SUPPLEMENTAL NOTES AND FIGURES

SUPPLEMENTAL NOTES

Note S1. Uncorrected PacBio reads in the assembly

We examined the impact of uncorrected PacBio reads on our assembly. PacBio uncorrected ≥10 kb and corrected reads were mapped to the best assembly with BLASR (Chaisson and Tesler 2012). Only approximately 0.9 Mb of the best assembly was covered by these reads added in the last assembly step by PBJelly (File S3).

Note S2. PacBio only assemblies

We also attempted to assemble the gorilla Y chromosome using only PacBio reads and the HGAP (Chin et al. 2013) and MHAP (Berlin et al. 2015). HGAP assemblies were generated with smrtpipe.py v1.87.139483 and (BETA) HGAP version 3 with default settings and genome size of 60 Mb. The computation was spread across a computational cluster with >3 nodes per assembly, resulting in final computational time of <9 days for each assembly. Since we used SLURM, a cluster management and job scheduling system not natively supported by SMRT analysis software, several technical modifications needed to be introduced to job specifications files and Celera specification file (runCA.spec).

We found the assembly results to be sensitive to the length threshold used for error-correction by HGAP. The default threshold, calculated by HGAP was 18 kb; however, we also tried several lower thresholds (Table S5). As expected, lower thresholds produced more corrected reads and larger final assemblies as compared with higher thresholds. However, the assemblies with lower thresholds significantly overestimated the size of the gorilla Y chromosome (up to 233 Mb for the lowest threshold, whereas the expected size is ~60 Mb (Gläser et al. 1998). Based on inflated sizes of ampliconic gene families observed in the non-18kb threshold assemblies during subsequent analysis (data not shown), we suspect that this is due to redundancy (e.g. contigs contained within other contigs). Therefore we conclude that the assembly with an automatic length threshold is more accurate than the assemblies with lower thresholds.

Additionally, we performed another PacBio only assembly using PBcR pipeline. Self-correction was performed with a probabilistic overlapping algorithm MHAP (Berlin et al. 2015) and error-corrected reads were assembled with Celera (version wgs-8.3rc2) (Myers et al. 2000) using following command:

```
/bin/PBcR -length 500 -partitions 63 -threads 63 -noclean -l ${name} -s
/nfs/brubeck.bx.psu.edu/scratch5/monika/MHAP_runs/pacbio.spec -fastq ${reads}
genomeSize=25000000
```

We used PacBio subreads after filtering against female (keeping hits to X chromosome), this resulted in 1,033,011 reads that were subject to error-correction.

Note S3. Experimental validations and rearrangement identification

Using BLASR we identified 425 contigs that map to human and/or chimpanzee Y chromosome but do not map to the gorilla female reference genome with 80% identity threshold. We randomly selected 50 out of these 425 contigs for experimental validations of exclusive presence in the male gorilla genomic DNA. 18 of them have to be disregarded from the final scoring since they were still present in female genomic DNA due to the pseudoautosomal/X-degenerate and/or autosomal location (80% identity threshold used).

QUAST is a tool that can identify breakpoints of an assembly relative to a reference genome. We ran QUAST on our assembly twice, once using the human Y (GRCh38) and once using the chimpanzee Y (gi|326910934|gb|DP000054.2| Pan troglodytes chromosome Y) as a reference. QUAST flagged 2,068 and 1,380 breakpoints, respectively. We further merged breakpoints from the same dataset that occurred within 100 bp of each other, obtaining 1,729 and 1,125 breakpoints, respectively. We next identified breakpoints that occurred in both datasets within 1,000 bp of each other, taking the leftmost one, to get 686 breakpoints. Scaffolds with breakpoints were then used as a target, to which raw reads from Illumina PE, MP, and PacBio long reads (>10-kb) were mapped using the bowtie2 (Langmead and Salzberg 2012) and BLASR aligners (Chaisson and Tesler 2012). Bedtools coverage was used to determine breakpoints where 50 bp upstream and downstream region of a breakpoint was supported by a minimum of 5x coverage of PacBio raw reads and 10x coverage of Illumina raw reads. This identified 121 candidate breakpoints for experimental validation. Apart from these 121 breakpoints, we found 41 breakpoints, which were supported by a minimum of 5x coverage of only PacBio raw reads without any Illumina support. To design suitable primers, two further conditions were imposed -- the breakpoints had to be at least 1,000 bp distant from the contig edges and contain at least 100 contiguous unique bases. Thus, we obtained 1,000 bp flanking regions from 29 breakpoints to test experimentally. Additionally, we also retrieved 1,000 bp flanking regions from 13 breakpoints supported by only PacBio reads and not Illumina reads.

Experimental validations of detected breakpoints were performed via PCR amplifications of gorilla male DNA (or WGA flow-sorted gorilla Y DNA) using the following conditions: initial denaturation at 94°C for 2 min, 30 cycles of denaturation at 94°C for 1 min, annealing at primer specific temperature for 30 sec, extension at 72°C for 45 sec to 2 min; followed by a final extension at 72°C for 5 min. PCR reaction mixtures consisted of: 20 ng of DNA, 1 unit of ChoiceTaq DNA polymerase (Denville Scientific), 10x PCR buffer, 1.5 uM $MgCl_2$ (Denville Scientific), 500 uM dNTPs (Roche), 1.25 uM of each primer, and water brought to the final volume of 25 uL.

Note S4. Gene prediction

We applied the *ab initio* gene finding tool AUGUSTUS (Stanke et al. 2008) on our best assembly. Using AUGUSTUS (Stanke et al. 2008), we were able to predict candidate genes on the gorilla Y. We used human genes as the training set for AUGUSTUS. AUGUSTUS predicted 565 partial or complete genes. These included 21 Y-specific non-PAR protein-coding genes and eight PAR genes (Table S20), confirming the validity of this approach. We limited our analysis to nine other predicted genes (1) expressed in gorilla testis, (2) aligning to neither the gorilla pseudogenes (Cortez et al. 2014) nor the gorilla female genome; (3) aligning to the homologous human RefSeq genes; and (4) whose translated sequences aligned to protein databases (Swissprot, trEMBL or nr). Among these nine predicted genes, four genes (homologs of *KAL1*, *GGT1*, *TPTE*, and *FRG1*) were of particular interest (Table S20) because they were supported by the RNA-seq data and differred substantially from the female genome homologs. However, Y chromosome homologs of *KAL1* and *FRG1* were found to be pseudogenes, and Y chromosome homologs *GGT1* and *TPTE* contained only partial sequence compared to their autosomal homologs (Table S20).

Note S5. Palindrome identification

Extracting human palindromes. First, to recover from the best assembly the sequences homologous to human palindromes P1-P8, we extracted palindrome arms and spacers based on a dotplot analysis (gepard-1.30) from the GRCh38 release of the human genome. The lengths of arms and spacers extracted corresponded to those previously reported (Skaletsky et al. 2003) with the exception of palindrome P2 (Fig. S7A, Table S14). This palindrome exhibited highly repetitive structure near the arm-spacer boundary, which might explain the lowest recovery of this palindrome in our assembly. For initial screening, we used the same script as for testing the gene presence for long palindromes homologous to human (see Methods, Fig. 2C). In a complementary analysis, we mapped 288,512,424 paired-end reads from our Y-flow sorted reads and 126,746,822 reads from the male genomic library on the repeat-masked arms and spacers (one arm and one spacer per palindrome). We then mapped the best assembly to the same reference (Table S14).

Coordinates of extracted palindromes from the human assembly (GRCh38):

Palindrome	Start coordinate	End coordinate
P1	23359067	26311550
P2	23061889	23358813
P3	21924954	22661453
P4	18450291	18870104
P5	17455877	18450126
P6	16159590	16425757
P7	15874906	15904894
P8	13984498	14058230

Extracting chimpanzee palindromes. We extracted all 19 palindromes from the chimpanzee reference (gi|326910934|gb|DP000054.2) from *p* arm to *q* arm, keeping tandem palindrome array in the proximity of centromere consisting of six palindromes as a single unit.

Coordinates of extracted palindromes, the homology to human palindromes shown in bold (if present):

Palindrome	Start coordinate	End coordinate	Human homologs
C1	1,759,451	2,053,069	
C2	2,298,081	2,984,818	
C3	**3,587,737**	**3,925,944**	**P1 and P2**
C4	**4,669,973**	**5,439,450**	**P5**
C5-C10_array	8,627,160	10,826,862	
C11	11,035,665	11,649,875	
C12	**12,627,411**	**12,938,743**	**P1 and P2**
C13	**13,315,586**	**14,060,184**	**P5**
C14	14,773,730	14,999,930	
C15	15,471,778	16,034,847	
C16	16,455,342	16,769,765	
C17	**21,569,785**	**21,663,227**	**P8**
C18	**23,510,295**	**23,539,175**	**P7**
C19	**23,799,408**	**24,569,752**	**P6**

There are 12 chimpanzee-specific palindromes (in comparison to human Y chromosome) that form two groups based on sequence similarity.
group1: C2 ~ C11 ~ C15
group2: C1 ~ C5-C10 ~ C14 ~ C16

Extracting gorilla-specific palindromes. For new, gorilla-specific palindromes that are fully contained within a single scaffold, we developed a script that aligns our best assembly against reverse complement of the same assembly (using blastn, BLAST 2.2.29+). Then we parsed the hits in order to identify a palindrome-like pattern -- both arms at least 3,000 bp long mapping to each other with the identity at least 95%. This resulted in the initial dataset of 60 candidates. Subsequently, we analyzed arms, a spacer, and flanking regions of such palindromes (two best hits to the female using BLASR, best blast hit against nucleotide database online, percentage of repeat-masked regions). First, we manually filtered out all palindromes that could originate from the autosomal contamination (if majority of hits for arms, spacer and two flanking regions were mapping to the female and the best blast hit was not of the Y chromosome origin). Second, we removed palindromes with a spacer longer than each arm and those with repeat content per arm and spacer over 95%, as well as cases when one repeat family dominated the whole arm (e.g. those dominated by LINE elements). We additionally excluded cases not supported by at least one PacBio read (spanning the arm-spacer-arm boundary).

Palindrome read depth analysis

All paired-end reads from flow-sorted material (including those resulting from MP reads) were mapped to the assembly with bwa mem (Li and Durbin 2009) using default parameters. The number of reads mapped to each scaffold was then divided by the scaffold length.

We identified scaffolds belonging to each palindrome as follows: gorilla assembly was split into 1-kb windows and for each window, the best alignment to human/chimpanzee was retrieved. All windows with the alignments >800 bp contributed to the final set of scaffolds per each palindrome. All boxplots were plotted using R package boxplot {graphics} with outline=FALSE.

Note S6. Optimal Illumina sequencing amount

In order to reduce the dataset for assembly and make computations more manageable, *in silico* normalization was performed on the post-RecoverY reads with Trinity (Haas et al. 2013). This reduced the dataset by ~80%, and the resulting short reads were assembled and then augmented with PacBio reads to obtain the best assembly. Thus, it was concluded that we initially sequenced at an unnecessary high depth (227x and 250x for PE and MP libraries, respectively, Table 1), and that 80% of the Y chromosome-specific data could be removed to obtain our best assembly. It is then suggested that 50x coverage of each of paired-end and mate pair Y-specific data is a reasonable estimate for sequencing of enriched (flow-sorted) data. This is also in keeping with the recommendation of 100x coverage of combined datasets for short read assembly (Gnerre et al. 2011).

Note S7. Selection tests

We used the codeml module of PAML (version 4.8) (Yang 1997) to detect branch-specific differences in the nonsynonymous-to-synonymous rate ratios and to test for positive selection acting on 16 X-degenerate genes and four ampliconic genes shared by human, chimpanzee, gorilla, and macaque. Coding sequences of Y-chromosomal X-degenerate genes (named "Y") and their X homologs (named "X") were retrieved from GenBank and aligned using ClustalW (Larkin et al. 2007). The consensus sequence has been generated for each multi-copy ampliconic gene family for each of the species. The phylogenies were generated with the Neighbor-Joining method (Saitou and Nei 1987) (with 1,000 bootstrap replicas) as implemented in Mega6 (Tamura et al. 2013). We excluded chimpanzee pseudogenes (that had no active genes in chimpanzee) from the analysis. First, for each gene, the one-ratio model (assuming the same nonsynonymous-to-synonymous rate ratio ωn for the entire tree) was compared with the two-ratio model (assuming that the branch-specific ratio ωs is different from the background ratio ωo). When the difference between the two models was significant, this indicated that the synonymous-to-nonsynonymous rate ratio was different for the branch tested.

For X-degenerate genes, among the gene-branch combinations with significantly higher nonsynonymous-to-synonymous rate ratio as compared to the background, 10 such ratios were greater than 1 (Table S18). Notably, five of such cases represented genes on the chimpanzee Y chromosome (*DDX3Y, EIF1Y, PRKY, SMCY* and *SRY*; Table S18). This could be due to adaptation or relaxed selective constraints acting upon these genes in the chimpanzee lineage. To test for positive selection, for each of these 10 cases we compared a model with estimated branch-specific ω against an alternative model with branch-specific ω fixed at 1. Significant differences were observed in only one case, and this difference became non-significant after Bonferroni correction for multiple testing (Table S18).

Similar analysis was performed for ampliconic genes (Table S19). In two cases (for *CDY* in chimpanzee and *RBMY* in macaque), the branch-specific ω were significantly higher than the background ω and also higher than 1. However, when comparing a model with estimated branch-specific ω against an alternative model with branch-specific ω fixed at 1, the results were non-significant providing no evidence for positive selection.

REFERENCES

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**: 623–630.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**: 238.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563–569.

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**: 488–493.

Gläser B, Grützner F, Willmann U, Stanyon R. 1998. Simian Y chromosomes: species-specific rearrangements of DAZ, RBM, and TSPY versus contiguity of PAR and SRY. *Mammalian Genome* **9**: 226–231.

Gnerre S, MacCallum I, Przybylski D. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* **108:** 1513–1518.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–1512.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biology* **14**: R47.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al. 2000. A whole-genome assembly of Drosophila. *Science* **287**: 2196–2204.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**: 406–425.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**: 825–U2.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637–644.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* **30**: 2725–2729.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555–556.

SUPPLEMENTARY FIGURES

Figure S1. Bivariate flow karyotype of the gorilla male cell line with all the chromosome clusters identified. The Y chromosome peak is clearly distinguished.

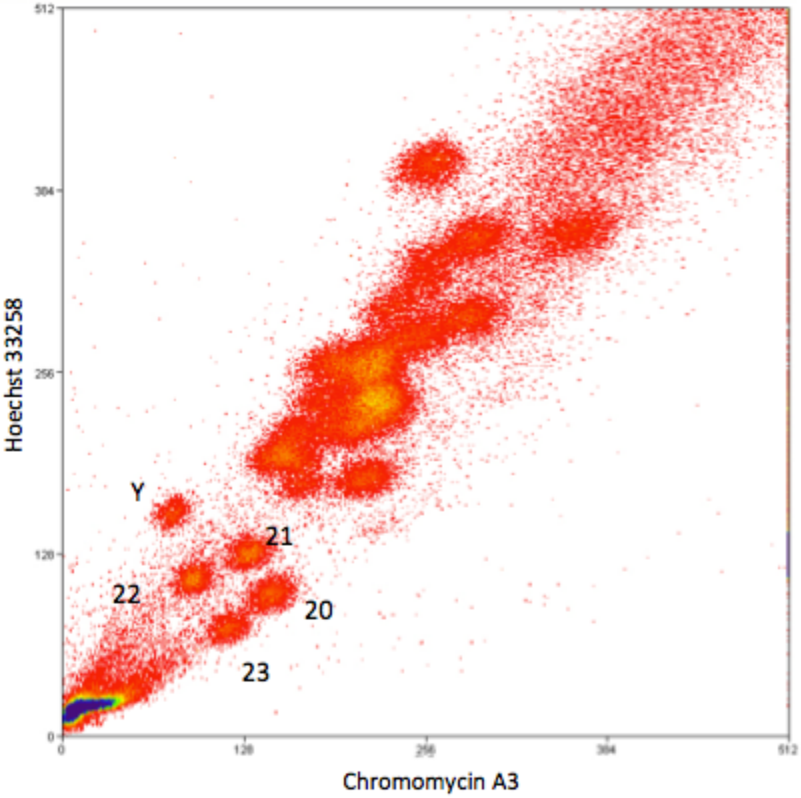


Figure S2. The evaluation of the quality, quantity, and size distribution of the whole-genome amplified (WGA) and debranched (see Methods) flow-sorted gorilla Y DNA. (**A**) MassRuler High Range DNA Ladder (Thermo Scientific). (**B**) Lambda phage DNA, 20 ng. (**C**) Lambda phage DNA, 120 ng. (**D**) Negative control. (**E**) WGA of human DNA. (**F, I**) Human genomic DNA, 50 ng. (**G**) WGA flow-sorted gorilla Y DNA, 50 ng. (**H**) Debranched WGA human DNA, 50 ng. (**J**) Debranched WGA flow-sorted gorilla Y DNA, 50 ng.

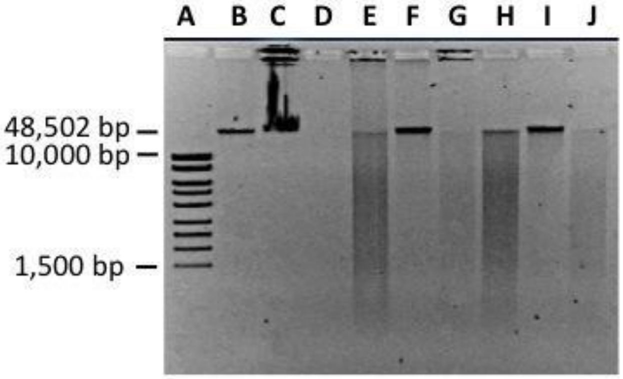


Figure S3. PCR screening of the whole-genome amplified flow-sorted gorilla Y DNA with four Y-chromosomal markers, as an initial test of Y-chromosome enrichment. (**A**) ***DAZ* amplified with 18219FY and 19599RY** (Makova and Li 2002). A1: 1 kb DNA ladder (NEB); A2: Negative control for PCR; A3: positive control, human male gDNA; A4: negative WGA control; A5: positive control, WGA human male gDNA; A6: WGA flow-sorted gorilla Y DNA. (**B**) ***DBY* amplified with DBY7-9-F1 and DBY7-9-R1** (Goetting-Minesky and Makova 2006). B1: 100 bp DNA ladder (NEB); B2: Negative control for PCR; B3: positive control, human male gDNA; B4: WGA gorilla Y DNA. (**C**) ***UTY* amplified with three primers: UTX/UTY (an X/Y universal primer), UTY (a Y specific primer amplifying a short fragment) and UTX (an X specific primer amplifying a longer fragment)** (Villesen and Fredsted 2006). C1: 100 bp DNA ladder (NEB); C2: Negative control for PCR; C3: positive control, human male gDNA; C4: negative WGA control, water; C5: positive control, WGA human male gDNA; C6: WGA gorilla Y DNA. (**D**) ***ZFY* amplified with ZFXntlF1 and ZFXntlR1** (Goetting-Minesky and Makova 2006). D1: 100 bp DNA ladder (NEB); D2: Negative control for PCR; D3: positive control, human male gDNA; D4: negative WGA control, water; D5: WGA gorilla Y DNA.

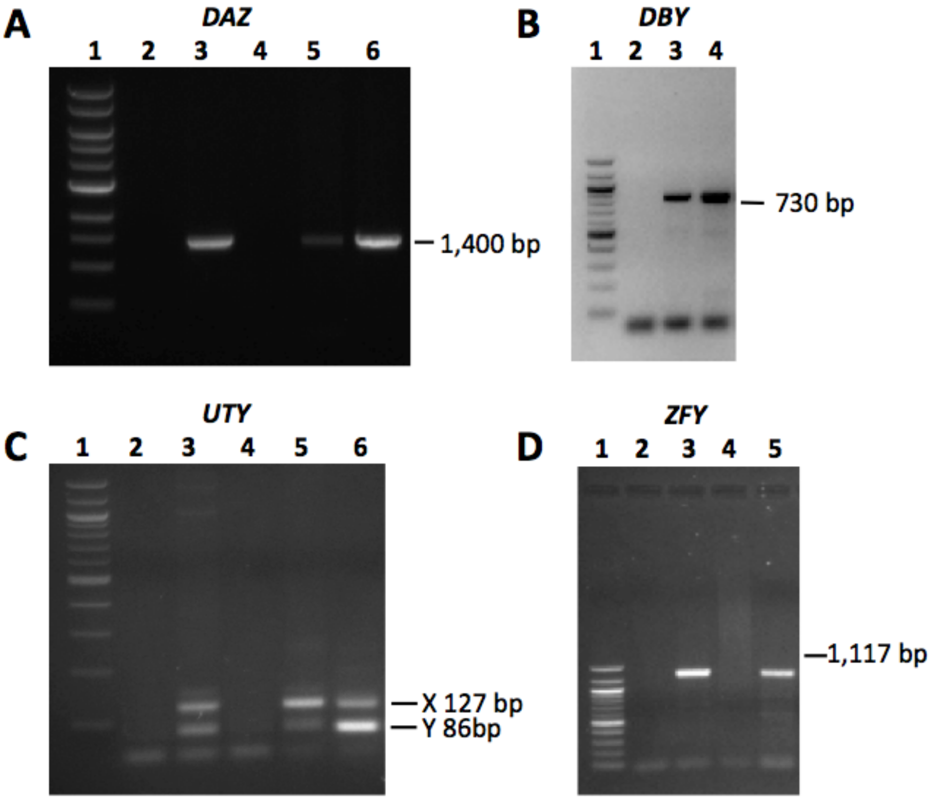


Figure S4. PacBio data generated. (**A) A histogram of the number of base pairs generated per SMRT cell (in Mb) using P4-C2 (blue) and P5-C3 (red) PacBio chemistries.** (**B) The distribution of subread lengths for both chemistries.** The X-axis represents subread lengths and the left Y-axis subread counts. The fitted line links subread length to the number of Mb greater than subread length. The median subread length was 4171 bp and 4269 bp for P4-C2 and P5-C3, respectively. (**C) The read quality profile for both chemistries.** The X-axis represents read accuracy and the left Y-axis read counts. The peak is at the 86% accuracy (14% estimated error rate). Plots (B) and (C) are generated directly by smrtanalysis-2-3-0 software from PacBio.

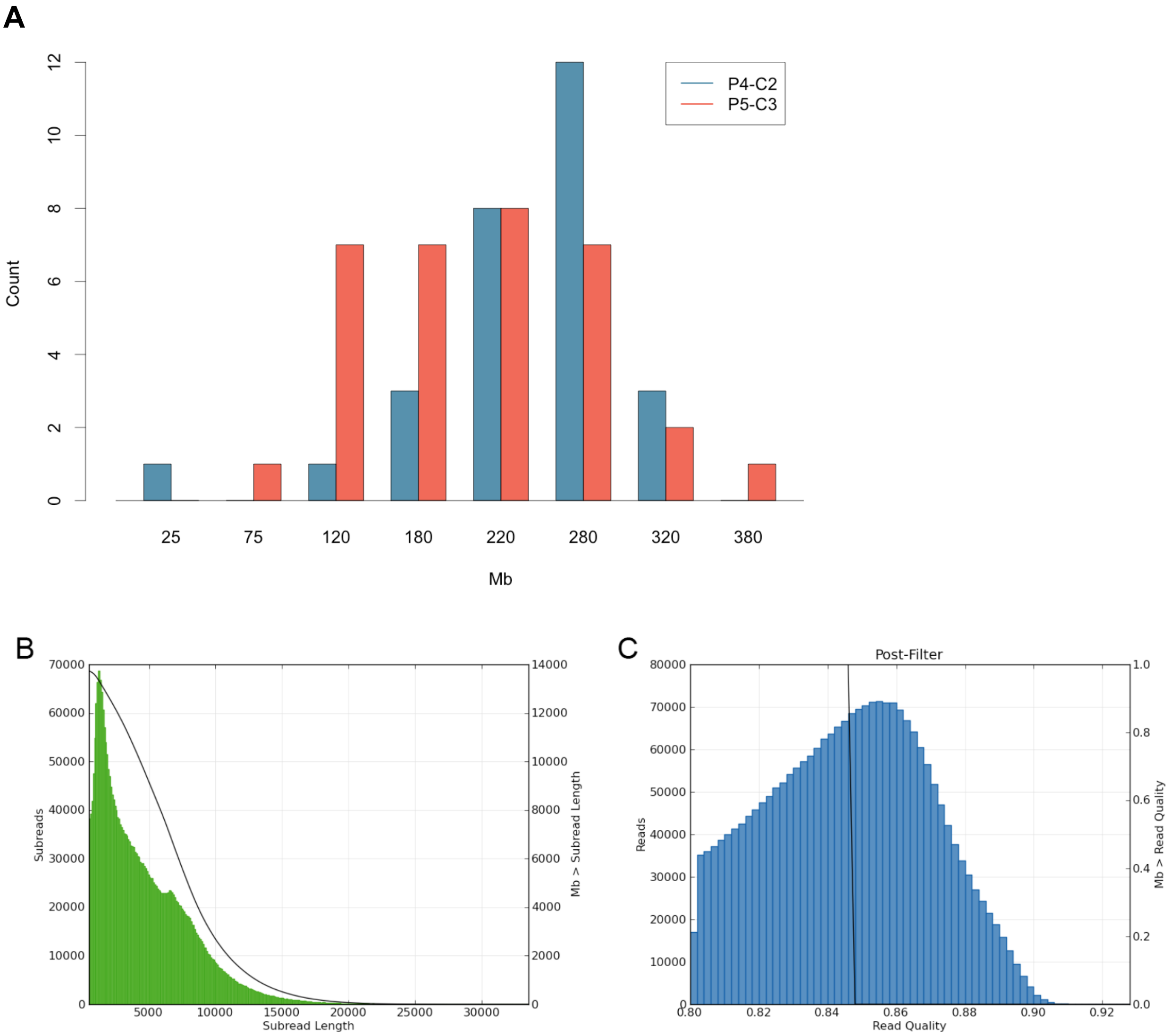


Figure S5. Examining WGA bias by depth analysis. In order to elucidate potential artifacts of WGA, we examined the depth profiles of X-degenerate and ampliconic scaffolds in the best assembly**.** These scaffolds contain exons from X-degenerate or ampliconic genes identified by aligning gorilla transcripts from (Cortez et al. 2014), gorilla coding sequences from (Goto et al. 2009), and gorilla transcripts assembled here from testis RNAseq data (see details in Methods). These exons have >95% identity (average 99.8%) to the gorilla transcripts and have full coding potential. All paired-end reads -- either from flow-sorted or from genomic material -- were mapped to the Repeatmasked best assembly with bowtie2, allowing for up to 8 mappings per each read (bowtie2 --fr -p 63 --no-unal -k 8). The number of reads mapped to each scaffold was then divided by the scaffold length and evaluated separately for reads originating from (A) flow-sorted material (including WGA step) and (B) male genomic reads (no WGA). The tight range of read depth for X-degenerate scaffolds and high similarity of (A) and (B) profiles suggests absence of WGA bias. A total of 26 ampliconic and 45 X-degenerate scaffolds are plotted (R package boxplot {graphics}, outline=FALSE).

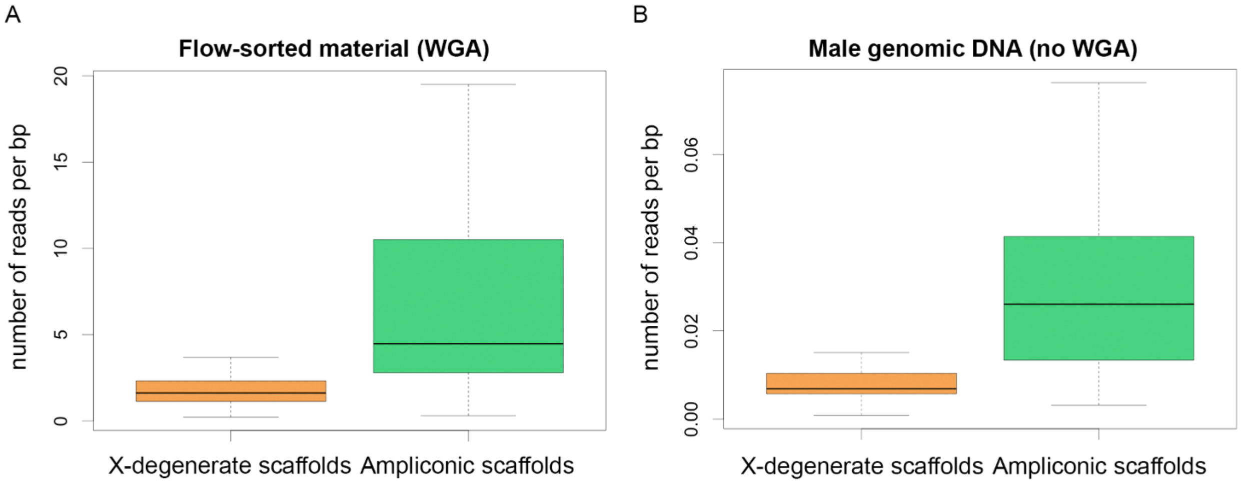


Figure S6. Overrepresentation of the non-Y chromosomes in the debris from flow-sorting. We mapped all PacBio reads to the gorilla female reference (gorGor3.1.74) and normalized the counts per chromosome lengths. The area of the yellow circles is proportional to the amount of the reads mapped and the location of the circles indicates GC content (y-axis, Ns were excluded from the calculation) and chromosome size (x-axis). The GC content and size of Y chromosomes of gorilla, chimpanzee (Pan), and human are plotted as orange triangles. The most overrepresented chromosomes are X (likely due to mapping errors due to its homology to Y), chromosome 10, and two small chromosomes similar in size to that of Y chromosomes (21 and 22). The reads were mapped using blasr using -bestn 1 setting and each line of the output was counted as one hit.

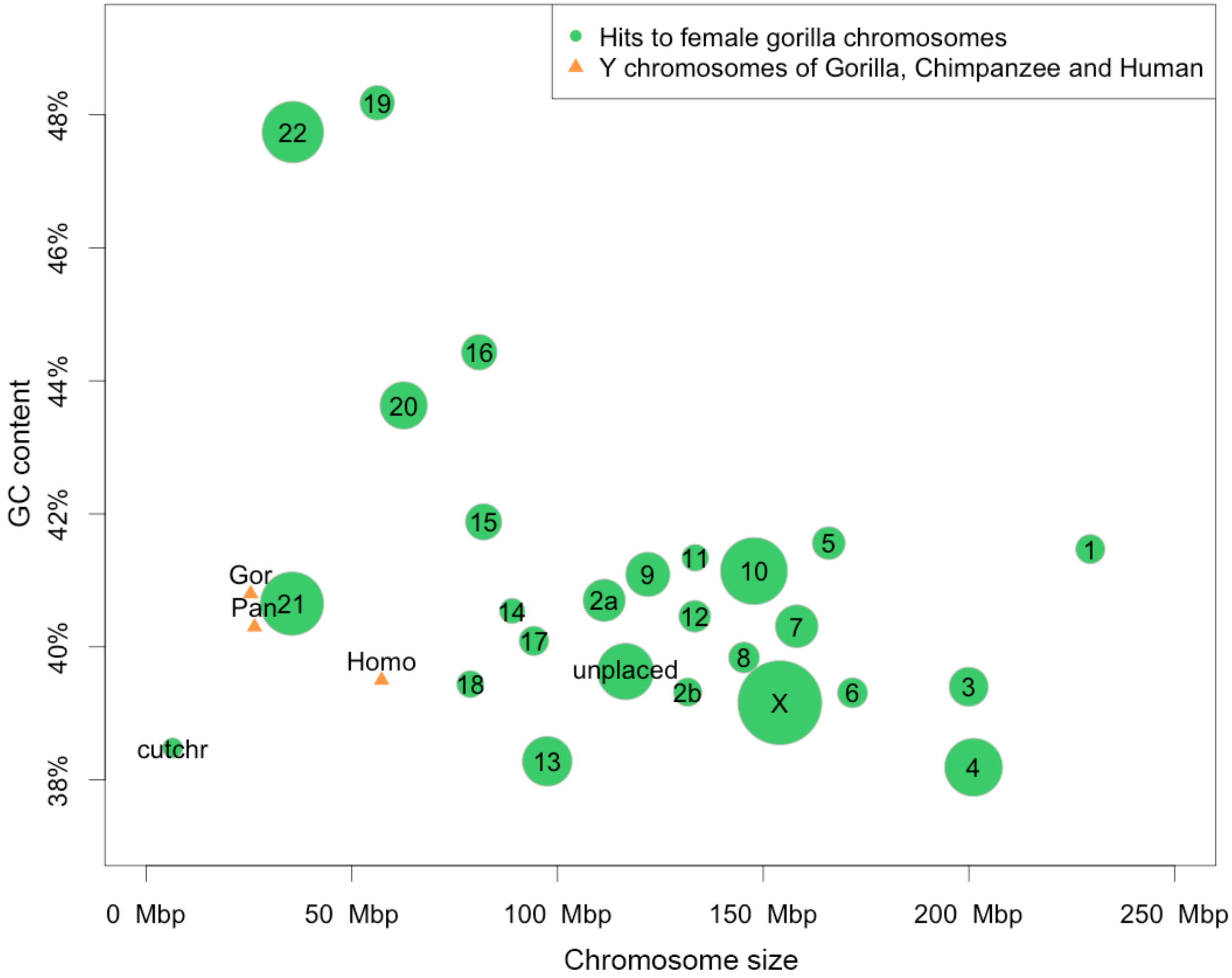
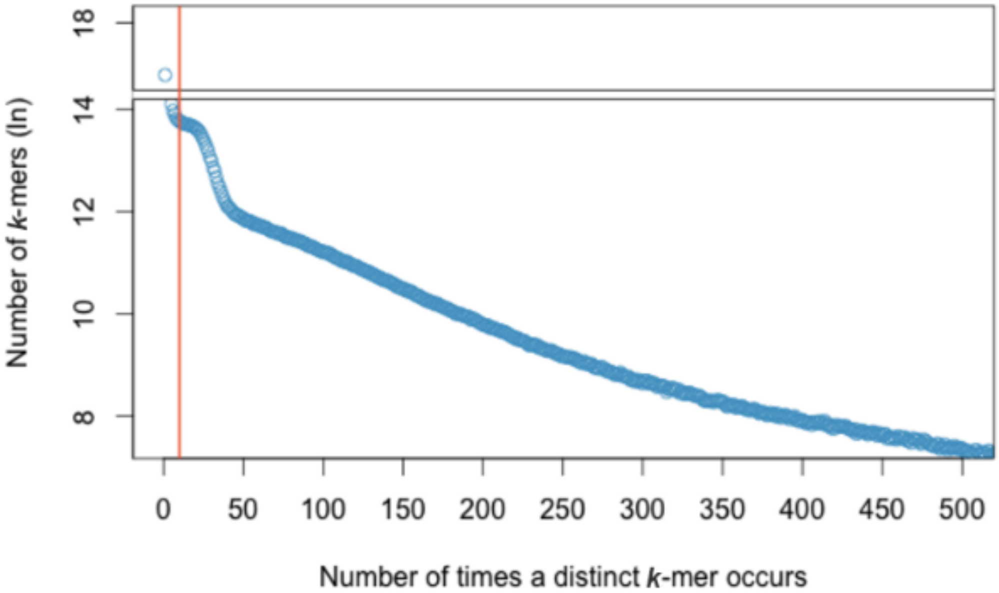


Figure S7. *K*-mer abundance distribution for gorilla flow-sorted Y mate pair data, (see Methods). Two libraries were evaluated. **(A) Library GY19** The threshold for RecoverY was chosen at 10x. **(B) Library MPGY.** The threshold for RecoverY was chosen at 50x.

A



B

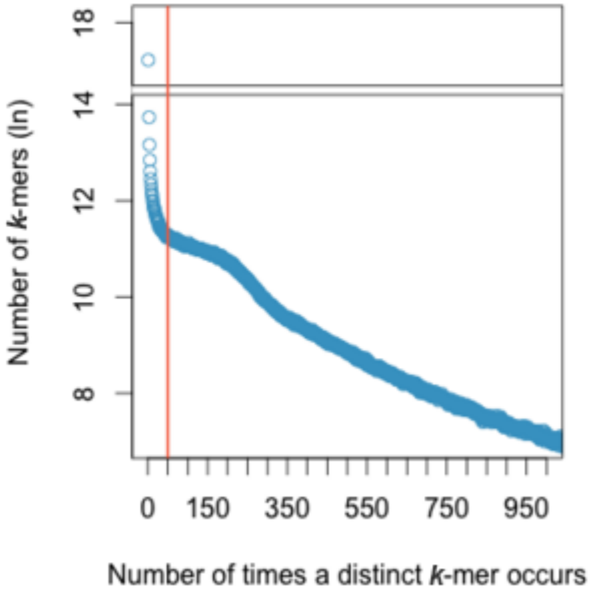
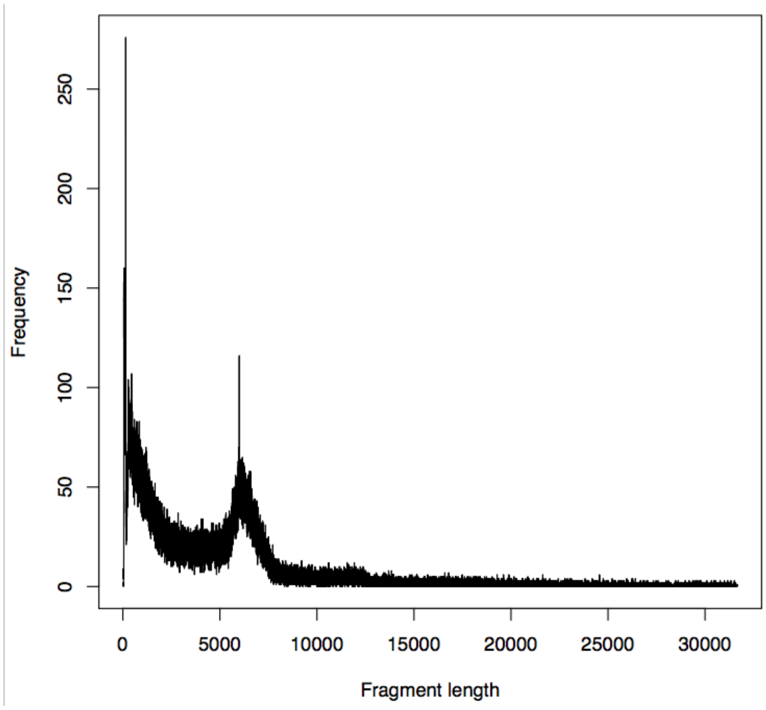
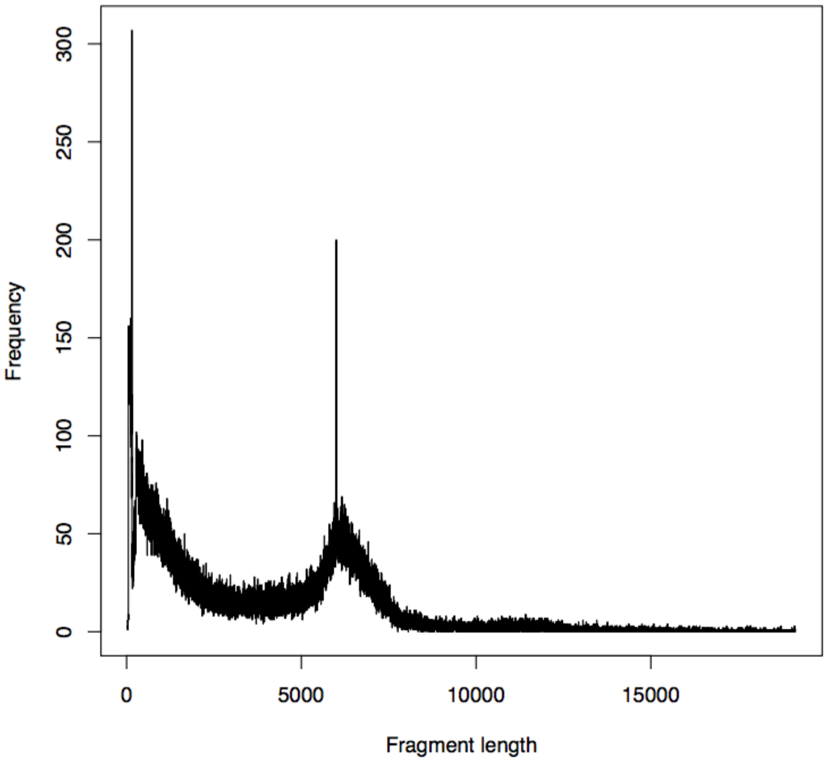


Figure S8. A. The insert size distribution of MP data mapping to the best assembly. Using our best assembly as a reference, we mapped reads from an MP library utilized in the scaffolding step. This analysis was performed using the assembly validation tool REAPR (Hunt et al. 2013), which reported that 22.95% of our data mapped with an insert size below 1,000 bp (indicating potential paired end contamination still remaining after running NxTrim); 10.57% of our data mapping at an insert size above 10 kb, which could indicate chimeric scaffolds; 66.48% mapped with an insert size of 1 kb to 10 kb. **B. The insert size distribution of MP data mapping to SSPACE scaffolds** (without PacBio data added) and **C. The insert size distribution of MP data mapping to BESST scaffolds.** Scaffolds from SSPACE were evaluated using mapping of MP and comparison to an independent scaffolder, BESST (Sahlin et al. 2014).

A



B


Frequency
0
50
100
150
200
250
300
0
5000
10000
15000
Fragment length

C

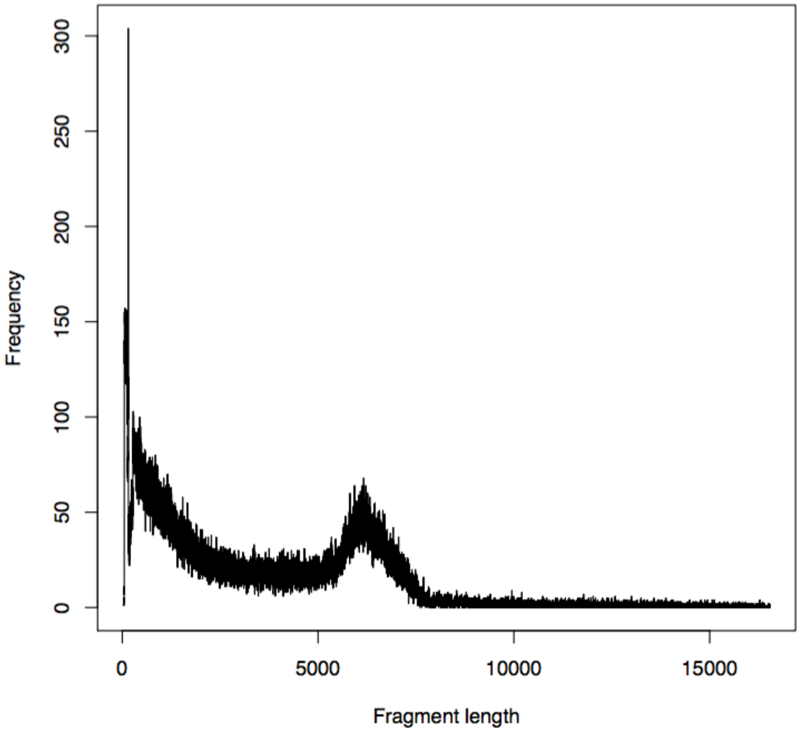
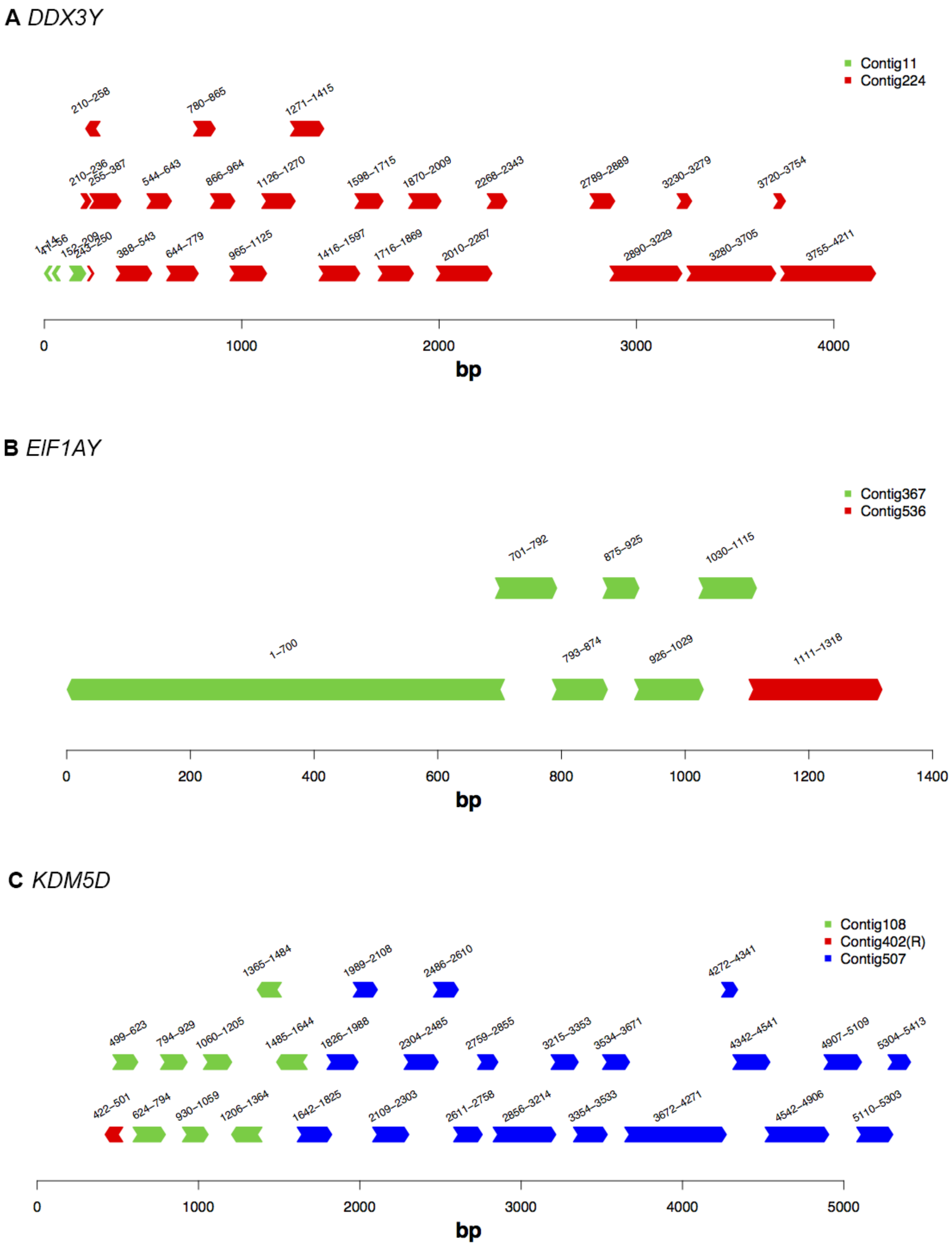
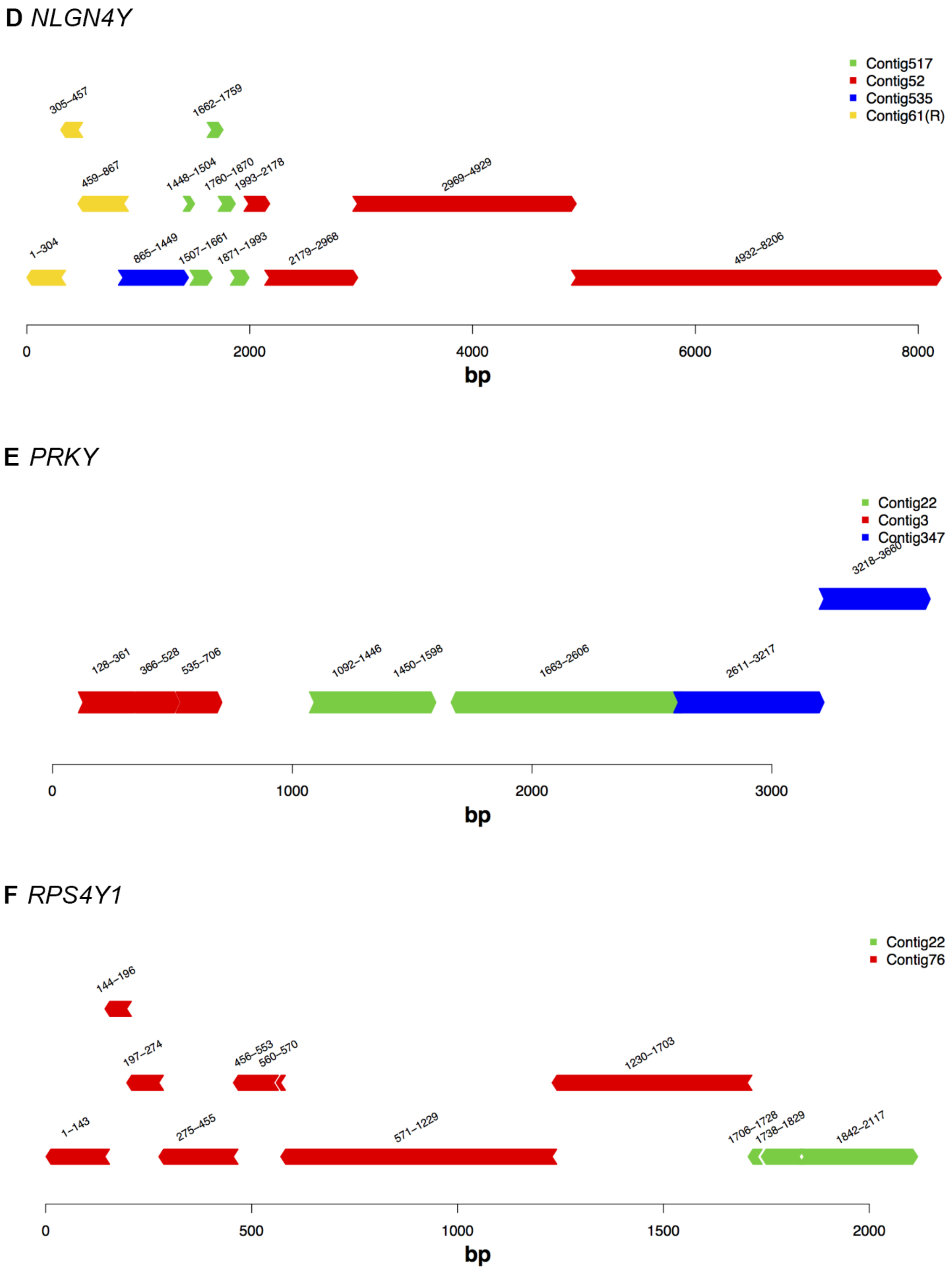

Frequency
0
50
100
150
200
250
300
0
5000
10000
15000
Fragment length

Figure S9. Transcript-guided assembly scaffolding. The alignment of gorilla transcripts to the assembly produced hits to different scaffolds for 11 single-copy protein-coding genes. The X-axis is the position in the transcript and the arrows correspond to alignments with >95% identity (the coordinates of the transcripts are on the top and the direction of the arrow reflects the orientation of the alignment in a scaffold). Different scaffolds are indicated by different colors. R means that this contig should be reverse complemented to conserve the direction of the transcript. The plots were generated after removing multiple hits inside the scaffold or in the other scaffold. (**A**) ***DDX3Y*** (2 scaffolds merged for a final scaffold length of 256,893 bp), (**B**) ***EIF1AY*** (2 scaffolds merged for a final scaffold length of 48,627), (**C**) ***KDM5D*** (3 scaffolds merged for a final scaffold length of 160,034 bp), (**D**) ***NLGN4Y*** (4 scaffolds merged for a final scaffold length of 513,187 bp), (**E**) ***PRKY*** (3 scaffolds merged for a final scaffold length of 571,209 bp), (**F**) ***RPS4Y1*** (2 scaffolds merged for a final scaffold length of 317,430 bp), (**G**) ***TBL1Y*** (4 scaffolds merged for a final scaffold length of 779,490 bp), (**H**) ***USP9Y*** (9 scaffolds merged for a final scaffold length of 612,541 bp), (**I**) ***UTY*** (11 scaffolds merged for a final scaffold length of 1,175,566 bp), (**J**) ***ZFY*** (4 scaffolds merged for a final scaffold length of 296,714 bp). Single copy ampliconic genes: (**K**) ***PRY*** (3 scaffolds merged for a final scaffold length of 477,886 bp). The plots for transcript recovery and contig scaffolding were generated with a modified version of the function plotGenes in the Sushi package (Phanstiel et al. 2014).

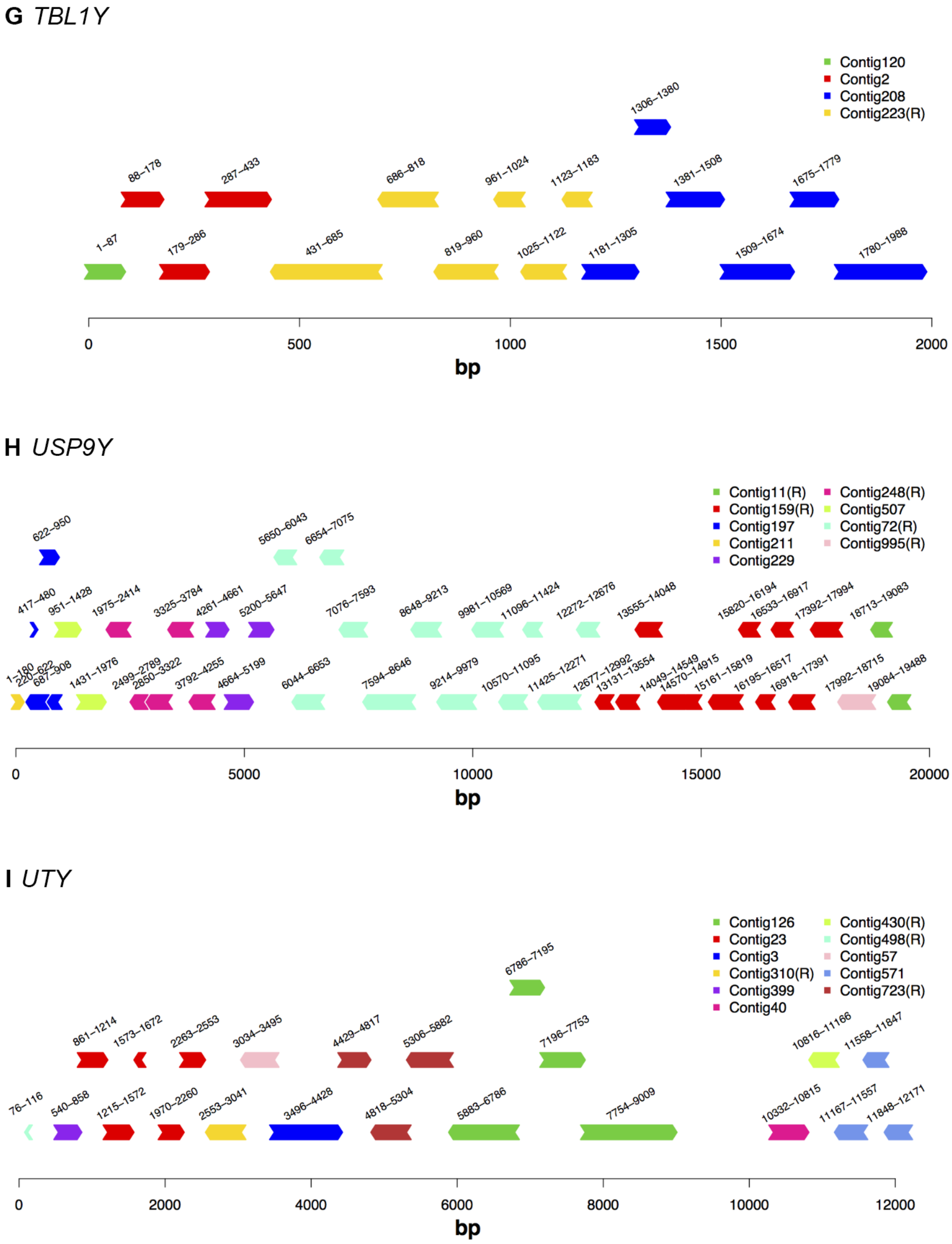

A DDX3Y
Contig11
Contig224
210–258
780–865
1271–1415
210–236
255–387
544–643
866–964
1126–1270
1598–1715
1870–2009
2268–2343
2789–2889
3230–3279
3720–3754
1–14
41–56
152–209
243–250
388–543
644–779
965–1125
1416–1597
1716–1869
2010–2267
2890–3229
3280–3705
3755–4211
0
1000
2000
3000
4000
bp
B EIF1AY
Contig367
Contig536
701–792
875–925
1030–1115
1–700
793–874
926–1029
1111–1318
0
200
400
600
800
1000
1200
1400
bp
C KDM5D
Contig108
Contig402(R)
Contig507
1365–1484
1989–2108
2486–2610
4272–4341
499–623
794–929
1060–1205
1485–1644
1826–1988
2304–2485
2759–2855
3215–3353
3534–3671
4342–4541
4907–5109
5304–5413
422–501
624–794
930–1059
1206–1364
1642–1825
2109–2303
2611–2758
2856–3214
3354–3533
3672–4271
4542–4906
5110–5303
0
1000
2000
3000
4000
5000
bp



D *NLGN4Y*

E *PRKY*

F *RPS4Y1*


G TBL1Y
🟩 Contig120
🟥 Contig2
🟦 Contig208
🟨 Contig223(R)
1306–1380
88–178
287–433
686–818
961–1024
1123–1183
1381–1508
1675–1779
1–87
179–286
431–685
819–960
1025–1122
1181–1305
1509–1674
1780–1988
0
500
1000
1500
2000
bp
H USP9Y
🟩 Contig11(R)
🟥 Contig159(R)
🟦 Contig197
🟨 Contig211
🟪 Contig229
🟪 Contig248(R)
🟨 Contig507
🟩 Contig72(R)
🟥 Contig995(R)
622–950
5650–6043
6654–7075
417–480
951–1428
1975–2414
3325–3784
4261–4661
5200–5647
7076–7593
8648–9213
9981–10569
11096–11424
12272–12676
13555–14048
15820–16194
16533–16917
17392–17994
18713–19083
1–180
220–622
687–908
1431–1976
2499–2789
2850–3322
3792–4255
4664–5199
6044–6653
7594–8646
9214–9979
10570–11095
11425–12271
12677–12992
13131–13554
14049–14549
14570–14915
15161–15819
16195–16517
16918–17391
17992–18715
19084–19488
0
5000
10000
15000
20000
bp
I UTY
🟩 Contig126
🟥 Contig23
🟦 Contig3
🟨 Contig310(R)
🟪 Contig399
🟪 Contig40
🟨 Contig430(R)
🟩 Contig498(R)
🟥 Contig57
🟦 Contig571
🟥 Contig723(R)
6786–7195
861–1214
1573–1672
2263–2553
3034–3495
4429–4817
5306–5882
7196–7753
10816–11166
11558–11847
76–116
540–858
1215–1572
1970–2260
2553–3041
3496–4428
4818–5304
5883–6786
7754–9009
10332–10815
11167–11557
11848–12171
0
2000
4000
6000
8000
10000
12000
bp

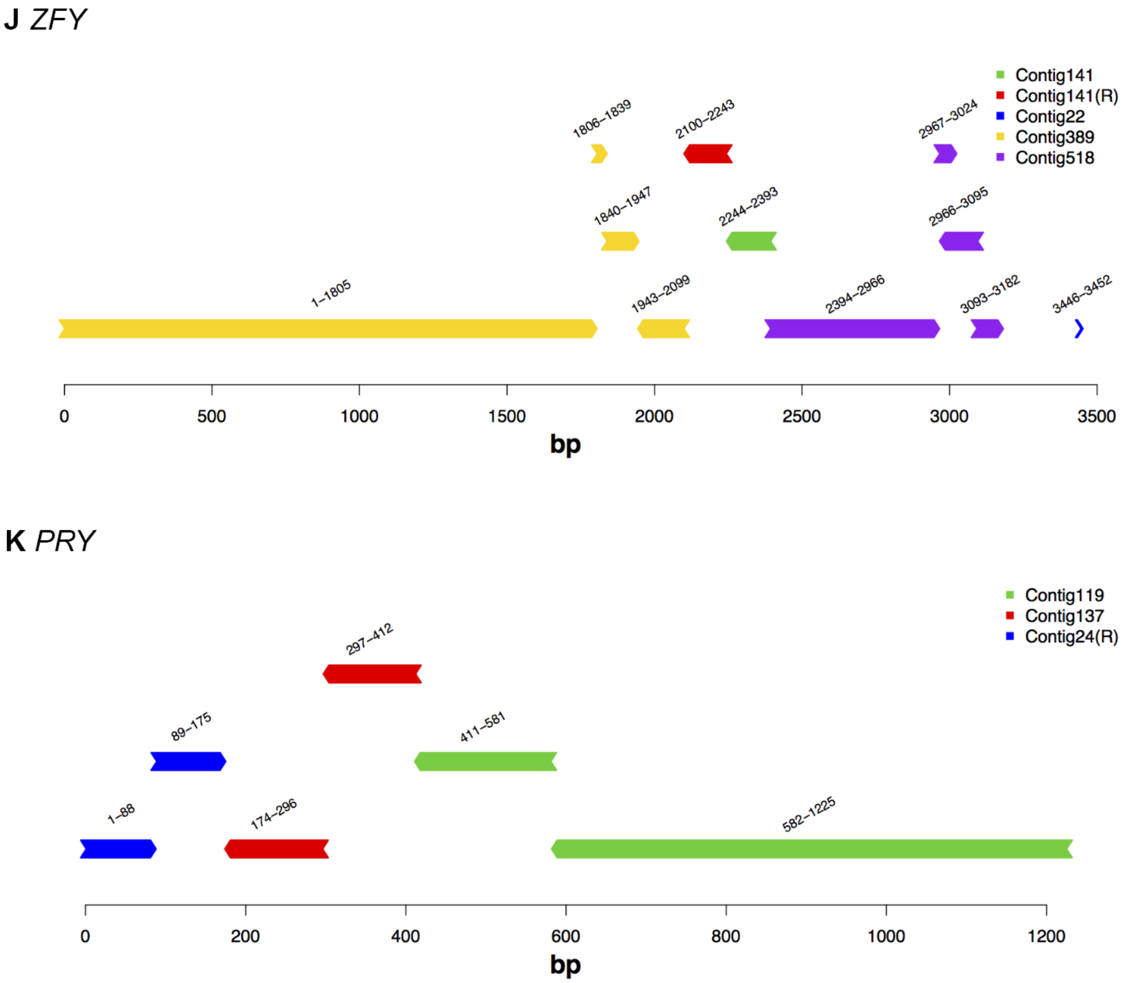
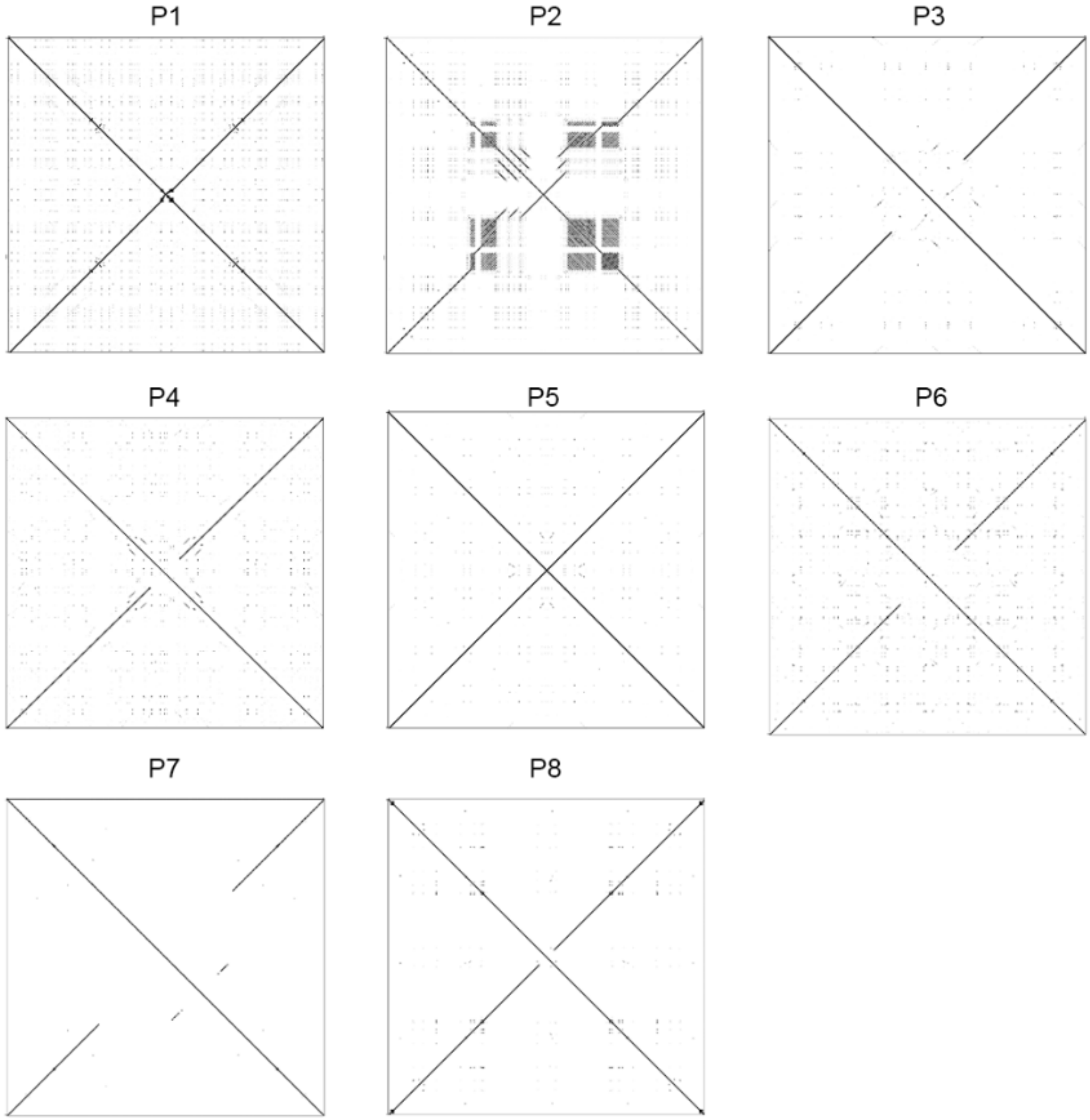

J ZFY
Contig141
Contig141(R)
Contig22
Contig389
Contig518
1806–1839
2100–2243
2967–3024
1840–1947
2244–2393
2966–3095
1–1805
1943–2099
2394–2966
3093–3182
3446–3452
0
500
1000
1500
2000
2500
3000
3500
bp
K PRY
Contig119
Contig137
Contig24(R)
297–412
89–175
411–581
1–88
174–296
582–1225
0
200
400
600
800
1000
1200
bp

Figure S10. Dotplots of human palindromes extracted from the reference human assembly (hg38). Word length for visualization was set to 40 for P1, P3 and P5, and 20 for P2, P4, P6, P7 and P8. Palindrome P2 exhibited highly repetitive structure near the arm-spacer boundary.



Figures S11A-D. A comparison of palindrome recovery between human, chimpanzee and gorilla and depth analysis. **(A, B) Presence of sequences homologous to human and chimpanzee palindromes.** We mapped extracted human and chimpanzee palindromes to human, chimpanzee and gorilla assemblies using BWA (Li and Durbin 2009) with seed length=5 to increase sensitivity. The heatmaps illustrate how eight human palindromes and 12 chimpanzee-specific palindromes were recovered in the assemblies. The heatmap only captures that at least one copy is present but does not evaluate possible fragmentation or copy number. **(C, D) The results of the read depth analysis.** From left to right: the number of reads per bp for all gorilla scaffolds (ALL), scaffolds carrying X-degenerate genes (DEG), scaffolds carrying ampliconic genes (AMP, see legend of Fig. S4), and scaffolds corresponding to human and chimpanzee palindromes (see Note S5).

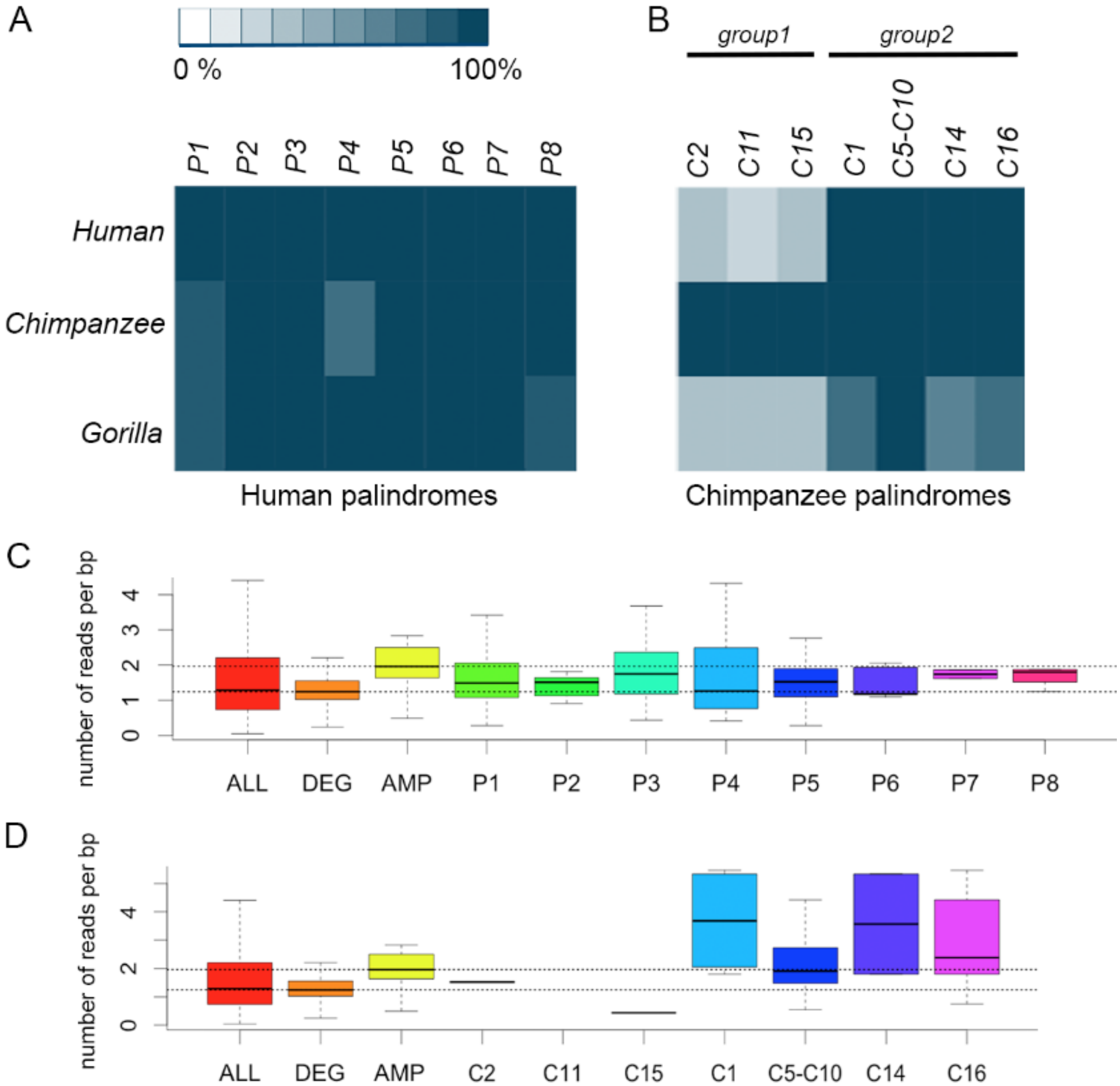

A
0 %
100%
P1
P2
P3
P4
P5
P6
P7
P8
Human
Chimpanzee
Gorilla
Human palindromes
B
group1
group2
C2
C11
C15
C1
C5-C10
C14
C16
Chimpanzee palindromes
C
number of reads per bp
0
1
2
3
4
ALL
DEG
AMP
P1
P2
P3
P4
P5
P6
P7
P8
D
number of reads per bp
0
2
4
ALL
DEG
AMP
C2
C11
C15
C1
C5-C10
C14
C16

Figure S12. Within-scaffold gorilla-specific palindromes. New gorilla-specific palindromes that are fully present within one scaffold. The length of the arms (left axis) and spacers (right axis) is visualized by the length of the horizontal bars (bp). Colors on the left axis (red, orange, green) represent the identity of the corresponding arms, gray marks the proportion of the sequences that is RepeatMasked for both arms and spacers. A palindrome marked with a star does not have homologous sequence on the chimpanzee Y chromosome and is present only partially on human Y chromosome. The dotplot for each palindrome is shown (computed using 10-bp windows).

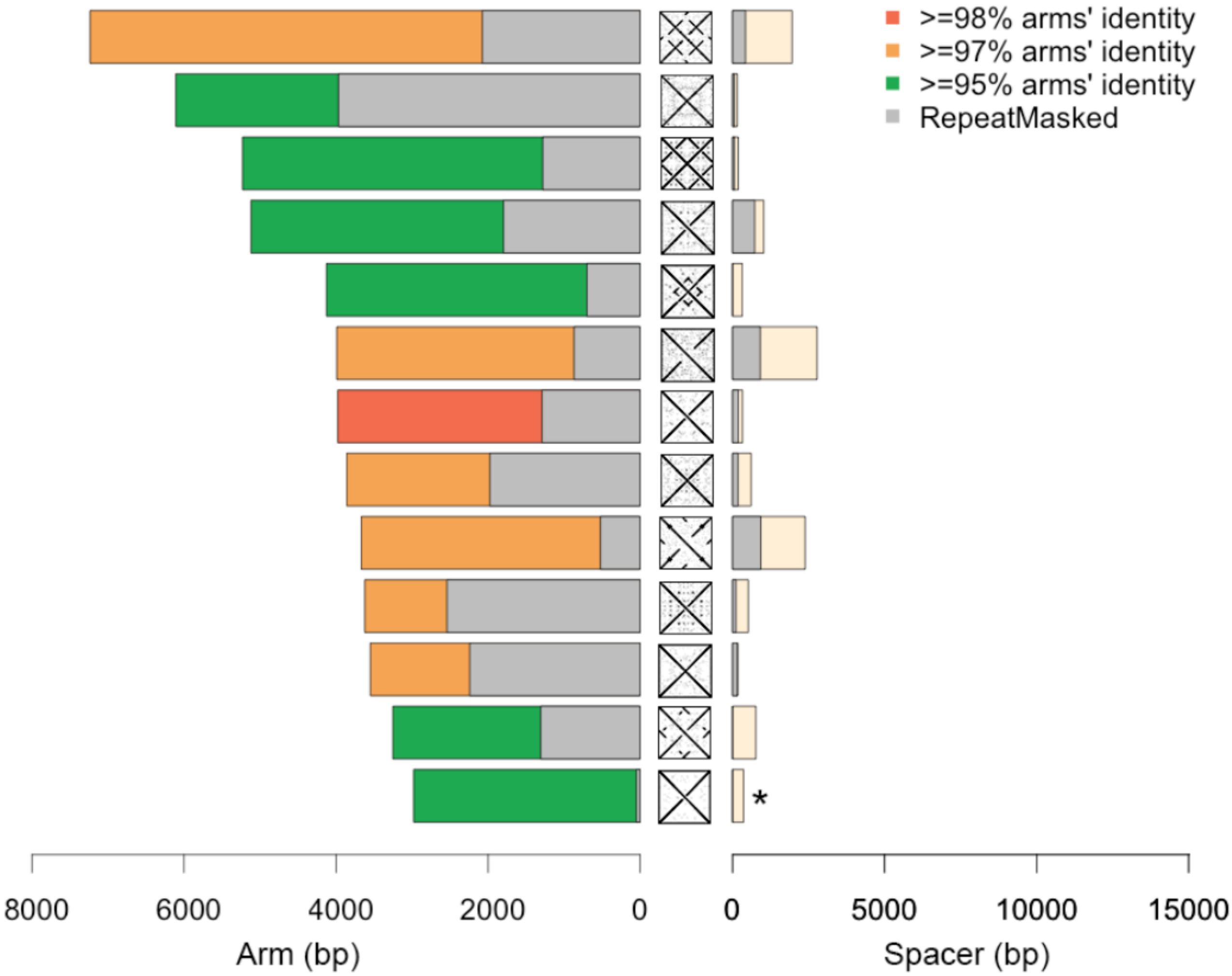
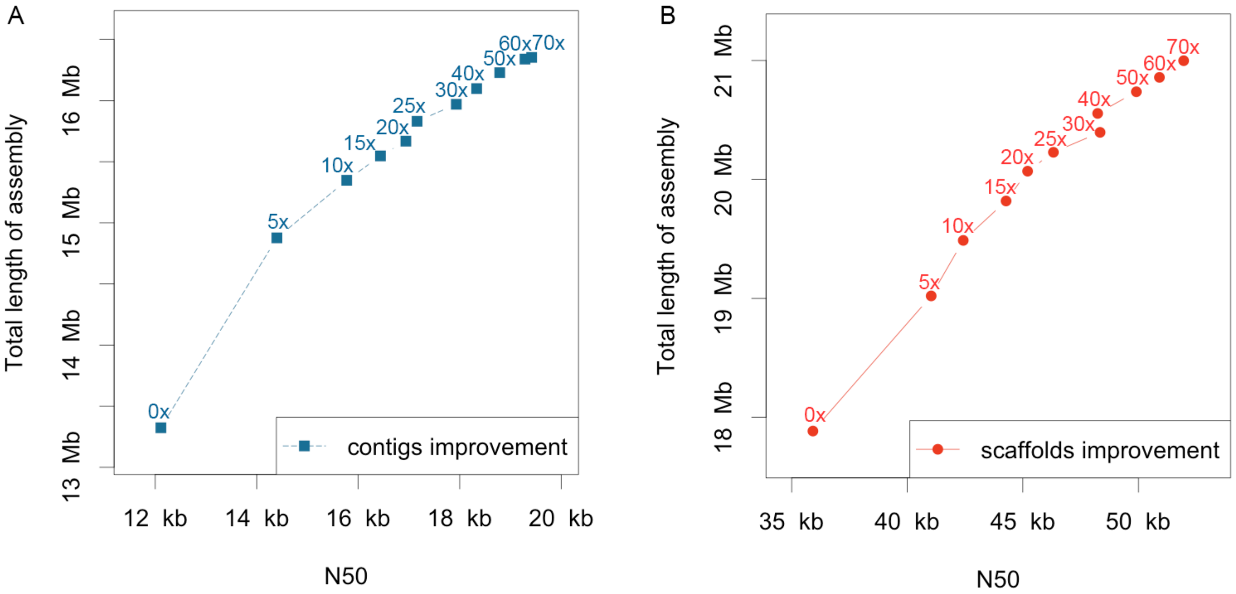


Figure S13. Progressive improvements to the Illumina-only assembly with PacBio data. (A) The effect of increasing coverages on the assembly quality. All PacBio reads were prefiltered against female (autosomes were filtered out) and the expected coverage was calculated assuming the size of gorilla Y of 60 Mb. The initial assembly from SPAdes (0x coverage, no PacBio data) was used to improve contigs and scaffolded assembly after SSPACE (0x coverage, no PacBio data) was used to improve scaffolds. The numbers above each point indicate the coverage used for the same initial assembly. We observe a significant jump in N50 around 5x-10x coverage, suggesting that even very few SMRT cells already provide cost-effective improvement of Illumina assemblies. (**B) Dependence of read lengths on assembly quality.** Reads from eight SMRT cells were sorted according to increasing read length and then partitioned into five bins such that each bin had an equal number of basepairs. The first bin (leftmost X-axis) contained a large number of short reads, whereas the last bin (rightmost X-axis) contained a small count of very long reads. PBJelly was then run five separate times, using a different bin each time, using the same Illumina-only assembly as the starting point. Although a higher N50 and assembly length for the bin with longest reads were expected, this experiment demonstrates that the increase in assembly length for the long read bin vs. the short read bin is as much as 6.8 Mb of sequence (a 36% increase).



C

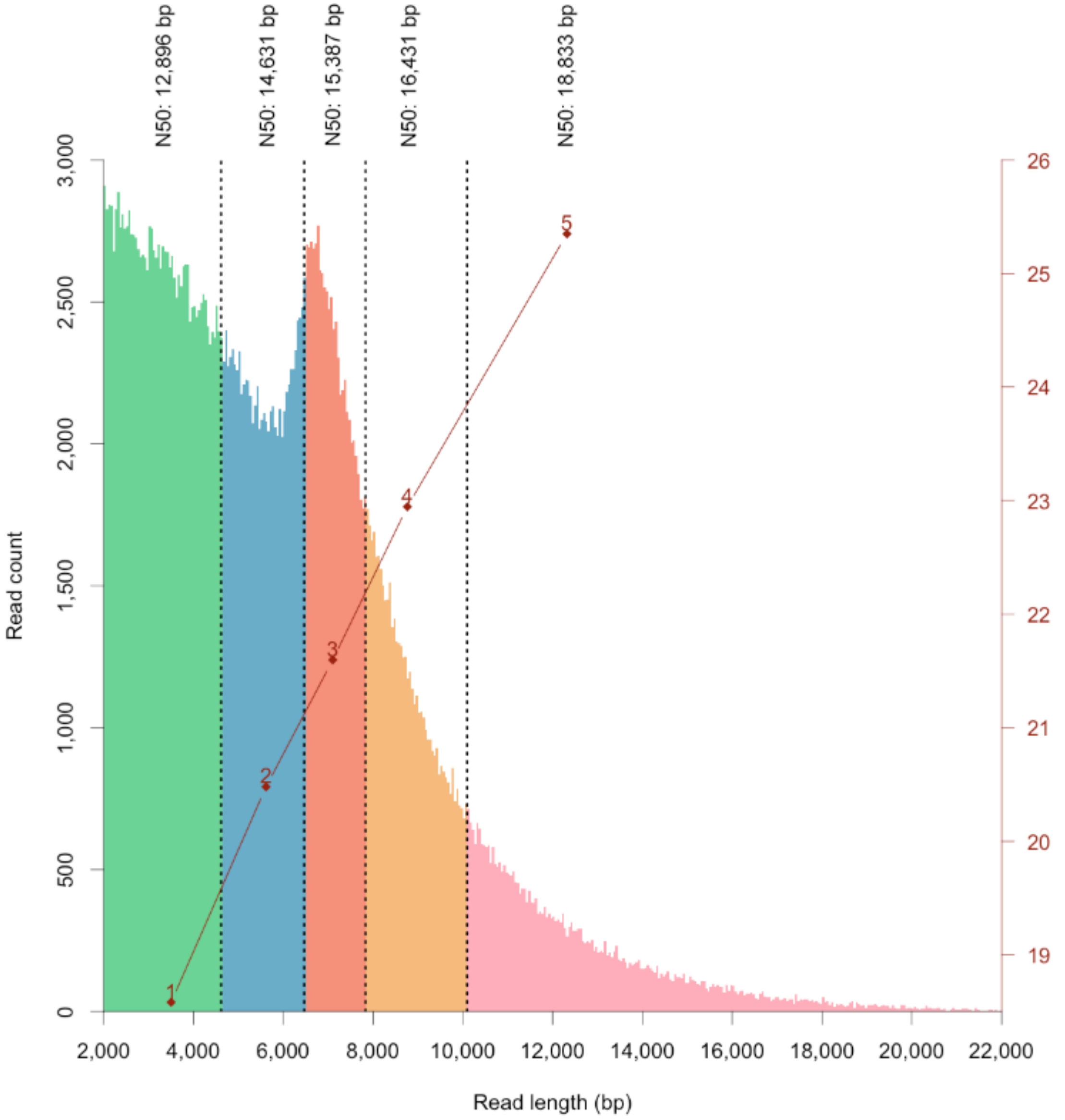
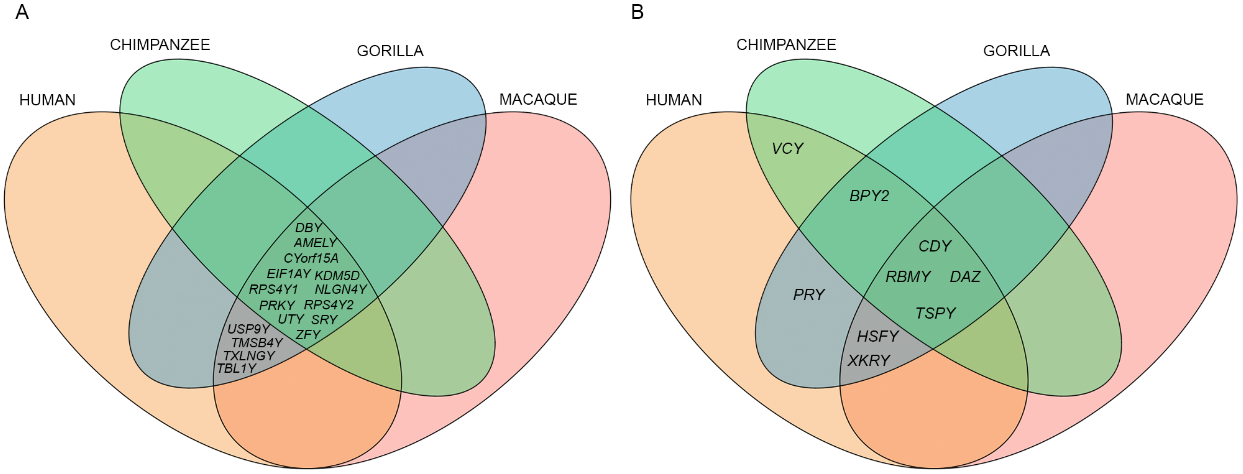

N50: 12,896 bp
N50: 14,631 bp
N50: 15,387 bp
N50: 16,431 bp
N50: 18,833 bp
Read count
3,000
2,500
2,000
1,500
1,000
500
0
26
25
24
23
22
21
20
19
1
2
3
4
5
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
20,000
22,000
Read length (bp)

Figure S14. Gene content comparison among sequenced primate Y chromosomes. (A) X-degenerate genes. (B) Ampliconic genes. Four species were included: Human (Orange), Chimpanzee (Green), Gorilla (Blue) and Macaque (Pink).



REFERENCES

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**: 488–493.

Goetting-Minesky MP, Makova KD. 2006. Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. *J Mol Evol* **63**: 537–544.

Goto H, Peng L, Makova KD. 2009. Evolution of X-Degenerate Y Chromosome Genes in Greater Apes: Conservation of Gene Content in Human and Gorilla, But Not Chimpanzee. *J Mol Evol* **68**: 134–144.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Makova KD, Li WH. 2002. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**: 624–626.

Phanstiel DH, Boyle AP, Araya CL, Snyder MP. 2014. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**: 2808–2810.

Villesen P, Fredsted T. 2006. Fast and non-invasive PCR sexing of primates: apes, Old World monkeys, New World monkeys and Strepsirrhines. *BMC Ecol* **6**: 8.