Supplementary Notes

Note 1: Clone selection

We used three strategies to identify clones for sequencing. The first strategy utilized the human Y chromosome sequence as a scaffold. Initially, 23 pools of overgo probes spanning the entire human MSY were used to screen the CHORI-251 male chimpanzee BAC library. This approach enabled the assembly of nearly the entire X-degenerate region[1] and identified 54 BACs from the ampliconic region. The remaining 106 ampliconic and non-X-degenerate single copy BACs were identified using one of two other strategies: 1) chromosome walking using end sequences from sequenced chimpanzee BACs and 2) identification of Y chromosomal chimpanzee-specific sequence in the whole genome shotgun data[2]. A fosmid library, constructed from the same chimpanzee as CHORI-251, was utilized to fill gaps. There are five BACs in the ampliconic sequence that are from the RPCI-43 library, which was constructed from a closely related animal.

Our finished clone assembly consists of eight large contigs. However, there are eight small gaps within these contigs – two in the X-degenerate region[1] and six in the ampliconic contigs. Sequence data from homologous regions in human indicate that the X-degenerate gaps are approximately 14 and 69 kb in size and each of the ampliconic gaps is likely less than 30 kb in size. Interphase FISH results are consistent with these estimates (Supplementary Fig. 4).

Note 2: Estimating completeness of clone-based chimpanzee MSY sequence

We obtained 384,310 454-sequencing reads (with an average read length of 99 bp) using whole-genome amplified flow-sorted chimpanzee Y chromosomes as starting material. We enriched for high quality reads using the following criteria[3]: 1) removal of

sequence reads that contained one or more unknown bases (or Ns), 2) removal of sequences that were shorter than 80 bp or longer than 105 bp, 3) removal of sequences that had average quality scores of less than 25. The remaining 199,099 reads were then masked for repeats using RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). Only reads that were <50% repeat-containing were used for further analysis – 102,659 in total. These reads were used to BLAST the GenBank non-redundant database. Only the top match for each read was recorded. Sequences whose best match was to a bacterial or fungal genome were dismissed as environmental contaminants (4,336 in total). The remaining 92,281 hits were categorized as follows: 1) > 94% match to chimpanzee Y chromosome, 2) > 94% match to human or chimpanzee X chromosome or autosome (note: GenBank contains limited autosomal or X-chromosome sequence from chimpanzee), or 3) < 94% match to human or chimpanzee genome or best match to non-human or non-chimpanzee genome. The total number of hits for each category are shown below.

Target	Number of hits	% of Total
Y chromosome	58,302	59.3
X + autosomes	39,315	40.0
Unknown	628	0.639

Sequences that match with > 94% identity to autosomes or the X-chromosome are assumed to be derived from contaminants in the flow-sort. The amount of autosome and X chromosome contamination is within the range expected because of the extensive overlap of the Y-chromosome peak with the chromosome debris in this particular flow karyotype (not shown). We cannot exclude the possibility, however, that the these matches represent very recent transpositions to the chimpanzee Y chromosome. If this were the case, then some of the autosome hits would be physically clustered. There are no substantially sized regions on the autosomes or the X chromosome that are overrepresented in the 454 sequence, however.

The sequences classified as unknown could represent chimpanzee Y chromosome sequence missing from our assembly. Even if we assume that all of the unknown sequence is actually Y chromosome sequence, then we are missing only about 1.06% of the sequence (z-test 95% C.I. 0.983-1.15). This is likely an overestimate, however, because a significant proportion of the reads are likely chimpanzee-specific autosomal or X-chromosomal contaminants not represented in GenBank.

Note 3: Comparative sequence analysis

To calculate the amount of sequence in each species that is not present in the other, the data from the chimpanzee vs. human MSY dot-plot (Fig. 2) was used. All gaps in the plot that were greater than 20 kb in size were totaled on each axis. The amount of unshared sequence for each species was taken as the gap total. The amount of shared sequence was then calculated by subtracting the gap total from the total length of sequence.

To calculate the average lengths of alignable intervals between chimpanzee and human in X-degenerate vs. ampliconic sequence, the data from the chimpanzee vs. human MSY dot-plot (Fig. 2) was used. Alignable intervals were defined as segments of sequence alignments with 100% identity within 200 bp windows that are not interrupted by an inversion breakpoint or a gap of > 50 kb.

[1] Hughes, J. F. et al., Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437** (7055), 100 (2005).

[2] The Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437** (7055), 69 (2005).

[3] Huse, S. M. et al., Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8** (7), R143 (2007).