

File S5

Supplementary material

Method for haplotype block reconstruction

The first step in the construction of our genetic map is to remove sites where very few of the sperm have SNP calls. When very few sperm samples were sequenced at a site, it becomes difficult to detect recombination events between a pair of sites since they may be no sperm sequenced at both sites. Once these sites are removed, the remaining polymorphic sites that are adjacent to each other in the reference assembly are clustered into blocks that show no evidence of recombination, and then these blocks are phased in order to minimize the number of recombination events. Whether a recombination event occurs between two sites can be easily checked by counting the number of haplotypes that occur at each site. If only two haplotypes occur, then there is no evidence that a recombination event occurred between these sites. Then they are joined together into a longer haplotype block. Each haplotype block is extended to include all adjacent sites for which there is evidence of only two haplotypes. When a third haplotype is detected, a new block is created and the procedure is repeated. Once all the haplotype blocks are established, they are phased to minimize the number of recombination events occurring between adjacent haplotype blocks. This can be easily accomplished since each haplotype block has only two potential phases, and we assume that each haplotype block is in the phases that gives a lower number of recombination events with the upstream block. Since haplotype blocks are constructed under the assumption that sites adjacent to each other in the reference assembly are separated by very little genetic distance, it is not expected a large number of recombination events occurring between adjacent blocks. When this does occur, we can mask the downstream haplotype block which has an excessive number of recombination events and

ignore this haplotype block for phasing purposes. Finally, the adjacent haplotype blocks that are not affected by recombination events need to be placed into the same parental or maternal phases so that the MST software, which assumes that the parental and maternal identity of markers are known, can be used to construct a genetic map.