TEXAS BIOMEDICAL RESEARCH INSTITUTE

The S. mansoni genome v6

Linkage based assessment of the new assembly

Frédéric Chevalier – Winka Le Clec'h – Marina McDew-White – Tim ANDERSON

8/9/2016

Contents

Context

In July 2016, the Sanger Institute shared with us the v. 6 of *Schistosoma mansoni* genome. This assembly was generated using data from PacBio sequencing and optical mapping to further improved the assembly of the Illumina reads.

We and our collaborators from the IHPE laboratory (Université de Perpignan Via Domitia, France) conducted a total of 4 genetic crosses in order to generate F2 progeny. We used marker segregation in these crosses (a) to identify misplaced part of the *S. mansoni* genome v. 5 and (b) to benchmark the new v.6 assembly and (c) to provide additional information to further improve the v6 assembly.

This document describes the crosses, the methodologies used for generating the genetics maps and the list of markers that can be used to further refine the assembly.

I. Genetic crosses: origin, preparation and data processing

1. Origin of the schistosome populations

We used four populations of distinct origins:

- SmBRE: Brazilian population from Recife
- SmLE: Brazilian population from Belo Horizonte
- SmEG: Egyptian population
- SmOM: Omani population

Four crosses were performed and individuals from F2 progeny were selected:

- Cross 1: SmLE_1 x SmEG_4, 42 F2 individuals
- Cross 2: SmBRE_4 x SmLE_19, 82 F2 individuals
- Cross 3: SmOM (F0Hh1) x SmOM (F0fh1), 65 F2 individuals
- Cross 4: SmOM (F0Hs3) x SmOM (F0fs13), 71 F2 individuals

2. Library preparation

For each cross, we prepared exome capture libraries (using SureSelect from Agilent) using the F0 parents, the F1 and the F2 progeny. We obtained DNA from F0s and F1s (except cross 1) from adult worms without whole genome amplification (WGA). For Cross 1 we obtained DNA of F1s and F2s following WGA (GenomiPhi V2 DNA Amplification Kit) of single miracidia preserved on FTA cards. DNA of F1s and F2s of all the other crosses was obtained using WGA of cercariae previously extracted with Chelex. We sequenced exome libraries on HiSeq 2500 (16 to 24 sample by lane).

3. Bioinformatic processing

Before performing alignment, chromosomes and scaffolds of the v. 6 assembly were renamed to avoid rewriting parts of our scripts. The renaming was done with the following command:

```
sed -r "s/SMv6_([0-9])/SC_\1/g ; s/SMv6_C/Chr_/g ; s/SMv6_(ZW|W)/Chr_W/g ;
s/([0-9])_([0-9])/\1.un.SC_\2/g ; s/SMv6_m/m/g" SMAN_V6.fa >>
SMAN_V6_renamed.fa
```

This command removes SMv6_ prefix from scaffolds and chromosomes, adds SC_ prefix to unassigned unassembled scaffolds, renamed ZW and W in Chr_W, and adds un.SC to assigned unassembled scaffolds (chr. 1 and chr. 5). These new names are used in the marker file attached to this report.

Sequencing reads were checked for quality using FastQC, aligned against the *S. mansoni* reference genome (v6 renamed) using BWA (v0.7.12) and SAMtools (v1.2). Realignment around indels was performed using GATK (v3.3.0) and PCR duplicates were marked with Picard (v1.136).

We called variants using UnifiedGenotyper from GATK (v3.3.0) independently for each cross. Genotypes and total read depth of each variant for each sample were extracted using vcftools (v0.1.14).

II. Genetic maps of *S. mansoni* genomes

We constructed genetic maps for each cross using R/qtl package. Our R script first filters variants on five criteria:

- Only sites supported by 20 reads in F2s were selected
- Only sites with less than 20% of missing data in F2s were selected
- Only sites with alternative fixed alleles in the parents were selected
- Only sites showing a Mendelian inheritance between F0s and F1s were selected
- Only bi-allelic sites were kept

Here are the variant numbers obtained by cross:

- Cross 1: 3,034
- Cross 2: 12,802
- Cross 3: 6,133
- Cross 4: 7,914

We transformed variant tables for each cross to a suitable format for the R/qtl package. All scaffolds were considered as a single chromosome named SC. For each cross, LOD scores and recombination fractions between pairwise combinations of markers were computed using the `est.rf` command and plotted using the `plot.rf` command from R/qtl. An example of genetic map based on LOD scores and recombination fractions for the genome v. 5 and v. 6 from the same cross is showed on figure 1. Genetic maps related to the other crosses are attached to this report (see Genetic_maps.zip).

V. 6 is substantially improved relative to v. 5 (see genetic maps drawn from cross 2 – Figure 1). All chromosomes showed almost perfect assembly except some very small parts (see chromosome 1 and ZW). The locations of unassembled scaffolds in the genome is also shown.

III. Assembly improvements from our genetic maps

The LOD scores used to draw the genetic maps can also be used to map the positions of the misplaced or unassembled scaffolds. By finding the most strongly linked markers (the combination showing the highest LOD score), we can find the likely position where these markers should be inserted.

To obtain an informative set of markers pairs, we selected pairs that showed the highest LOD scores then we excluded pairs on two criteria:

- Redundancy: when the same pair shows up twice.
- Proximity: when two markers were neighboring or when they were less than 1 Mb apart.

This was done independently for each cross then data were merged to obtain a final table (SNP.reposition.table_crosses.csv). This table contains six columns:

- source_scaffold: scaffold (or chromosome) on which the marker is identified
- marker_pos: position of the marker on the source scaffold
- target_scaffold: scaffold (or chromosome) to which the source marker is linked
- linked_marker_pos: position to where the source marker is linked
- lod_score: LOD score of the pair
- cross: from what cross the pair is identified

For instance, line 4 of the table reads:

```
Chr_7    2532900    SC_018          50296    14.956836438383       SmBRE4_m x SmLE19_f F2A
```

This means that the marker at position 2,532,900 on chr. 7 is linked to position 50,296 on scaffold 18 with a LOD score of 14.96 and this information come from the cross 2. Hence, the marker on scaffold 18 is linked to chromosome 7 and most probably belongs on this chromosome.

By having contiguous markers that map to the same position on the target chromosome, we can determine the location of the scaffold (or pieces of scaffolds) in the chromosome.

This table also contains information about misplaced parts of the chromosomes (e.g., see line 40 of the table where a chr. 6 marker should be relocated to chr. 4). In this case, there is probably a very small piece of the chromosome that is misplaced. Bioinformatic criteria could then be used to determine the precise insertion location and size of the misplaced fragment.

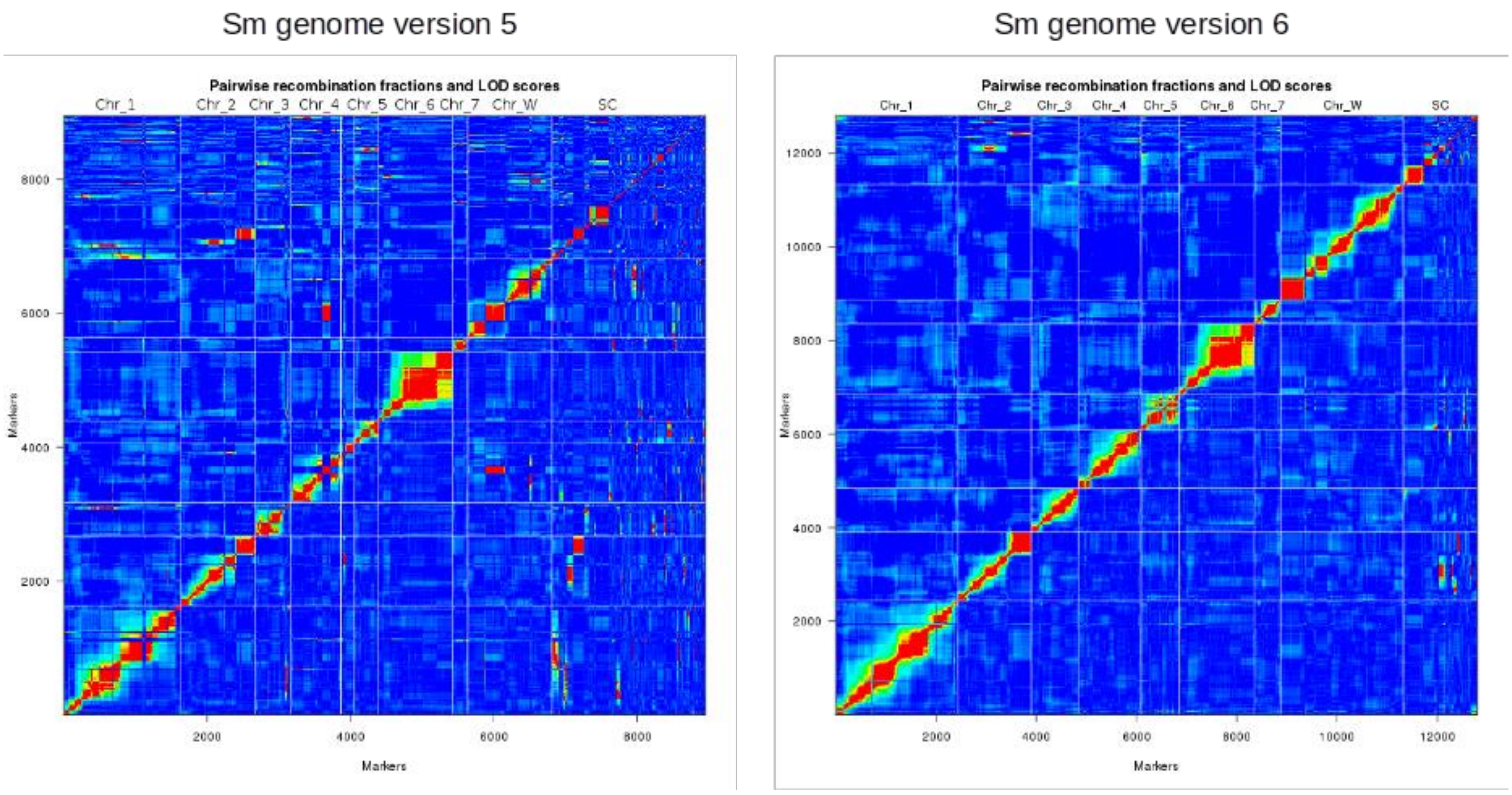

Figure 1 - Genetic maps of the v. 5 and 6 *S. mansoni* genomes. The genetic maps shown are for cross 2. V6 maps for the other three crosses are provided in the appendix. Chr: chromosome; SC: unsassemble scaffolds (which can be assigned to chromosome or not). NB. The v.5 map as fewer markers, because only exon markers are used, while the v.6 map also utilises markers outside the exomes. We will send the full map for v. 5 when this has been completed, but we don't expect this to alter the main result.

In some cases, a marker is relocated quite close by on the same chromosome (e.g., see line 10 of the table). This is probably an artifact due to missing data (as the value of the LOD score can be influenced by the number of individuals). Such reassignments should be treated with caution or ignored.

Finally we examined the number of relocations and the contribution of each cross. We counted 200 unique scaffold/chromosome relocations leading directly or indirectly to the integration of these scaffolds into chromosomes. These include 66 (parts of) scaffolds relocated to other scaffolds, 109 (parts of) scaffolds relocated to chromosomes and 25 chromosome fragments (including assigned scaffolds) relocated to other chromosomes (or assigned chromosomes). These relocations account for a total length of 29,2 Mb (~8% of the genome) assuming they are entirely integrated in the genome. Evidence for most relocations comes from just one of the four crosses; however, many relocations are supported by linkage evidence from multiple crosses (Figure 2). Figure 3 shows that cross 4 contributes to half the relocations and cross 2 for a third. By increasing the number of crosses we increase the number of relocations.

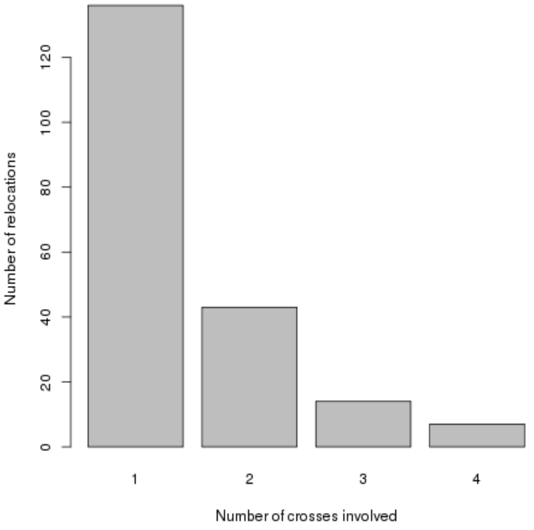

Figure 2. Number of relocations supported by one, two three or all four crosses

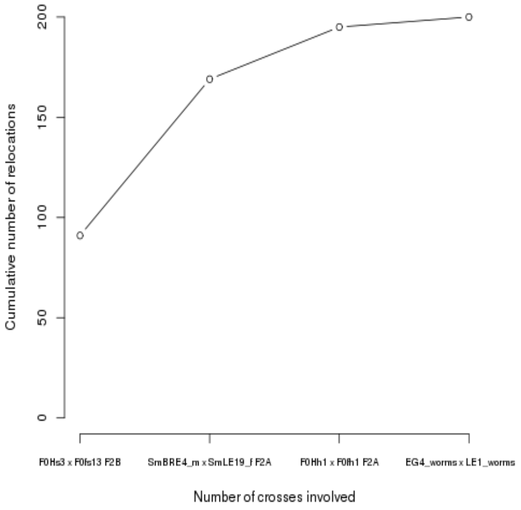

Figure 3. Cumulative numbers of relocations inferred from the four crosses

Conclusion

Our linkage-based benchmarking of the v. 6 *S. mansoni* genome strongly validates the accuracy of the new assembly. PacBio and optical mapping greatly improved the assembly relative to v. 5. In addition, linkage information allowed us to identify errors in the assembly and assign many unaligned or misaligned scaffolds to chromosomes.