Island-specific evolution of a sex-primed autosome in a sexual planarian

Sex chromosomes evolve from homologous autosomes that acquire sex-determining genes and lose their ability to recombine. As such, sex chromosome evolution and recombination suppression are closely associated. However, because direct evidence of such homologous autosomes primed for evolution into sex chromosomes is difficult to capture, little is known about the molecular signatures associated with the evolution of recombination suppression.

The freshwater planarian *Schmidtea mediterranea*, an important model organism for studies of regeneration, exists as asexual and sexual reproductive strains. The sexual strain is distributed mostly in Tunisia and on the islands of Sardinia, Corsica and Sicily. The sexual strain of the planarian, indigenous to Tunisia and several Mediterranean islands, is a hermaphrodite. Here we isolate individual chromosomes and use sequencing, Hi-C and linkage mapping to assemble a chromosome-scale genome reference. The linkage map reveals an extremely low rate of recombination on chromosome 1. We confirm suppression of recombination on chromosome 1 by genotyping individual sperm cells and oocytes. We show that previously identified genomic regions that maintain heterozygosity even after prolonged inbreeding make up essentially all of chromosome 1. Genome sequencing of individuals isolated in the wild indicates that this phenomenon has evolved specifically in populations from Sardinia and Corsica. We find that most known master regulators of the reproductive system are located on chromosome 1. We used RNA interference to knock down a gene with haplotype-biased expression, which led to the formation of a more pronounced female mating organ. On the basis of these observations, we propose that chromosome 1 is a sex-primed autosome primed for evolution into a sex chromosome.

Chromosome-scale genome assembly

To transform the 481 scaffolds into a chromosome-scale genome reference, we carried out chromosome sequencing (ChrSeq) of a laboratory strain, as well as chromatin proximity sequencing by Hi-C (Fig. 1a). To do so, we dissected individual chromosomes from mitotic cells using laser capture and amplified and sequenced each chromosome individually (Fig. 1b). We examined the sequencing depth of each scaffold across multiple samples of the same and different chromosomes to ensure reproducibility and specificity, respectively (Extended Data Fig. 1). Overall, we successfully amplified and confidently assigned 740 Mb of the 774-Mb genome to one of four chromosomes (Supplementary Table 1).

We used ChrSeq information and chromatin interaction data generated by Hi-C to correct and connect the individual scaffolds within a chromosome into a chromosome-scale genome, hereafter referred to as Smed_chr_ref_v1. The Hi-C data analysed using the SALSA scaffolding algorithm resolved the 481 scaffolds into 57 super-scaffolds and 104 singletons. TheChrSeq data indicated that 3 of the 57 super-scaffolds were disjointed inter-chromosomal fragments, consistent with the Hi-C contact heat map (Fig. 1c). We split the disjointed super-scaffolds, merged the scaffolds into chromosomes, and ordered and oriented these alleles and investigate the reasons underpinning the persistence of heterozygosity in *S. mediterranea*, a detailed assembly of all four chromosomes is needed.
all scaffolds within the chromosomes using Juicebox visualization software\(^4\) (Methods). Chromosome assignments by ChrSeq alone (Supplementary Table 1) and by Hi-C alone were inconsistent for only 3 of 384 (0.8\%) scaffolds (Supplementary Table 2). We manually assigned these three scaffolds to chromosomes according to the Hi-C data. Hi-C also detected 26 inter-chromosomal or intra-chromosomal assembly errors (Supplementary Table 2) in the previous assembly\(^26\), 5 of which were confirmed by ChrSeq to be inter-chromosomal disjointed scaffolds (Supplementary Table 2). The final genome assembly, Smed_chr_ref_v1 (Fig. 1c), had four chromosomes with a total size of 764 Mb, which is 98.4\% of that reported previously\(^26\). Of the 1.6\% of the previous assembly not contained in Smed_chr_ref_v1 by Hi-C, about half (52.6\%) of the scaffolds could not be assigned to a chromosome or were assigned to two chromosomes by ChrSeq, indicating lower assembly quality (Supplementary Table 3). Moreover, 33.0\% and 12.4\% of the scaffolds had chromatin interaction signals with structurally complex regions of chromosomes 1 and 2, respectively (Extended Data Fig. 2), suggesting that they might be alternative assemblies or repetitive sequences. In our new assembly, Smed_chr_ref_v1, the two ends of chromosome 4 were capped by >1,000 copies of the telomere repeat TTAGGG, indicating high assembly quality.

To validate the linearity of the chromosomes in Smed_chr_ref_v1, we generated a linkage map. We crossed two divergent laboratory strains of *S. mediterranea*, S2F10b and D5, to produce an *F\(_2\)* population, and we genotyped individual worms with RADseq\(^3\). Eighty markers that were evenly distributed and genotyped in at least 98\% of the *F\(_2\)* segregants (91 of 93) were used to establish four linkage groups (Fig. 1d and Supplementary Table 4) representing the four chromosomes. The ordering of the 80 markers in the linkage map is consistent with Smed_chr_ref_v1, independently supporting the quality of our chromosome-scale genome assembly. This highly contiguous and complete genome assembly and linkage map together facilitate further genetic and epigenetic investigations of the functions of the genome in this model planarian.

### Chromosome 1 recombination suppression

We next re-examined the heterozygous alleles that define the J and V haplotypes in the newly assembled genome. We found that 87.7\% of the genetic markers that remained heterozygous were located on chromosome 1, spanning 333 Mb at a density of 30,148 variants per 10 Mb (Fig. 2a). The remaining 12.3\% of the heterozygous markers were located on the other three chromosomes at a density of 3,274 variants per 10 Mb and probably correspond to differences between highly similar copies of repetitive elements rather than true polymorphisms (Supplementary Table 5). All *F\(_2\)* worms (*n* = 93) in this study were heterozygous for chromosome 1 but not for chromosomes 2–4 (Supplementary Table 6), which is consistent with the previous study\(^24\). Hence, we concluded that the J/V haplotypes were on chromosome 1.

Our linkage map revealed an extremely low rate of recombination on chromosome 1 of only 0.5 cM for the entire chromosome (Fig. 1d). This is particularly notable because, at 333 Mb, chromosome 1 was the largest of the four chromosomes, containing more than 40\% of the genome.
To directly examine whether chromosome 1 can recombine, we sequenced 45 single sperm cells and 28 single oocytes from J/V line S2 (Fig. 2d). Gamete sequencing is preferred because recombination events in hatchlings can be selected for by differential fertilization or embryonic lethality. We identified 3,197 single-nucleotide variants (SNVs) on chromosome 1 and 3,312 SNVs on chromosome 3, covering 99% of the length of each chromosome (Supplementary Table 7). The SNVs were distributed at a similar density across 20-Mb windows, with coefficients of variation of 0.38 and 0.31 for chromosomes 1 and 3, respectively (Supplementary Table 7). We observed that 98% of the sperm cells (44 of 45) and 93% of the oocytes (26 of 28) had no crossovers on chromosome 1. By contrast, most sperm cells and oocytes had crossovers on chromosomes 2, 3 and 4 (Fig. 2b, Extended Data Fig. 3a, b and Supplementary Table 7). We thus concluded that recombination on chromosome 1 was strongly suppressed.

Consistent with this conclusion, we found that during prophase I, when other chromosomes had numerous crossovers, chromosome 1 formed a ring structure in the oocyte in J/V worms but not in J/J worms (Fig. 2c and Extended Data Fig. 4). Fluorescence in situ hybridization (FISH) with telomere probes suggested that crossovers between the two homologous pairs of chromosome 1 occurred only in regions close to the telomeres, leading to the observed ring conformation rather than the side-by-side pairing observed for chromosomes 2, 3 and 4. Furthermore, Hi-C analysis showed that chromosome 1 had three putative inversions, each >20 Mb in size (Fig. 2d); such inversions can cause crossover suppression32–34. The rest of the genome had only one large inversion of approximately 10 Mb on chromosome 2 (Extended Data Fig. 2b).

Island-specific evolution of chromosome 1
To investigate the genetic diversity of S. mediterranea in its natural environment and determine whether chromosome J/V heterozygosity occurs throughout the entire species, we used RADseq31 to sample the genomes of 70 sexual individuals from Sardinia, Corsica, Sicily and Tunisia and 2 asexual individuals from Menorca1 (Fig. 3a). To look for genetic relationships between the individuals, we determined clustering of relatedness measured by identity-by-state pairwise distances35 and identified two superclusters. Animals from Sicily were closely related to those from Tunisia, and animals from Sardinia were closely related to those from Corsica (Fig. 3b). Phylogenetic clustering (Extended Data Fig. 5), fixation index (FST) values (Extended Data Fig. 6) and analyses with Structure (Extended Data Fig. 7) further supported this observation.

The relatedness of the populations from Sardinia and Corsica suggests that they may share genome characteristics that differ from those of the populations from Sicily and Tunisia. Indeed, animals in the populations from Sardinia and Corsica had greater heterozygosity on chromosome 1 than expected under Hardy–Weinberg equilibrium (Fig. 3c), whereas the heterozygosity on the other three chromosomes
in these populations and on all four chromosomes in the other populations closely followed expectations. By analysing the J/V haplotype markers in wild populations, we found that animals from Sardinia and Corsica were heterozygous J/V, whereas those from Sicily and Tunisia were homozygous J/J. Moreover, the animals from Sardinia (n = 28) had many sites that were heterozygous in more than 80% of the individuals and were distributed along the length of chromosome 1 except near the ends; conversely, the animals from Sicily (n = 27) had very few such heterozygous sites (Fig. 3d, e and Supplementary Table 8). Few such heterozygous sites were observed on chromosomes 2, 3 and 4 in either population (Fig. 3d, e and Extended Data Fig. 3c, d). These analyses suggest that chromosome 1 specifically evolved and diverged on the islands of Sardinia and Corsica.

A sex-primed chromosome

To gain insight into the island-specific suppression of recombination on chromosome 1, we examined the genes located on this chromosome, which contained 39% of all annotated genes. Five of the seven known master regulators of the reproductive system were found on chromosome 1, including nanos5, nhr-1 (refs. 6, 7), npy-8 (ref. 8), npyr-1 (ref. 10) and CPEB-2 (ref. 11) (ophis12 and boule-2 (ref. 15) were the two exceptions) (Fig. 4a). Knockdown of any one of these master regulators leads to depletion of both male and female reproductive tissues5,6,9–12. The presence of these genes with crucial roles in sexual development on chromosome 1 suggests that chromosome 1 integrity is important for the maintenance of sexual reproduction. Indeed, the asexual lineage of S. mediterranea, which is devoid of any reproductive organs13, had a translocation from chromosome 1 to chromosome 3 and probably evolved through loss of function of one or more of these genes.

Chromosome 1 also contained three of the four known master regulators of male or female reproductive tissues (CPEB-1 (ref. 12), onecut15 and Bic-C; dmd-1 (ref. 16) was the exception) (Fig. 4a). The CPEB-1 gene is specifically required for the development of oocytes and yolk glands17. Loss of onecut and Bic-C expression leads to a ‘no-testes’ phenotype without affecting the ovary18,19. The presence of a female-determining gene on a chromosome that does not recombine provides an ideal foundation for the evolution of a sex chromosome. A loss-of-function mutation in the gene on one of the two homologous chromosomes would turn the chromosome with the mutation into a male-determining chromosome. Similar logic applies to a male-determining gene.

Haplotype-specific expression of sex-specific genes is a predicted signature of a sex-primed autosome. To test this prediction, we examined the expression of 20 genes in sexually mature adult worms and in the male copulatory organ, the penis papilla, in a J/V line. The 20 genes were chosen because they were characterized as having sex-specific function and/or expression in the literature and include the 8 master regulators mentioned above6,12 (Supplementary Table 9). Of these 20 genes, 13 contained at least one heterozygous variant in the coding sequence, which allowed us to determine whether these genes showed biased expression from the J or V haplotype (Extended Data Fig. 8). We found that a gelsolin-like protein20, Smegl, was expressed predominantly...
We used FISH with hybridization chain reaction to confirm that the expression of Smed-glp is highly enriched in both the penis papilla and the bursa canal (Extended Data Fig. 9), which is a female organ used to receive sperm from mating partners. To examine the function of Smed-glp, we used RNA interference (RNAi) to knock down its expression in young hatchlings. After 8 weeks of feeding, both the knockdown and control hatchlings reached sexual maturity. Notably, the bursa canal was much more pronounced in the knockdown animals (Fig. 4c, d). No obvious morphological differences were observed in the male copulatory organ. These results suggest that the J allele of Smed-glp may have a greater role in preventing over-development of the female bursa canal (Fig. 4e), probably by controlling the number of organ-specific muscle fibres. This is consistent with a previous observation that J/J individuals show higher egg production, suggesting a larger contribution to female reproduction (Fig. 4b in ref. 24; Supplementary Information).

We have thus identified a chromosome in the hermaphroditic planarian S. mediterranea that does not recombine, is enriched in master regulators of reproductive systems, shows allele-specific expression of sex-related genes, and contains genes orthologous to those on the sex chromosome of Schistosoma mansoni (Supplementary Tables 9 and 10, and Supplementary Information), which is the only known Platyhelminthes species with differentiated sex chromosomes. These observations collectively led us to propose that chromosome 1 of S. mediterranea is primed for evolution into a sex chromosome.

Chromosome 1 of S. mediterranea marks an intriguing system for studying sex chromosome evolution. Planarians from the islands of Sardinia and Corsica are heterozygous for the J and V haplotypes, whereas those in Sicily and Tunisia exist as J/J homozygotes. In laboratory crosses involving J/V lines, all hatchlings were J/V heterozygotes. No V/V planarians have been identified either in nature or in laboratory crosses. We genotyped single zygotes from crosses between J/V individuals and determined that homozygous zygotes (J/J or V/V) exist (Extended Data Fig. 10a and Supplementary Information). We propose that early embryonic lethality leads to the loss of J/J and V/V adults in both nature and laboratory crosses. Such lethality may arise from degeneration of coding sequences on the J and V haplotypes as a consequence of crossover suppression. Indeed, chromosome 1 has an elevated rate of transposable elements and mutations introducing stop codons relative to the rest of the genome (Supplementary Table II) as well as higher nonsynonymous substitution rates (Supplementary Table 13 and Extended Data Fig. 10b). As a consequence, the J and V haplotypes each carry unique sets of functional genes that are silenced or truncated on the other haplotype (Supplementary Table II).

The presence of three nested inversions on chromosome 1 suggests the possibility that recombination suppression on this chromosome may have evolved in a stepwise manner similar to that reported for the human X and Y chromosomes. In support of this possibility, we observed evidence for three evolutionary strata corresponding to the inverted regions. We used PacBio genome sequencing data to identify long reads that bridged the inversion breakpoints, and we validated the three inversions identified by Hi-C (Supplementary Table 12 and Supplementary Information). We observed that the rates of heterozygous sites and synonymous substitutions were elevated within the three inversions in a pattern consistent with the evolutionary strata (Fig. 2a, Supplementary Tables 5 and 13, and Extended Data Fig. 10b, c). We used deep whole-genome sequencing data of individuals from laboratory crosses to estimate the de novo mutation rate in sexual planarians to be approximately 1.0 × 10⁻⁸ mutations per nucleotide per generation. On the basis of this estimate, it is likely that inversions 3, 1 and 2 evolved approximately 450,000, 320,000 and 260,000 generations ago, respectively.

Although it is not possible to know how chromosome 1 will evolve, our findings provide a snapshot of an incipient stage that supports the
hypothesis that sex chromosomes evolve from homologous autosomes that acquire sex-specific roles and cease to recombine\(^1\). The locked J/V heterozygous system may facilitate this process by maintaining sex-specific alleles in the planarian population before the evolution of dioecy\(^2\). We propose that the planarian chromosome I haplotypes provide a unique opportunity to directly examine the molecular characteristics of a sex-primed autosome.

**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/s41586-022-04757-3.
**Methods**

**Planarian husbandry, RNAi and phenotyping**

Sexual planarians were fed organic beef liver once a week. All animals used for experiments were selected randomly. For RNAi, hatchlings around 2 weeks old from strain S2F10 were used. RNAi food was prepared by mixing 1 µg of double-stranded RNA with 10 µl of liver paste. To examine the RNAi phenotypes, photographs were taken with a stereomicroscope (Zeiss) when the animals were freely swimming. The lengths of the worms or bursa canal were measured with ImageJ software by multiple researchers in a double-blind experimental design. The ratio of the bursa canal to the whole worm was used to mitigate potential variations in worm size and their degree of relaxation caused by their soft bodies.

**Chromosome sequencing**

Chromosomes were collected from multiple animals of one clonal line, S2, maintained in the laboratory by amputation and regeneration. Chromosome spreads were prepared on nuclease-free membrane slides (Zeiss) according to a previously developed protocol except that, at the last step, the tissues were dissociated into single nuclei and placed onto the slides without squashing with a coverslip. Single chromosomes were identified under a ×40 lens and were collected into the caps of single PCR tubes by PALM MicroBeam laser microdissection (Zeiss). The collected chromosomes were spun down with a tabletop centrifuge in 4 µl PBS, and the DNA in the pellets was amplified with a REPLI-g Single Cell Kit (Qiagen) for sequencing on a MiSeq or HiSeq 3000 sequencing system (Illumina).

**Chromosome-scale genome assembly**

A Hi-C sequencing library was prepared from multiple animals of the S2 strain using the enzyme DpnII. Sequenced reads were aligned to dd_Smes_g4.fasta with bwa mem (version 0.7.17)\(^{1}\). An assembly file was prepared from the SALSA\(^{1}\) outputFINAL.fasta with juicebox scripts (Phase Genomics). The .hic file was prepared by run-assembly-visualizer. sh from three-dimensional de novo assembly\(^{44}\). The two files were loaded into Juicebox\(^{4}\) for scaffold manipulation using split, merge, order and orient commands and for chromosome assembly. The modified assembly file (chromosome-scale) was converted to fasta (Smed_chr_ref_v1) with juicebox assembly_converter.py.

**Hi-C library construction**

A Hi-C library was generated using Phase Genomics Proximo Animal Kit version 3.0. Approximately four worms were finely chopped and were then cross-linked for 20 min at room temperature with end-over-end mixing in 1 ml of Proximo cross-linking solution. The cross-linking reaction was terminated with quenching solution for 15 min at room temperature, again with end-over-end mixing. The quenched tissue was rinsed once with 1× chromatin rinse buffer (CRB), transferred to a liquid nitrogen-cooled mortar and ground to a fine powder. The powder was resuspended in 700 µl Proximo lysis buffer 1 and lysed with glass beads for 20 min at room temperature on a vortex mixer. A low-speed spin was used to clear the large debris, and the chromatin-containing supernatant was transferred to a new tube. After a second spin at high speed, the supernatant was removed, and the pellet containing the nuclear fraction of the lysate was washed with 1× CRB. After removal of the 1× CRB wash, the pellet was resuspended in 100 µl Proximo lysis buffer 2 and incubated at 65 °C for 15 min. Chromatin was bound to the recovery beads for 10 min at room temperature. The beads were placed on a magnetic stand and washed with 200 µl of 1× CRB.

The chromatin bound on the beads was resuspended in 150 µl Proximo fragmentation buffer, and 2.5 µl of Proximo fragmentation enzyme was added. The reaction was incubated for 1 h at 37 °C, cooled to 12 °C and then incubated with 2.5 µl of finishing enzyme for 30 min. After the addition of 6 µl of Stop Solution, the beads were washed with 1× CRB and were resuspended in 100 µl of Proximo ligation buffer supplemented with 5 µl of proximity ligation enzyme. The proximity ligation reaction was incubated at room temperature for 4 h with end-over-end mixing. To this, 5 µl of reverse cross-linking enzyme was added, and the reaction was incubated at 65 °C for 1 h.

After reversing the cross-links, the free DNA was purified with recovery beads, and the Hi-C junctions were bound to streptavidin beads and washed to remove unbound DNA. The washed beads were used to prepare paired-end deep sequencing libraries using Proximo library preparation reagents.

**Oocyte and sperm cell sequencing**

Sperm cells were released from sexually mature S2 strain animals into calcium- and magnesium-free buffer (1% BSA). The cell dissociation solution was placed onto a slide and examined under a phase-contrast microscope to identify single sperm cells. Oocytes were released from egg capsules. Single sperm cells or oocytes were transferred into single PCR tubes for amplification with a REPLI-g Single Cell Kit (Qiagen). RADseq libraries and whole-genome libraries were prepared and sequenced on a HiSeq 3000 or NovaSeq S2 sequencing system (Illumina). RADSeq sequencing data were analysed as described in the linkage map section. Whole-genome sequencing data were analysed as described in the recombination section.

**Synteny analysis**

For candidate genes, the protein sequences of genes of interest were obtained from *Schneideria* specimens. Such genes were aligned to *S. mansoni* genomes\(^{45}\) using the Protein to Nucleotide BLAST (tblastn) tool. To systematically examine the synteny of the whole genome, we used SonicParanoid (version 1.3.8)\(^{46}\) to identify one-to-one protein orthologues between *S. mediterranea* and *S. mansoni*. The *S. mansoni* genome assembly (V9) and its protein annotations are available at https://zenodo.org/record/5149023#.Ybk9jn3MK3I.

**Linkage map**

A J/J line, D5, was crossed to a J/V line, S2F8b, to build an F2 population of 93 animals. Genomic DNA was extracted from a fragment of each animal using an Easy-DNA gDNA Purification Kit (K180001, Thermofisher). Sequencing libraries for RADseq were prepared according to the procedures of Adapterama II\(^{50}\) with a few modifications. Genetic variants were identified using Stacks (version 2.41)\(^{46,47}\). All variants were filtered with VCFtools (version 0.1.14)\(^{46}\) to remove insertions and deletions and to select biallelic SNVs. Clusters of markers located within 200 bp were removed because they were likely to correspond to repetitive elements. Markers disobeying Mendelian segregation were also removed. Only markers that were homozygous in both parents were used. The linkage map was built with R/QTL\(^{45}\).

**Quantifying recombination**

Sequencing reads from S2 and its oocytes and sperm cells were aligned to Smed_chr_ref_v1 with bwa mem (version 0.7.17). Genetic variants were jointly called using the Genome Analysis Toolkit (GATK, version 4.1.4) with GenomicsDB and GenotypeGVCFs.\(^{50}\) Biallelic heterozygous markers in S2 were further filtered by removing abnormal markers, including those with no segregation of two alleles in the gametes, clusters of markers in close proximity (<200 bp apart) and markers of heterozygosity in sperm cells. The J and V haplotypes were manually phased for oocytes without recombination and were phased by MPR.genotyping\(^{51}\) for all gametes. The MPR.genotyping package was also used to impute or correct missing or erroneous genotypes. The final genotype bins were used to identify and visualize recombination with customized R code. Quantification of recombination was focused on crossovers between long tracks of haplotypes along a chromosome. Putative gene conversion events such as short tracks of haplotype switches encompassing <1% of the chromosome length were not included.
Gene expression analysis
To examine the gene content on chromosome 1, transcriptome data were downloaded from the NCBI Sequence Read Archive (SRA). To examine genes related to sexual development, transcriptomes from sexual adults, sexual juveniles and sexual adults with nhr-1 RNAi were used. To examine stem cell-enriched genes, transcriptomes from sorted X1 cells and CIW4 were used.

All sequencing data were aligned to dd_Smed_v6 (ref. 27) with bwa mem (version 0.7.17). Differential gene expression was analysed with DESeq2 (version 1.26.0) (ref. 31). Expression was quantified at the transcript level with kallisto (version 0.44.0) (ref. 32) and was imported and summarized to gene-level count matrices by tximport (ref. 33).

Haplotype-specific expression
To examine the haplotype-specific expression of critical regulators of the reproductive system, mRNA was extracted from six sexually mature animals of the S2J line and was analysed as three biological replicates, with two animals pooled into each replicate. Nine penis papillae were dissected from nine sexual adult animals of the same line and were analysed as three biological replicates, with penis papillae from three animals pooled into each replicate. All mRNA samples were extracted on the same day and were processed at the same time for library preparation and sequencing to minimize technical variation. Libraries for RNA-seq were prepared with a Clontech SMARTer Stranded Total RNA-seq (Pico) Kit. The workflow consisted of converting total RNA to cDNA and then adding adaptors for Illumina sequencing through PCR. The PCR products were purified, and ribosomal DNA was depleted. The cDNA fragments were further amplified with primers universal to all libraries. Lastly, the PCR products were purified again to yield the final cDNA library. Different adapters were used to multiplex samples in one lane. Sequencing was performed on a NovaSeq 6000 with a 150-bp paired-end-read run. Data quality checks were conducted using the Illumina Sequencing Analysis Viewer. Demultiplexing was performed using Illumina Bcl2fastq version 2.17.

All sequencing data were aligned to dd_Smed_v6 (ref. 27) with bwa mem (version 0.7.17). To ensure accuracy, haplotype-specific expression of the 14 genes of interest was manually examined using the Integrative Genomics Viewer. J or V allele counts were identified for each biallelic variant in the exons. For a particular gene, the allele counts were aggregated from all variants on all exons for the J and V haplotypes. The allele counts were then subjected to binomial testing and Bonferroni correction to determine whether the observed allele bias was statistically significant. $P > 1$ after Bonferroni correction was set to 1.

De novo map, parameter optimization and phylogenetic inference
All datasets were run through the de novo pipeline as implemented in Stacks version 2.52 (refs. 46,47). First, paired-end reads were demultiplexed and filtered for quality using PROCESS-RADTAGS and were truncated to a length of 135 bp. Individual reads with Phred scores below 30 or uncalled bases were discarded (96.9% of reads passed the quality filters). Optimal parameters were identified following the guidelines of ref. 48 by running multiple iterations of the de novo pipeline and varying just one parameter with each new iteration on a subset of 13 samples from the same population, (Sic_mar), following recommendations (ref. 49). We varied the minimum stack depth ($m$) between 1 and 6 (m1–m6), the number of mismatches allowed between stacks ($M$) between 1 and 10 (M1–M10), and the number of mismatches allowed to merge catalogue loci (n) while keeping all other parameters constant (m3, M2 and n0). We then compared the number of polymorphic assembled loci across samples using a sample representation cut-off of 80% (r80) and gain or loss of polymorphic loci with each new iteration. Once $m$ and $M$ were optimized, we assessed $n$ by evaluating the change in the number of polymorphic loci for $n = M - 1, n = M$ and $n = M + 1$. RAD loci were then assembled using the denovo_map.pl wrapper in Stacks, and the following parameters were set: $m = 3, M = 2$ and $n = 3$.

Assembled loci that were present in 75% of all individuals were kept from POPULATIONS, a minor allele frequency (MAF) filter of 0.04 (–min-maf) was used to filter out singleton SNPs that could mask population structure and a maximum observed heterozygosity (–max-obs-het) filter of 0.99 was used to remove potentially paralogous loci (ref. 50). Additionally, to build the phylogenetic tree, we concatenated all RADseq loci after filtering (–phylip-var-all). Phylogenetic trees were built by maximum likelihood using RAxML (ref. 50), starting from a random seed and applying a GTR+G substitution model and 1,000 bootstrap replicates. The sample from Menorca (Sme7-5) was used as an outgroup to root the tree.

Population structure analysis
For this analysis, we excluded the outgroup sequence and ran POPULATIONS to retain loci present in all populations (–pop 10) and 75% of the individuals present in each population (–r 0.75). On the basis of the loci that passed our filtering criteria, a random whitelist of 1,000 loci was generated and again run through POPULATIONS with the same criteria but retaining the first SNP at each locus (–write-single-snp). The output was exported into Structure format, and Structure version 2.3.4 (ref. 50) was used to infer population structure with 10,000 chains as burn-in and 100,000 MCMC chains with 20 iterations for $K = 1–11$. The resulting files were run through Structure Harvester (ref. 51), and the optimal $K$ was determined.

DNA FISH
To examine tissue expression of the genes gip and smedwi-1, we used probes and buffers for third-generation ISH chain reaction purchased from Molecular Instruments. Sexually mature planarians were treated with 7.5% N-acetyl-L-cysteine (Sigma-Aldrich) for 10 min and were then fixed in 4% paraformaldehyde (Electron Microscopy Sciences; 16% solution diluted 1:4 in PBS) for 20 min at room temperature. The repeat located near centromeres in Fig. 2c had the following sequence: TCTGGACTGGAAATTTTTTAATCTTTATAGGTGTATCTGTCAAATTTTATTGGTTCTATACATTTATATTCTCGAATTCTTCAGCTACATGAGCTTAAAAATGGCATACATAACGGGCTCGTCAAACACAAGTCAT. Figure 2a had the following sequence: AAAATCATAACGGGCTCGTCAAACACAAGTCAT. Hybridization chain reaction and RNA FISH
To examine tissue expression of the genes gip and smedwi-1, we used probes and buffers for third-generation ISH chain reaction purchased from Molecular Instruments. Sexually mature planarians were treated with 7.5% N-acetyl-L-cysteine (Sigma-Aldrich) for 10 min and were then fixed in 4% paraformaldehyde (Electron Microscopy Sciences; 16% solution diluted 1:4 in PBS) for 20 min. The copulatory apparatus was dissected into a 1.5-mL RNA-free tube. The rest of the procedures followed the hybridization chain reaction RNA FISH protocol of Molecular Instruments for whole-mount mouse embryos.

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

Data availability
The authors confirm that all data underlying the findings are fully available without restriction. All sequencing data are available from the NCBI SRA database (accession number PRJNA731187). The chromosome-scale genome assemblies for sexual S. mediterranea (including phased genomes and genomes with repetitive elements) are openly available on the Planosphere (https://planosphere.stowers.org/), PlanMine (https://planmine.mbpcl.mpg.de/planmine/begin.do), GenBank (GCA_022537955.1) and Zenodo (https://doi.org/10.5281/
Zenodo.5807415 databases. We used publicly available NCBI PacBio sequencing data (accession numbers SRX2700681–SRX2700684) and planarian transcriptome data (accession numbers SRR2658118–SRR2658125, SRR2658134–SRR2658141, SRR3473955–SRR3473957, SRR3629945–SRR3629952, SRR6351185–SRR6351188, SRR6351201–SRR6351204, SRR6351213–SRR6351216, SRR6363910–SRR6363927 and SRR6364586–SRR6364588).

Code availability

Custom code for data analysis can be openly accessed on the Zenodo database at https://doi.org/10.5281/zenodo.5807415.

Acknowledgements

We thank Longhua Guo or Leonid Kruglik for discussions on demographic history. We acknowledge Life Science Editors for professional editing services. Confocal laser-scanning microscopy was performed at the Advanced Light Microscopy/Spectroscopy Laboratory and the Leica Microsystems Center of Excellence at the California Nano Systems Institute at UCLA with funding from NIH Shared Instrumentation grant 1S10OD025017 and NSF Major Research Instrumentation grant CHE-0722519. This work was supported by grants from the Howard Hughes Medical Institute (L.K.) and the Helen Hay Whitney Foundation (L.G.).

Author contributions

L.K. and L.G. conceptualized the study. L.G. coordinated the project, conducted experiments and wrote the original draft. L.K. and L.G. conceptualized the project. L.G. coordinated the project, and the Helen Hay Whitney Foundation (L.G.).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03564-6

Correspondence and requests for materials

should be addressed to Longhua Guo or Leonid Kruglik.

Peer review information

Nature thanks Bo Wang 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 | Chromosome assignment by ChrSeq.**

(A–C) Chromosome sequencing depth of different scaffolds is consistent with linkage group assignment. Linkage groups: lg11 = chr1, lg1+lg4 = chr2, lg7 = chr3, lg17 = chr4. (D–F) Chromosome sequencing depth of different scaffolds is consistent with chromosome assignment in the genome assembly smed_chr_ref_v1. Chromosome sequencing samples: chromosome 1 (A, D); chromosome 2 (B, E); chromosome 3 (C, F).
Extended Data Fig. 2 | Putative inversions in the planarian genome detected by chromatin contact heatmap. (A) Chromosome 1 chromatin contact heatmap showing 3 potential inversions (dashed-line rectangles) visualized in Juicebox. (B) A complex structural variant on chromosome 2. Green outline: dd_Smes_g4 scaffolds. Blue outline: chromosome boundaries.
Extended Data Fig. 3 | Recombination in gametes and heterozygosity distribution in the wild on chromosomes 2 and 4. (A-B) Number of recombinations in oocytes and sperm on chromosomes 2 and 4 summarized as a violin plot (A) and a box plot (B). The median number of recombinations on chromosome 2 is 3 for sperm and 2 for oocytes, derived from a total of 49 sperm and 38 oocytes. The median number of recombinations on chromosome 4 is 2 for sperm and 1 for oocytes, derived from a total of 50 sperm and 26 oocytes.

| Chr2 sperm | Chr2 oocytes | Chr4 sperm | Chr4 oocytes |
|------------|-------------|------------|-------------|
| min = 0    | min = 0     | min = 0    | min = 0     |
| max = 7    | max = 7     | max = 7    | max = 3     |
| median = 3 | median = 2  | median = 2 | median = 1  |
| first quartile = 2 | first quartile = 0 | first quartile = 1 | first quartile = 1 |
| third quartile = 4 | third quartile = 4 | third quartile = 4 | third quartile = 2 |

(C-D) Very few sites are consistently heterozygous in more than 80% of the wild isolates on chromosomes 2 and 4 in Sicily (C) and in Sardinia (D).
Extended Data Fig. 4 | Representative images of late prophase I oocytes with ring-shaped chromosome 1. More than 100 oocytes from > 40 ovaries and > 20 sexual strains were examined for the presence of a ring-shaped chromosome 1. The experiment was repeated >10 times independently with similar results. Chromosome 1 always forms a ring in J/V strains in late prophase I oocytes. Shown in the figure are chromosomes from 8 different oocytes at late prophase I. Arrows point to the ring-shaped chromosome 1.
Extended Data Fig. 5 | Phylogenetic tree of wild isolates of Schmidtea mediterranea. The tree is rooted with one asexual animal from Menorca (Sme7-5). The tree on the left (A) includes two animals from SAR and SIC with higher genome heterozygosity than the cohorts from the same collection sites. Phylogenetic relations agree with the previous designation of eastern and western populations1. SAR: Sardinia; COR: Corsica; SIC: Sicily; TUN: Tunisia.
Extended Data Fig. 6 | Fixation index ($F_{ST}$) analysis of wild isolates. Pairwise population $F_{ST}$ for all collection sites from Sardinia (SAR), Corsica (COR), Menorca (MEN), Sicily (SIC), and Tunisia (TUN). The size and darkness of the blue circles (top right) correspond to the values of $F_{ST}$ (bottom left).
Evanno analysis supports a $K=2$ split, dividing the samples geographically into Eastern and Western groups. With $K=3$, a Tunisian signal was found to be present in all other 3 groups (light blue).

Extended Data Fig. 7 | STRUCTURE analysis of wild isolates. Evanno analysis supports a $K=2$ split, dividing the samples geographically into Eastern and Western groups. With $K=3$, a Tunisian signal was found to be present in all other 3 groups (light blue).
Extended Data Fig. 8 | Datasets for the examination of haplotype-specific expression. The sequencing reads from genomes (J-haplotype oocytes, V-haplotype oocytes, diploid S2 worms) and transcriptomes (pp, ww from the diploid S2 worms) were aligned to Smed_chr_ref_v1. Shown in the figure are reads for the gene CPEB-2. Two variants with J and V alleles are marked by colored bars. In the genome of the adult worm S2 (i.e., DNA), both J and V alleles are present at equal ratios. In the genome of oocytes (J or V haplotype, i.e., DNA), the variants have only one allele. These observations validate the heterozygous variants. In the transcriptomes (i.e., RNA) of penis papillae (pp) and whole worms (ww) from S2 adults, both J and V alleles are present. In the three transcriptomes of pp, J alleles were more highly expressed. In the three transcriptomes of ww, the J:V allele ratios vary and do not significantly depart from equal expression.
Extended Data Fig. 9 | Specific expression of smed-glp in the male reproductive organ (penis papillae) and female reproductive organ (bursa canal). Expression of the glp gene (in red color) and the stem cell marker smedwi-1 (in green color) in the dissected copulatory apparatus was visualized by Fluorescence in situ hybridization chain reaction. (A) White dashed lines outline penis papillae and bursa canal. Nuclei were stained by DAPI (in blue color). No glp expression in tissues surrounding penis papillae or bursa canal. Less than 1% of glp+ cells express the stem cell marker smedwi-1. (B) Cells expressing the two markers do exist in both bursa canal and penis papillae. This experiment was repeated 3 times independently with 6 sexually mature worms, producing similar results.
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | Embryonic lethality and synonymous divergence in Schmidtea mediterranea. (A) Two J/V lines with differential SNP markers on chromosome 3 were crossed. Zygotes were collected for genotyping. Heterozygous chromosome 3 markers validate the occurrence of fertilization. All three genotypes of chromosome 1 (J/J, J/V and V/V) were observed in zygotes. (B–C) Non-synonymous divergence (B) and Synonymous divergence (C) for all heterozygous sites in the J/V strain S2 and their distribution in different chromosomes summarized as boxplots. Chromosome 1 has the highest mean (red dots) and median dN or dS values. The min, max, median, mean, first quartile, and third quartile of dN (B) for each chromosome are as following: chr1 = (0, 0.00052, 0, 0.00017, 0, 0.0021), chr2 = (0, 0, 0, 0.000092, 0, 0), chr3 = (0, 0, 0.000031, 0, 0), and chr4 = (0, 0, 0, 0.000088, 0, 0). The min, max, median, mean, first quartile, and third quartile of dS (C) for each chromosome are as following: chr1 = (0, 0.024, 0.0022, 0.006, 0, 0.0098), chr2 = (0, 0, 0, 0.00021, 0, 0), chr3 = (0, 0, 0, 0.00093, 0, 0), and chr4 = (0, 0, 0, 0.00026, 0, 0). (D) Synonymous divergence for all heterozygous sites in the J/V strain S2, and their distribution in the three chromosome 1 inversions and the rest of the genome, summarized as boxplots. Mean (red dots) and median dS decreases in the following order: inversion 3 (inv3), inversion 1 (inv1), inversion 2 (inv2), and the rest of the genome (Not INV). One way ANOVA and Turkey Honest Significant Differences tests showed that dS were significantly different between all 3 inversions (inv2-inv1: p.adj = 0.005; inv3-inv1: p.adj < 1e-06; inv3-inv2: p.adj < 1e-06). The min, max, median, mean, first quartile, and third quartile of dS for each inversion are as following: inversion 1 = (0, 0.026, 0.0034, 0.0064, 0, 0.001), inversion 2 = (0, 0.021, 0.0026, 0.0053, 0, 0.0083), inversion 3 = (0, 0.036, 0.0084, 0.0091, 0, 0.015), and non-inversion = (0, 0, 0.0005, 0, 0). A total of 8317 genes from chromosome 1, 6568 genes from chromosome 2, 3394 genes from chromosome 3, and 1834 genes from chromosome 4 were plotted for non-synonymous divergence (B) and synonymous divergence (C). A total of 6633 genes from inversion 1, 754 genes from inversion 2, 298 genes from inversion 3, and 12428 genes from the rest of the genome were plotted for synonymous divergence (D).
Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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

- □ □ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

- □ □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

- □ □ The statistical test(s) used AND whether they are one- or two-sided

  *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*

- □ □ A description of all covariates tested

- □ □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

- □ □ 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)

- □ □ For null hypothesis testing, the test statistic (e.g. f, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

  *Give P values as exact values whenever suitable.*

- □ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

- □ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

- □ □ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), 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

ImageJ (Fiji, 1.53i51)

Data analysis

Iowa mem (version 0.7.17), SALSA, Juicebox, 3d-dna, sonic paranoid (v1.3.8), Stacks (version 2.41; 2.52), VCFtools (version 0.1.14), GATK (v.4.1.4.1), DEseq2 (version 1.26.0), kallisto (version 0.44.0), Bcl2fastq2 (version 2.17), RAWML-NG (version 0.9.0), STRUCTURE (version 2.3.4), custom codes [https://zenodo.org/record/5807415#.Yig0VRD Mk3]

For manuscripts utilizing custom algorithms or software that are not central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitLab). See the Nature Portfolio 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 description of any restrictions on data availability

- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The authors confirm that all data underlying the findings are fully available without restriction. All sequencing data are available from the NCBI SRA database (accession number PRJNA731187). The chromosome-scale genome assemblies for sexual Schistosoma mediterranea (including phased genomes and genomes with repetitive elements) are openly available at: planosphere(planosphere.sliders.org); planmine (planmine.mbl.bpc.mspg.de/planmine/begin.do); Zenodo (10.5281/zenodo.5807415). Publicly available Pacbio sequencing data from NCBI were used (accessions: SRRX700681-SRRX700689). Publicly available planarian transcriptome data from NCBI were used (accessions: SRR2658118 SRR2658125, SRR2658134 SRR2658144, SRR3479365 SRR3479357, SRR3629945 SRR3629952).
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

Life sciences study design

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

- **Sample size**
  The numbers of chromosomes, sperm, oocytes, wild isolates and animals used in each experiment range from 3 to more than 20. The numbers were chosen to ensure representative samples; the specific choices for each experiment were guided by practical considerations of reproducibility and feasibility.

- **Data exclusions**
  Failed sequencing samples were excluded if the DNA amplification failed or the sequencing resulted with little data. No data exclusions were performed on successful samples.

- **Replication**
  All experiments were repeated at least three times. All attempts to replicate the experiments were successful.

- **Randomization**
  All chromosomes, sperm, oocytes and animals were selected randomly. For experimental and control groups, animals were allocated randomly.

- **Blinding**
  The measurements of the length of female bursa canal or worms were double-blinded and repeated at least twice. In other experiments, blinding is not essential. For example, in the analysis of recombination or wild diversity, all gametes, animals, chromosomes were processed together as one dataset. The researchers were essentially blinded to the relevance of a particular variant, chromosome or animal.

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
- □ Antibodies
- □ Eukaryotic cell lines
- □ Palaeontology and archaeology
- □ Animals and other organisms
- □ Human research participants
- □ Clinical data
- □ Dual use research of concern

### Methods

- **n/a** Involved in the study
- □ ChIP-seq
- □ Flow cytometry
- □ MRI-based neuroimaging

Animals and other organisms

Policy information about: studies involving animals. ARRIVE guidelines recommended for reporting animal research

- **Laboratory animals**
  The animals used in the study are the sexual biotype of Schmidtea mediterranea, which are simultaneous hermaphrodites. Sexually mature adults (4 month old to 1 year old) were used.

- **Wild animals**
  No live wild animals were used. DNA samples previously extracted from wild animals were sequenced and analyzed.

- **Field-collected samples**
  Wild isolates of Schmidtea mediterranea were collected from the field and immediately fixed in 70% ethanol. DNA were extracted from these animals in earlier studies. Such DNA samples were used in the current study.

- **Ethics oversight**
  As Schmidtea mediterranea is a free-living invertebrate and harmless to humans/environments, no university ethical approval is needed. We were mindful to give the animals best housing conditions (e.g., constant optimal temperature, clean water, consistent feeding schedules, ample space) and best euthanasia practice (e.g., quick treatment with 10% bleach). We followed guidelines of ARRIVE in our reporting.

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