Open access • Posted Content • DOI:10.5281/ZENODO.5149023

Assembled chromosomes of the blood fluke Schistosoma mansoni provide insight into the evolution of its ZW sex-determination system — Source link

Sarah K. Buddenborg, Sarah K. Buddenborg, Alan Tracey, Duncan Berger ...+12 more authors

Institutions: Morgridge Institute for Research, Wellcome Trust Sanger Institute, University of Liverpool

Published on: 10 Jul 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Pseudoautosomal region, ZW sex-determination system, Pseudogene, Genome and Synteny

Assembled chromosomes of the blood fluke *Schistosoma mansoni* provide insight into the evolution of its ZW sex-determination system

Sarah K Buddenborg[1,2]*, Alan Tracey[1], Duncan J Berger[1], Zhigang Lu[1], Stephen R Doyle[1], Beiyuan Fu[1], Fengtang Yang[1], Adam J Reid[1], Faye H Rodgers[1], Gabriel Rinaldi[1], Geetha Sankaranarayanan[1], Ulrike Böhme[1,3], Nancy Holroyd[1], Matthew Berriman[1]*

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

[2] Morgridge Institute for Research, Howard Hughes Medical Institute, University of Wisconsin-Madison, WI, USA

[3] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

* Corresponding authors

ABSTRACT

Background

Schistosoma mansoni is a flatworm that causes a neglected tropical disease affecting millions worldwide. Most flatworms are hermaphrodites but schistosomes have genotypically determined male (ZZ) and female (ZW) sexes. Sex is essential for pathology and transmission, however, the molecular determinants of sex remain unknown and is limited by poorly resolved sex chromosomes in previous genome assemblies.

Results

We assembled the 391.4 Mb *S. mansoni* genome into individual, single-scaffold chromosomes, including Z and W. Manual curation resulted in a vastly improved gene annotation, resolved gene and repeat arrays, trans-splicing, and almost all UTRs. The sex chromosomes each comprise pseudoautosomal regions and single sex-specific regions. The Z-specific region contains 932 genes, but on W all but 29 of these genes have been lost and the presence of five pseudogenes indicates that degeneration of W is ongoing. Synteny analysis reveals an ancient chromosomal fusion corresponding to the oldest part of Z, where only a single gene—encoding the large subunit of pre-mRNA splicing factor U2AF—has retained an intact copy on W. The sex-specific copies of U2AF have divergent N-termini and show sex-biased gene expression.

Conclusion

Our assembly with fully resolved chromosomes provides evidence of an evolutionary path taken to create the Z and W sex chromosomes of schistosomes. Sex-linked divergence of the single U2AF gene, which has been present in the sex-specific regions longer than any other extant gene with distinct male and female specific copies and expression, may have been a pivotal step in the evolution of gonorchorism and genotypic sex determination of schistosomes.

KEYWORDS

sex chromosomes, schistosomiasis, centromere, gametologues, sex determination, gonochorism, sex chromosome evolution, spliced-leader trans-splicing, gene clusters

BACKGROUND

Schistosoma mansoni is one of three main schistosome species that causes schistosomiasis, a neglected tropical disease that affects ~240 million people worldwide [1]. Within the Phylum Platyhelminthes (flatworms), schistosomes are remarkable; while virtually all other flatworm families are hermaphrodites, family schistosomatidae are gonochoristic (separate sexes) and sexually dimorphic as adults. Sex is genetically determined with heterogametic females (2n=16, ZW) and homogametic males (2n=16, ZZ).

Adult female worms reside within the gynecophoric canal of adult males and the paired worms produce several hundred eggs a day. The eggs either traverse the intestinal wall to reach the lumen and be excreted in faeces or become trapped in host tissues, mainly liver and intestine, driving the pathology associated with schistosomiasis [2]. It has been postulated [3,4] that dimorphism and gonochorism in schistosomes is an evolutionary adaptation to their residence in the venous system, close to capillary beds of warm-blooded host species; a division of labor between the sexes enables both a muscular male body to move against the blood flow of large veins and a thin slender female body shape to deposit eggs in small venules, allowing their efficient exit. However, the adaptions required to develop this dimorphism are unclear, limited by a lack of understanding of sex-linked molecular mechanisms, including unresolved sex chromosomes.

Despite major advances in the quality and quantity of published genome assemblies, sex chromosomes that are limited to the heterogametic sex (W and Y) are underrepresented in the growing list of whole genome assemblies. These sex-specific chromosomes are usually present at a lower copy number than autosomes, and the problem of assembling them is compounded by difficult to resolve highly repetitive sequences and by genetic divergence between the sex chromosomes, such that they can vary along their lengths [5]. There are exceptions—notably the recent publication of the eel genome [6] included resolved centromeres, subtelomeric sequences and the highly repetitiive Y chromosome short arm containing no gaps—but other sex chromosome assemblies, such as the *Drosophila* Y chromosome [7] and *Gallus gallus* W chromosome [8], are in fragmented states and even the reference human Y chromosome assembly [9] lacks continuity between the heterochromatic and euchromatic regions.

Degeneration of sex-limited chromosomes (W or Y) often distinguishes them from the shared (Z or X) chromosomes. Along the W chromosome of schistosomes, extensive heterochromatinization and the accumulation of satellite repeats, has been described, including a large satellite repeat SM-ɑlpha [10]. Extensive gene loss, or pseudogene-formation is also expected but without an adequate W assembly, it has not previously been possible to comprehensively describe the W-specific gene and repeat content that may play an important role in sex determination.

The *S. mansoni* genome was first published as a draft assembly [11], followed by a more contiguous version (v5) three years later [12] that took advantage of high throughput short-read sequencing technology. At that stage, as much as 80% of the genome had been assigned to chromosomes but gaps were prolific and large regions remained unresolved. The Z and W sequences were assembled together into merged scaffolds, with multiple Z-specific sequences and almost no resolution of W-specific sequences. As part of a sustained commitment to produce a complete genome sequence, in the present study, we have significantly improved upon previous efforts using a combination of long-read sequencing technology, optical mapping and manual curation to generate a highly contiguous chromosome-scale assembly that includes a fully assembled Z chromosome and a contiguous representation of the highly repetitive W chromosome. Our fully resolved reference genome is a key pre-requisite for understanding the evolution of sexual dimorphism in schistosomes and exposes sex-linked protein-coding and non-coding genes tentatively involved in sex determination.

RESULTS

The chromosome-level genome of *Schistosoma mansoni*

Using a combination of PacBio long-read and Illumina short-read sequencing, optical mapping, fluorescent *in situ* hybridization (FISH), Hi-C, and manual curation, we have assembled complete chromosomes from the 391.4 Mb genome of *S. mansoni*, including resolution of its Z and W sex chromosomes. The assemblies of chromosomes 2, 5, 6 and 7 comprise single scaffolds with telomeric repeats at either end; the remaining 5 chromosomes are also single scaffolds with a telomere at one end and sub-telomeric sequence at the other (Figure 1a,b). The number of gaps has decreased by 96% from 8,640 in the previous assembly to just 356 (Table 1).

Table 1.

	v5	**v9**
Assembly size (Mb)	364.5	391.4
Gaps	8,640	356
Repeat Content (Mb)	191.8	213.2
Scaffolds		
Number	885	9
N50 (Mb)	32.1	52.8
N90 (Mb)	0.547	25
Largest (Mb)	65.5	89.1
Gene statistics		
Protein-coding genes	10,116	9,794
Novel genes*	-	810
Deleted genes	-	867
Pfam annotated	66.9%	70.3%

Transcript statistics		
Transcripts	11,075	14,031
Alternative splicing	6.9%	27.9%
Average exons per transcript	5.9	7.9

The total repeat content of the assembly is 213.2 Mb (Table S1), a 21.4 Mb increase compared with the previously published version [12], reflecting the ability of PacBio long-read sequencing to account for repetitive regions that were previously difficult to assemble. For instance, an array of rRNA genes known as the nucleolar organizer region (NOR) of chromosome 3 (Figure 1) was highly collapsed in the earlier assembly and is now fully resolved. Newly resolved repetitive regions also include arrays of tandemly duplicated protein-coding genes enabling us to obtain a more accurate count for genes previously thought to be present as single copies. Two striking examples are the major egg antigens IPSE (IL-4-inducing principle of *S. mansoni* egg) and omega-1. These genes, specifically expressed in the eggs, have been intensely studied due to their roles in immune-modulation, pathogenesis and mechanisms of egg translocation to the intestinal lumen [13–16]. IPSE and omega-1 transcripts are encoded by paralogous gene arrays of at least 13 and 7 gene copies, respectively. In fact, based on the depth of coverage of aligned sequencing reads, these numbers are likely to be even higher and may contain as many as 20 and 14 copies of IPSE and omega-1, respectively (Figure S2).

We extended the analysis to identify other clusters of genes with conserved functions. Across the genome, there are 44 clusters of genes sharing similar predicted functions based on their protein (Pfam) domains, more than twice the number of clusters and domain types as seen in the previous v5 published genome version (Table S2). Clusters of *S. mansoni* Kunitz protease inhibitors and elastases are striking. Eleven Kunitz protease inhibitors (PF00014) exist in a cluster and 25 copies of elastase (PF00089; trypsin) are found across two clusters. The well-studied SmKI-1 (Smp_147730 in v5; Smp_311660, Smp_311670, and Smp_337730 in v9), is known to be involved in defense mechanisms of *S. mansoni* within the mammalian host [17]. The elastases are an expanded group of serine proteases originally noted for their role in host skin penetration, but are also expressed in intra-molluscan stages, where they likely facilitate movement of the parasite through snail tissue [18,19].

Annotation improvements through manual curation

We have significantly improved upon previous gene annotations of the *S. mansoni* genome. Using Augustus [20] and extensive RNA-seq evidence (Table S3) for gene prediction, followed by extensive manual curation, the total number of genes has decreased from 10,116 to 9,794 (excluding genes on scaffolds that correspond to alternative haplotypes; Table S4), compared to the v5 genome. This is the lowest number of genes for any sequenced platyhelminth; for instance, the cestodes *Echinococcus multilocularis* [21] and *Hymenolepis microstoma* [22] have 10,663 and 10,139 genes, respectively. In spite of the modest net reduction in genes, a total of 3,610 updates to gene models from v5 to v9 have been made, including 810 new, 867 deleted, 344 merged, 189 multiple copies, 190 split, and 1,210 with large structural changes (defined as >20% of coding region affected; Figure S3; Tables S5-S6). Using BUSCO v3.0.2 [23], the *S. mansoni* protein set was estimated to be 95.3% complete based on the representation of eukaryota orthologs (full genome-level BUSCO results at Table S7).

Spliced leader (SL) trans-splicing is an mRNA maturation process where an independently transcribed SL exon is transferred to a pre-mRNA. SL sequences originate from SL genes 613 bp in length, consisting of a 36 bp exon sequence (position 144–181 bp) flanked by an upstream precursor sequence (1–143 bp) and a downstream intron (182–613 bp) (Figure S1). A ~1 Mb tandem array containing 41 full-length spliced-leader (SL) RNA genes has been resolved on chromosome 6 (Figure S1), together with an additional 109 partial gene sequences that contain the exon sequence only in the same array. On most other chromosomes, 1–4 SL gene fragments containing the exon sequence can also be found. Using RNA-seq data from all life cycle stages with an improved gene set (Table S3), we located SL receptor sequences in the transcripts of 6,641 genes in the primary assembly (i.e. no haplotypes), indicating that the majority of genes (66.3%) encode at least one trans-spliced isoform compared to 6.9% reported in the previous assembly (Table S8). This number is similar to the nematode *Caenorhabditis elegans* where ~70% are identified as being trans-spliced [24].

The complexity of gene structures has increased substantially; the average number of exons per gene has increased from 5.9 to 7.9 (Table 1) and 97.7% of transcripts have both 5' and 3' untranslated regions (UTRs) annotated (Table S9). Further, the proportion of genes with alternative splicing to generate distinct transcribed isoforms has increased from 6.9% to 27.9%. Systematic improvements to gene finding and gene structural changes have enabled a richer set of putative functions to be ascribed to the *S. mansoni* proteome, reflected in the 47 new protein (Pfam) domains to *S. mansoni* from new genes and 79 new Pfams domains annotated in genes with improved gene structure (Table S10).

Centromere motif conservation and divergence

S. mansoni chromosomes are monocentric [25], each with a cytologically distinguishable primary constriction (Figure 1a). The centromeric sequences are large repeat arrays that, on all chromosomes except 4 and Z, are highly conserved within a centromeric array and are between 93.1–98.5% similar to a 123 bp centromeric repeat proposed by Melters et al. [26] (Figure S4; Table S11). Between the centromeres of different chromosomes, the sequence conservation is more variable: 56% identity between the two most divergent centromere monomers (chromosomes Z and 6) and 100% identity between the centromeres of chromosomes 2 and 3 (Figure S4). The unit size is typical of the centromeric repeats of many other species [26], including the platyhelminth *Hymenolepis microstoma* [22]. The centromeric repeats for chromosomes 4 and Z have diverged from each other and from those of other chromosomes (Figure S4); their respective repeat units are 107 and 175 bp, and they are only 82 and 59% identical to the consensus from Melters et al. Centromeric repeats were previously estimated to comprise 0.48% of the genome (1.9 Mb) [26], but after including the divergent centromeres and estimating the degree to which all centromeric repeats were under-represented in the assembly based on mapped sequence coverage (from three PCR-free

Illumina libraries), we estimate that centromeres make up at least 1.15% (4.5 Mb) of the genome.

Architecture of the Z chromosome

Our new assembly includes a full-length 88 Mb Z chromosome that includes defined, recombining pseudoautosomal regions 1 (10.7 Mb) and 2 (42.9 Mb) and a non-recombining 33.1 Mb Z-specific region. In contrast to the previously published v5 assembly [12], where the Z chromosome was only partially resolved, the new sequence is 27.2 Mb larger with misassemblies corrected along its length, aided by the new long-range information that has been incorporated (Figure 2**)**. In particular, the sequence that is unique to the Z chromosome (i.e., the Z-specific region, or ZSR), is clearly visible based on the lower depth of coverage of resequencing reads mapped from heterogametic females. The ZSR is flanked by two regions that are common to both sex chromosomes, termed pseudoautosomal region (PAR) 1 and 2**.**

Based on the earlier assembly (v5), it was previously shown [27] that the Z chromosome comprises different sub-regions or strata that have evolved differentially in the African and Asian Schistosoma lineages from a common 'Ancestral' stratum that is common to both lineages. Using the v9 assembly as a reference, where the ZSR is now resolved as a 33.2 Mb continuous sequence (Figure 2; Table S12), we plotted coverage of mapped sequencing reads across Z chromosome orthologs from four schistosome species (*S. mansoni*, *S. rodhaini, S. haematobium, S. japonicum*) and the hermaphroditic trematode *Echinostoma caproni.* In contrast to the relatively uniform mapped coverage for *E. caproni*, the ZSRs for the *Schistosoma* species are clearly visible, with a 19.1 Mb Ancestral shared region (ZSR2; ZSR coordinates 13,993,393-33,063,208) that has extended more recently in different directions amongst the African (*S. haematobium*, *S. rodhaini*, *S. mansoni)* and Asian species *(S. japonicum).* It also appears that in the Asian *S. japonicum*, two inversions have resulted in orthologues changing position and, therefore, creating coverage anomalies near the ZSR boundaries. The more recent 14 Mb African stratum (ZSR1; ZSR coordinates 1 - 13,993,392) extends beyond the centromere but is shorter than the Ancestral stratum (ZSR2). In contrast to the single, contiguous Z-specific region in the v9 assembly, the v5 assembly contained two blocks of what we now know to be PAR fragments which were incorrectly located inside the sex-specific region. It was previously reported that blocks of sequence shared by Z and W are located in the large region of recombination repression (i.e. the ZSR) [28]; based on this observation, Hirai, Hirai, and LoVerde [29] proposed three inversions in homologous Z/W regions from Z to W occurred before heterochromatinization, followed by at least one more inversion. These conclusions do not hold true in v9 and can now be attributed to misassemblies in v5.

To gain further insight into the evolutionary origins of the ZSR, we looked at the relationship between the Z chromosome and the chromosomal sequences of distantly related tapeworms. We have previously shown that flatworm genome structure can be defined based on conserved chromosome synteny blocks [30] (Figure 3b). When orthologs of *S. mansoni* and tapeworms are compared, synteny is largely preserved between these blocks, even though collinearity is disrupted. It is evident that one end of the Z chromosome is highly related to chromosome 3 of *Echinococcus multilocularis* and the other end is highly related to chromosome 5. When taken in isolation, the orthology evidence equally supports an ancient fusion in the schistosome lineage or an ancient fission in the tapeworm lineage. However, the position of the junction between the chromosome synteny blocks coincides with the position of the Ancient stratum (Figure 3a), suggesting that a fusion in the schistosome lineage is likely to have played a role, resulting in suppressed recombination.

For neutral positions in the genome, the genetic diversity present is expected to reflect the number of copies of that region in the genome [31]. For the ZSR, the relative number of copies is 0.75 relative to autosomes (1.0), thus the diversity is expected to be lower than that of autosomes. Along the ZSR, we identified 352 genes in the African stratum and 580 in the Ancient stratum, which are flanked by 229 and 1,071 protein-coding genes in PAR1 and PAR2, respectively. We calculated the median nucleotide diversity (π) across the protein-coding genes of the autosomes and PARs and Z-specific regions (Figure 4; Table S13) using published genome variation data [32]. Across 50 kb windows, the autosomes have a median π range of 0.0026 to 0.0039. The PARs have a similar median π range to that of the autosomes at 0.0027 to 0.0032 in females and 0.0027 to 0.0034 in males suggesting that recombination between ZW and ZZ bivalents in the PARs is similar to that of the autosomal chromosomes. Also, the median π of the ZSR is significantly lower than that of the PARs for both males and females ($p<0.001$; Mann-Whitney test). We observed significantly lower π values in the Z African stratum when compared to the Z Ancestral stratum in both male and female samples ($p<0.001$; Mann-Whitney test), consistent with the effective population size of the Ancestral stratum being smaller for longer. The π values of the Z chromosome are close to that which would be expected in a neutral equilibrium with equal and constant male and female populations sizes ($\pi_Z/\pi_{Autosomes}$=0.75; [31] with $\pi_Z/\pi_{Autosomes}$=0.71 in males and $\pi_Z/\pi_{Autosomes}$=0.70 in females.

Assembling the W Chromosome

The W chromosome shares >50 Mb of sequence with the Z chromosome in the pseudoautosomal regions, PAR1 and PAR2, that flank a highly repetitive W-specific region (WSR) (Figure 5; Table S12). In the v5 assembly, the highly repetitive W-specific region could not be resolved beyond ~100 small and unordered contigs (1.1 Mb); by sequencing clonal females on multiple sequencing platforms, we resolved 22 repeat-rich W-specific scaffolds totalling 3.7 Mb (Figure S5). In many cases, the long reads used in our assembly were insufficient to fully span the arrays of repeats in the W chromosome. As a result, unique sequences are represented but the number of repeat units in many of the repeat arrays is vastly underestimated. After manual curation of the major repeat blocks, the W-specific assembly scaffolds were further ordered, oriented and linked by identifying as few as one, long PacBio subreads that spanned two consecutive blocks (Table S14). Metaphase FISH was also used to localize and orient three W-specific scaffolds that could not be placed through computational assembly methods (Figure S5).

Previous karyotype measurements from 22 female metaphase cells [33] showed the W chromosome to be approximately 14% longer than the Z chromosome, a figure we confirmed with our own measurements of 14.7% using 6 female metaphase cells (Table S15; Figure S6). In particular, a long repetitive region in the short (p) arm of the W chromosome accounts for much of this size difference and is responsible for the p-arm being ~40% of the W-chromosome length. Assuming a uniform density along the chromosome, relative measured lengths of other chromosomes with known assembly sizes (Figure S6), and genomic coverage of W-specific repeats (Table S16), we estimate the size of the W-specific region (WSR) to be ~46 Mb. However, given that this region is heterochromatic and, therefore, more densely packed, its true size could be much longer. We attempted to estimate the degree to which repetitive regions remain collapsed within the assembly by mapping high-coverage Illumina sequencing reads from adult females. Extrapolating the read depth across repetitive regions (Table S16; see next section for results on W repeats) and comparing it with the median coverage for the genome (Table S17; ERS039722), we estimate a length of 17.6 Mb for the W-specific region. Clearly the mapping approach is inaccurate for estimating the true size of these collapsed regions. In fact, there are many regions of repetitive sequence in W where very few Illumina reads are mapped, indicating that certain repeat motifs are underrepresented in the sequence data. So-called “dark” and “camouflaged” regions of genomes have previously been reported, where specific sequencing technologies perform poorly (e.g. short tandem repeats, duplicated regions, regions with high GC content, non-random fragmentation) [34,35].

Repeat classification and heterochromatinization of the W chromosome

Like the human Y chromosome, the *S. mansoni* W chromosome is largely heterochromatic with a large proportion of its length composed of satellite repeats. There are just three bands of euchromatin on the W chromosome (chevrons in Figure 5) [10,33]. Because some individual PacBio reads contained tandem arrays of the same repeat unit, we were able to assemble complete repeat units. Within the WSR constitutive heterochromatin, we characterized 36 unique repeats, named smw01-smw36 (Figure 5; Table S16). The 36 W-specific repeats comprise >95% of the assembled length of the W-specific region.

Of the 36 repeats, five (smw07, smw20, smw21, smw25, smw29) are related to the previously described 337 bp retrotransposable element SMαt-2 [36,37]. Although a variant of SMαt-2 has been previously published as female-specific (SMAlphafem-1; NCBI accession U12442), we found one complete copy (coordinates: 23,37,004–23,936,670; 92.3% identity, 99.7% coverage, e-value 9.07e-133) and 38 partial copies (>75.0% identity; >95.0% coverage) on the Z chromosome. Metaphase FISH has shown striking fluorescence of a SMαt-2-related probe hybridizing near the short arm euchromatic gap [33,37]. However, across the v9 genome, we found SMαt-2 repeats sporadically distributed on all autosomes and both sex chromosomes [38], but only as a large tandem array on the W chromosome, corresponding to the smw07 repeat found near the euchromatic band of the short arm [33].

Interestingly, 21 of the repeats can be grouped into five distinct families, where members within each family share at least 75.0% nucleotide identity, suggesting they may have evolved from a common ancestor including an SMα(aka SM-alpha and SMAlpha-fem) retrotransposon repeat family (smw03, 07, 20, 21, 29) (Table S16).

Gametologues and their possible role in schistosome sex determination

The ZSR contains a total of 932 protein-coding genes. Of these, only 33 have clear homologous copies (termed gametologues) on the W chromosome, all within the WSR (Table S18). Although there is some positional clustering, extensive rearrangements by inversions, repeat expansions and transposable elements have largely disrupted collinearity between the WSR and the ZSR. The more recent African stratum contains 31 of the gametologues. For two of these, the corresponding W-copies have duplicated; there are three copies of genes encoding DnaJ domain proteins (heat shock protein 40 member B6) and two copies encoding a hypothetical protein with no discernible conserved features. At least five of the gametologues in the African stratum have degenerated into pseudogenes on W that have not yet been lost.

Considering the longest transcript for each gene, the W gametologues have an average of 55 amino acids less per protein sequence than the Z gametologues (Table S19). Only three W gametologues (spliceosome-associated protein, Smp_310950; ENTH domain-containing protein, Smp_303540; splicing factor U2AF 35 kDa small subunit, Smp_348830) are longer than their Z counterparts. Most Z and W gametologues are highly similar with average amino acid identities of >80% across their entire lengths using the Needle Wunsch algorithm in the EMBOSS package [39]. Excluding the five W pseudogenes and their Z gametologues, the gametologue pair with the greatest divergence was Smp_348820 on W and Smp_031310 on Z (encoding 40S ribosomal subunit S26) with only 47.6% identity. However, as with other low-similarity pairs, it was not possible, even through manual curation, to rule out gene finding inaccuracies due to a lack of isoform-specific transcript data.

We used previously published sex- and stage-specific RNA-seq [40,41] to analyse differences in expression between the Z and W gametologue pairs (Figure 6). As expected, using unique mapping reads only for analysis, very few male reads mapped to the W gametologues. There were slight differences in the levels of expression between male and female samples for the Z gametologues, although RNA-seq coverage and replicate number from some of these samples were inadequate to enable robust analysis and interpretation. It has been shown one gametologue pair, encoding DnaJ homolog subfamily B member 4, have diel expression in males and females with the Z gametologue (Smp_336770) with the the Z gametologue cycling in adult females, males, and male heads, and the W gametologue (Smp_020920) cycling in females [42]. Expression of several W gametologues in female samples indicates possible stage-specific activity (such as Smp_317860, DnaJ heat shock protein family member B6) that is expressed in female larval cercariae and pre-dimorphic mammalian-stage schistosomula but not in adults; however, the Z gametologue to this gene, Smp_022330, shows consistent expression values across all stages.

There is an almost complete lack of gametologues in the Ancestral stratum, which is consistent with this part of the chromosome having become sexually differentiated earlier and degenerative processes thus having been underway for longer. Within this long tract of degenerated sequence, two gametologues are clear exceptions. The first of these is a long multi-exon gene on Z, encoding a protein with ankyrin repeats and helicase domains. The corresponding gametologue on W is a pseudogene with several frameshifts and missing exons (Figure S7a). The second gametologue is predicted to encode the large subunit of splicing factor U2 snRNP auxiliary factor (Smp_019690 on Z and Smp_348790 on W). Strikingly, the sequences are almost identical (>95%) for most of their lengths but have divergent N-terminal sequences. After correcting for an artifactual frameshift in the W chromosome consensus sequence (based on aligned RNA-seq reads; Figure S7b), the copy on W shares the single-exon structure but the first 125 aa share only 45% identity.

DISCUSSION

Our chromosome-scale assembly and curated annotation significantly extends the genetic resources for *S. mansoni*, and provides a more robust scaffold for genome-wide and functional genomic approaches for this important but neglected pathogen. It has enabled a greatly improved definition of the gene content, with the sequences of more than 25% of genes changed with >20% of coding region affected, and better resolution of those present in repetitive arrays, such as those encoding spliced leader RNA and stage-specific gene families. Amongst the gene families, many are known to encode highly abundant products —such as IPSE, omega-1, elastases, Kunitz protease inhibitors—that are important in host-parasite interactions. Major egg antigens omega-1 and IPSE are associated with a Th2 immune response in the host resulting in granulomatous inflammation around trapped parasite eggs

[43]. Given the critical role of the intestinal granuloma for the egg translocation from the blood vessels to the intestinal lumen [44], genome expansions of these genes might have represented a selective advantage.

A major advance is in the analysis of schistosome sex chromosome evolution. Our previous analysis of orthologue synteny across the flatworms showed that the *S. mansoni* Z chromosome corresponds to two or more chromosomes in tapeworms [30]. From those data alone, it was not possible to determine whether a chromosome fusion had occurred in the schistosome lineage or whether it was a fission in tapeworms. However, in several other taxa, including filarial nematodes and several lepidotera, a chromosomal fusion has underpinned the genesis of sex chromosomes [45,46]. We therefore speculate that a fusion has similarly occurred in the ancestral schistosome, creating a new pre-sex autosomal chromosome. The fusion event could have resulted in an isolated sex-determining locus that was advantageous to females and/or antagonistic to hermaphrodite worms. Consistent with this hypothesis, we show that the position of the putative fusion is within the oldest part of the Z-specific region of the chromosome and, within it, there is a single protein-coding ancestral gene (U2AF; splicing factor U2AF 65 kDa subunit) and a single pseudogene that are common to all African and Asian schistosomes. The alternative hypothesis to explain the observed synteny would require a fission at that position somewhere in the tapeworm lineage. This would have occurred prior to the formation of a sex determining region and the fission would, therefore, have played no role.

As one of two genomes found in the earliest-diverging part of the sex chromosomes, we identify the W gametologue encoding the pre-mRNA splicing factor U2AF 65 kDa subunit (Smp_348790) as a leading candidate gene for involvement in schistosome sex-determination. U2AF has been studied extensively in *Drosophila* for its association with the master sex-determining protein Sex-lethal (Sxl) [51] that is expressed exclusively in female flies. Sxl competes with U2AF binding to inhibit the splicing and translation of the *msl-2* gene (male-specific-lethal-2) [52,53]. Considering that sex is determined by inhibiton of U2AF binding to pre-mRNA in *Drosophila*, it is tempting to speculate that the *S. mansoni* female-specific W copy of U2AF may antagonise the activity of the Z copy to inhibit the splicing of one or more genes. Further implicating U2AF in sex determination, the sex-specific regions also contain a homolog of the U2AF 35kDa subunit. In many taxa, U2AF is a heterodimer composed of large and small subunits that are required for spliceosome assembly in order to remove intron sequences from pre-mRNAs. U2AF binds to the 3' splice site and polypyrimidine tract of introns in a complex with several other small nucleolar ribonucleoproteins (snRNPs) bound to the 5’ splice donor, commiting pre-mRNA to splicing (see review [50]). Our identification of U2AF2 is independently validated by Elkrewi et al. [49], who show using a search strategy based on the differential distribution of k-mers, that U2AF2 is the only intact gene in the ancient stratum of the ZSR.

How has sexual dimorphism evolved in schistosomatidae? The characterization of chromosomal fusions resulting the sex chromsomes, distinct evolutionary strata among closely related species, and the identification of U2AF allows us to propose a model of a model of the evolution of the schistosome sex chromosomes (Figure 8). At some point during the evolution of the Z and W sex chromosomes, the centromeric repeats diverged. It is not possible to know whether the centromere divergence occurred simply as a result of recombination or whether it played a more pivotal role in driving the suppression of recombination. Given the location of the centromere towards the far end of the more recent African stratum of the ZSR, the centromere divergence could have enabled a large expansion of the ZSR in the common ancestor of the African lineage of parasites. The high homology in amino acid sequence along with the conservation of functional domains between the gametologues suggests function has not changed between the gametologue pairs. Analysis of existing RNA-seq revealed sex- and stage-specific expression of the Z and W gametologues that could play a role in female-specific development. The duplication and triplication of two Z gametologues on W may be important in maintaining gene dosage or specialized female expression for those genes and is worthy of future study.

Although sexual dimorphism needs not rely on the existence of sex chromosomes and not all sexually dimorphic traits need to be linked to sex chromosomes [55], there must have been selective pressure to isolate sexually antagonistic and/or advantageous loci on non-recombining regions of sex chromosomes [56,57]. Unlike many species in which a master sex-determining gene triggers male or female development, the absence of a W chromosome-specific genes suggests that multiple sex-determining loci were isolated on the sex chromosomes to produce separate sexes. With this in mind, we hypothesize that the W-copy of U2AF is regulating other gametologues or even genes located on the autosomes to control the suppression of male or female function. Identifying downstream interactions of U2AF with other genes is a critical next step for uncovering the mechanisms involved in schistosome sex determination. For example, do posttranslational modifications or splicing of W gametologues by U2AF directly inhibit the activity of a male-promoting product or create a male-lethal product? Future studies are needed to understand the functional role the gametologues like U2AF play in schistosome sex biology.

CONCLUSIONS

S. mansoni is the most studied trematode and an accurate genome sequence underpins research into this important pathogen as well as enabling it to serve as a model for other trematodes. As the first species with completely assembled Z and W sex chromosomes, the *S. mansoni* genome provides a novel resource for studying other ZW organisms and is a crucial resource for future investigation into the sexual biology of schistosomes. The results presented provide a signfiicant advance toward understanding the evolution of sex chromosomes among the Schistosomatidae. As the agent of a prominent neglected tropical disease, understanding the evolutionary origins and molecular mechanism of sex determination in schistosomes may reveal new vulnerabilities to combat these parasites. The identification of the W-copy of U2AF as a candidate sex determining factor is clearly a major first step. This new assembly and annotation has already assisted in a broad range of studies on schistosomiasis including monitoring genetic diversity in field strains [32,58], the discovery of alleles under selection for resistance to the antihelmintic praziquantel [59], and the analysis of stage- and sex-specific epigenetic changes [60–62]. Future studies using this resource will undoubtedly continue to reveal novel biological insights into schistosome development, infection, host-parasite interactions, and pathogenicity.

METHODS

Parasite material

Schistosoma mansoni developmental stages

A summary of the parasite material for genome and transcriptome sequencing can be found in Table S17 and Table S3**,** respectively. Unless otherwise specified, the different *S. mansoni* developmental stages were collected following described protocols [63,64]. Unless otherwise noted, samples for RNA extraction were resuspended in 1 ml of TRIzol and stored at -80°C until a standard TRIzol RNA extraction method was performed. Genomic DNA was extracted using a standard phenol:chloroform DNA extraction method.

Sporocysts

Sporocysts were collected from Brazilian *B. glabrata* snails (BgBre) infected with 10 miracidia of their sympatric Brazilian *S. mansoni* (SmBre) strain. Secondary (daughter) sporocysts were dissected from 20 snails at 15 days and 4.5 weeks after infection. Following RNA extraction, DNA was removed with the Ambion® DNA-free™ Kit following the standard procedure and purified with the RNeasy® Mini Kit (QIAGEN).

Cercariae

At 4.5 weeks post exposure to 15-30 miracidia each, snails were washed, transferred to a beaker containing ~50 ml conditioned water, and placed under light to induce cercarial shedding. Cercariae were collected and water was replaced every 30 minutes for 2 hours. Cercariae were incubated on ice for 30 minutes and concentrated by centrifugation at 1500 x g for 30 minutes at 4°C.

Snails exposed to single miracidium each were tested for patent infection after 5 weeks by exposure to light to collect genomic DNA from pooled male and pooled female cercariae. Snails with patent infection were kept and exposed to light every three days. Cercariae collected from each snail were stored for DNA extraction. Sex of the cercariae was identified by PCR [65].

Schistosomula and adult worms

Briefly, water containing cercariae was filtered, cercariae were washed, and tails were sheared off by ~20 passes through a 22-G emulsifying needle. Schistosomula bodies were separated from the sheared tails by Percoll gradient centrifugation, washed, and cultured at 37°C under 5% CO_2.

Adult worms were collected by portal perfusion from experimentally-infected mice at 6, 13, 17, 21, 28 and 35 days post infection following methods previously described [66]. Clonal female or male adult worms were collected from mice infected with PCR-confirmed female or male cercariae, respectively, shed from single monomiracium-infected snails.

For RNA preparation, samples were thawed on ice and transferred to MagNA Lyser Green Beads (Roche Molecular Systems, Inc). The samples were homogenized using the FastPrep-24 instrument (MB Biomedicals, UK) for two 20 second pulses with a speed setting of 6. A standard TRIzol RNA extraction followed and RNA was concentrated using RNA Clean and Concentrator Kit (Zymo Research) according to the manufacturer's recommendations. RNA quality was assessed on the Bioanalyzer (Agilent) and samples with the highest quality were chosen for reverse transcription.

Miracidia

Livers were removed from hamsters 49 days post-infection with cercariae of the Liberian strain of *S. mansoni* and homogenised in PBS. The homogenate was centrifuged for 10 minutes at 5,500 x g at 4°C and the supernatant was discarded. The pellet was washed twice by resuspension in 0.9% NaCl followed by centrifugation as above. The pellet was resuspended in fresh conditioned water, exposed to light, and miracidia were collected. Miracidia were centrifuged for 30 minutes at 15,000 rpm at 4°C. Pelleted miracidia were resuspended in 100 µl TriFast (Peqlab) before storage at -80°C. The miracidia were allowed to thaw at room temperature before homogenisation with a polypropylene pestle, and snap frozen in liquid nitrogen. This was repeated twice more before TriFast was added to 500 µl. RNA was then extracted according to the manufacturer's instructions. Extracted RNA was quantified using a BioPhotometer plus (Eppendorf). RNA quality was assessed with the Bioanalyzer RNA 600 Pico Kit (Agilent).

Illumina and PacBio genome sequencing

Clonal male and female mate pair libraries (3 kb fragment size) were prepared from cercariae genomic DNA, following a modified SOLiD 5500 protocol adapted for Illumina sequencing [67]. Additionally, genomic DNA from clonal male and clonal female adult material was used to make separate PCR-free 400-550 bp Illumina libraries following previously described protocols [68], with the exception of using Agencourt AMPure XP beads for sample clean-up and size selection. Genomic DNA was precipitated onto beads after each enzymatic stage with an equal volume of 20% polyethylene glycol 6000 and 2.5 M sodium chloride solution. Beads were not separated from the sample throughout the process until after the adapter ligation stage. Fresh beads were then used for final size selection. Illumina libraries were sequenced on either a HiSeq 2000 or 2500 (Table S17).

Genomic DNA from *S. mansoni* clonal female adults was used to prepare a SMRTbell library following the Pacific Biosciences protocol ‘20 kb Template Preparation Using BluePippin Size-selection System’. The resulting library was used to produce 40 SMRT cells on the Pacific Biosciences RSII platform. We also prepared a PacBio library using genomic DNA from a pool of male cercariae from a snail monomiracidium-infection producing 28 SMRT cells on the Pacific Biosciences RSII platform (Table S17).

Optical mapping for genome assembly corrections and increased resolution

Female clonal cercariae were used to make agarose plugs using the CHEF Genomic DNA Plug Kit (Bio-Rad) following methods previously described [69]. High molecular weight *S. mansoni* genomic DNA was prepared by proteinase K lysis of trypsin-digested adults mixed with molten agarose set in plugs. DNA molecules were stretched and immobilized along microfluidic channels before digestion with the restriction endonucleases *BamHI* and *NheI*, yielding a set of ordered restriction fragments in the order that they occur within the genome.

The optical data was generated and analysed using the Argus Optical Mapping System from OpGen and associated MapManager and MapSolver software tools. As the *S. mansoni* genome is significantly larger than the 100 Mb cut-off suggested by OpGen for *de novo* assembly, OpGen’s GenomeBuilder software was used to generate targeted local optical map assemblies from the sequence contigs to provide additional mapping information. The median coverage of fluorescently-labelled molecules in the optical contigs from which consensus sequences were built was 30x. The raw data for each optical map contig were manually scrutinized using OpGen’s AssemblyViewer software, allowing us to validate accuracy (i.e. consistent coverage of ≥20x). Contigs with a visible dip in raw molecular coverage were discarded as assembly errors. This resulted in a set of manually curated, non-redundant optical contig consensus sequences that were generated near remaining scaffold gaps, rather than being generated to cover the whole genome, due to finite computational and analytical resources. Comparison of sequence contigs with validated optical contig consensus sequences allowed further scaffolding of the genome assembly and resolution of misassemblies as necessary in Gap5.

de novo assemblies and manual curation

We combined existing short read data [11,12] with additional Illumina data, long PacBio reads (Table S17), optical contigs, and genetic markers [70], to construct an intermediary genome assembly (version 7; GCA_000237925.3) that could be used by the public immediately while time-intensive manual curation took place. Misassemblies were corrected using long-read evidence, as well as optical map data and genetic markers [70]. Remaining gaps were filled using gap-filling software [71,72]. Genetic markers [70] and an updated genetic linkage map(unpublished data, Chevalier et al) were used to assign further scaffolds to chromosomes, and to aid improvement and validation of the rest of the assembly. Version 7 contains 10 chromosomal scaffolds (8 chromosomes plus two scaffolds whose coordinates are known in the W chromosome; 95.91% of scaffolded bases), 13 scaffolds assigned to an autosome with known coordinates (11 of these are primarily repetitive scaffolds), 20 W-specific scaffolds without chromosomal coordinates, 17 scaffolds not assigned to a chromosome, and one mitochondrial scaffold.

Following the v7 assembly submission, we further improved the assembly, particularly in assembling all W-specific contigs and in creating individual chromosomal scaffolds for both Z and W sex chromosomes. To assemble the W chromosome, we first produced separate *de* *novo* assemblies for Illumina and then used Spades [73]) and CANU [74]) to assemble PacBio genomic reads that did not map to the v7 assembly with >500bp of soft-clipping. Second, the *de novo* assemblies were screened against the NCBI NR database in order to screen out any non-*S. mansoni* sequences. New contigs were examined in Gap5 [75] for absence of mapped reads from a male Illumina library (PCR-free pooled male cercariae) and presence of mapped reads from the PCR-free pooled female cercariae Illumina library (Table S17). Manual improvement was performed in Gap5 [75]. Putative new W-specific contigs were examined for sequence similarity to the 22 existing W-specific scaffolds in v7 to determine unique W-specific contigs. All genomic reads (Table S17) were re-mapped to the new assembly and concordant soft-clipped sequences were extended. This process was continued iteratively until no further progress could be made, by which point all contigs terminated in tandem repeats. At this point, the PacBio subreads were surveyed to find long read evidence linking the W chromosome tandem repeats together (Table S16). This elucidated the order of the repeats and W-specific regions to construct a single W chromosome scaffold.

Z and W-specific chromosomal regions were determined from mapping coverage of PCR-free female Illumina libraries (Table S17) with ~22x coverage in the ZSR and ~44x coverage in the PARs, as expected in ZW females. Female-only libraries were used to manually identify gametologues on the W chromosome.

We resolved the haplotypic diversity that typically exists in genome assemblies by sequencing clonal parasites derived from single miracidium-infected snails. Haplotype genes were determined in Gap5 [75] by identifying genes with half coverage, and localisation to a single scaffold that is also half coverage, as compared to non-haplotype scaffolds. An erroneously classified W chromosome scaffold (SM_V7_W019) from v7 was re-classified as a chromosome 1 haplotype. Haplotypes are represented in 259 scaffolds (2.74% of scaffolded bases) (Table S4; DOI:10.5281/zenodo.5149023).

Metaphase fluorescent *in situ* hybridization (FISH) to confirm order of W-specific scaffolds

S. mansoni NMRI strain daughter sporocysts from *B. glabrata* snails were dissected at 29 days post exposure. Sporocysts were placed in 0.05% (0.5mg/ml) colchicine (Sigma-Aldrich) and titurated ~20 times using an 18G blunt-end needle. This single cell suspension was incubated at room temperature for 2-4 hrs to arrest cell division. Cells were spun at 500 x g for 5 min, incubated in nuclease-free water for 20 min at room temperature, and then preserved in ice-cold 3:1 methanol:acetic acid fixative.

Several primer sets were designed to amplify 15 kb-30 kb fragments using the 22 W-specific scaffolds identified post-v7. Fragments were amplified using either PrimeSTAR GXL polymerase (TaKaRa Bio) or LA Taq Hot Start Version Polymerase (TaKaRa Bio) per the manufacturer's instructions. The PCR products were run on an agarose gel and bands of the targeted size were cut and isolated using the QIAEX II Gel Extraction Kit (Qiagen). We successfully amplified sufficient DNA for labelling for scaffolds W005, W002, and W014 to confirm their order in the v9 assembly (Figure S5). Multiplex metaphase FISH and karyotyping were done following the procedures previously described [76].

Arima-HiC data to validate the *S. mansoni* v9 assembly

The Arima-HiC Kit for Animal Tissues (Arima Genomics; Material Part Numbers: A510008 Document Part Number: A160140 v00 Release Date: November 2018) was used following the manufacturer’s instructions with ~100 fresh female *S. mansoni* worms as input. An Illumina library was made using the Swift Biosciences Accel-NGSO 2S Plus DNA Library Kit, with the modified Arima Genomics protocol. The library was sequenced on the Illumina HiSeq X Ten platform resulting in high resolution with >260x coverage of the genome (Table S17). Arima-HiC data was aligned to the v7 assembly using BWA [77]; version 0.7.17). The HiC contact map was made with PretextMap (https://github.com/wtsi-hpag/PretextMap) and viewed in

PretextView (https://github.com/wtsi-hpag/PretextView) (Figure 1). Minor misassemblies and placement of previously 31 unplaced scaffolds were done manually in Gap5 [75].

Illumina RNA-seq and PacBio IsoSeq transcriptome sequencing across *S. mansoni* developmental stages

Illumina RNA-seq libraries were prepared with the TruSeq RNA Library Prep Kit following the manufacturer's protocol. The Smart-seq2 protocol [78] was followed as described to synthesize full length cDNA from 1 µg total RNA for PacBio IsoSeq full-length transcript sequencing. cDNA was amplified in 12 cycles PCR and size fractionated in SageELF electrophoresis system (Sage Science). One or more cDNA size fractions were pooled for the library preparation. For some samples, libraries were produced from more of the size fractions obtained from the SageELF, with the aim of reducing size bias in the PacBio RSII sequencing reads (Table S3).

Heterozygosity in Z and W sex chromosomes and nucleotide diversity in the Z chromosome

Genome-wide SNP calling was performed using GATK HaplotypeCaller with PCR-free Illumina genomic libraries (Table S17) and 7 previously published samples (12663_1_4, 12663_2_4, 7164_6, 7164_7, 7307_7, 7307_8, 8040_3) [32].

To calculate nucleotide diversity (π), median and mean autosomal coverage was calculated for all samples in the Crellen *et al.* data set [32]. Individuals with >10x median and mean coverage on Z and W chromosomes were retained (54 male and 61 female). Of these, the ZSR:PAR ratio was calculated. Individuals with >0.70 ZSR:PAR ratio and a PAR/ZSR <1.5 were designated as males and individuals with <0.70 ZSR:PAR ratio and a PAR/ZSR >1.5 were designated as females. This resulted in a data set consisting of 54 males and 61 females. We used PIXY (v.0.95.01) [79] to calculate π in 50 kb sliding, non-overlapping windows across each chromosome separately for male and female populations for the autosomes. Nucleotide diversity for the ZSR and PARs was calculated in 5 kb sliding, non-overlapping windows. We then calculated the bootstrapped (95%) confidence intervals for each population median using 1000 bootstrap samples of genomic windows for each population using previously published methods [58] (https://github.com/duncanberger/PZQ_POPGEN/blob/master/Figures/figure_2.md). We compared nucleotide diversity between ZSR and the PARs for male individuals testing for significance using an unpaired t-test.

W-repeat classification and quantification

Dot plots were generated for each repeat array on the W chromosome contigs to ensure that a representative repeat unit was selected from each visually distinct section of each repeat array. This process yielded 36 unique repeat unit sequences subsequently named smw01-smw36. The 36 repeat units were compared, pairwise, using blastn with a word size of 6 and dust off. For each comparison with an e-value <0.01, the percentage identity and bit score was recorded and plotted in a matrix plot to reveal similarities between repeat units that define repeat unit families (i.e. Sm-α).

An attempt was made to computationally quantify the W-repeats. Using female PCR-free Illumina data (sample 6520_5; Table S17), gDNA reads were mapped to 19 known single and multi-copy genes (e.g. SmVAL, omega-1) and to all 36 identified W-repeat sequences. Using bedtools coverage on 50 bp windows from the resulting bam file, the single-copy genes had a median coverage of 67 with a range of 54 to 72 and a median of median coverages of 67. SmVAL had double this (151x) and omega-1 had 10 times this (671x) as expected. Taking normal coverage to be 67x, W coverage should be half that at 33.5x. From this we calculated an estimated expected size for our W-repeats (Table S16).

Gene finding

Protein-coding genes

A new protein-coding gene set was produced for the v9 assembly from evidence-based predictions from Augustus [20] with Illumina and PacBio transcriptome reads (Table S3), followed by manual curation. Repeat Modeller v2.0.1 [80] and Repeat Masker v4.1.2 on sensitive mode [81] were run to identify, classify, and mask repetitive elements, including low-complexity sequences and interspersed repeats. The masked genome was then used for gene finding with Augustus v3.2.2 [82] with the following parameters designed to predict one or more splice-forms per gene: *--species=schistosoma2 --UTR=1 --alternative-from-evidence=1.* To predict better gene models and alternative splicing, we used extrinsic information as evidence (i.e. ‘exonpart’ and ‘intron’ hints in Augustus) based on Illumina short reads of all life stages except egg (set priority = 4 in the hints file), and PacBio Iso-seq reads of three life stages (male, female and schistosomula; priority = 40) (Table S3).

To facilitate the comparison of gene sets between assemblies, we also transferred the latest gene models from v5 (based on GeneDB in July 2017) to v7 using RATT [83] with the PacBio setting. The transferred gene models were then compared to those from *de novo* predictions using gffcompare v0.9.9d [84], to determine consensus or novel transcripts (blastn hit of <94% coverage or nucleotide identity <78% between the two assembly versions). When changes occurred compared to a previous gene model, namely an amino acid sequence had changed >20% in either identity or coverage as determined by blastp, or the gene was merged with another gene, or split into several new genes, a new identifier (starting with Smp_3) was assigned and the old Smp number(s) was kept as a previous systematic id (PSID). Otherwise, the previous v5 Smp identifier was transferred to the v7 gene model. Genes that were related to retrotransposons in v5, or not transferred by RATT to the v7 assembly, were not kept in the new gene set. From v7 to v9, gene models were transferred using Liftoff [85]. For gene models with structural changes compared to the v5 gene set, or potentially novel genes predicted by Augustus in the v7 and v9 assemblies, we have carefully inspected them and curated them in Web Apollo [86] (Tables S5, S6).

For functional annotation, blastp v2.7.0 against SwissProt was used to predict product information, and InterproScan v5.25 [87]) to predict product protein domains and Gene Ontology terms. For some genes their product information was preserved from the v5 gene set (taken from GeneDB) if the evidence code was not “Inferred from Electronic Annotation”.

Coverage of UTRs in the genome sequence was calculated as following: first we extract the 5’- and 3’- UTR annotations from the gff file, adding up the total UTR length for each transcript, and then for each gene, we took the transcript with longest UTR as a representative. Finally, all UTRs were summed up for calculating the coverage. Other feature statistics were calculated using Eval v2.2.8 [88].

To recover possible additional novel genes from Boroni *et al [89]*, the CDS/transcript sequences were obtained directly from the authors and aligned to the v9 gene set using blast, where genes with hits were considered as existing. For those without hits to current gene models, their sequences were aligned to the whole genome using blastn and PROmer [90]. Genes with hits to multiple scaffolds were discarded. For genes hitting to the same scaffold the overlapping hit regions were merged using “bedtools merge” and set as “exon” in a gff. All possible models were manually inspected in Apollo using the same RNA-seq tracks as in the publication. We found evidence for 8 of the 759 putative novel genes reported by Boroni *et al.* [89]) (Table S20).

We initially assessed genome completeness using BUSCO v3.0.2 [23]. Although only 85.8% complete eukaryota orthologs were found in the genome sequence (using “--mode genome”; Table S7), representation is expected to be considerably less than 100% in platyhelminths due to their phylogenetic distance from other species in the BUSCO databases [22]. It is known that BUSCO applied to genomic sequences underestimates the completeness of assemblies due to the difficulty of detecting complete genes in the assembly [91] providing further explanation for missing orthologs. As an alternative, we tested the completeness of our predicted gene models using BUSCO (“--mode proteins”) and recovered 95.3% complete eukaryota orthologs.

Transfer RNAs (tRNAs)

tRNAscan v.1.3.1 [92], was used to identify transfer RNAs (tRNAs) in the *S. mansoni* v9 assembly. The algorithm was run with default parameters except for “--forceow --cove”.

Long intergenic non-coding RNAs (lincRNAs)

In order to locate long intergenic non-coding RNAs in v9 of the *S. mansoni* genome assembly, we used RATT [83] to migrate previously generated annotation [93] from v5 to v9. To this end, we downloaded the published annotation as a GFF file, transformed it to EMBL file (as required by RATT) and proceeded to migrate the annotations using the “PacBio” setting of RATT. From a total of 7,029 lincRNAs annotated in v5, 6,876 transfers were made (6,874 unique, two duplications) and 273 lincRNAs were not transferred.

Spliced-leader RNAs (SL RNAs)

Using RNA-seq data (Table S3), we have located SL (spliced leader) sequences in 6,497 genes (Table S8) or 66.3% of all annotated genes in the primary assembly. SL sequences were identified using the canonical *S. mansoni* SL sequence AACCGTCACGGTTTTACTCTTGTGATTTGTTGCATG (Genbank M34074.1 [94]) and a custom in-house spliced leader detection script [95] (https://github.com/stephenrdoyle/hcontortus_genome/blob/5543173b7ee83b903d976931813d85f96f7a6e13/03_code/hcontortus_genome.section5_workbook.md). The script first trims a predefined SL sequence from the 5' end of RNA-seq reads allowing for a minimum length match with an allowed error rate of 10% using Cutadapt [96]. The trimmed sequences are extracted, sorted, and counted, making a sequence logo. The trimmed reads are mapped to the genome using HiSat2 [97] and a BAM file of the mapped trimmed reads is generated for visualisation. A BED file is also made of the splice site coordinates along with a WebLogo [98] of 20 bp surrounding the splice site. Finally, the script determines the coverage of splice sites with transcript starts, (200 bp upstream and 30 bp downstream of the annotated start codon) and internal CDSs, accounting for both misannotated and internal splice variants.

Following published methods [30], we looked for alternative SL sequences using a custom python script to identify reads that (a) aligned to annotated genes, or within 500 bp upstream, and (b) were soft-clipped by more than 5 bp at the 5′ end relative to the annotated gene. Soft-clipped sequences were clustered using CD-HIT-EST v4.7 [99] and only one prominent cluster was identified. Thus, the *S. mansoni* SL sequence appears to be highly conserved within the genome, and there is only one sequence with the abundance of the known SL sequence, occurring in around 10% of the randomly chosen RNA reads.

Gene clusters and gene density in the *S. mansoni* genome

To explore whether there are particular gene functions overrepresented on some chromosomes, we searched for genomically adjacent genes (>=3) with the same Pfam annotations. To investigate whether gene families that had been incorrectly collapsed in the v5 assembly and are now expanded in the v9 assembly, this analysis was performed for both v5 and v9 using Pfam annotations from InterproScan (see “Protein coding genes” section above). For clusters with at least 5 genes, the start coordinates of the first and last genes as well as the number of genes were indicated (Table S2).

IPSE and omega-1 were found to be multi-copy genes clustered in two tandem repeat regions. In order to compare how many bases of curated IPSE and omega-1 genes could be mapped to the v5 and v7 assemblies, we ran Exonerate with a max intron size of 1,500 bp for both IPSE and omega. The IPSE gene Smp_112110 was used in Exonerate, but for omega-1, the mRNA sequence was used because the omega-1 gene has a long and complex gene structure. GFF files were produced of mapped features for IPSE and omega-1 which served to illustrate how many copies of these genes could be annotated.

The IPSE v9 sequence is 199,167 bp with the equivalent v5 sequence is 86,067 bp. The gap in v9 is approximately 29 kb larger than the total of the gaps in v5 in this region. There are approximately 84 additional kilobases in v9 in this region mostly due to expansion of repeat sequence to give a closer representation of reality (Figure S2). Likewise, the omega-1 v9 sequence is 155,103 bp and the v5 sequence is 105,726 bp. There is a 29,982 bp increase due to a large gap in v9, leaving 19,395 bp of additional sequence mainly due to expansion of the repeat array.

Gene expression across different *S. mansoni* life stages and sexes

To explore gene transcript levels across different life stages and between males and females, previously published RNA-seq data [40,41] was used. Briefly, reads were mapped to *S. mansoni* v9 genome using STAR v2.4.2a [100]. Counts per gene and TPMs were summarised with StringTie v2.1.4 [101]. Mean TPM values were calculated for samples of the same life stage and sex and log-transformed. For gametologue expression, only unique mapping reads were used for quantification.

In comparing gene expression of gametologues on WSR and ZSR regions, the ACT genome browser [102] was used with PROmer version 3.07 [103] to show sequence similarity. A transposon inserted into the Smp_318710 pseudogenes was annotated based on PROmer sequence similarity to other transposons on ZSR. For Figure 7, bm_1, bm_2, bm_3 male libraries and bf_1, bf_2, bf_3 female libraries were used [104]. For Figure S7b, bf_1 was used.

Identification of centromeres and telomeres

A 123 bp tandem repeat motif was identified in *S. mansoni* by Melters *et al* [26] due to its high abundance (~1% of the genomic reads), relative to all other tandem repeats in the genome. The original consensus was derived from multiple chromosomes and an almost identical motif is present in chromosomes 1-3, 5-7, and W (Table S11). On both chromosome 4 and Z, single candidate tandem repeats were identified with broadly similar repeat lengths and sequences to previously described consensus motif [26].

We examined repeats in Gap5 [75], taking only the portion of the repeat with the centromere tandem repeat motif. Centromere size estimates (Table S11) were based on Illumina genomic sequencing from female clonally-derived cercariae (sample ERS039722 from Table S17) mapped to 1 representative repeat unit of each of the 8 centromere repeats. As a control, reads were also mapped to the 1st 121 bp of the genomic sequence covered by 12 known single copy genes. These 12 genes gave us a median coverage of 15x. From this we were able to extrapolate sizes for each of the 8 centromeric repeats which totalled 2.25 Mb.

A MAFFT/Jalview alignment was created from all centromere motif sequences [105] and a neighbor-joining tree was constructed using the ETE Toolkit Phylogenetic tree viewer [106] (Table S11)**.** Centromere motif sequence similarity was assessed using the alignment tool PRSS with the Smith-Waterman algorithm (https://embnet.vital-it.ch/software/PRSS_form.html [107,108].

Hirai and LoVerde [109] determined the sequence motif of schistosome telomeres (CCCTAA repeat) through FISH detection. In African schistosomes, the telomeric repeat sequence can be found in the heterochromatin and centromere of the W chromosome. Because it is theorized that *Schistosoma* originated in Asia (see review [110]), the African schistosomes experienced more gene shuffling than the Asian schistosomes, accounting for the presence of telomeric repeats outside the ends of the chromosomes [111].

Abbreviations

PAR: pseudoautosomal region

WSR: W-specific region

ZSR: Z-specific region

BUSCO: Benchmarking Universal Single Copy Orthologs

Mb: megabase

Kb: kilobase

bp: base pair

aa: amino acid

tRNA: transfer RNA

lincRNA: long intergenic non-coding

SL: spliced leader

SLTS: spliced leader trans-splicing

NOR: nucleolar organizer region; rDNA

Declarations

Ethics approval

To propagate the life cycle of the *Schistosoma mansoni* NMRI strain (Puerto Rican) and obtain different developmental stages of the parasite, BALB/c mice and susceptible BB02 strain *Biomphalaria glabrata* snails are routinely infected with parasites at the Wellcome Sanger Institute (WSI). The mouse infections were conducted under Home Office Project Licence No. 80/2596 and No. P77E8A062, and all protocols were presented and approved by the Animal Welfare and Ethical Review Body (AWERB) of the WSI. The AWERB is constituted as required by the UK Animals (Scientific Procedures) Act 1986 Amendment Regulations 2012. With the exception of sporocysts and miracidia, all life cycle stages were collected at the WSI.

Schistosoma mansoni SmBRE strain sporocysts dissected from infected BB02 *B. glabrata* snails were collected at The University of Perpignan laboratory which has permission A 66040 from both the French Ministère de l'agriculture et de la pêche and the French Ministère de l'Education Nationale de la Recherche et de la Technologie for experiments on animals and certificate for animal experimentation (authorization 007083, decree 87-848 and 2012201-0008) for the experimenters. Housing, breeding and animal care follow the national ethical requirements.

Schistosoma mansoni NMRI strain miracidia were collected at Justus-Liebig-University Giessen Institute for Parasitology. Animal experiments were approved by the Regional Council (Regierungspräsidium), Giessen, Germany (V54-19 c 20/15 c GI 18/10), which are in accordance with the European Convention for the Protection of Vertebrate Animals used for experimental and other scientific purposes (ETS No 123; revised Appendix A).

Consent for publication

Not applicable

Availability of data and materials

The primary genome assembly generated and analyzed during this study are available on the European Nucleotide Archive (ENA) website under project accession PRJEB13987. Additional assembled haplotypes, haplotype annotations, and primary assembly annotations can be found in permanent links at WormBase ParaSite under BioProject PRJEA36577 and at Zenodo DOI:10.5281/zenodo.5149023. All other data generated or analyzed during this study are included in this published article and its supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Wellcome Trust grants 098051 and 206194. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Authors' contributions

MB designed research, which was coordinated by NH; SKB and GR maintained the parasite life cycle and generated parasite material, SKB, DB, and GS prepared genomic DNA, RNA and transcriptome sequencing libraries; BF, FY, and SKB performed FISH experiments; SKB, AT, ZL, SD, DB, FR, AJR, UB analyzed data; SKB drafted the complete manuscript, with sections of text contributed by AT, AJR and ZL. All authors read and approved the final manuscript.

Acknowledgements

The *Schistosoma mansoni* genome project is funded by the Wellcome Trust through their core support of the Wellcome Trust Sanger Institute (grant 098051 and 206194). We thank Simon Claire and his team for technical assistance with the life cycle maintenance and collection of parasite material. We thank Tom Huckvale, Hayley Bennet, and Arporn Wangwiwatsin for technical support in generating RNA-seq and gDNA libraries. Thank you to Christoph Puethe for assisting in the ENA submission. A special thank you to Karl Hoffman and Josephine Ford-Thomas for supplying additional infected snails used for metaphase FISH, Paul Brindley and Victoria Mann for clonal worms, and Christoph Grunau and Benjamin Gourbal for providing *in vivo* sporocysts for RNA-seq.

Figure and table legends

Figure 1: **Ideograms of the *S. mansoni* chromosomes with HiC plots showing end-to-end chromosomal resolution.** (a) Representative ZZ (male) and ZW (female) *S. mansoni* metaphase spreads, karyotypes, and ideograms. The yellow arrowheads point to the nucleolar organizer region (NOR; rDNA). Grey regions are euchromatic DNA, black are constitutive heterochromatin (C-band) regions, blue is confirmed telomeric sequence, and light blue bands are confirmed sub-telomeric sequence. (b) HiC visualization plots from genome version 5 (left) and version 9 (right) with the yellow arrowhead pointing to the NOR in chromosome 3.

Table 1: **Genome-wide statistics for the *S. mansoni* haploid v9 assembly compared to the previous v5 assembly.** The v9 assembly size has grown considerably with the addition of 26.9 Mb. The number of gaps present between versions was reduced by 96%, the majority of which are now only present in the collapsed repeat regions of the W chromosome. The chromosomes are assembled into 9 scaffolds (autosomes 1-7, W, and Z). Characterization of SLTS (spliced leader trans-splicing) in the transcripts has increased our previous estimates of only 7% of transcripts being trans-spliced to over 72% in the current assembly. *Completely new, previously partial, or previously unannotated.

Figure 2: **Improvements in the Z-specific region of the Z chromosome between the previous v5 *S.mansoni* genome assembly and current v9 assembly.** Assemblies were compared using PROmer and visualized in ACT. The v5 assembly contained a partially resolved Z chromosome with misassemblies and between the Z-specific region (ZSR) and pseudoautosomal regions (PARs). Corrected inversions from v5 to v9 are shown in lighter blue. Coverage of mapped sequencing reads from female-only sequencing libraries highlight the ZSR as a region with approximately half the depth of coverage as pseudoautosomal regions.

Figure 3: **Z-specific regions of African and Asian *Schistosoma* spp. evolved differentially from an ancestral region that coincides with a likely fusion between chromosomes.** (a)

Evolutionary strata are defined through $\log_2$ genome coverage on the x-axis of one-to-one orthologs in four schistosomes and the hermaphroditic platyhelminth *Echinostoma caproni.* The African-specific stratum in dark purple defines Z-specific region 1 (ZSR1) of *S. mansoni* where approximately half coverage is seen in the African schistosomes *S. mansoni, S. rodhaini*, and *S. haematobium*. The Asian-specific stratum in green has reduced coverage specific only to *S. japonicum* with two possible inversions shown in green brackets. The ancestral *Schistosoma* stratum represents the schistosome orthologs ancestrally isolated to the Z sex chromosome between all schistosome species. (b) Tapeworm orthologs and chromosome synteny blocks show evidence of the fusion in the schistosome lineage between chromosomes 3 and 5 of the tapeworm *Echinococcus multilocularis.* Figure 3b adapted from Olson *et al* [30].

Figure 4: **Median nucleotide diversity (π; pi) across the protein-coding genes of the autosomes, PARs, and Z-specific regions using published genome variation data** [32]**.** Median nucleotide diversity (π; pi) was calculated separately for males (left) and females (right) in 50 kb windows (a,b) or 5 kb windows (c-f) across all protein-coding genes. Pi is shown for the autosomes (a,b), PARs and ZSRs (c,d) and combined autosomal regions and ZSRs (e,f).

Figure 5: **Detailed, annotated idiograms of the Z and W sex chromosomes.** (a) The true size of the W chromosome is approximately 14% larger than Z which can be accounted for in the large expansion of repeats in the WSR. All but 36 genes have been lost on the WSR with 5 of those being pseudogenes and 2 present in triplicate and duplicate. Chevrons mark the approximate location of 3 euchromatin bands in the WSR. (b) The assembled size of the WSR is ~6 Mb, less than its true size of ~46 Mb because of 36 collapsed repeats. (c) C-banding shows the alternating AT- and GC-rich DNA repeats present in the WSR.

Figure 6: **Illumina RNA-seq expression of the W and Z gametologues in adult paired and naive male and female *S. mansoni* worms.** Unique mapping of RNA-seq to the gametologues reveals relatively similar expression of the Z gametologues between males and females for most gametologues. As expected, the W gametologues show expression limited to the female

samples. Lines connect gametologue pairs between Z and W. In two cases, the W gametologue exists in triplicate or duplicate (see W gametologues Smp_317860, Smp_317870, Smp_348710 and Smp_318680, Smp_318860).

Figure 7: **A comparison of U2AF 65 kDa subunit gametologues on the Z and W chromosomes.** The gametologues of the large, 65 kDa subunit of U2AF (Z: Smp_019690; W: Smp_348790) are shown on ZSR (top) and WSR (bottom). Predicted transcript sequences in yellow. Sequence similarity was determined using PROmer and shows that the N-terminal region of the coding sequence (blue) is more diverse. The black arrow head highlights the position of a likely sequencing error on WSR which causes a frameshift, but which has been corrected in the gene model. Unnormalised coverage of RNA-seq reads is shown for female (bf_1, bf_2, bf_3) and male samples (bm_1, bm_2, bm_3). This highlights male expression on only the ZSR, with lower female coverage on ZSR and WSR as expected. Numbers above gene models indicate position on the contigs, numbers above RNA-seq coverage indicate maximum read depth.

Figure 8: **Hypothetical illustration of the schistosome Z and W sex chromosome evolution.** A chromosomal fusion between two ancestral schistosome autosomes occurred near the ZSR stratum boundary (see Figure 3) creating a new set of autosomes. Followed by, or in conjunction with, this fusion event, a male antagonistic and/or female advantageous locus was isolated on the pre-sex chromsomes (see Figure 6; potentially pre-mRNA splicing factor U2AF). The need to isolate the phenotypic effects of the gene(s) in this locus on the pre-W chromosome required recombination suppression (see Figure 4). Isolation of additional loci with sex-specific effects and elimination and/or pseudogenization of non-sex-specific coding loci is evidenced in Fig 5. Following initial recombination suppression, extensive heterochromatinization of W ensured long-term recombination suppression between W and Z sex-specific regions and resulted in the huge expansion of repeats in the W-specific region (Figure 5; Table S16)

Additional Files

Additional file 1: **Supplementary Tables S1 to S20**

Supplementary Figures Titles and Legends

Figure S1: **Array of spliced-leader RNA genes on chromosome 6 of the *S. mansoni* genome**. On chromosome six, a 62.6 kb locus exists containing 41 full-length spliced leader RNA genes (top track). An additional 109 partial gene sequences that contain the spliced leader exon sequence only exist in the same array (bottom track).

Figure S2: **Resolving the repetitive IPSE and omega-1 loci.** Genes in the (a) IPSE loci and (b) Omega-1 locus shown in v9 through gene model annotations (top tracks) and genomic coverage mapping (bottom tracks) with yellow boxes to connect gene annotations to genomic coverage. The annotations show the v9 gene models, some of which coincide with elevated read coverage. The histogram in the coverage plots show depth of read coverage and compared with the flanking regions, the depth is elevated in the IPSE and omega-1 loci suggesting these gene arrays are smaller in this assembly than their true size.

Figure S3: **Gene changes from genome v5 to v9 of *S. mansoni*.** There have been a total of 3,610 gene changes represented by 810 new, 867 deleted, 344 merged, 189 multiple copies, 190 split, and 1,210 structurally changed. The bar graph shows totals of different protein-coding gene changes in the primary assembly (i.e. no gene fragments, haplotypes, or pseudogenes).

Figure S4: **Alignment of the centromeric repeat sequences relatedness between all *S. mansoni* chromosomes.** MAFFT/Jalview alignment of a single centromeric repeat unit from each chromosome shows high similarity between chromosomes 1-3, 5-7, and W. Chromosomes 1-3, 5-7, and W are 93.1-98.5% identical to a 123bp centromeric repeat

proposed by Melters *et al [26].* The centromeric repeats for chromosomes 4 and Z are diverged from the other chromosomes.

Figure S5: **Validation of the assembly and placement of W-specific scaffolds using metaphase FISH.** Twenty-two W-specific scaffolds existed after computational and manual assembly. Scaffold W007 contained the junction from PAR1 into the WSR and scaffold W016 spanned the WSR into PAR2. The centromeric repeat for the W chromosome was in scaffold W002 (7.65-7.75 Mb) with the orientation of this contig inferred from alignment of centromere sequence in this scaffold. The remaining 8 scaffolds with gametologues (purple) and 11 scaffolds without gametologues (black), whose positions and orientations could not be determined using sequence data alone, were placed using metaphase FISH.

Figure S6: **Measurements of Z and W chromosomes from 6 female metaphase cells.** The W chromosome is approximately 14.7% larger than the Z chromosome based on measurements taken of the chromosomes from the metaphase figures shown. Measurements were taken using the measurement tool in Inkscape. This figure is consistent with previously published measurements from 22 female metaphase cells [33].

Figure S7: **Comparisons of ancestral region gametologues between ZSR and WSR.** (a) The Z gametologue Smp_158310 is clearly expressed in males (red RNA-seq coverage) and females (blue RNA-seq coverage), but the W gametologue Smp_318710 is not. Furthermore, the gene model is incomplete on WSR and there is a transposon inserted within the gene (red bar), resulting in a pseudogene. The genes are inverted between ZSR and WSR, indicated by the overlapping sequence similarity bars. (b) The genome sequence for the WSR gametologue of U2AF 65kDa (Smp_348790) subunit contains a single base insertion, suggesting a possible frameshift mutation. However, RNA-seq reads show that this is a sequencing error and the corrected gene model based on this data results in an N-terminal amino acid sequence more similar to, although still somewhat divergent from, the Z gametologue (Smp_019690).

References

1. Toor J, Alsallaq R, Truscott JE, Turner HC, Werkman M, Gurarie D, et al. Are We on Our Way to Achieving the 2020 Goals for Schistosomiasis Morbidity Control Using Current World Health Organization Guidelines? Clin Infect Dis. 2018;66:S245–52. Available from: http://dx.doi.org/10.1093/cid/ciy001

2. Colley DG, Bustinduy AL, Secor WE, King CH. Human schistosomiasis. Lancet. 2014;383:2253–64. Available from: http://dx.doi.org/10.1016/S0140-6736(13)61949-2

3. Loker ES, Brant SV. Diversification, dioecy and dimorphism in schistosomes. Trends Parasitol. 2006;22:521–8. Available from: http://dx.doi.org/10.1016/j.pt.2006.09.001

4. Platt TR, Brooks DR. Evolution of the schistosomes (Digenea: Schistosomatoidea): the origin of dioecy and colonization of the venous system. J Parasitol. 1997;83:1035–44. Available from: https://www.ncbi.nlm.nih.gov/pubmed/9406775

5. Tomaszkiewicz M, Medvedev P, Makova KD. Y and W Chromosome Assemblies: Approaches and Discoveries. Trends in Genetics. 2017. p. 266–82. Available from: http://dx.doi.org/10.1016/j.tig.2017.01.008

6. Xue L, Gao Y, Wu M, Tian T, Fan H, Huang Y, et al. Telomere-to-telomere assembly of a fish Y chromosome reveals the origin of a young sex chromosome pair. Genome Biol. 2021;22:203. Available from: http://dx.doi.org/10.1186/s13059-021-02430-y

7. Mahajan S, Wei KH-C, Nalley MJ, Gibilisco L, Bachtrog D. De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. PLoS Biol. 2018;16:e2006348. Available from: http://dx.doi.org/10.1371/journal.pbio.2006348

8. Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, et al. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. Nat Genet. 2017;49:387–94. Available from: http://dx.doi.org/10.1038/ng.3778

9. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003;423:825–37. Available from: http://dx.doi.org/10.1038/nature01722

10. Hirai H, Spotila LD, LoVerde PT. Schistosoma mansoni: chromosomal localization of DNA repeat elements by in situ hybridization using biotinylated DNA probes. Exp Parasitol. 1989;69:175–88. Available from: https://www.ncbi.nlm.nih.gov/pubmed/2753121

11. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, et al. The genome of the blood fluke Schistosoma mansoni. Nature. 2009;460:352–8. Available from: http://dx.doi.org/10.1038/nature08160

12. Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, et al. A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. PLoS Negl Trop Dis. 2012;6:e1455. Available from: http://dx.doi.org/10.1371/journal.pntd.0001455

13. Abdulla M-H, Lim K-C, McKerrow JH, Caffrey CR. Proteomic identification of IPSE/alpha-1 as a major hepatotoxin secreted by Schistosoma mansoni eggs. PLoS Negl Trop Dis. 2011;5:e1368. Available from: http://dx.doi.org/10.1371/journal.pntd.0001368

14. Everts B, Perona-Wright G, Smits HH, Hokke CH, van der Ham AJ, Fitzsimmons CM, et al. Omega-1, a glycoprotein secreted by Schistosoma mansoni eggs, drives Th2 responses. J Exp Med. 2009;206:1673–80. Available from: http://dx.doi.org/10.1084/jem.20082460

15. Costain AH, MacDonald AS, Smits HH. Schistosome Egg Migration: Mechanisms, Pathogenesis and Host Immune Responses. Front Immunol. 2018;9:3042. Available from: http://dx.doi.org/10.3389/fimmu.2018.03042

16. Hagen J, Young ND, Every AL, Pagel CN, Schnoeller C, Scheerlinck J-PY, et al. Omega-1 knockdown in Schistosoma mansoni eggs by lentivirus transduction reduces granuloma size in vivo. Nat Commun. 2014;5:5375. Available from: http://dx.doi.org/10.1038/ncomms6375

17. Ranasinghe SL, Fischer K, Gobert GN, McManus DP. Functional expression of a novel Kunitz type protease inhibitor from the human blood fluke Schistosoma mansoni. Parasit Vectors. 2015;8:408. Available from: http://dx.doi.org/10.1186/s13071-015-1022-z

18. Ingram JR, Rafi SB, Alegra Eroy-Reveles A, Ray M, Lambeth L, Hsieh I, et al. Investigation of the Proteolytic Functions of an Expanded Cercarial Elastase Gene Family in Schistosoma mansoni. PLoS Neglected Tropical Diseases. 2012. p. e1589. Available from: http://dx.doi.org/10.1371/journal.pntd.0001589

19. Buddenborg SK, Kamel B, Hanelt B, Bu L, Zhang S-M, Mkoji GM, et al. The in vivo transcriptome of Schistosoma mansoni in two prominent vector species, Biomphalaria pfeifferi and B. glabrata. 2019. Available from: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0007013

20. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33:W465–7. Available from: http://dx.doi.org/10.1093/nar/gki458

21. Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sánchez-Flores A, Brooks KL, et al. The genomes of four tapeworm species reveal adaptations to parasitism. Nature. 2013;496:57–63. Available from: http://dx.doi.org/10.1038/nature12031

22. Olson PD, Tracey A, Baillie A, James K, Doyle SR. Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. bioRxiv. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.04.08.031872v1.abstract

23. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol. Oxford Academic; 2017 [cited 2020 Nov 3];35:543–8. Available from: https://academic.oup.com/mbe/article-pdf/35/3/543/24367705/msx319.pdf

24. Allen MA, Hillier LW, Waterston RH, Blumenthal T. A global analysis of C. elegans trans-splicing. Genome Res. 2011;21:255–64. Available from: http://dx.doi.org/10.1101/gr.113811.110

25. Short RB, Grossman AI. Conventional Giemsa and C-Banded Karyotypes of Schistosoma mansoni and S. rodhaini. The Journal of Parasitology. 1981. p. 661. Available from: http://dx.doi.org/10.2307/3280440

26. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 2013;14:R10. Available from: http://dx.doi.org/10.1186/gb-2013-14-1-r10

27. Picard MAL, Cosseau C, Ferré S, Quack T, Grevelding CG, Couté Y, et al. Evolution of gene dosage on the Z-chromosome of schistosome parasites. Elife. 2018;7. Available from: http://dx.doi.org/10.7554/eLife.35684

28. Lepesant JMJ, Cosseau C, Boissier J, Freitag M, Portela J, Climent D, et al. Chromatin structural changes around satellite repeats on the female sex chromosome in Schistosoma mansoni and their possible role in sex chromosome emergence. Genome Biol. 2012;13:R14. Available from: http://dx.doi.org/10.1186/gb-2012-13-2-r14

29. Hirai H, Hirai Y, LoVerde PT. Evolution of sex chromosomes ZW of Schistosoma mansoni inferred from chromosome paint and BAC mapping analyses. Parasitol Int. 2012;61:684–9. Available from: http://dx.doi.org/10.1016/j.parint.2012.07.007

30. Olson PD, Tracey A, Baillie A, James K, Doyle SR, Buddenborg SK, et al. Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. BMC Biol. 2020;18:165. Available from: http://dx.doi.org/10.1186/s12915-020-00899-w

31. Wilson Sayres MA. Genetic Diversity on the Sex Chromosomes. Genome Biol Evol. 2018;10:1064–78. Available from: http://dx.doi.org/10.1093/gbe/evy039

32. Crellen T, Allan F, David S, Durrant C, Huckvale T, Holroyd N, et al. Whole genome resequencing of the human parasite Schistosoma mansoni reveals population history and effects of selection. Sci Rep. 2016;6:20954. Available from: http://dx.doi.org/10.1038/srep20954

33. Hirai H, Spotila LD, LoVerde PT. Schistosoma mansoni: chromosomal localization of DNA repeat elements by in situ hybridization using biotinylated DNA probes. Exp Parasitol. 1989;69:175–88. Available from: https://www.ncbi.nlm.nih.gov/pubmed/2753121

34. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. BioMed Central; 2019 [cited 2020 Oct 9];20:1–23. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1707-2

35. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, et al. Non-random DNA fragmentation in next-generation sequencing. Sci Rep. Nature Publishing Group; 2014 [cited 2020 Oct 9];4:1–6. Available from: https://www.nature.com/articles/srep04532

36. Drew AC, Brindley PJ. Female-specific sequences isolated from Schistosoma mansoni by representational difference analysis. Mol Biochem Parasitol. 1995;71:173–81. Available from: http://dx.doi.org/10.1016/0166-6851(95)00048-6

37. Spotila LD, Hirai H, Rekosh DM, Lo Verde PT. A retroposon-like short repetitive DNA element in the genome of the human blood fluke, Schistosoma mansoni. Chromosoma. 1989;97:421–8. Available from:

http://dx.doi.org/10.1007/BF00295025

38. Spotila LD, LoVerde PT, Rekosh DM. Analysis of two repeated DNA sequences of Schistosoma mansoni. UCLA symposium on Molecular and Cellular Biology, 6ol. 1987. p. 159–68.

39. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636–41. Available from: http://dx.doi.org/10.1093/nar/gkz268

40. Lu Z, Sessler F, Holroyd N, Hahnel S, Quack T, Berriman M, et al. A gene expression atlas of adult Schistosoma mansoni and their gonads. Sci Data. 2017;4:170118. Available from: http://dx.doi.org/10.1038/sdata.2017.118

41. Picard MAL, Boissier J, Roquis D, Grunau C, Allienne J-F, Duval D, et al. Sex-Biased Transcriptome of Schistosoma mansoni: Host-Parasite Interaction, Genetic Determinants and Epigenetic Regulators Are Associated with Sexual Differentiation. PLoS Negl Trop Dis. 2016;10:e0004930. Available from: http://dx.doi.org/10.1371/journal.pntd.0004930

42. Rawlinson KA, Reid AJ, Lu Z, Driguez P, Wawer A. Daily rhythms in the transcriptomes of the human parasite Schistosoma mansoni. bioRxiv. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.04.21.440693v2.abstract

43. Everts B, Perona-Wright G, Smits HH, Hokke CH, van der Ham AJ, Fitzsimmons CM, et al. Omega-1, a glycoprotein secreted by Schistosoma mansoni eggs, drives Th2 responses. J Exp Med. 2009;206:1673–80. Available from: http://dx.doi.org/10.1084/jem.20082460

44. Schwartz C, Fallon PG. Schistosoma “Eggs-Iting” the Host: Granuloma Formation and Egg Excretion. Front Immunol. 2018;9:2492. Available from: http://dx.doi.org/10.3389/fimmu.2018.02492

45. Foster JM, Grote A, Mattick J, Tracey A, Tsai Y-C, Chung M, et al. Sex chromosome evolution in parasitic nematodes of humans. Nature Communications. 2020. Available from: http://dx.doi.org/10.1038/s41467-020-15654-6

46. Sahara K, Yoshido A, Traut W. Sex chromosome evolution in moths and butterflies. Chromosome Res. 2012;20:83–94. Available from: http://dx.doi.org/10.1007/s10577-011-9262-z

47. Singh RK, Singh D, Yadava A, Srivastava AK. Molecular fossils “pseudogenes” as functional signature in biological system. Genes Genomics. 2020;42:619–30. Available from: https://doi.org/10.1007/s13258-020-00935-7

48. Picard MAL, Vicoso B, Bertrand S, Escriva H. Diversity of Modes of Reproduction and Sex Determination Systems in Invertebrates, and the Putative Contribution of Genetic Conflict. Genes. 2021. p. 1136. Available from: http://dx.doi.org/10.3390/genes12081136

49. Elkrewi M, Moldovan MA, Picard MAL, Vicoso B. Schistosome W-linked genes inform temporal dynamics of sex chromosome evolution and suggest candidate for sex determination. Mol Biol Evol. 2021; Available from: http://dx.doi.org/10.1093/molbev/msab178

50. Krämer A. The structure and function of proteins involved in mammalian pre-mRNA splicing. Annu Rev Biochem. 1996;65:367–409. Available from: http://dx.doi.org/10.1146/annurev.bi.65.070196.002055

51. Rudner DZ, Kanaar R, Breger KS, Rio DC. Mutations in the small subunit of the Drosophila U2AF splicing factor cause lethality and developmental defects. Proc Natl Acad Sci U S A. 1996;93:10333–7. Available from: http://dx.doi.org/10.1073/pnas.93.19.10333

52. Bashaw GJ, Baker BS. The msl-2 dosage compensation gene of Drosophila encodes a putative DNA-binding protein whose expression is sex specifically regulated by Sex-lethal. Development. 1995;121:3245–58. Available from: https://www.ncbi.nlm.nih.gov/pubmed/7588059

53. Kelley RL, Solovyeva I, Lyman LM, Richman R, Solovyev V, Kuroda MI. Expression of msl-2 causes assembly of dosage compensation regulators on the X chromosomes and female lethality in Drosophila. Cell. 1995;81:867–77. Available from: http://dx.doi.org/10.1016/0092-8674(95)90007-1

54. Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, et al. Sex determination: why so many ways of doing it? PLoS Biol. 2014;12:e1001899. Available from: http://dx.doi.org/10.1371/journal.pbio.1001899

55. Dean R, Mank JE. The role of sex chromosomes in sexual dimorphism: discordance between molecular and phenotypic data. Journal of Evolutionary Biology. 2014. p. 1443–53. Available from: http://dx.doi.org/10.1111/jeb.12345

56. Graves JAM. Sex chromosome specialization and degeneration in mammals. Cell. 2006;124:901–14. Available from: http://dx.doi.org/10.1016/j.cell.2006.02.024

57. Charlesworth D, Charlesworth B, Marais G. Steps in the evolution of heteromorphic sex chromosomes. Heredity. 2005;95:118–28. Available from: http://dx.doi.org/10.1038/sj.hdy.6800697

58. Berger D, Crellen T, Lamberton PHL, Allan F, Tracey A, Noonan JD, et al. Data release: Whole-genome sequencing of Schistosoma mansoni reveals extensive diversity with limited selection despite mass drug administration. 2021. Available from: https://zenodo.org/record/5045162

59. Le Clec'h W, Chevalier FD, Mattos ACA, Strickland A, Diaz R, McDew-White M, et al. Genetic analysis of praziquantel resistance in schistosome parasites implicates a Transient Receptor Potential channel. bioRxiv. 2021 [cited 2021 Aug 10]. p. 2021.06.09.447779. Available from: https://www.biorxiv.org/content/10.1101/2021.06.09.447779v1.abstract

60. Picard MAL, Vicoso B, Roquis D, Bulla I, Augusto RC, Arancibia N, et al. Dosage Compensation throughout the Schistosoma mansoni Lifecycle: Specific Chromatin Landscape of the Z Chromosome. Genome Biol Evol. 2019;11:1909–22. Available from: http://dx.doi.org/10.1093/gbe/evz133

61. Augusto R de C, de Carvalho Augusto R, Cosseau C, Grunau C. Histone Methylome of the Human Parasite Schistosoma Mansoni. RNA Technologies. 2019. p. 607–24. Available from: http://dx.doi.org/10.1007/978-3-030-14792-1_24

62. Roquis D, Taudt A, Geyer KK, Padalino G, Hoffmann KF, Holroyd N, et al. Histone methylation changes are required for life cycle progression in the human parasite Schistosoma mansoni. PLoS Pathog. 2018;14:e1007066. Available from: http://dx.doi.org/10.1371/journal.ppat.1007066

63. Mann VH, Morales ME, Rinaldi G, Brindley PJ. Culture for genetic manipulation of developmental stages of Schistosoma mansoni. Parasitology. 2010;137:451–62. Available from: http://dx.doi.org/10.1017/S0031182009991211

64. Lewis F. Schistosomiasis. Curr Protoc Immunol. 2001;Chapter 19:Unit 19.1. Available from: http://dx.doi.org/10.1002/0471142735.im1901s28

65. Grevelding CG, Kampkötter A, Kunz W. Schistosoma mansoni: sexing cercariae by PCR without DNA extraction. Exp Parasitol. 1997;85:99–100. Available from: http://dx.doi.org/10.1006/expr.1996.4129

66. Wangwiwatsin A, Protasio AV, Wilson S, Owusu C. Transcriptome of the parasitic flatworm Schistosoma mansoni during intra-mammalian development. bioRxiv. biorxiv.org; 2019; Available from: https://www.biorxiv.org/content/10.1101/757633v1.abstract

67. Park N, Shirley L, Gu Y, Keane TM, Swerdlow H, Quail MA. An improved approach to mate-paired library preparation for Illumina sequencing. Methods in Next Generation Sequencing. 2013;1. Available from: http://access.portico.org/stable?au=pgj2kqzq4j1

68. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009;6:291–5. Available from: http://dx.doi.org/10.1038/nmeth.1311

69. Kuk FK, Tyler RS, Rustad N, Harker LA, Tye-Murray N. Alternating current at the eardrum for tinnitus reduction. J Speech Hear Res. 1989;32:393–400. Available from: http://dx.doi.org/10.1044/jshr.3202.393

70. Criscione CD, Valentim CLL, Hirai H, LoVerde PT, Anderson TJC. Genomic linkage map of the human blood fluke Schistosoma mansoni. Genome Biol. 2009;10:R71. Available from: http://dx.doi.org/10.1186/gb-2009-10-6-r71

71. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 2010;11:R41. Available from: http://dx.doi.org/10.1186/gb-2010-11-4-r41

72. Nadalin F, Vezzi F, Policriti A. GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics. 2012;13 Suppl 14:S8. Available from: http://dx.doi.org/10.1186/1471-2105-13-S14-S8

73. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77. Available from: http://dx.doi.org/10.1089/cmb.2012.0021

74. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36. Available from: http://dx.doi.org/10.1101/gr.215087.116

75. Bonfield JK, Whitwham A. Gap5--editing the billion fragment sequence assembly. Bioinformatics. 2010;26:1699–703. Available from: http://dx.doi.org/10.1093/bioinformatics/btq268

76. Algady W, Louzada S, Carpenter D, Brajer P, Färnert A, Rooth I, et al. The Malaria-Protective Human Glycophorin Structural Variant DUP4 Shows Somatic Mosaicism and Association with Hemoglobin Levels. Am J Hum Genet. 2018;103:769–76. Available from: http://dx.doi.org/10.1016/j.ajhg.2018.10.008

77. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. Available from: http://dx.doi.org/10.1093/bioinformatics/btp324

78. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9:171–81. Available from: http://dx.doi.org/10.1038/nprot.2014.006

79. Korunes KL, Samuk K. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. bioRxiv. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.06.27.175091v1.abstract

80. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7. Available from: http://dx.doi.org/10.1073/pnas.1921046117

81. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. RepeatMasker. 2013-2015. Available from: <http://www.repeatmasker.org>

82. König S, Romoth LW, Gerischer L, Stanke M. Simultaneous gene finding in multiple genomes. Bioinformatics. 2016;32:3388–95. Available from: http://dx.doi.org/10.1093/bioinformatics/btw494

83. Otto TD, Dillon GP, Degrave WS, Berriman M. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res. 2011;39:e57. Available from: http://dx.doi.org/10.1093/nar/gkq1268

84. Pertea G. GffCompare. 2018.

85. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021. Available from: http://dx.doi.org/10.1093/bioinformatics/btaa1016

86. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013;14:R93. Available from: http://dx.doi.org/10.1186/gb-2013-14-8-r93

87. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. Available from: http://dx.doi.org/10.1093/bioinformatics/btu031

88. Keibler E, Brent MR. Eval: a software package for analysis of genome annotations. BMC Bioinformatics. 2003;4:50. Available from: http://dx.doi.org/10.1186/1471-2105-4-50

89. Boroni M, Sammeth M, Gava SG, Jorge NAN, Macedo AM, Machado CR, et al. Landscape of the spliced leader trans-splicing mechanism in Schistosoma mansoni. Sci Rep. 2018;8:3877. Available from: http://dx.doi.org/10.1038/s41598-018-22093-3

90. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002;30:2478–83. Available from: http://dx.doi.org/10.1093/nar/30.11.2478

91. Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, et al. Six reference-quality genomes reveal evolution of bat adaptations. Nature. 2020;583:578–84. Available from: http://dx.doi.org/10.1038/s41586-020-2486-3

92. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 2016;44:W54–7. Available from: http://dx.doi.org/10.1093/nar/gkw413

93. Vasconcelos EJR, daSilva LF, Pires DS, Lavezzo GM, Pereira ASA, Amaral MS, et al. The Schistosoma mansoni genome encodes thousands of long non-coding RNAs predicted to be functional at different parasite life-cycle stages. Sci Rep. 2017;7:10508. Available from: http://dx.doi.org/10.1038/s41598-017-10853-6

94. Rajkovic A, Davis RE, Simonsen JN, Rottman FM. A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni. Proc Natl Acad Sci U S A. 1990;87:8879–83. Available from: http://dx.doi.org/10.1073/pnas.87.22.8879

95. Doyle SR, Tracey A, Laing R, Holroyd N, Bartley D, Bazant W, et al. Genomic and transcriptomic variation defines the chromosome-scale assembly of Haemonchus contortus, a model gastrointestinal worm. Commun Biol. 2020;3:656. Available from: http://dx.doi.org/10.1038/s42003-020-01377-3

96. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 [cited 2020 Apr 6];17:10–2. Available from: http://journal.embnet.org/index.php/embnetjournal/article/view/200

97. Kim D, Langmead B, Salzberg SL. hisat2. Nat Methods. 2015; Available from:

https://ccb.jhu.edu/software/hisat2/data/HISAT2-first_release-Sept_8_2015.pdf

98. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188–90. Available from: http://dx.doi.org/10.1101/gr.849004

99. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2. Available from: http://dx.doi.org/10.1093/bioinformatics/bts565

100. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. Available from: http://dx.doi.org/10.1093/bioinformatics/bts635

101. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5. Available from: http://dx.doi.org/10.1038/nbt.3122

102. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics. 2005;21:3422–3. Available from: http://dx.doi.org/10.1093/bioinformatics/bti553

103. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:e1005944. Available from: http://dx.doi.org/10.1371/journal.pcbi.1005944

104. Lu Z, Sessler F, Holroyd N, Hahnel S, Quack T, Berriman M, et al. Schistosome sex matters: a deep view into gonad-specific and pairing-dependent transcriptomes reveals a complex gender interplay. Sci Rep. 2016;6:31150. Available from: http://dx.doi.org/10.1038/srep31150

105. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25:1189–91. Available from: http://dx.doi.org/10.1093/bioinformatics/btp033

106. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33:1635–8. Available from: http://dx.doi.org/10.1093/molbev/msw046

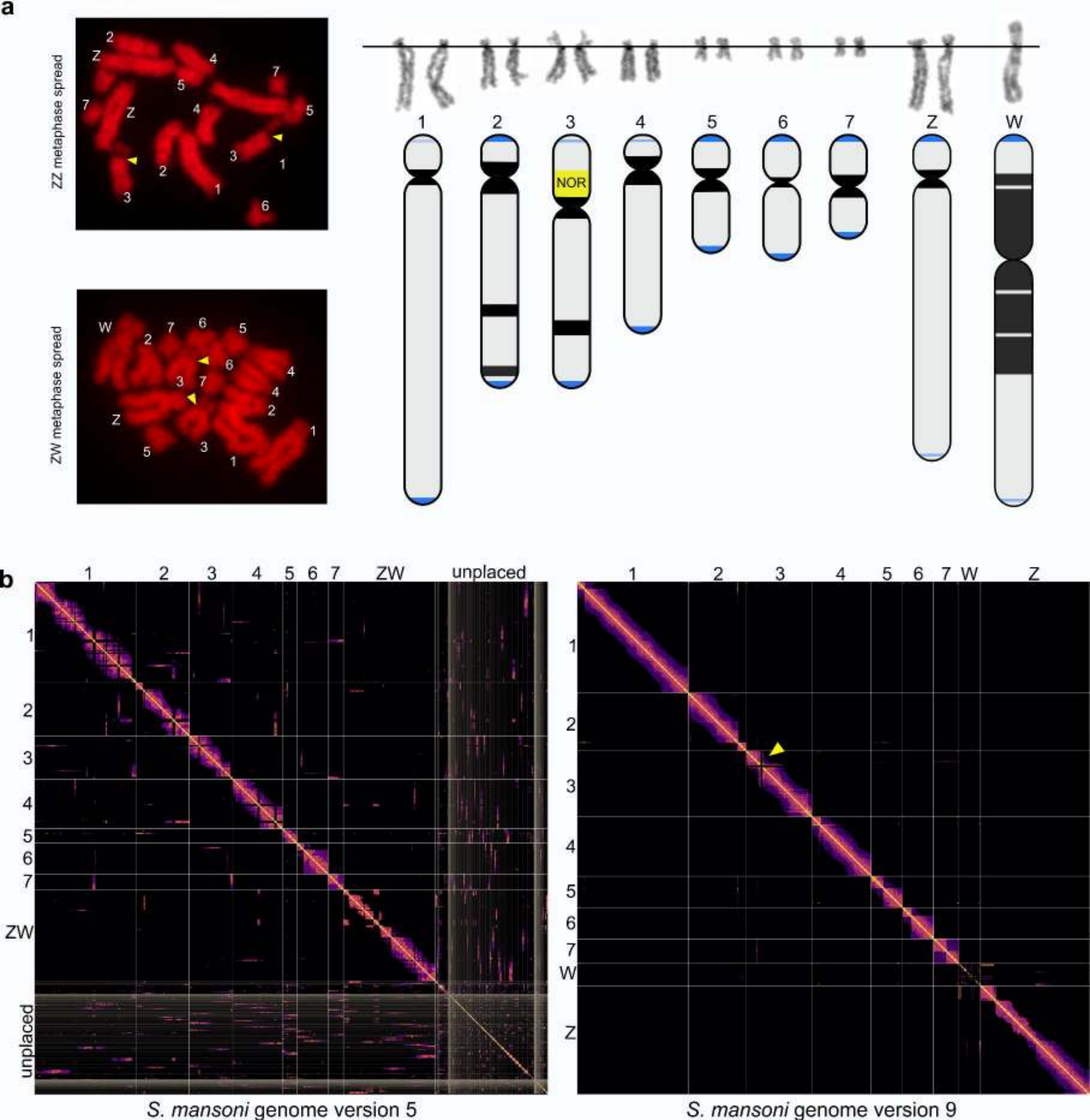
107. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85:2444–8. Available from: http://dx.doi.org/10.1073/pnas.85.8.2444

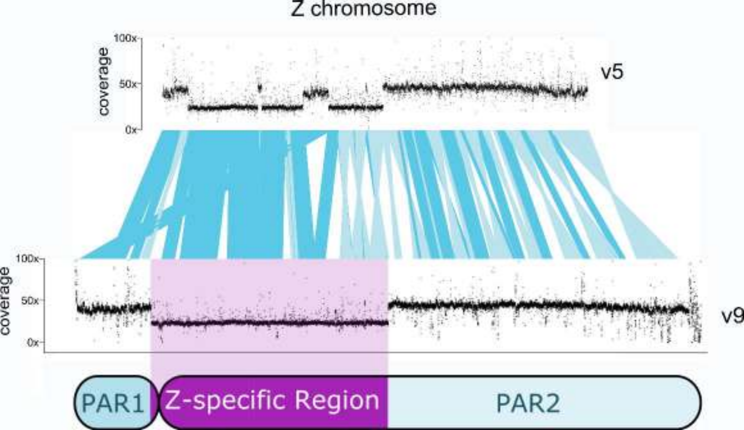
108. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences. 1990. p. 2264–8. Available from: http://dx.doi.org/10.1073/pnas.87.6.2264

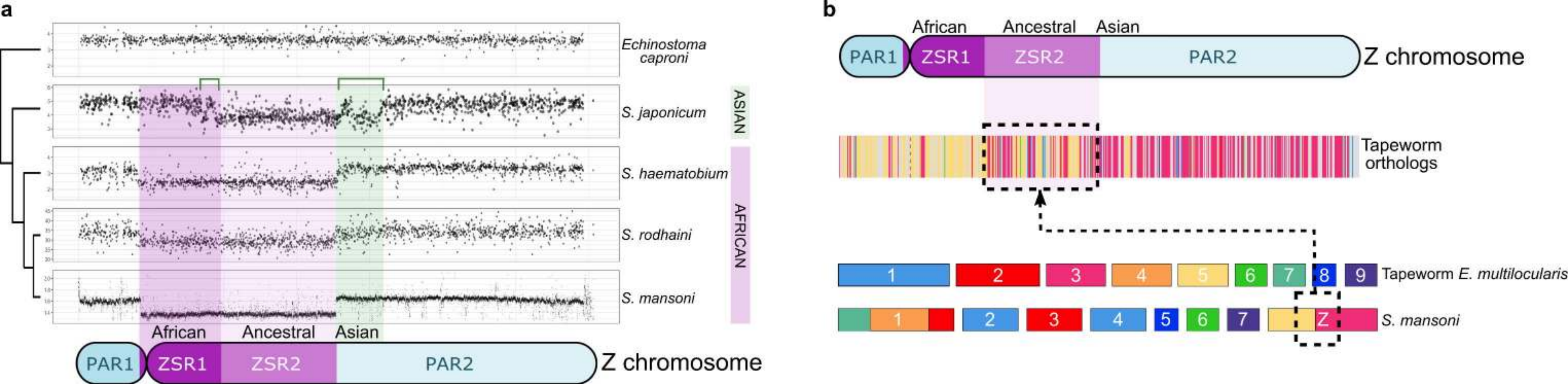
109. Hirai H, LoVerde PT. Identification of the telomeres on Schistosoma mansoni chromosomes by FISH. J Parasitol. 1996;82:511–2. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8636864

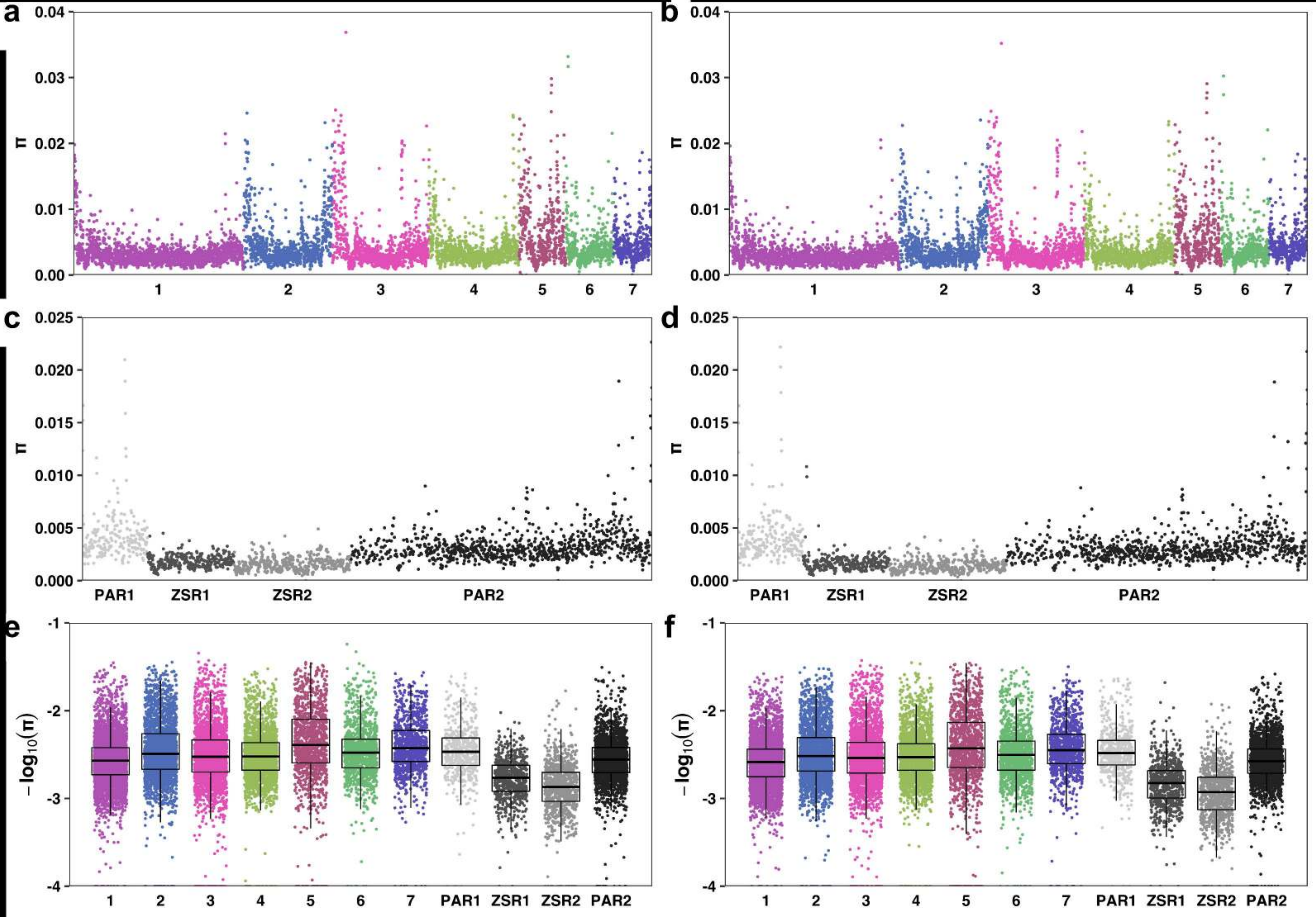
110. Lawton SP, Hirai H, Ironside JE, Johnston DA, Rollinson D. Genomes and geography: genomic insights into the evolution and phylogeography of the genus Schistosoma. Parasit Vectors. 2011;4:131. Available from: http://dx.doi.org/10.1186/1756-3305-4-131

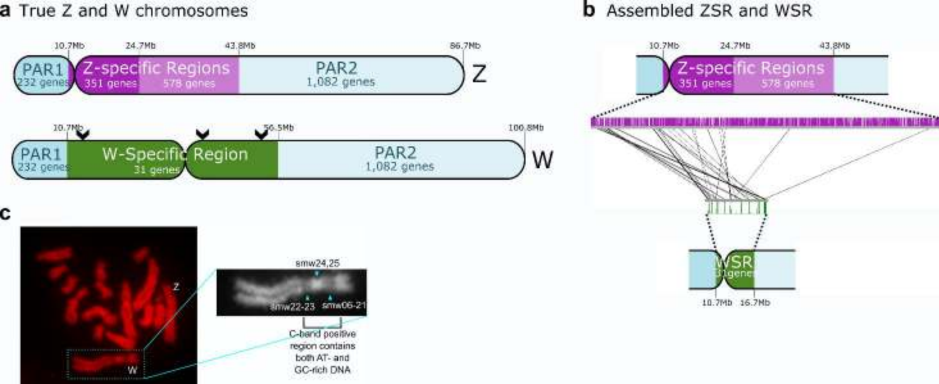
111. Hirai H. Chromosomal differentiation of schistosomes: what is the message? Front Genet. Frontiers; 2014;5:301. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2014.00301/full

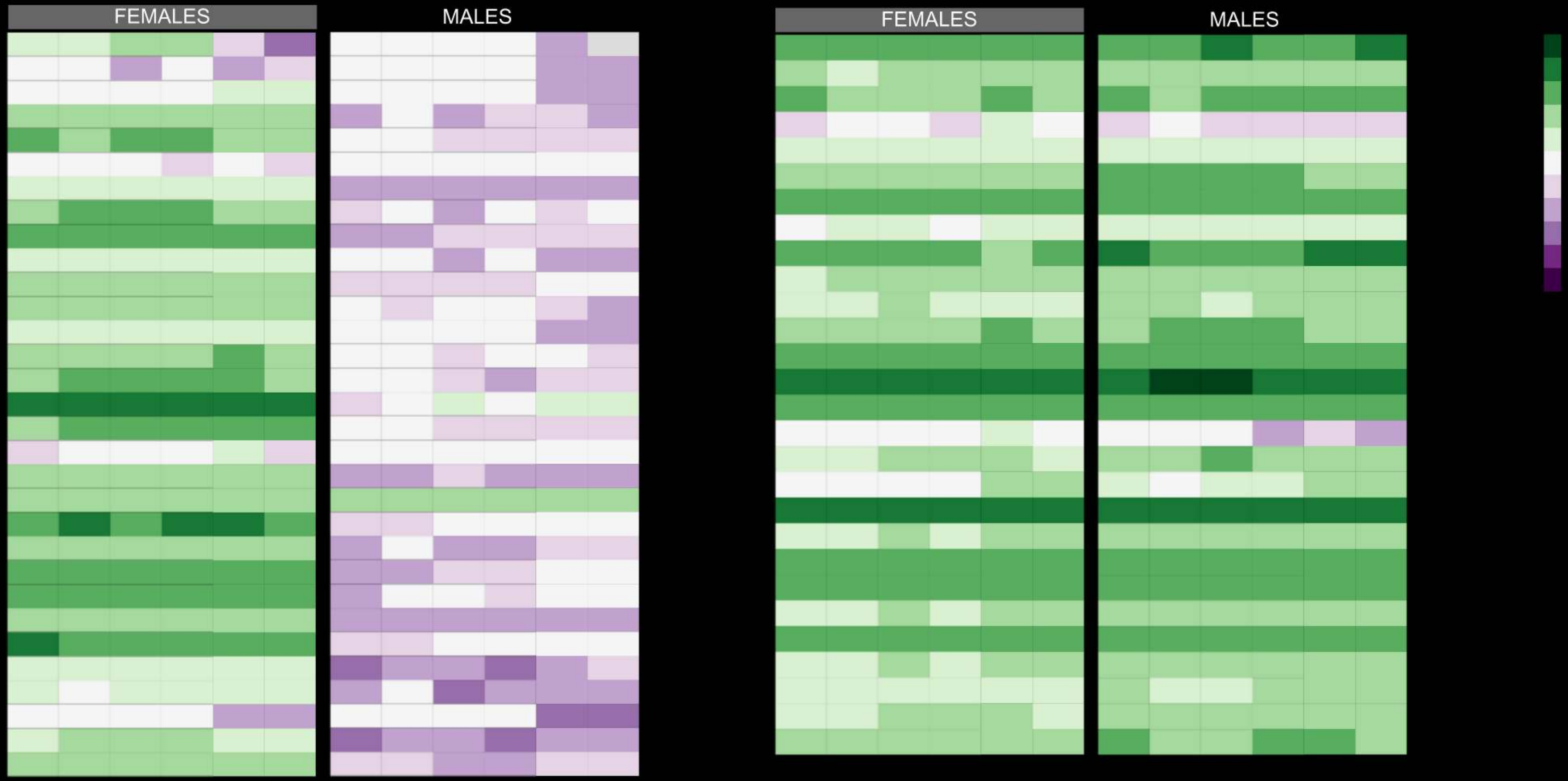

a
ZZ metaphase spread
2
Z
4
5
7
7
Z
4
5
3
1
2
1
3
6
ZW metaphase spread
W
2
7
6
5
6
4
3
7
4
2
Z
1
5
3
1
1
2
3
4
5
6
7
Z
W
NOR
b
1
2
3
4
5
6
7
ZW
unplaced
1
2
3
4
5
6
7
ZW
unplaced
S. mansoni genome version 5
1
2
3
4
5
6
7
W
Z
1
2
3
4
5
6
7
W
Z
S. mansoni genome version 9

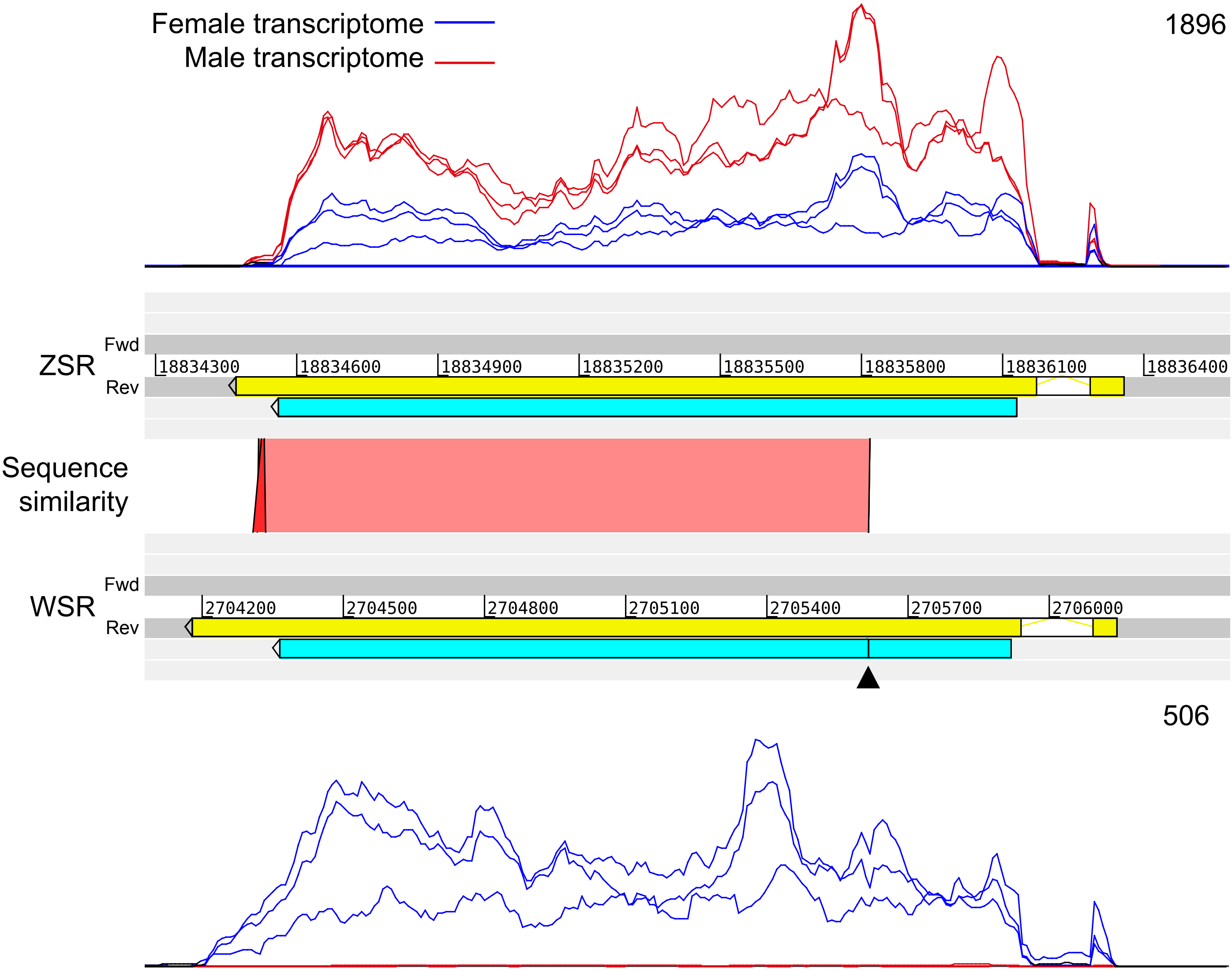

Z chromosome
100x
50x
0x
coverage
v5
100x
50x
0x
coverage
v9
PAR1
Z-specific Region
PAR2

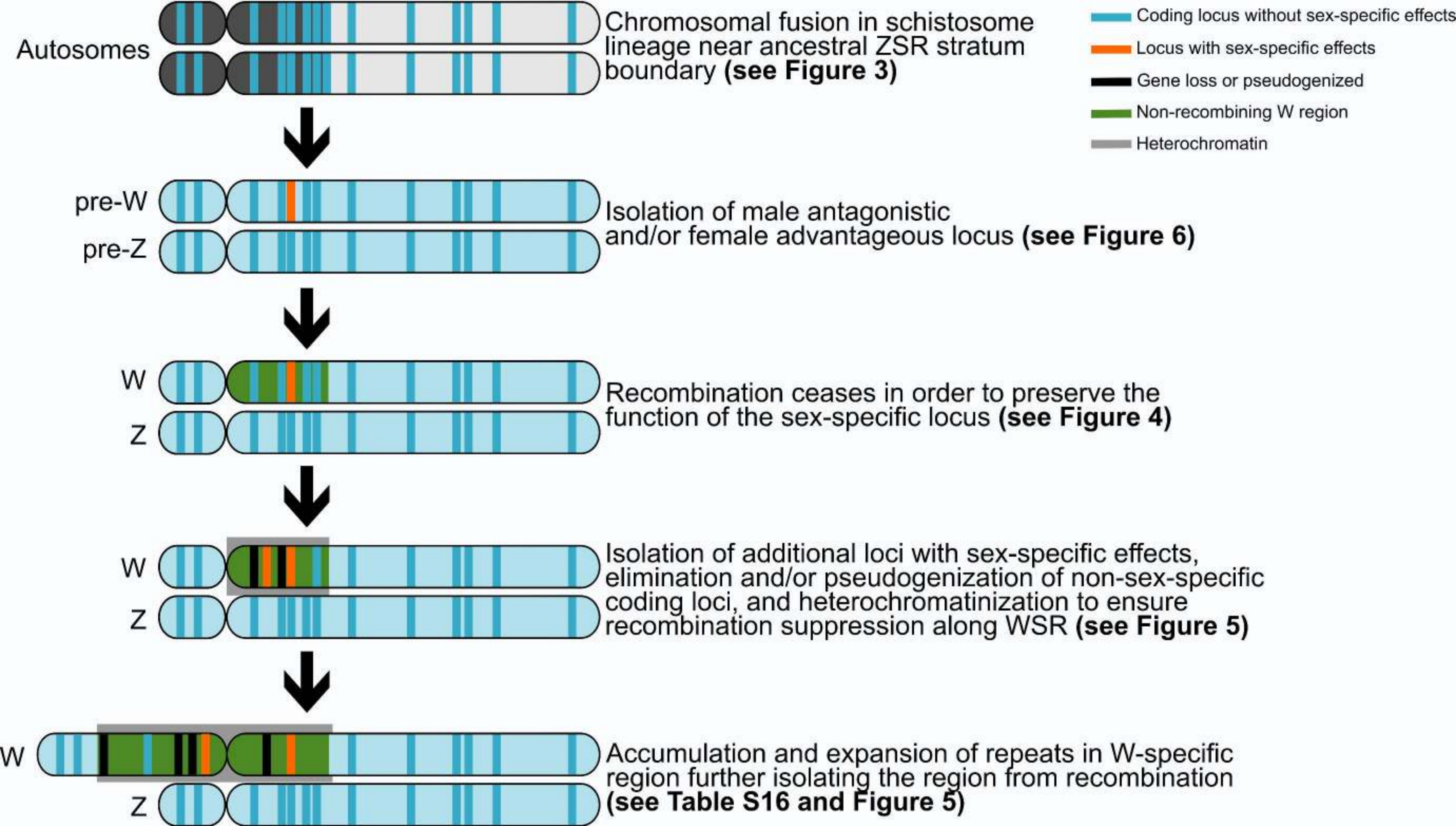

a
Echinostoma caproni
S. japonicum
S. haematobium
S. rodhaini
S. mansoni
ASIAN
AFRICAN
African
Ancestral
Asian
PAR1
ZSR1
ZSR2
PAR2
Z chromosome
b
African
Ancestral
Asian
PAR1
ZSR1
ZSR2
PAR2
Z chromosome
Tapeworm orthologs
1
2
3
4
5
6
7
8
9
Tapeworm E. multilocularis
1
2
3
4
5
6
7
Z
S. mansoni


a
b
c
d
e
f
π
0.00
0.01
0.02
0.03
0.04
1
2
3
4
5
6
7
0.000
0.005
0.010
0.015
0.020
0.025
PAR1
ZSR1
ZSR2
PAR2
-log10(π)
-1
-2
-3
-4


a True Z and W chromosomes
10.7Mb
24.7Mb
43.8Mb
86.7Mb
PAR1
232 genes
Z-specific Regions
351 genes
578 genes
PAR2
1,082 genes
Z
10.7Mb
56.5Mb
100.8Mb
PAR1
232 genes
W-Specific Region
31 genes
PAR2
1,082 genes
W
b Assembled ZSR and WSR
10.7Mb
24.7Mb
43.8Mb
Z-specific Regions
351 genes
578 genes
WSR
31genes
10.7Mb
16.7Mb
c
Z
W
smw24,25
smw22-23
smw06-21
C-band positive region contains both AT- and GC-rich DNA


FEMALES
MALES
FEMALES
MALES


Female transcriptome
Male transcriptome
1896
Fwd
ZSR
Rev
18834300
18834600
18834900
18835200
18835500
18835800
18836100
18836400
Sequence similarity
Fwd
WSR
Rev
2704200
2704500
2704800
2705100
2705400
2705700
2706000
506


Coding locus without sex-specific effects
Locus with sex-specific effects
Gene loss or pseudogenized
Non-recombining W region
Heterochromatin
Autosomes
pre-W
pre-Z
W
Z
W
Z
W
Z

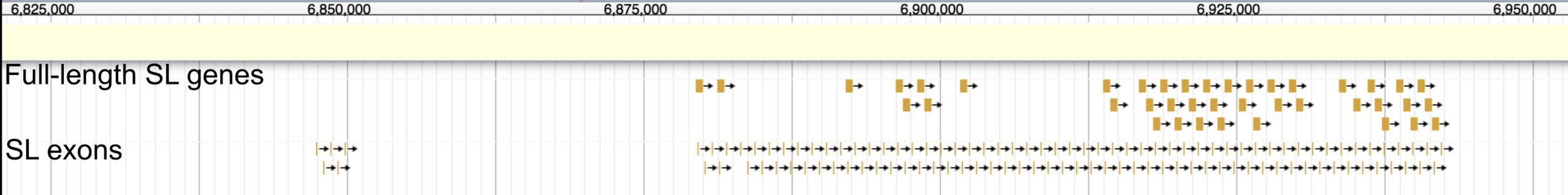
Chromosomal fusion in schistosome lineage near ancestral ZSR stratum boundary **(see Figure 3)**

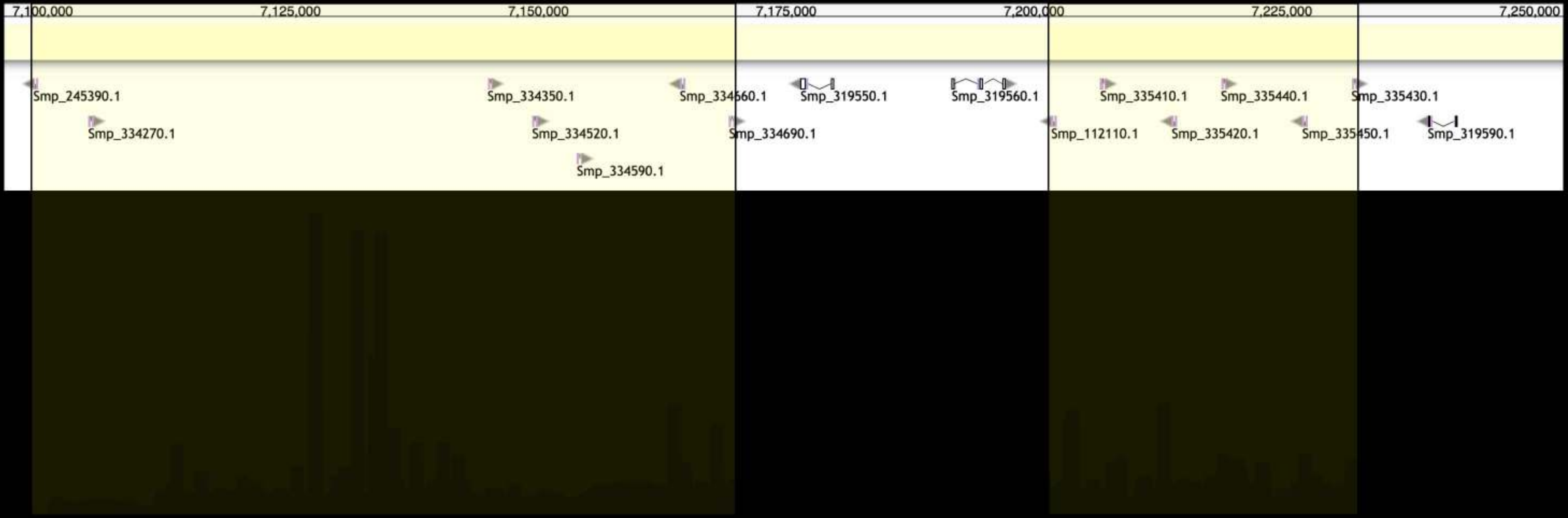
Isolation of male antagonistic and/or female advantageous locus **(see Figure 6)**

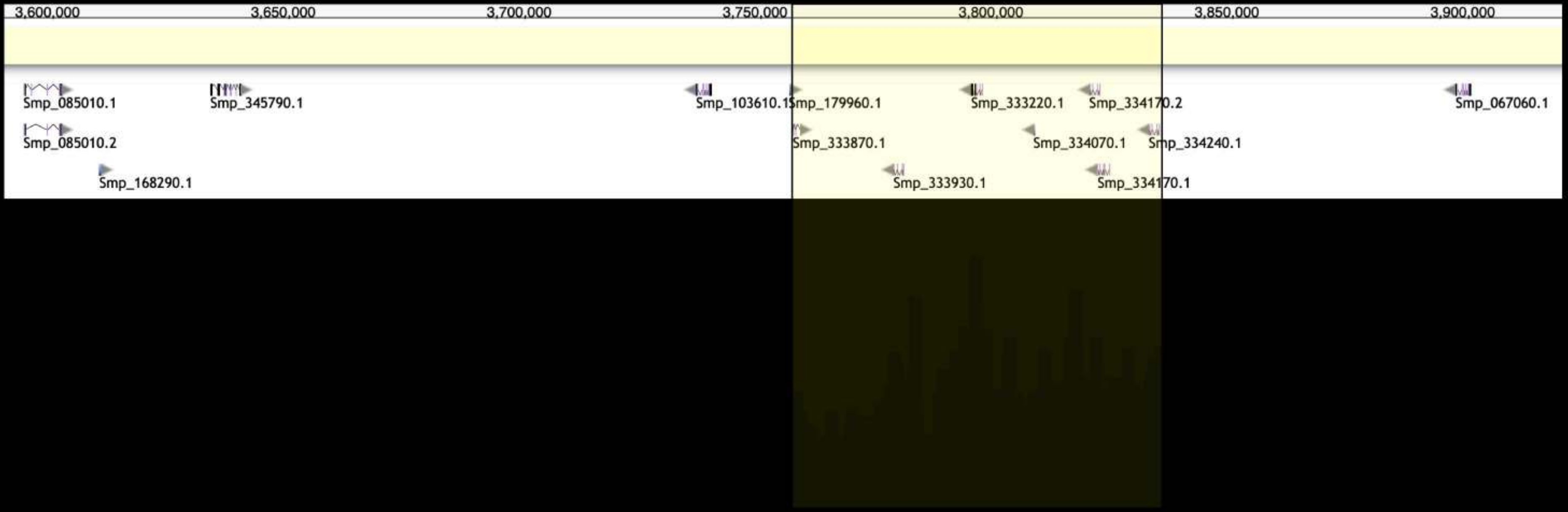
Recombination ceases in order to preserve the function of the sex-specific locus **(see Figure 4)**

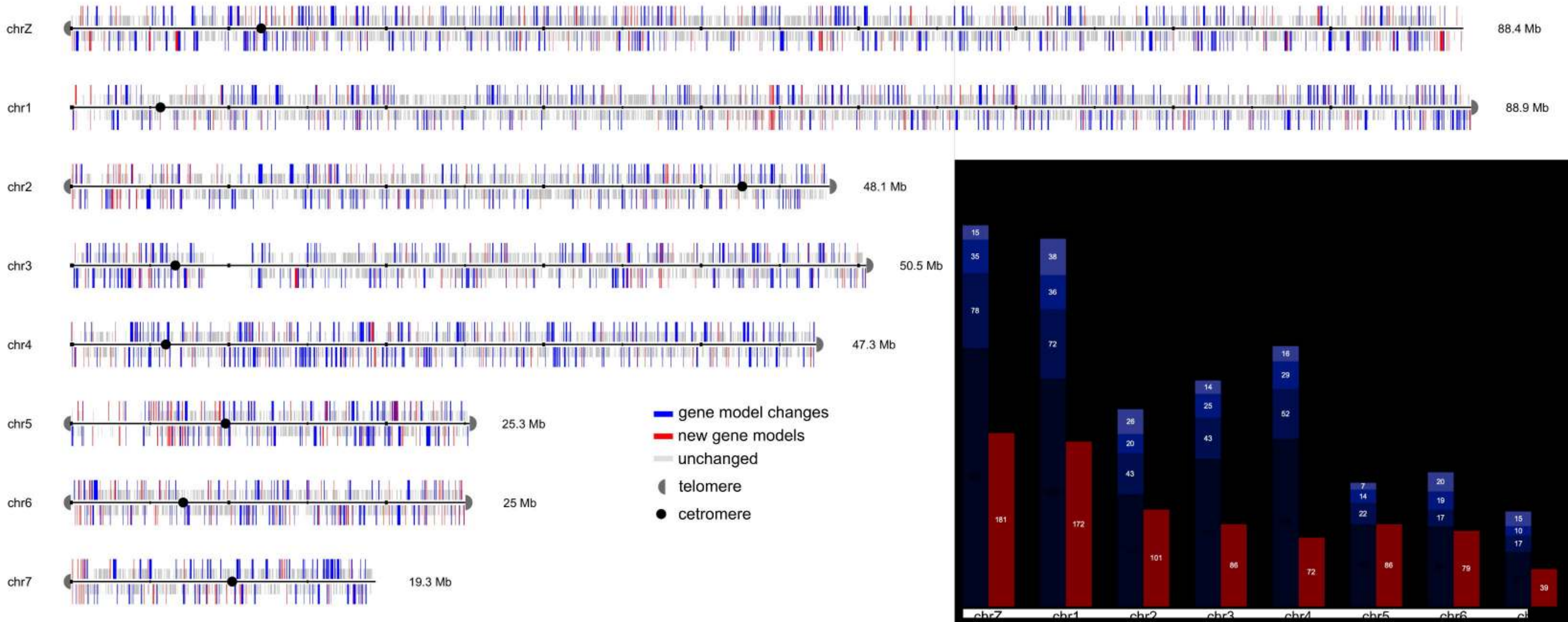
Isolation of additional loci with sex-specific effects, elimination and/or pseudogenization of non-sex-specific coding loci, and heterochromatinization to ensure recombination suppression along WSR **(see Figure 5)**

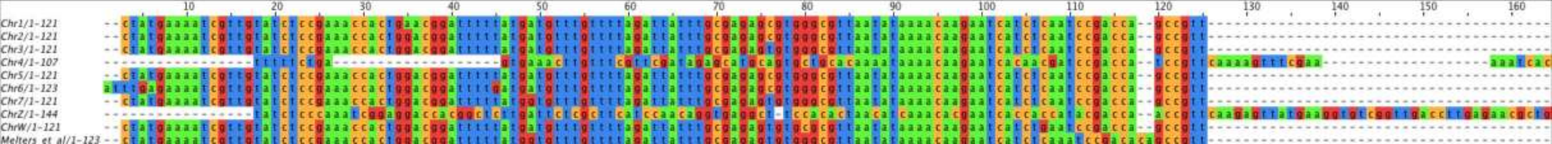
Accumulation and expansion of repeats in W-specific region further isolating the region from recombination **(see Table S16 and Figure 5)**

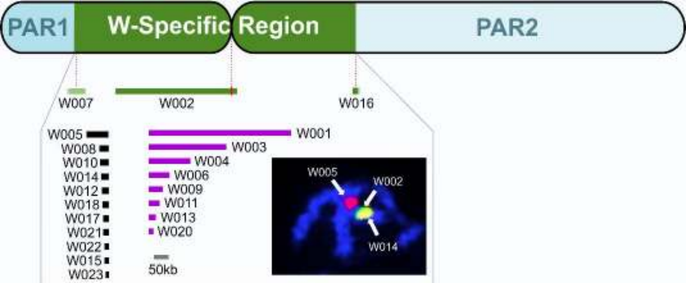

6,825,000
6,850,000
6,875,000
6,900,000
6,925,000
6,950,000
Full-length SL genes
SL exons


7,100,000
7,125,000
7,150,000
7,175,000
7,200,000
7,225,000
7,250,000
Smp_245390.1
Smp_334350.1
Smp_334560.1
Smp_319550.1
Smp_319560.1
Smp_335410.1
Smp_335440.1
Smp_335430.1
Smp_334270.1
Smp_334520.1
Smp_334690.1
Smp_112110.1
Smp_335420.1
Smp_335450.1
Smp_319590.1
Smp_334590.1

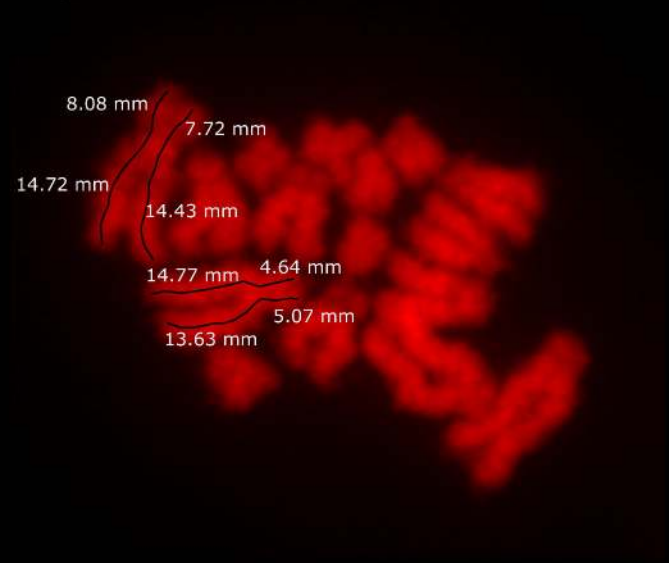

3,600,000
3,650,000
3,700,000
3,750,000
3,800,000
3,850,000
3,900,000
Smp_085010.1
Smp_345790.1
Smp_103610.1
Smp_179960.1
Smp_333220.1
Smp_334170.2
Smp_067060.1
Smp_085010.2
Smp_333870.1
Smp_334070.1
Smp_334240.1
Smp_168290.1
Smp_333930.1
Smp_334170.1


chrZ
88.4 Mb
chr1
88.9 Mb
chr2
48.1 Mb
chr3
50.5 Mb
chr4
47.3 Mb
chr5
25.3 Mb
chr6
25 Mb
chr7
19.3 Mb
gene model changes
new gene models
unchanged
telomere
cetromere
15
35
78
181
38
36
72
172
26
20
43
101
14
25
43
86
16
29
52
72
7
14
22
86
20
19
17
79
15
10
17
39
chrZ
chr1
chr2
chr3
chr4
chr5
chr6
ch

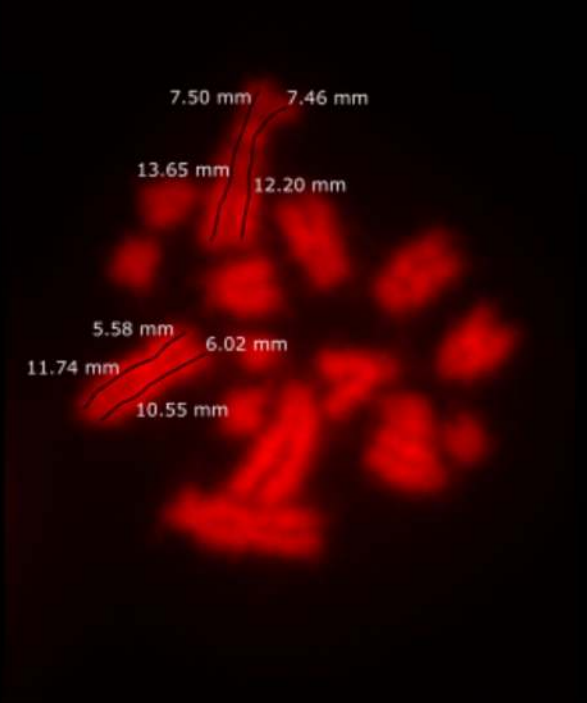

10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
Chr1/1-121
Chr2/1-121
Chr3/1-121
Chr4/1-107
Chr5/1-121
Chr6/1-123
Chr7/1-121
ChrZ/1-144
ChrW/1-121
Melters et al//1-123


PAR1
W-Specific Region
PAR2
W007
W002
W016
W005
W008
W010
W014
W012
W018
W017
W021
W022
W015
W023
W001
W003
W004
W006
W009
W011
W013
W020
50kb
W005
W002
W014

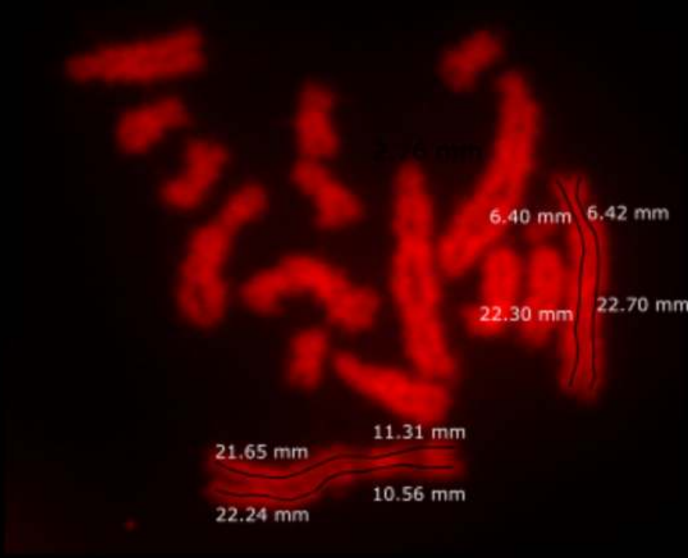
"Figure 1"



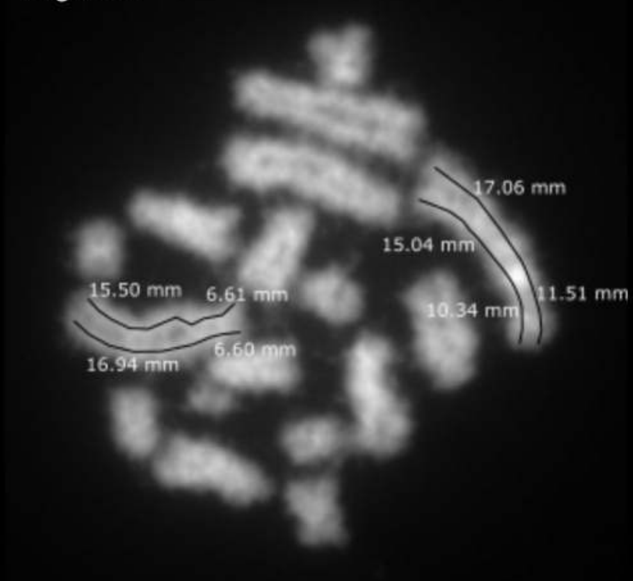
"Figure 5"



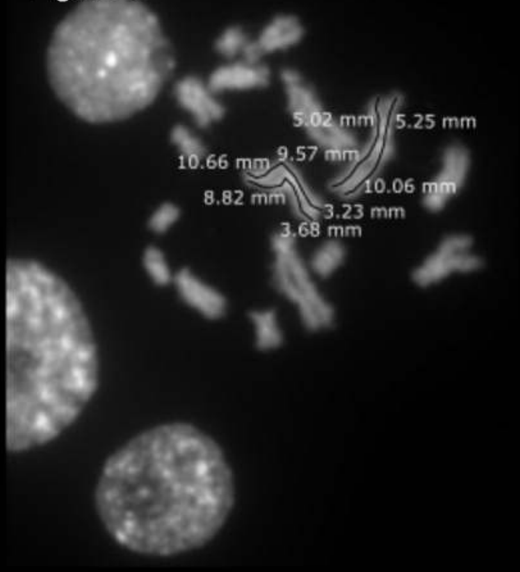
"Figure 7"

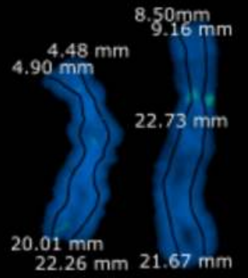


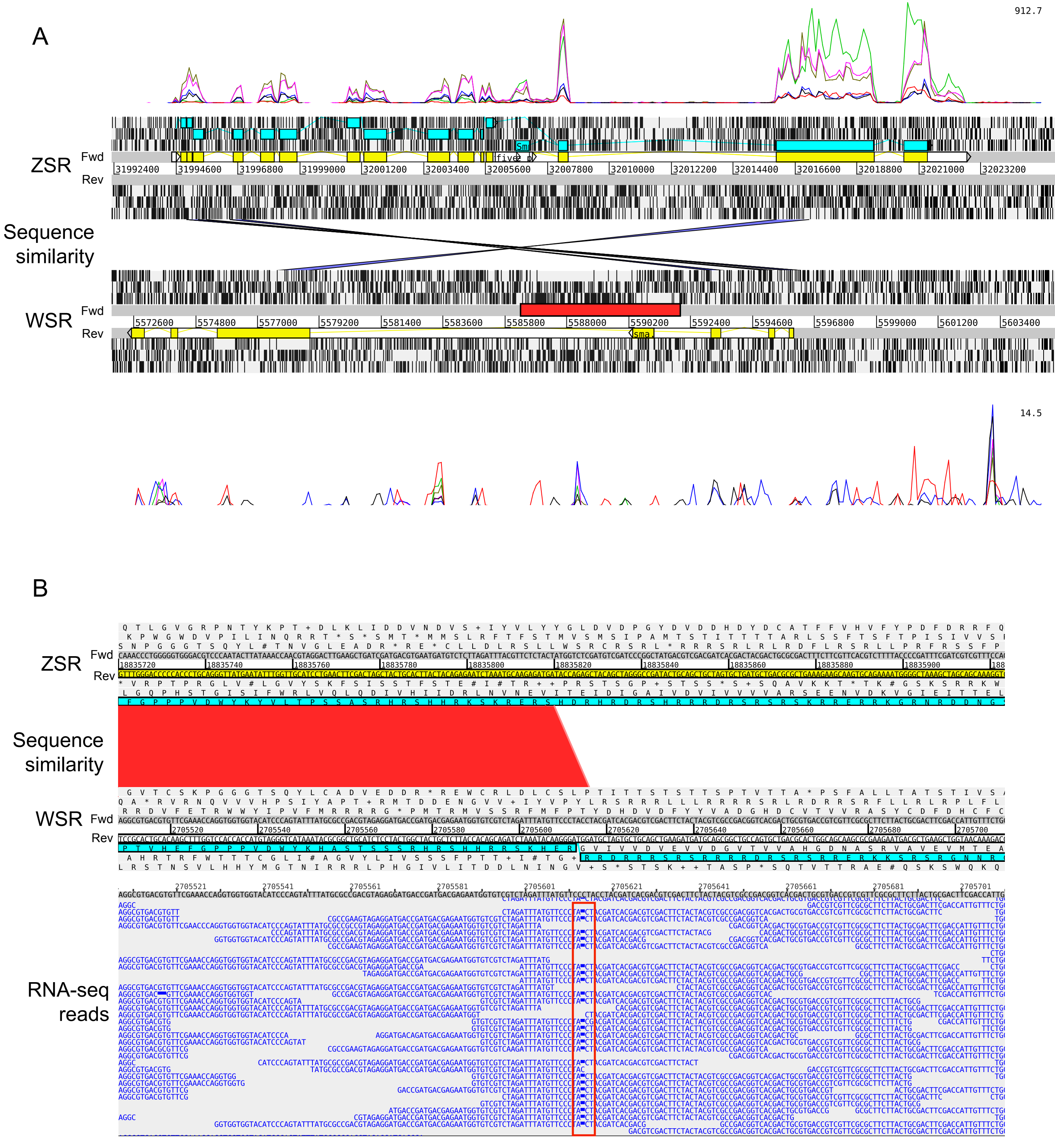
"Figure 8"



'Figure 9"






A
912.7
ZSR
Fwd
Rev
Sm
five p
31992400
31994600
31996800
31999000
32001200
32003400
32005600
32007800
32010000
32012200
32014400
32016600
32018800
32021000
32023200
Sequence similarity
WSR
Fwd
Rev
5572600
5574800
5577000
5579200
5581400
5583600
5585800
5588000
5590200
5592400
5594600
5596800
5599000
5601200
5603400
sma
14.5
B
ZSR
Fwd
Rev
18835720
18835740
18835760
18835780
18835800
18835820
18835840
18835860
18835880
18835900
Sequence similarity
WSR
Fwd
Rev
2705520
2705540
2705560
2705580
2705600
2705620
2705640
2705660
2705680
2705700
RNA-seq reads
2705521
2705541
2705561
2705581
2705601
2705621
2705641
2705661
2705681
2705701