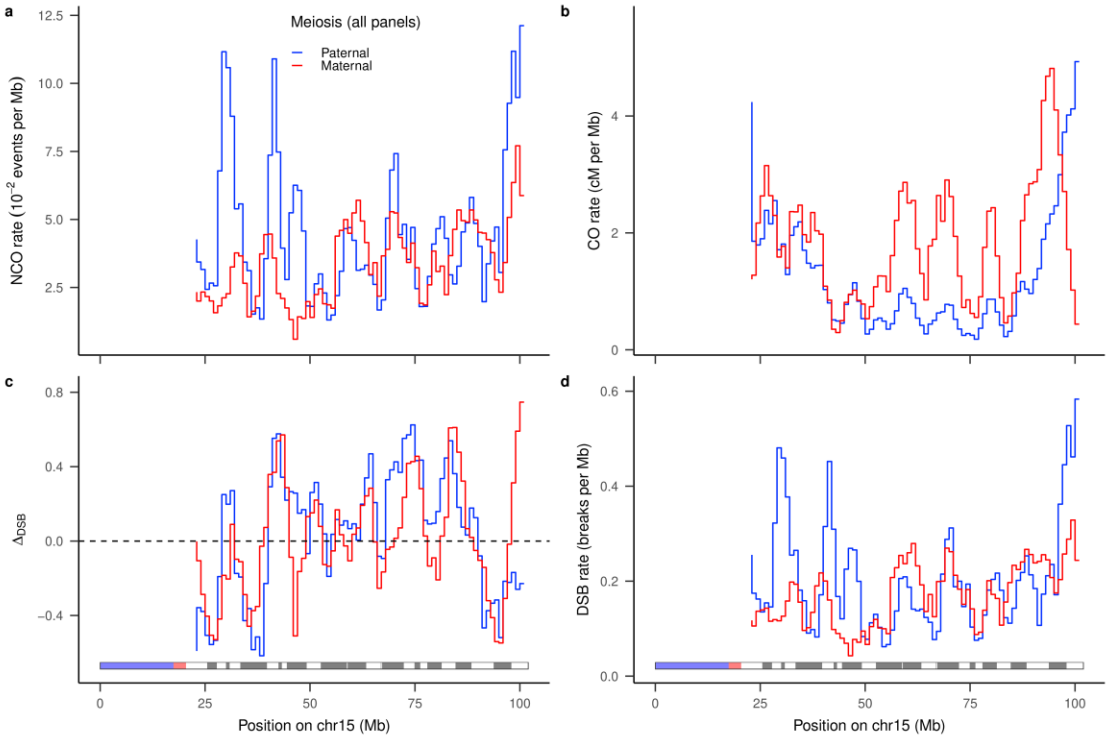
m) Recombination maps – chromosome 13



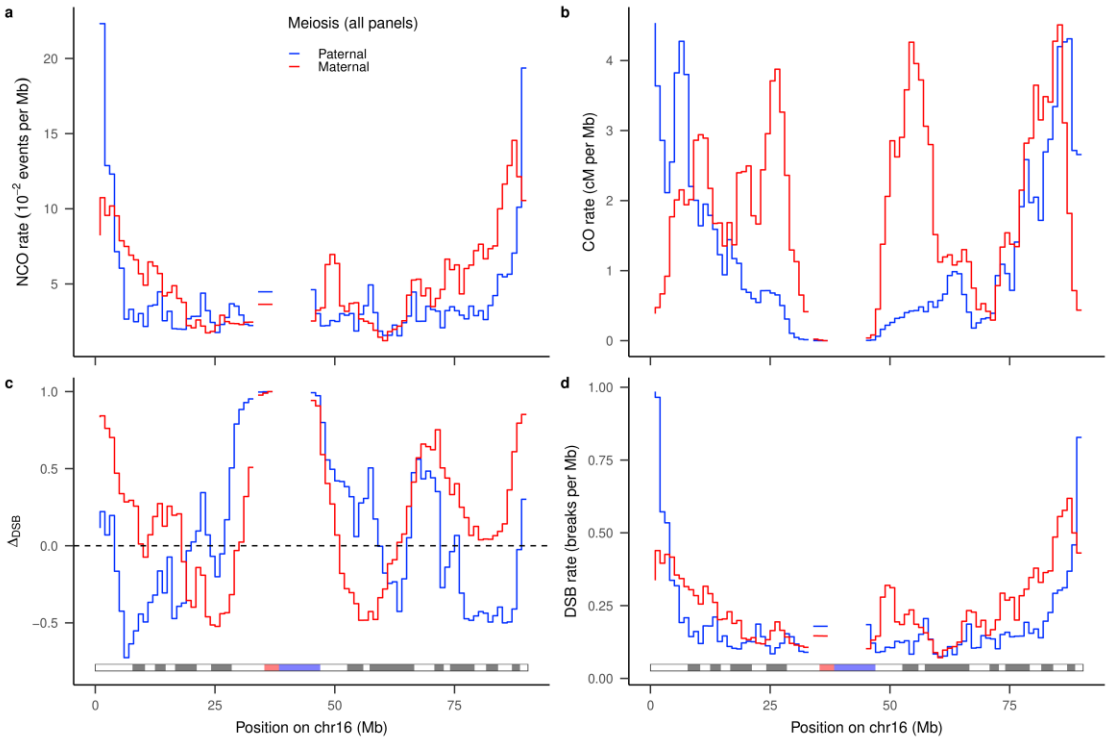
n) Recombination maps – chromosome 14



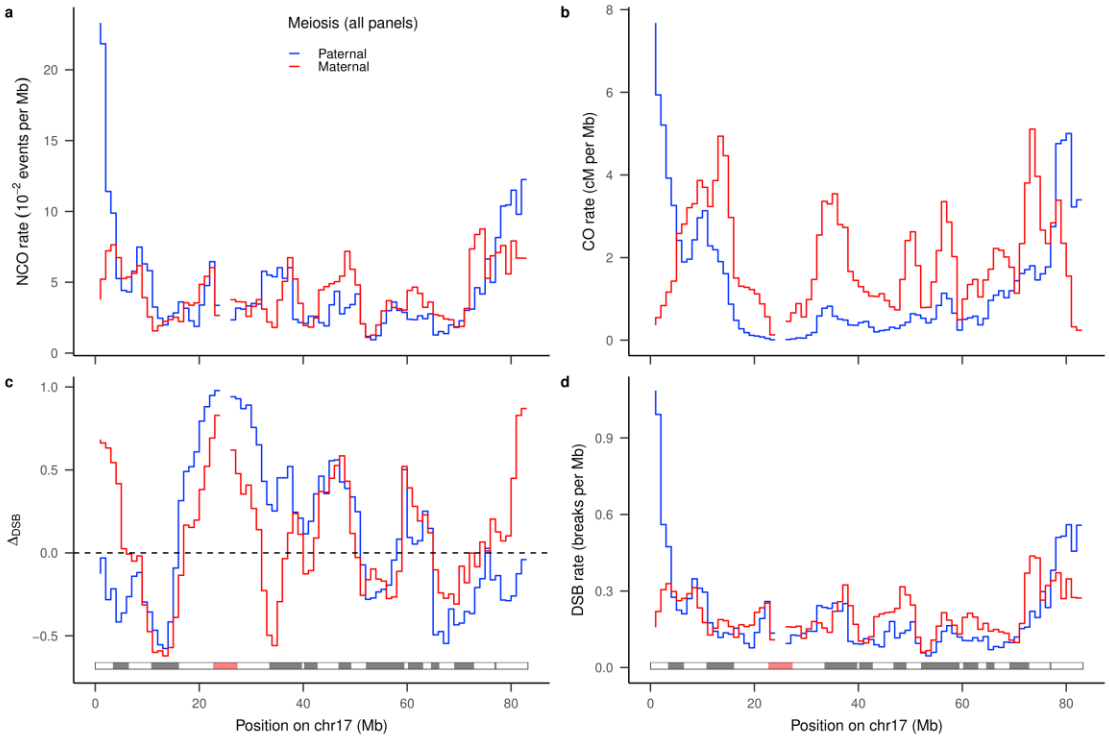
o) Recombination maps – chromosome 15



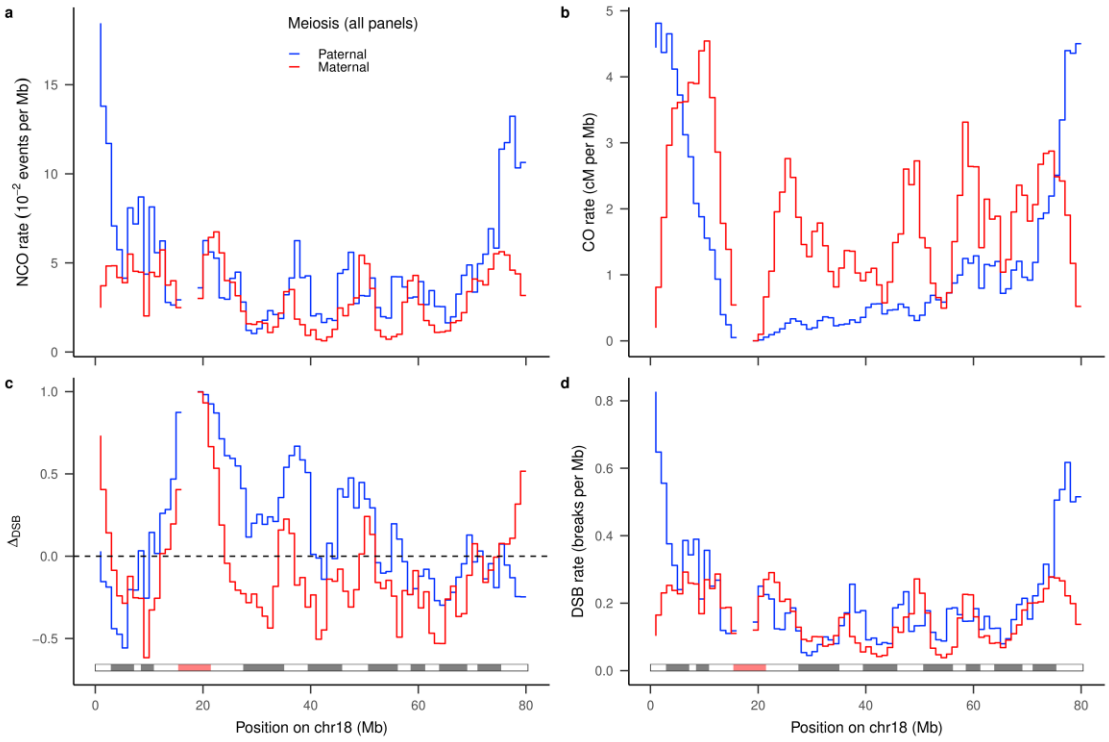
p) Recombination maps – chromosome 16



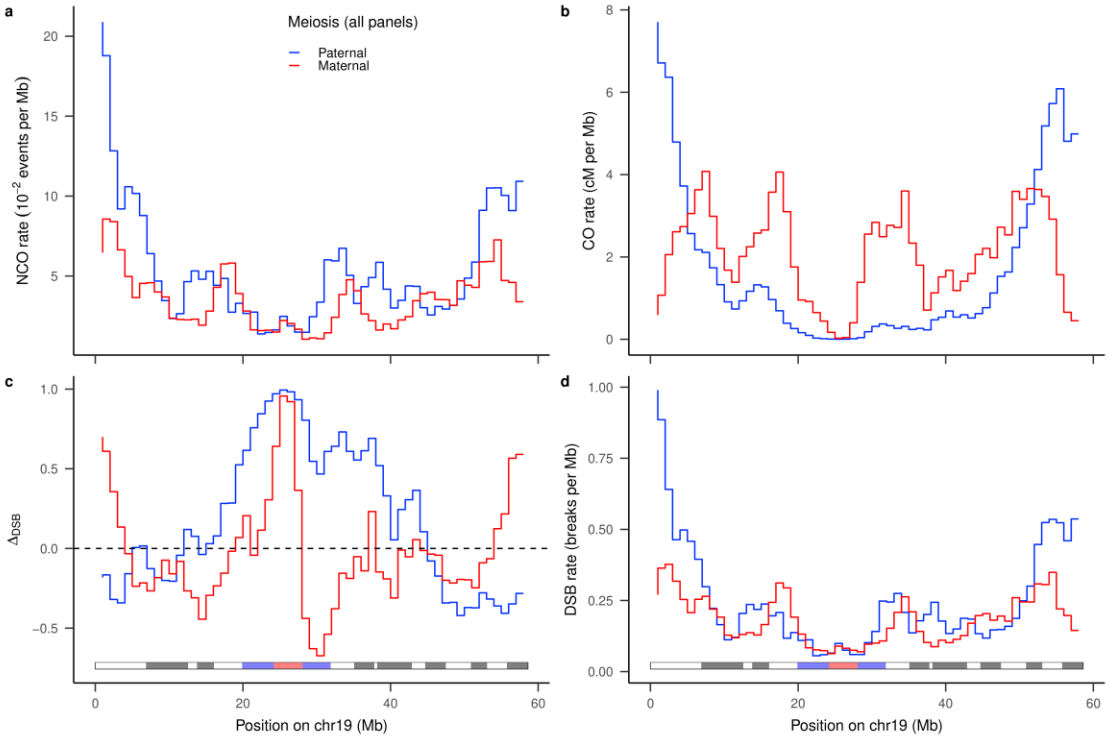
q) Recombination maps – chromosome 17



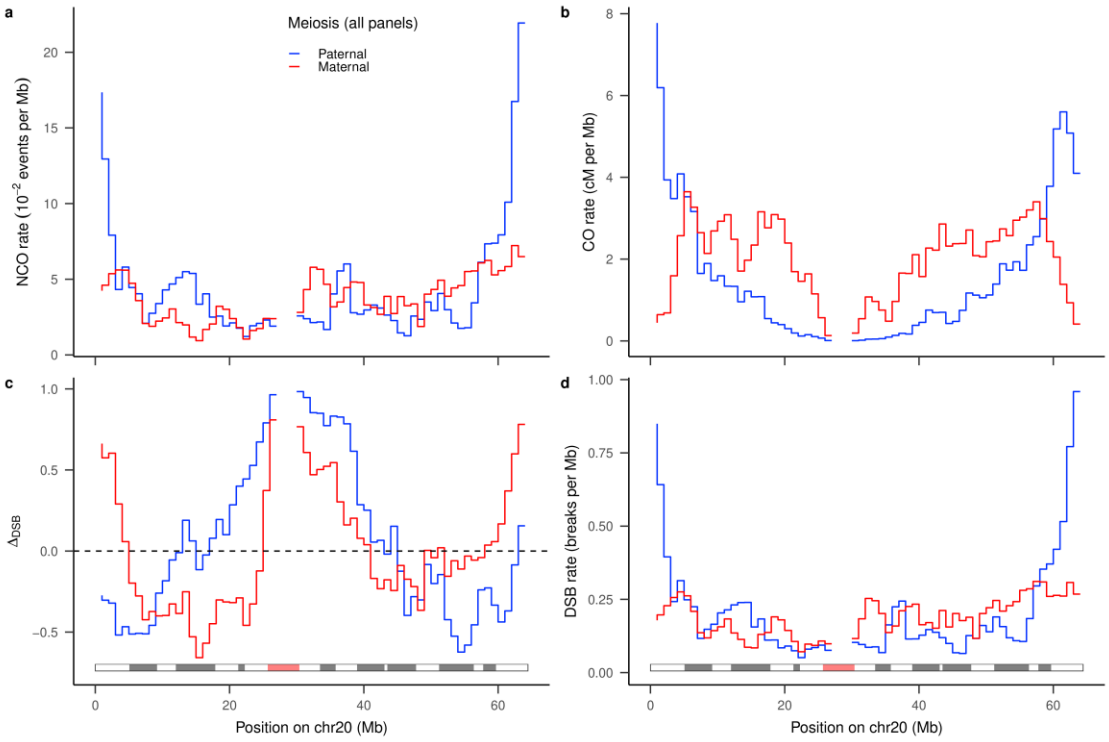
r) Recombination maps – chromosome 18



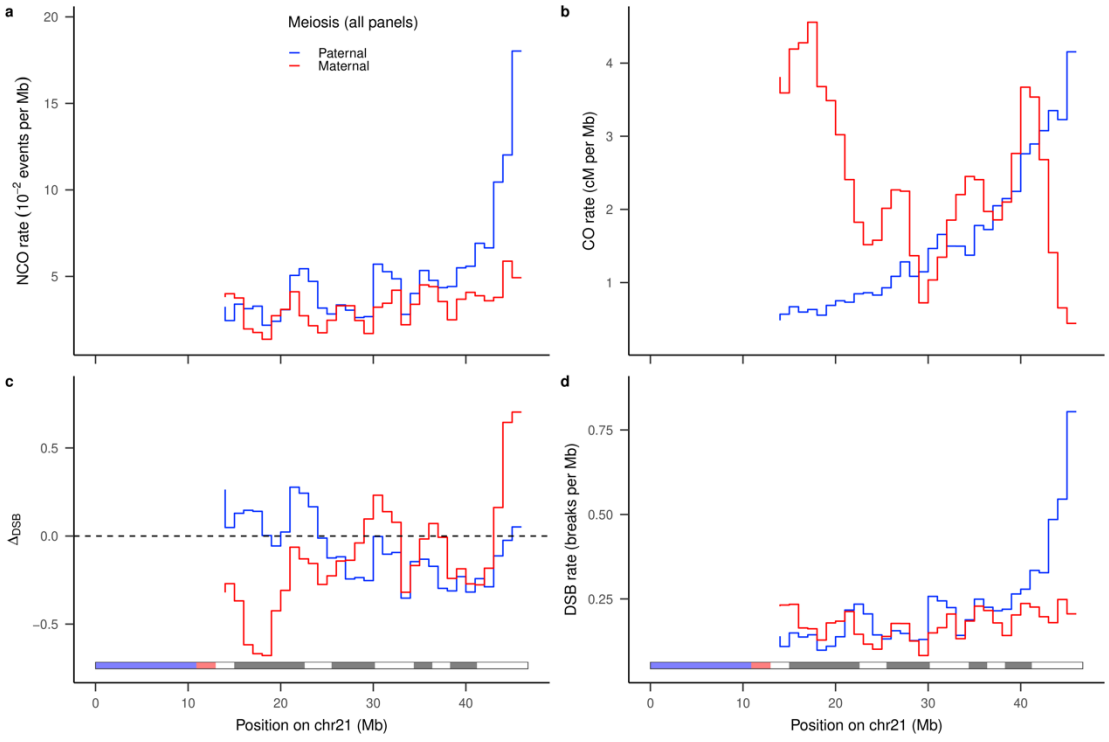
s) Recombination maps – chromosome 19



t) Recombination maps – chromosome 20



u) Recombination maps – chromosome 21



v) Recombination maps – chromosome 22

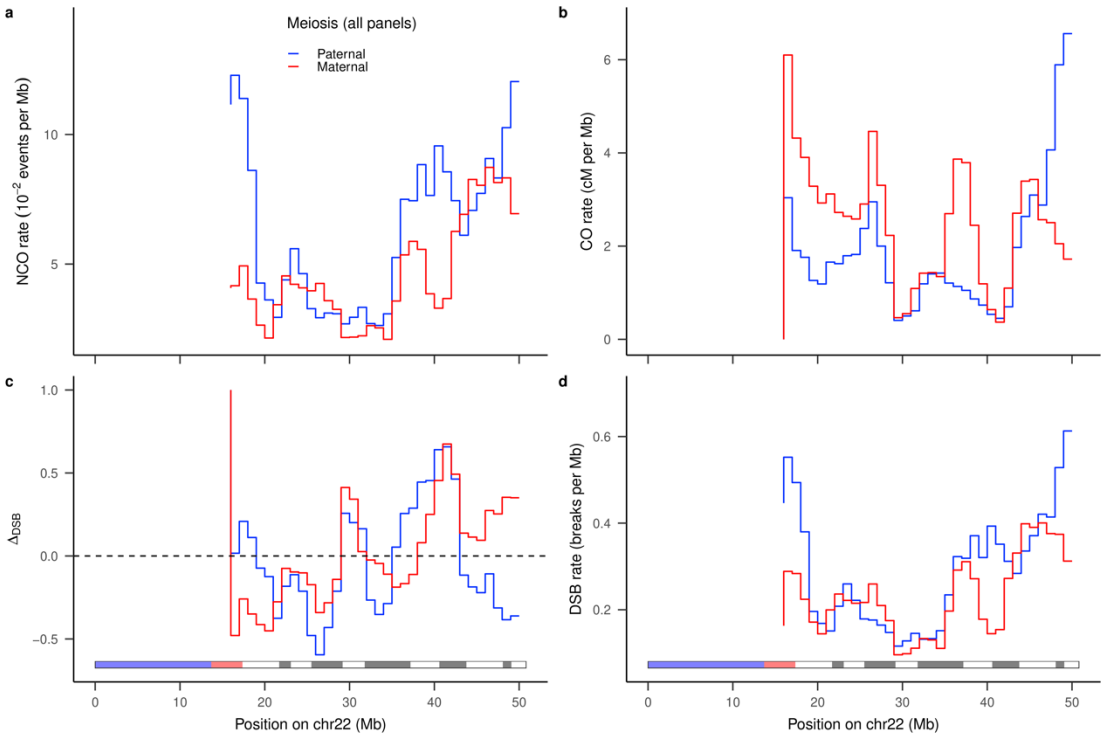
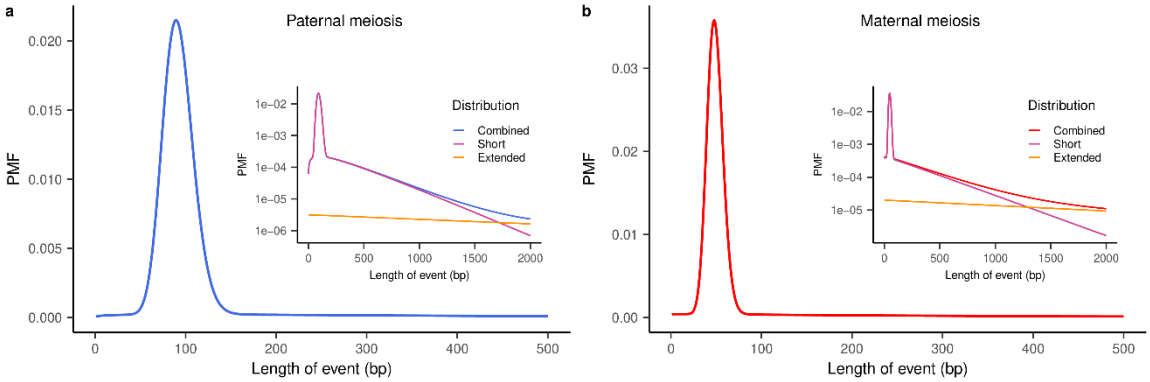
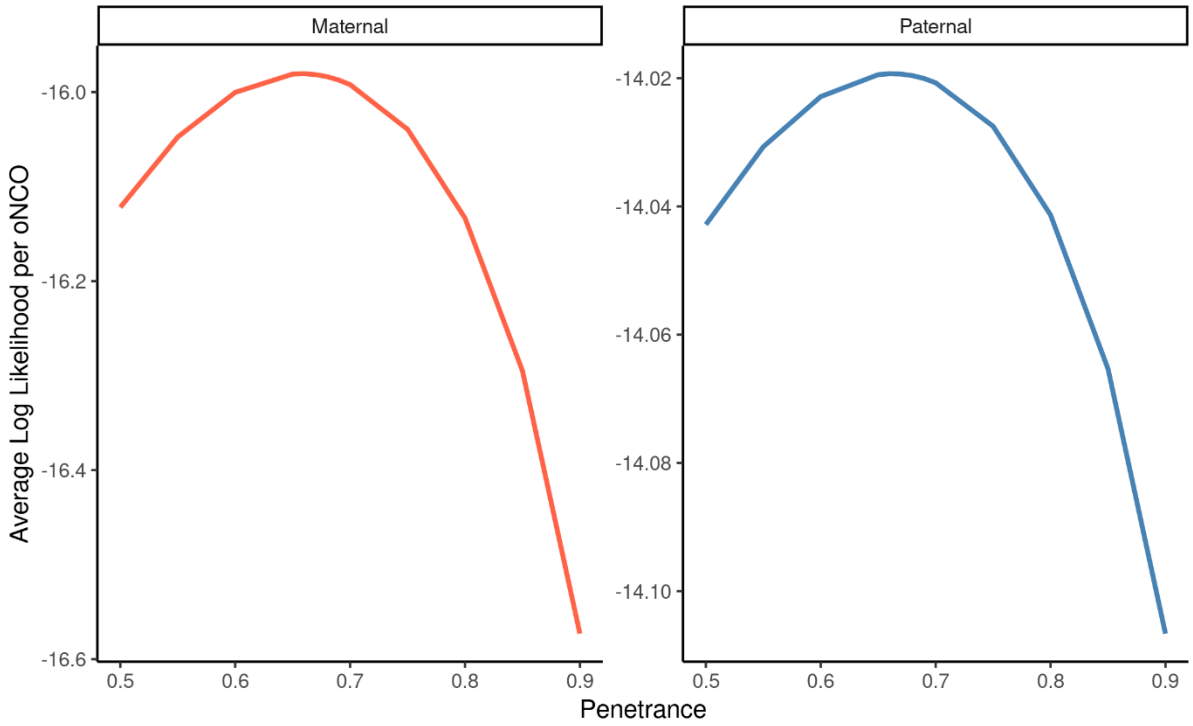

a
Meiosis (all panels)
Paternal
Maternal
NCO rate (10^{-2} events per Mb)
10
5
b
CO rate (cM per Mb)
6
4
2
0
c
Δ_{DSB}
1.0
0.5
0.0
−0.5
0
10
20
30
40
50
Position on chr22 (Mb)
d
DSB rate (breaks per Mb)
0.6
0.4
0.2
0
10
20
30
40
50
Position on chr22 (Mb)

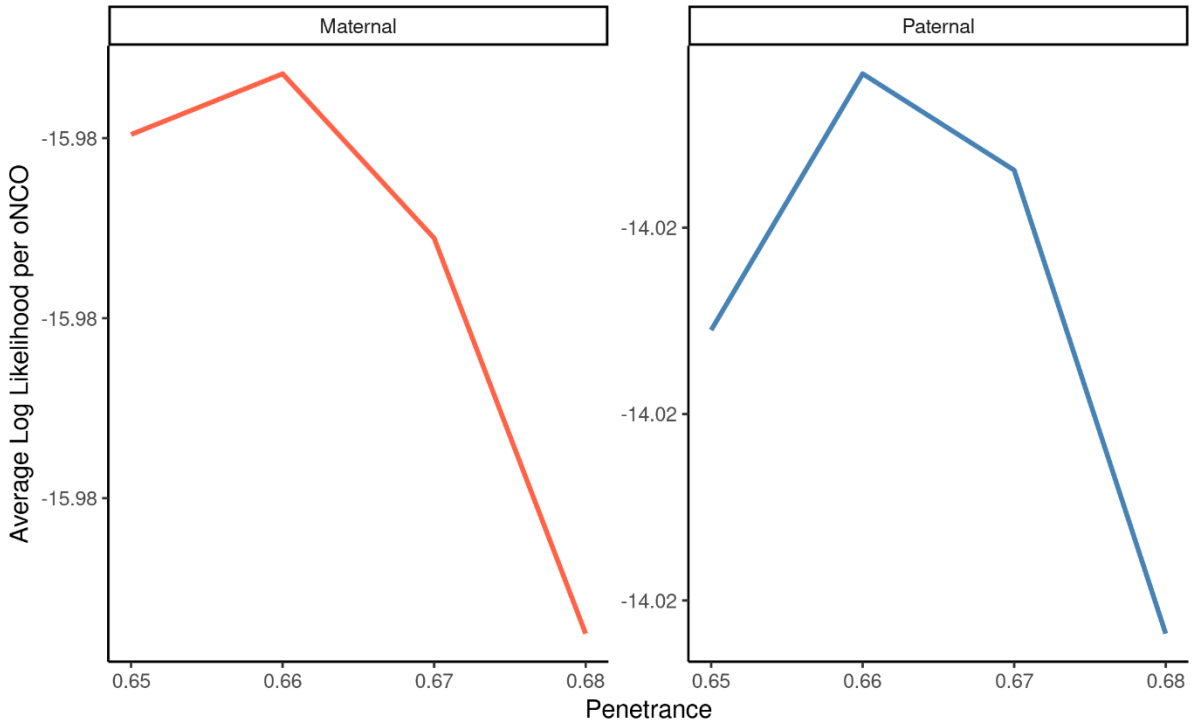
Fig. S2. NCO LENGTH DISTRIBUTION



Length distributions for NCOs. a| paternal NCOs, b| maternal NCOs. The inserts depict the distributions on a log-scale and include the distributions for the short and the extended NCOs.

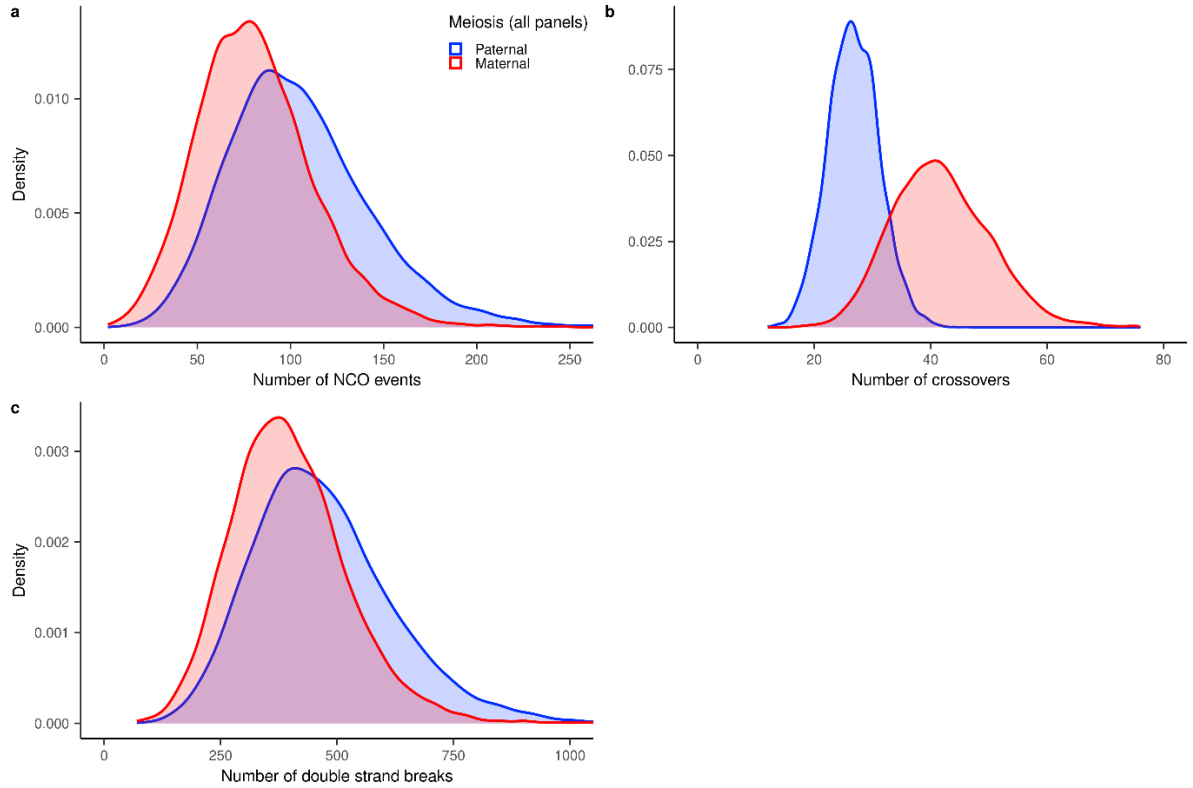
Fig. S3. PENETRANCE AVERAGE LOG LIKELIHOOD PER ONCO





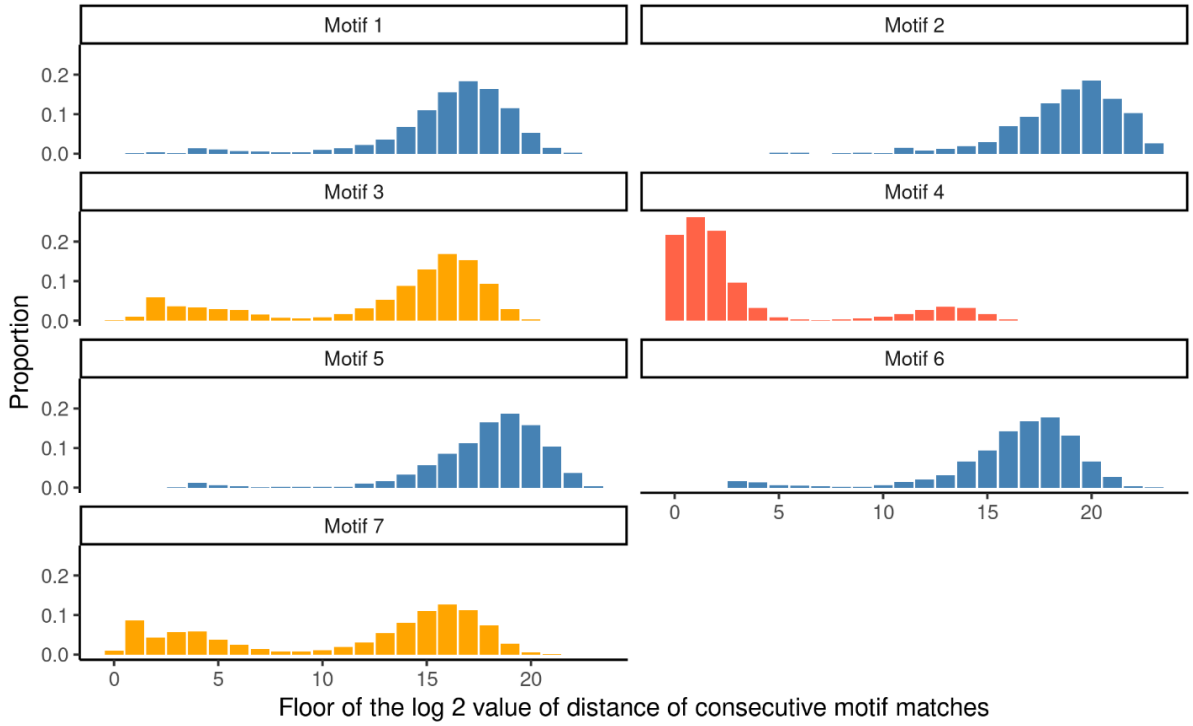
Average Log likelihood per oNCO of models produced by NCOurd[23] for various values of penetrance for maternal (red) and paternal (blue) oNCOs.

Fig. S4. DISTRIBUTION OF RECOMBINATION COUNT ESTIMATES PER MEIOSIS



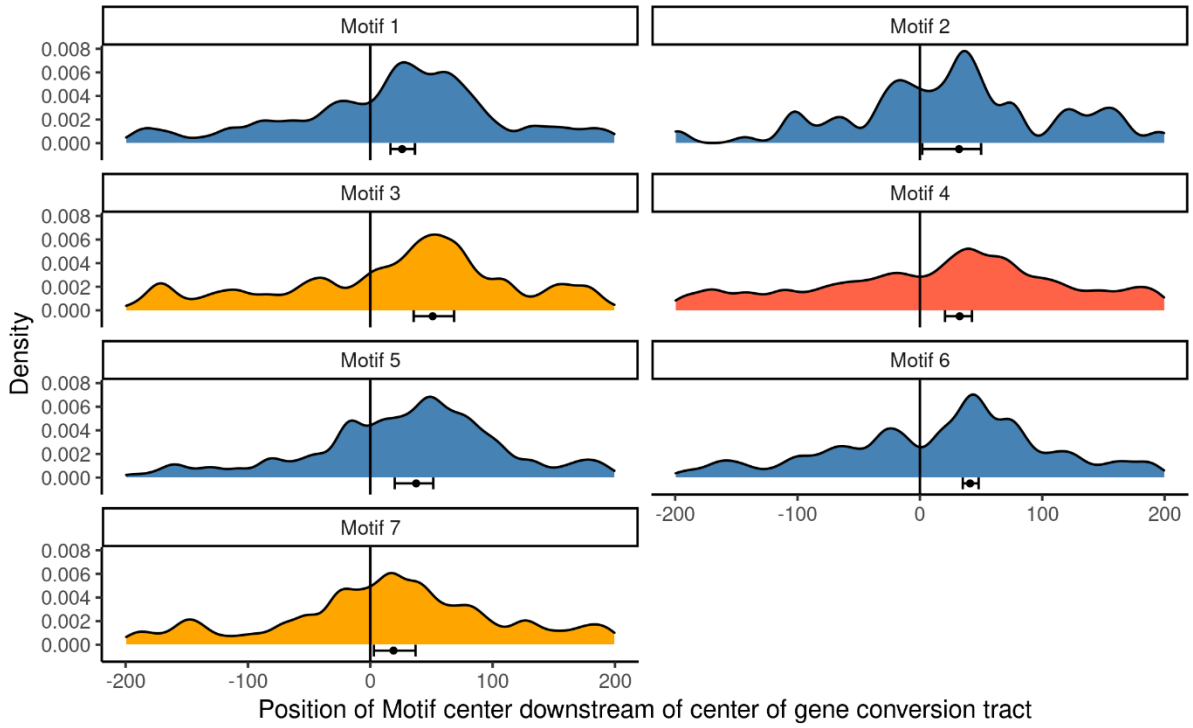
The estimated distributions of the number of recombinations. a| NCOs per offspring, b| COs per offspring, and c| DSBs per meiocyte.

Fig. S5. DISTRIBUTION OF CONSECUTIVE MOTIVE MATCH DISTANCES



The graphs show the histogram of the distance between consecutive matches of PRDM9 motifs as computed for the GRCh38 reference genome. Primary motifs are colored blue. The motifs colored orange and red have multiple close consecutive matches and are less reliable for estimating the distance to oNCO center.

Fig. S6. DISTRIBUTION OF MOTIF MATCHES NEAR oNCO CENTER

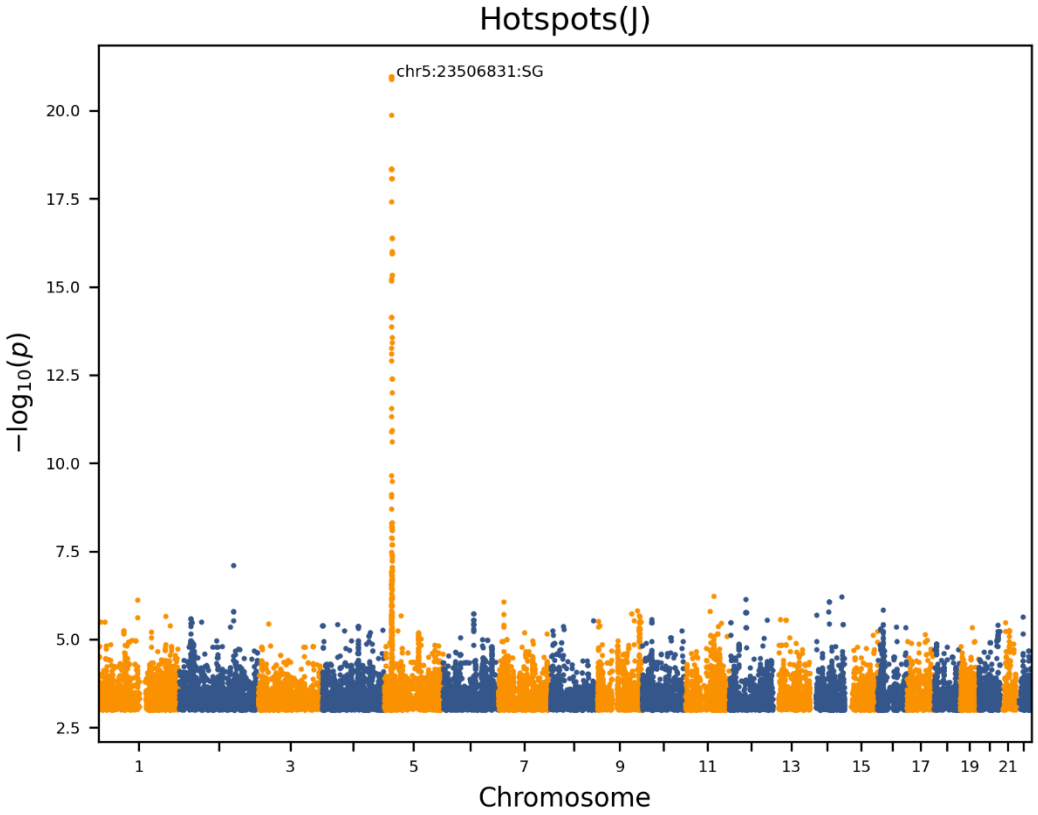


Kernel density plot for locations of motif matches with respect to oNCO center. The density is the average over all data points of a pdf of a Gaussian distribution with standard deviation 9 centered at the data point. Primary motifs are indicated in blue. The orange and red colors indicate motifs that have multiple close consecutive matches. Placement of median with 95% CI are shown below the motif match densities.

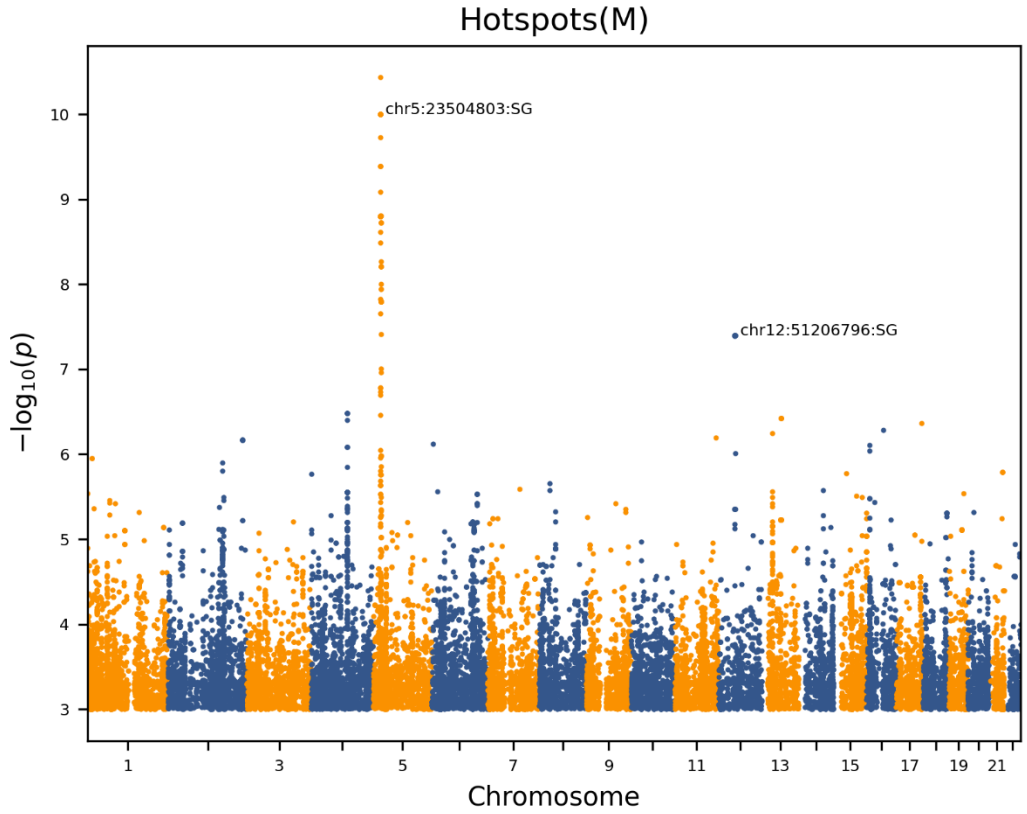
Fig. S7. ASSOCIATION RESULTS: HOTSPOTS

Manhattan plots for hotspot usage phenotype.

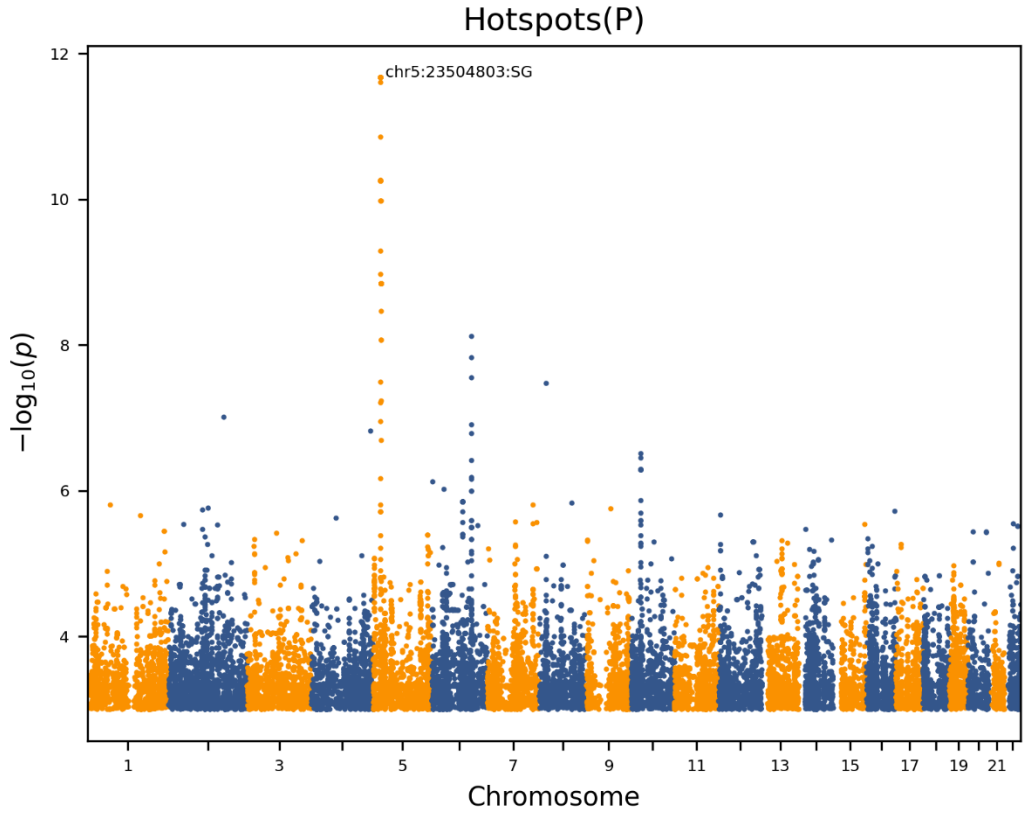
a) Maternal and paternal meioses analyzed jointly



b) Maternal meiosis only



c) Paternal meiosis only



Tables

Table S1. STUDY COHORT INFORMATION

Number of sequenced offspring	Number of families	Total number of meioses
2	1,464	5,856
3	434	2,604
4	114	912
5 or more	120	1,468
Total	**2,132**	**10,840**

The table shows the breakup of the study cohort in terms of family size.

Table S2. SEQUENCE MARKER INFORMATION

							Average number of markers per kilobase		
Chr	Begin	End	Size	SNPs	Indels	Total	All	Paternal MPPs	Maternal MPPs
chr1	898,818	248,920,024	248,021,206	652,030	49,887	701,917	2.830	0.244	0.247
chr2	28,228	241,910,802	241,882,574	708,277	53,383	761,660	3.149	0.279	0.282
chr3	58,666	198,051,828	197,993,162	604,429	45,659	650,088	3.283	0.299	0.303
chr4	68,894	189,873,983	189,805,089	612,778	48,568	661,346	3.484	0.322	0.325
chr5	50,887	181,271,941	181,221,054	547,730	41,962	589,692	3.254	0.294	0.298
chr6	202,452	170,597,405	170,394,953	518,948	41,660	560,608	3.290	0.299	0.303
chr7	55,363	159,331,791	159,276,428	466,665	35,257	501,922	3.151	0.281	0.285
chr8	220,692	145,075,197	144,854,505	469,134	33,056	502,190	3.467	0.311	0.315
chr9	203,822	138,123,971	137,920,149	362,148	25,833	387,981	2.813	0.246	0.250
chr10	97,942	133,622,502	133,524,560	414,885	30,448	445,333	3.335	0.295	0.300
chr11	198,510	135,075,871	134,877,361	403,074	30,095	433,169	3.212	0.286	0.290
chr12	79,119	133,240,522	133,161,403	399,819	30,980	430,799	3.235	0.283	0.287
chr13	18,928,389	114,332,455	95,404,066	310,647	24,618	335,265	3.514	0.324	0.328
chr14	19,868,402	106,880,170	87,011,768	273,220	21,145	294,365	3.383	0.294	0.298
chr15	23,450,901	101,854,166	78,403,265	233,484	17,482	250,966	3.201	0.278	0.283
chr16	40,914	90,056,698	90,015,784	257,370	16,113	273,483	3.038	0.249	0.254
chr17	150,509	83,161,768	83,011,259	212,833	16,024	228,857	2.757	0.215	0.218
chr18	131,903	80,257,805	80,125,902	242,684	18,876	261,560	3.264	0.296	0.301
chr19	259,575	58,583,759	58,324,184	170,878	13,223	184,101	3.157	0.214	0.217
chr20	80,457	64,281,925	64,201,468	189,982	13,470	203,452	3.169	0.268	0.272
chr21	14,108,572	46,672,824	32,564,252	113,237	8,480	121,717	3.738	0.343	0.347
chr22	16,797,575	50,740,328	33,942,753	106,002	7,405	113,407	3.341	0.254	0.258
Total			2,775,937,145	8,270,254	623,624	8,893,878			

The range and number of markers per chromosome in the markerset used for the study. The number of MPPs is the average for all offspring in this study.

Table S3. LIKELIHOOD RATIO TEST RESULTS FOR LENGTH DISTRIBUTIONS

Number of components	Degrees of freedom	P-value – Paternal	P-value - Maternal
1	2	-	-
2	5	$<2.2\cdot10^{-308}$	$<2.2\cdot10^{-308}$
3	8	$8.24\cdot10^{-43}$	$2.20\cdot10^{-102}$
4	11	$2.19\cdot10^{-6}$	$1.34\cdot10^{-11}$
5	14	1.000	$1.81\cdot10^{-10}$
6	17		1.000

Likelihood ratio test comparison of NCOurd solutions for the length distribution of NCOs. The p-values shown for n components indicate the chi-square p-values comparing the solutions with $n-1$ and n components.

Table S4. NCO LENGTH DISTRIBUTION RESULTS

	Component	Percent NCOs	Percent oNCOs	Mean NCO length	Sigma
Paternal NCOs					
Short	1	87.77 (54.12, 95.43)	60.09 (38.44, 70.10)	91 (72, 105)	16.6 (9.6, 49.0)
	2	11.09 (3.59, 38.36)	25.93 (16.32, 34.62)	374 (105, 770)	325 (41, 447)
	Short	**98.86 (94.12, 99.38)**	**86.01 (69.30, 88.85)**	**123 (94, 135)**	
Extended	3	0.89 (0.44, 5.53)	8.42 (6.01, 22.98)	2.9 (0.8, 6.3)	2.9 (0.7, 5.0)
	4	0.24 (0.10, 0.37)	5.57 (3.12, 7.79)	23.2 (17.5, 39.2)	21.6 (17.5, 25.5)
	Extended	**1.14 (0.62, 5.88)**	**13.99 (11.15, 30.70)**	**7.2 (1.8, 11.8)**	
Maternal NCOs					
Short	1	77.22 (71.63, 87.26)	18.65 (14.94, 24.42)	48 (33, 67)	8.8 (6.0, 28.8)
	2	15.89 (6.76, 20.38)	23.18 (14.32, 25.31)	363 (237, 681)	358 (75, 452)
	Short	**93.11 (90.07, 94.90)**	**41.83 (33.86, 44.98)**	**102 (71, 125)**	
Extended	3	5.26 (3.49, 7.73)	30.45 (20.07, 34.36)	2.6 (1.5, 3.2)	2.6 (0.9, 3.1)
	4	1.06 (0.84, 2.27)	16.13 (14.37, 26.70)	17.4 (7.5, 19.6)	7.4 (5.0, 10.3)
	5	0.57 (0.40, 1.27)	11.60 (9.15, 21.29)	53.5 (33.7, 60.8)	23.7 (23.0, 25.5)
	Extended	**6.89 (5.10, 9.93)**	**58.17 (55.02, 66.14)**	**9.1 (6.8, 10.8)**	

Percentage breakup and distribution parameters of the components in the NCO length distributions. The numbers indicate the results computed with the full cohort with 95% confidence intervals shown in parentheses, computed by bootstrapping the length distribution computation (Methods). The length scales shown for Short and Extended events are computed from the length scales of each individual component weighted by the component percentage.

Table S5. ALLELE SELECTION BIAS

	SNPs (GC bias) (%)	Indels (Insertion bias) (%)

	Converted	Unconverted	Converted	Unconverted
By oNCO length and SNP conversion type				
Paternal meiosis				
All oNCOs	61.9 (60.6-63.3)	47.6 (44.2-51.9)	44.9 (42.1-47.8)	48.8 (40.3-58.6)
Short	63.2 (62.3-64.2)	52.0 (47.3-56.3)	45.3 (42.4-48.4)	59.8 (48.1-72.7)
Extended	56.5 (53.0-60.8)	43.9 (39.8-49.4)	43.4 (38.6-47.7)	41.3 (31.6-52.8)
Transitions	61.9 (60.4, 63.3)	47.2 (43.7, 51.7)		
Transversions	62.3 (60.4, 64.3)	49.8 (42.9, 57.1)		
Maternal meiosis				
All oNCOs	65.5 (64.7-66.2)	68.0 (66.6-69.5)	62.5 (60.4-64.4)	74.6 (70.9-77.9)
Short	65.6 (64.8-66.4)	67.2 (65.2-69.1)	60.4 (57.9-63.0)	72.7 (67.8-77.7)
Extended	65.4 (64.4-66.3)	68.3 (66.6-70.1)	63.8 (61.5-65.8)	75.3 (71.1-78.9)
Transitions	65.1 (64.2, 65.9)	67.9 (66.3, 69.6)		
Transversions	67.1 (65.8, 68.5)	68.3 (66.0, 70.6)		
By haplotype segment size in MPPs				
Paternal meiosis				
1	65.3 (64.3-66.3)	52.3 (44.4-61.7)	42.6 (39.4-45.8)	66.7 (38.5-91.7)
2	60.9 (58.4-63.4)	50.0 (40.4-60.9)	45.5 (36.5-55.4)	73.3 (50.0-93.8)
3	50.2 (44.8-55.4)	49.5 (38.3-61.4)	63.3 (45.2-78.6)	46.2 (23.5-100.0)
4	48.8 (41.6-56.1)	47.7 (36.7-60.4)	68.2 (47.6-87.1)	40.0 (10.5-88.9)
5	48.6 (41.1-55.4)	44.1 (34.2-60.9)	53.8 (25.0-81.8)	50.0 (0.0-100.0)
6	54.3 (43.8-65.5)	44.0 (33.3-60.4)	54.5 (25.0-83.3)	33.3 (0.0-100.0)
7	57.9 (48.7-67.1)	55.4 (43.2-73.5)	37.5 (0.0-80.0)	80.0 (33.3-100.0)
8	42.3 (31.8-52.6)	41.2 (20.0-57.1)	60.0 (0.0-100.0)	50.0 (0.0-100.0)
9	54.9 (40.0-76.2)	41.4 (31.0-52.0)	20.0 (0.0-100.0)	50.0 (0.0-100.0)
10	49.3 (39.2-59.5)	35.7 (28.6-42.9)	50.0 (0.0-100.0)	100.0 (100.0-100.0)
11-20	51.8 (46.7-56.4)	48.0 (41.3-54.3)	53.8 (37.5-69.1)	32.1 (13.6-50.0)
21-50	51.1 (43.0-60.2)	42.2 (33.8-51.5)	48.5 (29.2-65.6)	40.0 (18.2-61.5)
51-100	43.4 (33.3-53.2)	54.9 (54.9-54.9)	52.6 (40.0-100.0)	50.0 (50.0-50.0)
Maternal meiosis				
1	73.2 (72.3-74.0)	81.5 (79.3-83.7)	57.6 (53.9-61.0)	85.3 (76.4-93.7)
2	68.5 (66.9-70.0)	75.2 (72.9-77.4)	70.8 (65.4-75.8)	75.5 (65.5-83.9)
3	68.0 (65.8-69.9)	74.8 (72.0-77.3)	68.4 (61.5-74.7)	85.5 (76.6-93.1)
4	69.1 (66.7-71.3)	71.3 (68.3-74.2)	76.6 (68.7-83.7)	76.8 (66.7-86.1)
5	66.7 (63.8-69.4)	65.3 (61.8-69.0)	68.5 (60.9-75.9)	81.2 (71.9-89.7)
6	65.3 (62.4-68.1)	65.9 (62.1-69.5)	78.8 (69.7-87.1)	75.0 (62.7-85.7)
7	66.7 (63.3-70.0)	63.0 (58.3-67.7)	71.2 (60.7-81.9)	67.9 (53.0-82.4)
8	61.1 (57.0-64.9)	64.5 (58.9-69.8)	77.5 (68.7-86.4)	76.9 (63.8-88.9)
9	61.3 (57.4-65.2)	65.9 (59.3-71.8)	67.3 (53.6-81.2)	79.3 (65.9-92.3)
10	60.8 (56.6-64.8)	67.2 (58.8-74.7)	67.5 (55.0-80.5)	90.9 (66.7-100.0)
11-20	57.7 (55.8-59.6)	60.4 (57.7-63.3)	61.8 (54.8-68.8)	67.4 (58.1-77.4)
21-50	53.7 (51.8-55.5)	51.9 (47.5-55.7)	45.9 (39.8-52.3)	57.1 (46.5-68.9)
51-100	51.9 (49.1-54.7)	50.4 (47.0-54.1)	52.6 (41.4-62.1)	47.8 (33.3-57.1)

Allele selection bias for different oNCO lengths and for different haplotype segment sizes within oNCOs. The table shows the bias computed for the full cohort with 95% confidence intervals shown in parentheses, computed by bootstrapping (Methods). Segment lengths are measured in number of MPPs. The table cells are colored based on the bias. The cell color is amber if there is significant positive bias (>50%), bluish if there is significant negative bias (<50%), and gray where there is no bias or the bias is not significant. The Bonferroni corrected threshold for statistical significance is $0.05/182 = 2.7 \cdot 10^{-4}$.

Table S6. PEARSON CORRELATION BETWEEN CO AND NCO MAPS

Chromosome	Paternal	Maternal
chr1	0.57 (0.44,0.70)	0.39 (0.29,0.48)
chr2	0.76 (0.63,0.83)	0.27 (0.13,0.40)
chr3	0.66 (0.56,0.75)	0.62 (0.53,0.70)
chr4	0.62 (0.54,0.82)	0.68 (0.53,0.78)
chr5	0.82 (0.72,0.88)	0.36 (0.15,0.56)
chr6	0.69 (0.38,0.90)	0.23 (-0.01,0.55)
chr7	0.55 (0.41,0.68)	0.24 (0.08,0.41)
chr8	0.64 (0.54,0.77)	0.41 (0.27,0.55)
chr9	0.69 (0.55,0.78)	0.38 (0.25,0.52)
chr10	0.69 (0.60,0.79)	0.57 (0.46,0.67)
chr11	0.77 (0.63,0.87)	0.53 (0.40,0.63)
chr12	0.69 (0.57,0.78)	0.44 (0.27,0.57)
chr13	0.69 (0.52,0.80)	0.32 (0.17,0.44)
chr14	0.47 (0.29,0.63)	0.26 (0.14,0.42)
chr15	0.57 (0.32,0.73)	0.25 (0.10,0.41)
chr16	0.52 (0.36,0.66)	0.17 (-0.03,0.37)
chr17	0.84 (0.69,0.91)	0.11 (-0.15,0.35)
chr18	0.79 (0.68,0.87)	0.41 (0.23,0.58)
chr19	0.89 (0.83,0.93)	0.40 (0.17,0.63)
chr20	0.77 (0.64,0.89)	-0.01 (-0.23,0.22)
chr21	0.77 (0.63,0.86)	-0.01 (-0.26,0.32)
chr22	0.61 (0.30,0.80)	0.45 (0.30,0.62)
Genome-wide	0.68 (0.65,0.71)	0.36 (0.32,0.40)

The table shows the Pearson correlation (r) between the NCO and CO recombination maps. The numbers shown indicate the correlation for the map for the full cohort with 95% confidence limits shown in parentheses, computed by bootstrapping 1000 samples (Methods).

Table S7. DISTANCE STATISTICS FOR MOTIF NEAR ONCO DISTRIBUTION

Motif	Number of oNCOs	Median (95% CI)
Primary motifs		
Combined	2437	36.0 (26.5, 41.0)
Motif 1	1000	26.0 (16.5, 36.5)
Motif 2	173	32.0 (2.0, 50.0)
Motif 5	382	37.5 (20.0, 51.5)
Motif 6	882	41.0 (35.0, 48.0)
Secondary motifs		
Motif 3	1260	51.0 (35.5, 68.5)
Motif 4	5997	32.5 (20.5, 42.5)
Motif 7	1286	19.0 (3.0, 37.0)

The table shows the number of oNCO having a motif match for each motif and median placement of motif match centers downstream of the oNCO centers with 95% confidence intervals shown in parentheses (Supplementary note 0).

Table S8. PERMUTATION TEST FOR DNM RATES

Distance bin	Genome-wide	In C>G enriched region	Other region
Paternal			
0kb - 1kb	1.31 (0.00,5.45)	1.11 (0.00,16.23)	1.35 (0.00,5.79)
1kb - 3kb	1.12 (0.00,3.72)	1.08 (0.00,9.48)	1.12 (0.00,4.02)
3kb - 40kb	1.05 (0.64,1.52)	1.09 (0.00,2.65)	1.05 (0.60,1.55)
40kb - 100kb	1.03 (0.70,1.36)	1.03 (0.18,2.24)	1.02 (0.68,1.37)
Maternal			
0 - 1kb	1.18 (0.00,8.55)	0.88 (0.00,10.91)	1.28 (0.00,10.38)
1kb - 3kb	1.05 (0.00,5.16)	0.99 (0.00,8.96)	1.07 (0.00,6.82)
3kb - 40kb	0.98 (0.54,1.65)	0.96 (0.17,2.56)	0.98 (0.51,1.73)
40kb - 100kb	0.97 (0.61,1.47)	0.97 (0.28,2.19)	0.96 (0.59,1.50)

Results of permutation test for rate elevation near oNCO. The table shows mean mutation rate elevation with 95% confidence intervals when DNMs are randomly permuted among the probands with the constraint that a DNM must not be assigned to a sibling. In the few cases where that happens, we discard the DNM. The mean value and confidence intervals are computed from 1000 different permutation samples.

Table S9. DNM RATE ELEVATIONS NEAR NCOS

	All NCOs	Extended NCOs	Short NCOs	COs[1]
Paternal - autosomes				
Enrichment within 0 to 1 kb	142.32 (105.77-183.20)	105.34 (70.41-147.17)	142.38 (105.83-183.28)	41.5 (33.2, 52.0))
Enrichment within 1 to 3 kb	10.38 (3.70-19.13)	12.29 (4.46-21.61)	10.38 (3.70-19.12)	6.91 (4.76, 10.1)
Enrichment within 3 to 40 kb	1.91 (1.21-2.71)	1.99 (1.18-3.03)	1.91 (1.21-2.71)	1.05 (0.82, 1.35)
Enrichment within 40 to 100 kb	1.28 (0.89-1.70)	1.39 (0.90-1.92)	1.28 (0.89-1.70)	
Maternal – autosomes				
Enrichment within 0 to 1 kb	125.48 (65.74-197.04)	113.89 (64.02-174.52)	126.68 (63.96-203.59)	58.4 (44.0, 77.4)
Enrichment within 1 to 3 kb	48.23 (21.67-78.99)	61.42 (30.32-97.20)	46.87 (19.71-78.84)	11.9 (7.42, 19.2)
Enrichment within 3 to 40 kb	37.01 (29.31-44.92)	38.72 (31.36-46.15)	36.83 (28.86-45.08)	2.21 (1.60, 3.06)
Enrichment within 40 to 100 kb	4.93 (3.32-6.78)	5.31 (3.45-7.62)	4.90 (3.22-6.86)	
Maternal – autosomal within C>G enriched regions				
Enrichment within 0 to 1 kb	119.06 (35.42-228.69)	107.51 (36.58-201.13)	120.45 (33.61-233.97)	
Enrichment within 1 to 3 kb	84.84 (30.68-151.28)	144.57 (64.55-244.07)	77.63 (22.82-145.04)	
Enrichment within 3 to 40 kb	91.39 (72.71-111.82)	93.17 (75.23-111.64)	91.18 (71.25-112.24)	
Enrichment within 40 to 100 kb	12.39 (7.70-18.01)	15.31 (9.24-23.09)	12.04 (7.31-17.90)	
Maternal – autosomal outside C>G enriched regions				
Enrichment within 0 to 1 kb	128.31 (45.74-215.79)	116.03 (48.00-197.95)	129.54 (40.79-220.55)	
Enrichment within 1 to 3 kb	30.59 (1.70-61.63)	16.57 (2.18-35.55)	31.99 (1.63-64.48)	
Enrichment within 3 to 40 kb	10.79 (6.19-16.53)	11.64 (7.25-17.10)	10.71 (5.85-16.78)	
Enrichment within 40 to 100 kb	1.95 (1.02-3.13)	1.58 (0.88-2.47)	1.99 (1.01-3.25)	

Elevation of mutation rates near NCOs and COs[1]. Distances are computed for each DNM to the nearest oNCOs belonging to the same proband as the DNM. To estimate the number of phased DNMs in each distance bin we use a Bayesian approach with a Beta distribution prior where the parameters of the prior correspond to the total number of DNMs expected from each parent for the distance bin in question based on the number of oNCOs considered. We also compute the elevation of mutation rates within and outside the C>G enriched regions[9]. Here we compute the parameters of the prior using region specific mutation rates which are computed from the complete DNM dataset. The numbers shown correspond to the elevation for the complete cohort with 95% confidence intervals shown in parentheses, computed by bootstrapping (Methods).

Table S10. NUMBER OF DNMS NEAR ONCOS

	Distance from oNCO center			
	0 - 1kb	**1kb - 3kb**	**3kb - 40kb**	**40kb - 100kb**
DNMs near paternal oNCOs - number of oNCOs = 12,812				
Paternal origin	50	4	25	23
Maternal origin	1	0	5	8
Unknown origin	22	8	18	25
Total DNMs	**73**	**12**	**48**	**56**
Expected paternal DNMs	0.53	1.06	19.7	31.9
Expected maternal DNMs	0.16	0.32	5.89	9.55
DNMs near maternal oNCOs - number of oNCOs = 15,589				
Maternal origin	18	18	213	63
Paternal origin	4	2	15	18
Unknown origin	12	8	147	53
Total DNMs	**34**	**28**	**375**	**134**
Expected maternal DNMs	0.22	0.43	8.00	13.0
Expected paternal DNMs	0.66	1.33	24.5	39.7
Within C>G enriched regions - number of oNCOs = 2,753				
Maternal origin	7	14	170	52
Paternal origin	1	1	4	6
Unknown origin	2	4	97	33
Total DNMs	**10**	**19**	**271**	**91**
Outside C>G enriched regions - number of oNCOs = 12,836				
Maternal origin	11	4	43	11
Paternal origin	3	1	11	12
Unknown origin	10	4	50	20
Total DNMs	**24**	**9**	**104**	**43**

The number of DNMs found near oNCOs. Distance is measured from the center of the oNCO to the location of the DNM. The expected number of DNMs is computed based on the total number of oNCOs for the whole cohort, using age specific and regional mutation rates for each proband. The rates are computed from linear regression parameters (Table S26). The age corresponds to the age of the parent at birth of proband and whether oNCO is within or outside the C>G enriched area.

Table S11. MUTATION CLASS BREAKUP OF PHASED DNMS

Mutation class	Genome-wide		Within 100kb of NCO		Within 1kb of NCO	
Mutation class	Phased DNMs	Class breakup (%)	Phased DNMs	Class breakup (%)	Phased DNMs	Class breakup (%)
Paternal						
C>A	22933	8.71 (8.59,8.84)	6	5.88 (1.85,10.2)	1	2 (0,6)
C>G	22833	8.67 (8.55,8.8)	9	8.82 (3.77,14.9)	3	6 (0,12.3)
C>T	56119	21.3 (21.1,21.6)	21	20.6 (11.4,29.6)	8	16 (6.45,25.6)
CpG>TpG	44180	16.8 (16.6,17)	21	20.6 (13.1,29.4)	12	24 (13.2,37)
T>A	15985	6.07 (5.97,6.17)	6	5.88 (1.94,10.5)	5	10 (2.38,18.4)
T>C	62756	23.8 (23.6,24.1)	26	25.5 (16.8,35.2)	12	24 (12.5,37.5)
T>G	16833	6.39 (6.3,6.49)	3	2.94 (0,6.25)	1	2 (0,6)
Indel	21592	8.2 (8.07,8.32)	10	9.8 (3.88,16.4)	8	16 (6,28.3)
Maternal						
C>A	4963	7.02 (6.91,7.13)	23	7.37 (4.55,10.4)	0	0 (0,0)
C>G	5337	7.55 (7.41,7.68)	116	37.2 (31.6,43.2)	6	33.3 (14.3,54.5)
C>T	17758	25.1 (24.8,25.4)	70	22.4 (17.7,26.9)	3	16.7 (0,37.5)
CpG>TpG	12383	17.5 (17.3,17.8)	11	3.53 (1.66,5.69)	2	11.1 (0,29.4)
T>A	4047	5.72 (5.63,5.82)	23	7.37 (4.7,10.4)	1	5.56 (0,18.2)
T>C	16173	22.9 (22.6,23.1)	37	11.9 (8.45,15.4)	3	16.7 (0,38.1)
T>G	3376	4.77 (4.69,4.85)	23	7.37 (4.37,10.4)	0	0 (0,0)
Indel	6682	9.45 (9.31,9.61)	9	2.88 (1.18,4.93)	3	16.7 (0,33.3)
Maternal - within C>G enriched region						
C>A	997	8.22 (8.09,8.37)	19	7.82 (4.45,11.5)	0	0 (0,0)
C>G	1925	15.9 (15.5,16.3)	99	40.7 (33.6,48.2)	4	57.1 (20,100)
C>T	2715	22.4 (22.1,22.7)	54	22.2 (16.8,27.4)	2	28.6 (0,66.7)
CpG>TpG	1273	10.5 (10.3,10.7)	4	1.65 (0.37,3.52)	0	0 (0,0)
T>A	807	6.66 (6.52,6.8)	18	7.41 (4.43,10.9)	0	0 (0,0)
T>C	2761	22.8 (22.4,23.1)	26	10.7 (6.94,14.7)	1	14.3 (0,50)
T>G	822	6.78 (6.65,6.92)	21	8.64 (5.15,12.2)	0	0 (0,0)
Indel	823	6.79 (6.68,6.91)	2	0.82 (0,2.17)	0	0 (0,0)
Maternal - outside C>G enriched region						
C>A	3966	6.77 (6.66,6.88)	4	5.8 (1.45,11.8)	0	0 (0,0)
C>G	3412	5.82 (5.73,5.92)	17	24.6 (15.2,33.3)	2	18.2 (0,40)
C>T	15043	25.7 (25.4,26)	16	23.2 (14.1,32.1)	1	9.09 (0,33.3)
CpG>TpG	11110	19 (18.7,19.2)	7	10.1 (3.61,19.3)	2	18.2 (0,50)
T>A	3240	5.53 (5.43,5.63)	5	7.25 (1.45,14.3)	1	9.09 (0,30.8)
T>C	13412	22.9 (22.6,23.2)	11	15.9 (8,25)	2	18.2 (0,50)
T>G	2554	4.36 (4.29,4.44)	2	2.9 (0,7.7)	0	0 (0,0)
Indel	5859	10 (9.84,10.2)	7	10.1 (3.33,17.7)	3	27.3 (0,53.9)

Tally of phased DNMs and breakup by mutation class. For the NCO-proximal data we only include DNMs that are phased to the same parent as the proximal NCO. The numbers shown for class breakup correspond to full cohort data shown in column 2, with 95% confidence intervals shown in parentheses, computed by bootstrapping (Methods).

Table S12. MATERNAL AGE EFFECTS

	Estimates for mothers at age		Change per decade	
	20 years	**40 years**	**effect**	**P-value**
Genome-wide				
oNCOs (n)	2.33 (2.24, 2.42)	3.84 (3.70, 3.97)	0.75 (0.66, 0.85)	$5.68 \cdot 10^{-52}$
Short	0.99 (0.94, 1.03)	1.58 (1.52, 1.65)	0.30 (0.25, 0.34)	$3.19 \cdot 10^{-37}$
Component 1	0.44 (0.42, 0.46)	0.71 (0.68, 0.74)	0.13 (0.11, 0.16)	$1.40 \cdot 10^{-32}$
Component 2	0.55 (0.52, 0.57)	0.88 (0.84, 0.91)	0.17 (0.14, 0.19)	$3.14 \cdot 10^{-38}$
Extended	1.34 (1.29, 1.40)	2.25 (2.17, 2.34)	0.45 (0.39, 0.51)	$6.61 \cdot 10^{-48}$
Component 3	0.71 (0.68, 0.74)	1.16 (1.12, 1.21)	0.23 (0.19, 0.26)	$1.44 \cdot 10^{-41}$
Component 4	0.38 (0.36, 0.40)	0.62 (0.59, 0.65)	0.12 (0.10, 0.14)	$2.88 \cdot 10^{-28}$
Component 5	0.26 (0.24, 0.28)	0.47 (0.44, 0.50)	0.11 (0.08, 0.13)	$8.50 \cdot 10^{-21}$
NCOs (n)	66.6 (61.8, 71.5)	107 (101, 114)	20.3 (15.7, 24.9)	$5.73 \cdot 10^{-7}$
Short	62.1 (57.5, 66.6)	99.7 (93.7, 106)	18.8 (14.5, 23.1)	$6.61 \cdot 10^{-7}$
Component 1	51.4 (47.6, 55.2)	82.6 (77.7, 87.6)	15.6 (12, 19.2)	$6.49 \cdot 10^{-7}$
Component 2	10.7 (9.84, 11.5)	17.1 (16, 18.1)	3.21 (2.44, 3.98)	$1.01 \cdot 10^{-6}$
Extended	4.58 (4.24, 4.92)	7.55 (7.09, 8.01)	1.49 (1.16, 1.81)	$4.30 \cdot 10^{-7}$
Component 3	3.51 (3.25, 3.77)	5.74 (5.39, 6.08)	1.11 (0.864, 1.36)	$4.99 \cdot 10^{-7}$
Component 4	0.71 (0.66, 0.75)	1.15 (1.09, 1.22)	0.22 (0.18, 0.27)	$1.23 \cdot 10^{-7}$
Component 5	0.37 (0.31, 0.43)	0.65 (0.57, 0.72)	0.14 (0.08, 0.20)	$1.58 \cdot 10^{-4}$
Extended NCOs (% of total)	6.87 (6.65, 7.1)	7.04 (6.79, 7.3)	0.09 (−0.11, 0.28)	$3.65 \cdot 10^{-1}$
Crossovers (n)	41.1 (40.7, 41.4)	42.3 (41.8, 42.8)	0.59 (0.22, 0.97)	$1.89 \cdot 10^{-3}$
NCO/CO ratio in meiocyte	3.25 (3.02, 3.48)	5.08 (4.78, 5.38)	0.91 (0.69, 1.13)	$9.99 \cdot 10^{-7}$
Δ_{DSB}	−0.09 (−0.11, −0.06)	0.14 (0.11, 0.17)	0.11 (0.09, 0.13)	$9.20 \cdot 10^{-8}$
Double strand breaks per meiocyte (n)	349 (329, 368)	514 (488, 539)	82.5 (64, 101)	$4.97 \cdot 10^{-7}$
oNCOs in C>G enriched region (%)	13.8 (12.5, 15.0)	24.6 (22.9, 26.3)	5.4 (4.2, 6.7)	$4.53 \cdot 10^{-17}$
oNCOs in DMC1 hotspots	0.68 (0.64, 0.72)	0.69 (0.63, 0.74)	0.003 (−0.04, 0.04)	$8.83 \cdot 10^{-1}$
oNCOs in DMC1 hotspots (%)	29.0 (27.6, 30.4)	15.9 (14.0, 17.8)	−6.5 (−7.9, −5.1)	$<2 \cdot 10^{-16}$
oNCOs in maternal CO hotspots	0.78 (0.74, 0.83)	0.92 (0.86, 0.98)	0.068 (0.022, 0.11)	$3.6 \cdot 10^{-3}$
oNCOs in maternal CO hotspots (%)	32.8 (31.4, 34.3)	22.1 (20.1, 24.1)	−5.4 (−6.9, −3.9)	$6.4 \cdot 10^{-13}$
Inside C>G enriched region				
oNCOs (n)	0.27 (0.24, 0.30)	0.91 (0.86, 0.95)	0.32 (0.28, 0.35)	$2.25 \cdot 10^{-74}$
NCOs (n)	5.94 (4.33, 7.56)	18 (15.6, 20.3)	6.01 (4.36, 7.66)	$4.04 \cdot 10^{-6}$
Extended NCOs (% of total)	7.71 (7.06, 8.36)	8.05 (7.37, 8.73)	0.17 (−0.369, 0.709)	$5.05 \cdot 10^{-1}$
Crossovers (n)	4.05 (3.96, 4.13)	4.32 (4.20, 4.45)	0.14 (0.05, 0.23)	$2.21 \cdot 10^{-3}$
NCO/CO ratio	2.96 (2.23, 3.7)	8.4 (7.33, 9.47)	2.72 (1.98, 3.46)	$3.97 \cdot 10^{-6}$
Δ_{DSB}	−0.10 (−0.16, −0.04)	0.40 (0.35, 0.46)	0.25 (0.20, 0.30)	$7.79 \cdot 10^{-8}$
Double strand breaks (n)	31.8 (25.4, 38.3)	80.5 (70.9, 90)	24.3 (17.7, 30.9)	$3.67 \cdot 10^{-6}$
Outside C>G enriched region				
oNCOs (n)	2.06 (1.98, 2.14)	2.93 (2.81, 3.05)	0.43 (0.35, 0.52)	$1.09 \cdot 10^{-22}$
NCOs (n)	60.9 (56.6, 65.1)	88.7 (83.2, 94.2)	13.9 (9.88, 17.9)	$7.13 \cdot 10^{-6}$
Extended NCOs (% of total)	6.79 (6.57, 7.02)	6.83 (6.57, 7.1)	0.02 (−0.18, 0.22)	$8.19 \cdot 10^{-1}$
Crossovers (n)	37.0 (36.7, 37.4)	38.0 (37.5, 38.4)	0.45 (0.11, 0.80)	$8.96 \cdot 10^{-3}$
NCO/CO ratio	3.29 (3.06, 3.53)	4.68 (4.37, 4.98)	0.69 (0.471, 0.91)	$1.85 \cdot 10^{-5}$
Δ_{DSB}	−0.08 (−0.11, −0.05)	0.09 (0.06, 0.13)	0.09 (0.06, 0.11)	$4.36 \cdot 10^{-6}$
Double strand breaks (n)	317 (300, 334)	431 (409, 453)	56.7 (40.6, 72.8)	$5.85 \cdot 10^{-6}$

Linear regression fits for recombination data vs. mother's age at birth of proband, computed genome-wide as well as inside and outside C>G enriched regions. For the parameters that pertain only to oNCOs or COs the regression is computed using data where each proband is included separately. Other parameters are computed for groups where the cohort is split up based on maternal age with each group spanning two years. The linear regression for those parameters is weighted with the inverse of the standard deviation obtained from bootstrap sampling in each age bin. Confidence intervals and p-values are computed from the linear regression model using Student's t-distribution. The genome-wide numbers for oNCOs and NCOs are computed by summing up the contribution from inside and outside the C>G enriched region. DMC1 annotation indicates the percentage of annotated oNCOs per proband.

Age dependence and confidence intervals for the DMC1 annotation[11] and maternal CO hotspots[1] are computed from a mixed model linear regression implemented in R[68] with the lmerTest package[69].

Table S13. ESTIMATES FOR ONCOS AND GENE CONVERSIONS

	Paternal	Maternal	Joint
Number of meioses	5,420	5,420	10,840
Total informative markers	4,229,340,533	4,288,524,590	8,517,865,123
oNCOs	12,948	15,712	28,660
Short	10,839.0	4,605.2	15,444.3
Extended	2,109.0	11,106.8	13,215.7
Converted SNPs	15,955	43,108	59,063
In short oNCOs	10,859.8	4,637.6	15,497.4
In extended oNCOs	5,095.2	38,470.4	43,565.6
Unconverted SNPs	1,621	13,461	15,082
In short oNCOs	7.2	1.5	8.7
In extended oNCOs	1,613.8	13,459.5	15,073.3
Converted indels	1,154	2,545	3,699
In short oNCOs	803.0	279.9	1,082.9
In extended oNCOs	351.0	2,265.1	2,616.1
Unconverted indels	121	802	923
In short oNCOs	0.0	0.0	0.0
In extended oNCOs	121.0	802.0	923.0
Total converted	17,109	45,653	62,762
Gene conversion rate (MPPs/M)	4.05 (3.83-4.32)	10.6 (10.2-11.2)	7.37 (7.09-7.67)
Converted markers per oNCO	1.32 (1.26-1.4)	2.91 (2.79-3.03)	2.19 (2.12-2.27)
GC bias on SNPs (%)			
Converted markers	61.9 (60.6-63.3)	65.5 (64.7-66.2)	64.5 (63.8-65.2)
In short oNCOs	64.5 (63.5-65.4)	69.3 (68.2-70.3)	65.9 (65.2-66.6)
In extended oNCOs	56.6 (54.1-59.3)	65.0 (64.1-65.8)	64.0 (63.1-64.9)
Unconverted markers	47.6 (44.2-51.9)	68.0 (66.6-69.5)	65.8 (64.2-67.3)
In short oNCOs	-	-	-
In extended oNCOs	47.5 (44.0-51.7)	68.0 (66.6-69.5)	65.8 (64.2-67.3)
Insertion bias on indels (%)			
For converted markers	45.0 (42.1-47.8)	62.5 (60.4-64.4)	57.0 (55.3-58.7)
In short oNCOs	43.4 (40.3-46.4)	49.1 (45.1-53.0)	44.9 (42.3-47.4)
In extended oNCOs	48.5 (44.4-52.9)	64.1 (62.0-66.2)	62.0 (60.1-64.0)
For unconverted markers	48.8 (40.3-58.6)	74.6 (70.9-77.9)	71.2 (67.5-74.6)
In short oNCOs	-	-	-
In extended oNCOs	48.8 (40.3-58.6)	74.6 (70.9-77.9)	71.2 (67.5-74.6)

Summary results for gene conversions and oNCOs showing 95% confidence intervals in parentheses where applicable. Confidence intervals are computed using bootstrapping (Methods). Where we indicate results for both short and extended oNCOs we have used the length distribution results computed with the complete cohort.

Table S14. ESTIMATES FOR ONCOS AND GENE CONVERSIONS – LFM

	Paternal	Maternal	Joint
Number of meiosis	717	717	1,434
Total informative markers	447,473,328	449,630,476	897,103,804
oNCOs	1,444	1,394	2,838
Short	1,275.8	644.1	1,919.0
Extended	168.2	749.9	919.0
Converted SNPs	1,596	4,368	5,964
In short oNCOs	1,305.8	673.7	1,979.5
In extended oNCOs	290.2	3,694.3	3,984.5
Unconverted SNPs	113	1,606	1,719
In short oNCOs	0.0	0.3	0.3
In extended oNCOs	113.0	1,605.7	1,718.7
Converted indels	116	244	360
In short oNCOs	90.5	31.6	122.1
In extended oNCOs	25.5	212.4	237.9
Unconverted indels	12	105	117
In short oNCOs	0.01	0.13	0.14
In extended oNCOs	11.99	104.87	116.86
Total converted	1,712	4,612	6,324
Gene conversions rate (MPPs/M)	3.83 (3.55-4.12)	10.26 (8.75-12.03)	7.05 (6.19-7.92)
Converted markers per oNCO	1.19 (1.15-1.22)	3.31 (2.89-3.77)	2.23 (1.99-2.50)
GC bias on SNPs (%)			
Converted markers	65.7 (63.4-68.0)	65.0 (62.6-67.3)	65.1 (63.2-67.0)
In short oNCOs	66.9 (64.1-69.4)	68.9 (65.9-71.9)	67.5 (65.4-69.5)
In extended oNCOs	60.7 (55.0-66.2)	64.2 (61.5-66.9)	63.9 (61.4-66.7)
Unconverted markers	57.4 (41.9-90.0)	66.4 (62.6-70.7)	65.5 (61.8-69.4)
In short oNCOs	-	-	-
In extended oNCOs	57.4 (41.9-90.0)	66.4 (62.6-70.7)	65.5 (61.8-69.4)
Insertion bias on indels (%)			
Converted markers	53.3 (43.5-62.6)	59.4 (53.5-65.5)	57.3 (51.9-62.3)
In short oNCOs	49.5 (39.1-59.3)	41.0 (28.9-55.0)	47.1 (39.6-54.6)
In extended oNCOs	66.7 (48.6-82.1)	62.2 (55.8-68.7)	62.7 (55.9-69.2)
Unconverted markers	-	73.5 (63.2-83.3)	72.6 (63.8-81.6)
In short oNCOs	-	-	-
In extended oNCOs	-	73.4 (63.0-83.3)	72.6 (63.8-81.6)

Same statistics as in Table S13 using the large-family method (LFM) of our earlier publication[24] and selecting families from the same pool of sequenced individuals as the current study uses. Where we indicate results for both short and extended oNCOs we have used the length distribution results computed with the complete cohort.

Table S15. oNCOs/MARKER DETECTION COMPARISON WITH LFM

	Current study		LFM	
	Paternal	**Maternal**	**Paternal**	**Maternal**
Overlapping oNCOs				
Number of oNCOs	671	788	671	790
Shared markers	787	2277	787	2277
Markers only seen in current study				
Unphased in LFM	1	1		
Untested in LFM	23	155		
Markers only seen in LFM				
Both parents heterozygous			32	743
Phred score too low			40	558
Unresolved markers			1	24
Total number of markers	**811**	**2433**	**860**	**3602**
oNCOs that do not overlap				
Number of oNCOs	1227	1375	772	609
Shared markers but the oNCO is complex in LFM		29		
Markers only seen in current study				
Unphased in LFM	494	913		
Untested in LFM	1087	2527		
Markers only seen in LFM				
Both parents heterozygous			381	513
Phred score too low			458	322
Unresolved markers			37	84
Total number of markers	**1581**	**3469**	**876**	**919**
Breakup of unresolved markers				
Marker quality				
Fail read quality checks			14	18
Fail Phred check in siblings			11	28
Unqualified oNCOs				
Within oNCOs that don't meet quality criteria			8	13
Within overlapping oNCOs			0	10
Haplotype classification mismatch				
Complex crossovers			0	25
Other mismatch			5	20
Total			38	108

Comparison of results from the current study and the large-family method (LFM)[24] performed within a total of 712 meioses, where three-generational data was available. No gene conversions were found within the current study which were disproven with the LFM. Note, that the number of overlapping maternal oNCOs is higher in the LFM as large oNCOs in the current study may overlap two distinct oNCOs in the LFM. A total of 5201 converted markers found in the current study can either not be phased with the LFM or go untested as the transmitted genotype cannot be confirmed. In the LFM 3,193 converted markers are found which are not shared with the current study. Most of those are either cases where both parents are heterozygous or where the Phred score of the markers in the parents,

or the child are too low. A few however are classified as unresolved and the last section of the table indicates the reasons why those markers did not qualify as converted in the current study.

Table S16. COMPARISON OF ESTIMATES IN SMALL AND LARGE FAMILIES

	Paternal	Maternal	Joint
Meioses (n)			
Small	2,928	2,928	5,856
Large	2,492	2,492	4,984
All	5,420	5,420	10,840
Informative markers (n)			
Small	2,150,609,830	2,184,517,298	4,335,127,128
Large	2,078,730,703	2,104,007,292	4,182,737,995
All	4,229,340,533	4,288,524,590	8,517,865,123
oNCOs (n)			
Small	6,441	8,187	14,628
Large	6,507	7,525	14,032
All	12,948	15,712	28,660
Gene-converted markers (n)			
Small	8,816	24,132	32,948
Large	8,293	21,521	29,814
All	17,109	45,653	62,762
Gene conversion rate (per million MPPs)			
Small	4.10 (3.76-4.56)	11.03 (10.30-11.78)	7.59 (7.16-8.02)
Large	3.99 (3.75-4.26)	10.22 (9.60-10.85)	7.12 (6.76-7.53)
All	4.05 (3.83-4.32)	10.6 (10.2-11.2)	7.37 (7.09-7.67)
P-value small vs. large	0.612	0.096	0.124
Markers per oNCO (n)			
Small	1.37 (1.26-1.51)	2.95 (2.77-3.12)	2.25 (2.13-2.37)
Large	1.27 (1.21-1.35)	2.86 (2.71-3.01)	2.12 (2.02-2.23)
All	1.32 (1.26-1.4)	2.91 (2.79-3.03)	2.19 (2.12-2.27)
P-value small vs. large	0.196	0.444	0.118
GC bias of converted SNPs (%)			
Small	61.0 (59.0-63.1)	65.0 (63.9-66.0)	63.9 (62.9-64.9)
Large	62.9 (61.3-64.4)	66.0 (64.9-67.1)	65.1 (64.2-66.0)
All	61.9 (60.6-63.3)	65.5 (64.7-66.2)	64.5 (63.8-65.2)
P-value small vs. large	0.178	0.208	0.074
GC bias of unconverted SNPs (%)			
Small	46.8 (42.6-52.2)	66.9 (64.6-69.2)	64.4 (62.1-66.7)
Large	49.1 (43.5-55.4)	69.3 (67.5-71.0)	67.5 (65.5-69.3)
All	47.6 (44.2-51.9)	68.0 (66.6-69.5)	65.8 (64.2-67.3)
P-value small vs. large	0.572	0.108	0.04
Insertion bias of converted indels (%)			
Small	44.0 (40.2-47.9)	62.2 (59.3-64.8)	56.5 (54.2-58.8)
Large	46.0 (41.6-49.9)	62.8 (59.9-65.7)	57.5 (55.0-60.0)
All	45.0 (42.1-47.8)	62.5 (60.4-64.4)	57.0 (55.3-58.7)
P-value small vs. large	0.472	0.692	0.56
Insertion bias of unconverted indels (%)			
Small	52.2 (42.2-64.9)	72.6 (67.4-77.4)	69.0 (63.9-73.8)
Large	38.7 (21.4-59.3)	77.0 (72.0-81.4)	74.0 (68.7-78.7)
All	48.8 (40.3-58.6)	74.6 (70.9-77.9)	71.2 (67.5-74.6)
P-value small vs. large	0.23	0.192	0.18

The table summarizes the oNCO and gene conversion data for small families (2 children) and large families (more than 2 children). Estimates are shown for gene conversion rates, markers per oNCO, GC bias for SNPs, and insertion bias for indels. Also shown is the p-value of the difference between the estimates for small and large families. Where applicable, 95% confidence intervals are shown in parentheses, computed using 1000 bootstrap samples (Methods).

Table S17. DESCRIPTION OF CHIP TYPES AND SEQUENCING TECHNOLOGY

Chipcode	Count	Chipcode/Equipment	Count
Omni chips		**HumanHap chips**	
omni2.5.41Multi	426	cnv370qv3	302
omni5-4v1B	699	hh1mdv3	567
omniexp24A	1347	hh610qv1	660
omni2.5.4v1B	2458	hh1mv1	727
omniexp12multih	2843	hh300v2	6728
omni2.5.8v1A	4142	cnv340v1	13975
omniexp24B	6693	hh300v1	15866
omni1mqv	11052	**Sequencing equipment**	
omni23medc	12857	HiSeq	2115
omniexp24C	13882	HiSeq X	40131
omniexpr12h11	19861	NovaSeq	22296
omniexpr12h	31873	GAII	277
omniexp24	35168	WPGA	2131

Chip types and sequencing equipment used for genotyping of individuals in the Icelandic cohort. A subset of 1732 individuals are genotyped on both Omni and HumanHap chips. No individuals are sequenced only on GAII, only 19 are sequenced only on WPGA and 956 are sequenced on both HiSeq/HiSeqX and NovaSeq equipment.

Table S18. DATASETS USED TO RESTRICT/EXCLUDE MARKER REGIONS

	Dataset	Comment
1	GiaB tier 1 regions of HG002[89]	Markers restricted to this set.
2	gnomAD-SV[90]	Regions with frequency > 1% removed
3	Eichler SVs[91]	Regions with frequency > 1% removed
4	GiaB Dark Genome[92]	Removed completely
5	SVs in the HG002 annotated set[92]	Removed completely
6	SVs discovered by Beyter et. al. [93]	Regions with frequency > 1% removed

List of genomic datasets used to restrict or exclude markers from consideration in this study. The removed regions are known to either harbor structural variants or be problematic with respect to variant calling of Illumina sequence reads.

Table S19. GENE CONVERSION RATES BY NUMBER OF MPPS PER SEGMENT

Largest segment	oNCOs	Percentage of oNCOs	Converted markers	Percentage of converted markers	Contribution to conversion rate	Cumulative conversion rate
Paternal oNCOs						
1	11,713	90.46	11,746	68.65%	2.78	2.78
2	842	6.5	1,704	9.96%	0.40	3.18
3	153	1.18	477	2.79%	0.11	3.29
4	65	0.5	317	1.85%	0.08	3.37
5	39	0.3	207	1.21%	0.05	3.42
6	21	0.16	136	0.79%	0.03	3.45
7	15	0.12	109	0.64%	0.03	3.48
8	11	0.08	103	0.60%	0.02	3.50
9	6	0.05	52	0.30%	0.01	3.51
10	5	0.04	64	0.37%	0.02	3.53
11-20	49	0.38	827	4.83%	0.20	3.72
21-50	24	0.19	790	4.62%	0.19	3.91
50-100	5	0.04	577	3.37%	0.14	4.05
Total	**12,948**	**100.00**	**17,109**	**100.00%**	**4.05**	
Maternal oNCOs						
1	11,488	73.12	11,612	25.44%	2.71	2.71
2	1,573	10.01	3,413	7.48%	0.80	3.50
3	630	4.01	2,171	4.76%	0.51	4.01
4	424	2.7	2,176	4.77%	0.51	4.52
5	276	1.76	1,786	3.91%	0.42	4.93
6	196	1.25	1,701	3.73%	0.40	5.33
7	164	1.04	1,684	3.69%	0.39	5.72
8	146	0.93	1,717	3.76%	0.40	6.12
9	122	0.78	1,542	3.38%	0.36	6.48
10	82	0.52	1,085	2.38%	0.25	6.74
11-20	383	2.44	7,059	15.46%	1.65	8.38
21-50	181	1.15	6,098	13.36%	1.42	9.80
50-100	38	0.24	2,490	5.45%	0.58	10.38
101-	9	0.06	1,119	2.45%	0.26	10.65
Total	**15,712**	**100.0**	**45,653**	**100.00%**	**10.65**	

Contribution to gene conversion rate based on oNCO segment size. Here oNCOs are grouped based on the largest number of MPPs in a contiguous segment (whether converted or not) within the oNCO.

Table S20. GC-BIAS OF SNPS FLANKING ONCOS

	Paternal				Maternal			
Distance bin (bp)	**CG/AT**	**CG**	**GC bias**	**P-value**	**CG/AT**	**CG**	**GC bias**	**P-value**
]0,10]	59	29	49.2	1.00	64	41	64.1	0.03
]10,30]	220	125	56.8	0.05	162	83	51.2	0.81
]30-100]	828	443	53.5	0.05	577	305	52.9	0.18
]100-300]	1907	974	51.1	0.36	1389	727	52.3	0.09
]300-1000]	4111	2049	49.8	0.85	3595	1928	53.6	$1.4 \cdot 10^{-5}$
]1000-3000]	4740	2388	50.4	0.61	5690	2994	52.6	$8.2 \cdot 10^{-5}$
]3000-10000]	4197	2113	50.3	0.67	6148	3201	52.1	$1.2 \cdot 10^{-3}$

The table shows the GC-bias of the two non-gene-converted SNPs that flank oNCOs, grouped by their distance from the nearest gene-converted marker of the oNCO. Also shown is the p-value of the GC-bias computed with a

*binomial test using the number of SNPs (column **CG**) where C or G is transmitted and the total number of SNPs (column **CG/AG**).*

Table S21. NCOURD PENETRANCE ESTIMATES

A)

Parent type	Non-boundary markers	Gene converted	Penetrance
Paternal	2772	1860	67.10%
Maternal	29221	19843	67.90%
All	21993	21703	67.80%

B)

Penetrance	Paternal Log Likelihood	Maternal Log Likelihood
0.5	-181559.77	-249583.36
0.55	-181402.91	-248431.84
0.6	-181301.76	-247701.05
0.65	-181257.53	-247401.55
0.66	**-181255.76**	**-247396.33**
0.67	-181256.42	-247410.47
0.68	-181259.64	-247444.47
0.69	-181265.46	-247498.82
0.7	-181274.00	-247574.43
0.75	-181361.62	-248306.01
0.8	-181540.51	-249759.56
0.85	-181850.73	-252262.23
0.9	-182384.68	-256566.70

A) Penetrance as the fraction of gene converted of non-boundary oNCOs markers. B) Log likelihood of model produced by NCOurd for various values of penetrance.

Table S22. ESTIMATES FOR NUMBER OF NCOS & DSBS

	Paternal	Maternal
NCOs per gamete	105 (95.9, 125.0)	81.6 (66.7, 103)
Short	104 (94.2, 123.7)	76.0 (61.4, 97.6)
Component 1	92.2 (59.6, 114)	63.05 (50.15, 87.02)
Component 2	11.6 (3.9, 40.7)	13.0 (5.8, 16.8)
Extended	1.2 (0.7, 6.4)	5.6 (4.7, 7.7)
Component 3	0.94 (0.49, 5.96)	4.3 (3.1, 6.0)
Component 4	0.25 (0.12, 0.39)	0.86 (0.76, 1.82)
Component 5	NA	0.46 (0.36, 0.99)
COs per gamete	26.8 (26.7, 26.9)	41.7 (41.5, 42.0)
DSBs per meiocyte	474 (438, 554)	410 (350, 496)
NCO/CO ratio in meiocyte	7.84 (7.16, 9.33)	3.91 (3.20, 4.94)

Estimates for the number of autosomal NCOs per gamete based on the component-wise oNCO count for the whole cohort and the average event detection fraction per proband. The results for the short, extended, and total NCOs are computed as the sum of the event counts for each component. The COs count comes from the same cohort as the oNCOs. The number of DSBs per meiocyte is computed from the NCO and CO counts as $4 \cdot NCO + 2 \cdot CO$*. 95% confidence intervals are shown in parentheses, computed using 1000 bootstrap samples of the length distribution (Supplementary note 0). Same sampling is used for the confidence intervals of the COs.*

Table S23. CHROMHMM ANNOTATIONS

Priority order	ChromHMM annotation	Description	Annotation group
1	TssA	Active TSS (transcription start site)	TssA-TssFlnkU
2	TssFlnkU	Flanking TSS Upstream	TssA-TssFlnkU
3	TssBiv	Bivalent/Poised TSS	Bivalent (Tss,Enh)
4	EnhBiv	Bivalent Enhancer	Bivalent (Tss,Enh)
5	TssFlnkD	Flanking TSS Downstream	TssFlnkD-TssFlnk
6	TssFlnk	Flanking TSS	TssFlnkD-TssFlnk
7	EnhA1	Active Enhancer 1	EnhAG
8	EnhA2	Active Enhancer 2	EnhAG
9	EnhG1	Genic enhancer1	EnhAG
10	EnhG2	Genic enhancer2	EnhAG
11	EnhWk	Weak Enhancer	EnhAG
12	Tx	Strong transcription	Tx
13	TxWk	Weak transcription	TxWk
14	ReprPC	Repressed PolyComb	ReprPC
15	ReprPCWk	Weak Repressed PolyComb	ReprPCWk
16	Het	Heterochromatin	Het
17	ZNF/Rpts	ZNF genes & repeats	ZNFRpts
18	Quies	Quiescent/Low	Quies

The table shows the ChromHMM annotations states[40] and their grouping in the analysis of enrichment of recombination events within regulatory regions of the genome. The priority order indicates how annotation is assigned to regions where merged samples do not agree. Lower priority order items take precedence over those with higher priority order. For testis the samples that were merged were BS01715, BS01718, and BS01719. For ovary the samples that were merged were BS01399, BS01401, BS01402, and BS01403.

Table S24. PHENOTYPE ASSOCIATION RESULTS

Phenotype	P-value	Effect	rsID	Marker	MAF	gene
Hotspots (J)	1.11×10^{-21}	-0.602	rs2973614	chr5:23506831:SG	3.17	*PRDM9*
Hotspots (P)	2.11×10^{-12}	-0.606	rs2973613	chr5:23504803:SG	3.17	*PRDM9*
Hotspots (M)	9.96×10^{-11}	-0.598	rs2973613	chr5:23504803:SG	3.17	*PRDM9*
Hotspots (M)	4.03×10^{-8}	1.742	Rs765589892	chr12:51206796:SG	0.15	*POU6F1*

The table shows GWAS results for the quantitative hotspot usage phenotype (Supplementary note 0) run under an additive model. Here: J = joint phenotype, P = paternal phenotype, M = maternal phenotype.

Table S25. DOWNSTREAM BIAS FOR PRDM9 MOTIF MATCHES

Motifs	Downstream	All	p-value	Bias (95% CI)
Primary motifs				
Joint	1416	2436	$1.1 \cdot 10^{-15}$	58.1% (56.1%, 60.1%)
Paternal	821	1361	$2.6 \cdot 10^{-14}$	60.3% (57.7%, 62.9%)
Maternal	595	1075	$5.0 \cdot 10^{-4}$	55.3% (52.3%, 58.3%)
All motifs				
Joint	5964	10978	$1.3 \cdot 10^{-19}$	54.3% (53.4%, 55.3%)
Paternal	3116	5583	$3.9 \cdot 10^{-18}$	55.8% (54.5%, 57.1%)
Maternal	2848	5395	$4.4 \cdot 10^{-5}$	52.8% (51.4%, 54.1%)

The table shows the count of motif matches near oNCOs indicating separately the number of downstream matches to highlight the downstream bias for motif matches. Confidence intervals (95%) and p-values are computed with a binomial test.

Table S26. AGE DEPENDENT DNMS RATES

	Inside C>G enriched regions		Outside C>G enriched regions	
	Transmitted DNMs at 20-years	**Change (per year)**	**Transmitted DNMs at 20-years**	**Change (per year)**
Paternal	3.45 (3.15,3.75)	0.13 (0.12,0.14)	36.6 (35.7,37.5)	1.38 (1.35,1.4)
Maternal	1.83 (1.55,2.11)	0.13 (0.12,0.14)	11.4 (10.6,12.1)	0.28 (0.26,0.3)

Poisson regression results[9] for the age dependence of DNM rates inside and outside the C>G enriched regions. The numbers within parentheses represent 95% confidence intervals.