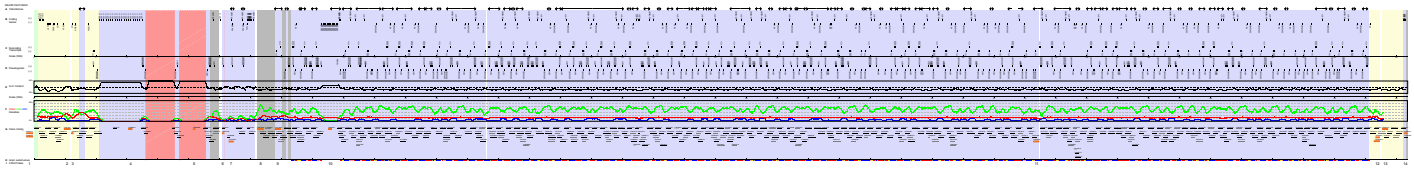
SUPPLEMENTAL MATERIAL

Sequence analysis in *Bos taurus* reveals pervasiveness of X-Y arms races in mammalian lineages

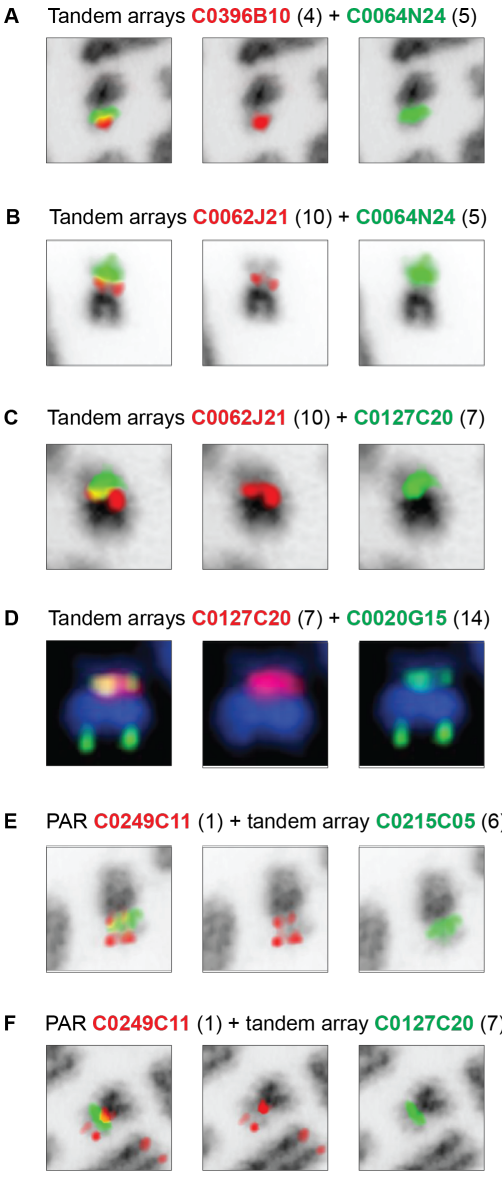
Jennifer F. Hughes, Helen Skaletsky, Tatyana Pyntikova, Natalia Koutseva, Terje Raudsepp, Laura G. Brown, Daniel W. Bellott, Ting-Jan Cho, Shannon Dugan-Rocha, Ziad Khan, Colin Kremitzki, Catrina Fronick, Tina A. Graves-Lindsay, Lucinda Fulton, Wesley C. Warren, Richard K. Wilson, Elaine Owens, James E. Womack, William J. Murphy, Donna M. Muzny, Kim C. Worley, Bhanu P. Chowdhary, Richard A. Gibbs, David C. Page

Note: Supplemental Tables and Supplemental Files are separate from this file and can be downloaded separately.

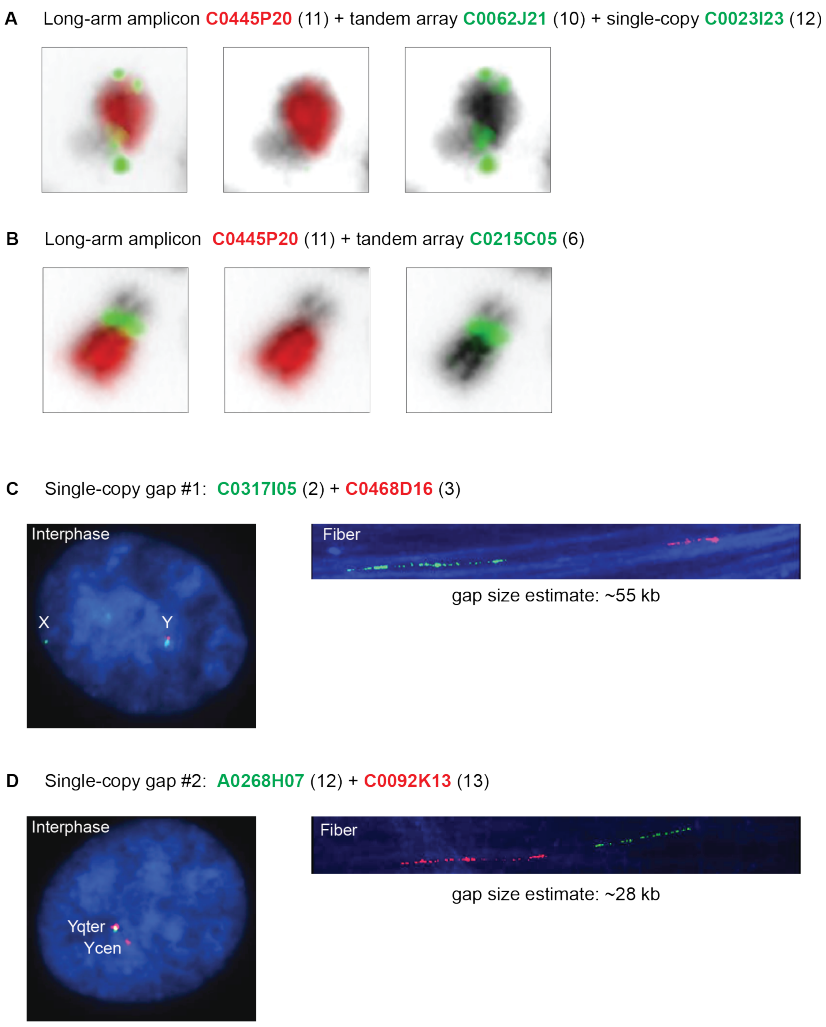
SUPPLEMENTAL FIGURES



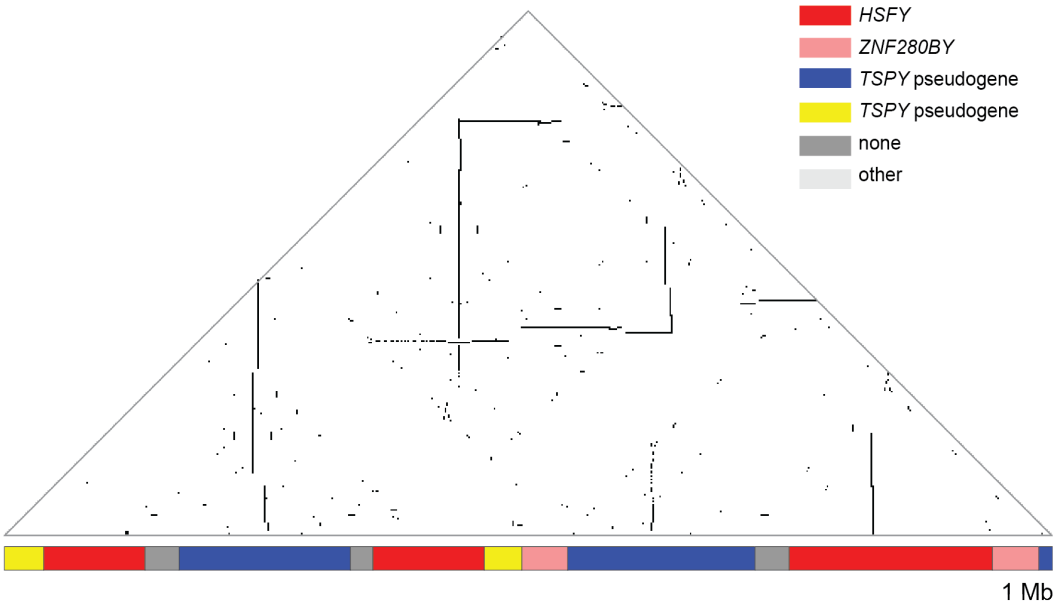
Supplemental Fig. S1. Annotated sequence of the bull MSY.
(Full size figure available at http://pagelabsupplement.wi.mit.edu/BP2020/map.pdf). Background colors represent sequence classes: pseudoautosomal (green); ancestral single-copy (yellow); ampliconic (blue); other (gray); and heterochromatic (pink stripes). Gaps shown in white. Bottom bar of red, pink, blue, and yellow stripes represents the substructure of each long arm repeat unit. All sequence features and BACs drawn to scale. (A) Black arrows represent large palindromes. (B) Positions of all intact protein-coding genes. Plus (+) strand above, minus (-) strand below. (C) Positions of noncoding transcripts. (D) Positions of pseudogenes. (E) G+C content (%) calculated in a 100-kb sliding window with 1-kb steps. (F) SINE, LINE, and ERV densities (%), calculated in a 200-kb sliding window with 1-kb steps. (G) Sequenced MSY BACs. Each bar represents the size and position of one BAC clone, labeled with the library identifier. BAC clones with prefix "C" are from the CHORI-240 library. BAC clones with prefix "A" are from the BTDAEX library. Black bars represent BAC clones used for the tiling path, dark gray bars represent other finished BAC clones, light gray bars represent unfinished BAC clones. Bars outlined in orange indicate BACs used as FISH probes. (H) Major repeat units in bull long-arm amplicon are color-coded as indicated in Figure 2 of main text. (I) Numerical labels for BAC FISH probes used in Figure 1 – figure supplements 2 and 3.



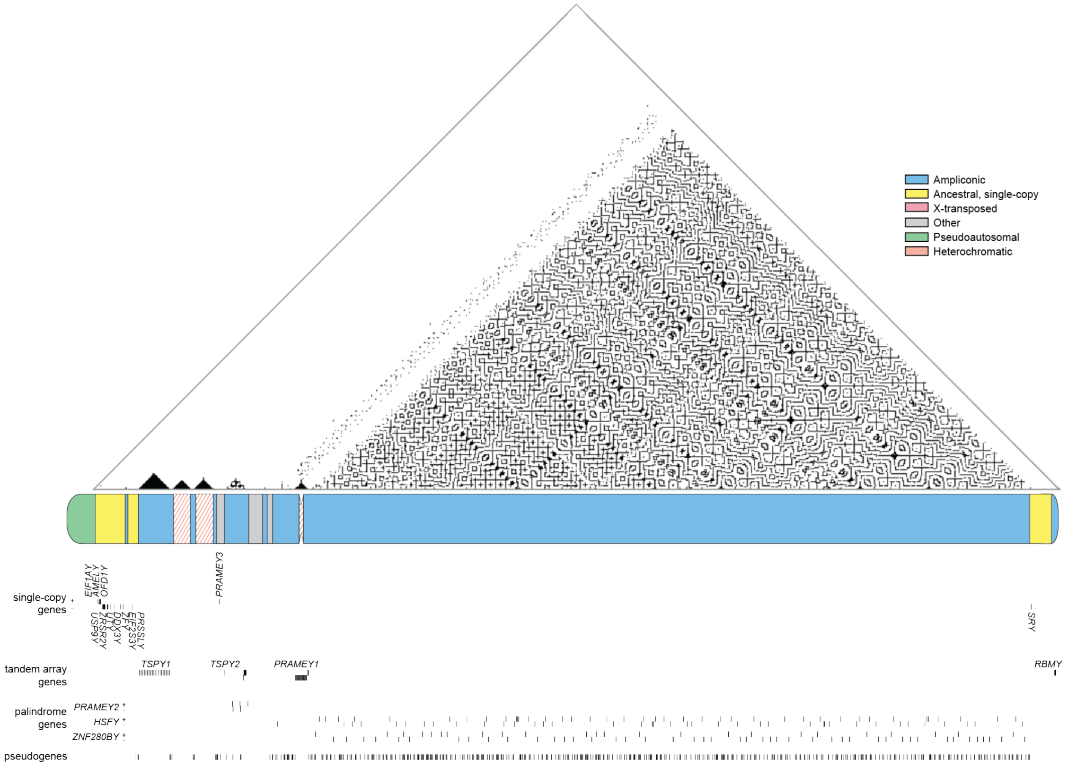
Supplemental Fig. S2. Extended metaphase FISH analysis to determine relative positions of tandem arrays.
Chromosomes are derived from a male *Bos taurus* (Hereford) cell line. BACs used as probes are indicated; numbers in parentheses after BAC names correspond to numerical labels in Supplementary File 1. 20 cells were examined for each experiment; representative images are shown. For each row, left, middle, and right images show merged, red, and green FISH signals, respectively. (A) C0396B10 is located in *TSPY* array; C0064N24 is located in heterochromatic array. Results indicate that both arrays are on Yp, and *TSPY* array is more distal than heterochromatic array. (B) C0062J21 is located in *PRAME1* array; C0064N24 is located in heterochromatic array. Results indicate that *PRAME1* array is very close to centromere, and heterochromatic array is distal to *PRAME1* array. (C) C0062J21 is located in *PRAME1* array; C0127C20 is located in *PRAME2* array, near *TSPY2* array. Results confirm that *PRAME1* array is very close to centromere, and *PRAME2* and *TSPY2* arrays are distal to *PRAME1* array. (D) C0127C20 is located in *PRAME2*/*TSPY2* array; C0020G15 is located in *RBMY* array. Results show that *RBMY* array is located on distal Yq, and *RBMY* probe cross-hybridizes with *PRAME2/TSPY2* array (due to a 3-kb region of homology between *RBMY* and *TSPY2* repeat units). (E) C0249C11 is located in pseudoautosomal region (PAR); C0215C05 is an unanchored BAC located near *PRAME2* array; C0401M09 is located in single-copy sequence near centromere. Results show that C0215C05 is located between PAR and C0401M09. (F) C0249C11 is located in PAR; C0127C20 is located in *PRAME2*/*TSPY2* array; C0401M09 is located in single-copy sequence near centromere. Results show that *PRAME2/TSPY2* arrays are located between PAR and C0401M09.



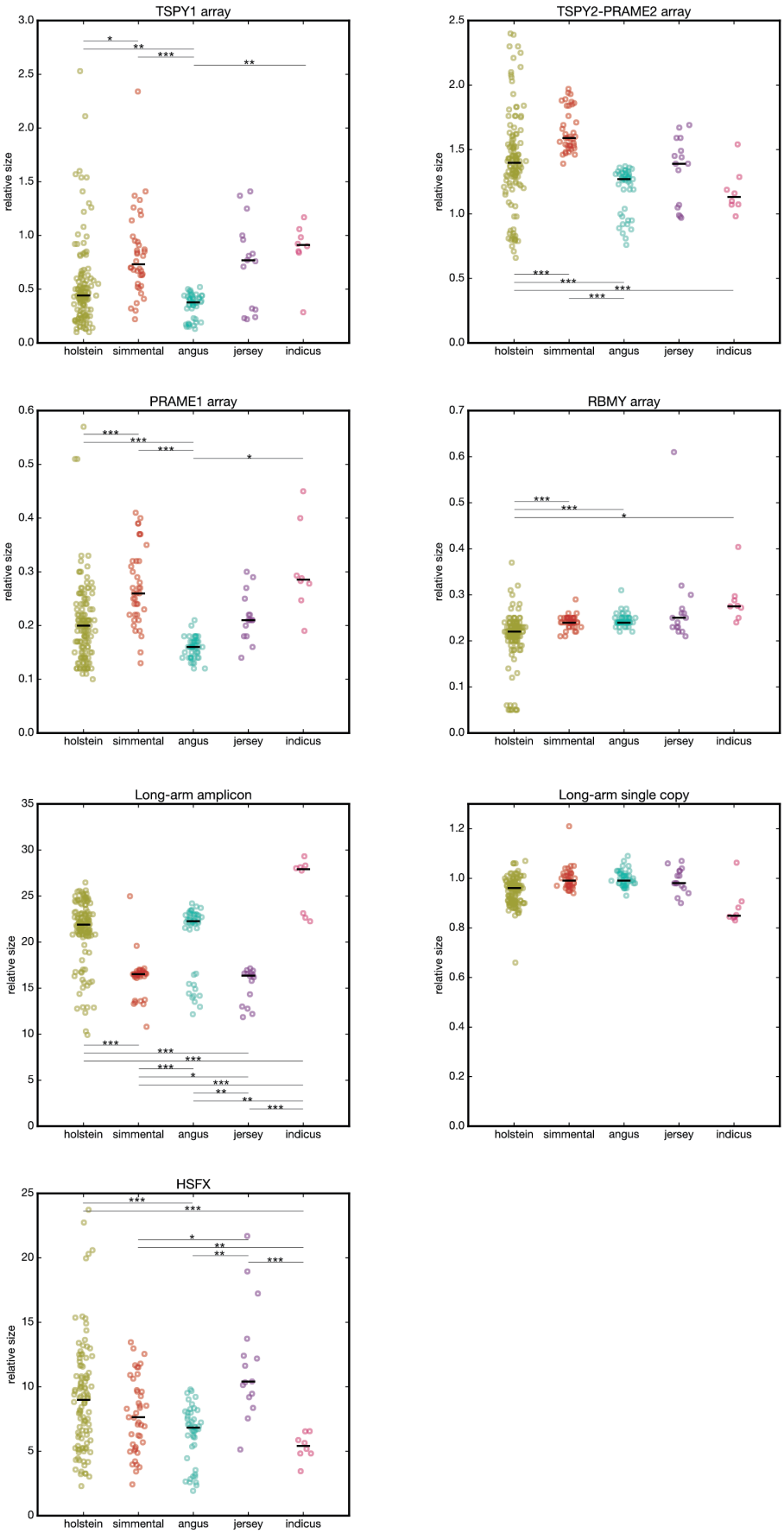
Supplemental Fig. S3. Extended metaphase FISH analysis to determine range of long-arm amplicon and interphase and fiber FISH analysis to estimate gap sizes. Chromosomes are derived from a male *Bos taurus* (Hereford) cell line. BACs used as probes are indicated; numbers in parentheses after BAC names correspond to numerical labels in Supplementary File 1. (A-B) Metaphase FISH experiments 20 cells were examined for each experiment; representative images are shown. For each row, left, middle, and right images show merged, red, and green FISH signals, respectively. (A) C0445P20 is located in long-arm amplicon; C0062J21 is located in *PRAME1* array; C0023I23 is located in single-copy sequence in distal Yq. Results indicate that long-arm amplicon extends across the entirety of Yq, from *PRAME1* array near centromere to distal Yq. (B) C0445P20 is located in long-arm amplicon; C0215C05 is located near *PRAME2* array. Results confirm that long-arm amplicon extends across the entirety of Yq. (C-D) Interphase FISH experiments. Estimations of gap sizes using FISH are only feasible for non-repetitive regions because this procedure requires sequence-specific, single-copy probes. 50-100 cells were examined for each experiment; representative images are shown. (C) At left, interphase FISH shows close proximity of C0317I05 and C0468D16. At right, fiber FISH-estimated distance between C0317I05 and C0468D16 is ~55 kb, based on Watson-Crick formula (1 μm ≈ 2.941 kb). Measurements were averaged across eight fibers. (D) At left, interphase FISH shows close proximity of A0268H07 and C0092K13. At right, fiber FISH-estimated distance between A0268H07 and C0092K13 is ~28 kb, based on the Watson-Crick model (2.9 kb per μm). Measurements were averaged across 42 fibers.



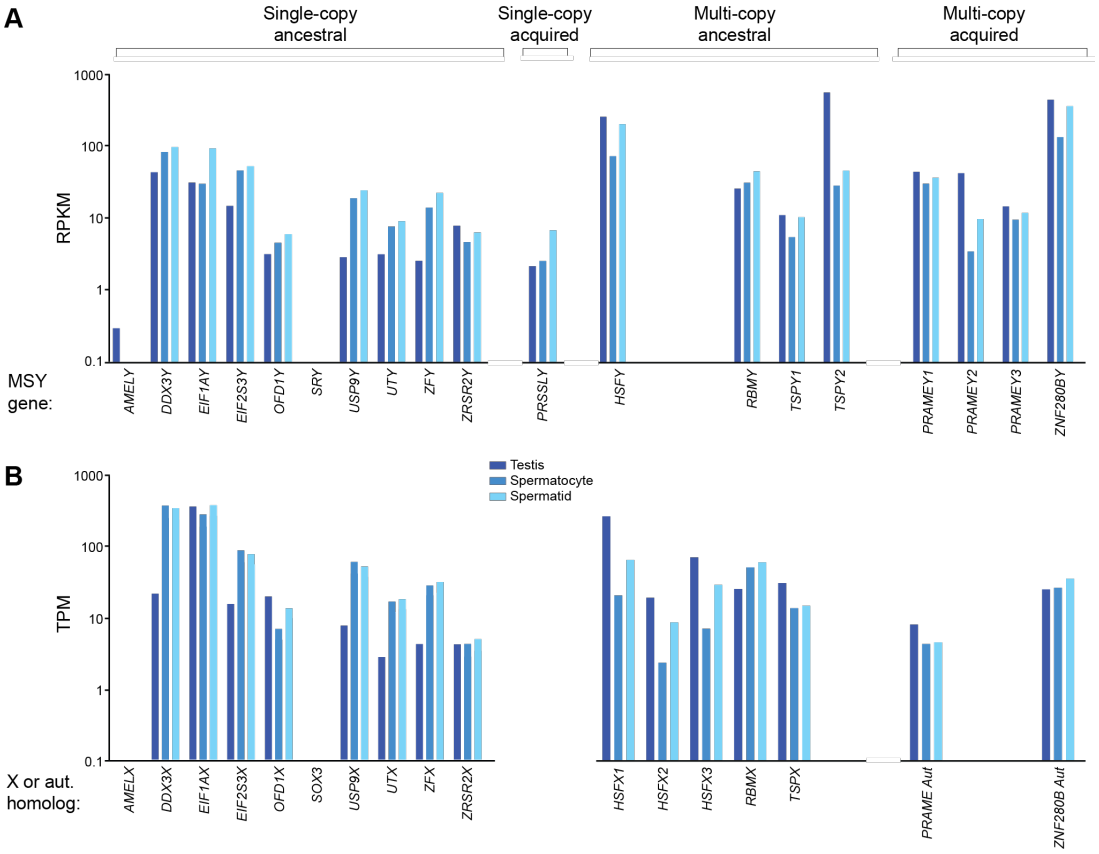
Supplemental Fig. S4. Long-arm amplicon substructure in bull MSY. Triangular dot plot of DNA sequence identity within 1-Mb region of bull Y long-arm. Each dot represents 100% intrachromosomal identity within a 100-bp window. Direct repeats appear as horizontal lines, inverted repeats as vertical lines, and palindromes as vertical lines that nearly intersect the baseline. Below plot, organization of repeat units, color-coded as indicated.



Supplemental Fig. S5. Map of bull MSY genes in relation to ampliconic repeat architecture. Triangular dot plot of DNA sequence identity within the bull MSY. Each dot represents 100% intrachromosomal identity within a 200-bp window. Direct repeats appear as horizontal lines, inverted repeats as vertical lines, and palindromes as vertical lines that nearly intersect the baseline. Below plot, schematic representation of Chromosome is shown. Sequence classes are color-coded as indicated. Other: single-copy male-specific sequences that are not homologous to the X Chromosome. Below diagram, single-copy and multicopy protein-coding genes are shown in black; pseudogenes are shown at bottom.



Supplemental Fig. S6. Copy number variability within MSY amplicons and *HSFX* among 234 bulls representing four different breeds.
Relative read depth mapping is shown for five MSY ampliconic regions (*TSPY1* array, *TSPY2-PRAME2* array, *PRAME1* array, *RBMY* array, and long-arm amplicon), *HSFX* on the X Chromosome, and a Y-Chromosome long-arm single-copy region as a control. Each graph shows number of reads mapping to a given region normalized by number of reads mapping to 1-Mb short-arm single-copy region on the Y Chromosome (for MSY amplicons) or the X Chromosome (for *HSFX*). Each circle represents data for one animal. Data for different breeds (Holstein, Simmental, Angus, and Jersey) are grouped separately. Black horizontal line indicates mean for each breed. Multiple-testing-corrected P-values of statistically significant differences between the means are indicated as determined by Welch t-test; **P* <0.05, ***P*<0.01, ****P*<0.001.



Supplemental Fig. S7. Gene expression analysis in purified male germ cells. Analysis includes RNA-seq datasets from whole testis and purified germ cell fractions containing pachytene spermatocytes and round spermatids. Expression for MSY genes (A) and their X or autosome homologs (B) was estimated in transcript per million (TPM) units. TPM values are plotted on a log10 scale. The analysis includes three replicates for whole testis and single samples for each purified germ cell fraction.

SUPPLEMENTAL METHODS

Localization of the bull Y centromere

Our assembly of the bull MSY does not include the centromere, but we determined the approximate location of the centromere from several lines of evidence:

1) FISH analysis (see main text experimental procedures) using a BAC probe derived from the *PRAMEY1* array placed this sequence very near or overlapping the centromere, as visualized by a constriction in metaphase chromosomes (Figure 1).

Figure 1. Metaphase FISH experiment on male *Bos taurus* (Hereford) cell line using BAC C0062J21 as a probe (shown in red).

2) We performed serial immunocytochemistry (with antibodies against centromere proteins) and DNA FISH to more precisely visualize the bull Y centromere. These experiments confirmed that the *PRAMEY1* array is adjacent to the centromere (Figure 2). In brief, cells were harvested and slides were prepared and immunostained as previously described (Jeppesen, 2000, Goshima et al., 2003) using primary antibody (anti-human centromere/kinetochore protein; Antibodies Incorporated, Davis, CA) and secondary antibody (anti-human IgG H&L fluorescein-conjugated, donkey polyclonal; Rockland Antibodies & Assays, Gilbertsville, PA). After imaging, the same slides were prepared for FISH analysis by treatment with RNase for one hour at 37ºC and denaturation in 70% FA lysis buffer for two minutes at 70ºC. Probe hybridization was performed overnight using standard procedures.

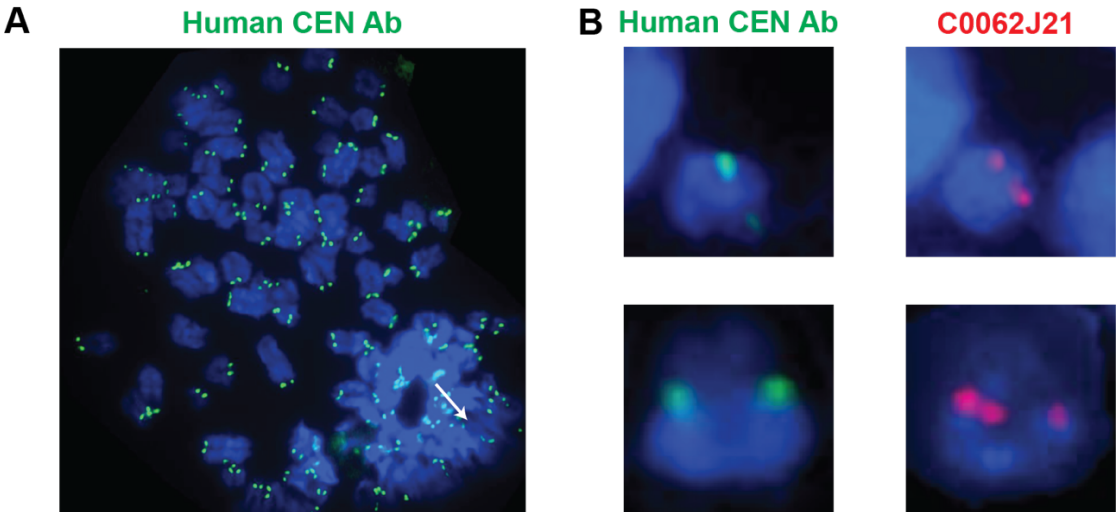


Figure 2. Chromosomes are derived from a male Bos taurus (Hereford) cell line. (A) Immunostaining with anti-human centromere protein (CEN) on metaphase spread. Y chromosome is indicated by arrow. (B) CEN immunostaining followed by FISH on metaphase spreads. BAC probe for FISH is located near presumed centromere. Images are cropped to highlight Y chromosome. Immunostaining and FISH images were captured separately for the same set of slides because co-staining is not possible (centromeric proteins are removed during DNA denaturation, which is required for FISH). Precise addresses for individual cells were noted so that both immunostaining and FISH images could be directly compared within the same cell.

3) We tested 23 Y-specific markers on a male *Bos taurus* 25,000-rad radiation hybrid (RH) panel constructed using DNA samples collected from the CHORI-240 library donor, and found that the retention frequency of markers immediately adjacent to the *PRAME1Y* array was dramatically higher than the remaining markers (Figure 3, see also Supplemental Table S2) – consistent with these sequences conferring the ability to segregate to both daughters in meiosis.

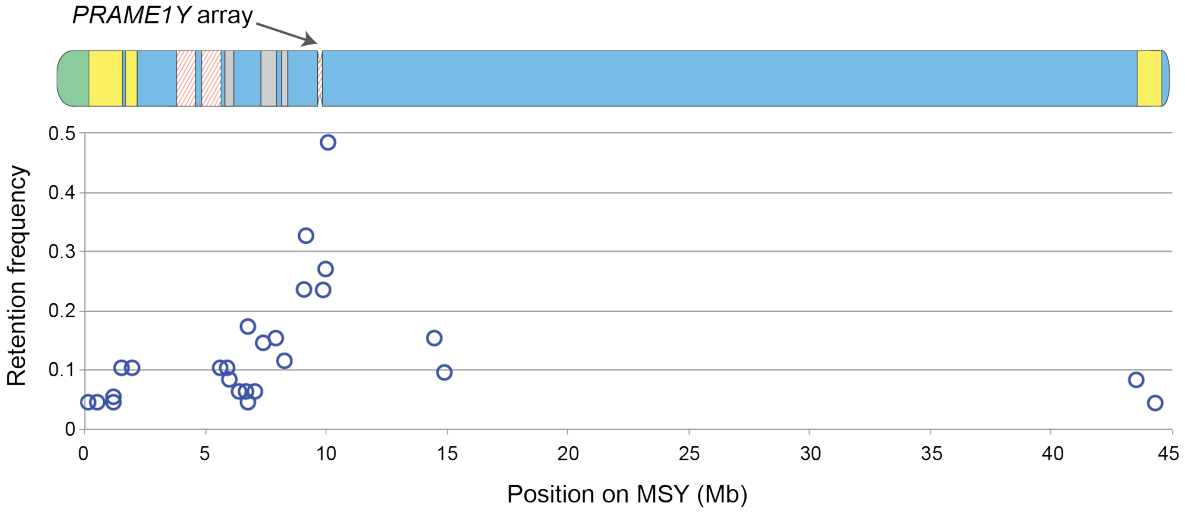


Figure 3. Retention frequencies and chromosomal locations of 25 STS markers tested on RH panel.

4) We employed ChIP-seq analysis (Vafa and Sullivan, 1997) to attempt to isolate bull Y centromere sequences directly. ChIP experiments were performed as previously described (Lee et al., 2006). Anti-CENP-A antibody (Abcam Inc., Cambridge, MA) was used for immunoprecipitation of chromatin prepared from a male *Bos taurus* (Hereford) cell line. Immunoprecipated DNA was sequenced using the Illumina MiSeq platform, generating 250-bp paired end reads. As a control, a MiSeq run was performed on purified DNA from whole-cell extract (WCE). All sequence reads enriched in the ChIP dataset (55-mers with >100-fold enrichment compared to WCE dataset) were assembled using SPADES (Bankevich et al., 2012). We identified six dominant classes of satellite sequences that corresponded to known *Bos taurus* satellite sequences in RepBase (www.girinst.org/repbase; BTSAT1-6). We generated long-range PCR products specific for each satellite class to use as FISH probes on male *Bos taurus* cells (Supplemental File S2). Surprisingly, each of these satellite-containing probes hybridized to all or most centromeres in the genome except the X and Y chromosomes (Figure 4). Although the sequence composition of the Y-chromosome centromere remains unknown, it appears to bear no resemblance to autosomal centromeres, which is a common feature of Y centromeres in mammals (Alexandrov et al., 2001, Pertile et al., 2009).

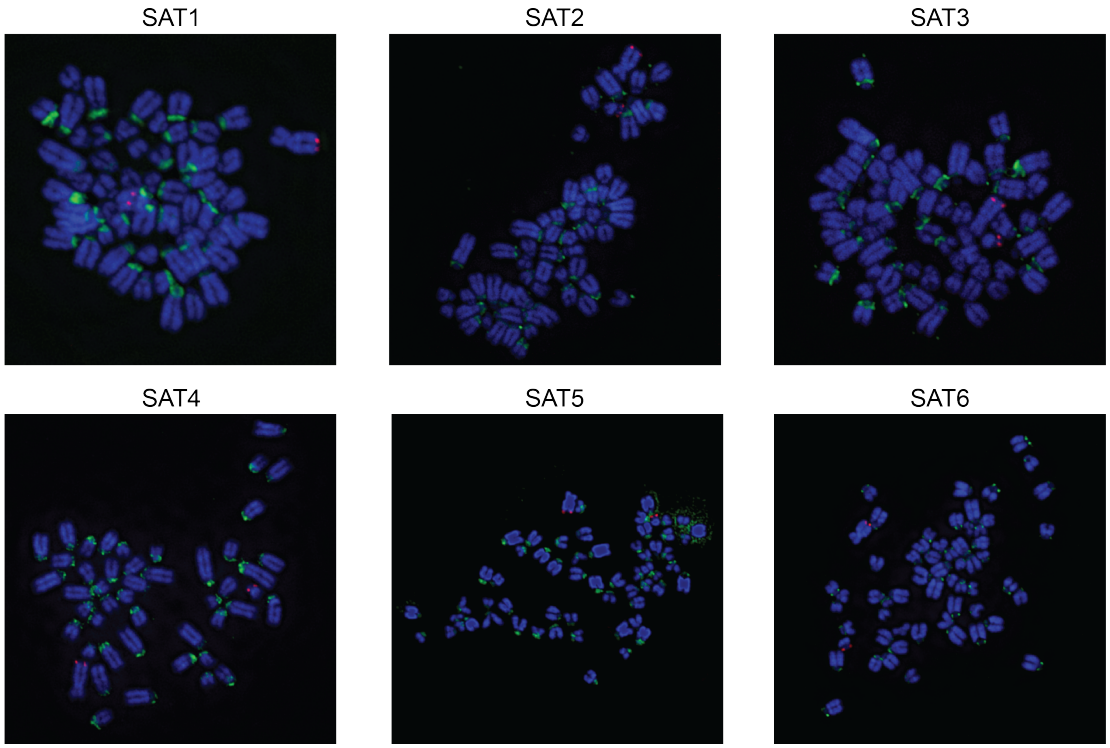


Figure 4. Two-color FISH experiments were performed using probes corresponding to BTSAT1-6 (green) combined with a PAR-derived BAC probe to mark the X and Y chromosomes (red). Results show that BT-SAT probes hybridize to centromeres of multiple autosomes, but not the X and Y chromosomes.

REFERENCES

Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. 2001. Alpha-satellite DNA of primates: old and new families. *Chromosoma,* 110**,** 253-66.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol,* 19**,** 455-77. doi:10.1089/cmb.2012.0021.

Goshima, G., Kiyomitsu, T., Yoda, K. & Yanagida, M. 2003. Human centromere chromatin protein hMis12, essential for equal segregation, is independent of CENP-A loading pathway. *J Cell Biol,* 160**,** 25-39. doi:10.1083/jcb.200210005.

Jeppesen, P. 2000. Immunofluorescence in cytogenetic analysis: method and applications. *Genet Mol Biol,* 23**,** 1107-1114.

Lee, T. I., Johnstone, S. E. & Young, R. A. 2006. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols,* 1**,** 729-748. doi:Doi 10.1038/Nprot2006.98.

Pertile, M. D., Graham, A. N., Choo, K. H. & Kalitsis, P. 2009. Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome Res,* 19**,** 2202-13. doi:10.1101/gr.092080.109.

Vafa, O. & Sullivan, K. F. 1997. Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. *Curr Biol,* 7**,** 897-900.