EXTENDED EXPERIMENTAL PROCEDURES

Details of BAC selection and sequencing / Single-haplotype iterative mapping and sequencing

In the first phase, we aimed to identify non-overlapping BACs for seed contigs. 8 STSs (364, 365, 372, 373, 379, 383, 384, 386, File S1) were designed using the sequence of one typical BAC (RP24-507D23) which was available before the project started. All BACs with end sequences matching RP24-507D23 were tested with the appropriate STSs. We identified 121 BACs containing unique variants, which were selected for sequencing, and were used to seed the initial contigs. In the second phase, we aimed to expand contigs. Unique SFVs were identified at the end of each contig, and all BACs with end sequences best matching this contig and which would extend the contig were tested for the SFV. Out of clones matching these criteria, we selected for sequencing one clone with at least 30 kb overlap and which would add most new sequence. In some cases, the overlap proved not to be real, in which case the BAC was treated as a new contig, and SFVs were again identified. This process was iterated until no more new BACs could be found. The number of contigs, and therefore points of extension, initially increased to 175 in the first round of iteration, but then gradually coalesced to 19 final contigs. Both phases of BAC selection depended heavily on availability of high quality BAC end sequences.

Calculation of sequence accuracy

The initial error rate estimated for BAC sequencing and assembly is 1 in 50 kb. However, as 65% of our sequence is covered redundantly by two BACs, we were able to identify and resolve all discrepancies in redundantly covered regions, so that the error rate for

these regions is zero. Therefore, the final error rate is estimated to be 0.35 * 1/50000 + 0.65 * 0 = 1/143000, or 1 in 143 kb.

Sequence assembly and gap-filling

To create a model assembly for analyses, the 19 sequence contigs were ordered and oriented by RH mapping, and joined with estimated gap sizes. Gaps that fell within the two regions of outstanding identity (7 Mb of 99.999% identity and 4.5 Mb of 99.995% identity) were filled with sequence from the corresponding segment of the other almost-identical repeat unit.

Estimation of degree of repetitiveness of the chromosome

We walked across the sequence with a 500-bp window of sequence, with step 1. We determined the number of times each 500-bp word appears elsewhere in the sequence. Therefore, if a word appears in the sequence *n* times, it appears an additional *n-1* times. We took the average of the additional number of times each 200-bp word appears as a measure of repetitiveness of the sequence.

Estimation of total Rbmy gene copy number and array size

We obtained about 370 kb of sequence containing 10 intact ORFs for *Rbmy*. As sequence obtained across the *Rbmy* array was incomplete, we estimated the total *Rbmy* gene copy number and array size as follows. We performed hybridization with probes for *Rbmy* on the RP24 library, and identified 38 *Rbmy*-positive BACs. The RP24 library has an average clone size of 155 kb, and average genome coverage of 10.8X or Y chromosome coverage of 5.4X. Thus, we calculated the total size of the *Rbmy* array to be approximately 1.1 Mb. Assuming an average repeat unit size of 37 kb, and that each

repeat unit contains an intact ORF for *Rbmy*, we estimate that there are 30 copies of *Rbmy*.

Identification of genes and transcription units

Srsy

We consider *Srsy* to have protein-coding potential, as it has a sizable open reading frame translating into 364 amino acids. We do note, however, that it has relatively low levels of transcription in the tissue panel we examined, compared with typical protein-coding genes.

LOC102642016 and *LOC102641986*

LOC102642016 (Accession no. KJ780361) and *LOC102641986* (KJ780362) were initially identified from a Celera alternate assembly unplaced contig, assembled from whole genome shotgun sequence from mixed Mus musculus strains (NW_001034423.1). Markers for both genes were strictly male specific when tested on C57BL6 male and female genomic DNA, leading us to conclude that both genes were located on the Y chromosome. We were unable to identify both genes within the CHORI-36 and RPCI-24 C56BL6/J BAC libraries, although we were able to identify them in BAC 45G04 from CHORI-29. Nevertheless, we confirmed the integrity of both genes in C56BL/6 by sequencing from C57BL/6Tac adult testis cDNA. We additionally determined the location of both genes by DNA FISH and RH mapping. DNA FISH using BAC probe CH29-45G04 shows that both genes are located on the distal tip of the Y chromosome short arm. RH mapping shows that both genes are linked to the RH marker located most distally on the mouse Y short arm. Because the Celera contig containing the two genes

also contains telomeric sequence, we infer that both genes are on the distal tip of the short arm, adjacent to the telomere.

Transcription units AK006152 and KC170991

KC170991 and *AK006152* are two related testis transcripts which are 94% identical. We identified *KC170991* in our genomic sequence, but not *AK006152*, which was previously described. However, we were able to assemble and identify both transcripts from 454 sequence reads of cDNA generated from C57BL/6JTac adult testis. RH mapping confirms the presence of *AK006152*, and places it at the distal tip of the short arm, close to *KC170991*.

Sequence differences between C57BL/6J and C57BL/6Tac

Discrepancies between the sequence, which was obtained from C57BL/6J, and the RH map, which was obtained from C57BL/6Tac, suggested that approximately 6 Mb of sequence that was present in one copy in C57BL/6J was present in two copies in C57BL/6Tac. The most parsimonious explanation for the difference was that a deletion had occurred in C57BL/6J, following a prior duplication in the common ancestor of the two strains. We confirmed this using quantitative PCR to measure the copy number of the putative deleted region compared to regions identical between the two strains.

Electronic fractionation plot

Intrachromosomal similarity, or percent identity to other MSY sequences, was determined by using custom Perl code that used BLAST to compare all 5 kb sequence segments, in 2 kb steps, to the entire remainder of the MSY sequence.