

Insights into variation in meiosis from 31,228 human sperm genomes

<https://doi.org/10.1038/s41586-020-2347-0>

Received: 1 May 2019

Accepted: 23 March 2020

Published online: 3 June 2020

 Check for updates

Avery Davis Bell^{1,2}, Curtis J. Mello^{1,2}, James Nemesh^{1,2}, Sara A. Brumbaugh^{1,2}, Alec Wysoker^{1,2} & Steven A. McCarroll^{1,2}

Meiosis, although essential for reproduction, is also variable and error-prone: rates of chromosome crossover vary among gametes, between the sexes, and among humans of the same sex, and chromosome missegregation leads to abnormal chromosome numbers (aneuploidy)^{1–8}. To study diverse meiotic outcomes and how they covary across chromosomes, gametes and humans, we developed Sperm-seq, a way of simultaneously analysing the genomes of thousands of individual sperm. Here we analyse the genomes of 31,228 human gametes from 20 sperm donors, identifying 813,122 crossovers and 787 aneuploid chromosomes. Sperm donors had aneuploidy rates ranging from 0.01 to 0.05 aneuploidies per gamete; crossovers partially protected chromosomes from nondisjunction at the meiosis I cell division. Some chromosomes and donors underwent more-frequent nondisjunction during meiosis I, and others showed more meiosis II segregation failures. Sperm genomes also manifested many genomic anomalies that could not be explained by simple nondisjunction. Diverse recombination phenotypes—from crossover rates to crossover location and separation, a measure of crossover interference—covaried strongly across individuals and cells. Our results can be incorporated with earlier observations into a unified model in which a core mechanism, the variable physical compaction of meiotic chromosomes, generates interindividual and cell-to-cell variation in diverse meiotic phenotypes.

One way to learn about human meiosis has been to study how genomes are inherited across generations. Genotype data are available for millions of people and thousands of families; crossover locations are estimated from genomic segment sharing among relatives and from linkage-disequilibrium patterns in populations^{2,4,7,9,10}. Although inheritance studies sample only the few gametes per individual that generate offspring, such analyses have revealed that average crossover numbers and crossover locations associate with common variants at many genomic loci^{3–6,11,12}.

Another powerful approach to studying meiosis is to directly visualize meiotic processes in gametocytes, which has made it possible to see that homologous chromosomes usually begin synapsis (their physical connection) near their telomeres^{13–15}; to observe double-strand breaks, a subset of which progress to crossovers, by monitoring proteins that bind to such breaks^{16,17}; and to detect adverse meiotic outcomes, such as chromosome missegregation^{18,19}. Studies based on such methods have revealed much cell-to-cell variation in features such as the physical compaction of meiotic chromosomes^{20,21}.


More recently, human meiotic phenotypes have been studied by genotyping or sequencing up to 100 gametes from one person, demonstrating that crossovers and aneuploidy can be ascertained from direct analysis of gamete genomes^{22–26}. Despite these advances, it has not yet been possible to measure multiple meiotic phenotypes genome-wide in many individual gametes from many people.

Development of Sperm-seq

We developed a method ('Sperm-seq') with which to sequence thousands of sperm genomes quickly and simultaneously (Fig. 1). A key challenge in developing Sperm-seq was to deliver thousands of molecularly accessible-but-intact sperm genomes to individual nanolitre-scale droplets in solution. Tightly compacted²⁷ sperm genomes are difficult to access enzymatically without loss of their DNA into solution; we accomplished this by decondensing sperm nuclei using reagents that mimic the molecules with which the egg gently unpacks the sperm pronucleus (Extended Data Fig. 1a–d). We then encapsulated these sperm DNA 'florets' into droplets together with beads that delivered unique DNA barcodes for incorporation into the genomic DNA of each sperm; we modified three technologies to do this (Drop-seq²⁸, 10× Chromium Single Cell DNA, and 10× GemCode²⁹, the latter of which was used to generate the data in this study) (Extended Data Fig. 1e, f). We then developed, adapted and integrated computational methods for determining the chromosomal phase of the sequence variants of each donor and for inferring the ploidy and crossovers of each chromosome in each cell.

We used this combination of molecular and computational approaches to analyse 31,228 sperm cells from 20 sperm donors (974–2,274 gametes per donor), sequencing a median of roughly 1% of the haploid genome of each cell (Extended Data Table 1). Deeper sequencing allows detection of roughly 10% of a gamete's genome.

¹Department of Genetics, Harvard Medical School, Boston, MA, USA. ²Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.

 e-mail: averydavisbell@gmail.com; mccarroll@genetics.med.harvard.edu

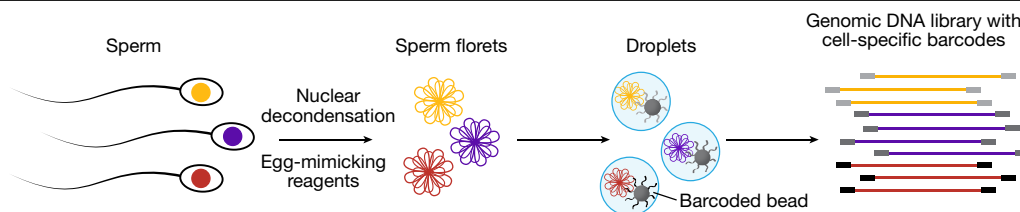


Fig. 1 | Overview of Sperm-seq. Schematic showing our droplet-based single-sperm sequencing method.

Sperm-seq enabled us to infer the haplotypes of donors along the full length of every chromosome: alleles from the same parental chromosome tend to appear in the same gametes, so the coappearance patterns of alleles across many sperm enabled us to assemble alleles into chromosome-length haplotypes (Extended Data Fig. 2a and Methods). In silico simulations and comparisons with kilobase-scale haplotypes from population-based analyses indicated that Sperm-seq assigned alleles to haplotypes with 97.5–100% accuracy (Extended Data Fig. 2b, c and Supplementary Notes).

The phased haplotypes determined by Sperm-seq allowed us to identify cell ‘doublets’ from the presence of both parental haplotypes at loci on multiple chromosomes (Extended Data Fig. 2d–f and Methods). We also identified surprising ‘bead doublets’, in which two beads’ barcodes reported identical haplotypes genome-wide through different single-nucleotide polymorphisms (SNPs), and thus appeared to have been incorporated into the same gamete genome (Extended Data Fig. 3a, b, Methods and Supplementary Methods). Bead doublets were useful for evaluating the replicability of Sperm-seq data and analyses (Extended Data Fig. 3c–e), which is usually impossible to do in inherently destructive single-cell sequencing.

Recombination rate in sperm donors and cells

We identified crossover (recombination) events in each cell as transitions between the parental haplotypes we had inferred analytically (Methods). We identified 813,122 crossovers in the 31,228 gamete genomes (Extended Data Table 1). Crossover locations were inferred with a median resolution of 240 kilobases (kb), with 9,746 (1.2%) inferred within 10 kb (Extended Data Table 1 and Supplementary Notes). Analysis of bead doublets indicated high accuracy of crossover inferences (Extended Data Fig. 3e). Estimates of crossover rate and location were robust to downsampling to the same coverage in each cell (Extended Data Fig. 4 and Supplementary Methods).

The recombination rates of the 20 sperm donors ranged from 22.2 to 28.1 crossovers per cell. This is consistent with estimates from other methods^{3,5,6,10–12,24,26}, but with far more precision at the individual-donor level (95% confidence intervals of 22.0–22.4 to 27.9–28.4 crossovers per cell) owing to the large number of gametes analysed per donor (Extended Data Table 1 and Extended Data Fig. 5a). Individuals with higher global crossover rates had more crossovers on average on each chromosome (Extended Data Fig. 5b). We generated genetic maps for each of the donors from their 25,839–62,110 observed crossovers; these maps were broadly concordant with a family-derived paternal genetic map⁶ (Extended Data Fig. 5c, d, Supplementary Notes and Supplementary Methods).

Much more variation was present at the single-cell level: cells routinely contained 17 to 37 crossovers (1st and 99th percentiles, median across donors), with a standard deviation of 4.23 across cells (median across donors), versus a standard deviation of 1.53 across donors’ crossover rates. Among gametes from the same donor, gametes with fewer crossovers in half of their genome tended to have fewer crossovers in the other half of their genome (Pearson’s $r = 0.09$, two-sided $P = 8 \times 10^{-54}$ with all gametes from all donors combined after within-donor normalization) (Supplementary Notes). This relationship, predicted by

earlier observations in families⁵ and spermatocytes²¹, suggests that the crossover number on each chromosome is partly shaped by factors that act nucleus-wide.

Crossover location and interference

All 20 donors shared a tendency to concentrate their crossovers in the same regions of the genome, with large concentrations of crossovers in distal regions, as expected from earlier analyses of families^{4,6,9,11,30}, and more modest shared enrichments in many centromere-proximal regions (Fig. 2a and Extended Data Fig. 6). Guided by these empirical patterns, we divided the genome into ‘crossover zones’, each bounded by local minima in crossover density (Extended Data Fig. 6b and Supplementary Methods). These zones are much larger-scale than fine-scale-sequence-driven crossover hotspots^{7,31–33}, which the spatial resolution of most crossover inferences was not well-suited for analysing.

The crossover zones with the most variable usage across people were all adjacent to centromeres; individuals with high recombination rates used these zones much more frequently (Fig. 2a and Extended Data Fig. 6a; with simulated equal SNP coverage, Extended Data Fig. 4c, e). The relative usage of distal and proximal zones varied greatly among donors and correlated with donors’ recombination rates (Extended Data Fig. 7). These results were robust to alternative definitions of ‘distal’ versus ‘proximal’ (Extended Data Fig. 7c and Supplementary Notes).

Positive crossover interference causes crossovers in the same meiosis to be further apart than they would be if crossovers were independent events^{26,30,34,35}. The effect of crossover interference was visible in each of the 20 sperm donors (Extended Data Fig. 8 and Supplementary Methods). Crossover separation varied greatly among sperm donors and correlated inversely with recombination rate (Extended Data Fig. 7b)—results that were robust to chromosome composition and that applied similarly to same-arm and opposite-arm crossover pairs (Extended Data Fig. 7e, f and Supplementary Notes).

The extremely strong correlations of donors’ crossover rates with crossover locations and interference could arise from an underlying biological factor that coordinates these phenotypes, or could arise trivially from the fact that chromosomes with more crossovers would also tend to have crossovers more closely spaced and in more regions. To distinguish between these possibilities, we focused on data from the 180,738 chromosomes with exactly two crossovers (here called ‘two-crossover chromosomes’) (Supplementary Notes). Even in this two-crossover chromosome analysis, distal-zone usage (Fig. 2b) and crossover separation (Fig. 2c) correlated strongly and negatively with genome-wide recombination rate (additional control analyses are described in the Supplementary Notes and Extended Data Fig. 7d, g, h). These relationships indicate that a donor’s crossover-location and crossover-spacing phenotypes reflect underlying biological factors that vary from person to person, as opposed to resulting indirectly from the number of crossovers on a chromosome.

To test whether this covariation of diverse meiotic phenotypes also governs variation at the single-gamete level, we investigated whether cells with more crossovers than the average for their donor also exhibit the same kinds of crossover-spacing and crossover-location

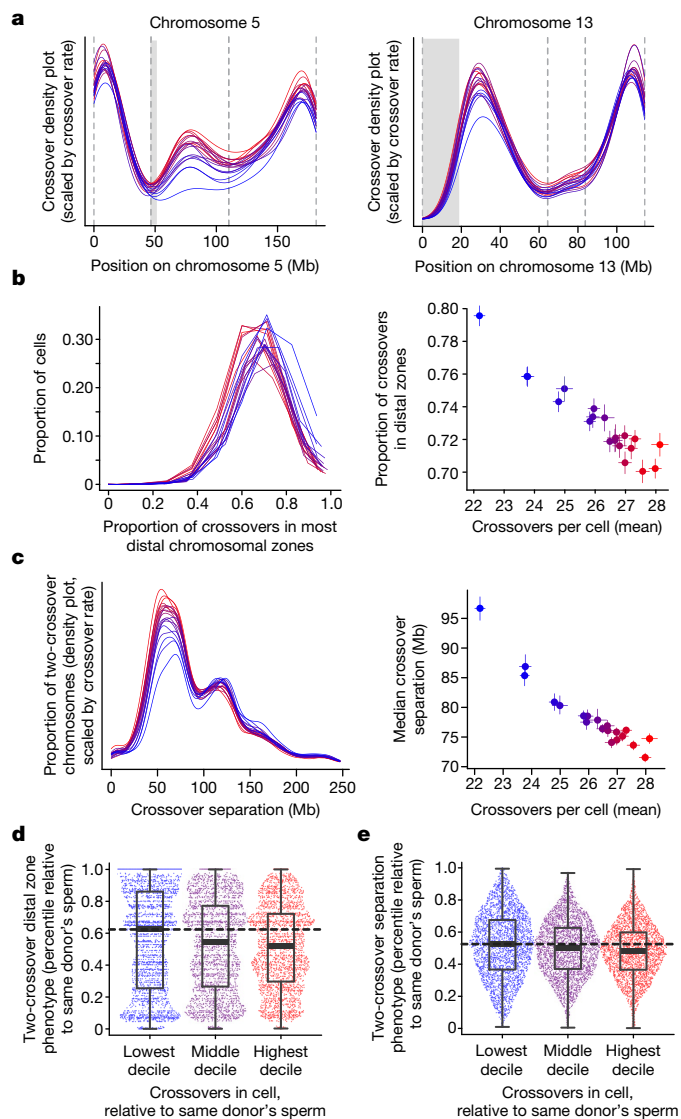


Fig. 2 | Variation in crossover positioning and crossover separation (interference). Colours indicate the crossover rate of donor or cell (blue, low; red, high). **a**, Crossover location density plots for two chromosomes (5 and 13) from each donor ($n = 20$). Dashed grey vertical lines show boundaries between crossover zones. Mb, megabases. **b–e**, Crossover positioning and separation (interference) on chromosomes with two crossovers. **b**, **c**, Interindividual variation among $n = 20$ sperm donors. Error bars show 95% confidence intervals. **b**, Left, per-cell proportion of crossovers in the most distal crossover zones (Kruskal–Wallis chi-squared = 1,034; $df = 19$; $P = 2 \times 10^{-207}$). Right, mean crossover rate versus the proportion of all crossovers (on two-crossover chromosomes) occurring in distal zones (Pearson's $r = -0.95$; two-sided $P = 8 \times 10^{-11}$). **c**, Left, density plot of separation between consecutive crossovers (Kruskal–Wallis chi-squared = 1,792; $df = 19$; $P < 10^{-300}$). Right, mean crossover rate versus median crossover separation on two-crossover chromosomes (Pearson's $r = -0.95$; two-sided $P = 7 \times 10^{-11}$). **d**, **e**, Among-cell covariation of crossover rate with distal zone use (**d**) or crossover interference (**e**). Phenotypes are analysed as percentiles relative to sperm from the same donor. Box plots: midpoints, medians; boxes, 25th and 75th percentiles; whiskers, minima and maxima. **d**, Single-cell distal-zone use (the proportion of crossovers on two-crossover chromosomes that are in the most distal zones) versus crossover rate (n cells per decile = 3,152, 3,080 and 3,101 for first, fifth and tenth deciles, respectively; Mann–Whitney $W = 5,271,934.5$; two-sided $P = 2 \times 10^{-9}$ between first and tenth deciles). **e**, Single-cell crossover separation (the median of all fractions of a chromosome separating consecutive two-crossover chromosome crossovers in each cell) versus crossover rate (Mann–Whitney $W = 148,548,161$, two-sided $P = 3 \times 10^{-53}$ between first ($n = 11,658$) and tenth ($n = 23,154$) deciles; all inter-crossover separations used in test).

phenotypes that donors with high crossover rates do (Supplementary Methods). Indeed, two-crossover chromosomes from cells with more crossovers tended to have closer crossover spacing and increased relative use of non-distal zones (Fig. 2d, e and Extended Data Fig. 7i, j; unnormalized results are in the Supplementary Notes). This result indicates that the correlated meiotic-outcome biases that distinguish people from one another also distinguish the gametes within each individual (see Discussion).

Chromosome and sperm donor aneuploidy

Aneuploidy generally arises from a chromosome missegregation that yields two aneuploid cells: one in which that chromosome is absent (a loss), and one in which it is present in two copies (a gain). Among the 31,228 gametes, we found 787 whole-chromosome aneuploidies and 133 chromosome arm-scale gains and losses (2.5% and 0.4% of cells, respectively) (Fig. 3a and Methods). All chromosomes and sperm donors were affected. The sex chromosomes and acrocentric chromosomes had the highest rates of aneuploidy, consistent with estimates based on fluorescence in situ hybridization analysis of chromosomes^{18,19} (Fig. 3b).

The 20 young (18–38-year-old) sperm donors, considered by clinical criteria to have normal-range sperm parameters, exhibited aneuploidy frequencies ranging from 0.010 to 0.046 aneuploidy events per cell (Fig. 3c and Extended Data Table 1). Permutation tests indicated that this 4.5-fold variation in observed aneuploidy rates reflected genuine inter-individual variation (one-sided $P < 0.0001$) (Supplementary Notes).

Under the prevailing model for the origins of aneuploidy, sperm with chromosome losses and gains should be equally common. However, we observed 2.4-fold more chromosome losses than chromosome gains (554 losses versus 233 gains; proportion test two-sided $P = 2 \times 10^{-30}$). This asymmetry did not appear to reflect technical ascertainment bias (Extended Data Fig. 9a and Supplementary Notes). This result is considered further in the Supplementary Discussion.

Errors in chromosome segregation can occur at meiosis I, when homologues generally separate, or at meiosis II, when sister chromatids separate. Because recombination occurs in meiosis I before disjunction but does not occur at centromeres, errors during meiosis I result in chromosomes with different (homologous) haplotypes at their centromeres, whereas sister chromatids nondisjoin in meiosis II have the same (sister) haplotype at their centromeres (Fig. 3a). (Sex chromosomes X and Y disjoin in meiosis I, and the sister chromatids of X and Y disjoin at meiosis II.) Encouragingly, for chromosome 21—the principal chromosome for which earlier estimates were possible—our finding of 33% meiosis I events and 67% meiosis II events matched previous estimates from trisomy 21 patients with paternal-origin gains³⁶.

Across all chromosomes, meiosis I gains and meiosis II gains had very different relative frequencies in different individuals and on different chromosomes (Fig. 3d, e). For example, sex chromosomes were 2.2 times more likely to be affected in meiosis I than meiosis II, whereas autosomes were 2.0 times more likely to be affected in meiosis II than meiosis I (proportion test two-sided $P = 1.3 \times 10^{-6}$). The lack of correlation between meiosis I and meiosis II vulnerabilities (Fig. 3d, e) indicated that meiosis I and II are differentially challenging to different chromosomes and to different people.

Although crossovers are required for proper chromosomal segregation³⁷ and seem to be protective against nondisjunction in maternal meiosis, in which chromosomes are maintained in diplotene of meiosis I for decades⁸, the relationship of crossovers to aneuploidy is less clear in paternal meiosis^{24,36,38–41}. We found that chromosome gains originating in meiosis I—when recombination occurs—had 36% fewer total crossovers than matched, well-segregated chromosomes did (Supplementary Methods), suggesting that crossovers protected against meiosis I nondisjunction of the chromosomes on which they occurred (Extended Data Fig. 9b and Supplementary Notes). No similar

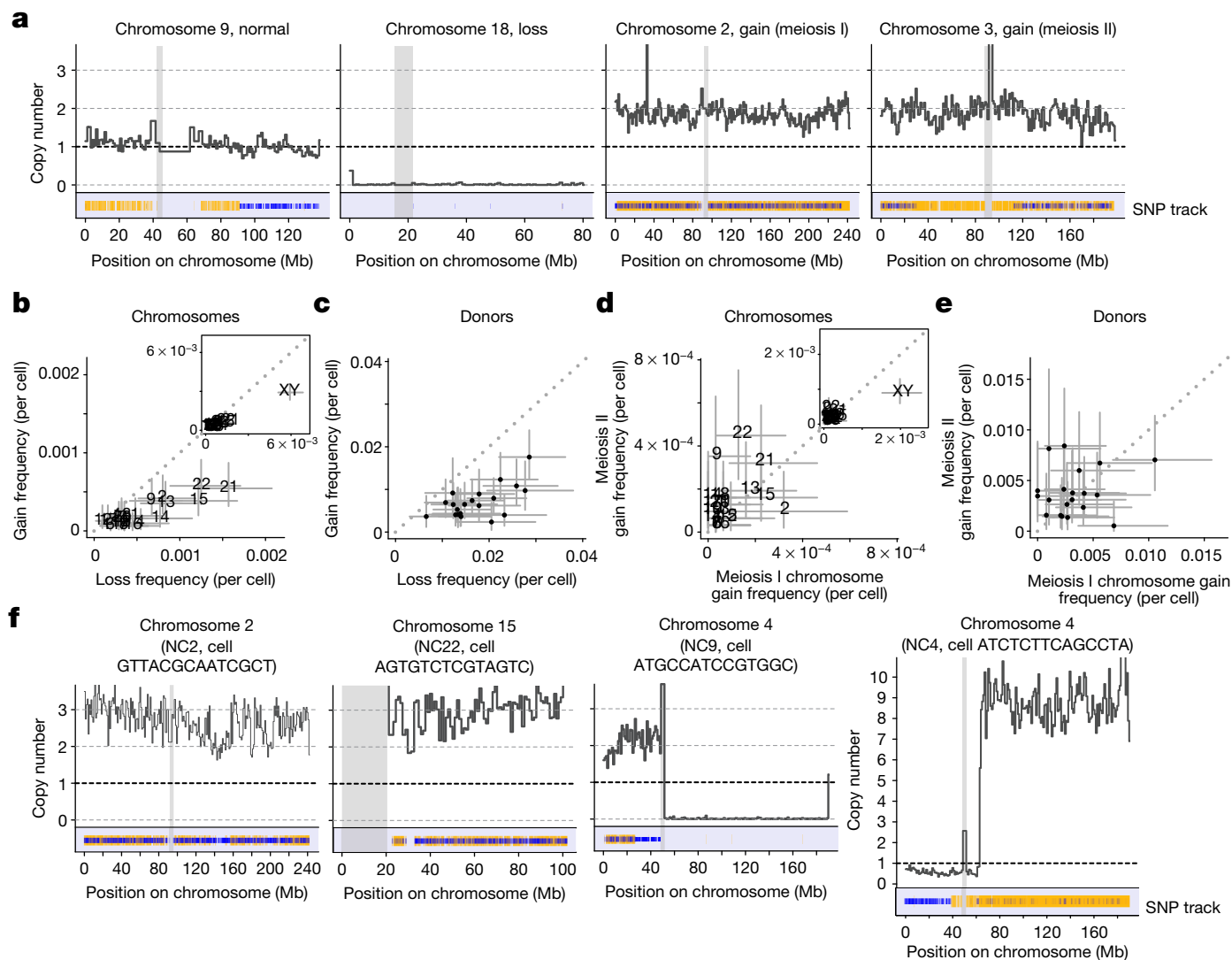


Fig. 3 | Aneuploidy in sperm from 20 sperm donors. **a**, Example chromosomal ploidy analyses. Thick dark grey line, DNA copy number measurement (normalized sequence coverage in 1-Mb bins); blue (haplotype 1) and yellow (haplotype 2) vertical lines, observed heterozygous SNP alleles, plotted with 90% transparency; grey vertical boxes, centromeres (based on the reference hg38 human genome). **b–e**, Frequencies (number of events divided by number of cells) of various aneuploidy categories. **b**, **d**, $n = 23$ chromosomes; **c**, **e**, $n = 20$ donors. Error bars are 95% binomial confidence intervals. **b**, Frequencies of whole-chromosome losses versus gains for each chromosome (excluding XY chromosomes, Pearson's $r = 0.88$, two-sided $P = 7 \times 10^{-8}$; including XY chromosomes (inset), Pearson's $r = 0.99$, two-sided $P < 10^{-300}$). **c**,

Per-sperm-donor aneuploidy rates (excluding XY (not shown), Pearson's $r = 0.51$, two-sided $P = 0.02$; including XY, Pearson's $r = 0.62$, two-sided $P = 0.003$). **d**, Frequencies of whole-chromosome gains occurring during meiosis I versus meiosis II for each chromosome (excluding XY, Pearson's $r = 0.32$, two-sided $P = 0.15$; including XY (inset), Pearson's $r = 0.85$, two-sided $P = 3 \times 10^{-7}$). **e**, Frequencies of whole-chromosome gains occurring during meiosis I versus II for each donor (excluding XY (not shown), Pearson's $r = 0.06$, two-sided $P = 0.80$; including XY, Pearson's $r = 0.17$, two-sided $P = 0.47$). **f**, Example genomic anomalies detected in sperm cells, plotted as in **a**. NC2, NC4, NC9 and NC22 signify individual sperm donors; cells are identified by 14-bp DNA barcode sequences.

relationship was observed for meiosis II gains (although the simulated control distribution for meiosis II is inherently less accurate; Supplementary Notes) or at other levels of aggregation (Extended Data Fig. 9b–d and Supplementary Notes).

Other chromosome-scale genomic anomalies

Many sperm had complex patterns of aneuploidy that could not be explained by the canonical single-chromosome missegregation event. We detected 19 gametes that had three, instead of one, copies of entire or nearly entire chromosomes (2, 15, 20 and 21; Fig. 3f and Extended Data Fig. 10a, b). Chromosome 15 was particularly likely to be present in two extra copies; in fact, sperm with three copies of all or most of chromosome 15 ($n = 10$) outnumbered sperm with two copies of chromosome

15 ($n = 2$) (Fisher's exact test versus Poisson two-sided $P = 2 \times 10^{-7}$) (Supplementary Notes).

Other gametes carried anomalies encompassing incomplete chromosomes. These included: one cell that gained the *p* arm of chromosome 4 while losing the *q* arm; cells with gains of two copies of a chromosome arm; and cells with losses of chromosome arms (Fig. 3f and Extended Data Fig. 10c, d). One cell carried at least eight copies of most of the *q* arm of chromosome 4 (Fig. 3f). This gamete—which we estimate contained almost a billion base pairs of extra DNA—carried both parental haplotypes of chromosome 4, though almost all of the roughly eight copies came from just one of the parental haplotypes (93% of observed alleles in the amplified region were haplotype 2). It is likely that diverse mutational processes generate these genomic anomalies (Supplementary Discussion).

Discussion

Interindividual variation in crossover rates has previously been inferred from SNP data from families^{2–7,9–12}. Here, highly parallel single-gamete sequencing has revealed that sperm donors with high crossover rates also exhibit closer crossover spacing, even when controlling for the number of crossovers actually made on a chromosome. On the basis of these analyses, we consider it most likely that interindividual variation in crossover interference is the true driver of variation in crossover rate and placement.

These same constellations of correlated meiotic crossover phenotypes—low interference, high rates and use of centromere-proximal zones—tended to characterize the same gametes from any donor. Cells with more crossovers in half of their genome tended to have more crossovers in the other half, to have made consecutive pairs of crossovers closer together in genomic distance—even when making just two crossovers on a chromosome—and to have placed proportionally more of their crossovers in nondistal chromosomal regions.

We considered what could cause these meiotic phenotypes to covary across chromosomes, in individual cells, and among people. The physical length of chromosomes during meiosis, which reflects their compaction, has been observed to vary up to twofold among individual spermatocytes while being strongly correlated across chromosomes in the same spermatocyte; spermatocytes with more-compacted chromosomes also generally have fewer incipient crossovers^{20,21,42}. A unifying model (Extended Data Fig. 11) explains the covariance of these meiotic phenotypes while providing a candidate mechanism for interindividual variation: cell-to-cell variation in the compaction of meiotic chromosomes—and person-to-person variation in the average degree of this compaction—would cause these phenotypes to covary in the manner observed in Fig. 2b–e.

Our enthusiasm for this model relies on several additional earlier observations (Extended Data Fig. 11). First, at a cellular level, crossover interference occurs as a function of physical (micrometre-scale) distance along the meiotic chromosome axis or synaptonemal complex, rather than as a function of genomic (base-pair) distance^{43–45}. Second, the first crossover on a chromosome is more likely to occur distally^{13–15}. Such a model also predicts a shared mechanism for sex differences in recombination rates and interindividual variation among individuals of the same sex: oocytes have a longer synaptonemal complex, more crossovers and decreased crossover interference (as measured in genomic distances) than spermatocytes, but have the same synaptonemal-complex length extent of crossover interference^{22,42,46,47}.

Human genetics research has revealed that recombination phenotypes are heritable and associate with common variants at many genomic loci^{3–6,11,12}. A recent genome-wide association study found that variation in crossover rate and placement is associated with variants near genes that encode components of the synaptonemal complex, which connects and compacts meiotic chromosomes, and with genes involved in the looping of homologues along the chromosome axis³. Our model predicts that inherited genetic variation at these loci may bias the average degree of compaction of meiotic chromosomes; the fact that this same property varies among cells from the same donor^{20,21} shows that variance is well-tolerated and compatible with diverse-but-successful meiotic outcomes.

The sharing of covarying phenotypes between the single-cell and person-to-person levels suggests that a core biological mechanism shapes both inter- and intra-individual (single-cell) variation in meiotic outcomes. Such parallelisms between cell-biological and human-biological variation could in principle exist in a wide variety of biological contexts.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2347-0>.

1. Broman, K. W. & Weber, J. L. Characterization of human crossover interference. *Am. J. Hum. Genet.* **66**, 1911–1926 (2000).
2. Coop, G., Wen, X., Ober, C., Pritchard, J. K. & Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**, 1395–1398 (2008).
3. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
4. Kong, A. et al. A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
5. Kong, A. et al. Common and low-frequency variants associated with genome-wide recombination rate. *Nat. Genet.* **46**, 11–16 (2014).
6. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
7. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
8. 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).
9. 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).
10. Cheung, V. G., Burdick, J. T., Hirschmann, D. & Morley, M. Polymorphic variation in human meiotic recombination. *Am. J. Hum. Genet.* **80**, 526–530 (2007).
11. Chowdhury, R., Bois, P. R., Feingold, E., Sherman, S. L. & Cheung, V. G. Genetic analysis of variation in human meiotic recombination. *PLoS Genet.* **5**, e1000648 (2009).
12. Fledel-Alon, A. et al. Variation in human recombination rates and its genetic determinants. *PLoS ONE* **6**, e20321 (2011).
13. Brown, P. W. et al. Meiotic synapsis proceeds from a limited number of subtelomeric sites in the human male. *Am. J. Hum. Genet.* **77**, 556–566 (2005).
14. Gruhn, J. R. et al. Correlations between synaptic initiation and meiotic recombination: a study of humans and mice. *Am. J. Hum. Genet.* **98**, 102–115 (2016).
15. Gruhn, J. R., Rubio, C., Broman, K. W., Hunt, P. A. & Hassold, T. Cytological studies of human meiosis: sex-specific differences in recombination originate at, or prior to, establishment of double-strand breaks. *PLoS ONE* **8**, e85075 (2013).
16. Baudat, F. & de Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Res.* **15**, 565–577 (2007).
17. Plug, A. W., Xu, J., Reddy, G., Golub, E. I. & Ashley, T. Presynaptic association of Rad51 protein with selected sites in meiotic chromatin. *Proc. Natl Acad. Sci. USA* **93**, 5920–5924 (1996).
18. Ioannou, D., Fortun, J. & Tempest, H. G. Meiotic nondisjunction and sperm aneuploidy in humans. *Reproduction* **157**, R15–R31 (2018).
19. Templado, C., Uroz, L. & Estop, A. New insights on the origin and relevance of aneuploidy in human spermatozoa. *Mol. Hum. Reprod.* **19**, 634–643 (2013).
20. Lynn, A. et al. Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science* **296**, 2222–2225 (2002).
21. Wang, S. et al. Per-nucleus crossover covariation and implications for evolution. *Cell* **177**, 326–338 (2019).
22. Hou, Y. et al. Genome analyses of single human oocytes. *Cell* **155**, 1492–1506 (2013).
23. Kirkness, E. F. et al. Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res.* **23**, 826–832 (2013).
24. Lu, S. et al. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338**, 1627–1630 (2012).
25. Ottoloni, C. S. et al. Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. *Nat. Genet.* **47**, 727–735 (2015).
26. Wang, J., Fan, H. C., Behr, B. & Quake, S. R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* **150**, 402–412 (2012).
27. Miller, D., Brinkworth, M. & Iles, D. Paternal DNA packaging in spermatozoa: more than the sum of its parts? DNA, histones, protamines and epigenetics. *Reproduction* **139**, 287–301 (2010).
28. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
29. Zheng, G. X. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
30. 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).
31. Berg, I. L. et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat. Genet.* **42**, 859–863 (2010).
32. 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).
33. Hinch, A. G. et al. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science* **363**, eaau8861 (2019).
34. Housworth, E. A. & Stahl, F. W. Crossover interference in humans. *Am. J. Hum. Genet.* **73**, 188–197 (2003).
35. Sun, F. et al. Human male recombination maps for individual chromosomes. *Am. J. Hum. Genet.* **74**, 521–531 (2004).
36. Oliver, T. R. et al. Investigation of factors associated with paternal nondisjunction of chromosome 21. *Am. J. Med. Genet. A* **149A**, 1685–1690 (2009).

37. Page, S. L. & Hawley, R. S. Chromosome choreography: the meiotic ballet. *Science* **301**, 785–789 (2003).
38. Sun, F. et al. The relationship between meiotic recombination in human spermatocytes and aneuploidy in sperm. *Hum. Reprod.* **23**, 1691–1697 (2008).
39. Ferguson, K. A., Wong, E. C., Chow, V., Nigro, M. & Ma, S. Abnormal meiotic recombination in infertile men and its association with sperm aneuploidy. *Hum. Mol. Genet.* **16**, 2870–2879 (2007).
40. Ma, S., Ferguson, K. A., Arsovska, S., Moens, P. & Chow, V. Reduced recombination associated with the production of aneuploid sperm in an infertile man: a case report. *Hum. Reprod.* **21**, 980–985 (2006).
41. Savage, A. R. et al. Elucidating the mechanisms of paternal non-disjunction of chromosome 21 in humans. *Hum. Mol. Genet.* **7**, 1221–1227 (1998).
42. Tease, C. & Hultén, M. A. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenet. Genome Res.* **107**, 208–215 (2004).
43. Wang, S., Zickler, D., Kleckner, N. & Zhang, L. Meiotic crossover patterns: obligatory crossover, interference and homeostasis in a single process. *Cell Cycle* **14**, 305–314 (2015).
44. Zhang, L., Liang, Z., Hutchinson, J. & Kleckner, N. Crossover patterning by the beam-film model: analysis and implications. *PLoS Genet.* **10**, e1004042 (2014).
45. Zhang, L. et al. Topoisomerase II mediates meiotic crossover interference. *Nature* **511**, 551–556 (2014).
46. Billings, T. et al. Patterns of recombination activity on mouse chromosome 11 revealed by high resolution mapping. *PLoS ONE* **5**, e15340 (2010).
47. Petkov, P. M., Broman, K. W., Szatkiewicz, J. P. & Paigen, K. Crossover interference underlies sex differences in recombination rates. *Trends Genet.* **23**, 539–542 (2007).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

A companion protocol for generating single-sperm libraries using the methods presented here is available via Protocol Exchange⁴⁸. Custom scripts (available via Zenodo⁴⁹) are referenced by name in the Methods sections describing the relevant analyses. Recombination and aneuploidy data generated by the methods described are also publicly available⁵⁰. All statistical analyses were performed in *R* unless otherwise noted. Details of further analysis methods are provided in the Supplementary Methods.

Sample information

Sperm samples from 20 anonymous, karyotypically normal sperm donors were obtained from New England Cryogenic Center under a 'not human subjects' determination from the Harvard Faculty of Medicine Office of Human Research Administration (protocols M23743-101 and IRB16-0834). Donors consented at the time of initial donation for samples to be used for research purposes. The 'not human subjects' determination was based on the use of discarded biospecimens for which research consent had been obtained, and on the fact that researchers had no interactions with the biospecimen donors and no access to identifiable information about the biospecimens. The reviewing committee also reviewed and approved our deposition of the data into a National Institutes of Health (NIH) repository. All experiments were performed in accordance with all relevant guidelines and regulations. (Specimens can be obtained from the New England Cryogenic Center upon Institutional Review Board (IRB) approval.) No statistical methods were used to pre-determine sample size. As no conditions or experimental groups were analysed for this study, no randomization or blinding was performed.

Samples arrived in liquid nitrogen in 'egg yolk buffer' or 'standard buffer with glycerol' (no further buffer information provided), and were aliquoted and stored in liquid nitrogen in the same buffers.

Per sperm-bank policy, donors are 18–38 years old at the time of donation and the precise age of donors is not released. Donor identifiers used here were created specifically for this study and are not linked to any external identifiers.

ddPCR to evaluate genome accessibility

To evaluate how often regions from two different chromosomes co-occurred (as would be expected from cells), we performed droplet digital polymerase chain reaction (ddPCR) with naked DNA, untreated sperm cells or sperm cells decondensed as described below but with variable heat incubation times. For each assay targeting each chromosome, we created a 20× assay mix by combining 25.2 µl of 100 µM forward primer (from IDT), 25.2 µl of 100 µM reverse primer (IDT) and 7 µl of 100 µM probe (IDT for fluorescein amidite (FAM)-labelled probes; Life Technologies for VIC-labelled probes) with 82.6 µl ultrapure water. We carried out ddPCR as previously described⁵¹, following section 3.2 steps 4–12, but with untreated sperm or sperm DNA florets as input instead of DNA.

For this analysis, we targeted chromosome 7 with an assay directed to intergenic region chr7:106552149–106552176 (hg38): forward primer sequence CGTAATGGGGCACAGGGATATA; reverse primer sequence CTGTGAGAGGTAGAGAATCGCC; probe sequence CACAGAGTCCATTTGCAGCACCTCAGT; probe fluorophore FAM. We targeted chromosome 10 with an assay for the *RPP30* gene at chr10:92631759–92631820: forward primer sequence GATTTGGA CCTGCGAGCG; reverse primer sequence GCGGCTGTCTCCACAAGT; probe sequence CTGACCTGAAGGCTCT; probe fluorophore VIC. We calculated the percentage of molecules expected to be linked from each reaction as previously described⁵².

Sperm cell library generation

We generated accessible sperm nuclei 'florets' using a combination of published decondensation protocols^{53,54} with some modifications.

Sperm aliquots containing more than 200,000 cells were thawed on ice and then washed by spinning for 10 min at 400g at 4 °C. The pellet was resuspended in 10 µl phosphate-buffered saline (PBS, Gibco/Life Technologies) and recentrifuged under the same conditions. The sperm pellet was resuspended in 2.5 µl of a sucrose buffer containing 250 mM sucrose (Sigma), 5 mM MgCl₂ (Sigma) and 10 mM Tris HCl (pH 7.5, Thermo Scientific). Sperm aliquots were submerged in liquid nitrogen and immediately quick-thawed by holding them in a warm fist; three such freeze–thaw cycles were performed.

Freeze-thawed sperm solution was combined with 22.5 µl decondensation buffer (113 mM KCl (Sigma), 12.5 mM KH₂PO₄ (Sigma), 2.5 mM Na₂HPO₄ (Sigma), 2.5 mM MgCl₂ (Sigma) and 20 mM Tris (Thermo Scientific) freshly supplemented with 150 µM heparin (sodium salt from porcine, Sigma catalogue number H3393) and 2 mM β-mercaptoethanol (Sigma)). The reaction was incubated at 37 °C for 45 min. To allow enzymatic DNA amplification, heparin was inactivated by mixing the sperm solution with 0.5 U heparinase I (Sigma H2519) by gently pipetting and incubating at room temperature for 2 h (ref. ⁵⁵).

The sperm solution was moved to ice, and sperm floret concentration was determined by diluting 1:100 with PBS and staining with 1× SYBR I (Thermo Scientific), then counting using the green fluorescence channel at 10× magnification.

Droplets were prepared using the following modifications to 10× Genomics' GemCode (version 1; ref. ²⁹) user guide revision C (in place of steps 5.1–5.3.9); all reagents come from the 10× Genomics GemCode kit. Ultrapure water was combined with 10,833 sperm to a final volume of 5 µl; 10,000 sperm were used for library generation. To each sperm sample was added 60 µl of a master mix containing 32.5 µl GemCode reagent mix, 1.5 µl primer release agent, 9.2 µl GemCode polymerase and 16.8 µl ultrapure water.

GemCode beads were vortexed at full speed for 25 s, and then diluted 1:11 with ultrapure water to a total volume of at least 90 µl per sample. Per 10× Genomics' GemCode's protocol, 60 µl of the sample-master mix combination was added to the droplet generation chip, followed by 85 µl of freshly pipette-mixed 1:11-diluted bead mixture and 150 µl of droplet generation oil.

Droplets were generated and processed by library generation following 10× Genomics' GemCode (version 1) user guide revision C (step 5.3.10 through to the end of section 6).

Sequencing and sequence data processing

We generated two libraries per sperm donor and additional libraries for four initial samples with low cell counts. We sequenced four or five libraries at a time on S2 200 cycle flow cells on an Illumina NovaSeq. The read structure was 178 cycles for read 1, 8 cycles for read 2 (index read one), 14 cycles for read 3 (index read two containing the cell barcode; later treated as the reverse read), and 5 cycles for read 4 (unused; included to fulfil the NovaSeq's paired-end requirement).

To convert the data to mapped binary alignment map (BAM) files with cell and molecular barcodes encoded as read tags, we used Picard Tools v.2.2 (<http://broadinstitute.github.io/picard>) and Drop-seq Tools v.2.2 (<https://github.com/broadinstitute/Drop-seq/releases>; see https://github.com/broadinstitute/Drop-seq/blob/master/doc/Drop-seq_Alignment_Cookbook.pdf for details on running many of the tools)²⁸.

Illumina binary base call (BCL) files were converted to unmapped BAM files using Picard's ExtractIlluminaBarcodes and IlluminaBasecallsToSam with read structure 178T8B14T (cell barcodes, present in the i5 index, were incorporated as read 2 for ease of downstream processing). BAMs were processed to include unique molecular identifiers (UMIs) and cell barcodes as read tags, and to exclude reads with poor-quality cell barcodes or UMIs; consequently, each read was retained as single-end with a 14-base-pair (bp) cell barcode stored in tag XC and a 10-bp molecular barcode/UMI stored in tag XM. The first 10 bp of read 1 were used as the UMI. First, DropSeq

Article

Tools' TagBamWithReadSequenceExtended was called with `BASE_RANGE = 1-14`, `BASE_QUALITY = 10`, `BARCODED_READ = 2`, `DISCARD_READ = true`, `TAG_NAME = XC`, `NUM_BASES_BELOW_QUALITY = 1`. Subsequently, TagBamWithReadSequenceExtended was called again with `BASE_RANGE = 1-10`, `BASE_QUALITY = 10`, `HARD_CLIP_BASES = true`, `BARCODED_READ = 1`, `DISCARD_READ = false`, `TAG_NAME = XM`, `NUM_BASES_BELOW_QUALITY = 1`. Finally, DropSeq Tools' FilterBAM was called with parameter `TAG_REJECT = XQ`.

Reads were aligned to hg38 using bwa mem⁵⁶ v.0.7.7-r441. BAMs were converted to FastQ using Picard's SamToFastQ, FastQ reads were aligned using bwa mem -M, and then unmapped BAMs were merged with mapped BAMs using Picard's MergeBamAlignment, with non-default options `INCLUDE_SECONDARY_ALIGNMENTS = false` and `PAIRED_RUN = false`. Reads were marked PCR duplicates using Drop-seq Tools' SpermSeqMarkDuplicates (part of Drop-seq tools v2.2 and above) with options `STRATEGY = READ_POSITION`, `CELL_BARCODE_TAG = XC`, `MOLECULAR_BARCODE_TAG = XM`, `NUM_BARCODES = 20000`, `CREATE_INDEX = true`. BAM files for all lanes and index sequences from the same sample were merged using Picard's MergeSamFiles before alignment and/or during duplicate marking with all BAMs given as input to SpermSeqMarkDuplicates.

Variant calling and sperm cell genotyping

For each donor, we pooled all reads from all libraries, including reads that did not derive from a barcode associated with a complete sperm cell. Using GATK v.3.7 (refs. ^{57,58}) in hg38, we followed GATK's best-practices documentation for base quality score recalibration; for genomic variable call format (gVCF) generation using HaplotypeCaller (in 'discovery' mode with `-stand_call_conf 20`); and for joint genotyping with GenotypeGVCFs. We filtered variants with `SelectVariants -selectType SNP` and `VariantFiltration (-filterExpression 'QD < 3.0')`. We then performed variant quality score recalibration (VQSR) following GATK's best practices, except that we excluded annotations MQ and DP (VariantRecalibrator with GATK provided resources; `-an QD, MQRankSum, ReadPosRankSum, FS and SOR`; `-mode SNP`; `-trustAllPolymorphic`; and tranches 90, 99.0, 99.5, 99.9 and 100.0). We applied tranche 99.9 recalibration using `ApplyRecalibration -mode SNP` and obtained the names of SNPs from SNP database (dbSNP) build 146 (ref. ⁵⁹) using `VariantAnnotator -dbSNP`. We filtered our sites to contain only those biallelic SNPs that were present in Hardy-Weinberg equilibrium in 1000 Genomes Phase 3 (ref. ⁶⁰) using `SelectVariants -concordance with a VCF containing only these sites (from GATK's resource bundle)`. We excluded SNPs in centromeric regions or acrocentric arms, as defined by the University of California Santa Cruz (UCSC)'s Genome Browser's cytoband track^{61,62} (<http://genome.ucsc.edu>; the same centromere boundaries were used in all analyses), and those in known paralogous regions⁶³. We selected only heterozygous SNPs using `SelectVariants -selectType SNP -selectTypeToExclude INDEL -restrictAllelesTo BIALLELIC -excludeFiltered -setFilteredGtToNocall -selectexpressions 'vc.getGenotype(""<sample name>"").isHet()'`.

We identified the SNPs present in each sperm cell and the allele that was present using GenotypeSperm (part of Drop-seq Tools v.2.2 and above). For downstream analyses, we generated a file with columns cell, pos and gt, with gt having the value 0 for the reference allele and 1 for the alternate allele for SNPs that had one or more UMIs covering only one base matching the reference or alternate allele (see our script `gtypesperm2cellsbyrow.R`).

Chromosome-scale phasing

We identified barcodes that were potentially associated with cells by plotting the cumulative fraction of reads associated with each ranked barcode and identifying the inflection point of this curve (Extended Data Fig. 1f). We then included only those barcodes that had substantial read depth on either the X or the Y chromosome but not both, as the vast majority of sperm cells should contain only one sex chromosome.

(We later added these barcodes back in before formally identifying and excluding cell doublets.)

To phase sperm donors' genomes, we used all quality-controlled heterozygous sites in these cell barcodes expected to correspond to sperm cells, excluding observations of SNPs for which the observed allele was not the reference or alternate allele in the parental genome, or for which more than one allele was observed. For each chromosome, we converted per-cell SNP calls into 'fragments' for input into the HapCUT phasing software^{64,65} by considering each consecutive pair of SNPs observed in a cell to be a fragment (see our script `gtypesperm2mf.R`). We then used HapCUT with parameter `-maxiter 100` to generate chromosomal phase. After identifying and removing cell doublets (see below), we repeated phasing with only non-doublet cell barcodes.

To validate our phasing method, we simulated single-cell SNP observations from known haplotypes, including 2% genotype errors and a variable percentage of cell doublets. In brief, sites were randomly sampled from one known haplotype of chromosome 17 until a crossover location was probabilistically assigned on the basis of the deCODE recombination map⁶, then sampled from the other haplotype (one crossover was simulated per cell). To simulate PCR or sequencing errors, 2% of the sites were randomly assigned to an allele. Doublets were simulated by combining two cells and retaining 70% of the observed sites at random. We performed five random simulations for each doublet proportion, for the mean proportion of sites 'observed' in each cell, and for the number of cells simulated, and then followed our phasing protocol using each simulation (see our script `simulatespermseqfromhaps.py`).

To further validate phasing, we used Sperm-seq data to phase one donor's genome and compared these phased haplotypes to this donor's Eagle^{66,67}-generated haplotypes. We compared the phase relationship between each consecutive pair of SNPs (identifying the proportion of switch errors between the two phased sets). We also compared the Sperm-seq allele-allele phase of all pairs of alleles in perfect linkage disequilibrium in 1000 Genomes Phase 3 (ref. ⁶⁰) in the populations matching the donor's ancestry.

Cell doublets

To identify cell barcodes associated with more than one sperm cell (cell doublets), we detected consecutively observed SNP alleles that appeared on different parental haplotypes, which could occur because of crossover, error, or the presence of two haplotypes in the same droplet (doublet). We ranked barcodes by the proportion of consecutive SNPs that spanned haplotypes by using all SNPs from all autosomes except the autosome with the most haplotype-spanning consecutive SNPs (so as to avoid mistakenly identifying cells with chromosome gains as doublets); this resulted in a clear inflection point wherein cell doublets had a quickly accelerating proportion of haplotype-spanning consecutive SNPs (Extended Data Fig. 2d-f). All cell barcodes below this inflection point (identified with the function 'ede' from the R package 'inflection' <https://CRAN.R-project.org/package=inflection>) were considered non-doublet (Extended Data Fig. 2f) (see our script `computeSwitchesandInfThresh.R`). Even though we specifically exclude the autosome with the most haplotype-spanning consecutive SNPs from doublet identification, any cells with multiple chromosome gains (especially more than two) or whole-genome diploidy would be excluded by this method.

Crossover events

We identified crossover events on all autosomes (but excluded the *p* arms of acrocentric chromosomes for which SNPs were excluded from analysis) by finding transitions between tracts of SNPs with alleles that match different parental haplotypes using a hidden Markov model written in R with package 'HMM' (<https://CRAN.R-project.org/package=HMM>). To ensure that we detected crossovers located near the ends of SNP coverage (subtelomeric

regions are frequently used for crossovers in spermatogenesis), we ran the HMM in both the forward-chromosomal and the reverse-chromosomal directions, with the start probability for one haplotype equal to 1 if the first two SNPs observed were of that haplotype. In addition to two states for parental haplotypes, we included a third ‘error’ state to capture cases in which a haplotype 1 allele is observed in a haplotype 2 region (and vice versa), for example, owing to PCR or sequencing error, gene conversion, or cases in which a small piece of off-haplotype ambient DNA was captured in a droplet. Crossovers were where one haplotype transitioned to another, or where one haplotype transitioned to the error state and then to the other haplotype. Crossover boundaries were the last SNP in the first haplotype and the first in the next. The key parameters for this algorithm are the transition probability between haplotypes (set to 0.001, from the per-cell median 26 crossovers divided by the per-cell median 24,710 heterozygous SNPs) and transition probability into and out of the ‘error’ state (we set the probability of transition into this state to 0.03 from either haplotype, as only a few percent of SNPs are off-haplotype; we set the probability of staying in error to 0.9 to allow for the occasional tract of SNPs from an ambient piece of off-haplotype DNA). Emission probabilities were 100% haplotype 1 alleles from haplotype 1, 100% haplotype 2 alleles from haplotype 2, and equal probability haplotype 1 or 2 alleles from the third ‘error’ state. Crossover calling was robust to a range of low transition probabilities (see our script `spseqHMMCOaller_3state.R`, which calls crossovers on one chromosome).

After aneuploidy identification, we marked aneuploid chromosomes as having no crossovers for all crossover analyses (absent chromosomes have no crossovers and crossovers are called differently on gained chromosomes, described below).

Identifying even-coverage cell barcodes

We used Genome STRiP v2.0 (<http://software.broadinstitute.org/software/genomestrip/>)^{68,69} to determine sequence read depth (observed number of reads divided by expected number of reads) in bins of 100 kb of uniquely mappable sequence across the genome in each sperm cell, using Genome STRiP’s default guanine–cytosine (GC) bias correction and repetitive region masking for reference genome gr38. We divided the read depth by two to obtain the read depth per haploid rather than diploid genome. Input to Genome STRiP was a BAM file containing only cells of interest, with read groups set to `<sample name>:<cell barcode>` (created using Drop-seq Tools’ `ConvertTagToReadGroup` with options `CELL_BARCODE_TAG=XC`, `SAMPLE_NAME=<name of sample/donor>`, `CREATE_INDEX=true`, and `CELL_BC_FILE=list of barcodes potentially associated with cells`, described above).

A minority of cell barcodes were associated with eccentric read depth across many chromosomes, with wave-like read depth vacillating between 0 and 2 or more. (We hypothesize that these cell barcodes were associated with sperm nuclei that did not properly decondense, such that some regions of the genome were more accessible than others, leading to undulating read depths across more- and less-accessible chromatin.) To identify and exclude such barcodes, we treated read depths across each chromosome as a time series and used Box–Jenkins autoregressive integrated moving average (ARIMA) modelling to model how read depth observations relied on their previous values and their overall averages (implemented via the R package ‘forecast’^{70,71}, excluding differencing). By visual inspection, we determined that chromosomes with certain ARIMA criteria were likely to have an undulating read depth, and that cell barcodes with five or more such identified chromosomes were likely to have eccentric read depths globally. We flagged individual chromosomes if: (1) the sum of the AR1 and AR2 coefficients was greater than 0.7, the AR1 coefficient was greater than 0.9, or the net sum of all AR and MA coefficients was greater than 1.25; and (2) either the net sum of AR and MA coefficients was greater than 0.4 or the intercept was less than 0.8 or greater than 1.2. If both criteria in

(2) were met, this signified an exceedingly odd chromosome, which we counted twice. Cell barcodes with five or more chromosomes flagged in this way were excluded from downstream analyses. (Because gains of large amounts of the genome cause artificially depressed read depths on nongained chromosomes, we manually examined any cells with a large range of ARIMA intercepts and more than five chromosomes denoted as unstable. Any such cells that had simply gained a large proportion of the genome—for example, three copies of chromosome 2—were included rather than excluded). We cross-referenced all cell exclusions with called aneuploidies, confirming that cells were not excluded simply on the basis of having lost or gained a chromosome (see our scripts `setupsgreaddepth.R`, `exclbadreaddepth_arima_1.R`, `exclbadreaddepth_initid_2.R`, and `exclbadreaddepth_finalize_3.R`).

Replicate barcodes (‘bead doublets’)

One sperm cell can be encapsulated in a droplet with more than one barcoded bead. To identify such cases, where pairs of sperm genomes were identical, we determined the proportion of SNPs that were of the same haplotype for each pair of barcodes. We imputed the haplotype of all heterozygous SNPs on the basis of the haplotype of surrounding observed SNPs and locations of recombination events, and compared SNP haplotypes across sperm cell pairs. SNP observations between boundaries of crossovers were excluded from analysis. Sperm cells shared on average 50% of their genomes, but a few sets of barcodes shared nearly 100% of their SNP haplotypes (Extended Data Fig. 3a). We considered these pairs to be ‘bead doublets’ or replicate barcodes. In all downstream analyses, we used only one barcode (chosen randomly) from a set corresponding to the same cell (see our scripts `imputeHaplotypeAllSNPs.R`, `compareSpermHapsPropSNPs.R`, `combineChrsSpermHapsPropSNPs.R`, and `curateNonRepBCList.R`).

Crossover zones

To define regions of recombination use, we found local minima of the density (built-in function in R) of all crossovers’ median positions across all samples on each chromosome. Minima were identified using the `findPeaks` function (from <https://github.com/stas-g/findPeaks>) on the inverse density with $m = 3$. Crossover zones run from the beginning of the chromosome (including the whole p arm for acrocentric chromosomes) to the location of the first local minimum, from the location of the first local minimum plus one base pair to the next local minimum, and so on, with the last zone on each chromosome ending at the chromosome end (see our script `findcozones_peaks.R`).

Aneuploidy and chromosome arm loss/gain

As described previously (see Methods section ‘Identifying even-coverage cell barcodes’), we used Genome STRiP (<http://software.broadinstitute.org/software/genomestrip/>)^{68,69} to determine read depth in each sperm cell in 100-kb bins. We located chromosomes or chromosome arms with aberrant read depth to identify aneuploidy.

We excluded genomic regions that had outlying read depths across all cells, defined as those with $P < 0.05$ in a one-sided one-sample t -test (looking for increased read depth) against the expected mean read depth of 2# (defined below). To identify gains of autosomes, we performed a one-sided one-sample t -test (expecting increased read depth in a gain) for each cell against the expected read depth for a gain of one copy, 2#. For each cell, this analysis compared the distribution of all bins’ read depth across a region of interest to the gain expectation 2#, and flagged any cells whose read depth distributions were not significantly different ($P \geq 0.05$). We used the same approach to identify losses, comparing a cell’s read depth distribution across bins to 0.1 and flagging any that were not significantly higher ($P \geq 0.05$).

The expected copy number for gains is 2, but the expected read depth for gains depends on the size of the chromosome: a library corresponding to a cell with a chromosome gain has more reads than would be in that same library without a gain. This phenomenon pulls the

Article

read depth down globally by increasing the total number of expected reads, causing the denominator in each read depth bin (the expected number of reads in that bin) to increase. Therefore, we computed a chromosome-specific critical read depth value for identifying gains: $2\# = 2 \times (\text{the proportion of the genome in base pairs coming from all chromosomes other than the tested one})$. For losses, we used 0.1 rather than 0 as the expected read depth, because a small number of reads generally align to every chromosome in every library.

For nonacrocentric chromosomes, we performed aneuploidy calling for the arms separately and for the whole chromosome. Because amplification of more than two copies of a chromosome arm could result in the whole chromosome passing the *P*-value threshold, we required a whole-chromosome event to pass the *P*-value threshold at the whole-chromosome level and to have a rounded read depth for both arms of 2 or more for a gain (or 0 for a loss). For the acrocentric chromosomes, only the *q* arm was considered, and any *q* arm gain or loss was considered to be a whole-chromosome event (unless investigated further).

For the sex chromosomes, we followed a similar statistical framework, but a loss was considered an aneuploidy only if both the X and the Y chromosomes were flagged as lost. A gain was called if both X and Y chromosomes were present (see our scripts `setupsreaddepth.R`, `idaneus_initialttests.R`, `curateaneudata_clean.R`, `getautosomalaneumatrix.R` and `getxykaryos_aneus.R` for aneuploidy calling and output formatting; see our scripts `curateAnFreqFromCodeMatrix.R`, `curateInitAnalyzeXYKaryos.R` and `combineAnFreq_AutXY.R` for conversion of outputs of aneuploidy calling to cross-donor aneuploidy frequency tables).

Division of origin for chromosome gains

To see when chromosome gains originated, we determined whether the centromeres of the multiple copies of the chromosomes were heterozygous and therefore from homologues, which typically disjoin in meiosis I, or homozygous and therefore from sister chromatids, which typically disjoin in meiosis II. We identified heterozygous regions for all cells using a hidden Markov model (HMM) in which the states are: (1) heterozygous (emitting either haplotype's alleles), or (2) homozygous (emitting only one haplotype's alleles), with the transition probability between the states equal to the recombination transition probability. For each gain, we determined whether heterozygous tracts overlapped the centromere. If a heterozygous tract started before the start of the centromere and ended after the end of the centromere, or started at the first SNP observed on an acrocentric chromosome or within the first ten SNPs and was more than ten SNPs long, then chromosome was classified as a meiosis I gain; if no heterozygous tract overlapped the centromere, it was classified as a meiosis II gain (see our scripts `getDiploidTracts_hmm.R`, `originOfGainID.R` and `curateOriginMultSamps.R`).

At the sex chromosomes, any XY sex chromosome gain derives from meiosis (X and Y are homologues), whereas an XX or YY gain derives from meiosis II (sister chromatids are duplicated).

Reporting summary

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

Data availability

Crossover and aneuploidy data (individual events and counts per donor and/or per cell), including the source data underlying Figs. 2, 3b–e and Extended Data Figs. 5–9, are available via Zenodo at <https://doi.org/10.5281/zenodo.2581570>. Raw sequence data are available in the Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra>) via the Database of Genotypes and Phenotypes (dbGaP) (<https://www.ncbi.nlm.nih.gov/gap/>) for general research use upon application and approval (study accession number phs001887.v1.p1).

Code Availability

Analysis scripts and documentation are available via Zenodo at <https://doi.org/10.5281/zenodo.2581595>.

48. Bell, A. D., Mello, C. J. & McCarroll, S. A. Sperm-seq wet lab protocol: sperm preparation and droplet-based sequencing library generation. *Protoc. Exch.* <https://doi.org/10.21203/rs.3.pex-823/v1> (2020).
49. Bell, A. D. et al. Analysis scripts for: Insights about variation in meiosis from 31,228 human sperm genomes. Zenodo, <https://doi.org/10.5281/zenodo.2581595> (2019).
50. Bell, A. D. et al. Recombination and aneuploidy data for: Insights about variation in meiosis from 31,228 human sperm genomes. Zenodo, <https://doi.org/10.5281/zenodo.2581570> (2019).
51. Bell, A. D., Usher, C. L. & McCarroll, S. A. Analyzing copy number variation with droplet digital PCR. *Methods Mol. Biol.* **1768**, 143–160 (2018).
52. Regan, J. F. et al. A rapid molecular approach for chromosomal phasing. *PLoS ONE* **10**, e0118270 (2015).
53. Montag, M., Tok, V., Liow, S. L., Bongso, A. & Ng, S. C. In vitro decondensation of mammalian sperm and subsequent formation of pronuclei-like structures for micromanipulation. *Mol. Reprod. Dev.* **33**, 338–346 (1992).
54. Samocha-Bone, D. et al. In-vitro human spermatozoa nuclear decondensation assessed by flow cytometry. *Mol. Hum. Reprod.* **4**, 133–137 (1998).
55. Taylor, A. C. Titration of heparinase for removal of the PCR-inhibitory effect of heparin in DNA samples. *Mol. Ecol.* **6**, 383–385 (1997).
56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
57. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
58. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
59. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
60. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
61. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
62. Tyner, C. et al. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* **45** (D1), D626–D634 (2017).
63. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
64. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
65. Selvaraj, S., Dixon, R. J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
66. Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
67. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
68. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
69. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
70. Hyndman, R. J. & Khandakar, Y. Automatic time series forecasting: the forecast package for R. *J. Stat. Softw.* **26**, 1–22 (2008).
71. Hyndman, R. et al. forecast: forecasting functions for time series and linear models. R package version 8.4, <http://pkg.robjhyndman.com/forecast> (2018).
72. Kauppi, L. et al. Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science* **331**, 916–920 (2011).
73. Kleckner, N., Storlazzi, A. & Zickler, D. Coordinate variation in meiotic pachytene SC length and total crossover/chiasma frequency under conditions of constant DNA length. *Trends Genet.* **19**, 623–628 (2003).
74. Revenkova, E. et al. Cohesin SMC1 beta is required for meiotic chromosome dynamics, sister chromatid cohesion and DNA recombination. *Nat. Cell Biol.* **6**, 555–562 (2004).
75. Zickler, D. & Kleckner, N. Meiotic chromosomes: integrating structure and function. *Annu. Rev. Genet.* **33**, 603–754 (1999).
76. Wang, S. et al. Inefficient crossover maturation underlies elevated aneuploidy in human female meiosis. *Cell* **168**, 977–989 (2017).
77. Blat, Y., Protacio, R. U., Hunter, N. & Kleckner, N. Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell* **111**, 791–802 (2002).

Acknowledgements We thank G. Genovese for suggestions on analyses; E. Macosko for advice on technology development; other members of the McCarroll laboratory, including C. Whelan, S. Burger and B. Handsaker, for advice; M. Daly, J. Hirschhorn, S. Elledge and S. Schilit for their insights; 10x Genomics for discussions about reagents; C. L. Usher and C. K. Patil for contributions to the text and figures; and those who commented on the preprint version of this article for their input. This work was supported by grant R01 HG006855 to S.A.M.; a Broad Institute NextGen award to S.A.M.; and a Harvard Medical School Program in Genetics and Genomics NIH Ruth L. Kirchstein training grant to A.D.B.

Author contributions A.D.B. and S.A.M. conceived and led the studies. A.D.B., S.A.M. and C.J.M. developed the experimental methods. A.D.B. and C.J.M. performed all experiments, generating all data. A.D.B. and S.A.M. designed the strategies for crossover and aneuploidy

analysis, and A.D.B. performed the crossover and aneuploidy analyses. A.D.B., J.N. and A.W. wrote the sequence and variant processing software, pipelines and analytical methods. A.D.B. wrote the software for crossover calling and analysis. A.D.B. and S.A.B wrote the software for aneuploidy calling. A.D.B. and S.A.M. wrote the manuscript with contributions from all authors.

Competing interests A.D.B. and S.A.M. are inventors on a United States Provisional Patent application (PCT/US2019/029427; applicant: President and Fellows of Harvard College), currently in the PCT stage, relating to droplet-based genomic DNA capture, amplification and sequencing that is capable of obtaining high-throughput single-cell sequence from individual mammalian cells, including sperm cells. A.D.B. was an occasional consultant from Ohana

Biosciences between October 2019 and March 2020. The other authors declare no competing interests.

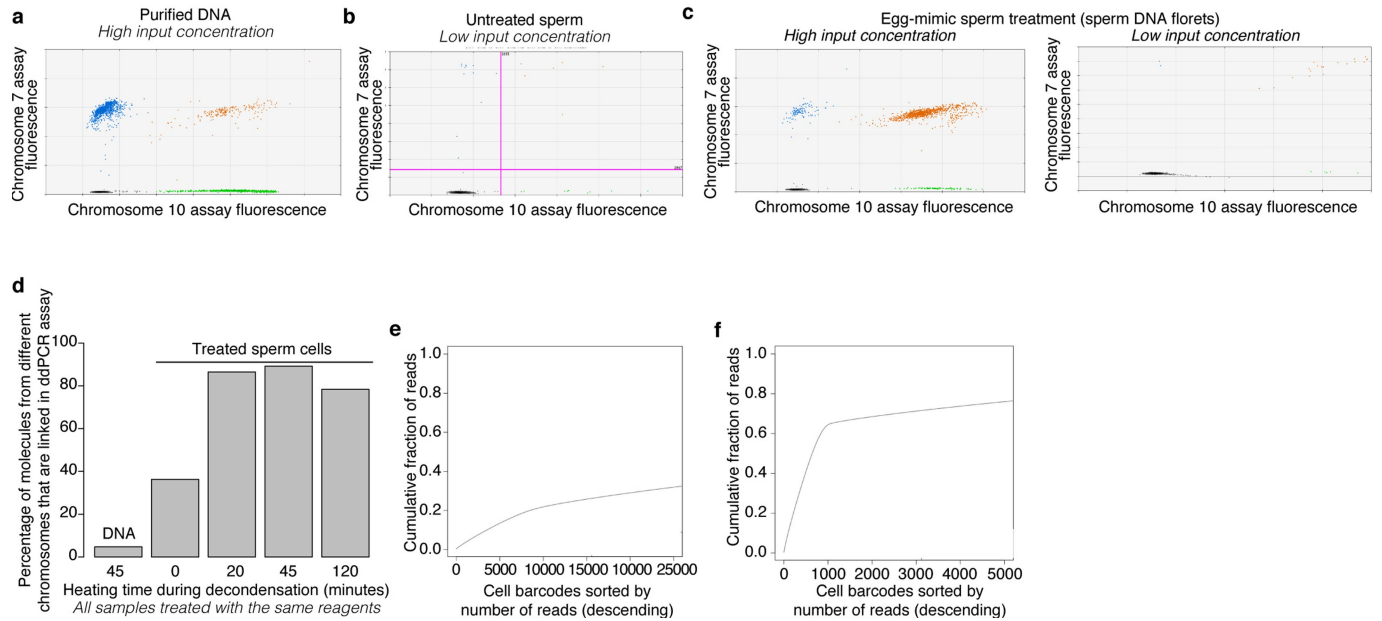
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2347-0>.

Correspondence and requests for materials should be addressed to A.D.B. or S.A.M.

Peer review information *Nature* thanks Donald Conrad, Beth Dumont, Augustine Kong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

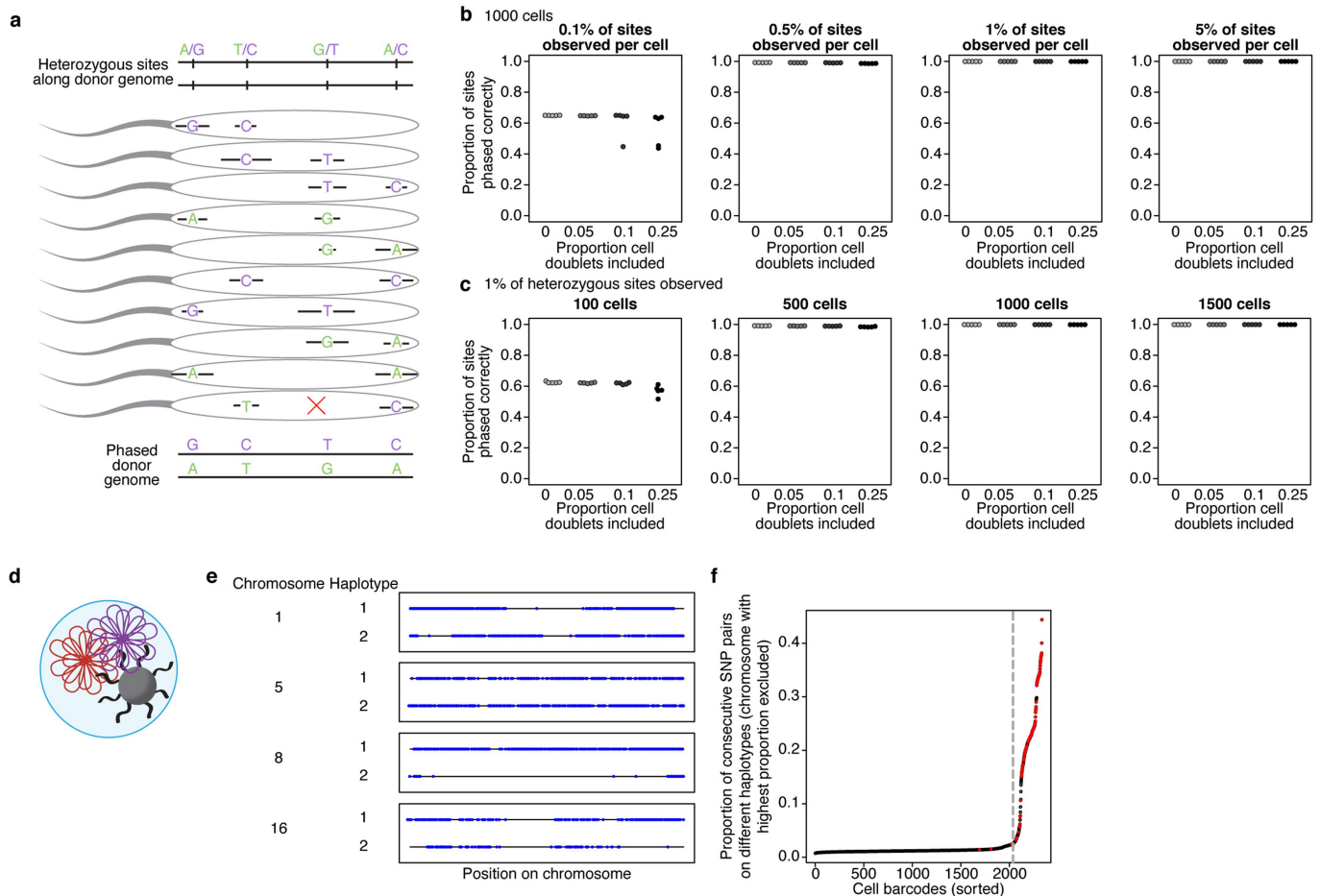
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Characterization of egg-mimic sperm preparation and optimization of bead-based single-sperm sequencing. a–c,

Two-channel fluorescence plots showing the results of ddPCR with the input template noted above each panel, demonstrating that two loci (from different chromosomes) are detectable in the same droplet far more often when sperm DNA florets (rather than purified DNA) are used as input. Each point represents one droplet. Grey points (bottom left) represent droplets in which neither template molecule was detected; blue points (top left) represent droplets in which the assay detected a template molecule for the locus on chromosome 7; green droplets (bottom right) represent droplets in which the assay detected a template molecule for the locus on chromosome 10; and brown points (top right) represent droplets in which both loci were detected. With a high concentration of purified DNA as input (**a**), comparatively fewer droplets contain both loci than when untreated (**b**) or treated (**c**) sperm were used as input. Sperm ‘florets’ treated with the egg-mimicking decondensation protocol had a much higher fraction of droplets containing both loci than did purified DNA (compare **a** with **c**, left) and had more-sensitive ascertainment and cleaner results (quadrant separation) than untreated sperm (compare **b** with **c**, right). The pink lines in **b** delineate the boundaries between droplets categorized as negative or positive for each assay. **d**, Optimization of sperm preparation: characterization of the effect of different lengths of 37 °C incubation of sperm cells treated with egg-mimicking decondensation

reagents on how often the loci on chromosomes 7 and 10 were detected in the same ddPCR droplet. The y-axis shows the percentage of molecules that are calculated to be linked to each other (that is, physically linked in input) for assays targeting chromosomes 7 and 10. Extracted DNA (‘DNA’, a negative control) gives the expected result of random assortment of the two template molecules into droplets. The 45-min heat treatment was used for all subsequent experiments in this study. **e, f**, Distribution of sequence reads across cell barcodes from droplet-based single-sperm sequencing. Each panel shows the cumulative fraction of all reads from a sequencing run coming from each read-number-ranked cell barcode; a sharp inflection point delineates the barcodes with many reads from those with few reads. Points to the left of the inflection point are the cell barcodes that are associated with many reads (that is, beads that are coencapsulated with cells); the height of the inflection point reflects the proportion of the sequence reads that come from these barcodes. Only reads that mapped to the human genome (hg38) and were not PCR duplicates are included. **e**, Data from an initial adaptation of 10× Genomics’ GemCode linked reads system²⁹, where a small proportion of the reads come from cell barcodes associated with putative cells. **f**, Data from the final, implemented adaptation of 10× Genomics’ GemCode linked reads system²⁹ for the same number of input sperm nuclei as in **e**. The x-axis in **f** includes five times fewer barcodes than in **e**.

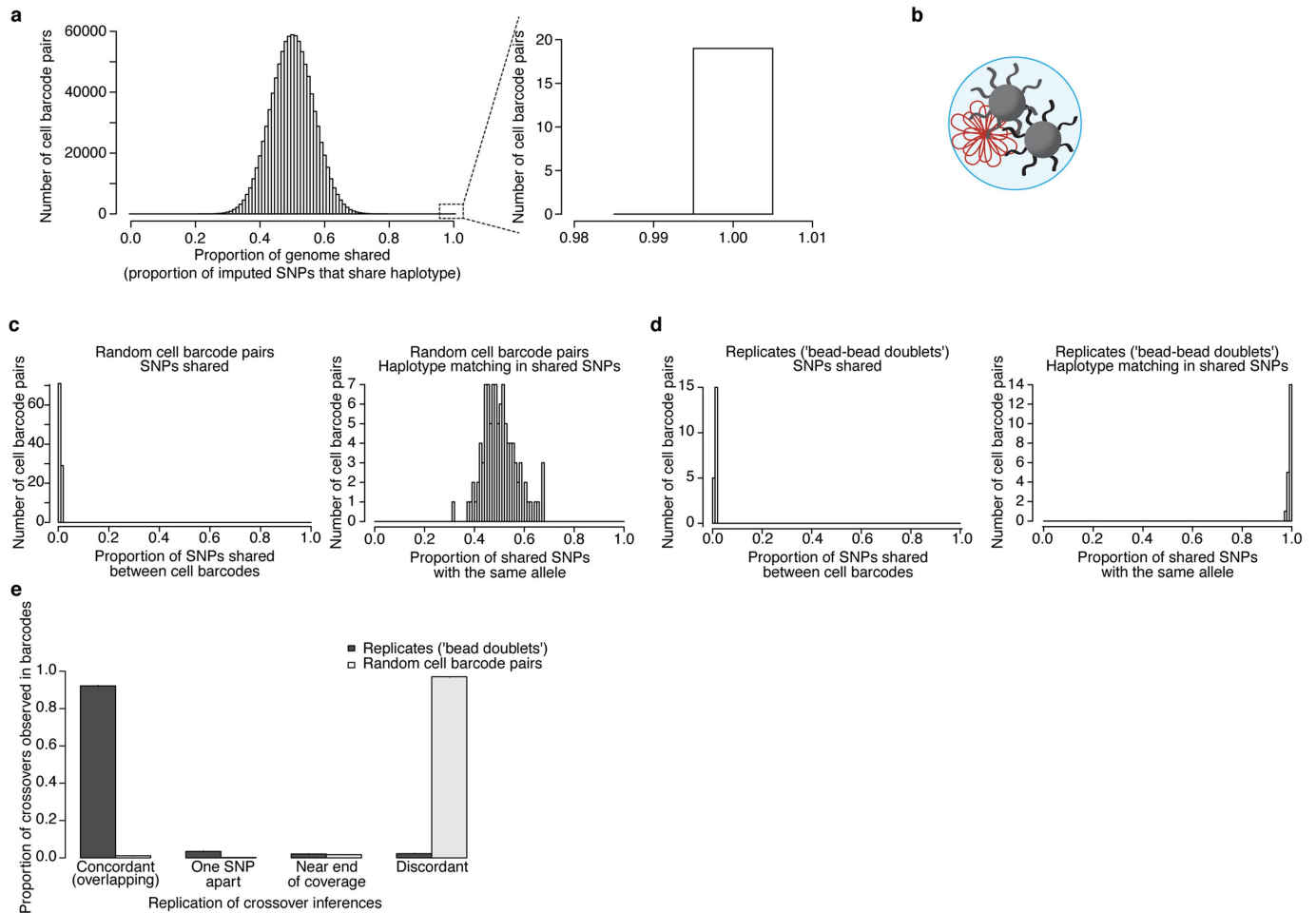


Extended Data Fig. 2 | Evaluation of chromosomal phasing and

identification of cell doublets. **a**, Phasing strategy. Green and purple denote the chromosomal phase of each allele (unknown before analysis). Each sperm cell carries one parental haplotype (green or purple), except where a recombination event separates consecutively observed SNPs (red X in bottom sperm). Because alleles from the same haplotype will tend to be observed in the same sperm cells, the haplotype arrangement of the alleles can be assembled at whole-chromosome scale (resulting in the phased donor genome).

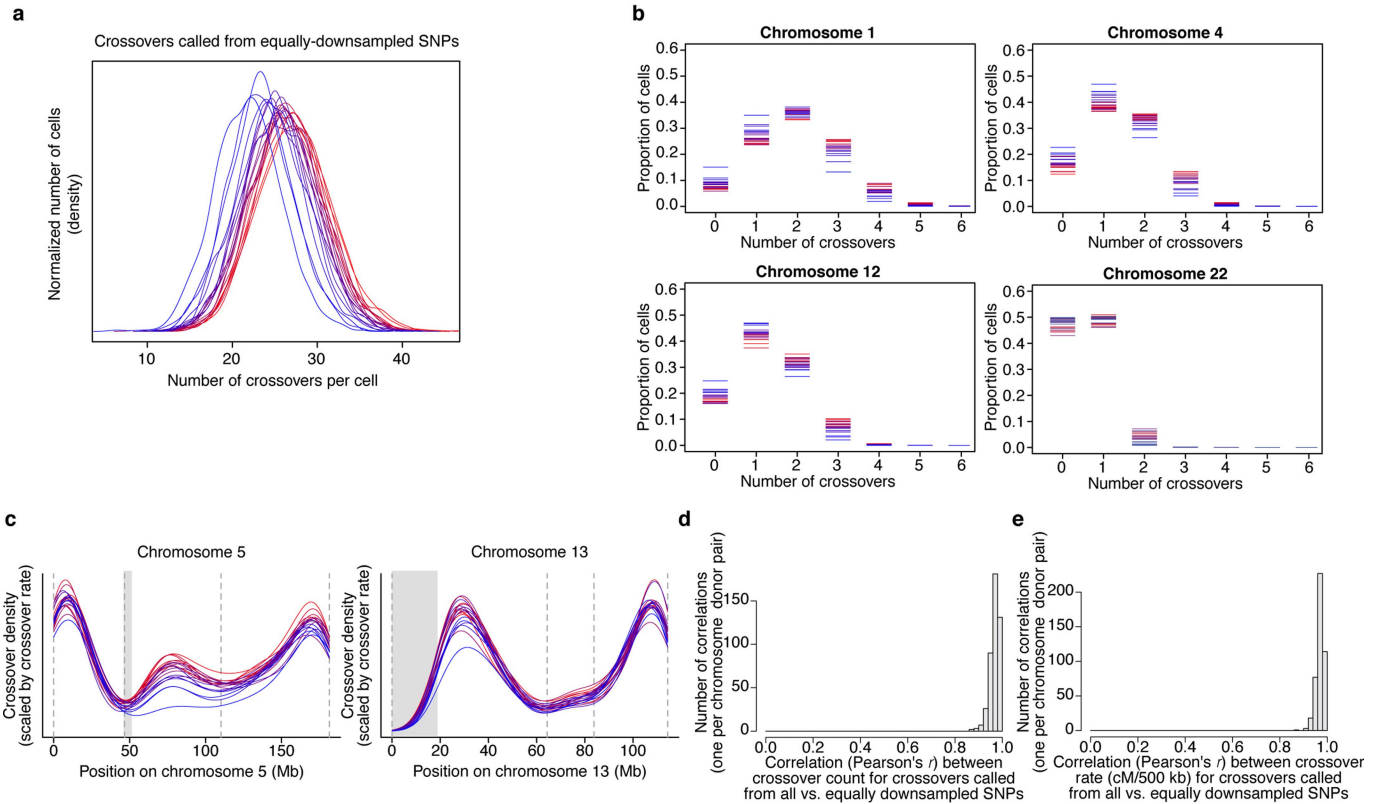
b, Evaluation of our phasing method using 1,000 simulated single-sperm genomes (generated from two a priori known parental haplotypes and sampled at various levels of coverage). Because cell doublets (which combine two haploid genomes and potentially two haplotypes at any region) can in principle undermine phasing inference, we included these doublets in the simulation (in proportions shown on the x-axis, which bracket the observed doublet rates). Each point shows the proportion of SNPs phased concordantly with the correct (a priori known) haplotypes (y-axis) for one simulation (five simulations were performed for each unique combination of proportion of cell doublets and percentage of sites observed). **c**, Relationship of phasing capability to number of cells analysed. Data are as in **b**, but for different numbers of simulated cells. All simulations had an among-cell mean of 1% of heterozygous sites observed. **d**, A cell doublet: when two cells (here, sperm DNA florets) are encapsulated together in the same droplet, their genomic

sequences will be tagged with the same barcode; such events must be recognized computationally and excluded from downstream analyses. **e**, Four example chromosomes from a cell barcode associated with two sperm cells (a cell doublet). Black lines show haplotypes; blue circles are observations of alleles, shown on the haplotype from which they derive. Both parental haplotypes are present across regions of chromosomes for which the cells inherited different haplotypes. **f**, Computational recognition of cell doublets in Sperm-seq data (from an individual sperm donor, NC11). We used the proportion of consecutively observed SNP alleles derived from different parental haplotypes to identify cell doublets; this proportion is generally small (arising from sparse crossovers, PCR/sequencing errors, and/or ambient DNA) but is much higher when the analysed sequence comes from a mixture of two distinct haploid genomes. We use 21 of the 22 autosomes to calculate this proportion, excluding the autosome with the highest such proportion (given the possibility that a chromosome is aneuploid). The dashed grey line marks the inflection point beyond which sperm genomes are flagged as potential doublets and excluded from downstream analysis. Red points indicate barcodes with coverage of both X and Y chromosomes (potentially X + Y cell doublets or XY aneuploid cells); black points indicate barcodes with one sex chromosome detected (X or Y). The red (XY) cells below the doublet threshold are XY aneuploid but appear to have just one copy of each autosome.



Extended Data Fig. 3 | Identification and use of 'bead doublets'. **a**, SNP alleles were inferred genome-wide (for each sperm genome) by imputation from the subset of alleles detected in each cell and by Sperm-seq-inferred parental haplotypes. For each pair of sperm genomes (cell barcodes), we estimated the proportion of all SNPs at which they shared the same imputed allele. A small but surprising number of such pairwise comparisons (19 of 984,906 from the donor shown, NC14) indicates essentially identical genomes (ascertained through different SNPs). **b**, We hypothesize that this arises from a heretofore undescribed scenario that we call 'bead doublets', in which two barcoded beads have coencapsulated with the same gamete and whose barcodes therefore tagged the same haploid genome. **c**, Random pairs of cell barcodes (here 100 pairs selected from donor NC10) tend to investigate few of the same SNPs (left), and also tend to detect the same parental haplotype on average at the expected 50% of the genome (right). **d**, 'Bead doublet' barcode pairs (here 20 pairs from donor NC10, who had the median number of bead

doublets, left) also investigate few of the same SNPs, yet detect identical haplotypes throughout the genome (right). Results were consistent across donors. **e**, Use of 'bead doublets' to characterize the concordance of crossover inferences between distinct samplings of the same haploid genome by different barcodes. The bead doublets (barcode pairs) were compared to 100 random barcode pairs per donor. Crossover inferences were classified as 'concordant' (overlapping, detected in both barcodes), as 'one SNP apart' (separated by just one SNP, detected in both barcodes), as 'near end of coverage' (within 15 heterozygous SNPs of the end of SNP coverage at a telomere, where the power to infer crossovers is partial), or as discordant. Error bars (with small magnitude) show binomial 95% confidence intervals for the number of crossovers per category divided by number of crossovers total in both barcodes (32,714 crossovers total in 1,201 bead doublet pairs; 67,862 crossovers total in 2,000 random barcode pairs; some barcodes are in multiple bead doublets or random barcode pairs).

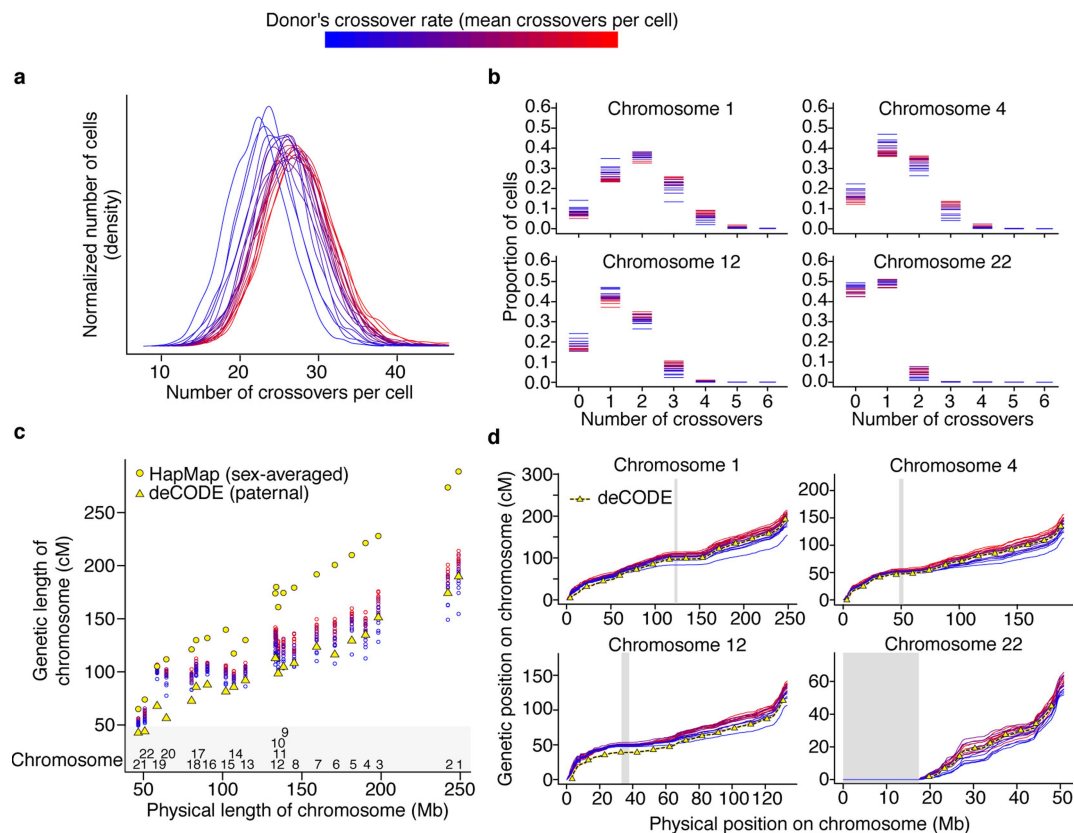


Extended Data Fig. 4 | Numbers and locations of crossovers called from downsampled data (equal number of SNPs in each cell, randomly chosen).

To eliminate any potential effect of unequal sequence coverage across donors and cells, we used downsampling to create datasets with equal coverage (numbers) of heterozygous SNP observations in each cell. Crossovers were called from these random equally sized sets of SNPs from all cells.

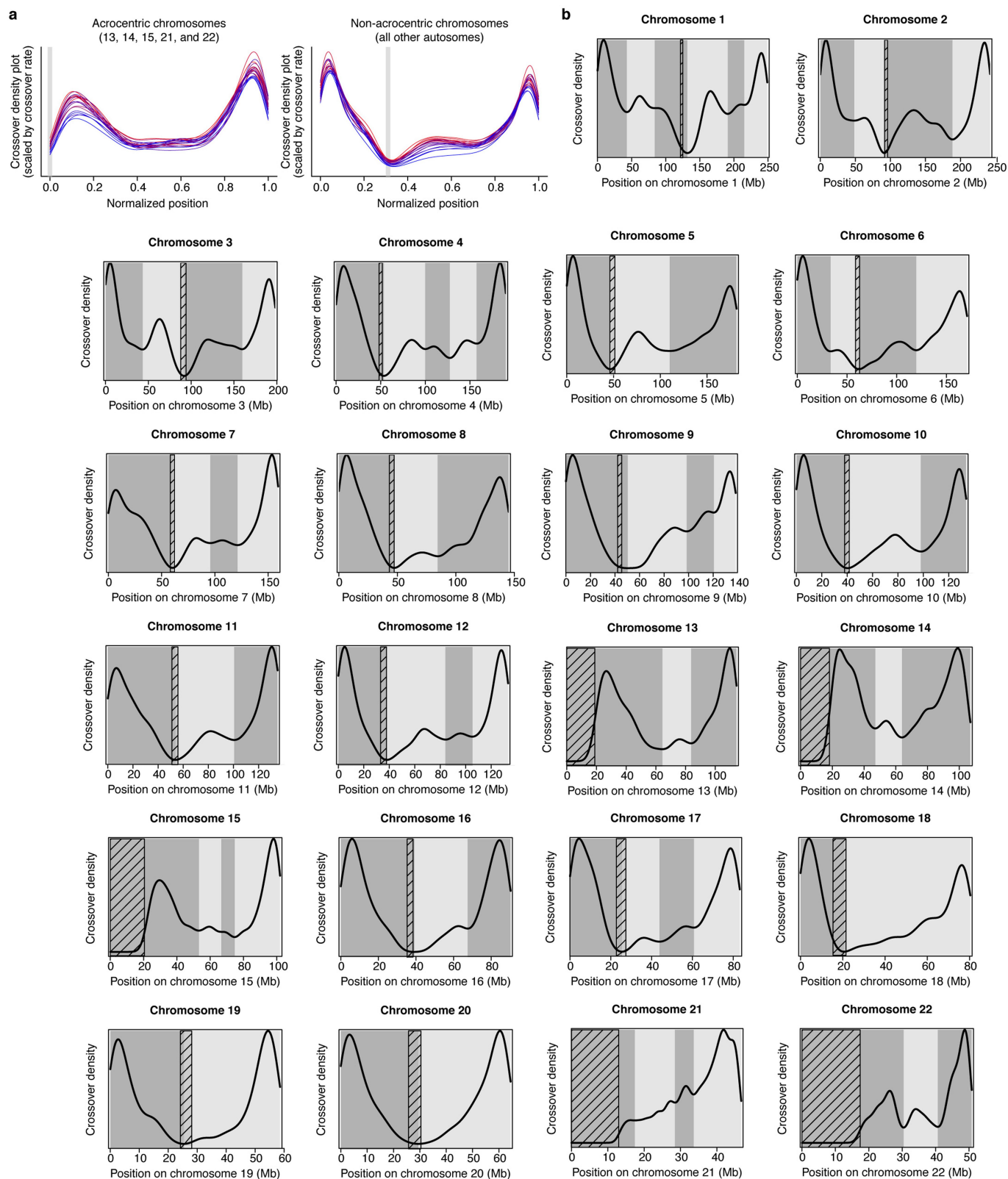
a, b, Crossover number per cell globally (**a**) and per chromosome (**b**) (785,476 total autosomal crossovers called from downsampled SNPs included, 30,778 cells included, aneuploid chromosomes excluded). **c**, Density plots of crossover location with crossover midpoints plotted and area scaled to be equal to the per-chromosome crossover rate. Grey rectangles mark centromeric regions; coordinates are in hg38. **d**, Similar numbers of crossovers were called from full data and equally downsampled SNP data: we performed correlation tests across cells for each donor and chromosome to compare the number of crossovers called from all data to the number of crossovers called

from equal numbers of randomly downsampled SNPs. The histogram shows Pearson's r values for all 460 (20 donors \times 23 chromosomes (total number plus number for 22 autosomes)) tests (n per test = 974–2,274 cells per donor as in Extended Data Table 1; all chromosome comparisons Pearson's $r > 0.83$; all two-sided $P < 10^{-300}$). **e**, Crossovers called from equally downsampled SNP data were in similar locations to those called from all data: we performed correlation tests comparing crossover rates in 500-kb bins (centimorgans (cM) per 500 kb) from all data versus equally downsampled SNP data for each donor and chromosome. The histogram shows Pearson's r values for all 460 (20 donors \times 23 chromosomes (genome-wide rate plus rate for 22 autosomes)) tests (n per test = number of 500-kb bins per chromosome (genome-wide: 5,739; chromosomes 1–22: 497, 484, 396, 380, 363, 341, 318, 290, 276, 267, 270, 266, 228, 214, 203, 180, 166, 160, 117, 128, 93, 101); all chromosome comparisons Pearson's $r > 0.87$, all two-sided $P < 10^{-300}$).



Extended Data Fig. 5 | Interindividual and intercell recombination rate from single-sperm sequencing. **a**, Density plot showing the per-cell number of autosomal crossovers for all 31,228 cells (813,122 total autosomal crossovers) from 20 sperm donors (per-donor cell and crossover numbers as in Extended Data Table 1; aneuploid chromosomes were excluded from crossover analysis). Colours represent a donor's mean crossover rate (crossovers per cell) from low (blue) to high (red). This same mean recombination rate derived colour scheme is used for donors in all figures. The recombination rate differs among donors ($n = 20$; Kruskal–Wallis chi-squared = 3,665; $df = 19$; $P < 10^{-300}$). **b**, Per-chromosome crossover number in each of the 20 sperm donors (data as in **a** but shown for individual chromosomes). **c**, Per-chromosome genetic map lengths for: each of the 20 sperm donors, as inferred from Sperm-seq data

(colours from blue to red reflect donors' individual crossover rates as in **a**); a male average, as estimated from pedigrees by deCODE⁶ (yellow triangles); and a population average (including female meioses, which have more crossovers), as estimated from HapMap data⁷ (yellow circles). The deCODE genetic maps stop 2.5 Mb from the ends of SNP coverage. **d**, Physical versus genetic distances (for individualized sperm donor genetic maps and deCODE's paternal genetic map) plotted at 500-kb intervals (in hg38 coordinates). Grey boxes denote centromeric regions (or centromeres and acrocentric arms). Sperm-seq maps are broadly concordant with deCODE maps (see the correlation test results in Supplementary Notes), except at subtelomeric regions that are not included in deCODE's map.

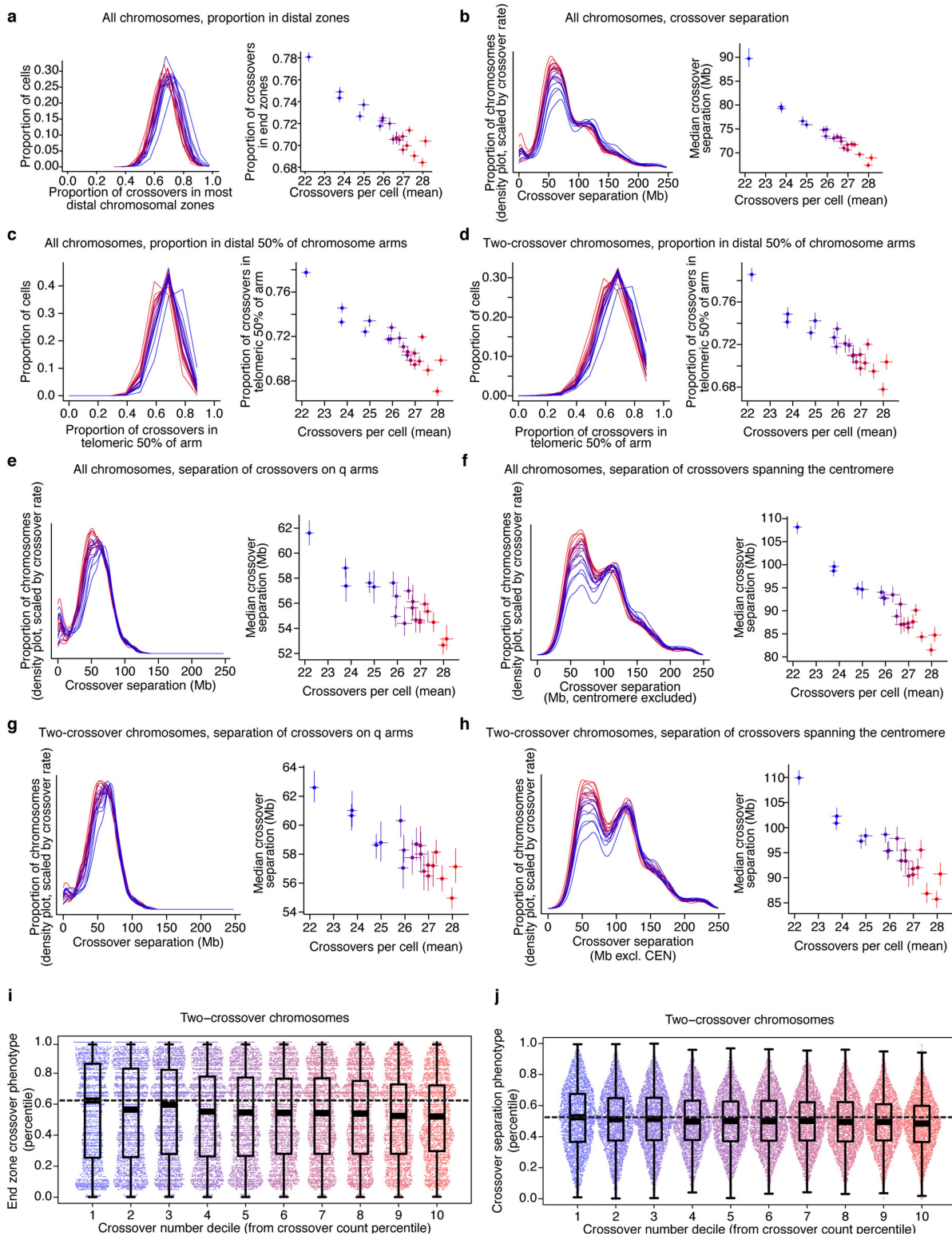


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Distributions of crossover locations along

chromosomes (in 'crossover zones'). a. Each donor's crossover locations are plotted as a coloured line; the colour indicates the donor's overall crossover rate (blue, low; red, high); grey boxes show the locations of centromeres (or, for acrocentric chromosomes, of centromeres and *p* arms). We used the midpoint between the SNPs bounding each inferred crossover as the position for each crossover in all analyses. To combine data across chromosomes, we show crossover locations (density plot) on 'meta-chromosomes' in which crossover locations are normalized to the length of the chromosome or arm on which they occurred. For acrocentric chromosomes, only the *q* arm was considered; for nonacrocentric chromosomes, the *p* and *q* arms were afforded space on the basis of the proportion of the nonacrocentric genome (in base pairs) they

comprise, with the centromere placed at the summed *p* arms' proportion of base pairs of these chromosomes. Crossover locations were first converted to the proportion of the arm at which they fall, and then these positions were normalized to the genome-wide *p* or *q* arm proportion. **b.** Identification of chromosomal zones of recombination use ('crossover zones') from all donors' crossovers for 22 autosomes. Density plots are shown of crossover location for all sperm donors' total 813,122 crossovers (aneuploid chromosomes excluded; the crossover location is the midpoint between SNPs bounding crossovers) along autosomes (hg38). Crossover zones (bounded by local minima of crossover density) are shown with alternating shades of grey. Diagonally hatched rectangles indicate centromeres (or centromeres and acrocentric arms).

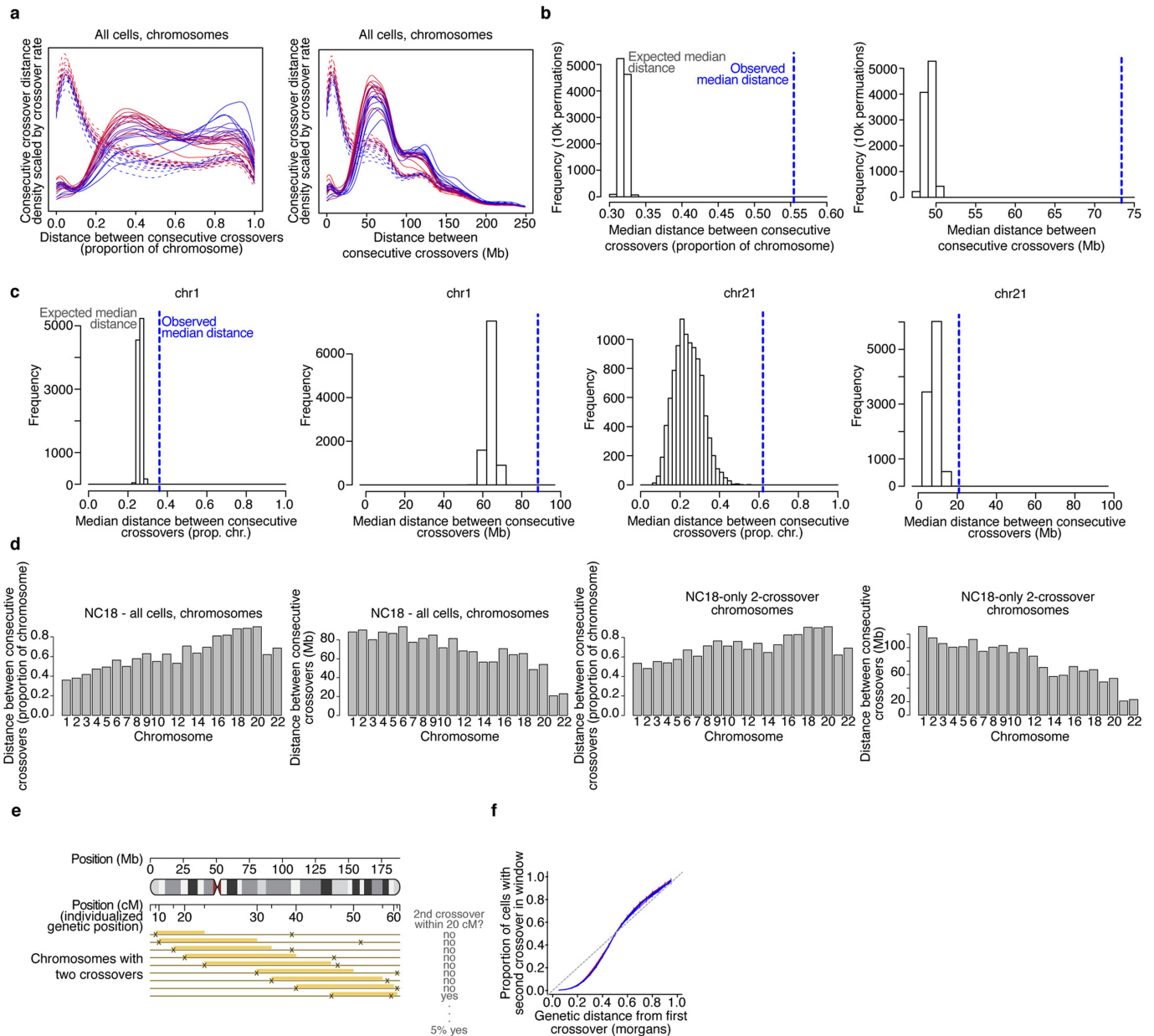


Extended Data Fig. 7 | See next page for caption.

Article

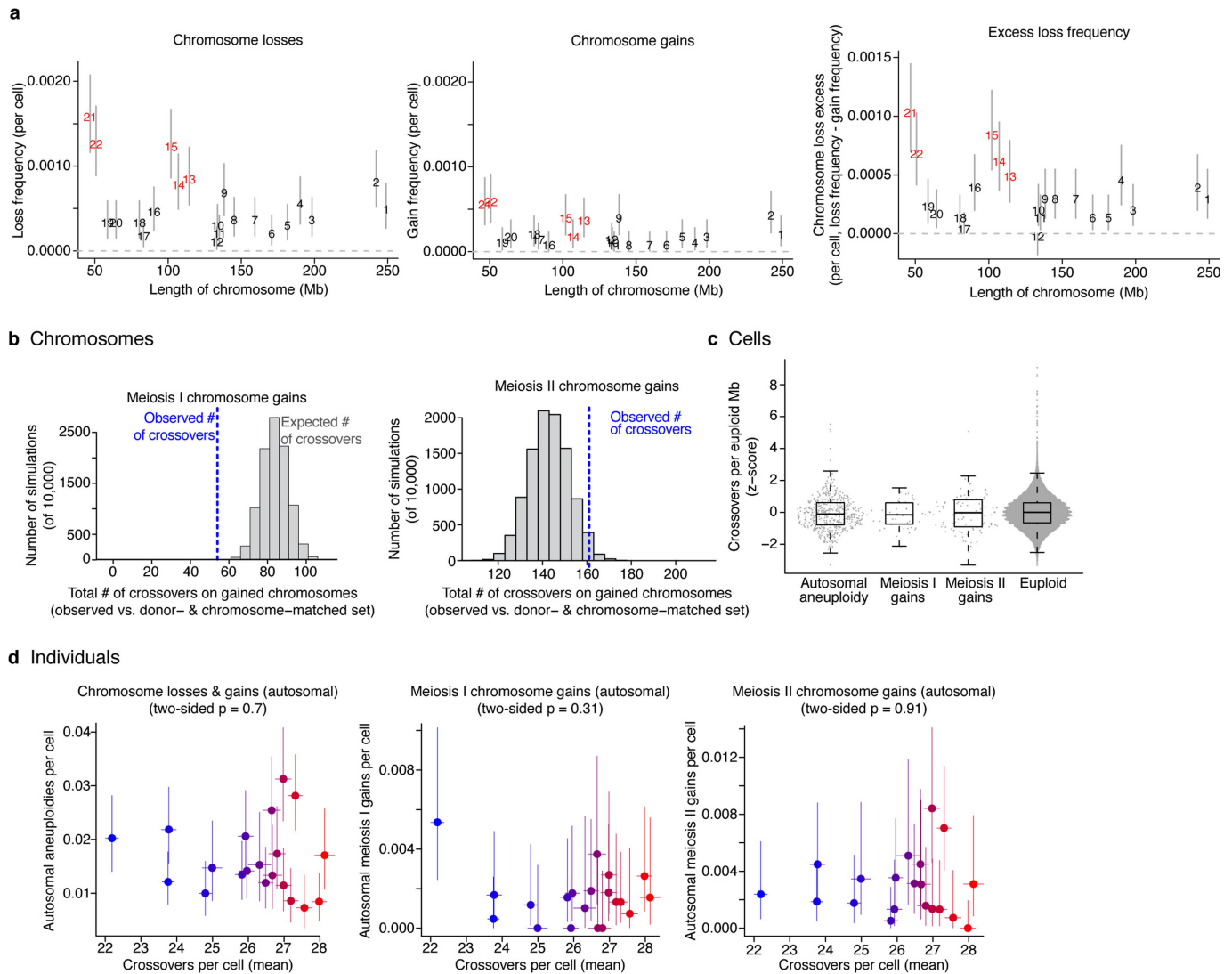
Extended Data Fig. 7 | Crossover placement in end zones, and crossover separation, varies in ways that correlate with crossover rate, among sperm donors and among individual gametes. Analyses are shown by donor (**a–h**; $n = 20$ sperm donors) or by individual gamete (**i, j**, $n = 31,228$ gametes). In **a–h**, the left panels show the phenotype distributions for individual donors, and the right panels show the relationship to the donors' crossover rates. To control for the effect of the number of crossovers, the analyses in **c, d** and **g–j** use 'two-crossover chromosomes'—chromosomes on which exactly two crossovers occurred. For scatter plots (**a–h**, right), all x-axes show the mean crossover rate and all error bars are 95% confidence intervals (y-axes are described per panel). **a, b**, Left, both the proportion of crossovers that falls in the most distal chromosome crossover zones (**a**) and crossover separation (**b**; a readout of crossover interference, the distance between consecutive crossovers in Mb) vary among 20 sperm donors (proportion of crossovers in end per-cell distributions among-donor Kruskal–Wallis chi-squared = 2,334, $df = 19$, $P < 10^{-300}$; all distances between consecutive crossovers among-donor Kruskal–Wallis chi-squared = 3,309, $df = 19$, $P < 10^{-300}$). The right panels show both properties (y-axes, total proportion of crossovers in distal zones and median crossover separation, respectively) versus the donor's crossover rate (correlation results for 20 sperm donors: proportion of all crossovers across cells in distal zones Pearson's $r = -0.95$, two-sided $P = 2 \times 10^{-10}$; Pearson's $r = -0.96$, two-sided $P = 1 \times 10^{-11}$). **c**, Results obtained from an alternative method for calculating the proportion of crossovers in the distal regions of chromosomes. The proportion of crossovers in the distal 50% of chromosome arms varies across donors (left, among-donor Kruskal–Wallis chi-squared = 2,209, $df = 19$, $P < 10^{-300}$) and negatively correlates with recombination rate (right, Pearson's $r = -0.92$, two-sided $P = 2 \times 10^{-8}$; the y-axis shows the actual proportion of crossovers in the distal 50%). **d**, As in **c**, but with the proportion of crossovers from two-crossover chromosomes occurring in the distal 50% of chromosome arms. Left, among-donor Kruskal–Wallis chi-squared = 1,058,

$df = 19$, $P = 2 \times 10^{-212}$; right, correlation with recombination rate Pearson's $r = -0.93$, two-sided $P = 4 \times 10^{-9}$. **e**, As in **b** but for consecutive crossovers on the *q* arm of the chromosome. Left, among-donor Kruskal–Wallis chi-squared = 346, $df = 19$, $P = 7 \times 10^{-62}$; right, correlation with recombination rate Pearson's $r = -0.90$, two-sided $P = 5 \times 10^{-8}$. **f**, As in **b** but for consecutive crossovers on opposite chromosome arms (that is, crossovers that span the centromere). Left, among-donor Kruskal–Wallis chi-squared = 1,554, $df = 19$, $P = 1 \times 10^{-300}$; right, correlation with recombination rate Pearson's $r = -0.96$, two-sided $P = 3 \times 10^{-11}$. **g**, As in **e** but for distances between consecutive crossovers on two-crossover chromosomes. Left, among-donor Kruskal–Wallis chi-squared = 181, $df = 19$, $P = 2 \times 10^{-28}$; right, correlation with recombination rate Pearson's $r = -0.88$, two-sided $P = 3 \times 10^{-7}$. **h**, As in **f** but for distances between consecutive crossovers on two-crossover chromosomes. Left, among-donor Kruskal–Wallis chi-squared = 930, $df = 19$, $P = 5 \times 10^{-185}$; right, correlation with recombination rate Pearson's $r = -0.92$, two-sided $P = 1 \times 10^{-8}$. **i, j**, Boxplots show medians and interquartile ranges with whiskers extending to 1.5 times the interquartile range from the box. Each point represents a cell. **i**, Within-donor percentiles showing the proportion of crossovers from two-crossover chromosomes that fall in distal zones, plotted against the crossover-rate decile. Groups are deciles of crossover rates normalized by converting each cell's crossover count to a percentile within-donor (all cells from all donors shown together; n cells in deciles = 3,152, 3,122, 3,276, 3,067, 3,080, 3,073, 3,135, 3,132, 3,090, 3,101, respectively (31,228 in total)). Because the initial data are proportions with small denominators, an integer effect is evident as pileups at certain values. **j**, Crossover interference from two-crossover chromosomes (showing the median consecutive crossover separation per cell). Each point represents the median of all percentile-expressed distances between crossovers from all two-crossover chromosomes in one cell (percentile taken within-chromosome); groupings and n values as in **i**.



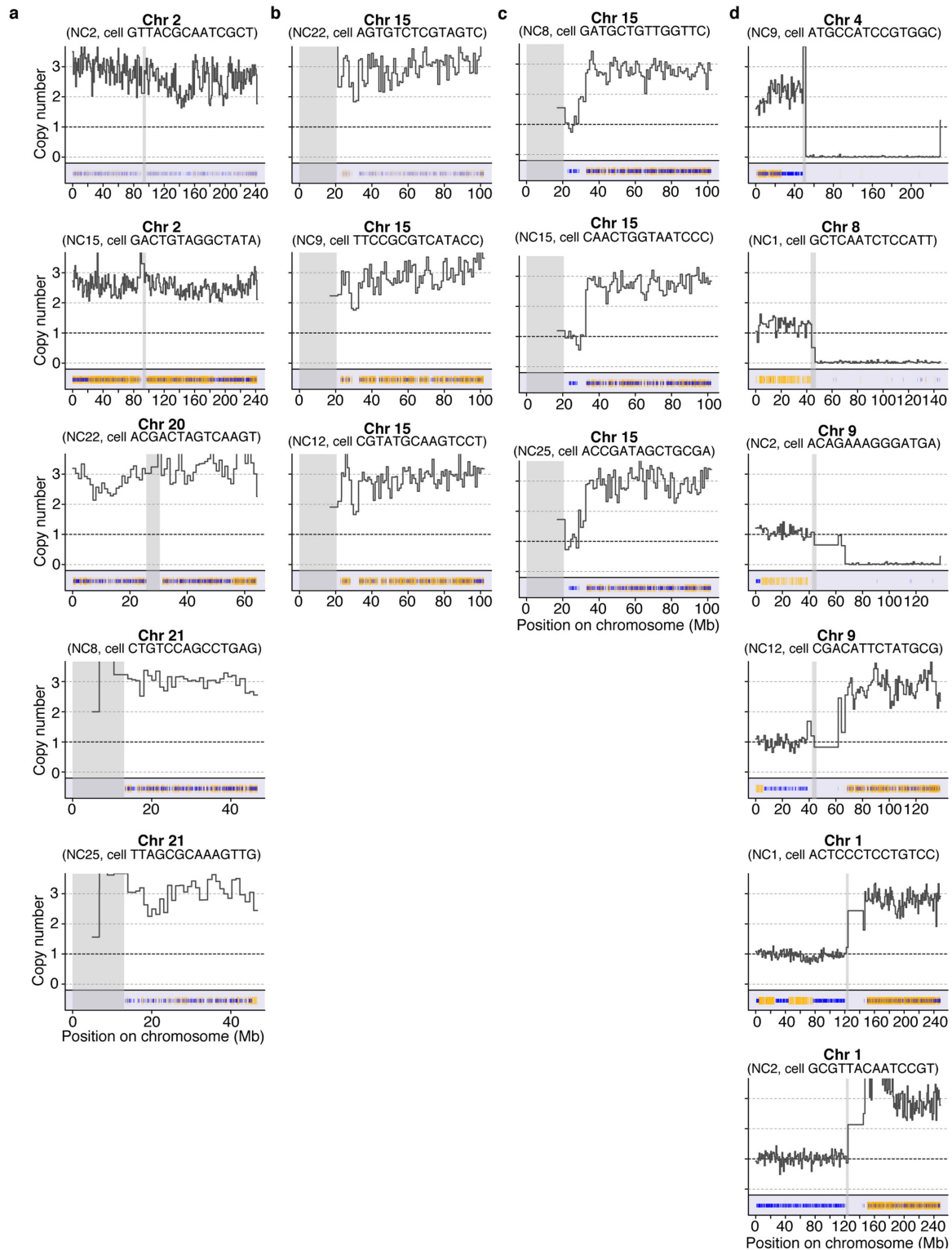
Extended Data Fig. 8 | Crossover interference in individual sperm donors and on chromosomes. **a**, Solid lines show density plots (scaled by donor's crossover rate) of the observed distance (separation) between consecutive crossovers as measured in the proportion of the chromosome separating them (left) and in genomic distance (right), with one line per donor ($n = 20$). Dashed lines show the distance between consecutive crossovers when crossover locations are permuted randomly across cells to remove the effect of crossover interference. **b**, The median of observed distances between consecutive crossovers for one donor (NC18, who had the tenth lowest recombination rate of 20 donors; blue dashed line) is shown along with a histogram of the medians of $n = 10,000$ among-cell crossover permutations (in both cases, the permutation one-sided P -value is less than 0.0001). The units are the proportion of the chromosome (left) and genomic distance (in Mb, right). **c**, Crossover separation on example chromosomes; plots and n values are as in **b**. Permutation one-sided $P < 0.0001$ for all chromosomes in all sperm donors except occasionally for chromosome 21, where especially few double

crossovers occur. **d**, Median distances between donor NC18's consecutive crossovers for each autosome for all inter-crossover distances (left two panels) and inter-crossover distances only from chromosomes with two crossovers (right two panels). Units are proportion of the chromosome or genomic distance. **e**, Diagram describing analysing crossover interference in individualized genetic distance (one 20-cM window is shown), using a donor's own recombination map. **f**, When parameterized using each donor's own genetic map, sperm donors' crossover interference profiles across multiple genetic distance windows (as shown in **e**) do not differ ($n = 20$ sperm donors; Kruskal-Wallis chi-squared = 0.22; $df = 19$; $P = 1$, using 20 estimates (cM distances) for each of 20 donors). Error bars show binomial 95% confidence intervals on the proportion of cells with a second crossover in the window given. This suggests that interindividual variation in crossover interference, although substantial when measured in base pairs, is negligible when measured in donor-specific genetic distance, pointing to a shared influence upon crossover interference and crossover rate.



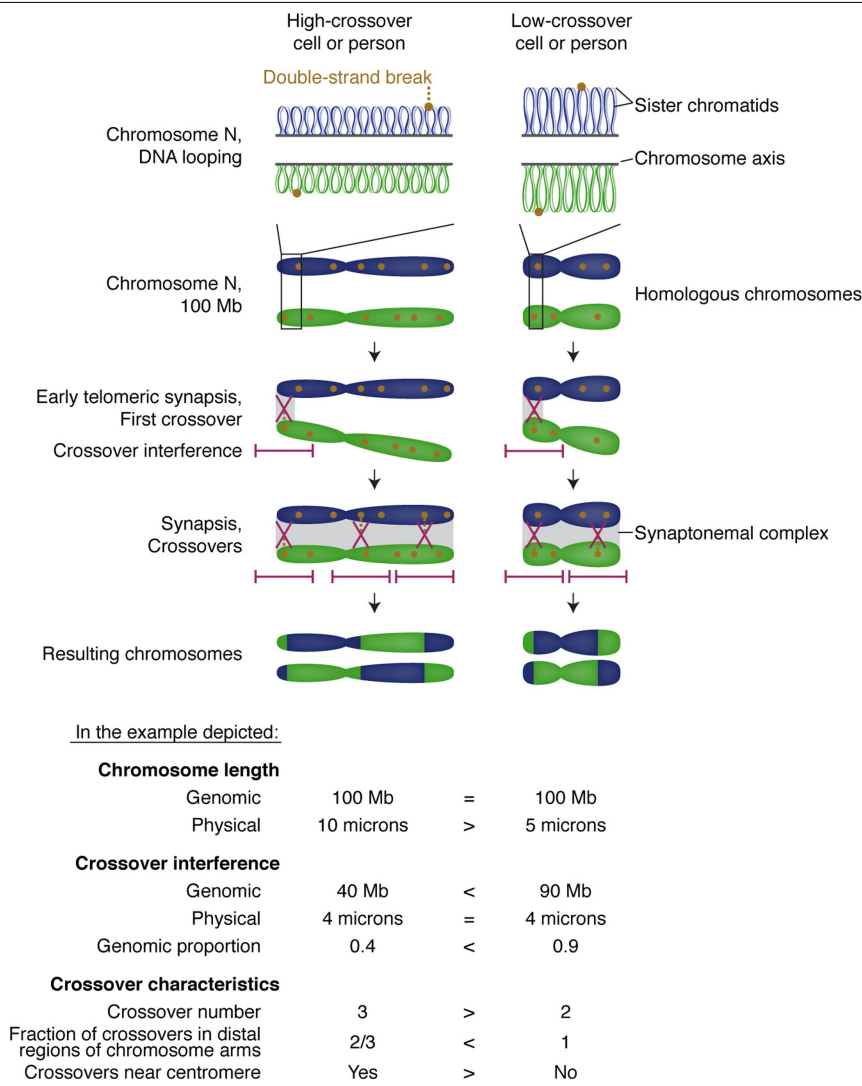
Extended Data Fig. 9 | Relationships of aneuploidy frequency to chromosome size and recombination. **a.** The across-donor per-cell frequency of chromosome losses (left) and gains (centre), plotted against the length of the chromosome (from reference genome hg38; for losses across $n = 22$ chromosomes, Pearson's $r = -0.29$, two-sided $P = 0.19$; and for gains across $n = 22$ chromosomes, Pearson's $r = -0.23$, two-sided $P = 0.30$). Right, the per-chromosome rate of losses exceeding gains (number of losses minus number of gains divided by number of cells) is plotted against the length of the chromosomes (across $n = 22$ chromosomes; Pearson's $r = -0.29$, two-sided $P = 0.19$). Red labels, acrocentric chromosomes. Error bars show 95% binomial confidence intervals on the per-cell frequency (number of events/number of cells, all 31,228 cells included). **b–d.** Relationship between aneuploidy frequency and recombination. Only autosomal whole-chromosome aneuploidies are included. **b.** Left, total number of crossovers on meiosis I nondisjoined chromosomes (blue line; chromosomes analysed, called as transitions between the presence of one haplotype and both haplotypes on the gained chromosome) compared with $n = 10,000$ donor- and chromosome-matched sets (35×2 chromosomes per set) of properly segregated chromosomes (grey histogram; permutation). Fifty-four total crossovers on meiosis I gains versus 84.2 mean total crossovers on sets of matched chromosomes; one-sided permutation $P < 0.0001$, for the hypothesis that gained chromosomes have fewer crossovers. Right, as left but for gains

occurring during meiosis II (71 meiosis-II-derived gained chromosomes of one whole copy from all individuals with fewer than five crossovers called on the gained chromosome). One-sided permutation $P = 0.98$ for meiosis II from $n = 10,000$ permutations, for the hypothesis that gained chromosomes have fewer crossovers; sister chromatids nondisjoined in meiosis II capture all crossovers whereas matched chromosomes do not: matched simulations and homologues nondisjoined in meiosis I capture only a random half of crossovers occurring on that chromosome in the parent spermatocyte. **c.** Crossovers per nonaneuploid megabase from each cell from each donor, split by aneuploidy status (n cells = 498, 50, 92, 30, 609, left to right; 'euploid' excludes cells with any autosomal whole- or partial-chromosomal loss or gain; 'gains' includes gains of one or more than one chromosome copy; Mann-Whitney test $W = 7,264,117, 722,191, 1,370,376$; two-sided $P = 0.07, 0.49, 0.66$ for all autosomal aneuploidies, meiosis I gains and meiosis II gains, respectively, all compared against euploid). Each cell is represented by one point; boxplots show medians and interquartile ranges with whiskers extending to 1.5 times the interquartile range from the box. **d.** Per-cell crossover rates versus per-cell rates of aneuploidy (left, loss and gain; middle and right, gain only, as only chromosome gain meiotic division can be determined); $n = 20$ donors (coloured by crossover rate). P -values shown are for two-sided Pearson's correlation tests. Error bars represent 95% confidence intervals on the mean crossover rate (x-axis) and on the observed aneuploidy frequency (y-axis).



Extended Data Fig. 10 | Additional examples of noncanonical aneuploidy events detected with Sperm-seq. This figure includes those shown in Fig. 3f. Copy number, SNPs, haplotypes and centromeres are plotted as in Fig. 3a. Donor and cell identities are noted above each panel. Coordinates are in the reference genome hg38. **a**, **b**, Chromosomes 2, 20, 21 (**a**) and 15 (**b**) are sometimes present in three copies in an otherwise haploid sperm cell.

c, A distinct, recurring triplication of much of chromosome 15, from around 33 Mb onwards but not including the proximal part of the *q* arm, also recurs in cells from three donors. **d**, Chromosome-arm-level losses (top three panels) and gains (including in more than one copy, bottom three panels, and a compound gain of the *p* arm and loss of the *q* arm, top panel).



Extended Data Fig. 11 | Single-cell and person-to-person variation in diverse meiotic phenotypes may be governed by variation in the physical compaction of chromosomes during meiosis. Previous work showed that the physical length of the same chromosome varies among spermatocytes at the pachytene stage of meiosis, probably by differential looping of DNA along the meiotic chromosome axis (for example, the left column shows smaller loops, resulting in more loops in total and in a greater total axis length compared with the right column, with larger loops)^{15,72–75}. This physical chromosome length is correlated across chromosomes among cells from the same individual^{21,76}, and correlates with crossover number^{15,20,21,42,73,76}. This length—measured as the length of the chromosome axis or of the synaptonemal complex (the connector of homologous chromosomes)—can vary by two or more fold among a human’s spermatocytes²¹. We propose that the same process differs on average across individuals and may substantially explain interindividual variation in recombination rate. On average, individual 1 (left) would have meiotic chromosomes that are physically longer (less compacted) in an average cell than individual 2 (right); one example chromosome is shown in the figure. After the first crossover on a chromosome (probably in a distal region of a chromosome, where synapsis typically begins in male human meiosis before spreading across the whole chromosome^{13–15}), crossover interference prevents

nearby double-strand breaks (DSBs) from becoming crossovers; however, DSBs that are far away can become crossovers (which themselves also cause interference). More DSBs are probably created on physically longer chromosomes, and crossover interference occurs among noncrossover as well as crossover DSBs⁷⁷. Crossover interference occurs over relatively fixed physical (micrometre) distances^{43–45,76}; these distances encompass different genomic (Mb) lengths of DNA in different cells or on average in different people owing to variable compaction. Thus, crossover interference tends to lead to a different total number of crossovers as a function of the degree of compaction, resulting in the observed negative correlation (Fig. 2c, e) of crossover rate with crossover spacing (as measured in base pairs). Given that the first crossover probably occurs in a distal region of the chromosome, this model can also explain the negative correlation (Fig. 2b, d) between crossover rate and the proportion of crossovers at chromosome ends. This figure shows the total number of crossovers, crossover interference extent, and crossover locations for both sister chromatids of each homologue combined; in reality, these crossovers are distributed among the sister chromatids, making these relationships harder to detect in daughter sperm cells and requiring large numbers of observations to make relationships among these phenotypes clear.

Extended Data Table 1 | Sperm donor and single-sperm sequencing characteristics and results

Donor	Ancestry*	Cells (number excluding cell and bead doublets)	Reads per cell (median, thousands)	Genome covered per cell (median, percent)	Heterozygous SNPs in genome (millions)	Unique heterozygous SNP alleles observed per cell (median, thousands)	Crossovers observed (total, thousands)	Crossovers per cell (mean)	Resolution of crossovers (kb, median)	Autosomal aneuploidy events (percent of cells) [†]	Sex chromosome aneuploidy events (percent of cells) [†]
Overall	--	31,228[‡]	211[§]	1.0	--	24.6[§]	813[‡]	26.11	240	1.6[§]	0.9[§]
NC1	Eur.	982	284	1.4	1.95	31.6	26	26.31	189	1.5	0.6
NC2	Eur.	1,680	163	0.8	1.98	18.2	37	22.19	307	2.0	0.7
NC3	Eur.	1,289	190	0.9	1.94	21.5	36	28.13	260	1.7	0.7
NC4	Eur.	1,482	243	1.1	1.98	26.8	40	26.98	243	1.1	0.5
NC6	Afr. Am.	1,370	154	0.8	2.53	23.8	38	27.57	253	0.7	0.3
NC8	As.	1,663	304	1.5	1.81	30.9	45	26.98	229	3.1	0.5
NC9	As.	1,894	245	1.2	1.79	25.6	53	27.98	231	0.8	1.5
NC10	As.	1,154	224	1.1	1.82	23.3	29	24.99	257	1.5	0.3
NC11	Eur.	1,930	202	1.0	1.92	22.8	50	25.82	242	1.3	0.4
NC12	Eur.	2,145	179	0.9	1.91	20.6	51	23.76	270	1.2	1.7
NC13	Eur.	1,514	259	1.2	1.92	28.3	41	27.19	202	0.9	1.0
NC14	Eur.	1,336	296	1.4	1.92	32.4	36	26.65	175	2.5	1.2
NC15	Eur.	1,702	211	1.0	1.93	23.2	42	24.80	268	1.0	0.9
NC16	Eur.	1,785	241	1.2	1.92	26.9	42	23.78	227	2.2	1.3
NC17	Eur.	1,504	220	1.0	1.94	23.8	39	25.92	250	2.1	0.7
NC18	Eur.	1,589	170	0.8	1.93	18.4	42	26.48	317	1.2	0.6
NC22	Afr. Am.	1,693	195	0.9	2.53	29.7	44	25.96	205	1.4	0.7
NC25	Afr. Am.	2,274	175	0.8	2.47	25.8	62	27.31	211	2.8	1.8
NC26	Afr. Am., As.	974	120	0.6	2.55	18.0	26	26.67	355	1.3	0.4
NC27	As. (?)	1,268	267	1.3	1.96	29.2	34	26.80	199	1.7	0.6

*Ancestry as provided by the sperm bank. Afr. Am., of African American ancestry; Eur., of European ancestry; As., of Asian ancestry; (?), conflicting ancestry information given.

[†]These numbers are the total number of aneuploidy events divided by the total number of cells multiplied by 100; cells can have more than one event.

[‡]Sum across all cells from all sperm donors.

[§]Median or mean across all individual cells from all sperm donors (31,228 measurements summarized).

^{||}Median or mean of aggregate metrics across samples (20 measurements summarized).

[¶]Median across all crossovers (813,122 measurements summarized).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| n/a | Confirmed |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	All data was collected via an Illumina NovaSeq, which generated sequencing BCL files. All subsequent data processing is outlined below.
Data analysis	Picard Tools v2.2 was used for sequence data processing (http://broadinstitute.github.io/picard/). BWA-MEM v0.7.7-r441 was used for alignment (http://bio-bwa.sourceforge.net/). GATK v3.7 was used to call variants (https://gatk.broadinstitute.org/). Custom code in Drop-seq Tools v2.2 was used to format sequencing data to include single-cell barcode information and to call heterozygous data in sperm genomes (https://github.com/broadinstitute/Drop-seq/releases). Genome STRIP v2.0 was used to ascertain read depth for sperm cells (http://software.broadinstitute.org/software/genomestrip/). HapCUT v1 was used for phasing (https://github.com/vibansal/hapcut). Custom code was written in R (www.r-project.org) for this study for data processing and to call and analyze recombination and aneuploidy events; it is available via Zenodo at http://dx.doi.org/10.5281/zenodo.2581595

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Crossover and aneuploidy data (individual events and counts per donor and/or cell), including source data underlying Figs. 2, 3b-e and Extended Data Figs. 5-9, are available via Zenodo, <http://dx.doi.org/10.5281/zenodo.2581570>. Raw sequence data are available in the SRA via dbGaP for general research use upon application and approval (study accession number phs001887.v1.p1).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We aimed to sequence ~1,000 sperm cells per individual donor to be able to detect rare events (aneuploidy occurs at a frequency of less than a few percent), and the actual number of cells sequenced per donor was a result of experimental fluctuations during library preparation. With this amount of sequencing per donor as a baseline, we then chose to sequence cells from 20 individuals to be able to observe inter-individual differences and to be able to detect trends with moderate power (for example, an analysis of 20 individuals is 80% powered to detect a Pearson's r of 0.58 at $p = 0.05$).
Data exclusions	No data were excluded from the analyses. For certain analyses, subsets of the data were used to examine questions about those subsets.
Replication	As this was a completely new data set, we did not seek to replicate our findings with a data set of the same type. We did compare our findings of crossover location and frequency to published estimates generated with other methods; our observations were consistent with earlier estimates (detailed in manuscript).
Randomization	N/A - we did not have experimental groups.
Blinding	N/A - we did not have experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

The 20 sperm sample biospecimens analyzed in this study came from from 20 anonymous, karyotypically normal sperm donors at the New England Cryogenic Center (NECC). Per sperm bank policy, donors are 18-38 years old at the time of donation; precise age of donors is not released to researchers. The donors are known only to NECC and not to the researchers. Donor identifiers used in the paper were created specifically for this study and are not linked to any external identifiers.

Recruitment

The biospecimens used were discarded samples from New England Cryogenic Center; the samples had been collected during sperm donation. Donors consented, at the time of donation, for biospecimens to also be used for research purposes. NECC routinely provides such samples to researchers for research that has been IRB-approved at the researchers' home institution.

The sperm donors might be subject to self-selection biases such as need for extra income and free time availability for donation, though any specific biases are unknown. If recombination or aneuploidy were correlated with any sperm donor recruitment or self-selection biases, the results of this study would reflect these underlying correlates. For example, the results are only reflective of donors aged 18-38 who had enough free time to act as sperm donors; if individuals of older ages or less free time had different meiotic phenotypes, they would not be captured in this study.

Ethics oversight

The Harvard Faculty of Medicine Office of Human Research Administration reviewed the research protocols (protocols M23743-101 and IRB16-0834) and determined that this research was "Not human subject research", based on the use of discarded biospecimens and the fact that researchers did not have interactions with the biospecimen donors. Harvard has also reviewed and approved our deposition of the data into an NIH respository; dbGaP will assure that access is provided only to researchers with legitimate research uses who agree not to try to re-identify donors based on data in the study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.