**SUPPLEMENTAL MATERIAL**

# Large palindromes on the primate X Chromosome are preserved by natural selection

Emily K. Jackson, Daniel W. Bellott, Ting-Jan Cho, Helen Skaletsky,
Jennifer F. Hughes, Tatyana Pyntikova, and David C. Page



*Supplemental Tables are separate from this file and can be downloaded separately.*

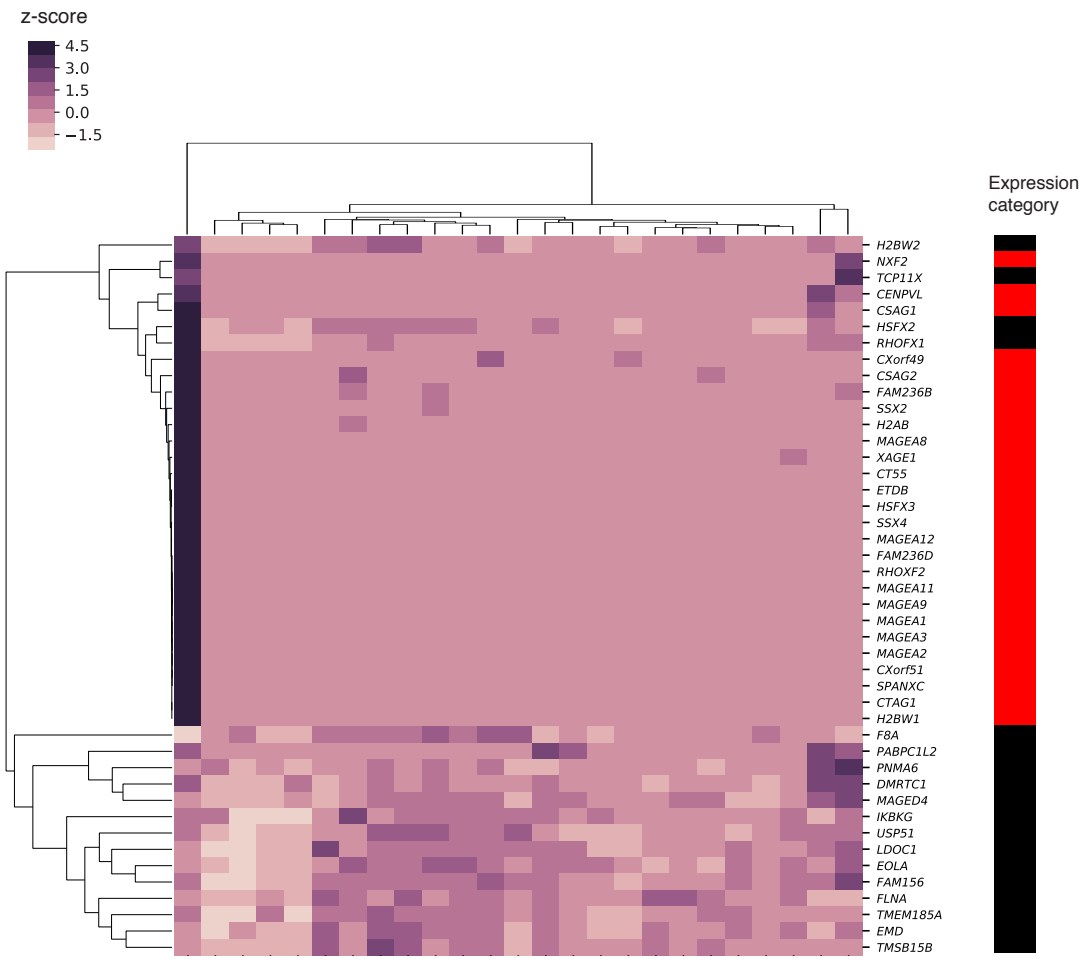
**Supplemental Note 1:** Treatment of human X-palindrome genes with conflicting annotations

There were three instances in which a gene in one arm of a palindrome was designated as protein-coding while the homologous sequence in the other arm was designated a pseudogene: *IKBKG* (protein-coding) and *IKBKGP1* (unprocessed pseudogene); *PNMA6A* (protein-coding) and *PNMA6B* (unprocessed pseudogene), and *PWWP4* (protein-coding) and novel gene ENSG00000224931 (processed pseudogene). We decided whether to include gene copies marked as pseudogenes in downstream analyses, i.e., whether their expression should be averaged with that of the corresponding protein-coding gene, as follows:

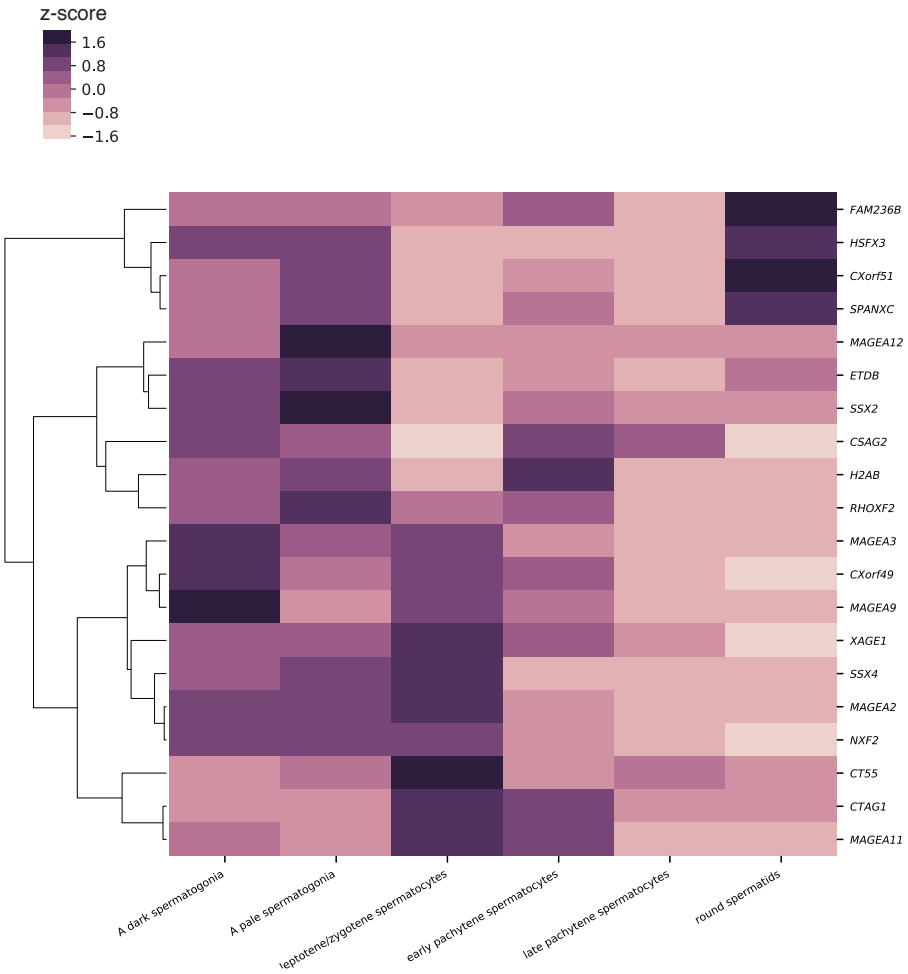
1) *PNMA6A* encodes a protein of 399 amino acids. *PNMA6A* and *PNMA6B* differ in their coding sequence by only a single missense substitution. The 3' UTR of *PNMA6B* is truncated, but the significance of this is unclear. Given that *PNMA6B* encodes an intact protein-coding sequence, we chose to include *PNMA6B* in downstream analyses.

2) *PWWP4* encodes a protein of 2061 amino acids. Novel gene ENSG00000224931 has a nonsense substitution, but contains a downstream start codon that would lead to translation of the terminal 1253 amino acids of *PWWP4.* Given that novel gene ENSG00000224931 encodes a protein encompassing more than half the length of the original protein, we chose to include novel gene ENSG00000224931 in downstream analyses.

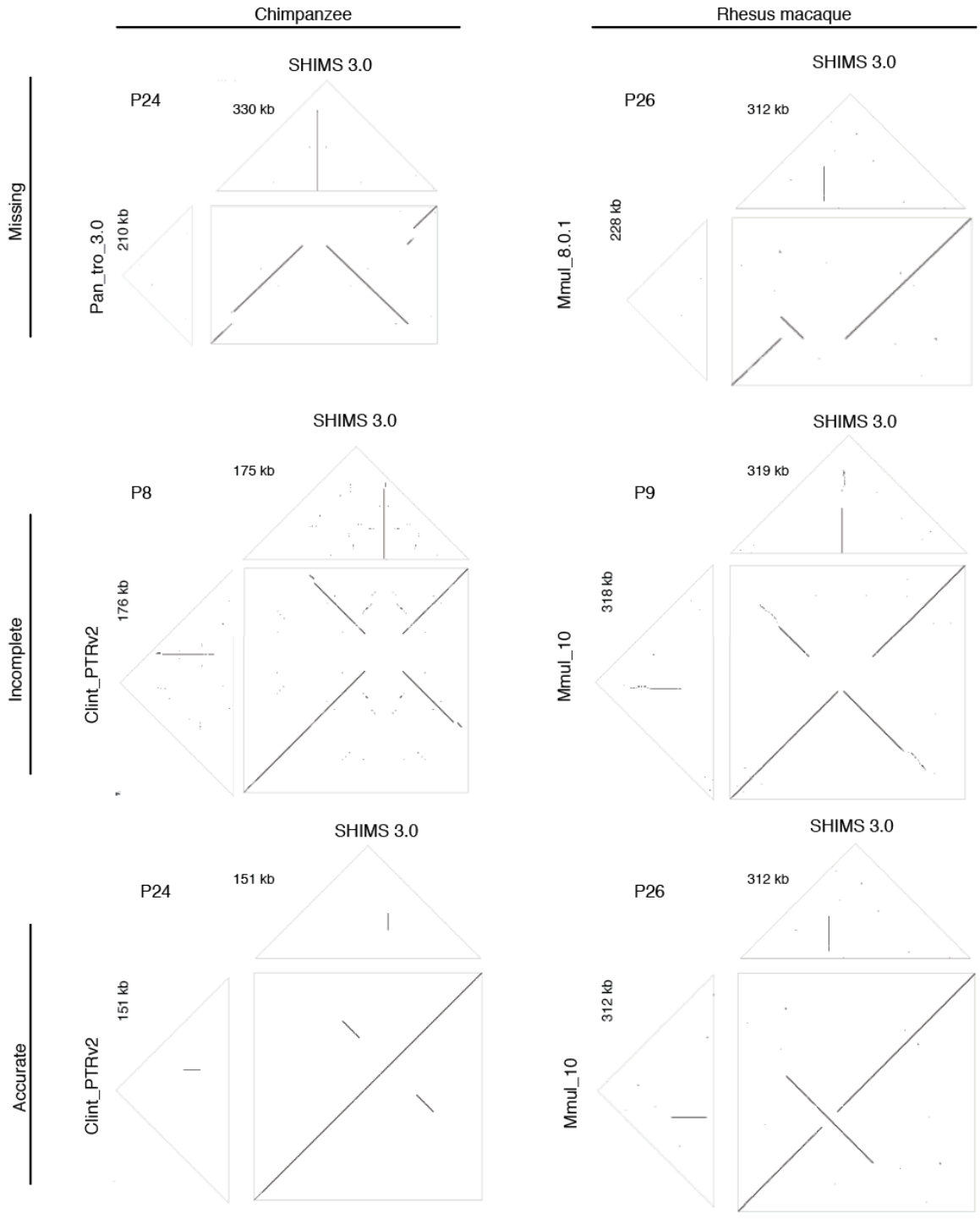
3) *IKBKG* encodes a protein of 419 amino acids. *IKBKGP1* is a well-characterized pseudogene lacking the promoter and first four exons of *IKBKG* (Aradhya et al. 2001); we therefore chose not to include it in downstream analyses.



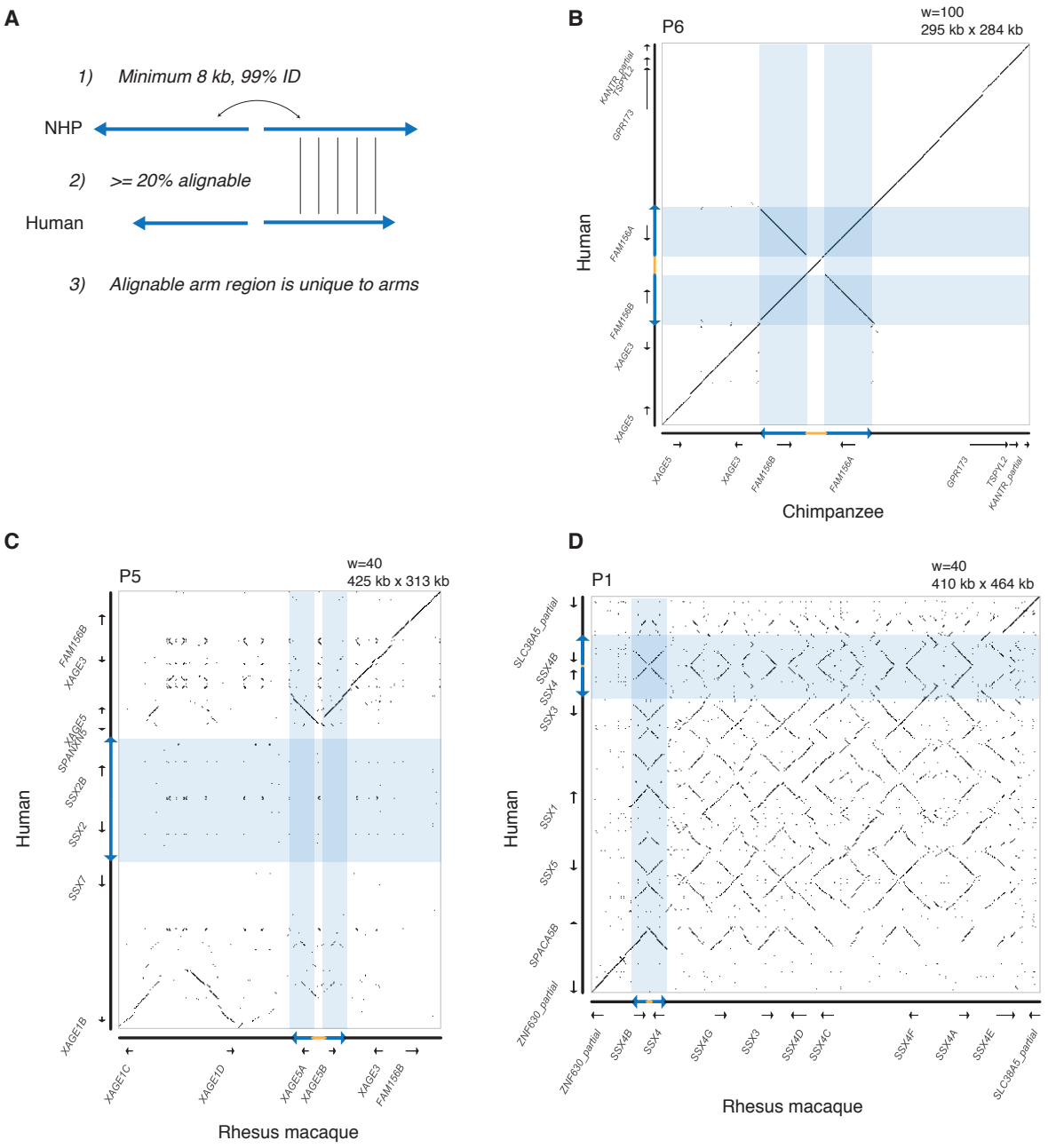

**Supplemental Figure S1.** Expression of human X-palindrome gene families. Heatmap shows z-score for each expressed gene (>2 TPM in at least one tissue) across 25 tissues from GTEx, with row and column order determined by hierarchical clustering. Expression category: Shows whether expression is testis-biased (red) or broad (black). Testis-biased: Minimum 2 TPM in testis, and testis accounts for >25% of log2 normalized expression summed across all tissues. Broad: All other expressed genes.



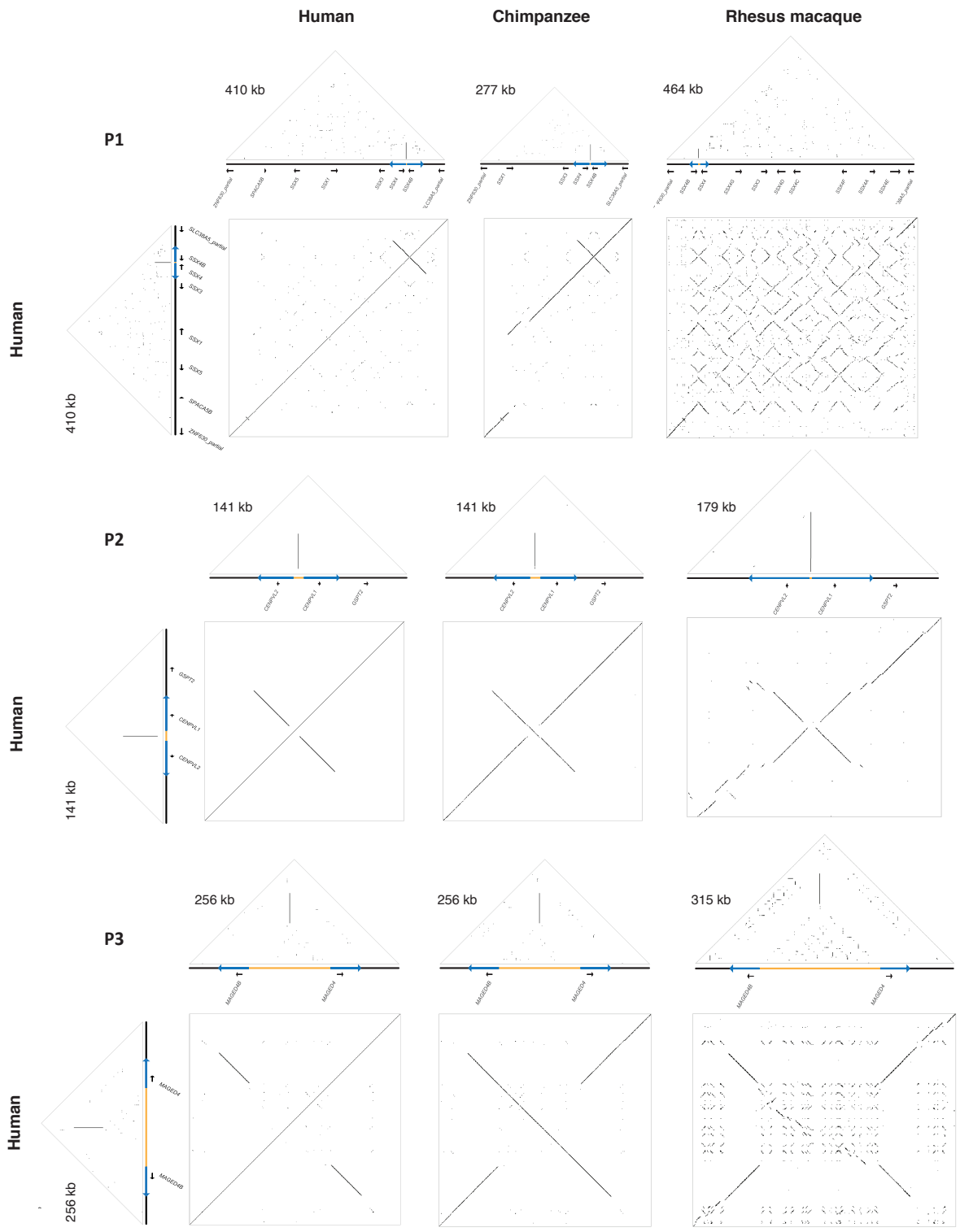

**Supplemental Figure S2.** Expression of human testis-biased X-palindrome gene families during spermatogenesis. Heatmap shows z-score for each expressed gene (>2 TPM in at least one spermatogenic stage) across six spermatogenic stages from Jan et al. 2017, with row order determined by hierarchical clustering.



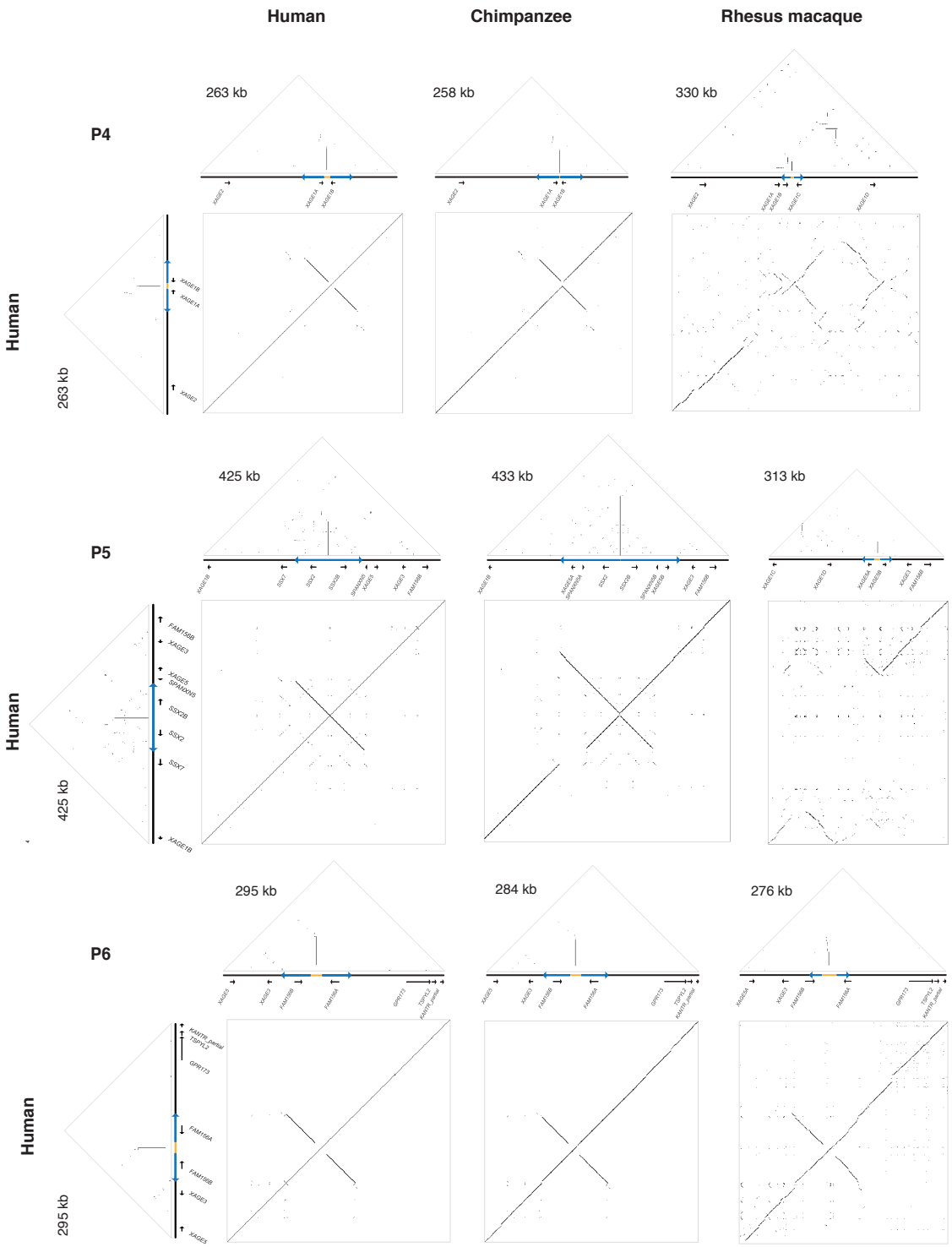

**Supplemental Figure S3.** Structural comparisons between palindromes in SHIMS 3.0 assemblies and existing X-Chromosome assemblies. Missing: No palindrome present in non-SHIMS 3.0 assembly. Incomplete: Part of palindrome present in non-SHIMS 3.0 assembly. Accurate: Full palindrome present in non-SHIMS 3.0 assembly. For accurate palindrome assemblies, note the presence in the square dot plot of an uninterrupted diagonal line and two arms that each map twice to the SHIMS 3.0 assembly. w=100 for all triangular and square dot plots.



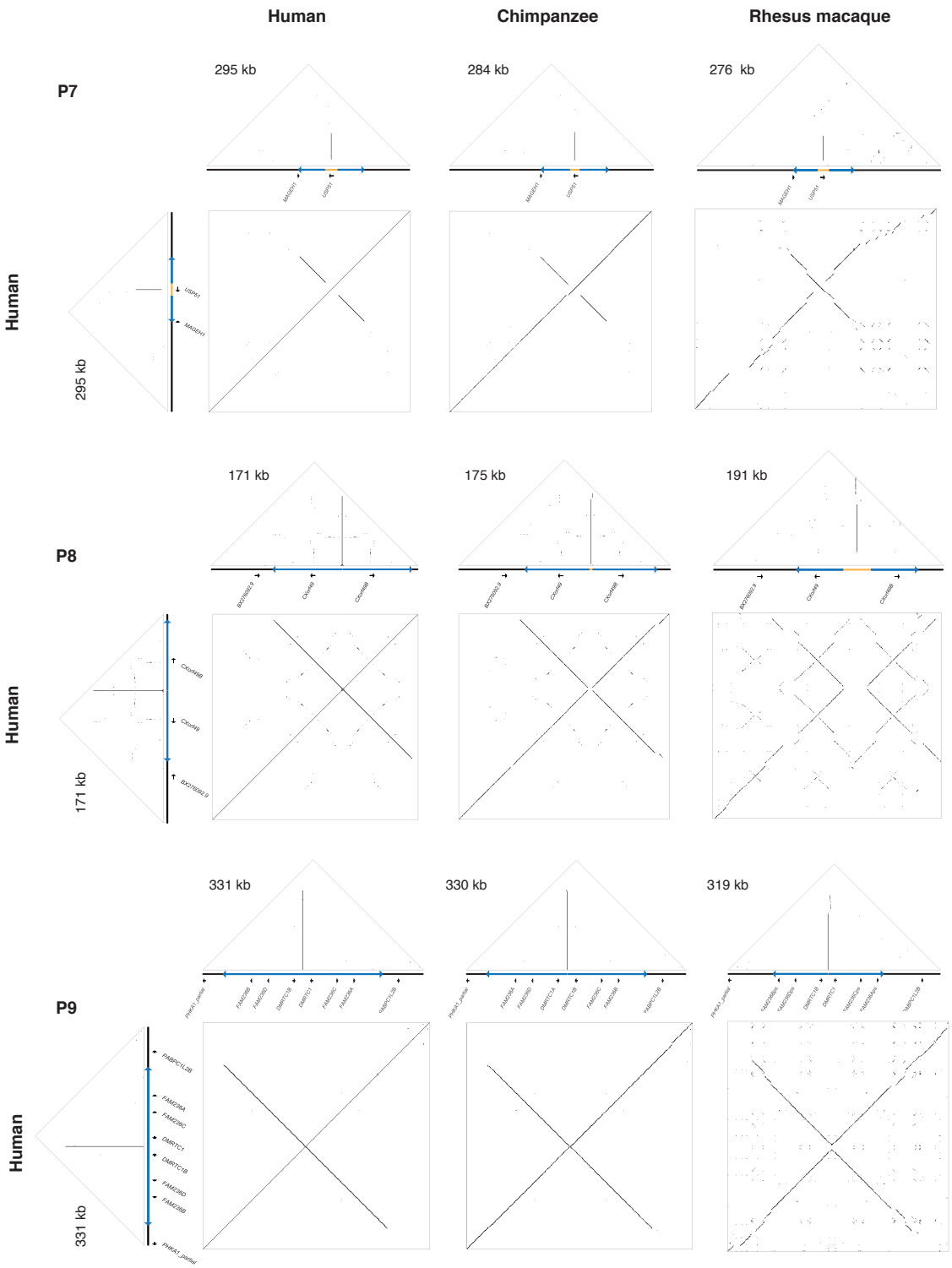

**Supplemental Figure S4.** Definition of orthologous palindromes. a) Criteria for defining orthologous palindromes. NHP = non-human primate (chimpanzee or rhesus macaque). Human and NHP palindrome arms were aligned with ClustalW, and required to have at least 20% alignment between species. Palindromes were excluded if the alignable region between palindrome arms mapped equally well to flanking sequence using reciprocal BLAST hits (>10% positions in high-quality hits mapping outside of palindrome arms). b) Example of an orthologous palindrome. Human palindrome arms map exactly twice to chimpanzee palindrome arms, and vice versa. c, d) Examples of palindromes that are not orthologous. c) Human palindrome arms have no orthologous sequence in rhesus macaque. Rhesus macaque palindrome arms correspond to flanking sequence in human. d) Rhesus macaque palindrome arms correspond equally well to more than two positions in human, and vice versa. Note that the region with the strongest orthology to rhesus macaque palindrome arms corresponds to flanking sequence in human.



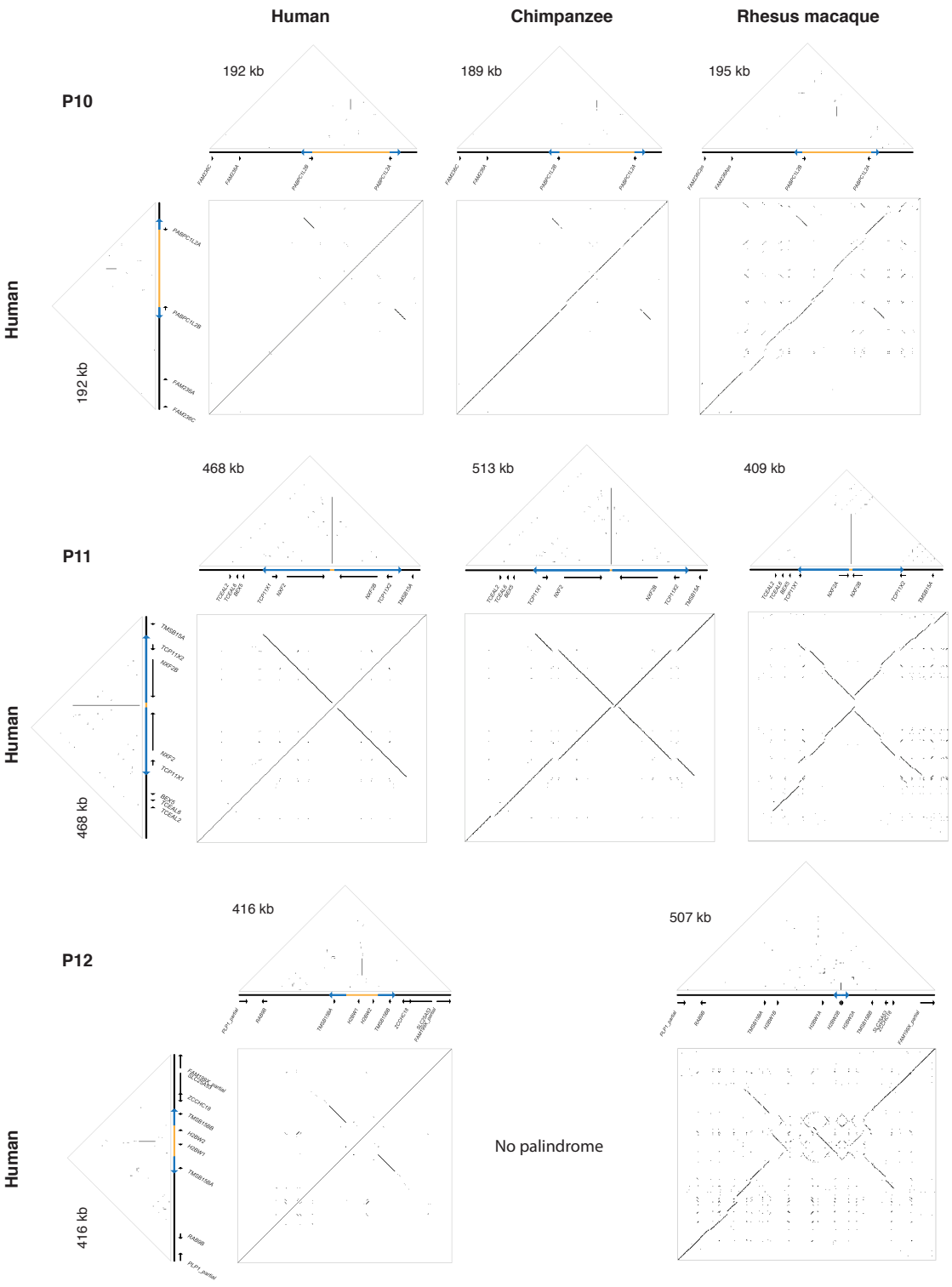

**Supplemental Figure S5.** Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



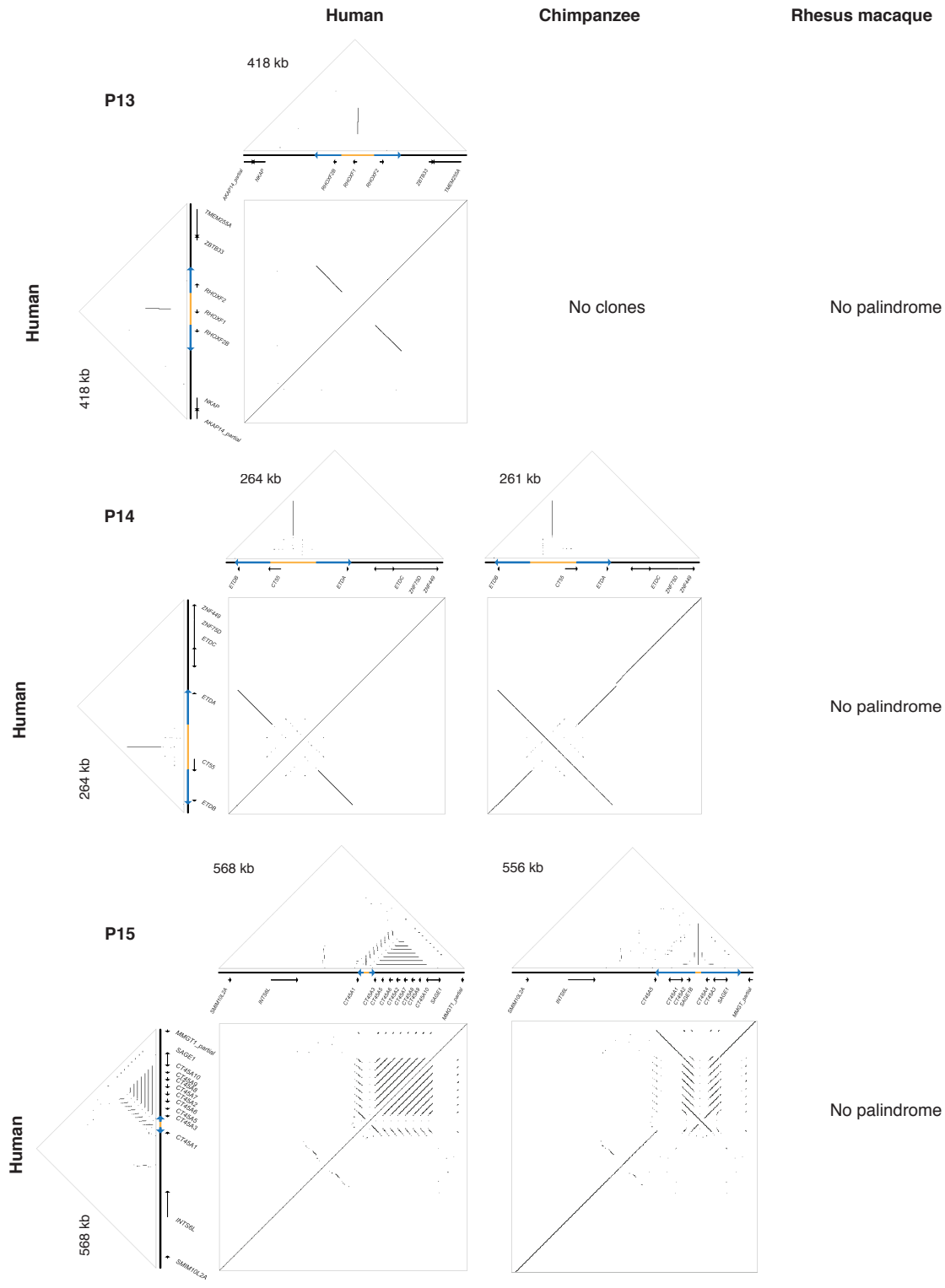

**Supplemental Figure S5.** Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



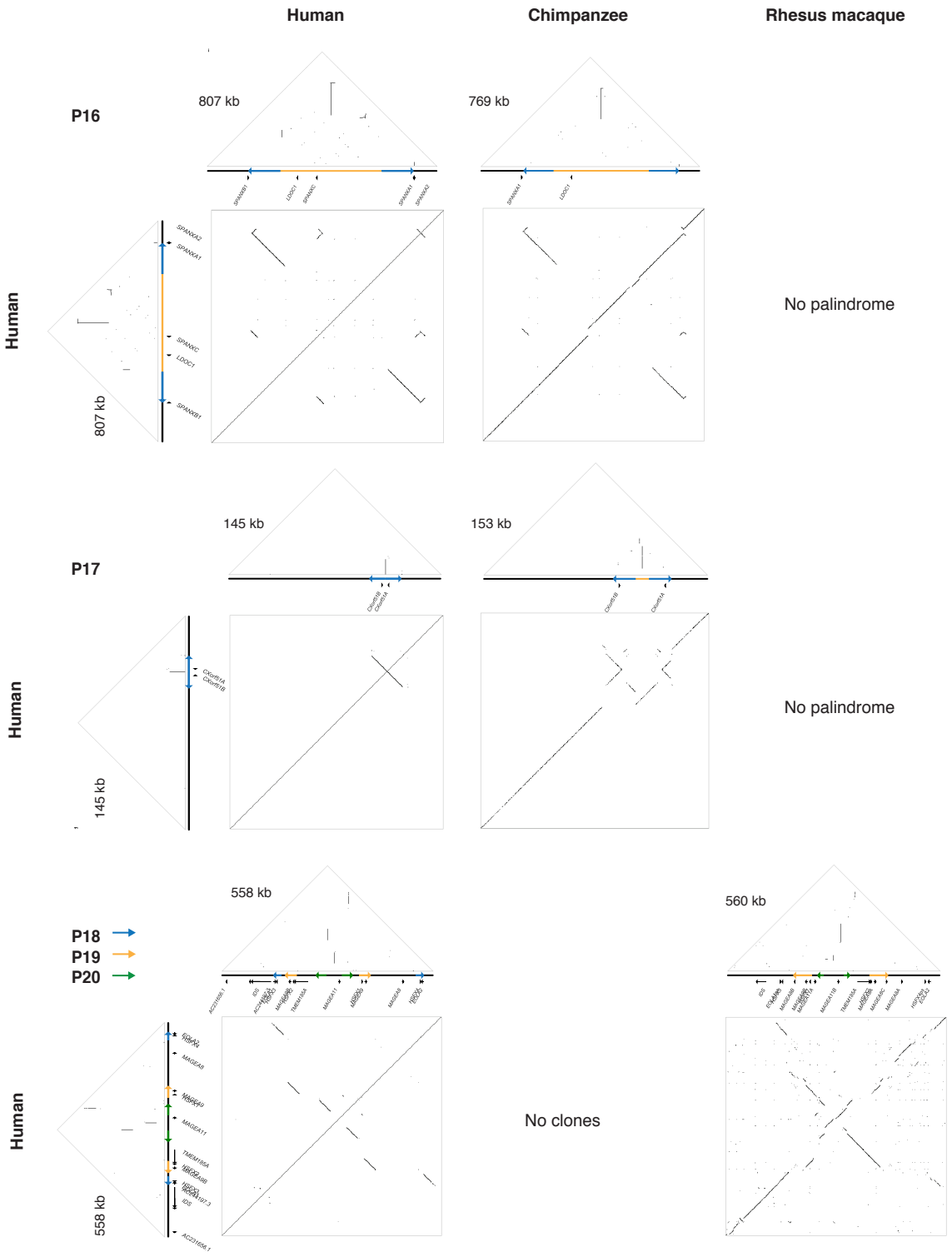

**Supplemental Figure S5.** Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



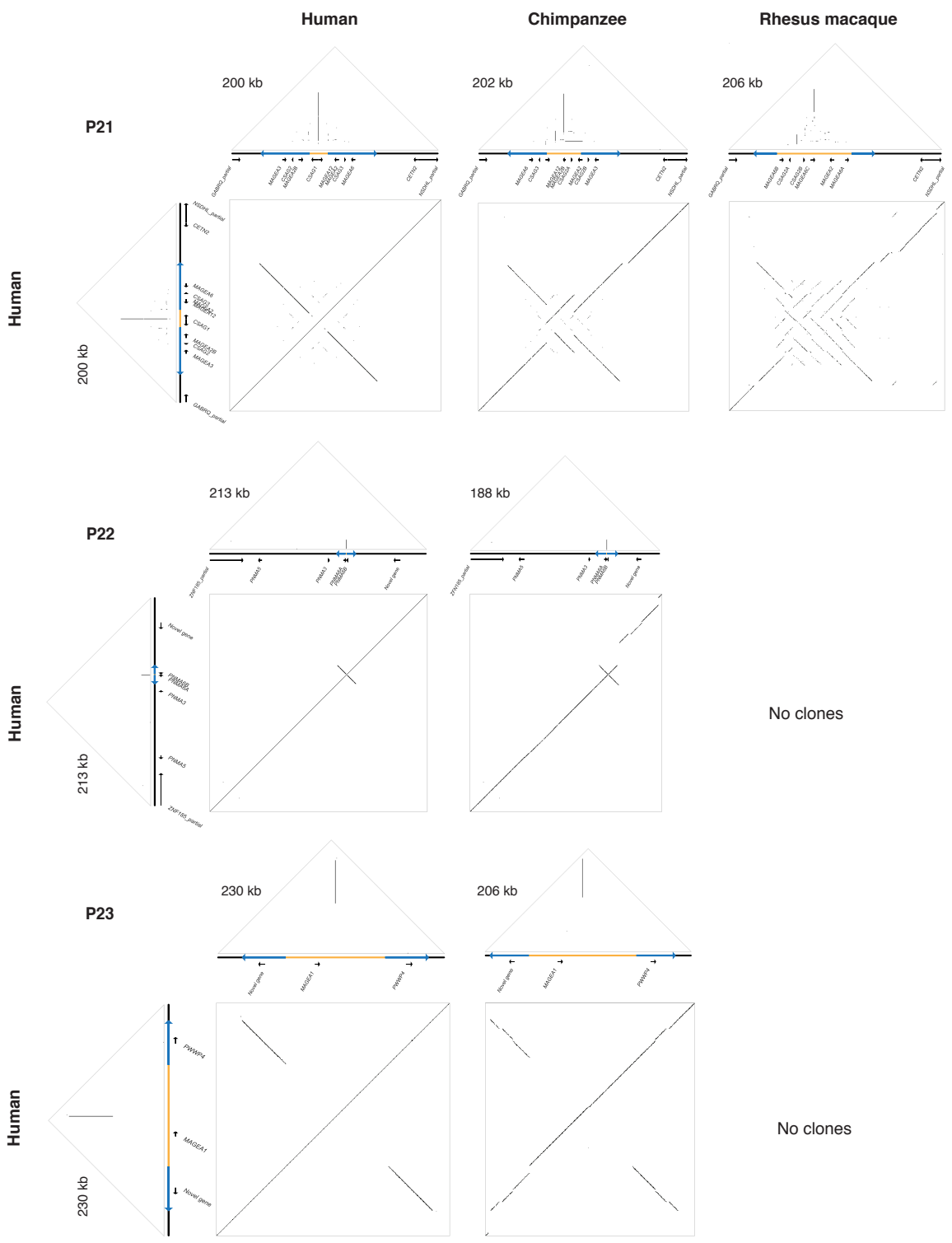

**Supplemental Figure S5.** Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



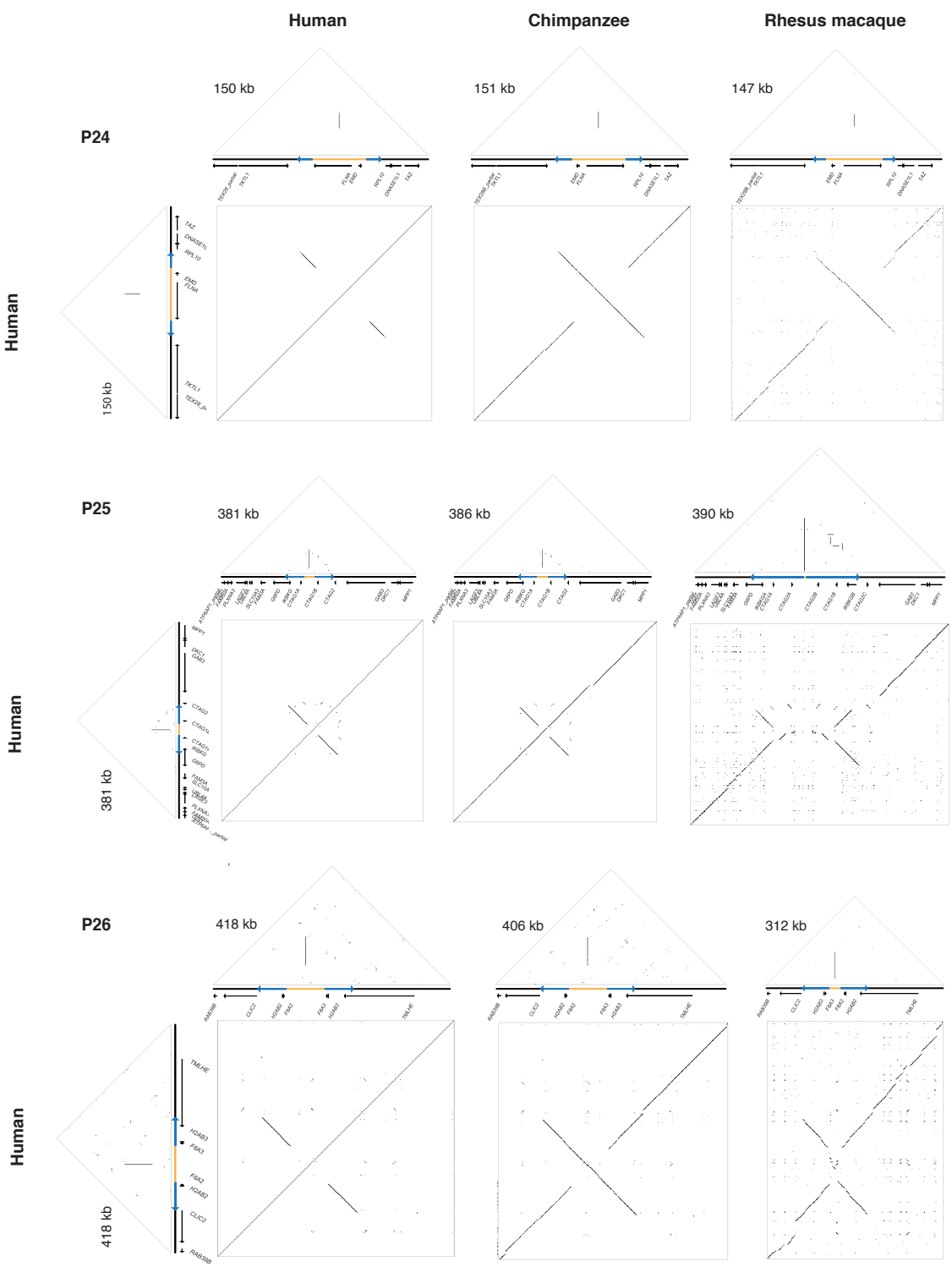

**Supplemental Figure S5.** Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



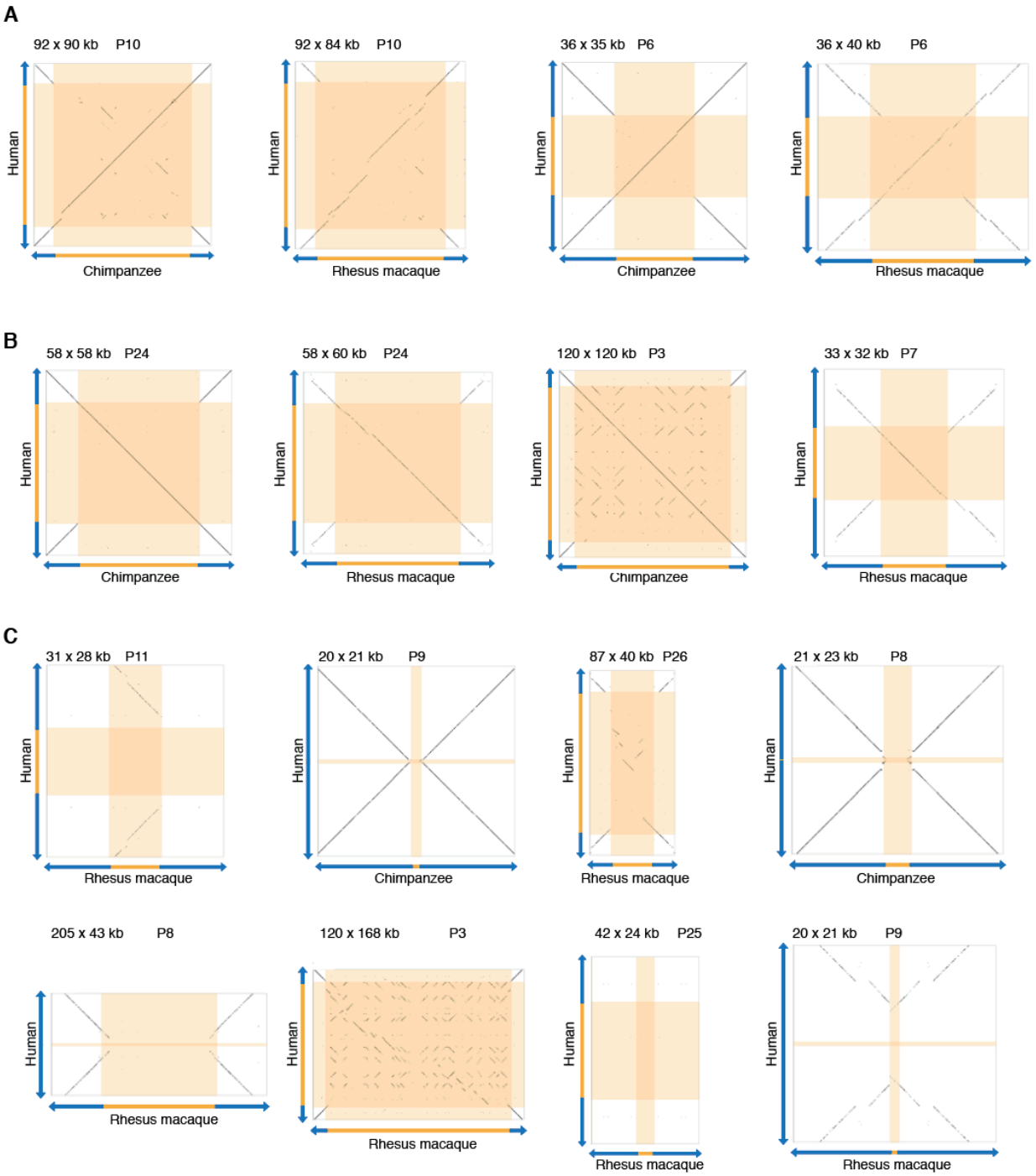

**Supplemental Figure S5.** Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



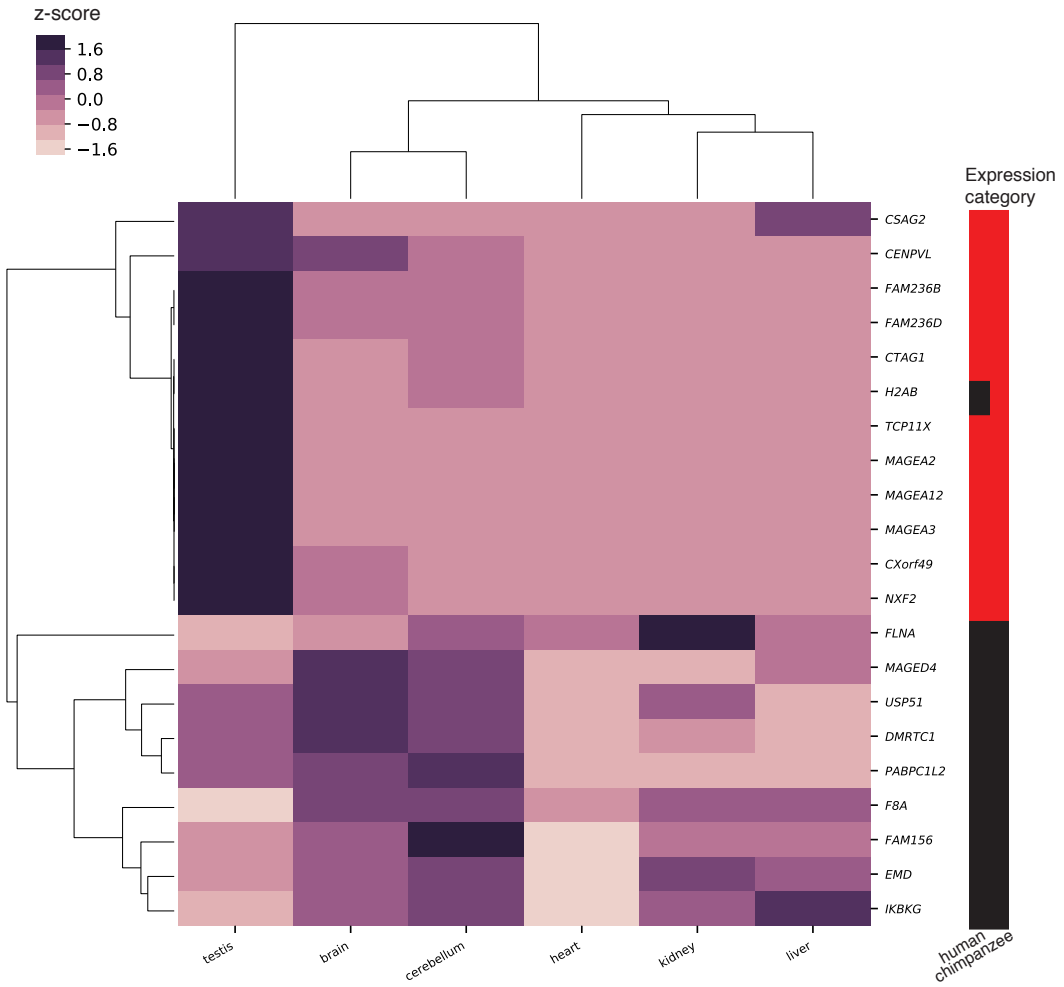

**Supplemental Figure S5.** Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



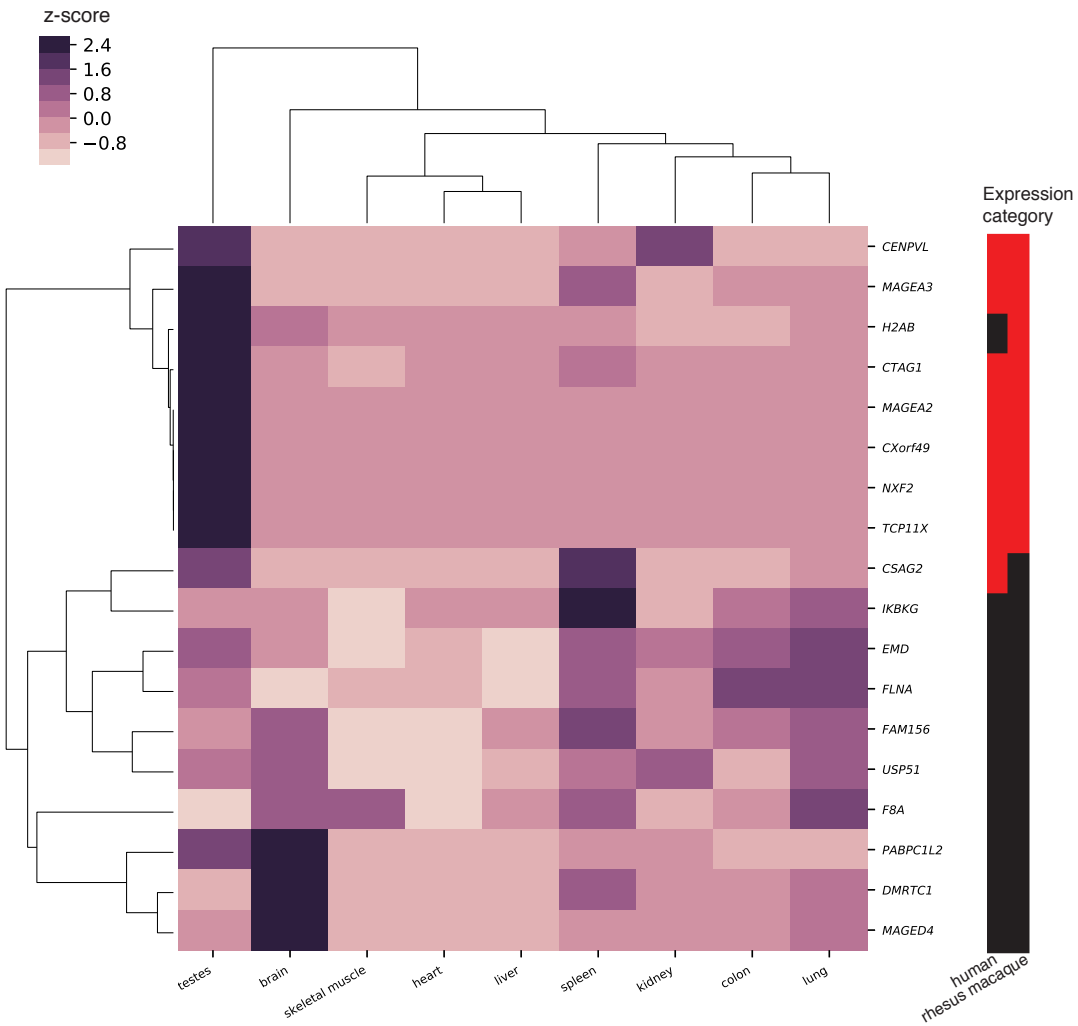

**Supplemental Figure S5.** Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



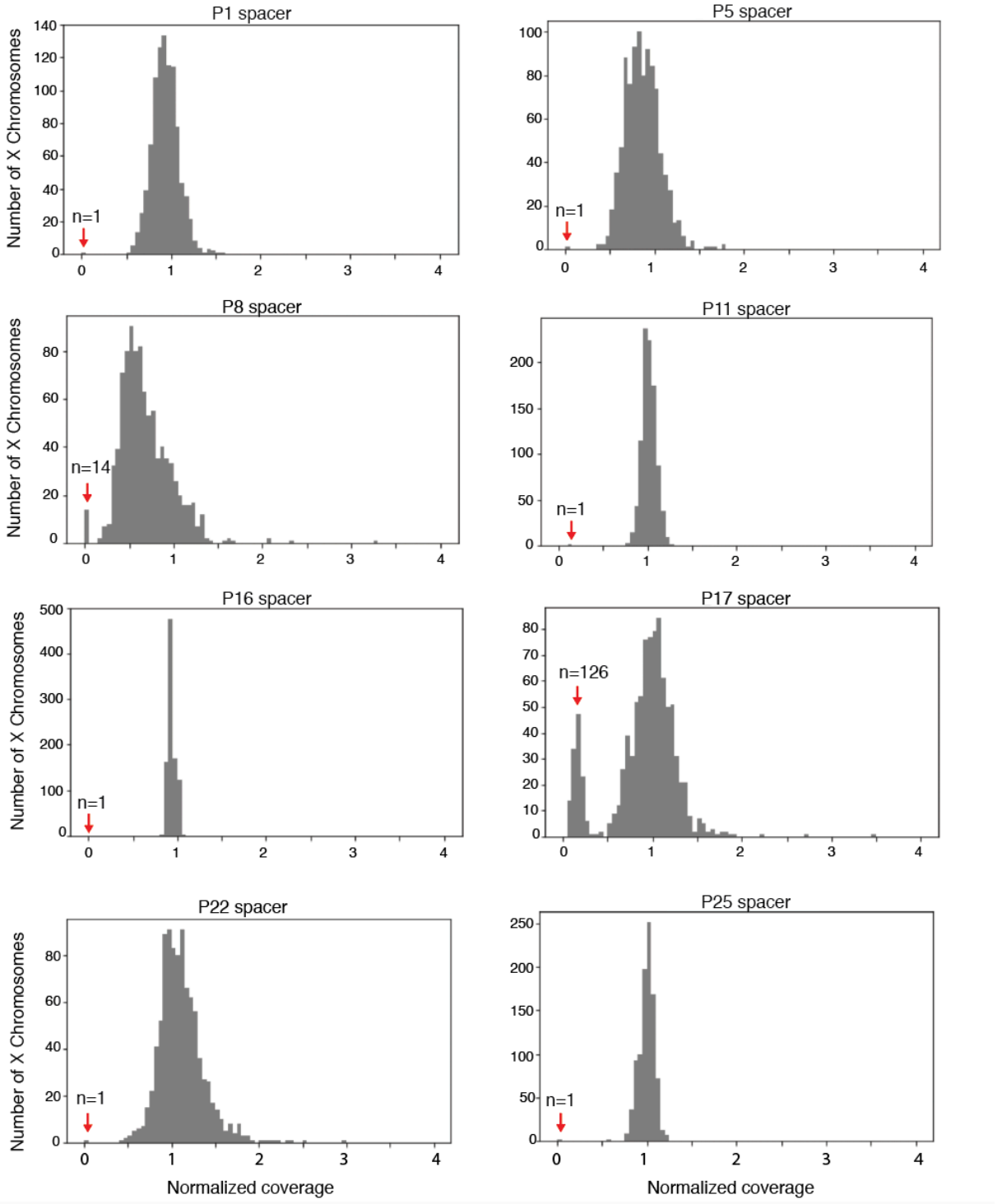

**Supplemental Figure S6**. Additional examples of spacer configurations in orthologous palindromes. All plots show the inner 10 kb of palindrome arms plus the spacer. w=40 for human vs. rhesus macaque comparisons; w=100 for human vs. chimpanzee comparisons. a) Human configuration, b) Inversions, c) Non-orthologous spacers.



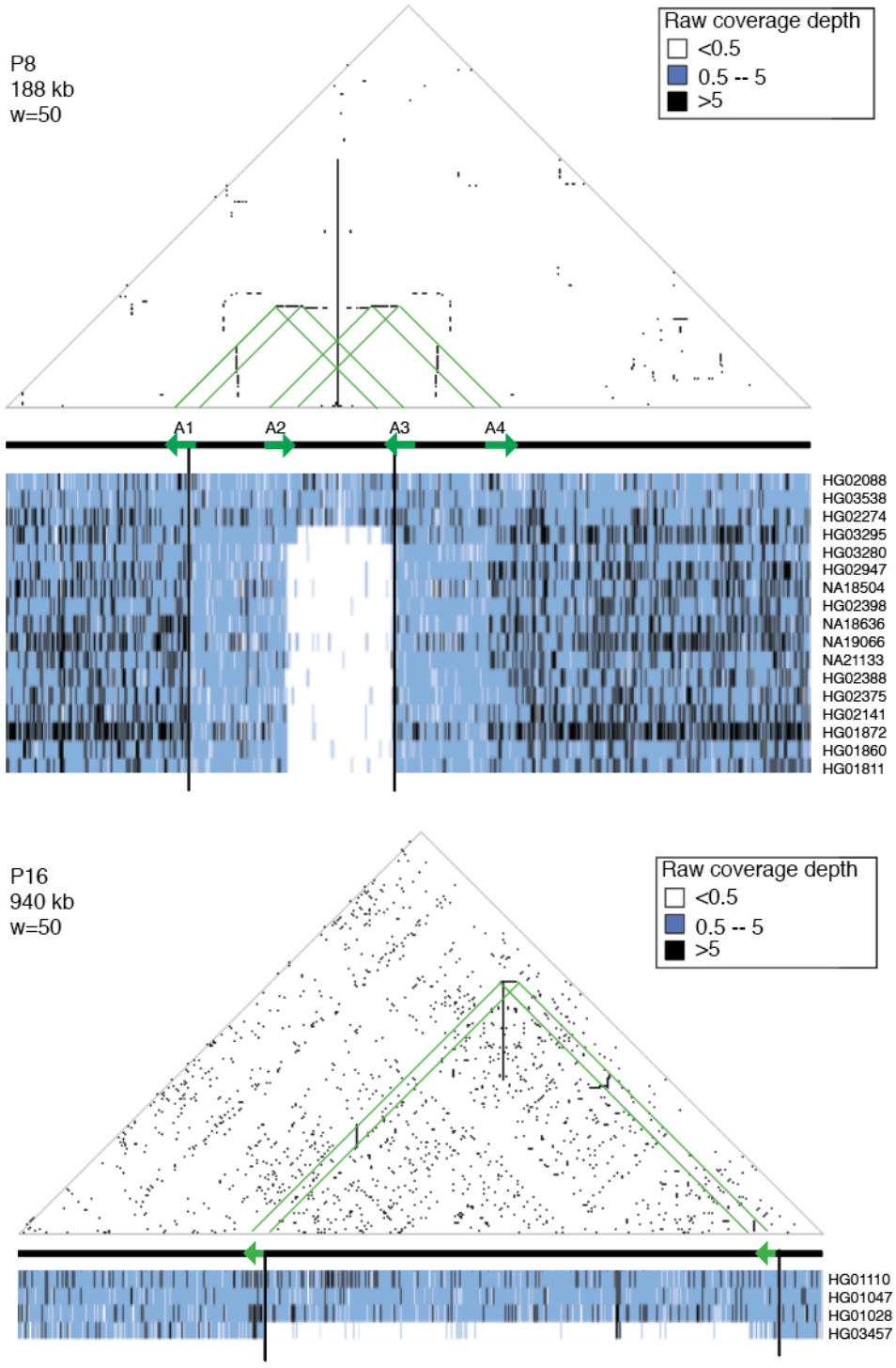

**Supplemental Figure S7.** Expression of gene families from palindromes shared by human, chimpanzee, and macaque in chimpanzee. Data from Brawand et al. 2011 was re-analyzed with kallisto. Each row shows averaged expression from one gene family. Row and column orders were determined by hierarchical clustering. Expression category: Shows whether expression is testis-biased (red) or broad (black) in the indicated species. Testis-biased: Minimum 2 TPM in testis, and testis accounts for >25% of log2 normalized expression summed across all tissues. Broad: All other expressed genes.



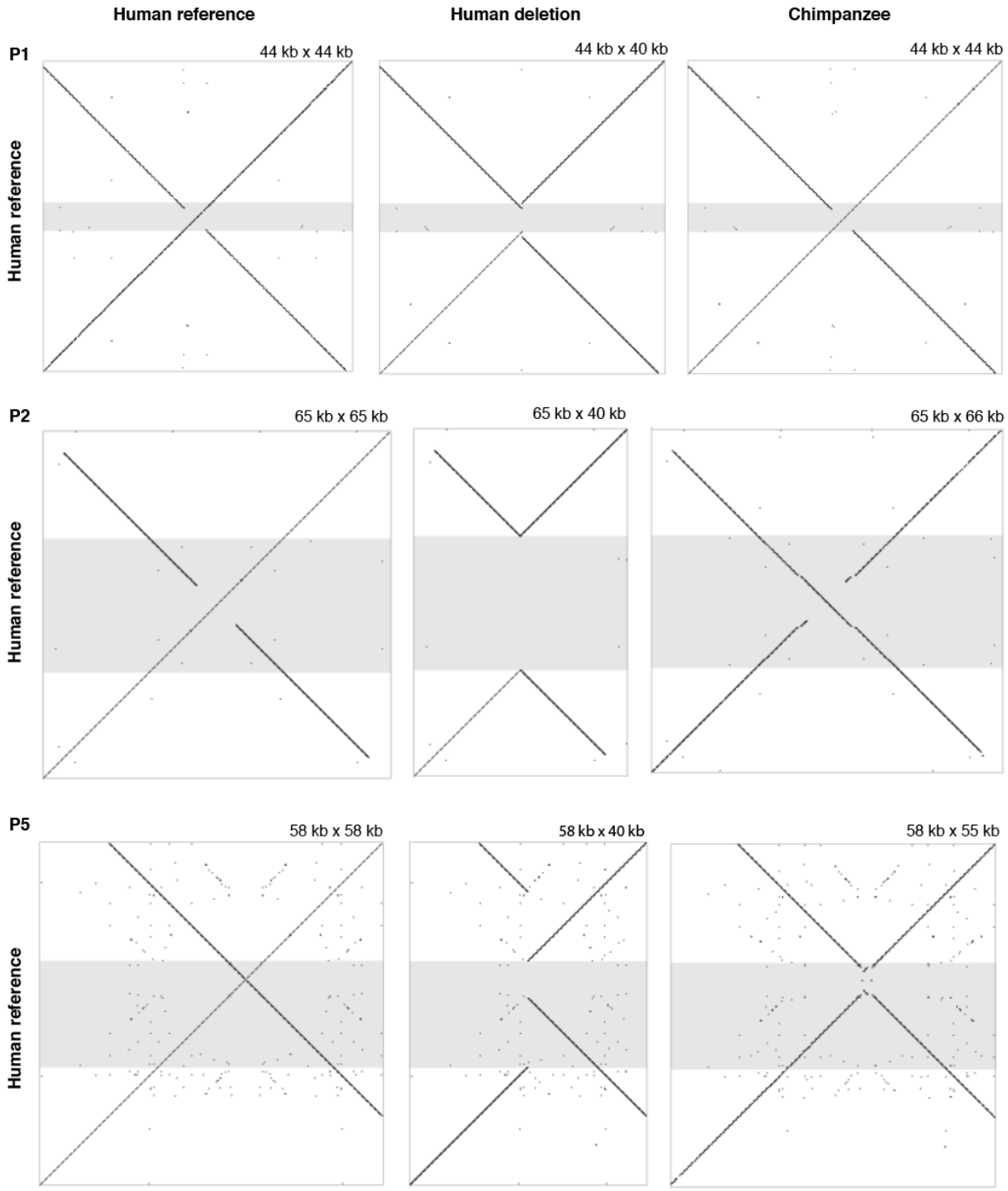

**Supplemental Figure S8.** Expression of gene families from palindromes shared by human, chimpanzee, and macaque in macaque. Data from Merkin et al. 2012 was re-analyzed with kallisto. Each row shows averaged expression from one gene family. Row and column orders were determined by hierarchical clustering. Expression category: Shows whether expression is testis-biased (red) or broad (black) in the indicated species. Testis-biased: Minimum 2 TPM in testis, and testis accounts for >25% of log2 normalized expression summed across all tissues. Broad: All other expressed genes.



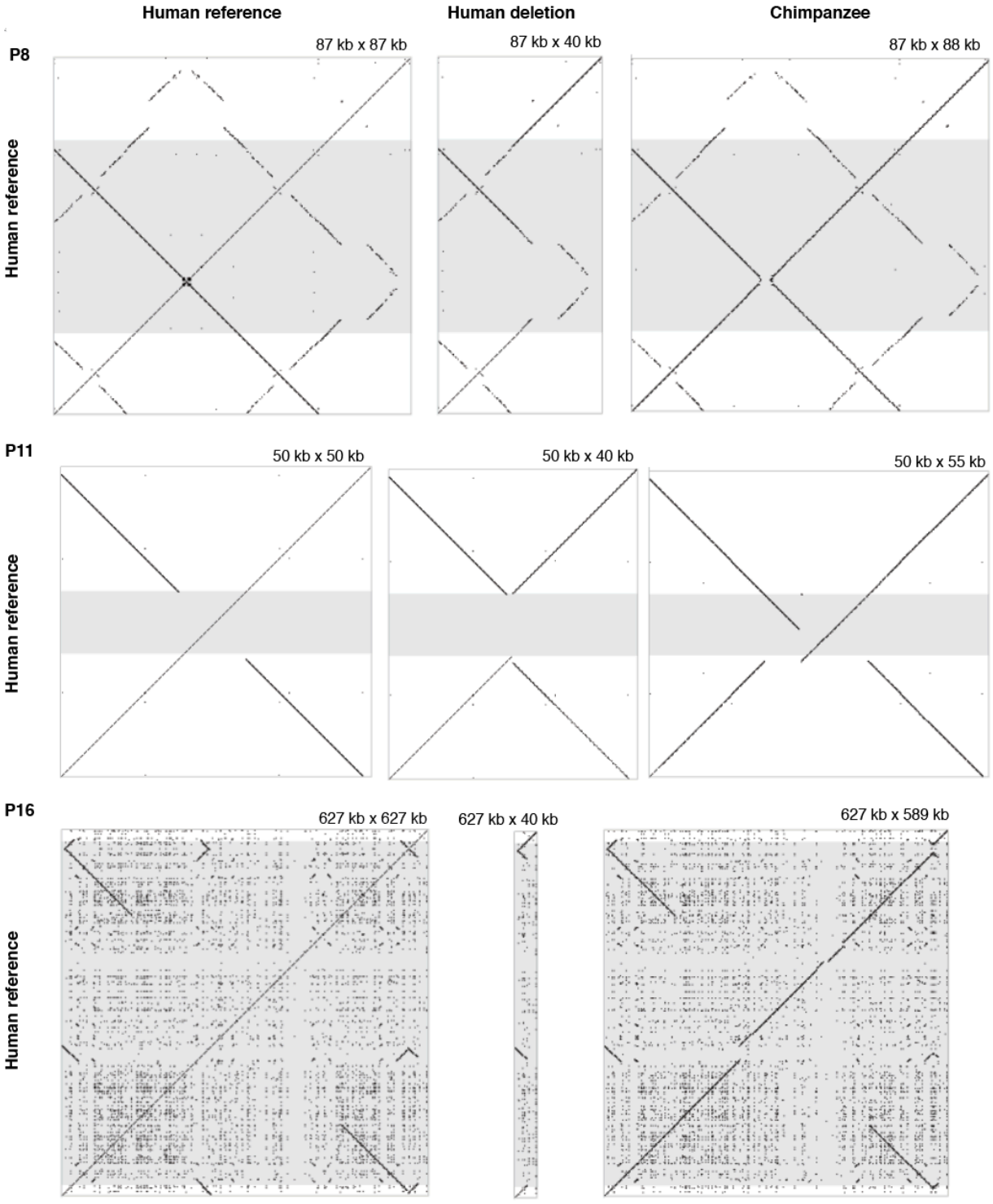

**Supplemental Figure S9**. Normalized coverage depths for eight palindrome spacers with at least one deletion in the 1000 Genomes dataset. Coverage depth for the ninth palindrome with spacer deletions, P2, is shown in Figure 6A.



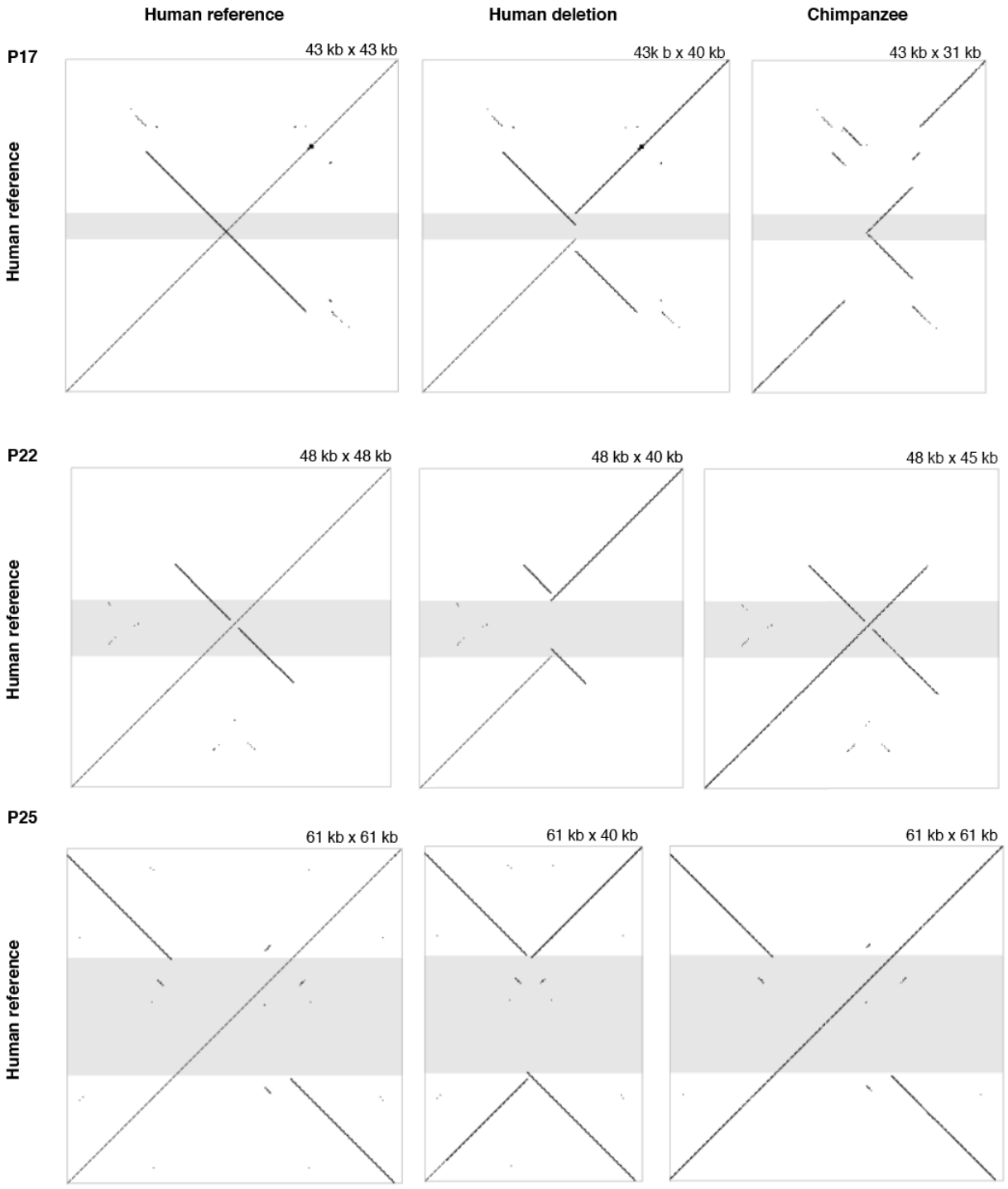

**Supplemental Figure S10.** Human spacer deletions with breakpoints within tandem repeats. Green arrows: Tandem repeats suspected to cause deletions through NAHR. In the case of P8, the deletion could have occurred with equal probability between arrows A1 & A3, or A2 & A4. Note that the suspected P8 deletion spans areas of no coverage (white) and reduced coverage (lighter blue, from approximately A1 to A2); the copy number of the lighter blue region is reduced from 2 to 1.



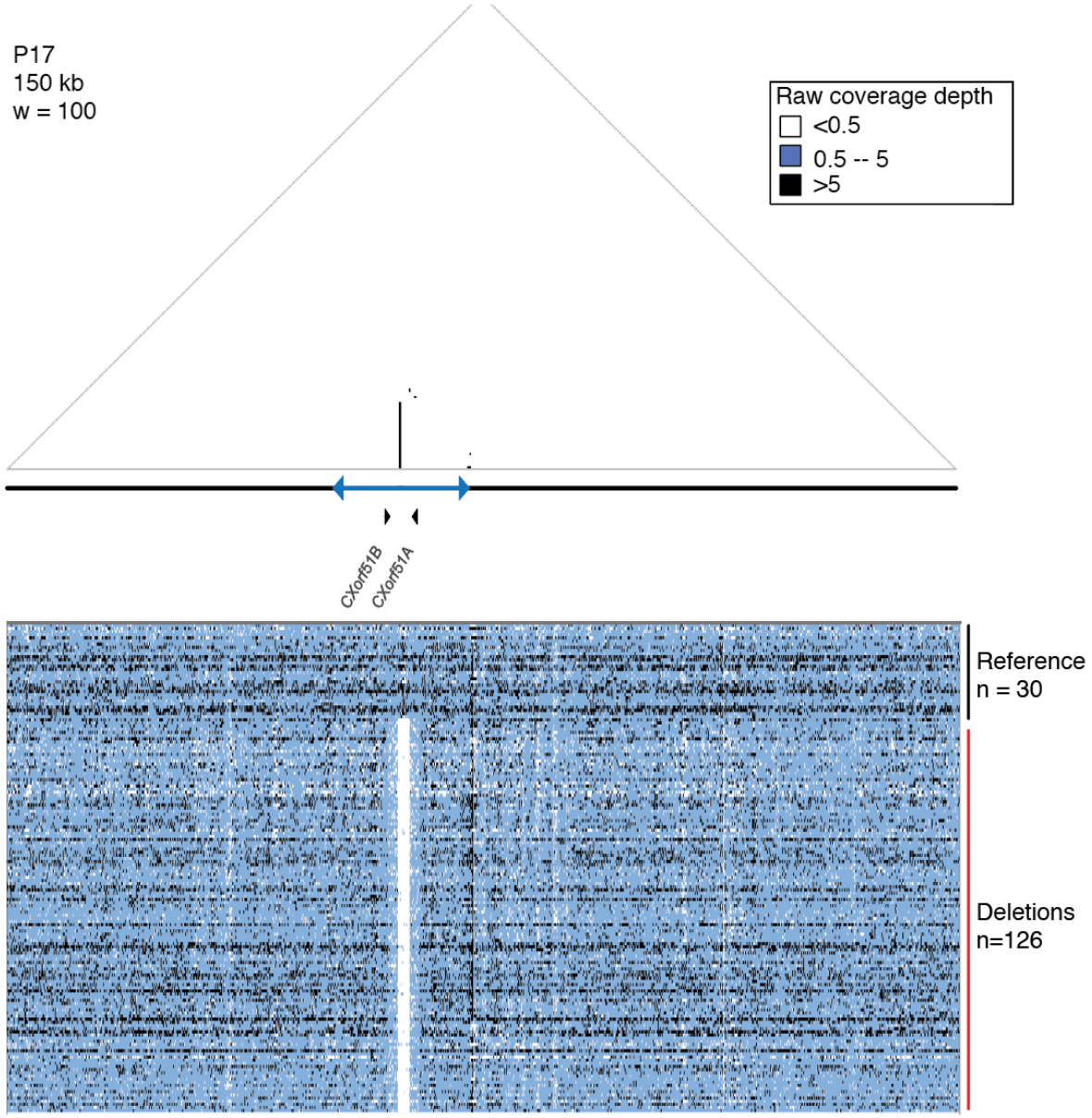

**Supplemental Figure S11.** Structural comparisons between human reference, human deletion, and chimpanzee for nine X palindromes with at least one spacer deletion. w=30 for all square dot plots. Position of the human X spacer deletion is highlighted in gray. For all nine palindromes, most or all of the of sequence absent in the human deletion is present in chimpanzee, confirming that the human structural polymorphism results from deletion rather than insertion.



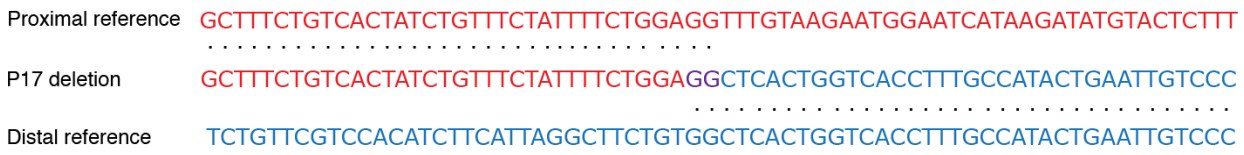

**Supplemental Figure S11.** Structural comparisons between human reference, human deletion, and chimpanzee for nine X palindromes with at least one spacer deletion. w=30 for all square dot plots. Position of the human X spacer deletion is highlighted in gray. For all nine palindromes, most or all of the of sequence absent in the human deletion is present in chimpanzee, confirming that the human structural polymorphism results from deletion rather than insertion.



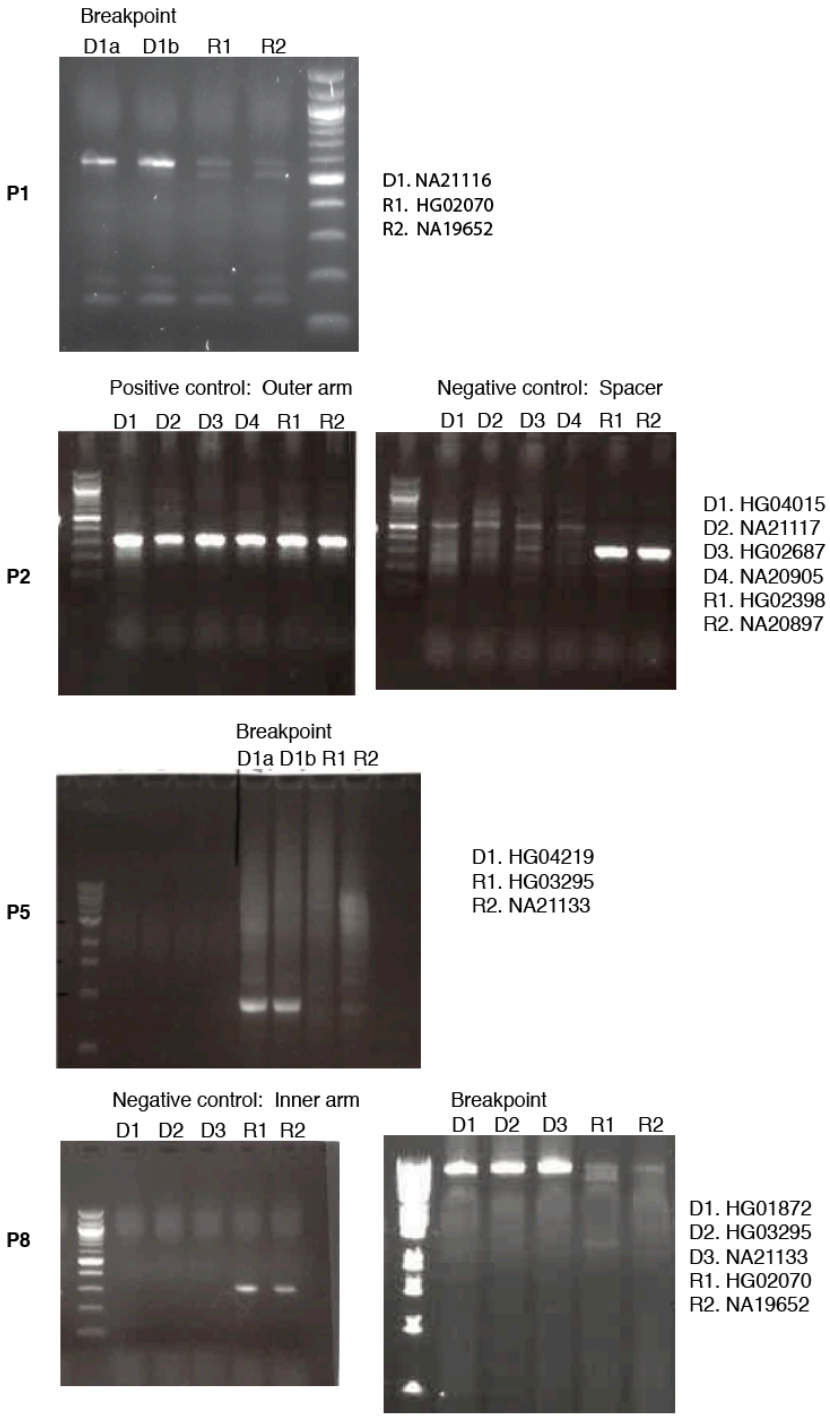

**Supplemental Figure S11.** Structural comparisons between human reference, human deletion, and chimpanzee for nine X palindromes with at least one spacer deletion. w=30 for all square dot plots. Position of the human X spacer deletion is highlighted in gray. For all nine palindromes, most or all of the of sequence absent in the human deletion is present in chimpanzee, confirming that the human structural polymorphism results from deletion rather than insertion.



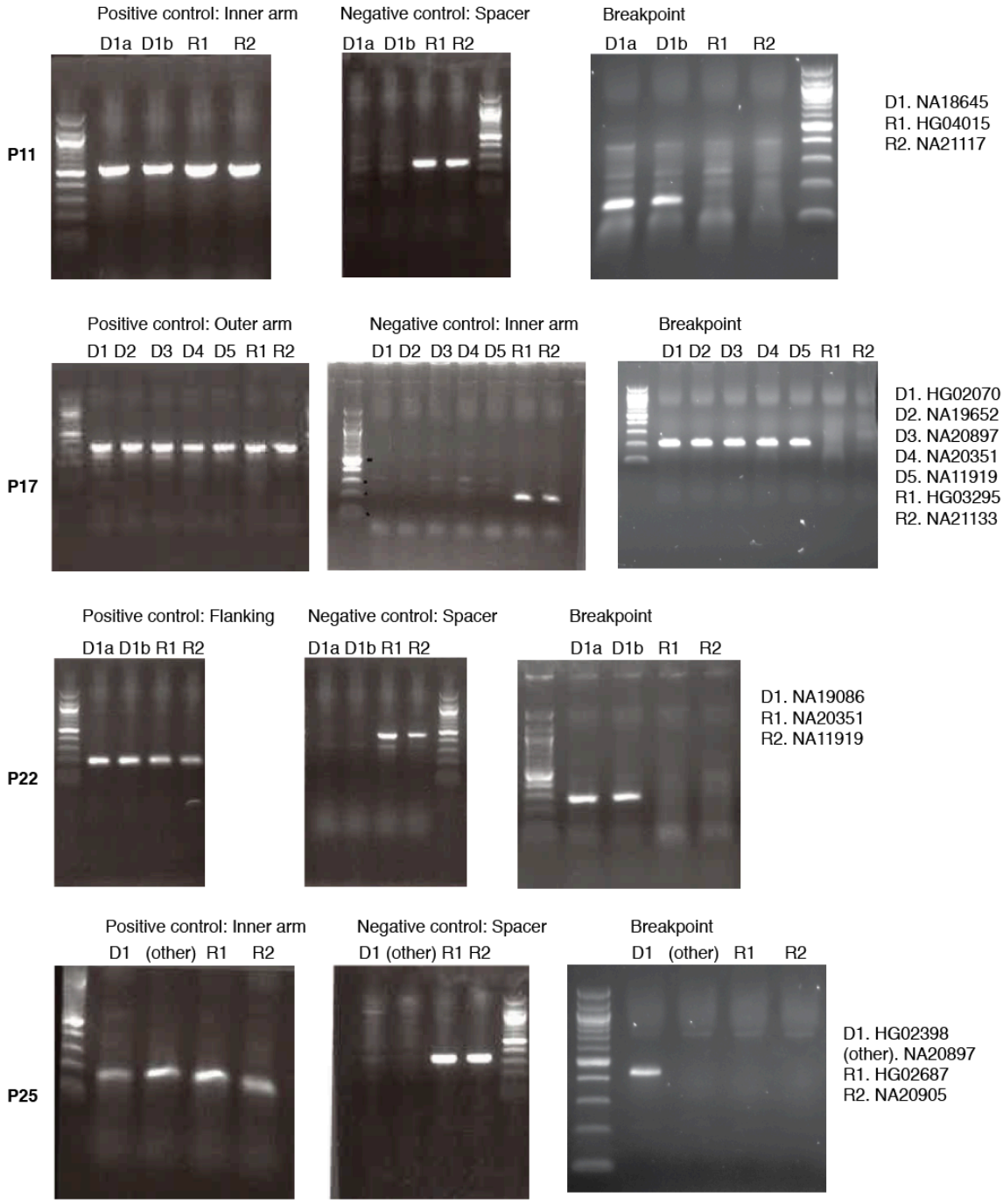

**Supplemental Figure S12**. Coverage depth for X Chromosomes with P17 spacer deletions. Tracks are shown for all 126 X Chromosomes with P17 spacer deletions, plus 30 randomly chosen X Chromosomes with the reference structure.




**Supplemental Figure S13.** Junction for P17 spacer deletion. Proximal reference: chrX: 146811296-146811365. Distal reference: chrX: 146814668-146814736. Two base pairs (GG, purple) overlap between the breakpoints.




**Supplemental Figure S14.** Verification of human X-palindrome spacer deletions. PCR primers were designed based on deletion breakpoints from split reads or, in cases where reads spanning the breakpoint could not be found, based on the estimated deletion breakpoints from visualization of coverage depth. Positive control: Sequence expected to be present in both reference samples and deletion samples. Negative control: Sequence expected to be present in reference samples, and absent in deletion samples. Breakpoint: Sequence expected to be present in deletion samples, and absent in reference samples. D = deletion sample, R = reference sample.




**Supplemental Figure S14.** Verification of human X-palindrome spacer deletions. PCR primers were designed based on deletion breakpoints from split reads or, in cases where reads spanning the breakpoint could not be found, based on the estimated deletion breakpoints from visualization of coverage depth. Positive control: Sequence expected to be present in both reference samples and deletion samples. Negative control: Sequence expected to be present in reference samples, and absent in deletion samples. Breakpoint: Sequence expected to be present in deletion samples, and absent in reference samples. D = deletion sample, R = reference sample.