Island-specific evolution of a sex-primed autosome in the planarian *Schmidtea mediterranea*

Longhua Guo^{1,6*}, Joshua Bloom^{1,6}, Daniel Dols Serrate², Eyal Ben David³, Olga T. Schubert^{1,6}, Kaiya Kazuma¹, Katarina Ho¹, Yubao Wei⁴, Daniel Leighton^{1,6}, James Boocock^{1,6}, Tzitziki Lemus Vergara^{1,6}, Marta Riutort², Alejandro Sánchez Alvarado^{5,6}, Leonid Kruglyak^{1,6*}

¹ Department of Human Genetics, Department of Biological Chemistry, and University of California Los Angeles, CA, USA

² Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, and Institut de Recerca de la Biodiversitat (IRBio), Barcelona, Catalonia, Spain

³ Department of Biochemistry and Molecular Biology, The Institute for Medical Research Israel-Canada, The Hebrew University-Hadassah Medical School, Ein Kerem, Jerusalem, Israel

⁴ Institute of Reproductive Medicine, Henan Provincial People's Hospital, Zhengzhou University, Zhengzhou, China

⁵ Stowers Institute for Medical Research, and, Kansas City, MO, USA

⁶ Howard Hughes Institute for Medical Research, Chevy Chase, MD, USA

* To whom correspondence should be addressed. E-mail: longhuagu@mednet.ucla.edu (LG), lkruglyak@mednet.ucla.edu (LK).

24

25 **Abstract**

26 The sexual biotype of the planarian *Schmidtea mediterranea* is a hermaphrodite
 27 indigenous to Tunisia and several Mediterranean islands. Here, we isolated individual
 28 chromosomes and used sequencing, Hi-C and linkage mapping to assemble a
 29 chromosome-scale genome reference. The linkage map revealed an extremely low rate
 30 of recombination on chromosome 1. We confirmed suppression of recombination on
 31 chromosome 1 by quantifying recombination events in individual sperm and oocytes. We
 32 showed that the extensive heterozygous regions, previously designated as J and V
 33 haplotypes, comprise essentially all of chromosome 1. Genome sequencing of individuals
 34 isolated in the wild indicated that this heterozygosity has evolved specifically in
 35 populations from Sardinia and Corsica. We found that chromosome 1 acquired many
 36 genes that determine the development of female and male reproductive systems, along
 37 with haplotype-specific expression of such sex specific genes. These molecular
 38 signatures and restricted recombination in a hermaphrodite led us to propose that
 39 chromosome 1 is a sex-primed autosome, providing direct molecular evidence for the
 40 traditional model of how sex chromosomes may have evolved from autosomes.

41

42
43 Sex chromosomes are thought to have evolved from homologous autosomes that
44 acquired sex-determining genes and lost their ability to recombine (1-10). As such, sex
45 chromosome evolution and recombination suppression are closely associated(1-10). Yet,
46 direct evidence of such homologous autosomes primed for evolution into sex
47 chromosomes is hard to capture, and little is presently known about the molecular
48 signatures associated with the evolution of recombination suppression.

49
50 The freshwater planarian *S. mediterranea* is an important model organism for studies of
51 regeneration (11, 12). It exists as asexual and sexual reproductive biotypes. The sexual
52 biotype is mostly distributed in Tunisia, and on the islands of Sardinia, Corsica, and Sicily
53 (13). The sexual biotype is a simultaneous hermaphrodite: it develops both male and
54 female reproductive systems in the same adult individual and obligately outcrosses to
55 fertilize other individuals (14, 15). We hypothesized that studying chromosome evolution
56 in a simultaneous hermaphrodite might provide insights into the early evolution of a
57 primitive sex chromosome.

58
59 *S. mediterranea* has 4 pairs of chromosomes, which are stably diploid. The genome is
60 reported to comprise 774 Mb assembled as 481 non-contiguous series of genomic
61 sequences or ‘scaffolds’ (16-18). In a previous study, we found that ~300 Mb of the
62 genome remains heterozygous even after extensive inbreeding of laboratory strains, and
63 that this phenomenon also occurs naturally in wild populations in Sardinia (14). The two
64 sets of heterozygous alleles were collectively named the J and V haplotypes. To define

the chromosomal locations of these alleles and investigate the reasons underpinning the persistence of heterozygosity in *S. mediterranea*, a detailed assembly of all four chromosomes was needed.

Chromosome-Scale Genome Assembly

To transform the 481 scaffolds (16) into a chromosome-scale genome reference, we carried out chromosome sequencing (ChrSeq) (19, 20) of a lab strain, as well as chromatin proximity sequencing by Hi-C (21, 22) (**Fig. 1A**). To do so, we dissected individual chromosomes from mitotic cells by 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 chromosome to ensure reproducibility and multiple samples of different chromosomes to ensure specificity (**fig. S1**). Overall, we successfully amplified and confidently assigned to one of the four chromosomes 740 Mb of the 774 Mb genome (**table S1**).

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 named *Smed_chr_ref_v1*. Hi-C data analyzed by the scaffolding algorithm SALSA (21) resolved 481 scaffolds into 57 super-scaffolds and 104 singletons. ChrSeq revealed that 3 of the 57 super-scaffolds were disjointed inter-chromosomal fragments, consistent with the Hi-C contact heatmap (**Fig. 1C**). We split the disjointed super-scaffolds, merged the scaffolds into chromosomes, and ordered and oriented all scaffolds within the chromosomes by using the visualization software Juicebox (23)

(Methods). Chromosome assignments by ChrSeq alone (**table S1**) and by Hi-C alone were inconsistent for only 3 of 384 (0.8%) scaffolds (**table S2**). We assigned these 3 scaffolds to chromosomes manually based on the Hi-C data. Hi-C also detected 26 inter-chromosomal or intra-chromosomal assembly errors (**table S2**) in the previous assembly (16), 5 of which were confirmed by ChrSeq to be inter-chromosomal disjointed scaffolds (**table S2**). The final genome assembly (Smed_chr_ref_v1, **Fig.1C**) has 4 chromosomes with a total size of 764 Mb, 98.4% of that reported previously (16). Of the 1.6% of the previous assembly that are not contained in Smed_chr_ref_v1 by Hi-C, around half (52.6%) of the scaffolds could not be assigned to a chromosome or were assigned to 2 chromosomes by ChrSeq, indicating their lower assembly quality (**table S3**); 33.0% and 12.4% of them had chromatin interaction signals with structurally complex regions of chromosome 1 and 2 (**fig. S2**), respectively, suggesting that they might be alternative assemblies or repetitive sequences. In our new assembly, Smed_chr_ref_v1, the two ends of chromosome 4 are 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 F2 population and genotyped individual worms with RADseq (24). Eighty markers that are evenly distributed and genotyped in at least 98% of the F2 segregants (91 of 93) were used to establish 4 linkage groups (**Fig. 1D, table S4**). These 4 linkage groups represent the 4 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 reexamined 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 heterozygous markers were located on the other 3 chromosomes at a density of 3,274 markers per 10 Mb; these likely correspond to differences between highly similar copies of repetitive elements rather than to true polymorphisms (**table S5**). All F2 worms (n=93) in this study were heterozygous for chromosome 1 but not for chromosomes 2–4 (**table S6**), consistent with our previous study (14). Hence, we conclude that the J/V haplotypes are on chromosome 1.

Our linkage map revealed an extremely low rate of recombination on chromosome 1 (only 0.5 cM for the entire chromosome; **Fig. 1D**). This is particularly remarkable because at 333 Mb, chromosome 1 is the largest of the 4 chromosomes, containing more than 40% of the genome.

To examine directly whether chromosome 1 can recombine, we sequenced 45 single sperm and 28 single oocytes from a J/V line, S2 (**Fig. 2B**). 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 on chromosome 3, covering 99% of the length of each chromosome (**table S7**). SNVs were distributed at a similar density across 20 Mb windows (coefficient of variation 0.38 and 0.31 for chromosome 1 and 3) (**table S7**). We observed that 98% (44/45) of the sperm and 93% (26/28) of the oocytes had no crossovers on chromosome 1; by contrast, only one sperm and two oocytes had no crossovers on chromosome 3 (**Fig. 2B**). We conclude from this data that recombination on chromosome 1 is strongly suppressed.

Consistent with our conclusion that recombination is suppressed on chromosome 1, 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**). Fluorescence *in situ* hybridization with telomere probes showed that crossovers between the two homologous pairs of chromosome 1 happened only in regions close to the telomeres, leading to the observed ring conformation rather than the side-by-side pairing seen for chromosomes 2, 3 and 4. Furthermore, Hi-C analysis showed that chromosome 1 has three putative inversions, each >20 Mb (**Fig. 2D**); such inversions are an established cause of crossover suppression (25-27). The rest of the genome has only one large inversion of ~10 Mb on chromosome 2 (**fig. S2**).

Island-Specific Evolution of Chromosome 1

To investigate the genetic diversity of *S. mediterranea* in its natural environment and to determine whether chromosome 1 J/V heterozygosity happens in the entire species, we

used RADseq (24) to sample the genomes of 70 sexual individuals from Sardinia, Corsica, Sicily and Tunisia and 2 asexual individuals from Menorca (13) (**Fig. 3A**). To look for genetic relationships between the individuals, we determined clustering of relatedness measured by identity-by-state pairwise distances (28) 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 (**fig. S3**) and a STRUCTURE analysis (**fig. S4**) further supported this population structure.

The relatedness of the populations from Sardinia and Corsica suggested 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 heterozygosity on the other three chromosomes in these populations and on all four chromosomes in the other populations closely followed expectation. By analyzing J/V haplotype markers in the wild populations, we found that the animals from Sardinia/Corsica were heterozygous J/V whereas those from Sicily/Tunisia were homozygous J/J. Moreover, the animals from Sardinia (n=28) contained many sites that were heterozygous in more than 80% of the individuals and that were distributed along the length of chromosome 1 except near the ends, whereas the animals from Sicily (n=27) had very few such heterozygous sites (**Fig. 3, D and E, left**) (**table S8**). Few such heterozygous sites were observed on chromosome 3 in either population (**Fig. 3, D and E, right**). 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 chromosome 1. We found that all currently characterized genes that are essential for development of both male and female reproductive systems were found on chromosome 1, even though differential gene expression analysis between juveniles and adults showed that adult-enriched genes were distributed equally among all 4 chromosomes (**fig. S5**). The critical regulatory genes—*nanos* (29), *nhr-1* (30, 31), *npy-8* (32), *npyr-1* (33), and *CPEB-2* (34)—were not clustered but distributed along the length of chromosome 1 (**Fig. 4A**). *Nanos*, *nhr-1* and *CPEB-2* were transcribed from both J and V haplotypes (**fig. S6**). No genetic variants that would allow determination of J- or V-haplotype-specific expression were found in *npy-8* and *npyr-1* transcripts. The presence of these genes on chromosome 1 and their crucial roles in sexual development indicate the importance of chromosome 1 integrity in the maintenance of sexual reproduction. The asexual lineage of *S. mediterranea*, which has a translocation from chromosome 1 to chromosome 3 and is devoid of any reproductive organs(12), likely evolved through loss of function of one or several of these genes.

We found that chromosome 1 has acquired many genes with male or female-specific functions. The *CPEB-1* gene on chromosome 1 (**Fig. 4B**) is specifically required for development of the female reproductive system (34). The presence of a female-determining gene on a chromosome that does not recombine provides the 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 leave the other chromosome with a unique role in determining the female sex. An extended search of the literature identified 13 additional genes with well characterized male or female-specific functions that are located on chromosome 1 (**table S9**) (29, 34). Two of these genes, Bic-C and sped-1, are in synteny with the ZW sex chromosomes of the parasitic flatworm, *Schistosoma mansoni* (35, 36). Loss of Bic-C expression leads to a “no testes” phenotype, without affecting the ovary (29).

To examine whether chromosome 1 has evolved a signature of a sex chromosome—haplotype-specific expression of male- or female-specific genes—we examined the expression of these 14 genes in the transcriptomes of sexually mature adult worms and of their male copulatory organ, the penis papilla (pp) (37), in a J/V line. Nine of the 14 genes had at least one heterozygous variant in the coding sequence, which allowed us to determine if these genes showed biased expression from the J or V chromosome. We found that a gelsolin-like gene (KY847538.1) was expressed predominantly from the J chromosome in both whole-worm and pp transcriptomes, with a larger effect in the male copulatory organ. Three genes (KY847529.1, KY847536.1, and Smed-MSY4) had biased expression in the whole worm transcriptome, and two genes (KY847532.1 and KY847539.1) had biased expression in the male copulatory organ (**Fig. 4B, table S9**). Haplotype-specific expression in 6 of 9 female- or male-related genes suggests that sex-specific *cis*-regulatory variation is maintained on chromosome 1.

We have identified a chromosome in the hermaphrodite planarian *S. mediterranea* that does not recombine and has acquired many genes with sex-specific functions, making it primed for evolution into a sex chromosome. The suppression of recombination on this chromosome has evolved specifically on the islands of Sardinia and Corsica; such a phenomenon is absent from the sampled populations of Sicily and Tunisia. Although we cannot know how chromosome 1 will evolve, our findings provide a snapshot that supports the hypothesis that sex chromosomes evolved from homologous autosomes that acquired sex-specific roles and ceased to recombine (1-8). The locked J/V heterozygous system may play a facilitating role in this process by maintaining sex-specific alleles in the planarian population prior to the evolution of dioecy (38). We propose that the planarian chromosome 1 haplotypes provide a unique opportunity to directly examine the molecular characteristics of a sex-primed autosome.

Methods

Chromosome sequencing

Chromosomes were collected from multiple animals of one clonal line, S2, maintained in the lab 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 dropped onto the slides without squashing with a coverslip (39). Single chromosomes were identified under a 40x lens and collected into caps of single PCR tubes by PALM MicroBeam laser microdissection (Zeiss). Collected chromosomes were spun down with

a tabletop spinner in 4ul of PBS, and the DNA in the pellets was amplified with REPLI-g Single Cell kit (QIAGEN) for sequencing on the 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, following the instructions of Phase Genomics Proximo Animal kit version 3.0 with few modifications. Sequenced reads were aligned to dd_Smes_g4.fasta (16) with bwa mem (version 0.7.17) (40). An assembly file was prepared from the SALSA (21) output FINAL.fasta with juicebox_scripts (Phase Genomics). The assembly and .hic files were loaded into Juicebox (23) for scaffold manipulation (i.e., split, merge, order, orient) and chromosome assembly. The modified assembly file (chromosome-scale) was converted to fasta (Smed_chr_ref_v1) with juicebox_assembly_converter.py.

Oocyte and sperm sequencing

Sperm were released from sexually mature S2 strain animals into CMFB (1% BSA) buffer. The cell dissociation solution was dropped onto a slide and visualized under a phase contrast microscope to identify single sperm. Oocytes were released from egg capsules. Single sperm or oocytes were transferred into single PCR tubes for amplification with the REPLI-g Single Cell kit (QIAGEN). RADseq libraries and whole genome libraries were prepared and sequenced on the HiSeq 3000 or NovaSeq S2 sequencing system (Illumina). RADseq sequencing data were analyzed as described in the linkage map

section. Whole-genome sequencing data were analyzed as described in the recombination section.

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 with Easy-DNA gDNA purification kit (K180001, ThermoFisher). Sequencing libraries for RADseq were prepared according to the procedures of Adapterama III(24) with few modifications. Genetic variants were identified with Stacks (version 2.41)(41, 42). All variants were filtered with VCFtools (version 0.1.14)(43) to remove insertions and deletions and to select biallelic single nucleotide variants. Clusters of markers that are located within 200bp were removed as they most likely correspond to repetitive elements. Markers that disobey Mendelian segregation were also removed. Only markers that are homozygous in both parents were used to build the linkage map with R/QTL(44).

Quantifying recombination

Sequencing reads from S2, its oocytes, and sperm were aligned to Smed_chr_ref_v1 with bwa mem (version 0.7.17). Genetic variants were jointly called by using the genome analysis toolkit GATK (v.4.1.4.1) with GenomicsDB and GenotypeGVCFs (45). Biallelic heterozygous markers in S2 were further filtered by removing abnormal markers: markers with no segregation of two alleles in the gametes, clusters of markers in close proximity (<200bp), and markers heterozygous in sperm. J and V haplotypes were manually phased with oocytes without recombination and phased by MPR.genotyping (46) with 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 (i.e., short tracks of haplotype switches encompassing < 1% of chromosome length) were not included.

Synteny analysis

Protein sequences of genes of interest were obtained from *S. mediterranea*. Such genes were aligned to *S. mansoni* genomes (36, 47) with tblastn.

Gene expression analysis

To examine gene contents on chromosome 1, transcriptome data were downloaded from NCBI Sequence Read Archive (SRA). To examine genes related to sexual development, transcriptomes from sexual adults (30, 34, 48), sexual juveniles (30), and sexual adults with nhr-1 RNAi (30) were used. To examine stem-cell-enriched genes, transcriptomes from sorted X1 cells and CIW4 were used (49, 50).

All sequencing data were aligned to dd_Smed_v6 (17) with bwa mem (version 0.7.17). Differential gene expression was analyzed with DEseq2 (version 1.26.0) (51). Expression was quantified at the transcript level with kallisto (version 0.44.0) (52), and imported and summarized to gene-level count matrices by tximport (53).

Haplotype-specific expression

To examine haplotype-specific expression of critical regulators of the reproductive system, mRNA was extracted from six sexually mature animals of a J/V line, S2, and analyzed as three biological replicates, with 2 animals pooled into each replicate. Nine penis papillae (pp) were dissected from nine sexual adult animals of the same line and analyzed as three biological replicates, with pp from 3 animals pooled into each replicate. All mRNA samples were extracted on the same day and processed at the same time for library preparation and sequencing to minimize technical variation. Libraries for RNA-Seq were prepared with Clonetechn SMARTer Stranded Total RNA-Seq (Pico) Kit. The workflow consists of converting total RNA to cDNA, and then adding adapters for Illumina sequencing through PCR. The PCR products are purified, and then ribosomal cDNA is depleted. The cDNA fragments are further amplified with primers universal to all libraries. Lastly, the PCR products are purified once more to yield the final cDNA library. Different adaptors were used to multiplex samples in one lane. Sequencing was performed on Illumina Novaseq 6000 with a paired-end-read 150 bp run. Data quality checks were done on Illumina SAV. Demultiplexing was performed with Illumina Bcl2fastq2 v 2.17.

All sequencing data were aligned to dd_Smed_v6 (17) with bwa mem (version 0.7.17). To ensure accuracy, haplotype-specific expression of the 14 genes of interest was manually examined in IGV. 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 J or V haplotypes. The allele counts were then subjected to binomial test and Bonferroni correction (54) to determine if the observed allele bias was statistically significant. P-values > 1 after Bonferroni correction were 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 (41, 42). First, paired-end reads were de-multiplexed and filtered for quality using PROCESS_RADTAGS and 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 Paris et al. (55) guidelines 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 (56). 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 catalog loci (*-n*) while keeping all other parameters constant (m3, M2 and n0). We then compared the number of polymorphic assembled loci across samples using the 80% sample representation cutoff (*r80*) and the gain or loss of polymorphic loci with each new iteration. Once *-m* and *-M* were optimized, we assessed *-n* by evaluating the change in number of polymorphic loci for $n = M - 1$, $n = M$ and $n = M + 1$. RAD loci were then assembled using the wrapper 'denovo_map.pl' in Stacks and the following parameters set: $m = 3$, $M = 2$ and $n = 3$. Assembled loci that were present in 75% of all individuals were kept from POPULATIONS, and a minor allele frequency (MAF) filter of 0.04 (*--min-maf*) was used to filter out singleton SNPs that could mask population structure, along with a maximum observed heterozygosity (*--max-obs-het*) filter of 0.99 to remove potentially paralogous loci (57). 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-NG

v0.9.0 (58), starting from a random seed and applying a GTR+G substitution model and 1000 bootstrap replicates. The sample from Menorca (Sme7-5_men) 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 (-p 10) and 75% of the individuals present in each population (-r 0.75). Based on the loci that passed our filtering criteria, a random whitelist of 1000 loci was generated and run through POPULATIONS once again, with the same criteria but retaining the first SNP at each locus (--write-single-snp). The output was exported in STRUCTURE format, and STRUCTURE 2.3.4 (59) 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 (60), and the optimal K was determined (61).

Acknowledgements

We thank Drs. Labib Rouhana, Giancarlo Bruni, and Stefan Zdraljevic for insightful discussions, Zain Kashif and Emily Warda for planarian husbandry, and Xinmin Li and the UCLA Technology Center for Genomics & Bioinformatics (TCGB) for next-generation sequencing. We thank Drs. Ying Zeng, Ekaterina Noskova, and Dobrynin Pasha 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 S10OD025017 and NSF Major Research Instrumentation grant CHE-0722519.

Funding

This work was supported by grants from the Howard Hughes Medical Institute (LK) and the Helen Hay Whitney Foundation (LG).

Competing interests

The authors have declared that no competing interests exist.

Data availability

The authors confirm that all data underlying the findings are fully available without restriction. All sequencing data will be available from the NCBI SRA database (accession number PRJNA731187). The chromosome scale genome assembly for sexual *Schmidtea mediterranea* is openly available at: <https://planosphere.stowers.org/>; <http://planmine.mpi-cbg.de/planmine/begin.do>

References

1. D. Bachtrog, A dynamic view of sex chromosome evolution. *Curr Opin Genet Dev* **16**, 578-585 (2006).
2. B. T. Lahn, D. C. Page, Four evolutionary strata on the human X chromosome. *Science* **286**, 964-967 (1999).
3. W. R. Rice, Evolution of the Y Sex Chromosome in Animals. *BioScience* **46**, 331-343 (1996).
4. B. Charlesworth, The evolution of chromosomal sex determination and dosage compensation. *Curr Biol* **6**, 149-162 (1996).
5. J. Hodgkin, Genetic sex determination mechanisms and evolution. *Bioessays* **14**, 253-261 (1992).

6. B. Charlesworth, The evolution of sex chromosomes. *Science* **251**, 1030-1033 (1991).
7. E. Jablonka, M. J. Lamb, The evolution of heteromorphic sex chromosomes. *Biol Rev Camb Philos Soc* **65**, 249-276 (1990).
8. J. J. Bull, *Evolution of sex determining mechanisms*. (The Benjamin/Cummings Publishing Company, Inc., 1983), pp. 316.
9. H. J. Muller, Genetic Variability, Twin Hybrids and Constant Hybrids, in a Case of Balanced Lethal Factors. *Genetics* **3**, 422-499 (1918).
10. D. Charlesworth, Evolution of recombination rates between sex chromosomes. *Philos Trans R Soc Lond B Biol Sci* **372**, (2017).
11. P. W. Reddien, A. Sanchez Alvarado, Fundamentals of planarian regeneration. *Annu Rev Cell Dev Biol* **20**, 725-757 (2004).
12. P. A. Newmark, A. Sanchez Alvarado, Not your father's planarian: a classic model enters the era of functional genomics. *Nat Rev Genet* **3**, 210-219 (2002).
13. E. M. Lazaro *et al.*, Schmidtea mediterranea phylogeography: an old species surviving on a few Mediterranean islands? *BMC Evol Biol* **11**, 274 (2011).
14. L. Guo, S. Zhang, B. Rubinstein, E. Ross, A. S. Alvarado, Widespread maintenance of genome heterozygosity in Schmidtea mediterranea. *Nat Ecol Evol* **1**, 19 (2016).
15. R. M. Zayas *et al.*, The planarian Schmidtea mediterranea as a model for epigenetic germ cell specification: analysis of ESTs from the hermaphroditic strain. *Proc Natl Acad Sci U S A* **102**, 18491-18496 (2005).
16. M. A. Grohme *et al.*, The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms. *Nature* **554**, 56-61 (2018).
17. H. Brandl *et al.*, PlanMine--a mineable resource of planarian biology and biodiversity. *Nucleic Acids Res* **44**, D764-773 (2016).
18. S. M. Robb, E. Ross, A. Sanchez Alvarado, SmedGD: the Schmidtea mediterranea genome database. *Nucleic Acids Res* **36**, D599-606 (2008).
19. A. Weise *et al.*, High-throughput sequencing of microdissected chromosomal regions. *Eur J Hum Genet* **18**, 457-462 (2010).
20. L. Ma *et al.*, Direct determination of molecular haplotypes by chromosome microdissection. *Nat Methods* **7**, 299-301 (2010).
21. J. N. Burton *et al.*, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119-1125 (2013).
22. E. Lieberman-Aiden *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
23. N. C. Durand *et al.*, Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99-101 (2016).
24. N. J. Bayona-Vasquez *et al.*, Adapterama III: Quadruple-indexed, double/triple-enzyme RADseq libraries (2RAD/3RAD). *PeerJ* **7**, e7724 (2019).
25. T. Dobzhansky, C. Epling, The Suppression of Crossing Over in Inversion Heterozygotes of *Drosophila Pseudoobscura*. *Proc Natl Acad Sci U S A* **34**, 137-141 (1948).
26. D. E. Miller *et al.*, The Molecular and Genetic Characterization of Second Chromosome Balancers in *Drosophila melanogaster*. *G3 (Bethesda)* **8**, 1161-1171 (2018).

27. D. E. Miller *et al.*, Rare recombination events generate sequence diversity among balancer chromosomes in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **113**, E1352-1361 (2016).
28. X. Zheng *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012).
29. Y. Wang, R. M. Zayas, T. Guo, P. A. Newmark, nanos function is essential for development and regeneration of planarian germ cells. *Proc Natl Acad Sci U S A* **104**, 5901-5906 (2007).
30. S. Zhang *et al.*, A nuclear hormone receptor and lipid metabolism axis are required for the maintenance and regeneration of reproductive organs. *bioRxiv*, 279364 (2018).
31. M. E. Tharp, J. J. Collins, 3rd, P. A. Newmark, A lophotrochozoan-specific nuclear hormone receptor is required for reproductive system development in the planarian. *Dev Biol* **396**, 150-157 (2014).
32. J. J. Collins, 3rd *et al.*, Genome-wide analyses reveal a role for peptide hormones in planarian germline development. *PLoS Biol* **8**, e1000509 (2010).
33. A. Saberi, A. Jamal, I. Beets, L. Schoofs, P. A. Newmark, GPCRs Direct Germline Development and Somatic Gonad Function in Planarians. *PLoS Biol* **14**, e1002457 (2016).
34. L. Rouhana, J. Tasaki, A. Saberi, P. A. Newmark, Genetic dissection of the planarian reproductive system through characterization of *Schmidtea mediterranea* CPEB homologs. *Dev Biol* **426**, 43-55 (2017).
35. M. Berriman *et al.*, The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352-358 (2009).
36. A. V. Protasio *et al.*, A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl Trop Dis* **6**, e1455 (2012).
37. T. Chong, J. M. Stry, Y. Wang, P. A. Newmark, Molecular markers to characterize the hermaphroditic reproductive system of the planarian *Schmidtea mediterranea*. *BMC Dev Biol* **11**, 69 (2011).
38. D. Charlesworth, S. David, *The evolution of dioecy and plant sex chromosome systems*. Sex Determination in Plants (Garland Science, ed. 1st, 1999).
39. L. Guo *et al.*, An adaptable chromosome preparation methodology for use in invertebrate research organisms. *BMC Biol* **16**, 25 (2018).
40. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
41. J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, W. A. Cresko, Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**, 3124-3140 (2013).
42. J. M. Catchen, A. Amores, P. Hohenlohe, W. Cresko, J. H. Postlethwait, Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda)* **1**, 171-182 (2011).
43. P. Danecek *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
44. K. W. Broman, H. Wu, S. Sen, G. A. Churchill, R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890 (2003).
45. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

46. W. Xie *et al.*, Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc Natl Acad Sci U S A* **107**, 10578-10583 (2010).
47. F. Luo *et al.*, An improved genome assembly of the fluke *Schistosoma japonicum*. *PLoS Negl Trop Dis* **13**, e0007612 (2019).
48. E. L. Davies *et al.*, Embryonic origin of adult stem cells required for tissue homeostasis and regeneration. *Elife* **6**, (2017).
49. A. Zeng *et al.*, Prospectively Isolated Tetraspanin(+) Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell* **173**, 1593-1608 e1520 (2018).
50. E. M. Duncan, A. D. Chitsazan, C. W. Seidel, A. Sanchez Alvarado, Set1 and MLL1/2 Target Distinct Sets of Functionally Different Genomic Loci In Vivo. *Cell Rep* **13**, 2741-2755 (2015).
51. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
52. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527 (2016).
53. C. Sonesson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).
54. Etymologia: Bonferroni correction. *Emerg Infect Dis* **21**, 289 (2015).
55. J. R. Paris, J. R. Stevens, J. M. Catchen, Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**, 1360-1373 (2017).
56. N. C. Rochette, J. M. Catchen, Deriving genotypes from RAD-seq short-read data using Stacks. *Nat Protoc* **12**, 2640-2659 (2017).
57. C. S. Stobie, C. J. Oosthuizen, M. J. Cunningham, P. Bloomer, Exploring the phylogeography of a hexaploid freshwater fish by RAD sequencing. *Ecol Evol* **8**, 2326-2342 (2018).
58. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453-4455 (2019).
59. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).
60. D. A. Earl, B. M. vonHoldt, STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361 (2012).
61. G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**, 2611-2620 (2005).
62. L. A. Ouellette, R. W. Reid, S. G. Blanchard, C. R. Brouwer, LinkageMapView-rendering high-resolution linkage and QTL maps. *Bioinformatics* **34**, 306-307 (2018).

Figure Legends

Figure 1 Chromosome-scale genome assembly (A) Schematic of using Hi-C, ChrSeq, and a linkage map to transform 481 scaffolds (16) into a chromosome-scale genome assembly. (B) Chromosomes from mitotic cells were spread onto membrane slides for laser capture and sequencing. Numbers denote the identity of chromosomes present in representative samples. (C) Contact heatmap of chromatin interactions visualized in Juicebox (23); black boxes denote the four chromosomes in the final assembly. (D) Linkage map visualized by LinkageMapView (62) shows lack of recombination on chromosome 1. Tick marks and labels denote genetic markers.

Figure 2 Recombination suppression and inversions on chromosome 1 (A) Distribution of heterozygous variants maintained in the S2 inbreeding pedigree along the four chromosomes; y-axis shows variant counts per 10 Mb. (B) Number of meiotic recombination events on chromosomes 1 and 3 in oocytes (red) and sperm (cyan). Dots represent individual gametes, summarized with violin plots. (C) Chromosome 1 in late prophase I in the oocytes from J/V worms (left and right panels) have fewer crossovers than those in oocytes from J/J worms (middle panel). Chromosomes in the right panel show FISH results with probes for telomeric repeats (TTAGGG, magenta) and centromere-like repeats (yellow). (D) Chromatin contact heatmap for chromosome 1. Dashed-line rectangles indicate locations of potential inversions. Arrows show chromatin contact signals with inversion regions from unassigned scaffolds.

Figure 3 Island-specific evolution of chromosome 1 heterozygosity (A) Collection sites of 72 wild isolates from Mediterranean islands and Tunisia. (B) Maximum likelihood

tree of 2 asexual isolates from Menorca (MEN_GOR) and 70 sexual isolates from Sardinia (SAR), Corsica (COR), Sicily (SIC) and Tunisia (TUN). Dots denote individual worms, with colors corresponding to their collection sites as noted in the legend. (C) Observed vs. expected heterozygosity for each chromosome (denoted by color) in each population (denoted by shape). Dashed line corresponds to equality between observed values and those expected under Hardy-Weinberg equilibrium and is shown as a guide to the eye; the departure of points corresponding to chromosome 1 in COR and SAR from this expectation is notable. (D-E) Number of sites heterozygous in >80% of the population per 10 kb window along chromosome 1 (left) and chromosome 3 (right) in Sardinia (D) and in Sicily (E).

Figure 4 Acquisition and haplotype-specific expression of sex-related genes

(A) A putative model of the structures of the J and V haplotypes of chromosome 1, with locations of critical genes for sexual development (cpeb-1 is female-specific). Arrows denote putative inversions. (B) Comparison of read counts for the J and V alleles of key genes in the transcriptomes of whole sexually mature worms and penis papillae (pp). Bonferroni-corrected P-values from a binomial test of equal expression are shown.

Supplemental Figure 1 Chromosome assignment by ChrSeq (A-C) Chromosome sequencing depth of different scaffolds are consistent with linkage group assignment. Linkage groups: lg1, lg11, lg17, lg4, lg7. (D-F) *de novo* assignment of 481 scaffolds to

chromosomes by sequencing depth. Chromosomes: chromosome 1 (A, D); chromosome 2 (B, E), chromosome 3 (C, F).

Supplemental Figure 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.

Supplemental Figure 3 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 populations (13). SAR: Sardinia; COR: Corsica; SIC: Sicily; TUN: Tunisia.

Supplemental Figure 4 STRUCTURE analysis of wild isolates.

Evanno analysis supports a K=2 split, dividing geographically the samples in an Eastern and a Western group. With K=3, a Tunisian signal was found to be present in all other 3 groups (light blue).

Supplemental Figure 5 Genes enriched in the sexual reproductive system were distributed equally on all 4 chromosomes

(A) Number of genes and total length of

each chromosome. Genes were based on dd_Smed_v6 annotation. (B) Differentially expressed genes and their localizations on chromosome 1 and the rest of the chromosomes. chr234/chr1: the ratio of genes on chr2, 3, and 4 to chr1. (C) Differentially expressed genes and their localizations on chromosome 1, 2, 3, and 4. Proportion: number of differentially expressed genes to the total number of genes per chromosome. (B-C) juvenile: comparing public transcriptome of juvenile animals to adult animals (30, 34, 48); nhr1 RNAi: comparing public transcriptome of nhr-1 RNAi animals to adult animals (30); pp: comparing transcriptome of penis papillae to whole worm, generated in current study; neoblast: comparing public transcriptome of sorted X1 cells to whole worm in CIW4 animals (49, 50).

Supplemental Figure 6 *Nanos*, *nhr-1* and *CPEB-2* transcribed from both J- and V-haplotypes The sequencing reads from genomes (J haplotype, V haplotype, S2) and transcriptomes (pp, ww) were aligned to Smed_chr_ref_v1. Shown above are reads for the gene CPEB-2. Three variants with J and V alleles are marked by colored bars. In the genome of the adult worm S2, both J and V alleles are present at equal ratios. In the genome of oocytes (J and V haplotype), the variants have only one allele. In the transcriptomes of penis papillae (pp) and whole worm (ww), both J and V alleles are present at equal ratios.

Table S1 Assigning 481 scaffolds to chromosomes with 17 sequenced chromosomes

Chromosome sample information (i.e., ploidy, identity and potential contamination), sequencing coverage and the chromosome assignment for each of the 481 scaffolds from the assembly dd_Smes_g4. In chromosome assignment, 0 (scaffolds that could not be assigned to chromosomes by chrseq); 5 (scaffolds that were assigned to two chromosomes by chrseq); 1-4 (chromosomes).

Table S2 Linking 481 dd_Smes_g4 scaffolds into chromosomes with Hi-C chromatin contact sequencing

dd_Smes_g4 scaffolds that were assigned or unassigned to Smed_chr_ref_v1 (chrAssembly), their identity in SALSA assembly of raw Hi-C sequencing data (salsa_scaffolds), and their assignment to chromosomes by chromosome sequencing (chrseq). In chromosome assignment, 0 (scaffolds that could not be assigned to chromosomes by chrseq); 5 (scaffolds that were assigned to two chromosomes by chrseq); 1-4 (chromosomes).

Table S3 Summary of the final chromosome scale genome assembly (Smed_chr_ref_v1)

Hi-C detected 26 assembly errors in dd_Smes_g4, 5 of which were confirmed by chromosome sequencing to be inter-chromosomal mis-joining. A total of 97 dd_Smes_g4 scaffolds were not assigned to Smed_chr_ref_v1.

Table S4 Genetic markers and distances in the linkage map established from a F2 mapping population

Linkage groups L.3, L.6 and L.8 only contained one genetic marker, which were not included in the table. Chromosomes 2 and 4 were split into two linkage groups each, likely due to the small size of the F2 mapping population.

Table S5 Distribution of Genetic variants that are maintained heterozygous in the inbreeding pedigree

Chromosomes were divided into 10Mb windows. Heterozygous variants were identified from the inbreeding pedigree from S2 to S2F9b.

Table S6 Chromosome 1 was heterozygous in all samples of a F2 population

The genetic markers were heterozygous in the J/V parent (parent_S2F10B_2A and parent_S2F10B_2B) and homozygous in the J/J parent (parent_D5-1). parent_S2F10B_2A and parent_S2F10B_2B are clones. Genotyping data from RADseq of 291 F2 samples were listed. The 291 F2 samples correspond to 93 unique segregants. 0/0: homozygous reference allele; 1/1: homozygous alternative allele; 0/1: heterozygous; ./.: missing data.

Table S7 Recombination in the gametes

Distribution of heterozygous variants identified in the J/V line S2 and used in the crossover assessment along chromosome 1 and 3, in 20Mb windows. Number of crossovers identified per gamete on chromosome 1 and 3.

Table S8 Chromosome 1 genotypes of wild isolates in Sardinia and Sicily

Genotyping data for 6 different collection sites in Sardinia and 2 different collection sites in Sicily were aggregated. Row 1 are names of individual animals. 0/0: homozygous reference allele; 1/1: homozygous alternative allele; 0/1: heterozygous; ./.: missing data.

Table S9 Allele specific expression of “male” and “female” genes

Genes with well characterized sex-related functions were identified from published work. Their locations on Smed_chr_ref_v1 and *Schistosoma mansoni* chromosomes were determined. The expressed J or V alleles for each of the 9 ‘male’ or ‘female’ genes were quantified in the transcriptomes of whole worms and isolated male copulatory organ, penis papillae.

Fig1

Fig2

A

B

C

D

Fig3

A

B

C

D

E

Fig4

B

Accession	J_worm	V_worm	J_proportion	P value (worm)	J_pp	V_pp	J_proportion	P value (pp)
KY847529.1	242	146	0.62	2.26E-05	295	297	0.50	1
KY847536.1	2451	2142	0.53	9.82E-05	42	29	0.59	1
Smed-MSY4	5398	5802	0.48	0.00251	60	81	0.43	1
KY847538.3	448	353	0.56	0.0159	6330	1926	0.77	9E-323
KY847532.1	135	145	0.48	1	2	18	0.10	0.00724
KY847539.1	142	168	0.46	1	446	341	0.57	0.0037
KY847534.1	11	9	0.55	1	0	0	NA	NA
Smed-elav2	182	155	0.54	1	0	0	NA	NA
Smed-bic-C	278	257	0.52	1	0	0	NA	NA

SupFig1

Chr1_b1_7a

Chr2_b1-4e

Chr3_b1-4b

bioRxiv preprint doi: <https://doi.org/10.1101/2021.05.24.445457>; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SupFig2

A

B

SupFig3

A

B

bioRxiv preprint doi: <https://doi.org/10.1101/2021.05.24.445457>; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SupFig4

SupFig5

A

	genes	length (bp)
chr1	12018	332698990
chr2	10296	265042666
chr3	5470	118597112
chr4	2932	47107876

B

	chr1	chr234	chr234/chr1
juvenile	1101	1763	1.601271571
neoblast	2261	3440	1.521450686
nhr1RNAi	844	1387	1.643364929

C

	chr1		chr2		chr3		chr4	
	number	proportion	number	proportion	number	proportion	number	proportion
juvenile	1101	0.092	990	0.096	517	0.095	256	0.087
nhr1RNAi	844	0.070	783	0.076	406	0.074	198	0.068
pp	240	0.020	208	0.020	100	0.018	66	0.023
neoblast	2261	0.188	1880	0.183	1036	0.189	524	0.179

SupFig6