nature communications

Article

https://doi.org/10.1038/s41467-025-65248-3

Long-read sequencing of primate testis and human sperm allows identification of recombination events in individuals

Received: 12 May 2025

Accepted: 7 October 2025

Published online: 24 November 2025

Check for updates

Peter Soerud Porsborg[1,11] ✉, Anders Poulsen Charmouh[1,11], Vinod Kumar Singh[2,3], Sofia Boeg Winge[4,5], Christina Hvilsom[6], Carmen Oroperv[2,3], Lasse Thorup Hansen[7], Juliana Andrea Berner[6], Marta Pelizzola[7], Sandra Laurentino[8], Nina Neuhaus[9], Asger Hobolth[7], Thomas Bataillon[1], Søren Besenbacher[1,2,3,12], Kristian Almstrup[4,5,10,12] & Mikkel Heide Schierup[1,12] ✉

Homologous recombination rearranges genetic information during meiosis, creating new combinations of the genome while also introducing mutations, and influencing GC content. Here we report direct detection of recombination events using highly accurate long-read sequencing from testis tissue of 16 individuals across six primate species and three human sperm samples. Based on methylation patterns, we classify sequencing reads as originating from either somatic or germline cells. We identify 2881 crossovers, 2314 simple gene conversions, and 555 complex events, and analyze their chromosomal distribution. Crossovers are more telomeric, showing stronger concordance with recombination maps than gene conversions. Human samples align with a double-strand break map, whereas other species differ, consistent with variation in PRDM9-directed breaks, although the recombination process is otherwise conserved. Gene conversion tracts are short and of similar length across species (mean 22–95 bp), implying that most non-crossover events are undetectable. We observe GC-biased gene conversion for both single and multiple-SNV events, including sites flanking crossovers. We infer longer gene conversion tracts associated with crossovers (318–688 bp) than with non-crossovers. Highly accurate long-read sequencing combined with methylation-based classification of reads to specific cell types provides a powerful way of studying recombination events in single individuals of any mammalian species.

Meiotic recombination is a ubiquitous cellular process, and the pathways involved are highly conserved. Yet, recombination rates can readily be subjected to selection, and recombination hotspots evolve rapidly both within genomes and across species. To gain a better understanding of how and why recombination evolves, we require a deeper mechanistic understanding of the pathways underpinning it, and the ability to quantify differences in recombination rates and patterns at the individual level.

In most species, proper meiotic chromosome segregation relies on programmed double-strand breaks (DSBs), where a minority of breaks are resolved as crossovers (COs) while the majority are subsequently repaired by one or both of the homologous strands

A full list of affiliations appears at the end of the paper. ✉e-mail: pepo@birc.au.dk; mheide@birc.au.dk

leading to two different types of non-crossovers (NCOs)[1]. In humans, a few hundred DSBs are formed in each meiosis primarily at specific hotspots associated with the presence of PRDM9 motifs[2–4]. In males, DSBs are enriched at the telomeres with further telomeric enrichment of DSBs resolved as COs[5]. DSBs are repaired by strand invasion from the homologous chromosome, and this process increases the probability of occurrence of de novo mutations[1,6,7]. In case of a heterozygous position, a mismatch will be induced, which is subsequently repaired by the mismatch repair (MMR) system. If repaired by a homologous sequence with a SNV, this will cause a visible NCO, also known as a gene conversion, and hereafter denoted (GCV). When a heterozygous position contains a weak (A,T) and a strong (C,G) allele, the conversion process is biased towards the strong allele[1,8,9]. Such GC-biased gene conversion (gBGC) is powerful in shaping genome evolution. It causes the equilibrium GC content to be higher in regions with high recombination rates and leads to an overall higher GC content than the expected equilibrium from mutational processes alone[5,10]. It is not yet clear whether CO or GCV events are most important for shaping the GC content.

Human CO patterns have been extensively studied from linkage disequilibrium (LD) patterns[11], pedigrees and direct sperm sequencing[12], with large differences in rates and positions observed between sexes[13], and also among individuals of the same species[14]. Comparatively less is known about GCVs since they are very hard to detect from LD patterns and require very accurate genotyping to detect from pedigrees[15]. Palsson et al.[1] recently reported a fine-scale map of NCOs from more than 9000 phased trios and estimated their tract length of males and females, separately. However, the number of events in each trio is limited, making it hard to quantify the interindividual variation in both CO and GCV events.

Highly accurate long-read sequencing offers a new avenue for detecting both COs and GCVs. The reads are sufficiently long (typically 10-20 kb) to allow phasing of the reads, and transfer between the paternal and maternal haplotypes can be observed directly when sequencing tissue that contains meiotic genomes (Fig. 1a). Because both sperm and testis tissue contain a substantial number of cells that have undergone meiosis, they can be directly informative about all CO and GCV events that occur during meiosis. Here, we use high-coverage HiFi sequencing to directly detect meiotic recombinations (COs and GCVs) in human sperm samples and in testis tissue from humans, chimpanzees, western gorillas, a lar gibbon, a guinea baboon, and a pig-tailed macaque. While different among species, GCVs are estimated to be very short in primates, and more uniformly distributed across the genome than the COs which are associated with much longer GCV tracts and contribute most significantly to the evolution of the GC content of the genome.

Results

Identification of recombination events from sperm samples and testis tissue

We sequenced three human sperm samples purified by density gradient centrifugation and testis tissue obtained from six men, four chimpanzees, three western gorillas, a lar gibbon, a guinea baboon, and a pig-tailed macaque to 26-90X genomic coverage using Pacbio HiFi sequencing (Methods, Table 1). For each sample, we first constructed a high-quality de novo genome assembly (see Methods) covering more than 95% of the genome with N50 contig sizes of 47.4-93.0 Mb across samples (see Supplementary Table 1). We then mapped all long reads back to this assembly, identified all high-confidence single-nucleotide variants (SNVs) and assigned variants to haplotypes, i.e., inferring the full phasing of all variants. Since reads typically have 10-30 SNVs, haplotype switch errors are exceedingly rare (see Methods). On this backbone, we interrogated all mapped reads for the occurrence of "variant shifts" between the haplotypes as a result of recombination events. We first filtered inferred recombination events located in acrocentric chromosomes and in regions of segmental duplication since these are expected to be enriched in interallelic recombination events[16]. After further filtering, we classified detected events into four types: COs, simple GCVs, complex events, and boundary cases (Supplementary Fig. 1, Fig. 1a).

The testis tissue comprises a mixture of somatic cells (e.g., blood, peritubular myoid, Leydig, and Sertoli cells), and germline cells (diploid spermatogonia, tetraploid primary spermatocytes, diploid secondary spermatocytes, haploid spermatids, and spermatozoa). Because we are only interested in recombination events affecting the germline, we developed an approach that uses CpG methylation patterns to identify the cellular origin of each sequencing read. Each sequencing read contains an average of 68.8-127.5 CpG sites, depending on sample read length distributions. From the kinetics information attached to HiFi reads it is possible to infer the methylation status for >80% of the CpG sites with great accuracy. We developed a classification method that, from the observed methylation pattern on a single PacBio HiFi read, estimates the probability of a read being of somatic or germline origin (see Methods). The method was trained on published methylation data from sorted cell types of human spermatogenesis[17] and somatic proxies (blood and neurons[18]). We can reliably classify more than 93% of the HiFi reads (except for sample HT60 (Human Testis, 60–65 year old donor) where only 72% of reads could be classified) to somatic versus germline (Supplementary Table 2, Methods, and Supplementary Note 1.1-1.5). The fraction of reads predicted to be germ cell-specific corresponded well to histological assessment of ongoing spermatogenesis in the samples (Supplementary Fig. 2, Supplementary Table 2). Based on this classification, we restricted the analysis of putative recombination events to the germline reads. This is particularly important for events where only a single SNV change between haplotypes occurs, which can also be caused by recurrent mutations in somatic cells which have higher mutation rates than germline cells[19,20]. Furthermore, it removes potential GCVs and other types of genetic exchange in cells that are not part of the germline.

To gain further confidence in the recombination calls, we manually curated all types of the remaining potential recombination events by visual inspection and obtained a high (>95%) interobserver concordance in the identification of false positives (see Methods). The number of likely false positives removed by this process varied between 7% and 77% (Supplementary Table 3). The curated set of events for each sample is shown in Table 1.

We next estimated whether the number of COs identified agreed with prior expectations. We can estimate how many COs we expect to identify in each sample from the mapped sequencing coverage, the detectability of a CO given that it occurred on the read, and the proportion of the germline reads that are from postmeiotic cells (see Methods). For COs, we have prior expectations of approximately 25 COs per haploid cell since the human male genomic map length is around 25 Morgans[21]. For detectability, we estimated by simulation on each sample the probability that a random CO on a read will cause an SNV pattern as the one we use to identify COs (Fig. 1a). The power of detection for CO events varies with the SNV density and read length, and thus by sample. We estimate across samples, between 18% (inbred baboon) and 66% (highly heterozygous macaque) of the COs should be detectable (Supplementary Fig. 3). Multiplying these estimates and the estimated fraction of postmeiotic cells for the testis tissue (based on prior reports[22–24] we used 75% of the germline reads), we infer that the number of expected COs for both sperm and testis samples (Table 1) are close to the number of COs detected (10-30% differences except for the testis tissue from HT55, see Supplementary Table 4), given a range of uncertainties. Part of this deviation could also be due to individual differences in the genetic map length. In HT55 and gibbon, we speculate that spermatogenesis is less efficient

https://doi.org/10.1038/s41467-025-65248-3

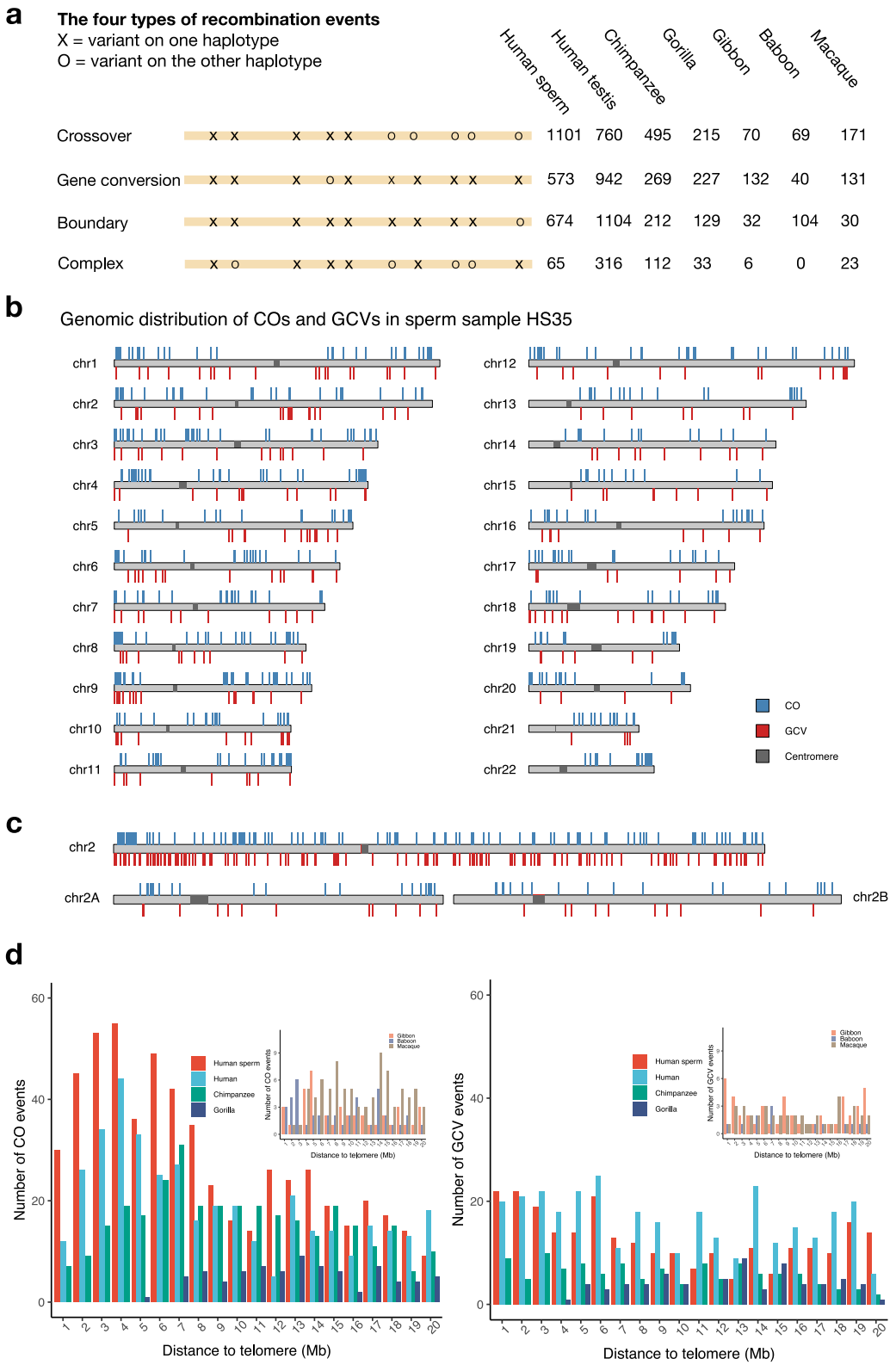

Fig. 1 | Genomic distribution of recombination events. a Polymorphism patterns indicating the different types of recombinations identified. O and X mark variants on the two different haplotypes. The total number of each type of event pooled by species and tissue type is shown to the right. **b** The genomic map (on the T2T-CHM13v2.0 assembly) of the identified crossover (CO; above the chromosome) and gene conversion (GCV; below) events for the human sperm sample HS35 (see Table 1). A similar map of all the other samples is shown in Supplementary Fig. 2. **c** Comparison of the distribution of all called events on chromosome 2 for humans and chromosome 2A and 2B for chimpanzees. **d** The number of COs and GCVs in the 20 Mbp regions closest to the telomere for each species and tissue type.

https://doi.org/10.1038/s41467-025-65248-3

Table 1 | Identification of recombination events in sperm and testis tissue

Name[a]	Species	E(read)	E(cov)	%germline	#SNV	ACRO	SD	CO	GCV	Complex	Boundary
HS25	Human	16362	26	0.993	1975563	17	117	288	173	34	136
HS35	Human	12840	55	0.98	2004809	29	148	501	241	10	378
HS50	Human	17148	39	0.988	1851268	19	107	312	159	21	160
HT20	Human	14302	43	0.535	1932854	2	42	66	68	15	69
HT45	Human	7951	47	0.571	1989078	11	127	194	278	63	232
HT50	Human	14514	43	0.362	1980285	4	93	53	188	153	135
HT55	Human	10616	90	0.26	1972726	2	38	38	74	25	60
HT60	Human	9388	101	0.788	2014968	5	152	345	273	54	449
HT65	Human	15325	44	0.448	1937874	1	53	64	61	6	59
CT15	Chimpanzee	15427	36	0.761	1920758	-	-	104	37	12	52
CT32	Chimpanzee	15118	44	0.789	1873245	-	-	150	65	15	88
CT28	Chimpanzee	15379	28	0.768	3449866	-	-	152	106	76	53
CT22	Chimpanzee	14311	28	0.761	2419571	-	-	89	61	9	19
GT22	Gorilla	10067	38	0.54	4380904	-	-	127	97	19	71
GT43	Gorilla	11015	31	0.463	4312722	-	-	71	92	12	50
GT21	Gorilla	14455	24	0.197	4280505	-	-	17	38	2	8
LT39	Gibbon	14842	25	0.502	4844406	-	-	70	132	6	32
BT15	Baboon	16050	28	0.625	981152	-	-	69	40	0	104
MT5	Macaque	15009	23	0.574	5815378	-	-	171	131	23	30

[a] The first letter denotes the species and the second the type of tissue, with T for Testis and S for Sperm. The numbers denote the age of the individual. For human samples, age is anonymised and rounded to the closest 5-year interval.
E(read) is the mean sequence read length. E(cov) is the mapped mean coverage to the haploid de novo assemblies of each sample before the classification of cell types. % germline is the proportion of reads being classified to be from germline cells by CpG methylation patterns (see methods). The number of events is restricted to germline reads. ACRO is the number of events detected in the p-arm of acrocentric chromosomes. SD is the number of events in regions of segmental duplication in humans. The last four columns are the number of different recombination events identified in germline cells.

than estimated (see also histology, Supplementary Fig. 2), i.e., a smaller proportion of the germ cells than expected are postmeiotic.

Likewise, the boundary cases, where an SNV at the end of a read changes haplotype, are expected to be a mixture of COs and GCVs in proportion to their estimated frequency. We can estimate the expected number of these that are COs and GCVs, respectively, and the sum of these matches well with the observed boundary cases (Supplementary Table 4). Together, our checks suggest that the detected sets of curated COs and GCVs are reliable.

Genomic distributions of crossovers and gene conversions

Next, we investigated the genomic distribution of CO and GCV events across the genome (Fig. 1b, sample HS35, see also Supplementary Fig. 1. The remaining samples are shown in Supplementary Fig. 2). COs seem particularly clustered towards the telomeres, whereas GCVs are more uniformly distributed. The human chromosome 2 originated from a fusion of two chromosomes in the human lineage (chromosomes 2A and 2B in the chimpanzee). When comparing the distribution on human chromosome 2 with chimpanzee chromosome 2A and 2B (Fig. 1c), we observe that the enrichment at the telomeres of the chimpanzee chromosomes is not found in humans at the position where they fused.

We also investigated the density of COs and GCVs across each chromosome after all human, gorilla and chimpanzee samples were pooled (omitting chromosome 2). Whereas the density of COs decreases with increasing chromosome size (Supplementary Fig. 4a, $P < 0.001$, linear regression of density of COs with chromosome size), the density of GCVs is not correlated with chromosome size (Supplementary Fig. 4b, $P = 0.10$). This results in a decreasing CO/GCV with increasing chromosome size (Supplementary Fig. 4c, $P = 0.00077$). Thus, the larger recombination rate per base pair on small chromosomes is due to a higher proportion of the DSBs resolved as COs as also reported by Palsson et al.[1].

Figure 1d shows that the observed number of CO events per Mb decreases rapidly away from the telomere for each type of sample (pooling observations by species and tissue). This effect is much weaker for inferred GCV events. The difference is mainly observable in human testis, human sperm, and chimpanzee testis, likely due to their increased sample size (see Discussion).

Since a male pedigree recombination map also reveals higher recombination rates near telomeres, we related the positions of COs to the recombination map by Halldorsson et al.[6] (smoothed at a 100 kb scale). Figures 2a and 2b show that the positions of the COs we identify are indeed highly enriched in regions with higher estimated recombination, both for human sperm and human testis samples. GCVs are also enriched, but to a much smaller extent, as expected, since COs are the basis for the recombination map.

Since COs and GCVs occur after programmed DNA DSBs, our results suggest that a non-random set of telomere proximal DSBs is resolved as COs. To test this more directly, we lifted a map of DSB hotspots from spermatocytes[2] to the T2T-CHM13v2.0 genome and studied the overlaps with the positions of our detected COs and GCVs. We separately recorded how many DSBs overlapped with the region between the jump SNVs (the polymorphisms flanking a CO event or a tract of a GCV event, Fig. 2c) and also a larger region between flanking SNVs of the CO since we, from the biased GCV results discussed below, suspected that some COs could be initiated outside of the jump SNVs (see Fig. 2c). We find that for most of the human samples, 40-60% of COs and GCVs overlap with a DSB hotspot, leading to an enrichment of around 30-fold compared to a random position. This enrichment is higher when including the flanking SNVs, supporting that some COs are initiated outside of the region marked by the jump SNVs but one SNV is then gene converted in the process (Fig. 2d). For two human samples, the enrichment is notably smaller. For HT50, this is likely because the individual is heterozygous for the PRDM9-A and PRDM9-B allele (all other human samples are homozygous PRDM9-A). For HT5,5

https://doi.org/10.1038/s41467-025-65248-3

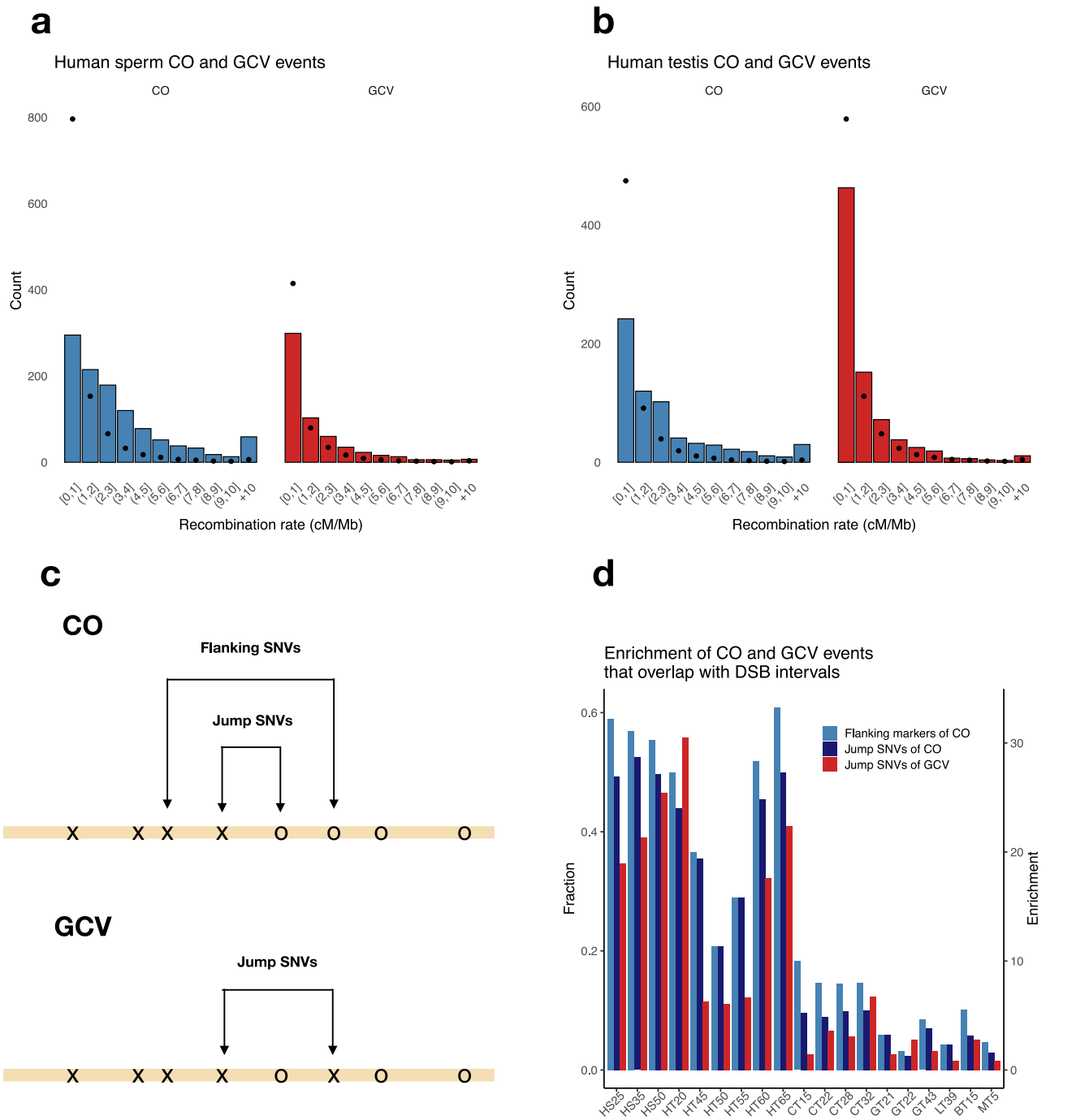

Fig. 2 | Double strand breaks and inferred recombination events. The recombination rate (from deCODE's recombination map[1,6] at 100 kb scale) for observed crossover (CO) and gene conversion (GCV) events pooled by human sperm samples (**a**) and human testis (**b**), respectively. Histograms indicate observed counts, while dots denote expectations under a uniform distribution of events along the genome. **c** Schematic illustration of flanking SNVs and jump SNVs for COs and GCVs. **d** The fraction of COs and GCVs in hotspots of double strand breaks (DSB). The corresponding enrichment factor of COs and GCVs in hotspots of DSB[2] is shown on the right Y-axis.

we ascribe the low enrichment to the poor spermatogenesis of this individual (Supplementary Fig. 2, Supplementary Table 2).

We do not see a similar strong enrichment between the human DSB hotspots and CO and GCV events for non-human testis samples, which is in agreement with the expectation that the DSB hotspots are not conserved between humans and other primates because DSBs are directed by different PRDM9 motifs[25,26]. Interestingly, the called events show greater enrichment in the four chimpanzee samples (Fig. 2d), suggesting a greater overlap between DSBs in humans and chimpanzees than in other more distantly related primates.

The short arms of the acrocentric chromosomes and segmental duplications have previously been shown to often engage in complex interallelic genetic exchanges[16]. In humans, for the acrocentric chromosomes, we identified 17–29 events in sperm and 1–11 in testis samples (Table 1), which are in proportion to the number of COs in the different tissues. They may cause translocations between acrocentric regions of different chromosomes, but this may not be detrimental since their rDNA genes are similarly oriented. However, for segmental duplications, there is an enrichment in the number of events in the testis samples (38–152) compared to the sperm samples

(107–148), as compared to the number of COs (relative proportion 0.33 versus 0.66, Table 1). This suggests that a significant fraction of the events in the segmental duplications in germline cells of the testis are deleterious and do not make it to the mature sperm perhaps because they occur between copies of the duplication, causing chromosomal translocations.

Gene conversion tract lengths and rates

Since GCV tracts are expected to be short and the density of SNVs is low, most NCO events do not move SNVs, and most of the inferred GCV only move one polymorphic site (82-93% across samples, see Supplementary Tables 4 and 5). Thus, instead of averaging the length of events detected, we used all inferred events from each sample to infer the lengths of GCVs using a probabilistic model (see Methods, and Charmouh et al.[27]). Briefly, the likelihood model infers an average tract length for GCVs from the number of GCVs moving successive (1, 2, 3,…) polymorphic sites and the observed distances between heterozygous sites in the genome of the individual investigated. The model assumes that the NCO tract lengths are from a single geometric distribution and estimates its mean (see refs. 15,28,29).

The estimated mean tract lengths are short for all samples, between 14 bp and 142 bp (Fig. 3a, Supplementary Table 6). The estimated tract lengths are significantly different among samples (Supplementary Fig. 5). Within species (considering sperm and testis, separately), there are no significant differences so this effect is due to differences among species. The short estimated tract lengths imply that across the genome, only 2-5% of NCOs move at least one SNV and hence are detectable as GCV events (Fig. 3b, Supplementary Table 6). We combined our estimates of mean tract length, rates of detection, and effective coverage of postmeiotic reads to estimate the overall probability that a bp is part of an NCO event in meiosis (Fig. 3c, Supplementary Table 6). Our estimates of 2-5 bp converted per 1 million bps align with previous estimates[1,15,30,31] and also suggest variation among individuals and species, as reported by Schweiger et al.[32].

These estimates are based on simple GCV events alone. We also observe between 0 and 153 complex events per sample. Complex events switch haplotypes several times (Table 1) and potentially represent more complex GCVs and/or COs with complex GCVs. In either case, they signify much longer tracts and haplotype switching than for the simple GCVs (see also below).

Biased gene conversion

Strand invasion associated with DSB repair will cause heteroduplex formation when the homologous chromosomes differ in the form of SNVs or indels, and these are expected to be repaired towards Cs and Gs causing gBGC, which is a pervasive evolutionary force that shapes GC content in the genome and is the cause of a universal positive correlation between recombination rates and GC content in genomes[9,10]. We examined the degree of GC bias in the different types of recombination events separately. First, for simple GCVs, we investigated the GC bias for events occurring in germline cells separately for single SNV and multiple SNV events since they have previously been shown to differ in GC bias in mice[28]. Overall, the human sperm data show significant gBGC ($P = 0.0012$, binomial test). For GCVs with a single SNV, we find an average gBGC of 54.21%+/−2.38% ($P = 0.042$, binomial test) (Fig. 4a) and a stronger gBGC of 67.35%+/−4.74% ($P = 0.00038$, binomial test) for GCVs with multiple SNVs (Fig. 4b). This is in contrast to the results for mice, where GCVs with multiple SNVs were found to be unbiased (see also Supplementary Figs. 3 and 4).

Next, we tested whether SNVs immediately flanking crossover events also showed signs of gBGC. We find a bias for all samples, with a significant bias for human sperm samples combined of 52.43% + /-1.15% ($P = 0.018$, binomial test) and for human testis samples combined of 52.98% + /-1.39 ($P = 0.017$, binomial test) (Fig. 4c). This suggests that strand invasion tracts associated with COs are longer than those causing the simple NCOs since the distance between SNVs is typically between 1 kb and 2 kb across samples (Supplementary Figs. 6 and 7), which is substantially longer than the estimated NCO tract length above. We calculated rough estimations of the strand invasion tract length associated with COs in the following way (see details in Methods): We assumed that COs were initiated uniformly between jump SNVs in a random direction and then calculated the probability that each GCV associated with the CO with a geometric distribution with mean s would include a jump SNV (see Methods). We then optimised s as a function of an assumed underlying bias (61.9% from Palsson et al.[1] and our empirical estimates (+/− 1 SE) in Fig. 4c). Figure 3d shows estimates pooled by tissue and species suggesting mean values in the range of 318-688 base pairs (see also Supplementary Table 8). For the guinea baboon and the pig-tailed macaque, this analysis is underpowered and results are not shown.

Our rough estimates agree with previous estimates of NCOs associated with COs of about 500 bps[28]. As a control, we also tested SNVs two positions away from the estimated CO point and found no significant GC-bias, but still with a point estimate of 51.2% (Supplementary Fig. 8). This suggests that some CO events indeed initiate outside of the interval marked by the SNVs that move from one haplotype to another. This also explains why we found a further enrichment of DSBs when including the region with the flanking SNVs of the jump SNVs (Fig. 2c).

SNVs participating in the complex GCV events also show a large GC-bias of 53.4% (95% confidence interval: [52.8%;54.1%]), (Fig. 4d). Since they affect several SNVs, they have the largest combined contribution to GC evolution. Among all events that we observed across samples, the net gain of Gs or Cs is 151 for GCVs, 254 for COs, and 759 for complex events. This suggests that simple gene conversions have a relatively small effect on the GC content of the genome and explains why crossovers are highly associated with GC content.

Discussion

We have shown that Pacbio HiFi reads allow precise mapping of COs and GCVs in post-meiotic cells of testis tissue and in sperm samples across Old World primate species, including humans. The ability to predict CpG methylation from HiFi kinetics allowed us to classify the reads coming from germline cells with great accuracy. The number of germline reads was directly linked to the estimated spermatogenesis efficiency detected in the neighbouring tissue by histological evaluation. The success of this approach likely reflects the striking difference in DNA methylation between the germline and somatic cells in the testis, as well as profound changes in DNA methylation during spermatogenesis[17]. This is important since mutational events mimicking both COs and GCVs are likely to occur in somatic cell types. The combined utility of long, very accurate sequencing and concurrent methylation calls hence, allows us to study the dynamics of recombination processes in great detail using testis tissue. It allows us to identify recombination events that appear deleterious to haploid cells and incompatible with fertility, such as the excess of events in segmental duplication, which we observe to a greater extent in human testis compared to human sperm. Importantly, our approach opens the possibility of studying recombination in a much broader range of individuals and species where extensive trios or purified sperm samples are not currently available.

Surveying six Old World monkey species, we find that the recombination process is evolutionary very well conserved. The telomeric enrichment of COs, the estimated lengths of gene conversion tracts associated with GCVs and COs have small quantitative differences among species and are qualitatively very similar. This broadly agrees with studies using very different methodology of *Papio anubis* (mean tract length of 24 bp[33]) and of *Macaca mulatta* (mean tract length of 155 bp[34]). Specifically, in humans, we find that DSBs induced by PRDM9 are responsible for most recombination events. A smaller

https://doi.org/10.1038/s41467-025-65248-3

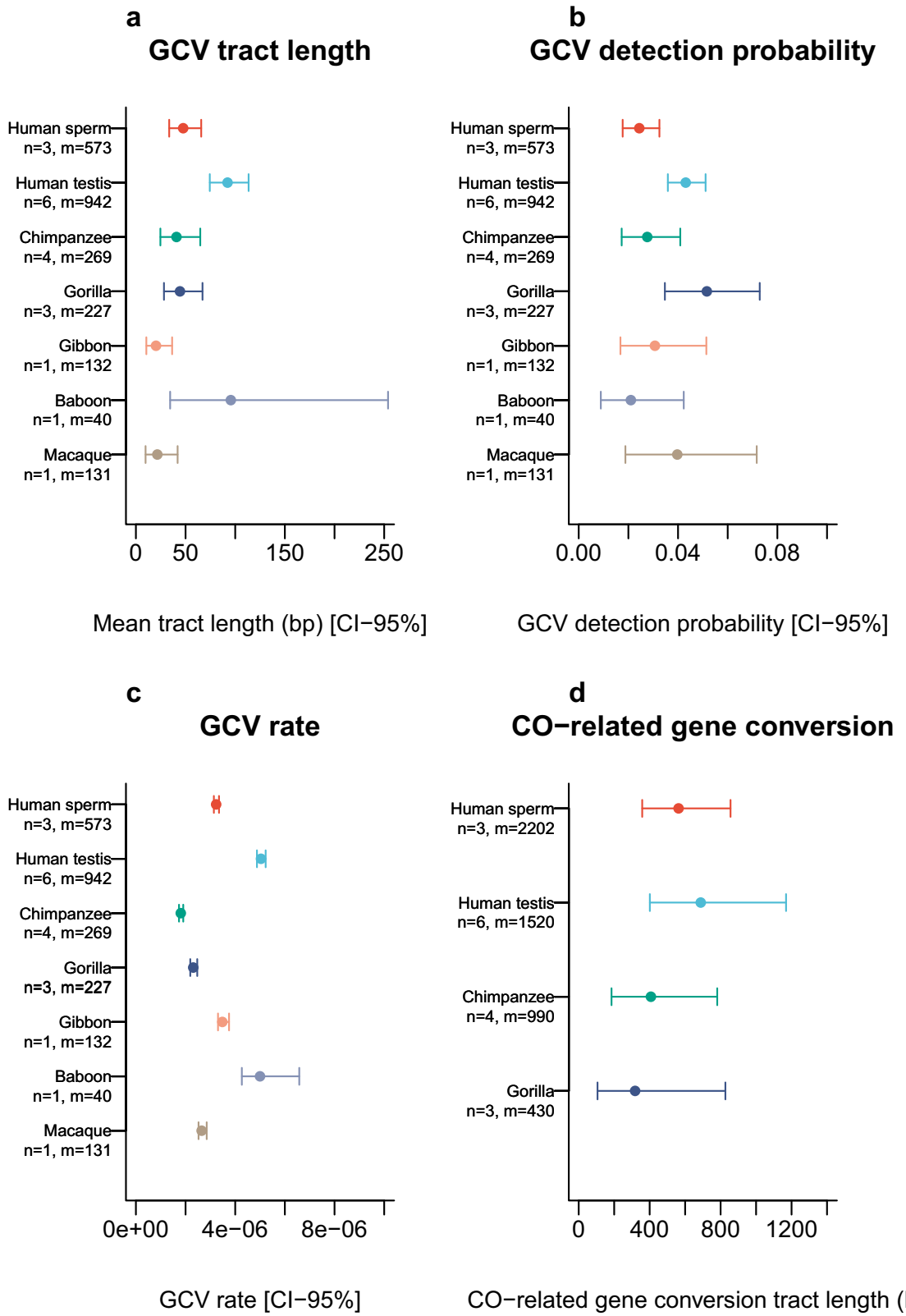

Fig. 3 | Gene conversion properties. In each panel, n denotes the number of individuals pooled and m denotes the number of SNVs analysed. **a** Estimated mean tract lengths (error bars denote 95% confidence intervals) from the empirical SNV density and the numbers of SNVs moved in GCVevents. **b** The estimated fraction GCV events which move one or more SNVs from one haplotype to the other (error bars denote 95% confidence intervals). **c** The estimated rate of GCV, i.e. the fraction of nucleotides involved in a GCV event in one generation (error bars denote 95% confidence intervals). **d** The estimated tract length of GCV-event co-occurring with CO events (error bars denote +/− SE, see methods for details).

https://doi.org/10.1038/s41467-025-65248-3

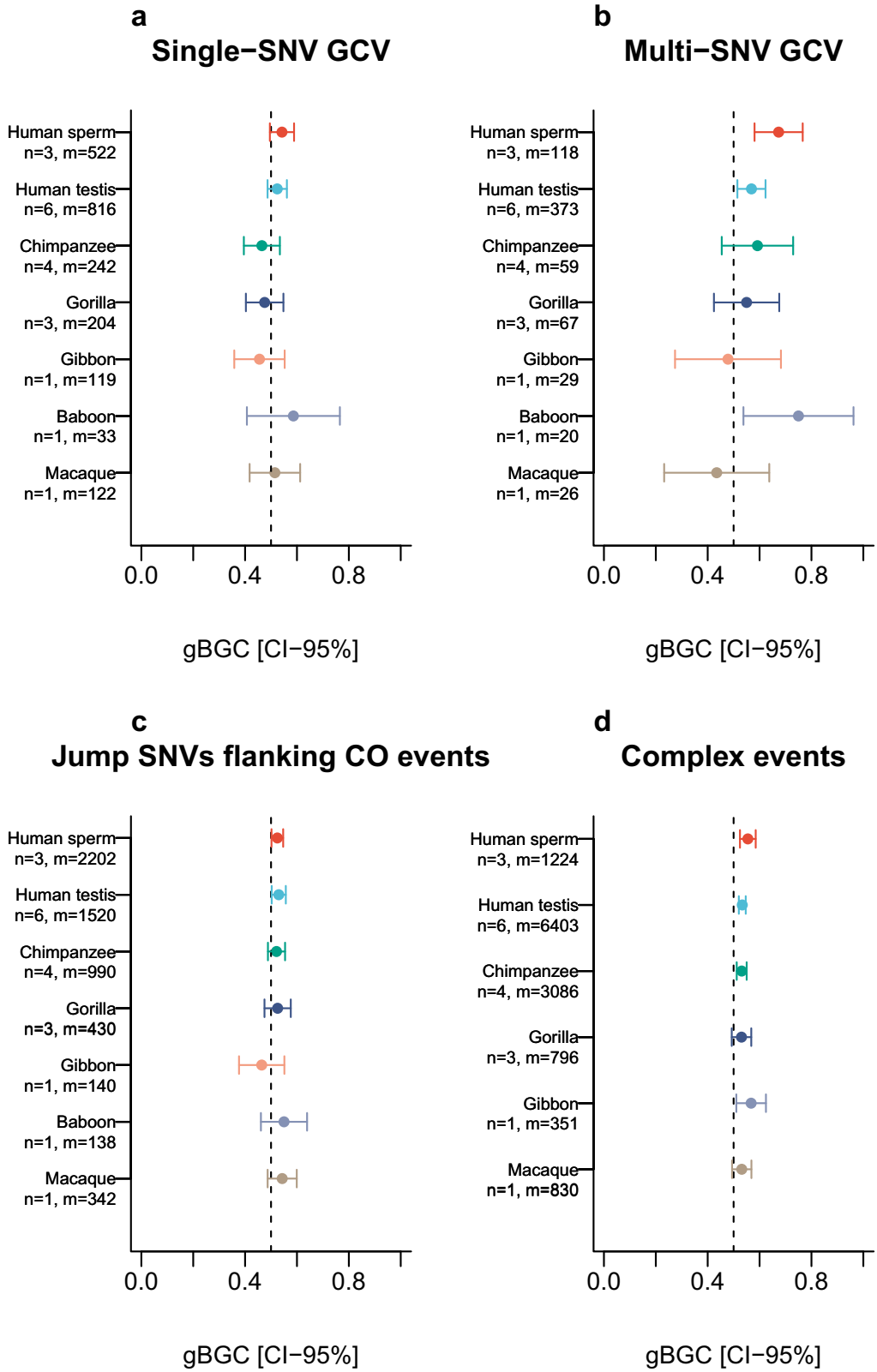

Fig. 4 | Estimates of the amount of gBGC with binomial 95% confidence intervals. In each panel, n denotes the number of individuals pooled and m denotes the number of SNVs analysed. The dashed line denotes no bias in whether a strong base (GC) or a weak base (AT) is donated. **a** The estimated amount of GC-biased gene conversion (gBGC) in gene conversions (GCVs) where a single SNV is moved from one haplotype to the other. (See **b**, GCVs where more than one subsequent SNV is moved. **c** SNVs in a complex GCV event (see Fig. 1a). **d** SNVs adjacent to an inferred CO, the jump SNVs (see Fig. 2b). For raw counts, SE and exact gBGC in a-d, see Supplementary Table 9-12.

fraction of DSBs (5-10%) are resolved as COs, and this process is biased towards the telomeres and smaller chromosomes. This implies that GCV events are less correlated to pedigree-based recombination maps. In the other species, an overall telomeric enrichment of COs is also observed. However, at a finer scale, corresponding to locations of DSBs in humans (average size 2 kb), we found very limited overlap, which is in agreement with non-human species having a divergent DNA-interacting zinc-finger motif of PRDM9 compared to humans[35,36].

We find a strong enrichment of COs near the telomeres in humans and chimpanzees, which are the species where we call most events. This enrichment is less clear in the other species. This may reflect a difference in how CO events are placed in these species or a lack of resolution. In gorillas, for example, subtelomeric repeats are more abundant[37] and this may cause a difference in CO placement.

Our estimates of GCV tract lengths from simple events are short (averaging between 20-100 bp across species), and conform closely to a geometric distribution. In contrast, from gBGC patterns, we estimate that tracts of exchange associated with COs are much longer, at least on the order of 500 bps. A special case of this is the complex GCV, which is also associated with gBGC and account for the largest effect on GC content. We cannot determine the maximum length of complex events due to the finite read lengths, but they could represent a mixture of complex COs and long GCV events as recently reported by refs. 1,29. Thus, the main effect of gBGC on GC content is not the resolution of simple GCV events but rather longer tracts associated with COs or complex events. This would explain why GC content correlates very well with pedigree-based recombination rates, even though GCV events do not correlate very well with recombination rates.

We find gBGC in GCV events containing multiple SNVs, suggesting that the repair of these involves template switching. Similarly, Gergelits et al.[38] (studying mice) found strong gBGC occurring at multi-SNV GCV events, while Li et al.[28] (studying interspecific mice hybrids) found no gBGC at multi-SNV events. While Gergelits et al.[38] and our study used GCV across the genome, Li et al. [28] focused on GCV found only around PRDM9-motifs, so there could be a difference in repair mechanisms after PRDM9-directed recombination and repair associated with other DSBs.

Besides studying recombination events in the testis, our approach of combining information from long sequencing reads opens new avenues of research. Accurate long-read sequencing combined with methylation calls could also be applied to study female meiosis, embryo development, as well as neoplastic transformation. Recent technological developments allowing less input material and lower prices also open for possibilities to perform similar lines of research on scarce material and on larger cohorts of samples.

Methods

Samples

DNA was isolated from human testis tissue obtained from six anonymous men (autopsies) and human sperm samples that were obtained from three anonymous sperm donors with good semen quality. As human tissue and sperm donors were anonymous, their exact age is unknown, but an approximate estimate is given. Finally, the study included testis tissue from seven great apes: four chimpanzees of the West African subspecies (*Pan troglodytes verus*) and three gorillas of the Western lowland subspecies (*Gorilla gorilla gorilla*). Additionally, testis tissue from a Lar Gibbon (*Hylobates lar*), a Guinea baboon (*Papio papio*), and a Pig-tailed macaque (*Macaca nemestrina*) were included. The samples from all apes and monkeys were collected post-mortem after death by natural causes or planned castration and the testis tissue was stored at −80 ^{0}C until DNA extraction. To evaluate spermatogenesis, a piece of the tissue was fixed in 10% neutral-buffered formalin for at least 16 hours, dehydrated and embedded in paraffin. Ongoing spermatogenesis with post-meiotic round spermatids in most tubules was verified by hematoxylin and eosin histological staining (Supplementary Fig. 2). Fixed tissue was not available from CT21, GT22, and GT43.

Samples are named according to species (H for human, G for gorillas, C for chimpanzees, L for Lar gibbon, B for baboon, and M for macaque), tissue (T for testis and S for sperm), as well as the age (estimated for human samples) as outlined in Table 1.

DNA extraction

DNA was isolated from testis tissue samples using the MagAttract® HMW DNA kit (Qiagen, Hilden, Germany) or the NanoBind PanDNA kit (Pacific Biosciences, Menlo Park, CA, USA) following the manufacturer's instructions.

The three sperm samples were initially allowed to liquify for 30 min at 37 °C and subsequently subjected to density gradient purification using PureSperm 50 (three aliquotes PureSperm 40 combined with one aliquot PureSperm 80; Nidacon, Gothenburg, Sweden) to avoid somatic cell contamination and enrich for mature sperm[39]. 2–3 mL of the raw sperm sample was gently added on top of 2 mL cold (4 °C) PureSperm 50 and centrifuged at 400 *g* for 15 min. The pellet, enriched for sperm, was subsequently washed with 5 mL phosphate-buffered saline (PBS). After centrifugation at 2200 *g* for 15 min, the pellet and 1 mL of the supernatant were stored in the freezer for subsequent DNA isolation. To isolate the DNA, an aliquot of the purified sperm sample was incubated with RLT buffer (Qiagen) and stainless steel beads on a shaker for 10 minutes to allow better access to the tightly packed DNA. The DNA was subsequently isolated using the automated Maxwell 16 system (SEV AS1010, Promaga, Madison, WI, USA).

The quantity and quality of the isolated DNA were evaluated by Nanodrop (ThermoFisher Scientific, Waltham, MA, USA), Qubit dsDNA HS or BR Assay Kits (ThermoFisher), and gel electrophoresis before library construction. All samples revealed DNA of high molecular weight.

PacBio HiFi sequencing

The DNA was subjected to PacBio HiFi sequencing at the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programmes of the Research Council of Norway and the Southeastern Regional Health Authorities.

In short, the sequencing libraries were prepared using the Pacific Biosciences protocol for HiFi library prep using SMRTbell® ExpressTemplate Prep Kit version 2.0 (Pacific Biosciences, Menlo Park, CA, USA). DNA was fragmented into 15-20 kb fragments using Megaruptor 3 (Diagenode, Denville, NJ, USA), and the final library was size-selected using BluePippin (Sage Science) with a 10 kb cut-off.

The final libraries were sequenced either on the Sequel II or the newer Revio instrument, and four samples (HT55, GT43, GT22, HS35) were sequenced on both instruments.

Samples sequenced on the Sequel II instrument (HT20, HT55, CT32, GT43, GT22, HS35, HS50) were sequenced with a 30 h movie time on 8 M SMRT cells using the Sequel II Binding kit 2.2 and sequencing chemistry v2.0 (Pacific Biosciences).

Samples sequenced on the Revio instrument (HT45, HT50, HT55, HT60, HT65, HS25, HS35, HS25, CT22, CT28, GT43, GT22, GT21, LT39, BT15, MT5) were sequenced with a 24 h movie time on 25 M SMRT cells using SMRTbell® ExpressTemplate Prep Kit version 3.0 and the Revio Polymerase kit.

CCS sequences were generated using the CCS pipeline (SMRT Link v10.2.0.133434 for Sequell II and v. 12.0.0.183503 for Revio reads). Parameters were set to include full base kinetics and full resolution of base qualities (the exact settings differ between Sequel II and Revio). Reads with at least 99% accuracy are retained as HiFi reads.

De novo assemblies
The de novo assemblies were built with *hifiasm* with its default parameters. The summary statistics of the assemblies are shown in Supplementary Table 1.

Mapping
Following the construction of the de novo assemblies, we map the reads from each individual against its de novo assembly using *pbmm2* with parameters -c 99 and -l 2740.

SNV calling
The SNVs were called using an in-house Python script based on pysam. The programme iterates through a BAM file and stops at a position where exactly two different nucleotides are present. Furthermore, the coverage in this position has to be greater than the 5th percentile and less than the 99.7th percentile of the coverage distribution. The programme also requires that no more than 10% of the reads can contain an indel at the given position and that the variant occurs in at least three reads.

Since PacBio HiFi reads accuracy is lower in repetitive regions, we filtered candidate SNVs that fall within local repetitive regions. Specifically, if the number of unique 4-mers in a 32-bp region is less than 19, we skip the position. We also skip the position if the candidate SNV occurs at the boundary of two homopolymers. For instance, if the sequence context is AAAA(T/A)TTTT, then it is difficult to assess whether the SNV is real or due to a sequencing artefact.

The last criterion for a position being an SNV is that it cannot occur within 15 bp of an indel column. We define an indel column as a position where more than 10% of the reads contain an indel.

For all reads at an SNV position, we assign a haplotype to each read. If the read contains the same base as the de novo assembly, the read is assigned haplotype 'X', whereas the read is assigned haplotype 'O', if it contains the alternative nucleotide. In rare cases, some reads contain an indel at the SNV position, and are assigned haplotype'n'. If a read is assigned 'O', but the base quality is less than QV35, or a small indel is within 10 bp in each direction, or a mismatch is within 10 bp in each direction, or a big indel (>10 bp) is within 250 bp in each direction, then the read is assigned haplotype 'o'. The distinction between 'O' and 'o' makes it possible to assess the quality of the haplotype assignment for all reads at all SNV positions. An identical assessment is performed for haplotype 'X'. Following the assignment of haplotypes, the programme continues until the last position in each contig.

Recombination calling
Candidate recombination events are called when a read contains high-confidence variants from both haplotypes (X and O). Subsequently, those events are divided into four basic types (see also Supplementary Fig. 1):

1. CO events are defined as reads that display a single jump of haplotype and where at least two SNVs from each haplotype are present (e.g., 'XXxOO').
2. NCO events are defined as an SNV string that contains exactly two haplotype jumps (e.g.'OoXOO').
3. The boundary events are defined as an SNV string with a single jump of haplotype (like the COs), but the jump occurs between the first or last two SNVs (e.g., OOooX).
4. The complex events are defined as an SNV string with more than two haplotype jumps.

To minimise the number of false positives, we exclude events that occur closer than 5000 bp to another event.

The overall pipeline is sketched in Supplementary Fig. 1, however, the final step in filtering the reads containing recombination events is specific to the human samples, where we exclude any recombination calls that land on the acrocentric short arms and any calls that overlap with segmental duplications.

To separate the events found in segmental duplications, we intersect the map of segmental duplications with the CHM13 coordinates of the called events[40].

SD and acrocentric annotations:
https://zenodo.org/records/7671779

Furthermore, for testis samples, we only record the recombination events observed in reads that are assigned as germline reads in Fig. 1a. For the classification of reads, see the section "Methylation-based classification of cell types" below.

Manual curation
Following the classification of germline read recombination events, all reads were manually curated using a script rendering relevant regions visible in IGV. Manual curation was done by PSP for all events and by APC, SB, and MHS for a subset of the events. From this, we calculated the interobserver concordance to be >0.95, with 5–20% of events deemed as false positives and removed from further consideration. The events scored as false positives were mainly due to repetitive regions of the genome, long distances between SNVs, or collapsed regions, where more than haplotype are present.

Inference of the distribution of GCV tract length from candidate read counts
We estimated GCV tract lengths and GCV rates using the framework of (Charmouh et al. 2024, in press). Briefly, we infer, using maximum likelihood, the mean of the best fitting geometric distribution of GCV tract length from the observed counts of reads that exhibit a footprint of GCV and by tallying how many reads convert/move 1, 2, 3, etc. SNVs in the read from an O to an X haplotype. This approach assumes that all GCV events are independent and that each GCV event induces a tract of physical length L, where L is modelled as a stochastic (random) variable geometrically distributed with expected mean 1/s. The data for each sample is summarised as the counts of reads n_i containing apparent GCV events that have moved in SNVs.

Using simulation, the method allows us to obtain the expected proportions of reads harbouring 1,2,3, etc. converted SNVs given the genome-wide SNV distribution of each sample as a function of some mean GCV tract length. By doing so, we account for differences in detectability induced by the fact that the SNV density underlying the two haplotypes varies hugely from region to region (Median distance: 347 bp, mean inter-SNV distance: 1173.07 bp, IQR: 812, see Supplementary Fig. 7).

Inference of CO-related GCV tract length using gBGC at CO-flanking SNVs
By definition, a gene conversion tract overlaps any converted SNV in a GCV event. We use the amount of gBGC at SNVs flanking CO events to estimate the length of the gene conversion tract co-occurring with CO events by comparing this gBGC to the amount of gBGC typically observed at GCV events. Specifically, we ask, given that GCV SNVs overlap with a GCV tract 100% of the time and assuming on average $gBGC_{GCV}$ = 61.95% gBGC (Palsson et al. 2025), which means tract length can produce enough overlap to explain the amount of $gBGC_{CO}$% gBGC observed at SNVs flanking CO events?

The following assumptions are made: For each CO event, the DSB has the same probability of occurring at any of the SNVs between the L bp that separate the jump SNVs. Second, the directionality of the tract is random. Third, tract lengths follow a geometric distribution with mean s. Given these assumptions, the probability $P(overlap|s,L)$ of a tract overlapping the right or left jump SNV at a CO becomes

$$P(overlap|s,L) = (1-(1-1/s)^{L-1})/((L-1)/s) \quad (1)$$

For each sample or pooled sample (see Fig. 3d), we investigated s values in the range [1, 15000] and selected as the estimated CO-related NCO tract length the s which resulted in an average of $(2(gBGC_{CO} - 0.5))/gBGC_{NCO}$ % overlap with a jump SNV across all CO events (factor of 2 is applied to account for the random directionality of the tracts). For each sample or pooled sample, we estimated s given $gBGC_{CO}$ and given $gBGC_{CO} \pm$ SE (see Fig. 3d).

Testing for differences in NCO tract length between samples

To test for differences in tract length between samples A and B, say, we employ a likelihood ratio test comparing the relative fit of two models to the data. Under Model0 (M_0), we assume that reads from both samples are independent observations of NCO where the induced conversion tract length is drawn from a single geometric distribution with mean length $L_0 = 1/S_0$. Under Model1 (M_1), we allow the induced conversion tract lengths to follow different geometric distributions with mean $1/S_A$ and $1/S_B$, respectively. The likelihood of the data **D** (counts in sample A and sample B) is maximised under both M_0 and M_1, and we use $G_{obs} = 2 \log (L_1(\mathbf{D})/(L_0(\mathbf{D}))$ to assess statistical significance. Under the null hypothesis that both samples follow an identical tract length, we have that G_{obs} is approximately X^2 distributed with 1 degree of freedom.

Using this framework, we can test for pairwise differences (e.g., between sample A and B) or extend the test to globally test for differences between all K samples. In that case, M_1 allows K geometric distributions with different means for each sample, and the test statistic G_{obs} is approximately X^2 distributed with K-1 degrees of freedom if M_0 is correct.

Given that the likelihood functions rely on the multinomial distribution of counts, and provided that the total number of counts is larger than 5, the X^2 approximation to the distribution of G_{obs} is expected to be very accurate (e.g. ref. 41, Chapter 8).

Methylation-based classification of cell types

This section gives a brief description of the method for classifying reads. A longer description can be found in Supplementary Note 1.

To train a probabilistic method for classifying reads as germline or non-germline, we required separate whole-genome methylation data from both germline and non-germline cell types. To estimate methylation levels in germline cells, we used methylation data from the four spermatogenic cell stages of spermatogenesis obtained from Sibert-Kuss et al.[17] and data from mature sperm obtained from Leitão et al.[42]. To represent the methylation levels in somatic cells, we used data from neurons and merged blood cell types obtained from Loyfer et al.[18].

The training data was aligned to hg38, and CpG sites with a coverage smaller than six in any of the seven training datasets (12.3% of all CpG sites of the genome) were discarded. The average depth of the remaining CpG sites in the training data can be seen in Supplementary Note 1.

The software tool Jasmine (https://github.com/PacificBiosciences/jasmine) was used to calculate methylation values for each CpG site in the HiFi reads. For each CpG site, i, in read, r, Jasmine assigns a probability, $Y_{r,i}$ that the CpG site is methylated. We use $Z_{r,i} \in \{0,1\}$ to denote the unknown true methylation status and assume that the observed methylation probabilities from methylated sites follow symmetric Beta distributions with shape parameters α and β:

$$P(Y_{r,i} = y | Z_{r,i} = 1) = f_{Beta}(y, \alpha, \beta) \quad (2)$$

$$P(Y_{r,i} = y | Z_{r,i} = 0) = f_{Beta}(y, \beta, \alpha) \quad (3)$$

The α and β values can differ between samples due to differences in the average number of consensus reads. To learn the α and β values for a given sample, we fit a mixture between the two symmetric functions:

$$f(x) = \omega f_{Beta}(x, \beta, \alpha) + (1 - \omega) f_{Beta}(x, \beta, \alpha) \quad (4)$$

Let C_r be a discrete random variable stating the (unknown) cell type for read r. Then $n_{m,i,c}$ and $n_{u,i,c}$ are the observed number of methylated and unmethylated reads, respectively, at CpG site in the training data from cell type c. As before $Z_{r,i}$ is a latent variable stating if the CpG site i is methylated in read r. We assume that $Z_{r,i}$ is Bernoulli distributed with a rate that depends on the cell type of the read:

$$P(Z_{r,i} = 1 | C_r = c) = (n_{m,i,c} + \omega)/(n_{m,i,c} + n_{u,i,c} + 1) \quad (5)$$

$$P(Z_{r,i} = 0 | C_r = c) = (n_{u,i,c} + (1 - \omega))/(n_{m,i,c} + n_{u,i,c} + 1) \quad (6)$$

Where the ω prior is the mixture proportion (from Eq. 3) estimated for this sample.

We can then determine the likelihood of an observed methylation value $Y_{r,i}$ given read r comes from cell type, c by combining Eqs. (2), (3), (5), and (6):

$$P(Y_{r,i} = y | C_r = c) = \sum_{z \in \{0,1\}} P(Y_{r,i} = y, | , Z_{r,i} = z) P(Z_{r,i} = z, | , C_r = c) = \quad (5)$$

$$\begin{aligned} & f_{Beta}(y, \beta, \alpha) \cdot ((n_{u,i,c} + (1 - \omega))/(n_{m,i,c} + n_{u,i,c} + 1)) \\ & + f_{Beta}(y, \alpha, \beta) \cdot ((n_{m,i,c} + \omega)/(n_{m,i,c} + n_{u,i,c} + 1)) \end{aligned} \quad (7)$$

To determine the log-likelihood that read r comes from cell type c, using the methylation status of all informative CpG sites on the read. We assume independence between the methylation levels of each CpG site and, accordingly, sum the site-specific log-likelihoods. We then classify a read as germline if the cell type with the highest log-likelihood is one of the five germline cell types.

By setting a minimum cutoff on the difference in log-likelihood between the most likely germline and the most likely non-germline cell type, we can increase the accuracy of the classification. Analyzing purely germline sperm samples and non-germline blood samples, we estimate that we can classify 72% of reads with an FPR of 0.05, and 21% with an FPR of 0.01. See Supplementary Note 1 for details.

GC-biased gene conversion

For all GCV reads that were approved during manual curation and which were classified as being of germline origin, we measured GC-biased gene conversion (gBGC). Since both haplotypes are known, we could determine which base was donated (i.e., copied from one haplotype onto the other) and which was erased. For each approved GCV read, we recorded which base was being donated and which was being erased. We counted all the times a strong base pair (G or C) was being copied while a weak base pair (A or T) was being erased, and vice versa. We then quantified gBGC as the number of conversions to strong bases and away from weak divided by the number of conversions from strong to weak plus the number of conversions from weak to strong. We defined as significant gBGC counts wherein the chi-squared value for gBGC was significantly higher than 0.5.

For GCVs, bases of all donated SNVs were recorded, which ranged from 1 to 29 bases depending on the event (Supplementary table 7). For CO events, bases of both jump SNVs were recorded. For complex cases, the bases of all SNVs were recorded. For boundary cases, the minority SNV was recorded, except for cases where no minority was present (e.g., XO, OX, see Fig. 1a), in which case both were recorded. In all cases, indels were excluded.

Ethics

Human testis tissue was obtained from anonymous men (autopsies) who consented to donate tissue for research purposes post-mortem. Human sperm samples were obtained from anonymous sperm donors in 2018 after consent and an ethical permit granted by the regional committee of the Capital Region in Denmark (H-17012149). The study was conducted in accordance to the criteria set by the Declaration of Helsinki.

The samples from all apes and monkeys were obtained from zoos within the EU and were collected post-mortem after death by natural causes or planned castration following the guidelines of the European Association of Zoos and Aquaria (EAZA). CT15 is EAZA studbook number 13342, CT22 is EAZA studbook number 12779, CT28 is EAZA studbook number 12786, and CT32 is EAZA studbook number 12295. GT21 is EAZA studbook number 1709, GT22 is EAZA studbook number 1435, and GT43 is EAZA studbook number 492, LT39 is EAZA studbook number 402, BT15 is EAZA studbook number 444 and MT5 is EAZA studbook number 775.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

According to Danish legislation, we are not allowed to deposit human sequencing data from anonymous human donors; however, all the summary data necessary to replicate the study are publicly available via https://github.com/PeterSoerud/recombination_calling and Zenodo [https://doi.org/10.5281/zenodo.16893843]. For other species, HiFi reads in fastq have been deposited in the ENA database under accession code PRJEB77177 [(https://www.ebi.ac.uk/ena/browser/view/PRJEB77177]. This also contains the de novo assemblies of non-human samples. The processed data are available at https://github.com/PeterSoerud/recombination_calling. In this paper, we have reused the following datasets for our classification analysis (see Supplementary Note for details): Genome-wide DNA methylation changes in human spermatogenesis: EGAS00001007449. The sperm epigenome does not display recurrent epimutations in patients with severely impaired spermatogenesis: PRJEB34432 A DNA methylation atlas of normal human cell types: EGAS00001006791 [https://ega-archive.org/datasets/EGAD00001009789]

Code availability

Scripts for calling recombination are available on GitHub at https://github.com/PeterSoerud/recombination_calling.This repository also contains details about all recombination events and is available on Zenodo [https://doi.org/10.5281/zenodo.16893843]. EM-seq/BS-seq data processing scripts are available on Github at https://github.com/vinodsinghjnu/BWA_METH_pipeline/ and on Zenodo [https://doi.org/10.5281/zenodo.17234585] The scripts used to perform the classification of reads are available on Github at https://github.com/vinodsinghjnu/Proj_GermReadsClassification/ and on Zenodo [https://doi.org/10.5281/zenodo.17233827]

References

1. Palsson, G. et al. Complete human recombination maps. *Nature* **639**, 700–707 (2025).
2. Pratto, F. et al. DNA recombination. Recombination initiation maps of individual human genomes. *Science* **346**, 1256442 (2014).
3. Brick, K., Pratto, F. & Camerini-Otero, R. D. After the break: DSB end processing in mouse meiosis. *Gene Dev.* **34**, 731–732 (2020).
4. 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).
5. Pratto, F. et al. Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell* **184**, 4251–4267.e20 (2021).
6. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
7. Hinch, R., Donnelly, P. & Hinch, A. G. Meiotic DNA breaks drive multifaceted mutagenesisin the human germ line. *Science* **382**, eadh2531 (2023).
8. Duret, L. & Galtier, N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).
9. Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**, 907–911 (2001).
10. Marais, G. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* **19**, 330–338 (2003).
11. McVean, G. A. et al. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
12. Hinch, A. et al. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science* **363**, eaau8861 (2019).
13. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
14. Kong, A. et al. Common and low-frequency variants associated with genome-wide recombination rate. *Nat. Genet.* **46**, 11–16 (2014).
15. Halldorsson, B. V. et al. The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* **48**, 1377–1384 (2016).
16. Vollger, M. R. et al. Increased mutation and gene conversion within human segmental duplications. *Nature* **617**, 325–334 (2023).
17. Siebert-Kuss, L. M. et al. Genome-wide DNA methylation changes in human spermatogenesis. *Am. J. Hum. Genet.* **111**, 1125–1139 (2024).
18. Loyfer, N. et al. A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).
19. Cagan, A. et al. Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
20. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* 1–6 (2021) https://doi.org/10.1038/s41586-021-03822-7.
21. 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**, 861–869 (1998).
22. Skakkebæk, N. E. & Heller, C. G. Quantification of human seminiferous epithelium. *Reproduction* **32**, 379–389 (1973).
23. Johnson, L. et al. Efficiency of spermatogenesis: a comparative approach. *Anim. Reprod. Sci.* **60**, 471–480 (2000).
24. Roosen-Runge, E. C. Quantitative investigations on human testicular Biopsies I. Normal testis. *Fertil. Steril.* **7**, 251–261 (1956).
25. Auton, A. et al. A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**, 193–198 (2012).
26. Stevison, L. S. et al. The time scale of recombination rate evolution in great apes. *Mol. Biol. Evol.* **33**, 928–945 (2016).
27. Poulsen, A. et al. Estimating gene conversion tract length and rate from pacbio hifi data. *Mol. Biol. Evol.* **42**, msaf019 (2025).
28. Li, R. et al. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nat. Commun.* **10**, 3900 (2019).
29. Hardarson, M. T., Palsson, G. & Halldorsson, B. V. NCOurd: modelling length distributions of NCO events and gene conversion tracts. *Bioinformatics* **39**, btad485 (2023).
30. Williams, A. L. et al. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* **4**, e04637 (2015).

31. Narasimhan, V. M. et al. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat. Commun.* **8**, 303 (2017).
32. Schweiger, R. et al. Insights into non-crossover recombination from long-read sperm sequencing. (2024) https://doi.org/10.1101/2024.07.05.602249.
33. 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**, evac040 (2022).
34. Versoza, C. J. et al. Novel insights into the landscape of crossover and noncrossover events in Rhesus Macaques (Macaca mulatta). *Genome Biol. Evol.* **16**, evad223 (2023).
35. Schwartz, J. J., Roach, D. J., Thomas, J. H. & Shendure, J. Primate evolution of the recombination regulator PRDM9. *Nat. Commun.* **5**, 4370 (2014).
36. Myers, S. et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**, 876–879 (2010).
37. Yoo, D. et al. Complete sequencing of ape genomes. *Nature* 1–18 (2025) https://doi.org/10.1038/s41586-025-08816-3.
38. Gergelits, V., Parvanov, E., Simecek, P. & Forejt, J. Chromosome-wide characterization of meiotic noncrossovers (gene conversions) in mouse hybrids. *Genetics* **217**, 1–14 (2020).
39. Lee, D. & Jee, B. C. Evaluation of normal morphology, DNA fragmentation, and hyaluronic acid binding ability of human spermatozoa after using four different commercial media for density gradient centrifugation. *Clin. Exp. Reprod. Med.* **46**, 8–13 (2019).
40. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
41. Grant, G. R. & Ewens, W. *Statistical Methods in Bioinformatics: An Introduction*. (Springer, 1989).
42. Leitão, E. et al. The sperm epigenome does not display recurrent epimutations in patients with severely impaired spermatogenesis. *Clin. Epigenet.* **12**, 61 (2020).

Acknowledgements

The authors are grateful to Bonnie Colville-Ebeling for helping with the collection of anonymous human testicular tissue, Ana Ricci C. G. Nielsen for tissue processing, Sissel Marie Bredesen for sperm purification, and Brian Vendelbo Hansen for helping with the DNA purification. We are also grateful to Veterinarian Imke Lüders, Münster Zoo, Germany, for helping collect great ape samples and Senior Engineer Ave Tooming-Klunderud from the Norwegian Sequencing Centre for the excellent help with the PacBio HiFi sequencing. We thank Regev Schweiger, Richard Durbin and Laurent Duret for fruitful discussions and information. The computational work has taken advantage of the genomeDK supercomputing infrastructure (genome.au.dk). The work was funded by the Novo Nordisk Foundation (grant# NNF21OC0069105). The authors have benefited from networking grants of COST Action CA20119 (ANDRONET) supported by the European Cooperation in Science and Technology (www.cost.eu).

Author contributions

M.H.S., T.B., S.B. and K.A. conceived the study. C.H., J.A.B., K.A. and S.B.W. collected the samples. S.B.W. and K.A. generated the data. P.S.P. built the de novo assembly, called recombination events, analysed the spatial distribution of recombination events and headed the manual curation of reads with input from M.H.S., T.B., A.P.C., A.H., M.P., C.O. and S.B. A.P.C. and L.T.H. developed and applied inference methods for estimating gene conversion and gBGC with input from P.S.P., M.H.S., T.B., S.B., and A.H. V.K.S., M.H.S, K.A. and S.B. developed and applied the read classifier with input and sorted germ cell data provided by N.N. and S.L. M.H.S. and P.S.P. wrote the first draft of the manuscript with substantial input from A.P.C., T.B., A.H., S.B., V.K.S., S.B.W., M.P. and K.A.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-65248-3.

Correspondence and requests for materials should be addressed to Peter Soerud Porsborg or Mikkel Heide Schierup.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. [A peer review file is available].

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Bioinformatics Research Centre, Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark. [2]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. [3]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. [4]Department of Growth and Reproduction, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark. [5]International Center for Research and Research Training in Endocrine Disruption of Male Reproduction and Child Health, Copenhagen, Denmark. [6]Copenhagen Zoo, Frederiksberg, Denmark. [7]Department of Mathematics, Aarhus University, Aarhus, Denmark. [8]Institute of Reproductive Genetics, Centre of Medical Genetics, University of Münster, Münster, Germany. [9]Centre of Reproductive Medicine and Andrology, Institute of Reproductive and Regenerative Biology, University of Münster, Münster, Germany. [10]Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark. [11]These authors contributed equally: Peter Soerud Porsborg, Anders Poulsen Charmouh. [12]These authors jointly supervised this work: Søren Besenbacher, Kristian Almstrup, Mikkel Heide Schierup ✉e-mail: pepo@birc.au.dk; mheide@birc.au.dk