www.sciencemag.org/content/352/6281/aae0344/suppl/DC1

Supplementary Material for

Long-read sequence assembly of the gorilla genome

David Gordon, John Huddleston, Mark J.P. Chaisson, Christopher M. Hill, Zev N. Kronenberg, Katherine M. Munson, Maika Malig, Archana Raja, Ian Fiddes, LaDeana W. Hillier, Christopher Dunn, Carl Baker, Joel Armstrong, Mark Diekhans, Benedict Paten, Jay Shendure, Richard K. Wilson, David Haussler, Chen-Shan Chin, Evan E. Eichler*

*Corresponding author. E-mail: eee@gs.washington.edu

Published 1 April 2016, *Science* **352**, aae0344 (2016)
DOI: 10.1126/science.aae0344

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/352/6281/aae0344/suppl/DC1)

Tables S2, S3, S7 to S9, S14, S16 to S20, S22 to S25, S28 to S32, S34, and S35 are provided as a separate Excel file.
Table S36 is provided as a separate Excel file.

Supplementary Text

1. Genome sequencing

1.1 DNA sample preparation and sequencing

Sequencing data was derived from a single female specimen of *Gorilla gorilla gorilla* (Susie, now residing at the Columbus Zoo and Aquarium), which served as the basis for the gorilla genome assembly (table S1). Peripheral blood was drawn from Susie during routine medical care at the Lincoln Park Zoo, Chicago (prior to her move to Columbus Zoo and Aquarium). DNA was isolated from whole blood samples using the Gentra Puregene Cell Kit (P/N: 158767). White blood cells were isolated from the buffy coat, lysed, protein precipitated out, and DNA prepared. Eluted DNA was stored at 4°C overnight for two days to resuspend the DNA pellet, with quality control performed by fluorimetry (Qubit, Life Technologies) and run on a gel to visualize genomic DNA fragmentation. Genomic libraries were prepared for DNA sequencing.

Pacific Biosciences (PacBio) Sequencing: We prepared five DNA fragment libraries (20-30 kbp inserts) using Megaruptor (Diagenode) shearing at the 35 kbp setting. Post SMRTbell preparation per the "Procedure and Checklist – 20 kbp Template Preparation Using BluePippin^{TM} Size-Selection System" (PacBio), libraries were size-selected with the BluePippin^{TM} system (Sage Science) at a minimum fragment length cutoff of 15 kbp. Single-molecule, real-time (SMRT) sequence data were generated using the PacBio RSII instrument with P6v2 polymerase binding and C4 chemistry kits (P6-C4) and run times of 6-hour movies. A total of 236 SMRT cells were processed yielding 83.7-fold (ROI/3.0 G) (74.8-fold aligned/3.0G) (100.4-fold raw/3.0G) whole-genome sequence (WGS) data. The average SMRT subread length was 12.9 kbp (10.2 kbp aligned) with a median subread length of 11.5 kbp (fig. S1).

Illumina sequencing: Libraries were generated with PCR-based and PCR-free protocols. PCR-based: gDNA was sheared using Covaris LE220 (duty cycle 10%, intensity 3, cycles/burst 200, time 100s) to an average size of ~700 bp. Sheared sample was taken directly to end repair (NEB Next End-Repair, New England BioLabs). After a column clean up (QIAquick PCR Purification Kit, Qiagen), fragments were A-tailed with Klenow (3'-5' exo-) (Roche) in the presence of dATP followed by an additional column purification. Y adaptor was added using T4 ligase (Enzymatics) followed by another column purification. The library was size-selected to 650-750 bp on an E-Gel EX 1% Agarose Gelgel (Invitrogen) followed by gel purification, then barcodes added with five cycles of PCR (Biorad IProof reagents) followed by a 1X AMPure XP bead wash to remove adapter dimer and small fragment contamination. The final library was quantified with Qubit (Life Technologies) and loaded on NextSeq (PE151). PCR-Free: gDNA was

sheared using Covaris LE220 with cycling conditions of 15% duty cycle, Peak Power 450W, Cycles/Burst 200, and Time 46s. The sheared DNA was processed using the Illumina TruSeq DNA PCR-Free LT Library Kit protocol to generate 550 bp inserts, which includes end repair, SPRI bead size selection, A-tailing, and Y-adapter ligation. Library concentration was measured by qPCR and loaded on MiSeq (PE151), NextSeq (PE151) and HiSeq (PE101) instruments to generate ~24-fold sequence coverage.

2. Genome assembly

2.1 *De novo* assembly with Falcon

SMRT sequence data from Susie were assembled using DALIGNER (*23*) and Falcon (https://github.com/PacificBiosciences/FALCON) for read overlap and string graph layout, followed by Quiver (*8*) to generate the consensus sequence. The Falcon assembly method operates in two phases: first, all reads are aligned against all reads to generate 97-99% accurate consensus sequences of overlapping reads; next, overlaps between the corrected longer reads are used to generate a string graph. Contigs are formed using the sequences of non-branching paths. Topology and coverage may be used for detection when branches result from allelic structural variation as opposed to collapsed duplication, in which case the sequence of one of the alleles may be incorporated into the assembly. Such regions of the graph, termed compound path regions, are recorded and may be reviewed later during analysis of structural variation. Sequencing errors and chimeric reads that are incorporated into a string graph will artificially fragment contigs when using a direct implementation of the string graph.

Two supplemental graph cleanup operations are defined to improve assembly quality by removing spurious edges from the string graph: tip removal (*25*) and chimeric duplication edge removal. Tip removal discards sequences with errors that prevent 5' or 3' overlaps. High-copy, long repeats in a genome can tangle the assembly graph significantly. We use 5'- and 3'-overlap counts to estimate the copy number and use it to filter out high-copy repeats to minimize graph tanglements. Chimeric reads or duplication regions in different parts of the genome may cause spurious edges in the assembly graph. Such spurious edges are detected using local graph topology (unitig repeat bridge) and removed. In a recent CHM1 assembly using 60X SMRT P6-C4 sequencing chemistry, for example, this step addressed 1,240 regions of the string graph that corresponded to 623 non-overlapping sequences in the genome. Of the 623 non-overlapping sequences, 479 overlap segmental duplications (average identity of 96.68%), with an 11-fold increase in segmental duplication bases over a random sample of similar regions.

We have modified the Falcon (v0.3.0) pipeline to increase performance, to ease detecting and restarting failed jobs, and to determine progress during assembly. One stage of the assembly process “LA4Falcon” is particularly time-consuming and IO intensive. We significantly increased the speed of the assembly process by copying needed information to a large number of faster local disks, doing the processing there, and then copying the results back to the large network disk, rather than doing the processing directly on the slower large network disk. We also modified the source code to use more efficient IO. In a controlled test, running LA4Falcon from Falcon v.0.2.2 with a sample dataset took 187

hours with unmodified code but 3.5 hours with our first set of changes—a factor of 53. Our changes were all designed to improve speed and did not affect the algorithms or output. These enhancements to Falcon are publicly available at http://eichlerlab.gs.washington.edu/publications/Gordon2016/FALCON-integrate.tar.gz.

2.2 Initial assembly statistics

We utilized Falcon (v.0.3.0) to generate an initial assembly "Susie3" of 3.1 Gbp with a contig N50 (half the assembly is in contigs greater than) of 10.02 Mbp. Susie3 assembled into 15,997 contigs, including 889 contigs >100 kbp (fig. S2), after error correction using Quiver (*8*). After resolving mis-assemblies (SM 2.4), Susie3 contained 16,073 contigs with a contig N50 of 9.6 Mbp. Based on alignment to human (GRCh38), we estimate that 98.87% of the euchromatic portion was assembled into 1,854 sequence contigs (fig. S3). Compared to the published gorilla sequence gorGor3 (*4*), Susie3 represents a decrease in assembly fragmentation: 433,861 to 16,073 contigs, a >96% reduction in total contig number (tables S2, S3; see separate Excel file).

2.3 Accuracy assessment and error correction

2.3.1 Initial accuracy of Susie3

We assessed the accuracy of the Susie3 assembly with three metrics: a) alignment identity of finished inserts from the BAC library of the gorilla Kamilah (CHORI-277) (table S6), b) the number of homozygous SNVs and indels called per base from whole-genome Illumina data for the same source gorilla Susie, and c) the proportion of protein-coding transcripts from GENCODE (v23 Basic) (*12*) that contain frame-altering variants when aligned to the gorilla assembly. Based on the results of our accuracy assessment, we applied an error correction protocol to the Susie3 assembly using Illumina data from a population of Western lowland gorillas.

We calculated alignment identity of BAC inserts against Susie3 using 19 previously sequenced CHORI-277 clones from GenBank that were annotated as "finished" and aligned to the gorilla reference with at least 120 kbp of sequence. Note: clones were derived from a large-insert library prepared from the Western lowland gorilla Kamilah—source of the published genome assembly. If more than one clone mapped to the same location in Susie3, we selected the clone with the smallest start coordinate to represent that locus. These clones were aligned to Susie3 with BLASR (*30*) to calculate an alignment identity of 99.66% (3,409 mismatches out of 995,249 aligned bases) (table S7; see separate Excel file). We corrected for allelic diversity between Kamilah and Susie by subtracting the number of variants expected between two Western lowland gorillas from the total mismatches based on the previously observed heterozygosity within the subspecies (1.6×10^{-3} - 2.4×10^{-3} variants per base (*10*)). Out of the total 995,249 aligned bases, we estimated that 1,592-2,389 variants between BACs and Susie3 were allelic. Assuming the remaining 1,020-1,817 mismatches are errors in the assembly, the overall accuracy of Susie3 is 99.82-99.9% for a corresponding accuracy or quality value (QV) score of 27-30 (table S7; see separate Excel file). Note: QV represents a per-base estimate of accuracy and is calculated as QV = -10log10(Pe) where Pe is the probability of error.

Second, we calculated the number of homozygous SNVs and indels per base from whole-genome Illumina data for Susie. We aligned ~14-fold coverage of 151 bp paired-end reads from a standard Illumina PCR-based library to Susie3 with BWA-MEM (*32*) (v0.7.3) and called high-quality variants (QUAL > 5) with FreeBayes (*33*) (v0.9.21) in contigs containing less than 75% satellite content and excluding regions of excess or depleted read depth. We then identified putative assembly errors by selecting variants with homozygous alternate genotypes in Susie. Using this approach, we identified 522,509 SNPs, 1,281,150 insertions, and 329,020 deletions for a total of 2,132,679 variants out of 2,823,479,459 bp assessed by FreeBayes. These variants correspond to an assembly accuracy of ~QV 31.

Finally, we measured the accuracy of gene models in Susie3 based on concordance with human gene models from GENCODE (*12*). To annotate gene models, we identified syntenic regions between gorilla and human by aligning all contigs from the gorilla assembly to the human reference assembly (GRCh38) with progressiveCactus (*34*). We used these syntenic alignments and TransMap to map human GENCODE transcripts to their corresponding sequence in the gorilla assembly (*14*, *15*). We identified potential assembly errors by flagging coding sequence modifications such as frameshifts, early stop codons, and coding insertions or deletions that disrupt the original gene models. Of 58,688 protein-coding transcripts assessed, 29,409 (50.1%) contained at least one such disruption.

2.3.2 Error correction

Insertions and deletions represented the majority of mismatches between Susie3 and BACs from Kamilah, Illumina reads from Susie, and transcripts from human (fig. S4). We attempted to resolve these systematic errors in the initial Falcon/Quiver assembly using population-based error correction with whole-genome Illumina data from six previously published Western lowland gorillas (*10*) and Susie (table S8; see separate Excel file). We aligned ~581 billion bases (~194-fold coverage) of paired-end Illumina reads from all seven genomes to Susie3 with BWA-MEM (v0.7.3) (*32*) and called SNVs and indels with FreeBayes (v0.9.21) (*33*). We called an initial set of 20,858,775 SNVs and indels in the 5,255 contigs with <75% satellite repeat content corresponding to 2.88 Gbp (93.6%) of the assembly (table S9; see separate Excel file). We excluded variants that occurred in 5,568 loci (~118 Mbp) that had been previously flagged for depleted or excess read depth to produce a set of 20,254,475 high-quality SNPs and indels. We found an excess of fixed insertions and deletions consistent with accuracy errors in the Susie3 assembly relative to gorGor3 (fig. S4).

We classified as assembly errors any SNPs or indels that had a homozygous alternate variant call in a) Susie and four or more other gorillas or b) all gorillas except for Susie. While there was no significant inflation of SNPs in Susie3 relative to gorGor3, we chose to apply the same majority rule to these variants to make Susie3 representative of the Western lowland gorilla subspecies—i.e., the "pan"-gorilla genome. Using these filters, we identified 2,310,692 variants to correct in Susie3, including 1,387,549 insertions, 333,886 deletions, 589,093 SNPs, and 164 complex variants. Of these ~2 million variants, 1,530,856 (66%) were homozygous for the alternate allele in all seven gorillas while an additional 468,868 (20%) were found in Susie and at least four other gorillas.

We corrected the Susie3 assembly sequence by replacing the reference sequence for all of these variants with their corresponding alternate alleles to produce a final reference assembly for release, Susie3.2. Both Susie3 and Susie3.2 will be released for comparison.

2.3.3 Accuracy assessment of Susie3.2

After error correction, we repeated our accuracy assessments with BAC alignments, homozygous alternate variants from Susie, and GENCODE transcript alignments. We calculated an alignment identity of 99.74% (8,326 mismatches out of 3,249,430 aligned bases) for the same 19 Kamilah BACs mapped to Susie3.2. After correcting for allelic diversity, we estimated the accuracy of Susie3.2 at 99.9-99.98% (QV 30-38). Similarly, we identified 224,867 homozygous variants between Illumina data from Susie and Susie3.2 out of 2,823,479,459 bp assessed corresponding to QV 41. Using BLAT to align GENCODE transcripts to the gorilla assembly, we found a decrease in alignment errors after error correction (fig. S5). Examples of restored open-reading frames are shown below (figs. S6, S7).

Overall, 1,600 of 3,464 contigs in Susie3 with at least one variant called (46%) are within the range of expected variation for Western lowland gorillas of up to 2.4 x 10^(-3) variants per base even prior to error correction (fig. S8). However, 312 contigs (9%) have an excess of variants per base (more than the Susie3 mean + 1 SD) that potentially indicates locus-specific assembly error or biologically-relevant variation (fig. S9). Of these 312 contigs, 112 (36%) have no alignments to GRCh38 and of the remaining 200 contigs, 135 (68%) map within human- or gorilla-specific segmental duplications or heterochromatic regions, including telomeres and centromeres (fig. S9). The remaining 1,552 contigs represent a second mode in the distribution of variants per base with more variation than expected based on previous studies (*10*) but close to the mean variation per contig for Susie3.

2.4 Misassembly analyses

2.4.1 Identifying regions of high variance in coverage

Regions in an assembly containing unusually high or low sequence coverage may be signatures of expanded or compressed repeats (*35*, *36*). We flagged these high/low-coverage regions in Susie3. Read depth is first smoothed using a 2 kbp sliding window. The upper and lower coverage cutoffs are determined by first sampling (without replacement) a million of the smoothed coverages. The first quartile, third quartile, and interquartile ranges are recorded. A window is marked as high coverage if the coverage is greater than the third quartile coverage plus two times the interquartile range. A window is marked as low coverage if the coverage is less than the first quartile coverage minus two times the interquartile range. To help reduce the number of falsely marked regions due to stochastic variation in coverage, a region of a contig is only flagged if it contains at least 10 kbp of consecutively marked windows. For Susie3, we applied thresholds at 119.15-fold and 21.9-fold average sequence read coverage per 2 kbp window for high and low coverage, respectively. We flagged 2,067 high-coverage (51.9 Mbp) and 3,501 low-coverage (65.6 Mbp) regions for a total of 117.5 Mbp (3.8% of Susie3).

2.4.2 Flagging potential misassemblies

Individual Susie3 contigs were further examined for potential misassemblies based on these stretches of sequence that contain large deviations in sequence coverage. If these repeats are located in different areas of the genome, then the assembler may erroneously bridge these distant areas together forming a chimeric contig. We resolved such chimeric contigs with the addition of long-range sequencing information, such as BAC or fosmid ends, that span the predicted breakpoint. If a BAC or fosmid does not span a potential breakpoint then it is deemed a site of potential misassembly.

We consider a contig misassembled if it satisfies two conditions: a) the contig contains a stretch of sequence greater than 10 kbp that has a high variance in sequence coverage and b) the stretch of sequence lacks concordant paired-end BAC and fosmid sequence support for at least one kilobase within the region. The end of contigs, by definition, will have reduced paired-end sequence support; thus, we exclude from this analysis sequence those occurring within 40 kbp from the ends of the contigs. Overall, 33 regions (1.39 Mbp) across 32 contigs larger than 100 kbp were marked as misassembled. Contigs were split at these breakpoints resulting in the 16,063 total contigs and a 9.6 Mbp N50.

The predicted misassembled contigs were compared against those identified by BLASR's whole-genome alignment. Alignments to GRCh38 were filtered by length and alignment similarity (5 kbp and 50%, respectively). We first compared our predicted misassemblies against the translocations identified by BLASR. Nine contigs containing translocations remained, three of which were correctly predicted as misassembled. Contig 000249F_quiver was marked as a translocation but contained alignments to GRCh38 chromosome 1 and chromosome 1 KI270711. Contig 000212F_quiver contained a 30 kbp alignment to GRCh38 chromosome 16 (87% similarity) within an approximately 10 Mbp alignment to chromosome 18. However, this 30 kbp stretch was completely contained within a segmental duplication and most likely an alignment artifact. Similarly, contig 000244F_quiver primarily aligns to chromosome 4 but contains a 38 kbp alignment to chromosome 1 (71.6% similarity) that lies within a segmental duplication. Contig 000884F_quiver contains an alignment to chromosomes X and Y with low similarity (81% and 73% similarity). These alignments are bridged by a segmental duplication.

The two remaining contigs 000002F_quiver and 000324F_quiver contained more prominent misassemblies. 000002F_quiver had a 13.7 Mbp alignment to GRCh38 chromosome 18 and a 19.1 Mbp alignment to chromosome 12. There was a sharp drop in sequence coverage bridging these two alignments; however, the span of this low-coverage region was roughly 7 kbp, below our 10 kbp cutoff for flagged regions. 000324_quiver contained alignments to GRCh38 chromosomes 10 and 6 of lengths 4.6 Mbp and 2.1 Mbp, respectively. Although there was no large change in sequencing depth across these alignments, there was a drop to 0 BAC and fosmid coverage. To improve the quality of Susie3, the contigs were further broken at the above regions resulting in 16,073 total contigs and a 9.6 Mbp N50.

2.5 Scaffold construction

Susie3 contigs greater than 100 kbp (889) were scaffolded by SSPACE (*37*) and Consed (*38*) using CH277 BAC and fosmid end pairs to bridge contigs. We excluded from this

analysis end sequence pairs that mapped to regions of high coverage (see above). To improve the accuracy of our scaffolds, any links to contigs marked as potentially misassembled were removed. After filtering, 101,431 and 2,470 intra- and inter-contig fosmid links and 104,150 and 2,956 intra- and inter-contig BAC links remained. Contigs marked as potentially misjoined were not used during scaffolding. SSPACE iteratively joined and oriented contigs that shared a minimum of two links (-k 2; -a 0.7).

552 scaffolds were produced by SSPACE with a scaffold N50 of 23.1 Mbp (a 2.4-fold increase). 422 (76.4%) scaffolds are comprised of a single contig. The longest scaffold increased from 36.2 Mbp to 110 Mbp. We evaluated the correctness of the scaffolds by aligning them to GRCh38. 2.15 Gbp of our multiple contig scaffolds aligned to the same chromosome, 1.76 Gbp of which were in a similar order and orientation as GRCh38. 405 Mbp lies on the same GRCh38 chromosome but contains a different ordering and/or orientation (fig. S10). 14.7 Mbp of our scaffolds contains alignments to multiple chromosomes. Upon investigation, these multi-chromosomal alignments lie within peri- and telocentric regions of the genome. We improve the quality of our Susie3 scaffolds by breaking the scaffolds at these locations. For completeness, we project our scaffolds onto the gorilla chromosomes (fig. S11).

2.6 A Golden Path (AGP) construction

Scaffolds were further oriented and ordered using GRCh38 to provide chromosomal resolution. The chromosomal AGP was created by first aligning the scaffolds to GRCh38 using NUCmer (*39*) (-mumref -l 60 -c 100) and then ordered using mummerplot (--layout). Scaffolds not aligned by NUCmer were aligned using BLASR (*30*). Breaks in AGP were introduced based on previously published large chromosomal rearrangements between human and gorilla (*4*, *40*, *16*, *41*) as follows:

1. HSA2 was a fusion of two chromosomes; thus, it was split into two chromosomes in Susie3:
    a. GGO2a: HSA2[0 - 111.9 Mbp] with an inversion corresponding to HSA2[94 Mbp - 111.9 Mbp]. GGO2a is then inverted so the short arm is first.
    b. GGO2b: HSA2[111.9 Mbp - 242 Mbp]
2. GGO4: Inversion of HSA4[49.3 Mbp - 70.0 Mbp]
3. GGO-specific HSA5/17 reciprocal translocation:
    a. GGO5: Inversion of HSA17[15.4 Mbp - 78.5 Mbp] followed by HSA5[79.9 Mbp - 180.7 Mbp]
    b. GGO17: HSA17[0 - 15.4 Mbp] followed by an inversion of HSA5[0 - 80 Mbp]
4. GGO7: Inversion of HSA7[76.5 Mbp - 102.0 Mbp]
5. GGO8: Inversion of HSA8[30.0 Mbp - 86.9 Mbp]
6. GGO9: Inversion of HSA9[0 - 70.0 Mbp]
7. GGO10: Inversion of HSA10[27.6 Mbp - 80.9 Mbp]
8. GGO12: Inversion of HSA12[21.2 Mbp - 63.6 Mbp]
9. GGO14: Inversion of HSA14[0 - 44.8 Mbp]
10. GGO18: Inversion of HSA18[0 - 14.9 Mbp]

The completed assembly consisted of 2.8 Gbp ordered and oriented across 24 chromosomes with 289.8 Mbp of unplaced sequence. The final AGP was plotted against GRCh38 using NUCmer (fig. S12).

3. Additional quality control

3.1 BAC/fosmid-end sequence validation

We assessed the quality of the underlying assembly by mapping BAC-end sequences (BES) (CH277) and fosmid-end sequences (FES) (CH1277) generated from the gorilla Kamilah against the PacBio assembly (Susie3) (tables S2, S10). Our analysis showed that 98.6% of the genome was supported by concordant best-paired BES and FES data. An analysis of aligned high-quality Sanger data (54.5 Mbp of PHRED > 35) from CH277 BES revealed high sequence identity 99.71% in Susie3 (99.88% in Susie3.2) (table S10). This is within allelic variance for the two gorilla genomes.

3.2 FISH validation

185 human BACs were previously mapped by FISH against gorilla chromosomes (*16*); 177 of these confirmed our associations between Susie3 contigs and human chromosomes and were in the same order in human as with FISH; four were in different order but same chromosome (one of these can be explained by an annotated duplication); and four were to different chromosomes (all of which can be explained by problems either with the BAC or the alignment).

4. Gap analysis

4.1 Gap closures

When we aligned Susie3 to the current gorilla genome reference (gorGor3) (*4*), we found that 94% of the 433,861 gorGor3 gaps were closed (figs. S13, S14), resulting in >164 Mbp of euchromatic sequence being added. Reanalysis of the gorGor3 gaps that were successfully closed revealed two types: gaps where additional intervening sequence could be placed (true gaps) and a minority where there was in fact no intervening sequence but the gaps had simply failed to be closed (false gaps). The former were enriched 3.8-fold for Alu repeats (table S11).

As the estimated gap sizes within gorGor3 increased, repeat content increased especially for segmental duplications (3- to 5-fold enrichment). For example, 10,959 gaps were in excess of 2 kbp; of these 21% (2,298) of the closed gaps mapped to segmental duplications, and 662 (6%) could not be closed (table S12). The average size of these open gaps was 8,629 bp with the largest gap annotated at 92,547 bp. Additionally, 629 (95%) of these open gaps overlapped segmental duplications in the human reference (GRCh37). The genomic context of these gaps represents the most difficult sequence to accurate assembly with any modern sequencing technology. 61.2% (548) of our largest contigs either begin or end within 10 kbp of an annotated segmental duplication or a region of high variation in sequence coverage. In addition to annotating human segmental duplications intersecting gap regions, gaps were annotated with previously identified

gorilla-specific and human segmental duplications. These gorilla-specific duplications were equally represented in open and closed gaps.

4.2 Sequence composition of Susie3 euchromatic gaps

Not all regions of the gorilla genome were closed within the SMRT genome assembly. Susie3 contigs were aligned using BLASR (https://github.com/mchaisso/blasr/tree/1.MC.rc44) to GRCh38 to determine the sequence composition of the remaining gaps in the assembly. After excluding heterochromatic regions (e.g., pericentromeric and subtelomeric), we considered 1,367 gaps comprising 40.8 Mbp of euchromatin. We examined the human and gorilla segmental duplication composition of these gaps using a sliding window of 10 Mbp (fig. S15). The estimated gap size (based on GRCh38) strongly correlated ($r^2 = 0.78$, $p < 2.2x10^{-16}$, Pearson correlation) with human and gorilla segmental duplication content.

5. Gene/exon analysis

5.1 Gene/exon content of closed gaps

We assessed the yield in gene content from gap closures in the gorilla reference by identifying human RefSeq exons that map within closed gap regions of Susie3. Of the 12,754 exons mapping within the gap regions, 11,105 (87%) were closed in the *de novo* assembly (table S12). As a result, we estimate based on human RefSeq gene models that the *de novo* assembly has completely resolved 3,473 of 3,696 of gorilla genes (94%), e.g., *CFH* (fig. S16). Another 52 genes are partially resolved but incomplete based on human annotation. Comparison of the percent of gorilla RNA-seq reads and transcripts aligned to the gorGor3 assembly versus the Susie3 assembly further confirms the improved representation in the Susie3 assembly (tables S13, S14; see separate Excel file for S14).

We annotated the Susie3 reference with gene models based on gorilla mRNA, human GENCODE models, and a Trinity (*42*) assembly that we created from the RNA-seq reads from gorilla iPS cells (GEO Accession: GSM1229060). We found an increased representation of full-length genes in Susie3 relative to the previous gorilla reference (gorGor3), including recovery of 3,370 gorilla iPS transcripts. While the full exon-intron structure of most genes was recovered, genes embedded within segmental duplications remained fragmented into multiple, often misassembled contigs, e.g., Notch2 (fig. S17).

For the RNA-seq alignments, the genomeGenerate function in STAR (*43*) was used to create a reference database for both gorGor3 and Susie3. Reads were then aligned using STAR with a maximum intron length of 50 kbp. Transcripts were aligned to the respective genomes using the gfClient/server version of BLAT (*44*).

5.2 Annotation and accuracy of gene models

We defined consensus gene models in the error-corrected gorilla assembly, Susie3.2, with Augustus(*13*) using TransMap alignments from the source annotation set GENCODE (v23 Basic) (59,638 protein-coding transcripts) as strong hints for the *de novo* discovery algorithm along with weak hints derived from published RNA-seq data from gorillas (*26*, *27*). We assessed the quality of both TransMap and Augustus transcripts with a series of binary classifiers representing assembly errors (resulting from annotated gaps in the

reference assembly or truncated transcript alignments) or alignment errors (resulting from alignment gaps, multiple alignments for the same transcript, or loss of intron information) (figs. S18-S22). Based on these binary quality classifiers, we attempted to pick the best transcript between TransMap and Augustus for the final consensus gene set. For each source TransMap transcript, if it did not fail any/all binary classifiers, we compared the alignment identity of the TransMap and Augustus transcripts to the original reference transcript and selected the transcript with the highest identity and coverage. If all transcripts for a given gene failed binary classification, we omitted potentially unreliable alternate isoforms and selected the best single transcript based on alignment identity to the reference transcript. Using this approach, we created 45,087 consensus models for protein-coding transcripts (19,633 genes) in Susie3.2 of which 43,127 of the 59,638 (72%) transcripts (15,879 genes) successfully passed binary classification and 15,801 (26%) transcripts (5,064 genes) failed quality classification resulting in a single best transcript for the consensus (fig. S19). We could not produce consensus models for an additional 719 transcripts (381 genes) that had no initial TransMap alignment. Of the 5,064 genes that failed quality classification, 437 (9%) map within segmental duplications in the human reference (GRCh38). Similarly, 62 of the 381 genes without TransMap alignments (16%) also map within duplications in the human reference.

The error-corrected gorilla assembly Susie3.2 significantly improves the contiguous representation of GENCODE genes compared to gorGor3. While Susie3.2 and gorGor3 have roughly the same number of alignment errors, the primary transcript errors in Susie3.2 are frameshifts, coding deletions (not in multiples of three), bad frames, and in-frame stops. In contrast, the primary errors in gorGor3 are incomplete transcript alignments, paralogous transcripts, differences in original intron boundaries, and bad frames. Due to the increased contiguity, Susie3.2 has significantly fewer assembly errors with 3,434 compared to 20,950 in gorGor3. Over 82% of protein-coding transcripts have >99.5% identity against Susie3.2 compared to ~52% in gorGor3 (fig. S18). These results are consistent with increased contiguity of sequence and more complete representation of segmental duplications in Susie3.2 compared to gorGor3.

Previous comparisons between the human and gorilla genomes were complicated by incomplete gene annotation. In order to investigate lineage-specific differences, we began by first comparing gene annotation between the human (GRCh38) and the two gorilla genome assemblies. Human GENCODE annotations were filtered to include only those with transcript support of 1, 2, or 3 (minimum one EST supporting a transcript). Augustus consensus annotations on Susie3 and Ensembl annotations on gorGor3 were filtered to eliminate any genes not present in the human annotations. All three assemblies share 15,490 genes (56.4% of human genes) (fig. S23). The human and Susie gorilla assemblies share an additional 10,854 genes that are absent from gorGor3. The Susie assembly thus recovers 96.0% of all human genes compared to the 56.8% recovered by gorGor3. Additionally, the human assembly has 992 genes that are not present in the gorilla assemblies. Of these human-specific genes 438 (44%) occur within segmental duplications and are therefore unlikely to be completely assembled in either gorilla assembly.

6. Repeat content

RepeatMasker (*28*) and Tandem Repeats Finder (*29*) were used to annotate repetitive regions within Susie3. 55% of Susie3 is annotated as repetitive. This analysis includes all 15,997 initially assembled contigs, of which only 724 were incorporated into AGP (fig. S24). 10,153 of the 10,235 (99.2%; 167.2 Mbp) of sequence contigs that could not be aligned to human GRCh38 consist of >75% satellite sequence. (Note: This is in contrast to the 2,958 (51.3%; 75.8 Mbp) sequence contigs that map to GRCh38 and show >75% satellite repeat composition.) The remaining 82 unaligned contigs with <75% satellite sequence comprise 412 kbp of sequence with a median length and coverage of 1,578 bp and 0-fold, respectively. These unaligned contigs lack any BES or FES support and are likely artifacts of the string graph assembler. The dominant satellite type in the unaligned contigs is pCht7, a telomeric satellite found in chimpanzee and gorilla (*9*). Among those contigs that are composed primarily of satellite sequence (>75%), the majority (95.3%) carry the subterminal satellite repeat pCht7.

6.1 Estimating heterochromatic content

In order to approximate the content of the heterochromatic satellite sequence of the gorilla genome, we reanalyzed repeat content of the all underlying SMRT sequence reads and compared it to the repeat composition of the assembled contigs. The reads were annotated with the corresponding satellite information from Susie3 and aggregated by satellite type (table S15). We estimate that 10% of the gorilla genome is composed of various classes of satellite sequence. pCht7 is the most prevalent repeat found in both the assembled Susie3 contigs and the aligned reads (50.7% and 40.0% of the satellite sequence, respectively). Because the median ratio of the satellite content of the reads is higher in Susie3 contigs (80.41) compared to the background satellites (71.74), it is clear that sequence reads containing satellites have been collapsed and are not being properly assembled. Almost none of these satellite-enriched contigs have been included in our AGP. By this metric, the satellites creating the greatest difficulty for assembly include ALRY-MINOR_PT (the minor repeat unit of chimpanzee alpha repetitive DNA), which shows a 2,916-fold excess in the reads compared to Susie3 contigs, and the general consensus centromeric alpha satellite ALR/Alpha, where there is a 1,801-fold excess when compared to assembled Susie3 contigs.

6.2 Putative functional macrosatellites

We defined satellites within the Susie3 assembly using Tandem Repeats Finder cataloging 8,595,075 tandem repeats (429 Mbp) of which 77.4% were previously flagged by RepeatMasker. Next, we determined which tandem repeats were potentially transcribed by aligning 32.6 million RNA-seq reads created from gorilla iPS cells to Susie3 using STAR (*43*). Alignments were filtered based on quality (SAMtools view -q 30) and whether the alignments spanned multiple exons. 1,817,918 (21.2%) tandem repeats overlapped RNA-seq alignments that span multiple exons. We further restricted our analysis to the largest tandem repeats (>5 kbp) in order to enrich in macrosatellites. This corresponds to 10,882 large tandem repeats comprising 238,818,615 bp (fig. S25). We provide a complete listing of these large repeats and indicate whether they are likely collapsed or full length based on read-depth analysis (table S16; see separate Excel file).

To identify novel macrosatellites, we aligned the large tandem repeats to GRCh38 using BLASR (-maxMatch 20 -bestn 1). The vast majority (10,317 or 94%) of the large tandem repeats did not align to GRCh38, only 22 of which contained spliced RNA-seq support. Two macrosatellites contained spliced RNA-seq support and Augustus annotations. The flanking regions of the macrosatellites were aligned to GRCh38 using BLASR. We identified a novel macrosatellite that occurred within the intron of the *YPEL1 (Yippee-like)* gene on HSA chromosome 22q11.2. This macrosatellite was not found in other human genome assemblies, GRCh38 or CHM1 (figs. S26-S28), strongly suggesting a gorilla-specific expansion. Although the precise function of this gene is not known, it is thought to be a target of bone morphogenetic proteins and play a role in normal craniofacial development (*45*). The remaining novel macrosatellite (25 bp unit length) was observed in GRCh38 and CHM1 but appeared to be greatly expanded in gorilla when compared to human (figs. S29, S30). In Susie3, for example, the macrosatellite contained roughly 250 repeat units, compared to the six repeat units in GRCh38. This macrosatellite occurred within an intron of *TSC1* (fig. S31)—a gene associated with tuberous sclerosis.

6.3 Macrosatellite composition of closed gaps

We also specifically examined the tandem repeat content of the closed gorGor3 gaps in Susie3 since our previous analysis of the human genome indicated that gaps would be enriched for these elements (*46*). 548,255 (17.8 Mbp) tandem repeats intersected with a closed gap (table S17; see separate Excel file). 128,459 (23.4%) of these tandem repeats overlap or are flanked with at least one spliced RNA-seq sequence (fig. S32). The longest tandem repeat mapped within a gene-rich region of chromosome 19, had a length of 15.3 kbp, and overlapped 241 spliced RNA-seq sequences, corresponding to three Augustus annotations for *GPI*—a gene that encodes a glucose phosphate isomerase protein (fig. S33). This protein is involved in the second step of glycolysis, aiding in the conversion of glucose-6-phosphate to fructose-6-phosphate. Mutations in this gene have been linked to chronic hemolytic anemia (*47*). It is worth noting that this macrosatellite is collapsed in GRCh38 but resolved in CHM1 (fig. S34). The second longest tandem repeat maps near the end of chromosome 14, was 8.95 kbp in length with a monomer unit length of ~500 bp (fig. S35), and corresponds to an expansion of 165 amino acid motif of the AHNAK2 protein—a gene implicated in FGF1 export and skeletal muscle function (*48*) (fig. S36). The long reads allowed these large repeat motifs to be properly sequenced and assembled.

6.4 Resolution of macrosatellites in Susie3

Although these results suggest that some large minisatellites and macrosatellites may be accurately assembled, our analysis of common satellites associated with heterochromatin showed that larger ones could not, especially when the sequence identity of the monomers was high. To investigate this further, we examined the resolution of some known functionally important macrosatellites. DXZ4 is a human X-linked macrosatellite composed of 3.0 kbp unit repeated between 12 and 100 copies (*49*). We aligned DXZ4 to Susie3 using BLASR. Only a single 3.0 kbp monomer unit was found within a Susie3 contig (fig. S37). The read depth of this 3.0 kbp monomer was 893.3-fold consistent with a collapsed repeat (12-fold) during assembly. Similarly, D4Z4 is a macrosatellite found on chromosome 4 that contains *DUX4* and reductions of copy number have been associated with fascioscapularhumeral muscular dystrophy. The D4Z4 macrosatellite is

composed of 3.3 kbp monomer repeated between 10 to 100 copies. We identified 2.62 units (8.7 kbp) of D4Z4 mapping to the end of a single Susie3 contig (fig. S38). Its high read-depth 312.1-fold suggests, once again, a collapse of an 11-copy macrosatellite. In contrast, another X-linked macrosatellite, X130, is found entirely within a single contig in Susie3 (*49*) (fig. S39). X130 spans roughly 70 kbp; however, it is less conserved than DXZ4 in terms of identity (68-85% compared to 99%) and has no obvious repeat unit. The difference in identity and repeat length allows Falcon to assemble through the macrosatellite. Thus, caution should be exercised in analyzing and interpreting this class of satellite. For each tandem repeat, we indicate the fold-sequence coverage as a metric of collapsed repeats in the accompanying tables.

7. Segmental duplication content

We analyzed both the original gorilla genome assembly (gorGor3) and the new assembly (Susie3) for segmental duplication content using a whole-genome analysis comparison pipeline (WGAC) developed previously (*50*). All putative segmental duplications in both assemblies were mapped to GRCh37 using UCSC's liftOver tool. In order to focus on *bona fide* gorilla segmental duplications (as opposed to assembly artifacts), we considered only those regions where sequence read depth from 31 different gorillas (27 Western lowland, 3 Eastern lowland and 1 Cross River) had predicted a gorilla segmental duplication (*51*). A region was deemed a gorilla segmental duplication if it had a copy number >= 2.5 in at least 27/31 of the gorillas (~90% of the population). We merged all duplicated regions that were 1000 bp from each other as described previously for the whole-genome sequence detection (WSSD) pipeline (*35*).

The analysis predicted 120 Mbp of segmental duplication within Susie3 when compared to 61.9 Mbp in gorGor3, although we should stress that detection of segmental duplications both by assembly (WGAC) and read depth (WSSD) does not imply that the gorilla segmental duplications are correctly assembled in either Susie3 or gorGor3. The largest gains occurred for the most highly identical duplications, especially in Susie3 for segmental duplications >98% identity (fig. S40a). In addition to the segmental duplication content, one of the greatest differences occurred in the length distribution of the duplicated sequences between the two assemblies. 80% of Susie3 segmental duplications exceeded 10 kbp in length (average size = 11 kbp) with 26 duplications >100 kbp. In contrast, the average size of segmental duplications in gorGor3 was only 3.5 kbp with only 23% of the duplications exceeding 10 kbp. We observed a complete absence of duplications >45 kbp in the gorGor3 assembly (fig. S40b). The smaller duplication sizes in the gorGor3 assembly represent duplications that are either collapsed or unresolved, reflecting the limitations of assemblies constructed from short-read and low-coverage Sanger sequencing.

7.1 Copy number variation analysis

Since segmental duplications can vary from 2 to ~25 copies in ape genomes, we assessed each assembly's performance with respect to copy number. A total of 8,324 loci were detected as fixed copy number gains in Western lowland gorillas relative to humans using read-depth profiles from gorilla whole-genome shotgun sequence (*17*). We mapped the human sequences with gorilla copy number gains to the gorGor3 and Susie3 assemblies

using BLASR (parameters -sam -bestn 100 -maxMatch 20) and detected sequences at least 1 kbp in length with at least 90% identity and where 80% of the human sequence length was represented. We found that segmental duplications were approximately three times more likely to be multicopy in Susie3 when compared to gorGor3 (i.e., a total of 619 events had a greater copy number in Susie3 compared to GRCh38, and 248 events were expanded in gorGor3 relative to GRCh38). We observe that while there are a greater number of duplications resolved in Susie3, they are biased towards shorter events (red, blue, and green points in figs. S41, S42).

We assessed agreement between the average copy number of duplicated sequences reported by Sudmant, P. et al. (*17*) and the copy number in the gorGor3 and Susie3 assemblies. While both assemblies had copy number variable sequences missing (figs. S42, S43), more sequences were resolved in the Susie3 assembly than gorGor3. In this particular analysis, 680 of the gorilla-duplicated sequences (35.5 Mbp) were not resolved in Susie3 but 2,537 sequences (69.0 Mbp) were not resolved in the gorGor3 assembly. In total, we find an additional 22,855,171 bases from duplicated loci in the Susie3 assembly not represented in gorGor3, and 4,047,093 similar bases in gorGor3. In summary, sequences that are duplicated in gorilla have more copies resolved in Susie3 than in gorGor3 but overall a significant fraction of segmental duplications remain unresolved in either assembly.

7.2 Copy number analysis of gene families

In addition to the copy number analysis of segmental duplications, we also focused on comparing gene family expansions in the two assemblies by mapping human gene annotations using BLASR and counting the number of expanded gene families in each assembly (see Methods at the end of this section). We identified 407 genes/pseudogenes expanded in Susie3 versus GRCh38, compared to 152 in gorGor3 using similar methodology (table S18; see separate Excel file). Only a small fraction of expanded genes were shared by both assemblies (fig. S44). An orthogonal approach based on sequence read depth confirms 216 of 407 (53.1%) of the expanded loci in Susie3, and 32 of 152 loci (21.1%) in gorGor3. We note that the average length of genes detected as expanded in either assembly (9-11 kbp) was significantly shorter than the average gene length (41.4 kbp) consistent with a bias for shorter genes as part of this analysis. As expected, both of these represent an underestimate compared to the gene families identified as expanded by read-depth analysis (*17*). For example, 1,521 genes reported as expanded in (*17*) were not observed as expanded in Susie3. In some cases, not a single instance of a full-length gorilla gene can be found (e.g., *C1QTNF3* and *AMACR*). In other cases when gene families were particularly large (e.g., *GOLGA6*, *NPIP*, etc.), we were unable to assign copy number differences discretely to particular members of the gene family.

In simpler cases, we were able to confirm expansions. For example, we partially resolved the carboxyl esterase gene family, *CES1* and *CES2*, and multiple copies of the olfactory receptor gene (n=43), PRAME cancer/testes expressed genes (n=44), immune response (n=22) and keratin production associated genes (n=10) (table S18; see separate Excel file). We found additional evidence of expansion of xenobiotic detoxification enzymes,

including alkaline phosphatastes *ALPI* and *ALPPL2*, aldo-keto reductase *AKR1B10*, and numerous members of the cytochrome P450 family.

Many of the duplicated genes we identified map to smaller contigs (<200 kbp) in Susie3. Of the 407 genes that have copy number expansions in Susie3, 68% (277/407) map to these short contigs. This represents a significant enrichment for duplicated genes in short contigs assuming a uniform distribution model for genes ($p < 0.05$, binomial). Moreover, these short contigs often show highly variable sequence coverage depth (fig. S45) indicative of collapsed sequences or lower consensus identity. We are able to confirm an instance of a gene, *WFDC2*, duplicated in the gorGor3 assembly not confirmed by read depth and the Susie3 assembly (fig. S46).

Methods: The sequences of human genes were extracted from GRCh38 using RefSeq coordinates (release r73), selecting the first sequence when multiple entries for the same locus exist. The sequences of genes were mapped to gorGor3, GRCh38, and Susie3 using BLASR (version rc43) with parameters "-maxMatch 20 -sam -bestn 20". Alignments with <75% sequence identity or <75% of the maximum scoring alignment were discarded. This permissive cutoff allows for the detection of more diverse gene families, at the expense of a less precise assignment of which gene in a family is expanded.

8. Structural variation detection

8.1 Insertions and deletions (indels)

We detected structural variation within the gorilla genome by mapping contigs in Susie3 to GRCh38 and validating using independent local alignments. A total of 2.76 Gbp of the human genome was mapped by contigs greater than 200 kbp, and 2.76 Gbp of local reassembly, with 2.72 Gbp in common. Regions of contigs with uncharacteristically low read support coverage (below 40X average coverage), or too high (above 120X coverage), were discarded as spurious. There were a total of 117,512 variants >=50 bp (58,621 insertions, mean 790 bp, and 58,891 deletions, mean 776 bp). Variants contained within alignments are base-pair resolved (N=117,450), and not base-pair resolved when split across alignments (N=102).

Of the 117,512 indel variants detected, 101,109 (86%) were not observed in previous studies (*16*, *17*, *52*), including 51,066 deletions with a mean size of 574 bp and 50,043 insertions with a mean size of 693 bp (fig. S47). A total of 37.1% inserted and 40.3% of the deleted sequences in Susie3 are annotated as mobile element insertions (MEIs), though the total proportion of bases annotated as MEIs are roughly equal (58.5% insertion and 58.3% deletion). A major difference in MEIs between gorillas and humans is PTERV1, an endogenous retrovirus shared in apes and Old World monkeys, but absent in the human genome (*18*). Southern blot analysis indicated over 100 copies of PTERV1 in gorilla genomes; although we only detect 32 copies of PTERV1 greater than 6 kbp present in gorGor3 (fig. S48). In contrast, 152 PTERV1 over 6 kbp were detected in Susie3. The total number of PTERV-1a fragments in gorGor3 is 266, although 59% of these are fragmented representations less than 2 kbp.

To illustrate the resolution of mobile elements in the gorilla genome by each assembly, we plotted the locations of PTERV insertions greater than 6 kbp and SVA greater than 2 kbp in the Susie3 and gorGor3 assemblies (fig. S49). There were 186 such PTERV and 426 SVA insertions in Susie3, and 24 such PTERV and 63 SVA insertions gorGor3. Assuming that the relative proportion of full-length SVA and PTERV elements is similar between the two gorillas, we estimate that Susie3 assembly resolved 6- to 8-fold more full-length retrotransposons.

To determine if the structural variants were fixed in the gorilla lineage, we assessed structural variant insertions and deletions for support in other Western lowland gorillas (table S19; see separate Excel file). Of the 117,410 indels called by whole-genome alignments of Susie3 to GRCh38 with base-pair resolution, 84,804 (72%) were supported by Illumina data from all six Western lowland gorillas (>5 properly paired reads spanning breakpoints with mapping quality >=30) while 102,453 variants (87%) were supported by at least four gorillas.

8.2 Inversions

We searched for inversions by mapping 20 kbp and 100 kbp tiled sequences covering Susie3 contigs to GRCh38 and examining for subsequences inside each tiled sequence that improve the alignment score when the subsequence is replaced by its reverse complement using the program screenInversions (www.github.com/eichlerlab/pacbio_variant_caller) with parameters -r --noClip -g 500 (table S20; see separate Excel file). We additionally searched for contigs with whole-genome alignments that are split into at least three sub-alignments where an internal subsequence is aligned in reverse orientation of flanking subsequences.

In total, we detect a merged set of 697 inversions (fig. S50). The average inversion length was 3,523 bp (SD 16,384 bp). Previous studies detecting structural variation using discordant end-sequenced BAC (*16*) and single-molecule sequencing (*3*) detected 12% of the inversions found in the Susie3 assembly (82 events; mean 18,072 bp; 44,380 bp SD). There are between 323 and 426 micro-inversions between human and chimpanzee (*53*, *54*). These estimates indicate that there are 2.2X-2.9X more human-gorilla inversions than human-chimpanzee inversions. This is higher than the 1.3- to 1.7-fold difference in evolutionary time between the human-chimpanzee and human-gorilla ancestor likely due to methodological differences and increased resolution when compared to previous studies (*53*, *54*).

We inspected the repeat content within 25 bp of either side of inversion breakpoints based on mapping to the human reference. For 453 of the 697 inversions (65%), an annotated repeat was present at both breakpoints. More specifically, 361 inversions (52%) had the same repeat class at both breakpoints with a significant enrichment of SINE and simple repeats and depletion of DNA and satellite repeats compared to the proportion of each repeat class present in the entire genome (table S21; $p = 0.000001467$, X-squared = 35.0553, df = 5).

8.3 Putative functional variants

Only 397 insertion/deletion structural variants (0.3%) intersected coding exons based on human RefSeq annotation (tables S22, S23; see separate Excel file). We investigated the subset of these potentially gene-altering variants that had not been observed in previous studies (*16*, *17*, *52*, *20*), mapped within unique regions of the human reference where structural variations can be most confidently called and were supported by Illumina data from all six Western lowland gorillas. After applying these filters, we identified 145 structural variants (76 deletions and 69 insertions) affecting 110 distinct RefSeq genes. Interestingly, 29 of these variants (20%) affected exons that were not present in all isoforms of the affected genes suggesting a potential for variation with a tissue-specific effect. Thus, 46 genes contain novel structural variants that alter exons and appear to maintain reading frames in Susie3. These affected genes can be classified into four major categories, including highly repetitive and variable genes like mucins and zinc fingers, sensory perception genes including *RP1L1* and olfactory receptors, transmembrane domains including *LILRB5*, and nucleotide binding proteins *MDN1* and *OBSCN*. Additionally, 7 inversions (1%) completely encompassed 14 human RefSeq genes, 8 inversions (1%) broke transcripts for 8 genes, and 206 inversions (30%) fell inside 203 distinct genes, including 2 genes for which the inversion broke an exon (*MAMLD1* and *PDE4DIP*) (table S24; see separate Excel file).

To identify coding sequence unique to gorilla, we searched for RNA-seq alignments from GenBank mRNAs, Trinity-assembled gorilla RNA-seq data (see above), and GENCODE-assembled transcripts within inserted regions in Susie3. A total of 138 RNA-seq marks from GenBank, 3,453 Trinity, and 31,202 GENCODE annotated exons alignments overlap inserted sequences. A 570 exon subset of the Trinity annotated exons and 3,543 exon subset of the GENCODE datasets have HUGO gene names (table S22; see separate Excel file). The well-characterized human-specific deletion of a single exon deletion of *CMAH* (*55*) is detected in the GenBank transcript alignments, though the sequence is absent from the gorGor3 assembly. Gorilla deletions: In total, 502 human exons and 745 untranslated regions overlap deletion events (table S23; see separate Excel file).

We searched for genes that show expansion of their coding regions mapping the splice start and endpoints from RefSeq in GRCh38 to Susie3. A total of 1,073 exons were expanded by at least 47 bp in Susie3 (table S25; see separate Excel file). Because some of these exon expansions may be a consequence of the improved sequence resolution of STRs and VNTRs using SMRT sequencing technology, we searched for evidence of these exon expansions in additional PacBio human genomes (e.g., sequenced hydatidiform mole CHM1). We found only 53 of these insertion events to have any overlap, indicating that the majority of these 1,020 events represent CDS expansions in the gorilla genome.

In order to identify a more complete spectrum of functional elements, we began by annotating 117,410 base-pair resolved structural variants (SVs) (86.5 Mbp) using GENCODE gene annotations for human GRCh38 (v22) after filtering for transcripts with support from at least one high-quality EST (table S26). Compared to the 384 variants

(0.3%) intersecting RefSeq coding exons in our original SV analysis, we identified 279 variants that intersected high-quality GENCODE coding exons (0.24%). To identify variants with putative effects on regulatory elements, we annotated those that intersected with deoxyribonuclease I hypersensitive (DNase I HS) sites associated with open chromatin, histone 3 lysine 4 trimethylation (H3K4Me3) signals associated with transcriptionally active regions/promoters, and histone H3 acetylated on lysine 27 (H3K27Ac) signals associated with enhancers (*12*). We annotated two sets of DNase I HS sites, including high-support DNase clusters that occur in >50% of the 95 published tissues and clusters from fetal brain tissue (*56*). We defined putative promoters (H3K4Me3) and enhancers (H3K27Ac) based on a signal >1 standard deviation across the entire genome for each one of the seven available sample/tissue combinations.

Based on this expanded annotation, we identified 2,205 variants (1.9% of events, 2.9% of bases) affecting coding or noncoding genic exons and 10,466 variants (8.9% of events, 14.2% of bases) affecting putative regulatory regions including DNase I HS clusters, promoters tagged by H3K4me3 signals, and enhancers tagged by H3K27Ac signals. Combined, we identified 12,196 gorilla variants (10.4% of events, 15.6% of bases) affecting functional genomic sequence. To determine what variation was specific to the gorilla lineage, we filtered these variants to include only those present in all other Western lowland gorillas (Illumina WGS, n=6) and absent in both human (n=1) and chimpanzee (n=1) genomes. Using these filters, we reduced the set to 2,450 distinct fixed, gorilla-specific variants (2.1% of events, 4.5% of bases), including 392 variants affecting genes and 2,151 variants affecting regulatory regions. These fixed, gorilla-specific variants affected 371 distinct genes. These genes were not significantly enriched for a specific function although small sets of genes (n=4-8) clustered into functional groups, including nucleic acid and protein transport, interleukin secretion, and cytokine secretion (DAVID GO (*57*, *58*)). At the chromosomal level we observed a slight enrichment of variants on human chromosomes 19 and 17 based on the genomic mean proportion of bases affected +/- one standard deviation (fig. S51).

To assess SV enrichment by functional element, we calculated the proportion of functional bases affected by dividing the number of nonredundant affected exonic or regulatory bases by the total bases for the corresponding functional category in the genome. Note, we excluded GENCODE biotypes labeled as "predicted", inactivated immunoglobulin pseudogenes, and relatively rare biotypes (those with <15,000 annotated exons). We tested significance of the enrichment by simulation (n=1,000,000 replicates of the SV distribution and recomputation of the observed statistics). Long intergenic noncoding RNAs (lincRNAs) were the most proportionally affected by SVs with 1.1% of annotated bases affected (fig. S52, table S27). The 5.6-fold enrichment was not, however, significant by simulation. In contrast, protein-coding genes and regulatory regions associated with promoters, enhancers, and DNase clusters were the least affected with 0.5% of combined genomic bases in those categories affected by SVs. The depletion of SVs (2- to 10-fold) was statistically significant ($p < 10^{-6}$) consistent with the action of purifying selection for these conserved functional elements.

Although most of this structural variation was never previously reported, we estimated what fraction of it could have, in principle, been identified in the original gorilla genome assembly (gorGor3). We remapped each putative functional SV back to the gorGor3 assembly by aligning 250 bp on either side of a deletion breakpoint (or the complete insertion sequence padded to 500 bp for events smaller than 500 bp) and requiring >=90% of the original sequence to align. Out of the 2.38 Mbp of functional sequence affected by fixed, GSVs, 1.36 Mbp (57%) could be genotyped in gorGor3. Thus, 43% of these fixed gorilla SVs still could not have been predicted in the original Illumina-based assembly even with prior knowledge the precise breakpoints.

As a final assessment of the potential importance of regulatory and coding SVs with respect the evolution of the gorilla and human genomes, we performed a more refined proximity analysis where we assessed the spatial correlation of SVs and annotated functional elements in the genome. For this analysis, we focused on the 2,323 fixed gorilla-specific structural variants (GSVs) between the human and gorilla genomes and tested for enrichment between GSVs and functional elements using the Genometricorr package (*59*). Permutation of the Jaccard distance revealed that there was a depletion of GSVs that overlapped with genes (empirical $p < 0.001$; relative KS p-value = 2.33e-15), with fetal CNV DNase sites (empirical $p < 0.001$; relative KS p-value < 2e-16), with H3K4me3 (empirical $p < 0.002$; relative KS p-value = 0.009) and with H3K27AC (empirical $p < 0.001$; relative KS p-value = 1.57e-9). These findings are consistent with purifying selection acting on functional regulatory elements.

In addition to measuring overlap, we quantified the spatial correlation between GSVs and the regulatory elements by comparing the midpoints of each functional element to the midpoint of the GSVs based on genomic coordinates. The midpoints of the GSVs overlapped significantly less with the protein-coding genes (fig. S53A) (projection test p-value = 6.03e-12) and with fetal DNase hypersensitive sites (fig. S53B) (projection test p-value = 4.3e-10) consistent with depletion of lineage-specific SVs over these sites. Interestingly, we observed a modest spatial enrichment (less than 100 bp) between GSVs and putative promoter and enhancer signatures (H3K4M3/H3K27AC (fig. S53C-D; projection test p-values: 0.041, 4.1e-15). As an independent validation of the GSV/H3K27AC correlation, we ran a 50 bp sliding window permutation test along the observed spatial distribution. We randomly shuffled the GSVs 1000 iterators for each window and measured the number of trials where there were more random GSVs within the window than observed. GSVs at a distance of 130-180 bp from H3K72AC were overrepresented (empirical p-value < 1e-3). We identified 327 protein-coding genes within 10 kbp of H3K4m3 marks that are overlapping or near GSVs (<100 bp) and 672 genes that had the same spatial pattern for H3K27AC marks. Of the 775 unique genes, nine overlapped with genes differentially expressed with CTCF binding changes between human and gorilla (*ADAMTS10*, *ALDH1L1*, *CDH1*, *COL5A1*, *GRK5*, *IGF2BP1*, *INSR*, *IQGAP2*, and *SRC*) (*4*) (tables S28, S29; see separate Excel file). We found an additional eight of the original differentially expressed genes were also affected by SVs detected with the Susie3 assembly and fixed in Western lowland gorillas (*AMOTL1*, *DUSP4*, *HNF1B*, *ITGB8*, *SGPP2*, *TCL1A*, *ZDHHC19*, and *ZNF607*). None of the SVs affected

protein-coding regions of these genes; however, six affected 3' UTRs and two affected exons of noncoding transcripts.

8.4 Gorilla deletions

To assess the accuracy of both detection and breakpoint structural variation definition, we specifically focused on a set of 760 deletion events that had been previously identified as fixed events based on a read-depth analysis of a population of Western lowland gorillas (*17*). There are 664 events entirely contained within Susie3 contigs at least 200 kbp in length. Of the 96 missed deletion events, 46 are found in segmental duplications in human representing an eightfold enrichment, consistent with the lack of resolution of segmental duplication architecture in Susie3. Of these, 616 (92.8%) events overlap at least 50% with the intervals of deletions discovered in the assembled contigs, but only 414 events (62.3%) have a more stringent (90%) overlap, indicating an alternative resolution of deletion breakpoints. The resolution offered by the contig level sequence allowed us to investigate the local architecture surrounding the deletion event. Three of the deletion events surround other genomic rearrangements; *CLC* and *SELV* flank an inversion, and *FAM75E1* and *LOC392364* (a pseudogene). This particular region of the gorilla genome has been subject to a complex series of structural changes (fig. S54). These were not resolved in the original gorilla assembly due the large number of gaps and the fact that the human genome was used to guide the order and orientation of contigs.

8.5 Gorilla copy number variant analyses

We also assessed the resolution of sequence architecture of sequences known to be copy number variable (CNV) in Western lowland gorillas. A total of 8,324 sequences were reported as CNV Western lowland gorillas in (*17*), but because these sequences were detected as variable read-depth profiles from gorilla high-throughput sequencing mapped to human, the organization of these duplicated sequences is not known (table S30; see separate Excel file). To determine the extent that the architecture of the CNV sequences are resolved in our assembly, we mapped the orthologous human sequences from the copy number map (*17*) to both Susie3 and gorGor3 and counted the number of alignments of each sequence that are over 1 kbp with at least 90% identity and 80% the human sequence length. There are 6,515 CNV sequences with an average of two or more copies in Western lowland gorillas. Of these, 3,391 have two or more copies in Susie3, whereas there are 2,112 duplicated CNV sequences in gorGor3, indicating an improvement in resolution of duplication architecture in Susie3 over gorGor3. There is a mean of 1.4 (4.5 SD) fewer copies of CNV sequences in Susie3 than in (*17*), indicating that resolution of highly duplicated sequences is incomplete though it is an improvement over gorGor3, which has on average 2.3 (4.7 SD) fewer copies than in (*17*). We also assessed the resolution of sequence architecture of known CNV sequences in Western lowland gorillas. A total of 8,324 sequences were reported as CNV Western lowland gorillas in (*17*), but because these sequences were detected as variable read-depth profiles from gorilla high-throughput sequencing mapped to human, the organization of these duplicated sequences is not known. To determine the extent that the architecture of the CNV sequences are resolved in our assembly, we mapped the orthologous human sequences from the copy number map (*17*) to both Susie3 and gorGor3 and counted the number of alignments of each sequence that are over 1 kbp with at least 90% identity and

80% the human sequence length. There are 6,515 CNV sequences with an average of two or more copies in Western lowland gorillas. Of these, 3,391 have two or more copies in Susie3, whereas there are 2,112 duplicated CNV sequences in gorGor3, indicating an improvement in resolution of duplication architecture in Susie3 over gorGor3. There is a mean of 1.4 (4.5 SD) fewer copies of CNV sequences in Susie3 than in (*17*), indicating that resolution of highly duplicated sequences is incomplete though it is an improvement over gorGor3, which has on average 2.3 (4.7 SD) fewer copies than in (*17*).

There are a total of 1,992 CNV sequences that have greater copy number in Susie3 than in (*17*) indicating the utility of *de novo* assembly in detecting the architecture of duplicated and polymorphic sequences. There are 891 expanded CNV sequences gorGor3, with the disparity between Susie3 and gorGor3 increasing as copy number increases (fig. S55). As expected, as the length of the CNV sequence decreases, the ability to resolve additional copies of the sequence increases, as indicated in fig. S56. We detected a total of 619 events that have a greater copy number in Susie3 compared to GRCh38. As a reference, we detected 248 events that were expanded in gorGor3 relative to Susie3 (table S18; see separate Excel file).

8.6 Comparative structural variation

We estimated the degree of structural rearrangements between the Susie3 assembly and the human reference (GRCh37) through whole-genome alignment by LASTZ (*60*) followed by alignment chaining and netting as previously described for the gibbon assembly (*5*). (Note: an earlier version of the human genome was used for this analysis because it served as a baseline for previous primate comparisons of large-scale evolutionary rearrangements.) The number of rearrangements between assemblies was measured by the number of collinear blocks present in the final alignment nets. We performed the same analysis for chimpanzee (panTro4), gorilla (gorGor3 and gorGor4), and gibbon (nomLeu3) assemblies for comparison. We found that the distribution of rearrangements in Susie3 is more consistent with the chimpanzee than the current gorilla assembly. Overall, Susie3 had marginally more rearrangements at each size threshold than chimpanzee while both assemblies had a significantly greater number of rearrangements than the current gorilla assembly (gorGor3) at all thresholds greater than 10 kbp (fig. S57). In concordance with (*5*), gibbon has significantly more rearrangements at all sizes than all other species. Interestingly, gorGor3 contains more rearrangements than Susie3 and panTro4 at the 10 kbp threshold and nearly as many as nomLeu3. This pattern highlights the relatively fragmented representation of the gorilla genome in gorGor3.

8.7 Lineage-specific gene variation

We identified GSVs and indels affecting genes using long-read data by selecting all SVs (indels >=50 bp) and small indels (3-49 bp) that disrupted a coding or noncoding exon in GENCODE gene annotations (v22). We additionally required these variants to be fixed, gorilla-specific and mapping outside of segmental duplications. Using these filters, we identified 3,915 distinct genes affected by variants in Susie. We compared our new set of genes disrupted by structural variation to previously published reports (e.g., Scally et al. (*4*) report 1,594 gorilla-specific gene deletions; Prado-Martinez et al. (*10*) report 125 genes affected by deleterious indels (>=3 bp) or large deletions in the gorilla lineage and

Sudmant et al. (*17*) report 262 genes affected by fixed deletions in the gorilla lineage). After excluding genes known to occur in segmental duplications where SV and indel calling is less reliable for both assemblies, these published data provide 1,607 gorilla-specific genes affected by SVs or indels. Only 283 putative disruptive genes were identified in both assemblies (after converting GENCODE names to RefSeq nomenclature to match the earlier studies (fig. S58). We further investigated genes predicted to be disrupted exclusively by Illumina genome sequence analysis (n=1,324). We found that 603 of these "genes" did not have a complete GENCODE annotation in Susie3. This included four disrupted genes detected by PacBio sequencing but did not have sufficient population support to be classified as fixed and gorilla-specific. The majority (485 genes) corresponded to predicted genes often with undescribed open-reading frame that could not be compared between GENCODE, Augustus and RefSeq. After excluding 10 genes mapping to the Y chromosome (unavailable for annotation in two female gorillas), we found 108 genes in the Illumina-only studies that could not be confirmed in the Susie3 assembly. Overall, our results suggest that the majority (95%) of the genic SVs detected by Susie3 are novel.

8.8 Human deletions

Previous research has identified sequences >=500 bp missing in the human reference assembly that are present in nonhuman primate references, including 5,361 (10,870,110 bp) in chimpanzee (panTro3), 5,260 (8,600,762 bp) in gorilla (gorGor3), and 19,412 (37,495,683 bp) in orangutan (ponAbe2; (*17*)). We assessed the presence of these sequences in Susie3 with high-quality BLASR alignments (alignment parameters: -bestn 1 -affineAlign -affineOpen 8 -affineExtend 0 -maxMatch 30 -sdpTupleSize 13; MAPQV >=30 and alignment identity >98%). Of the 5,260 gorilla sequences, 3,699 (70%) had an unambiguous representation in Susie3 with 538 (10%) mapping to Susie3 contigs that anchored to GRCh38. An additional 216 sequences (4%) had low mapping quality alignments. Correspondingly, 36% of chimpanzee sequences (1,909 with 382 anchored in GRCh38) and 9% of orangutan sequence (1,675 with 442 anchored in GRCh38) mapped unambiguously to Susie3. In the other nonhuman primates, sequences mapped with low quality corresponding to ~1% of chimpanzee and ~1% of orangutan. Thus, Susie3 captures 80% of gorilla sequence missing in the human reference, 38% of chimpanzee sequence, and 10% of orangutan sequence.

8.9 Analysis of major histocompatibility complex (MHC) classes I and II in Susie3

We identified two MHC genes deleted in Susie3: *HLA-J and HLA-F. HLA-J* is annotated as a noncoding pseudogene and only present in RefSeq and not in GENCODE. It is almost completely deleted (chr6:30002849-30009100). By contrast, the deletion of *HLA-F* is restricted to the (chr6:29722792-29722894) 5' UTR of its longest GENCODE transcript. Its three other transcripts are unaffected by the deletion and the affected transcript is not present in the RefSeq annotations. *HLA-F* is considered a nonessential HLA gene but was a progenitor to other more critical HLA genes (OMIM # 143110).

The Susie3 assembly also contains a 2.04 Mbp contig (000730F_quiver) with an inserted copy of a gene similar to *HLA-A* (fig. S59), as well as two additional smaller contigs (001838F_quiver, and 000737F_quiver) that also contain sequences similar to *HLA-A*. We observe that the average sequence coverage across the inserted *HLA-A* locus is 31-

fold for 00730F_quiver (gray bar) and 35.6-fold for contig 001838F_quiver. In contrast, the sequence flanking the locus is 62.8-fold coverage. These results suggest that two alternate *HLA-A* haplotypes have been assembled and are sufficiently diverse that allelic variation can be distinguished. *HLA-A* is one of the classic polymorphic class I genes thought to interact with both the T-cell receptor and killer immunoglobulin-receptors expressed on natural killer cells (OMIM #142800).

We identified the sequence in the gorilla assembly corresponding to the human MHC Class II locus where the MHC Class II genes cluster (chr6:32247048-32937214 in GRCh37) and assessed the content of the locus in gorilla (000630F_quiver:1752442-2557056 or 000630F_quiver_rc:249994-1054608). The MHC Class II gene cluster in gorilla is composed of 804,615 bp with a G+C composition of 40.63% compared to the genome wide G+C of 40.69%. More than half of the locus is repetitive (55%) with LINE1s representing the highest proportion of repeats (19% of all bases). The segmental duplication content of the locus has expanded in gorilla relative to human with 79,166 bp (~10%) of duplications compared to 53,084 (~8%). We identified 115 kbp of gorilla-specific sequence >10 kbp across three regions (15 kbp, 48 kbp, and 52 kbp, respectively) corresponding to 14% of the locus (Fig. 3C). Although the G+C composition of this sequence is similar to the entire locus at 40.56%, this sequence is slightly more repetitive than the entire locus at 58% repeat content with LINE1s still comprising the majority of the repeat sequence by base at 18%. Similarly, the duplication content is slightly higher at ~13%. Additionally, we compared the MHC Class II locus from Susie3 to the corresponding sequence in gorGor3 (fig. S60). We found 168 gaps in the gorGor3 sequence of which 8 mapped within 100 bp of 27 distinct genes annotated by non-gorilla RefSeq annotations. Of these 27 genes, 3 were MHC Class II genes.

To verify the organization of the MHC Class II region in Susie3, we constructed a tiling path of nine BACs from the diploid Kamilah BAC library (CHORI-277) across the region, sequenced and assembled each BAC with PacBio long reads, and assembled the entire locus into two supercontigs with Sequencher (*61*). All BACs spanning the locus were selected based on BAC-end sequence alignments to Susie3 and initially sequenced with the Nextera protocol on a MiSeq with 150 bp paired-end reads. We mapped read pairs from all BACs to Susie3 with BWA-MEM (0.7.3), called variants with FreeBayes (0.9.14), and clustered BACs per haplotype based on shared SNPs and indels. We selected a tiling path of nine BACs from the same haplotype cluster for PacBio sequencing. All BACs were assembled into single contigs with HGAP/Quiver (SMRT Analysis 2.3.0) (*3*) using an HGAP cutoff of 7,800 bp. Supercontigs were assembled in Sequencher (v.5.0) by a stepwise assembly of pairs. To accept a join between PacBio assemblies of BACs in the tiling path, we required an overlap of at least 100 bp with 99% identity across which the only mismatches allowed were homopolymer errors. Based on these parameters, we assembled six clones into an 863,324 bp contig from one haplotype (MHC Class II haplotype 1) and two clones into a 289,560 bp contig from the other haplotype (MHC Class II haplotype 2) with a single redundant clone excluded from the first haplotype. The majority of the MHC Class II haplotype 1 contig (863,218 bp) aligned to Susie3 with 99.6% identity while the entire MHC Class II haplotype 2 contig aligned with 99.9% identity. In contrast, 859,977 bp of the haplotype 1 contig aligned to

gorGor3 with 98.1% identity and 289,555 bp of the haplotype 2 contig aligned with 98.9% identity.

9. Evolutionary and population genetic analyses

9.1 Divergence

The two gorilla assemblies (Susie3 and gorGor3) as well as panTro4 were aligned to GRCh38. The average sequence divergence within a 1 Mbp window (step size 250 kbp) was 1.30%, 1.65%, and 1.60% for panTro4, gorGor3 and Susie3 (fig. S61), respectively. Comparing Susie3 and gorGor3 to human, we found a large fraction of 1 Mbp windows with greater divergence in gorGor3 (1.6% of all windows). The difference in means between gorGor3 and Susie3 are statistically significant using both parametric and nonparametric tests (fig. S61B). Regions with higher gorGor3-human divergence, compared to Susie3-human were enriched for Alu and GC content (Fig. 5B). Human chromosome 19 had the highest GC content and greatest divergence between the two gorilla assemblies (fig. S61B; table S31; see separate Excel file).

9.2 Population genetic analyses

To quantify how our new assembly affected population-based analyses, we aligned reads and called SNVs/indels for seven Western lowland gorillas against Susie3 and gorGor3. A greater fraction of the Illumina pair-end reads mapped, with higher mapping quality, to Susie3 (table S8; see separate Excel file). A similar number of SNVs were called on both assemblies (gorGor3: 15.6 million, Susie3: 14.9 million); however, Susie3 had >4X enrichment for insertions (table S8). In total, there were 17.6 million variants called against gorGor3 and 20.3 million called against Susie3 (table S8). The ratio of heterozygous genotypes was higher in gorGor3 (mean: 0.35) compared to Susie3 (mean: 0.33) (table S32; see separate Excel file). Next, we examined the average observed heterozygosity across both assemblies in 100 kbp windows. The average observed autosomal heterozygosity was 0.33 for gorGor3 (mu = 0.331; SD = 0.066) and 0.32 (mu = 0.316; SD = 0.057) for Susie3 (fig. S62). The largest difference in observed heterozygosity, between the assemblies, was on chromosome X where gorGor3's average heterozygosity was 0.27 and Susie3's heterozygosity was 0.23. Since estimates of heterozygosity are important for inferring population parameters, we sought to explain the difference in heterozygosity between the assemblies.

A plausible explanation for the increased gorGor3 heterozygosity is mapping errors in gorGor3 due to the underrepresentation of sequence. gorGor3 has over 400K gaps, many of which contain repetitive sequence. Mapping software errs on the side of sensitivity, meaning reads derived from underrepresented regions will be incorrectly placed. To test this idea, we lifted reads (including mate-pairs) from Coco (Western lowland female gorilla) that overlap heterozygous positions in gorGor3 and remapped them to both gorGor3 and Susie3. Consistent with our hypothesis, only 87% of the remapped heterozygous calls in gorGor3 were also heterozygous calls in Susie3 (tables S33, S34; see separate Excel file for S34). The remaining heterozygous calls in Susie3 have lower read depth compared to gorGor3, supporting the idea that mapping error is inflating heterozygosity in gorGor3 (fig. S63) (*62*).

We fit the PSMC model to Illumina data from four Western lowland gorillas mapped against Susie3, Susie3 with indel correction, and gorGor3 (Fig. 5C). The most recent estimates of the effective population size were 6.5k and 6.1k for Susie3 and gorGor3, respectively (SD: 0.980 thousand years ago [kya], 0.560 kya) (table S35; see separate Excel file). Consistent with previous studies, the population underwent a fourfold bottleneck 30-50 kya, reducing the population from ~40K to 6K (*17*, *63*). Climate change and disease are two potential explanations for the decrease in the numbers of Western lowland gorillas (*64–66*). The effective population size between gorGor3 and Susie3 are significantly different at ~50 kya and 5 million years ago (mya) (Fig. 4D). An excess of heterozygosity caused by mapping errors in collapsed repetitive sequence and segmental duplication explains divergence at ~5 mya. The Susie3 data suggest the severity of the recent ~50 kya bottleneck was underestimated by a factor of ~1.5, using the gorGor3 assembly. These findings stress the importance of using high-quality assemblies when fitting demographic models.

9.3 Other analyses

Divergence: A modified version of BLASR (https://github.com/mchaisso/blasr/tree/1.MC.rc44) was used to align each primate assembly to GRCh38. Divergence was calculated as the total number of single-base-pair differences between two aligned sequences within a 1 Mbp window divided by the number of aligned bases. The number of aligned bases accounts for gaps in the genomes, ambiguous bases ("N") and regions that could not be aligned between human and the primates. A sliding was used to partition the data into 1 Mbp bins with a 250 kbp step (https://github.com/zeeev/vcflib/wiki). Downstream analyses were carried out in [R].

Heterozygosity: Using the previously described raw FreeBayes "type=SNP" calls, we removed any entry where there were less than 10 alleles typed (AN > 9). Basic population statistics were calculated using GPAT++ (https://github.com/zeeev/vcflib/wiki/Basic-population-statistics-with-GPAT). Observed heterozygosity is explicitly calculated as the number of heterozygous genotypes divided by the number of callable genotypes at a variant site. Smoothing was also done with GPAT++ using a 100 kbp window with 25 kbp step.

The remapping experiment used SNP calls filtered in the previous heterozygosity analysis. We wrote a program that takes a VCF file and an individual BAM file and outputs an interleaved FASTQ file. If either mate-pair overlaps the POS filed in the VCF, the read-pair is emitted. We get reads that overlapped gorGor3 heterozygous SNP calls, remapped them, and used FreeBayes for variant calling. Basic statistics, including genotype counts and depth, were gathered using VCFLIB and GPAT++.

Demographic model/PSMC: The alignments of the previously described seven Western lowland gorillas were used. The standard PSMC pipeline was used, excluding sites with PHRED quality less than 20 and depth lower than 10 or higher than 100. A total of 800 bootstrap replicates were run, 100 for each gorilla and each assembly. The PSMC plotting script was used with the flag '-R' set to retain the parsed data. These data were collated and plotted using [R].

10. Data release

The Susie3 assembly, PacBio and Illumina sequencing data for Susie, and clone sequences have been deposited in the European Nucleotide Archive and GenBank under the project accession PRJEB10880. All structural variants and their detection in other gorilla genomes have been released in Table S36 (see separate Excel file).

11. Supplementary Figures

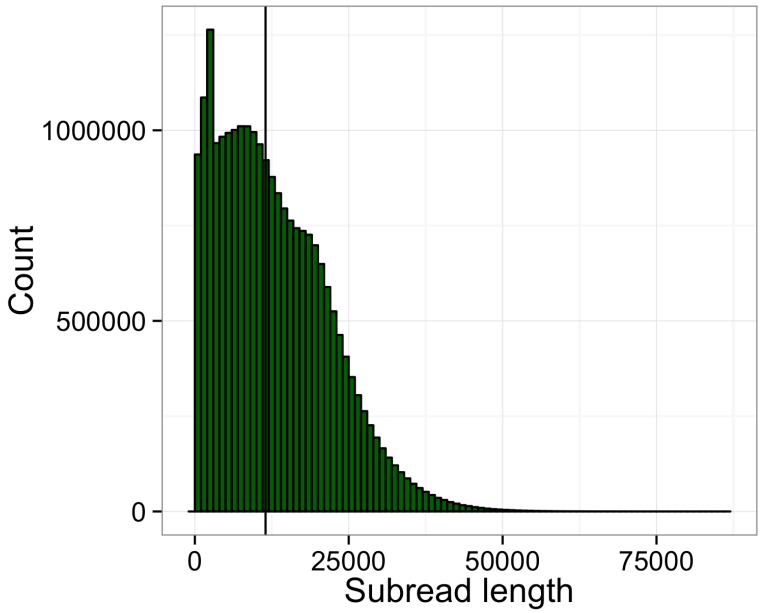

Figure S1. Histogram of PacBio SMRT P6-C4 subread lengths. A vertical line is drawn at the median subread length: 11.5 kbp.

Figure S2. Susie3 contig lengths. Susie3 contigs (N50 = 10.2 Mbp) are projected onto a human ideogram based on the human reference genome GRCh38. First two rows of black rectangles represent contigs >3 Mbp, the blue rectangles correspond to contigs <3 Mbp, and the brown rectangles indicate gorilla and human duplications (>100 kbp). Note the increase in contig fragmentation in regions of high segmental duplication content.

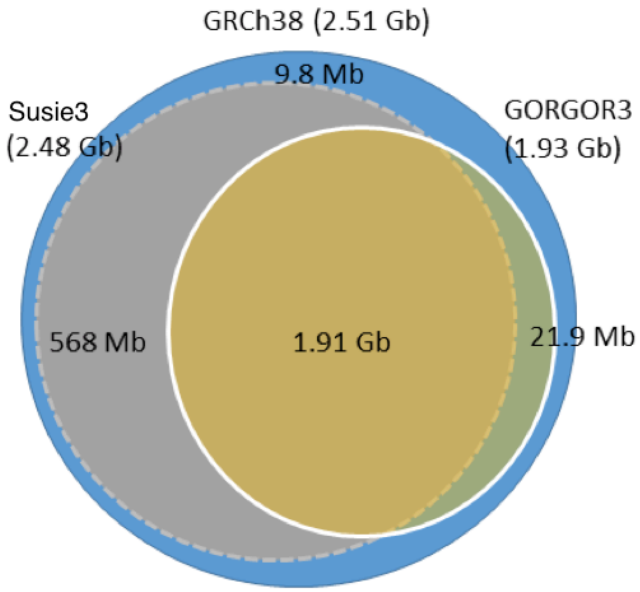

Figure S3. Euchromatin representation. Proportion of bases (non-centromeric, non-pericentromeric, nonhuman, gorilla segmental duplication) projected to human GRCh38 recovered by Susie3 vs. gorGor3.

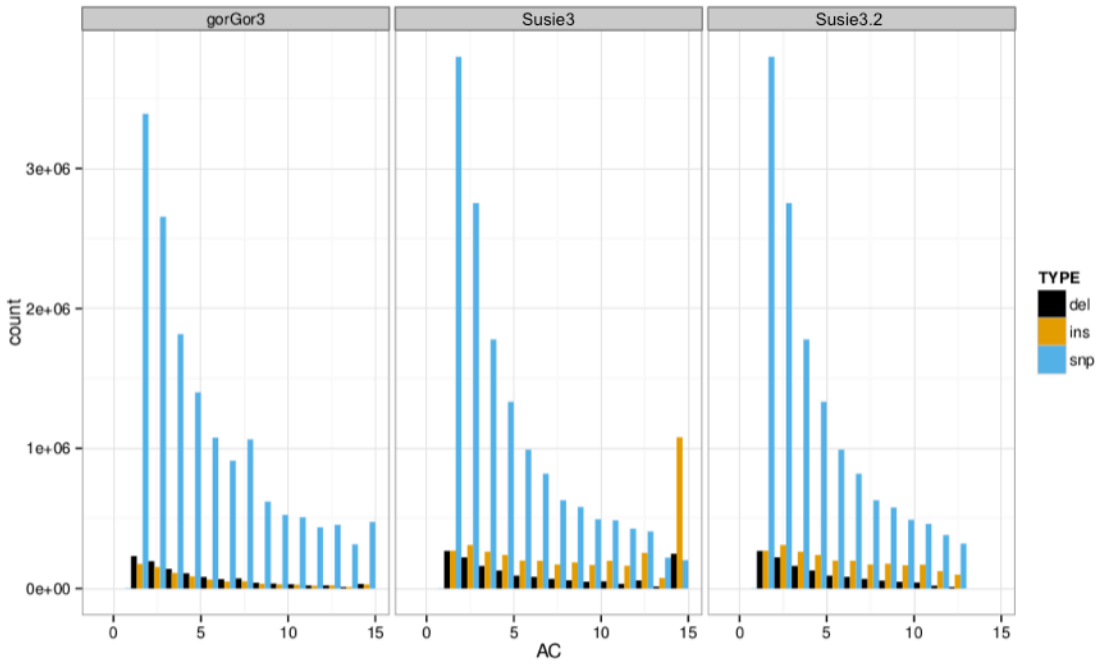

Figure S4. Distribution of allele counts for seven Western lowland gorillas by gorilla assembly and variant type. An excess of fixed insertions and deletions in Susie3 is consistent with accuracy errors in the assembly.

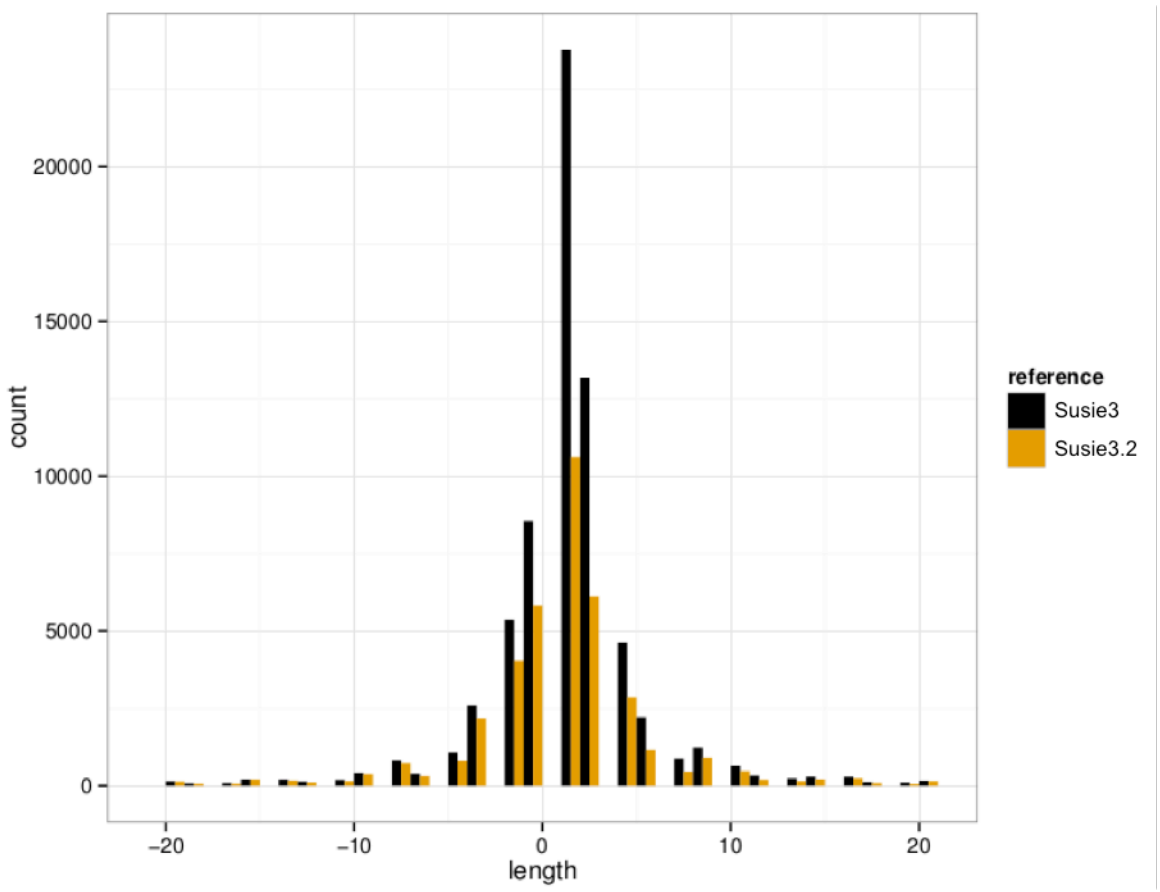

Figure S5. Distribution of insertion and deletion lengths for GENCODE (v23 Basic) transcripts aligned to Susie3 references with BLAT before and after error correction. Indel lengths with multiples of three are not shown. Susie3 is the uncorrected reference. Susie3.2 is corrected by SNPs and indels either from Susie and four or more gorillas or all gorillas except Susie.

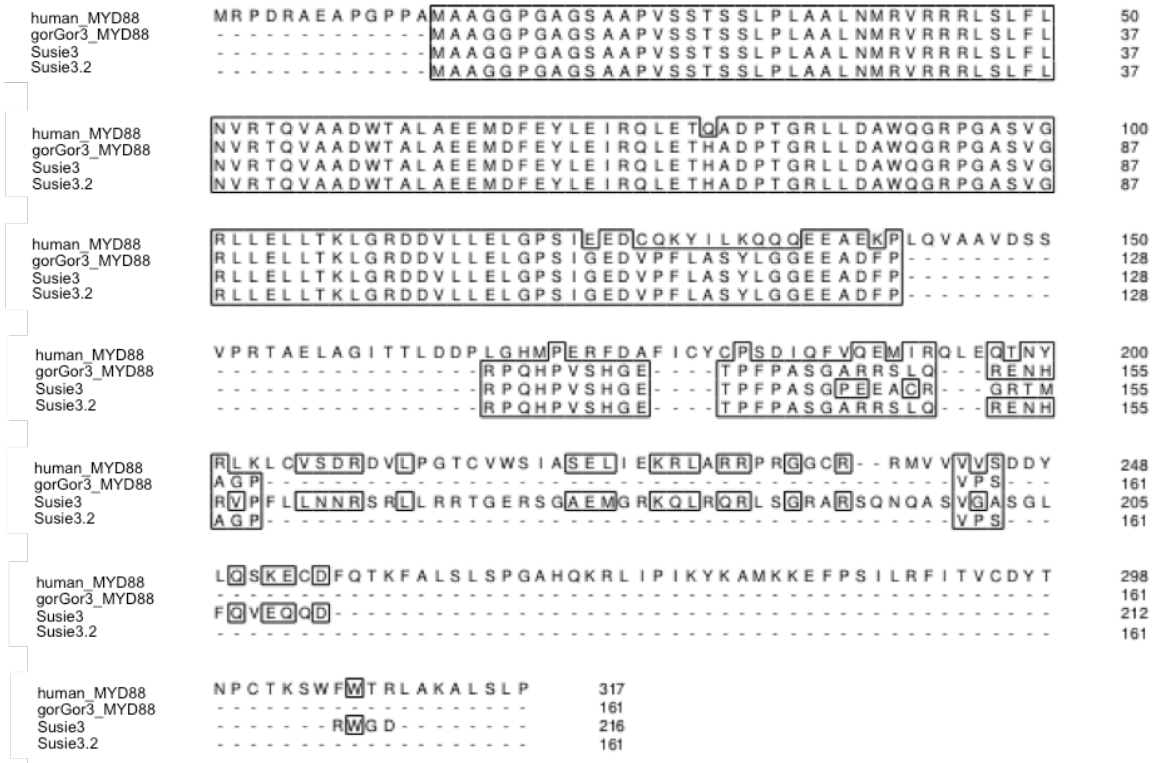

Figure S6. Multiple sequence alignment of MYD88 (human, gorGor3, Susie3, and an error-corrected Susie). A number of false missense mutations were resolved.

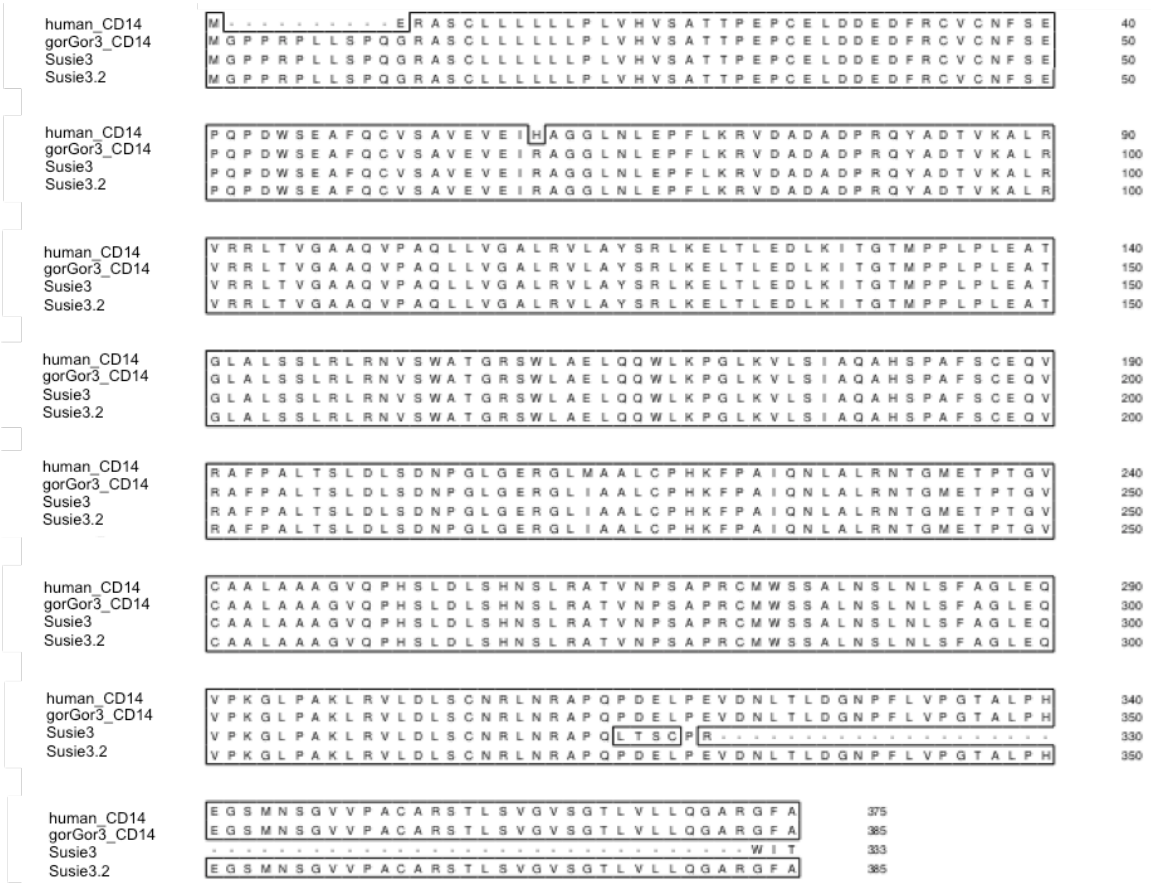

Figure S7. Multiple sequence alignment of CD14 (gorGor3, Susie3, and an error-corrected Susie3). An erroneous nonsense mutation was resolved.

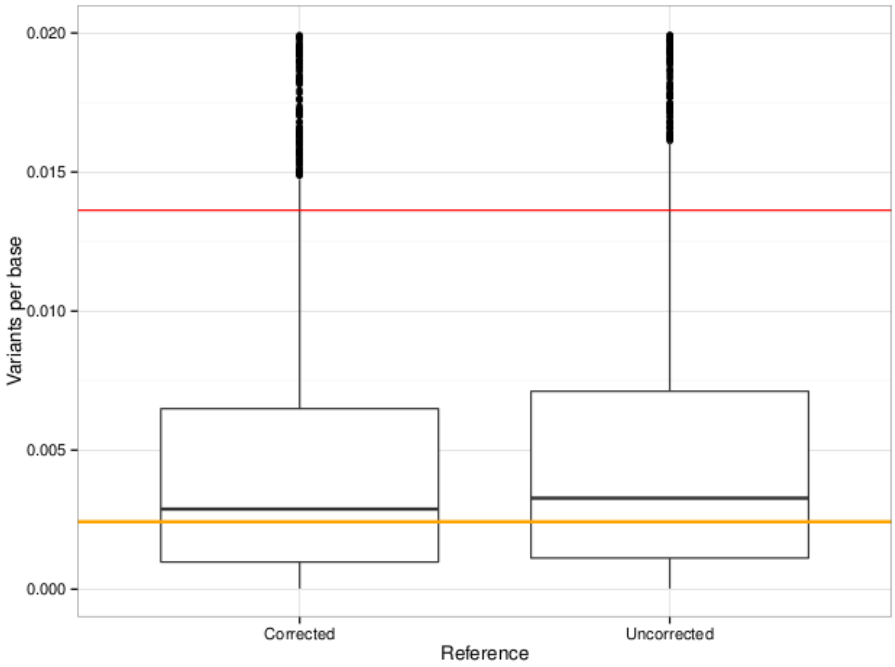

Figure S8. Distribution of variants per base across all contigs assessed for SNPs and indels shown in Susie3 before and after error correction. The range of expected variants per base from (*10*) for the same samples is shown in orange (y = [0.002362, 0.002486]). The threshold for excess variants per base is shown in red (y = 0.012).

Figure S9. Genomic distribution of 200 Susie3 contigs with excess variants per base (> mean + 1 SD) as projected onto GRCh38. Not shown are 112 contigs that have no alignment to GRCh38. The majority of the contigs shown (68%) map within human- or gorilla-specific segmental duplications or heterochromatic regions including telomeres and centromeres.

Figure S10. Susie3 scaffolds projected against the human genome (GRCh38). Scaffolds that are aligned in the same GRCh38 chromosome with the same order and orientation are black. Scaffolds colored gray contain contigs aligned to the same GRCh38 chromosome, but in different order and/or orientation. Unscaffolded contigs are colored in blue, and gorilla and human segmental duplications >100 kbp are colored in green.

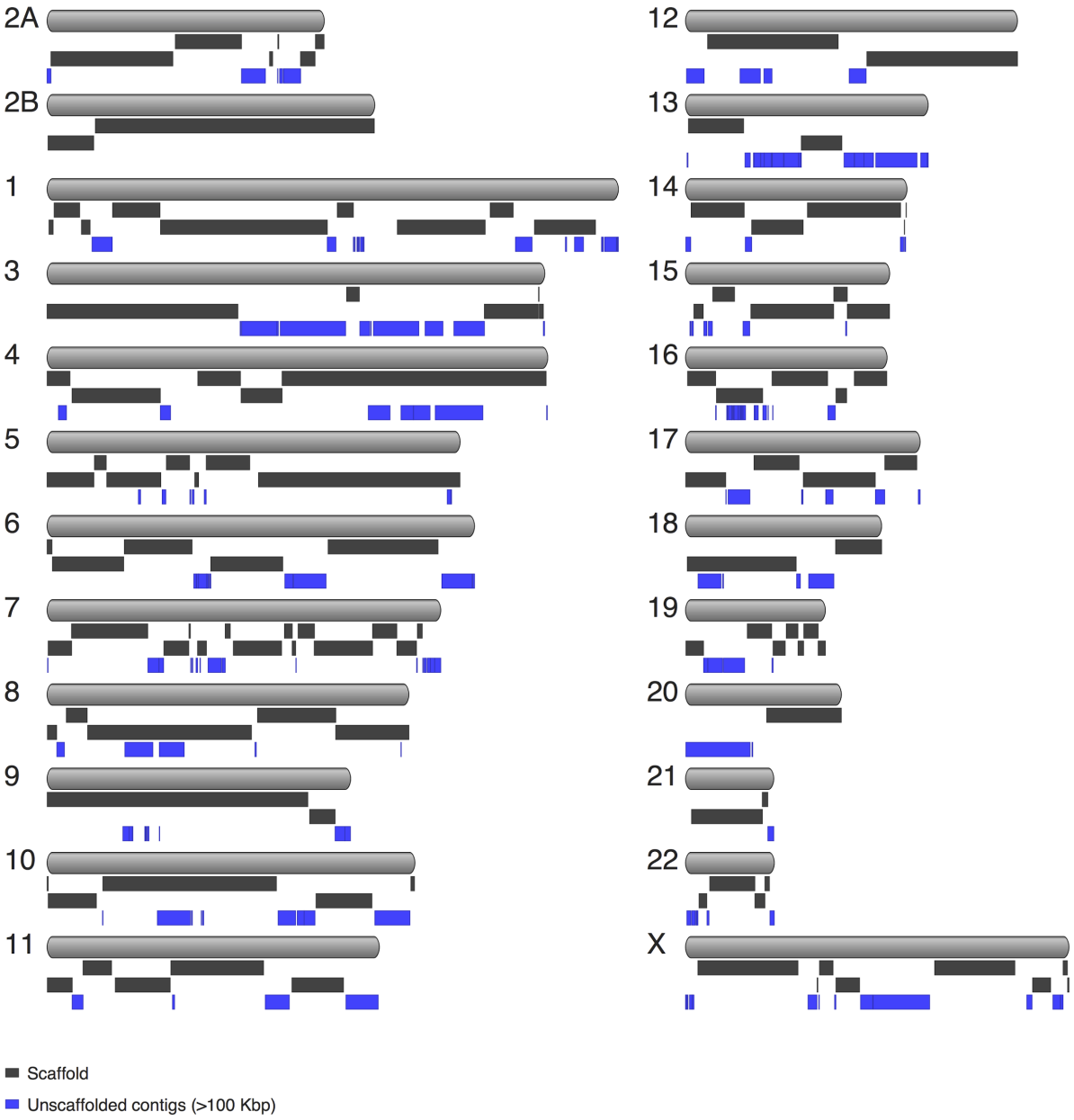

Figure S11. Scaffolds (N50 = 23.1 Mbp) mapped to gorilla chromosomes. The first two rows of black rectangles represent scaffolds, and the blue rectangles correspond to unscaffolded contigs (>100 kbp).

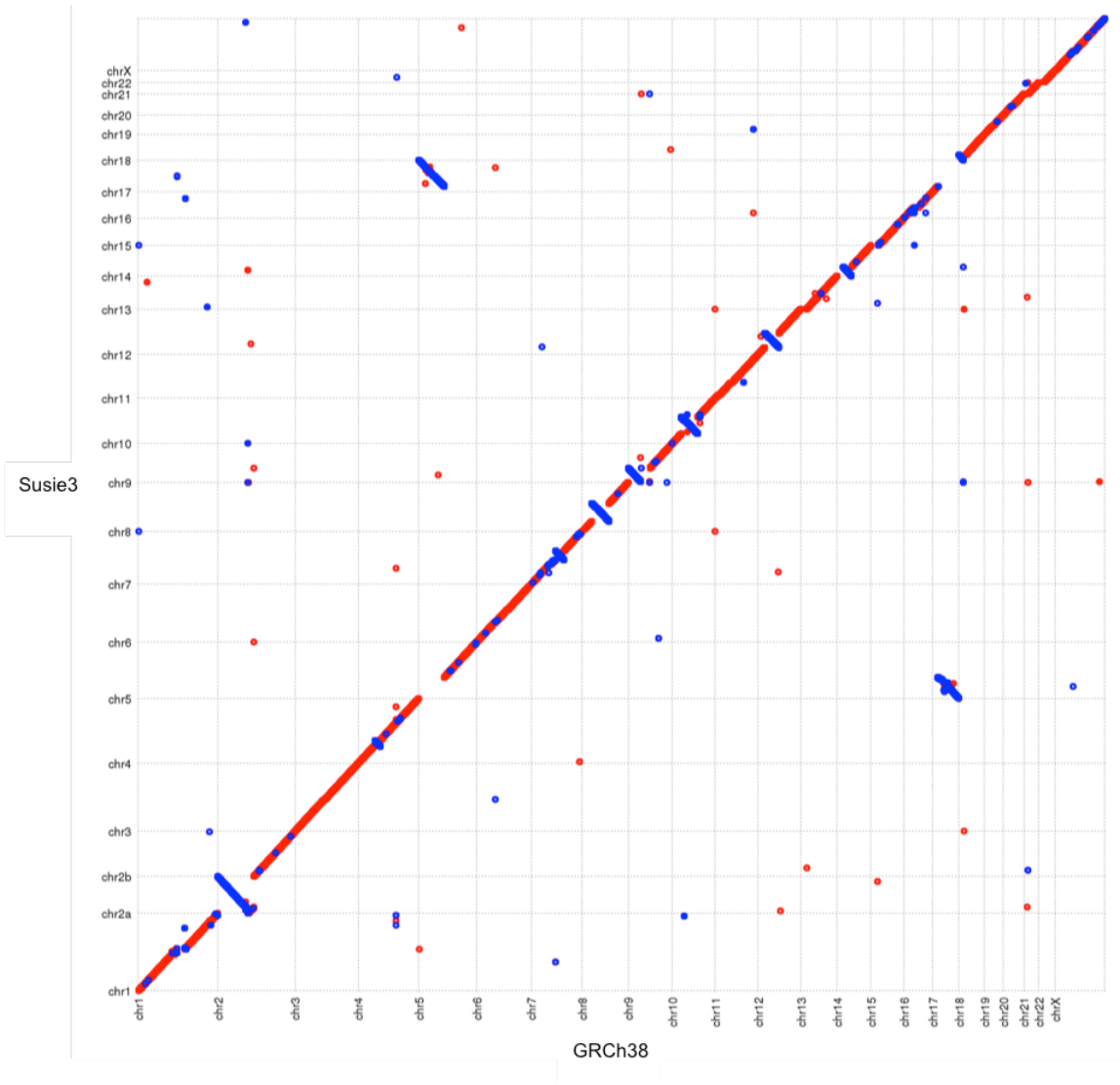

Figure S12. Large-scale differences between human and gorilla. Susie3 AGP aligned against GRCh38. Alignments were generated using NUCmer –mumref –l 100 –c 1000.

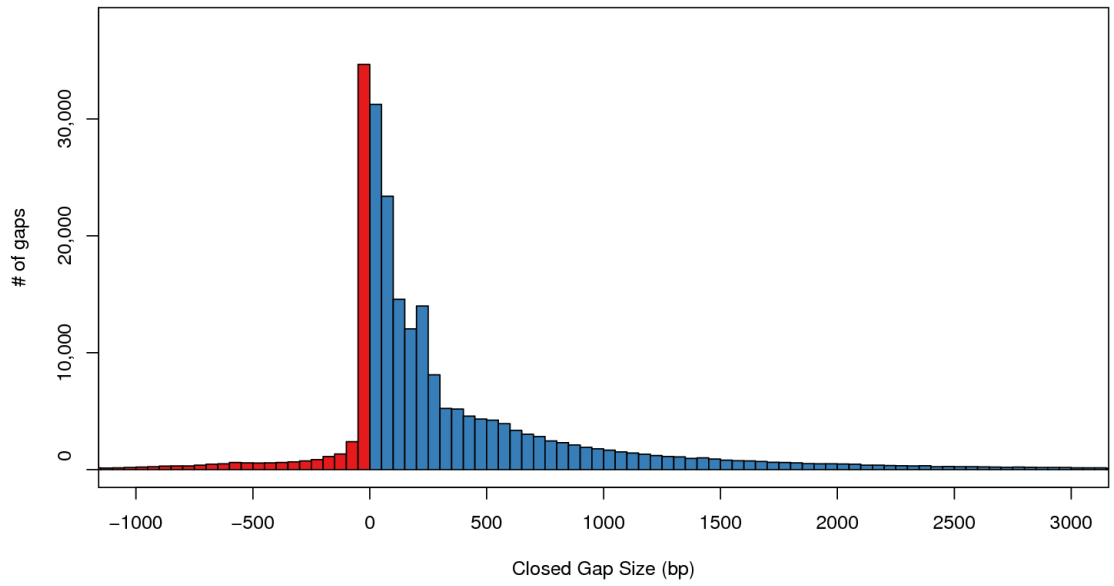

Figure S13A. Length distribution (bin size 50 bp) of gaps in gorGor3 closed by Susie3. True gaps (blue, gap size >= 0 bp) are those where sequence is missing in gorGor3 but present in Susie3. False gaps (red, size < 0 bp) are those where sequence is not missing from gorGor3 even though it is annotated as such. Gap sizes are measured as follows: 2,000 bp of flanking sequence each side of the putative gorGor3 gap are mapped to the spanning Susie3 contig. The gap size is the distance between these two mapped locations in Susie3. False gap sizes mean that these two flanking sequences map in the opposite order between Susie3 and gorGor3 and that the two sides of the gap in gorGor3 actually overlap each other. We restricted this analysis to gaps where the two flanking sequences map to the same strand of the same Susie3 contig and that contig aligns to the gorGor3 chromosome and position of the putative gap.

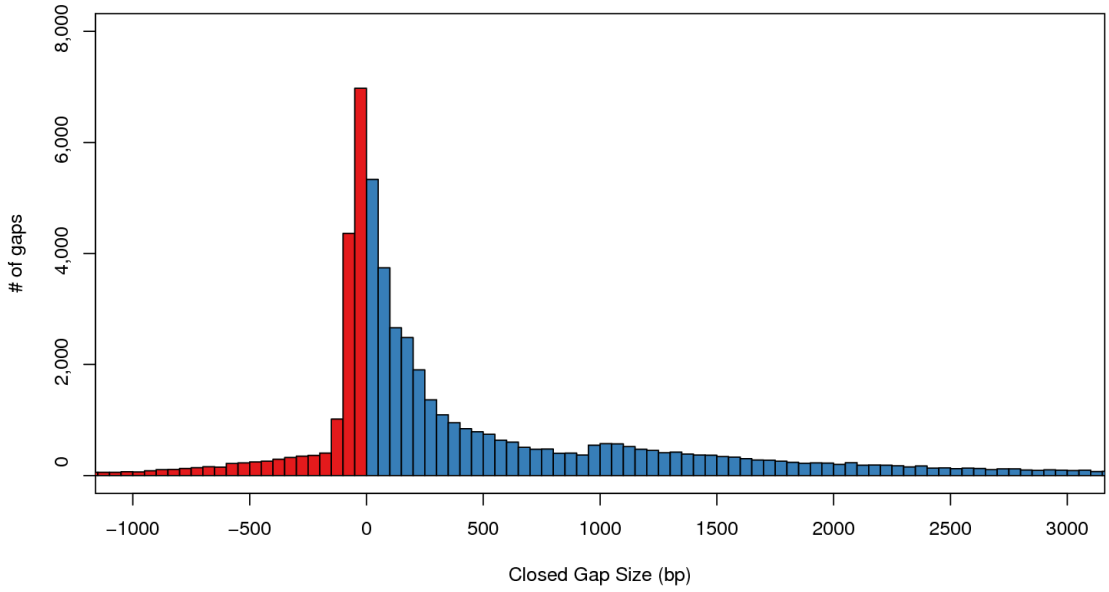

Figure S13B. Gaps closed in Susie3 (compared to gorGor4). See fig. S13A legend for description. 68% (123,818) of all gorGor4gaps (181,717) were closed based on comparison to Susie3. gorGor4 has been released (NCBI) under accession GCA_000151905.3, but the methods used to improve this version of the assembly are unknown.

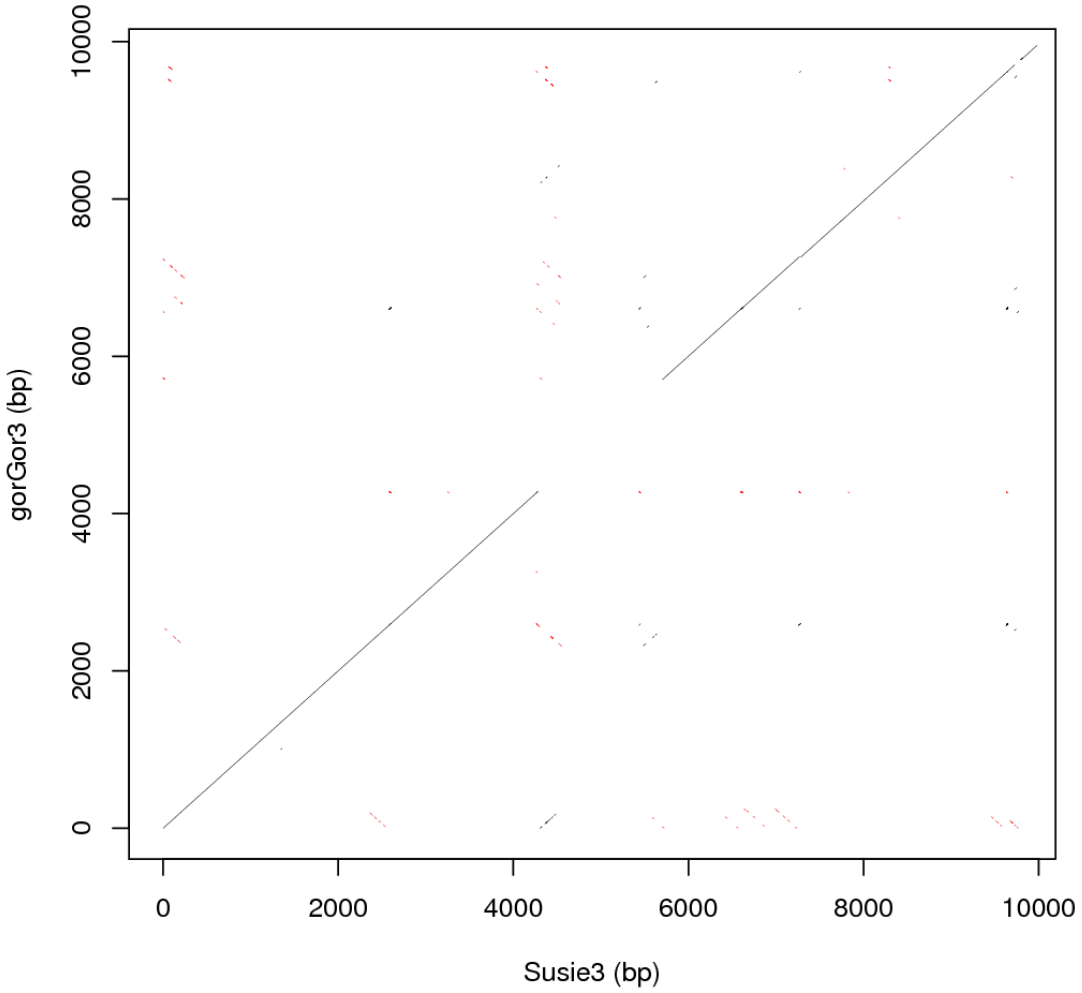

Figure S14A. A gap in gorGor3 (closed by Susie3) is indicated by the break in the line. gorGor3's gap was 1,426 bp while Susie3 showed that there were 1,424 bp of sequence at this location. 58% of "true gaps" (see fig. 13A legend) (gaps in which sequence is missing from gorGor3) of actual size 1,000 bp to 2,000 bp look like this.

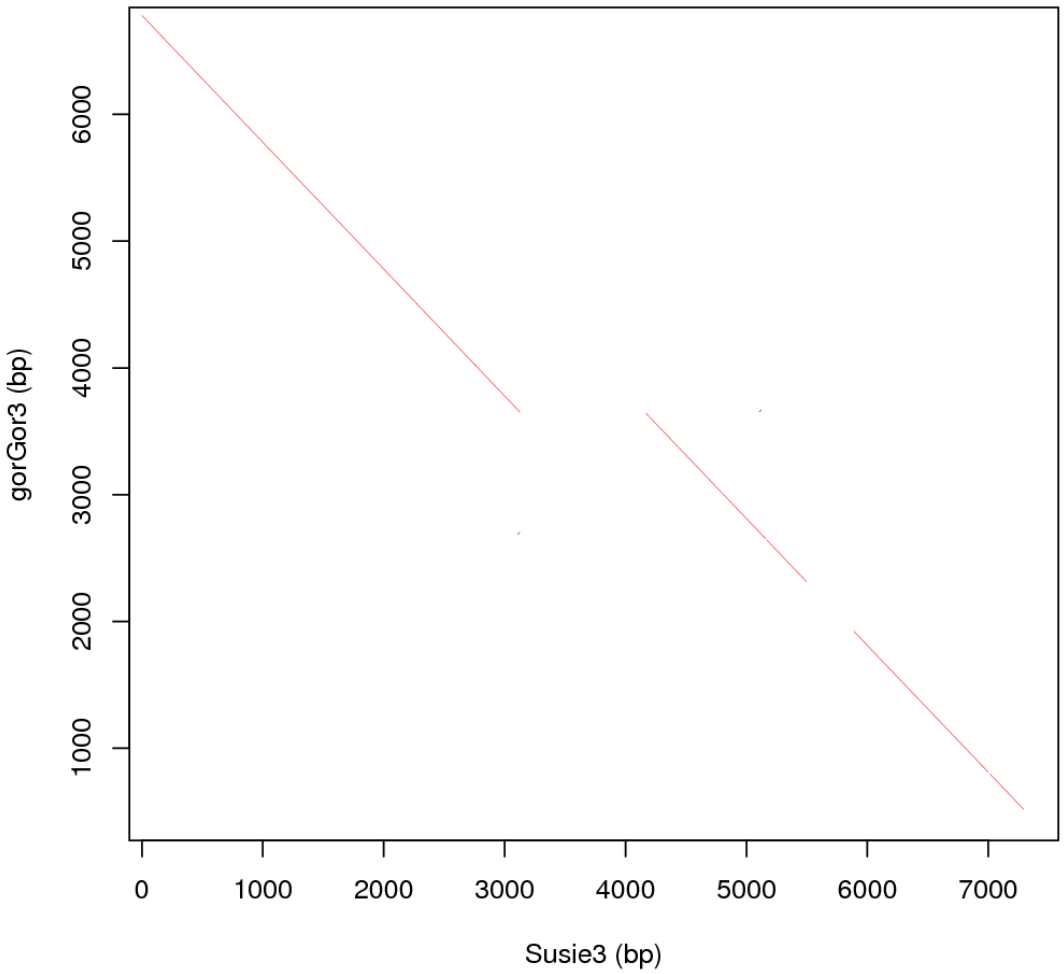

Figure S14B. Two gaps in gorGor3 (closed by Susie3) are indicated by breaks in the lines. In the large gap near the middle of the figure, gorGor3 has gap size 10, which is much smaller than the amount of Susie3 sequence (1,042 bp in this example) so the lines are offset. 22% of true gaps of actual size 1,000 bp to 2,000 bp have plots in which the lines are offset like this.

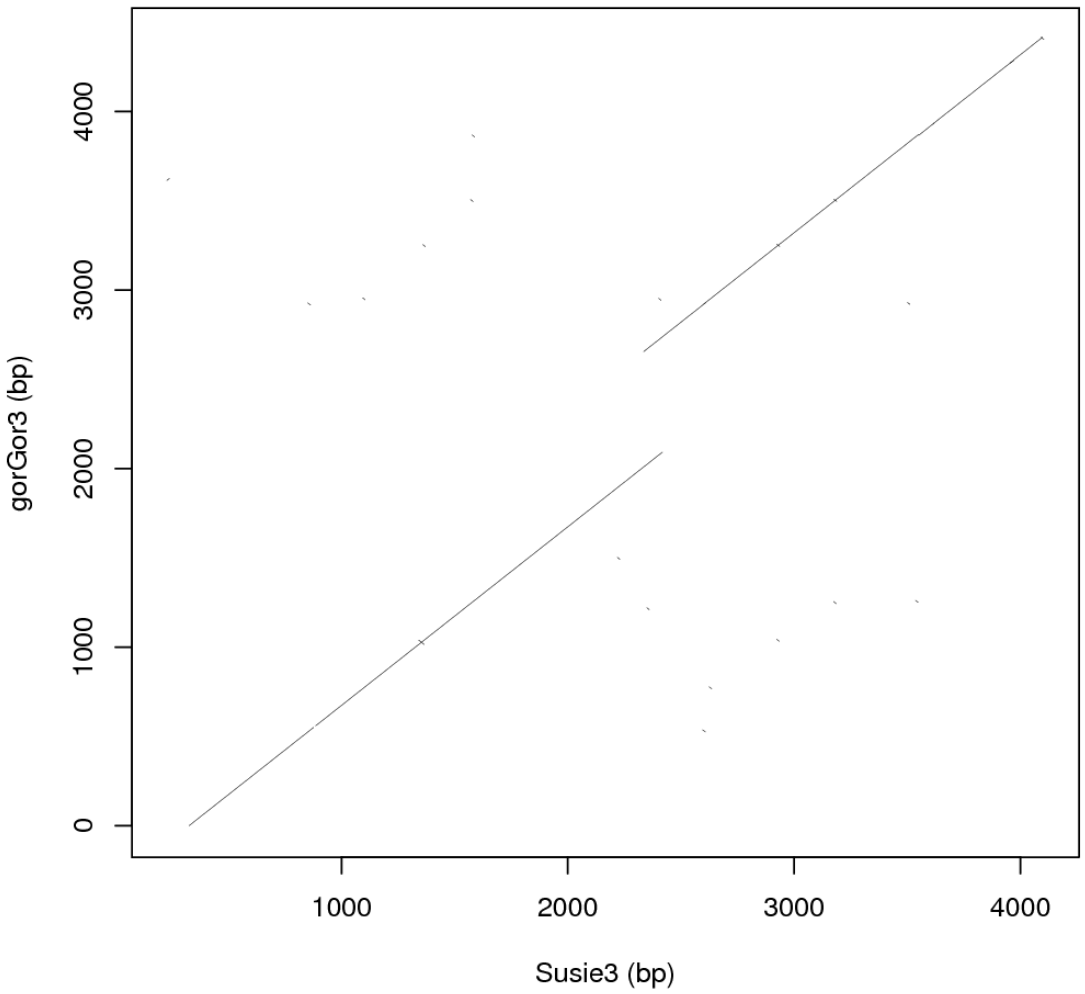

Figure S14C. One type of false gap where Ns (unknown nucleotides) and additional sequence are present in gorGor3, but not in Susie3. The sequence flanking the gap is not duplicated as in fig. S14D. About 21% of false gaps have sequence structures similar to this.

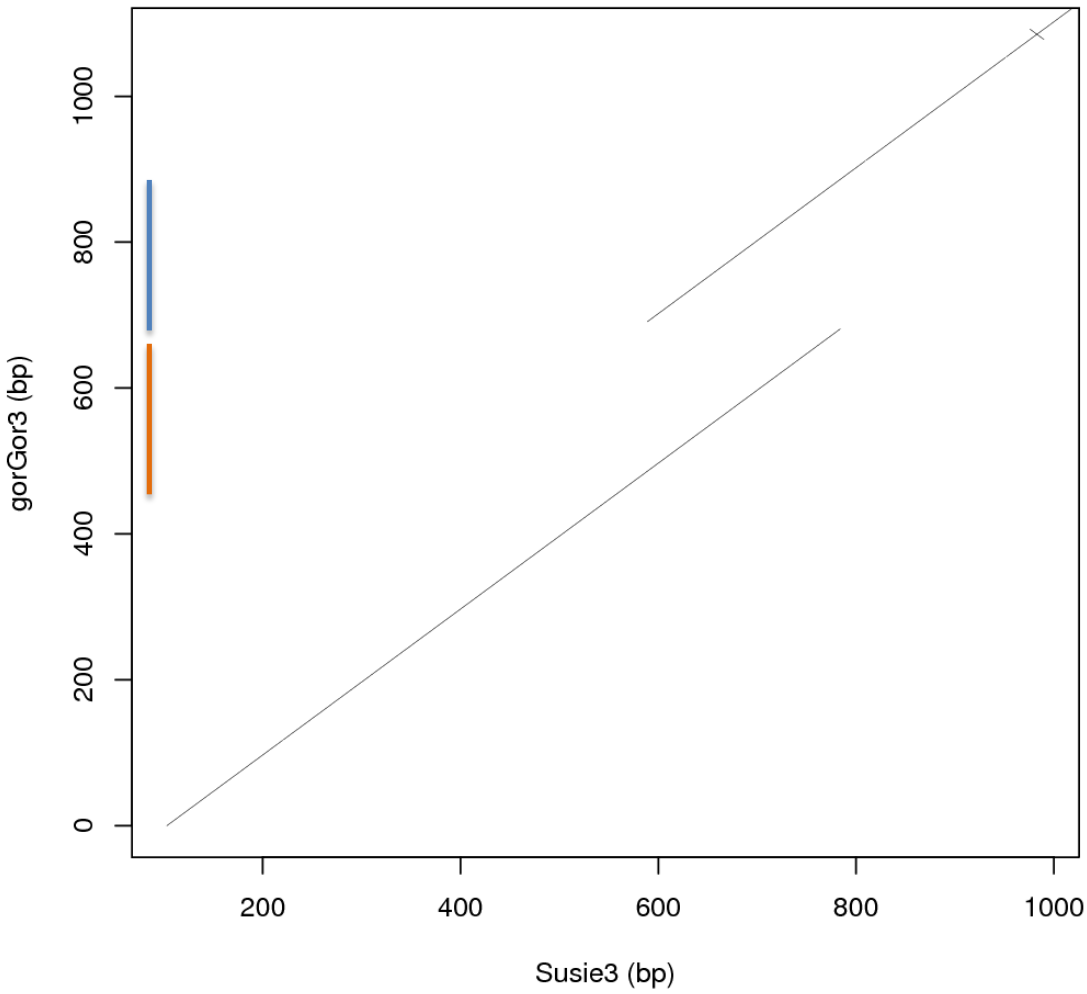

Figure S14D. An example of a false gap. About 58% of false gaps have similar plots indicating that a sequence in Susie3 (horizontal axis) has been artifactually duplicated in gorGor3 (the blue and orange vertical lines). Note gorGor3 had a gap size of 10 in this particular example, but in fact there is no missing sequence based on Susie3.

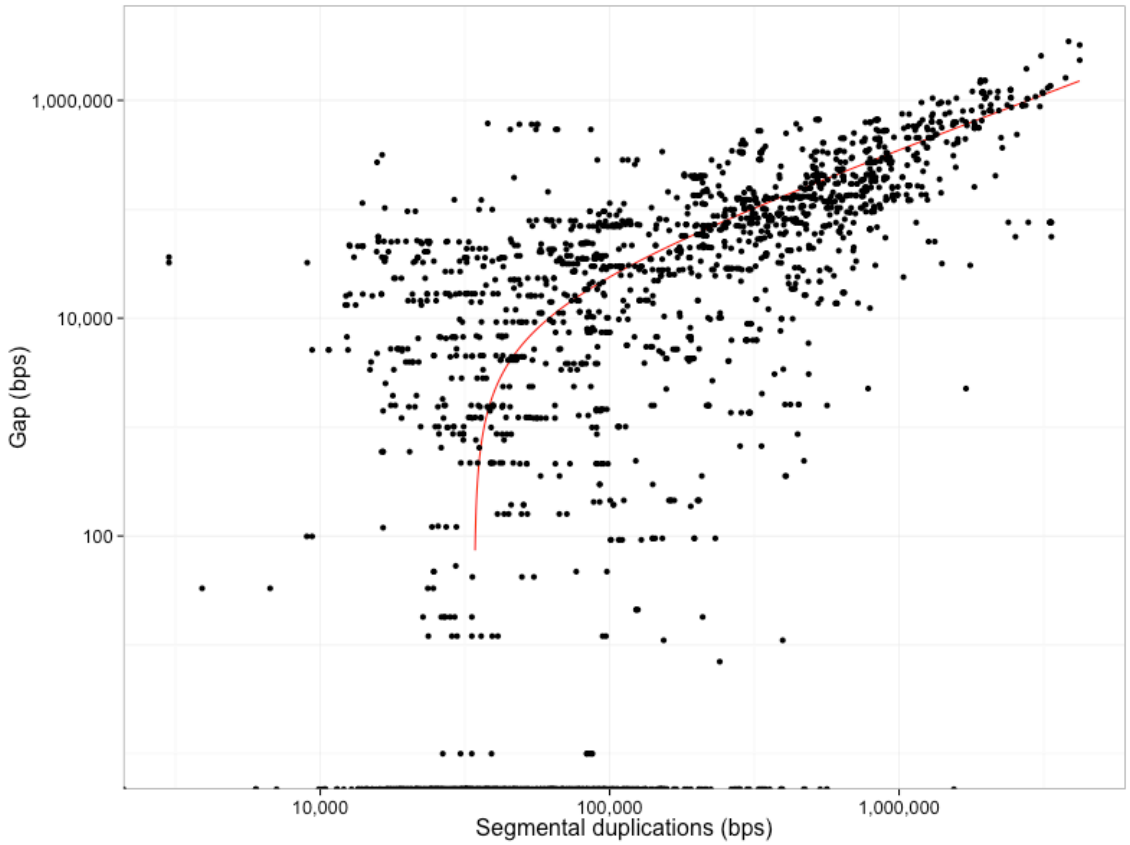

Figure S15. Human and gorilla segmental duplication composition for open gaps in Susie3 (inferred by aligning Susie3 contigs to GRCh38). Gaps and segmental duplication sequence were computed over a 10 Mbp window sliding in 1 Mbp increments.

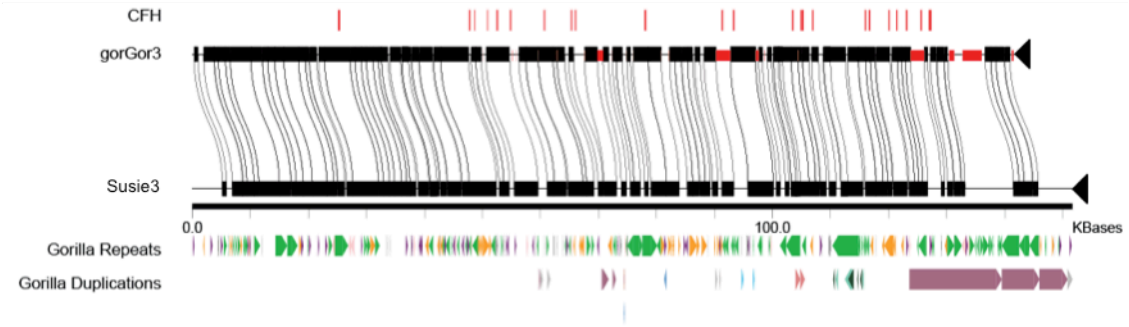

Figure S16. *CFH* resolved in Susie3 by closing gaps in gorGor3. Gene annotations are lifted over from GRCh38 to gorGor3. Red bars on gorGor3 sequence indicate gaps in the assembly. Alignments between gorGor3 and Susie3 are based on Miropeats (*31*). Exons for the gene *CFH* are shown above the gorGor3 sequence in red ticks.

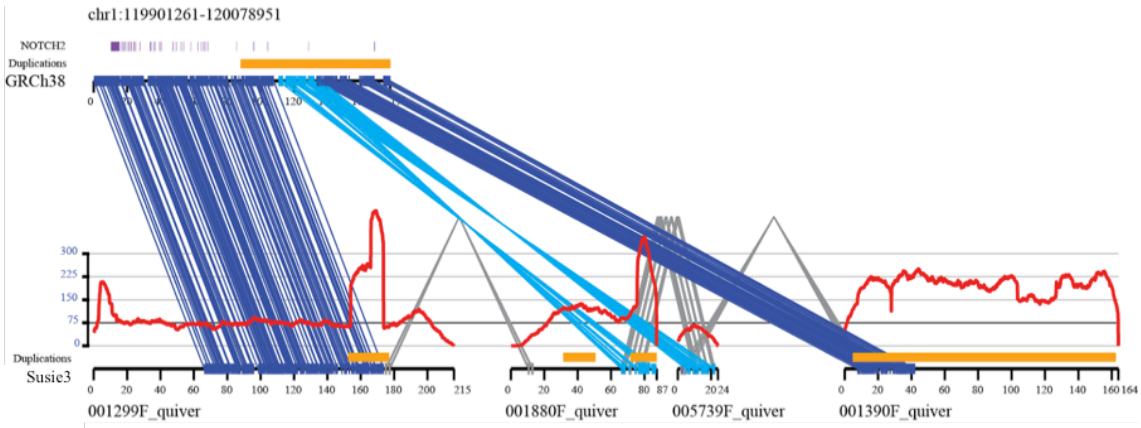

Figure S17. Misassembly of NOTCH2 in Susie3 compared to GRCh38 based on Miropeats alignment (*31*)**.** GRCh38 sequence is shown at top with NOTCH2 exons annotated in purple and segmental duplications annotated in orange. Dark blue lines between GRCh38 and Susie3 contigs indicate alignment of colinear sequence while light blue lines indicate inverted sequence in Susie3 relative to the human reference. Inter-contig alignments in Susie3 are shown in gray. Read depth of SMRT sequence from Susie aligned to Susie3 is shown in red with mean depth of 75 indicated by a black horizontal line. Regions flagged in Susie3 as duplicated based on excess read depth are shown in orange.

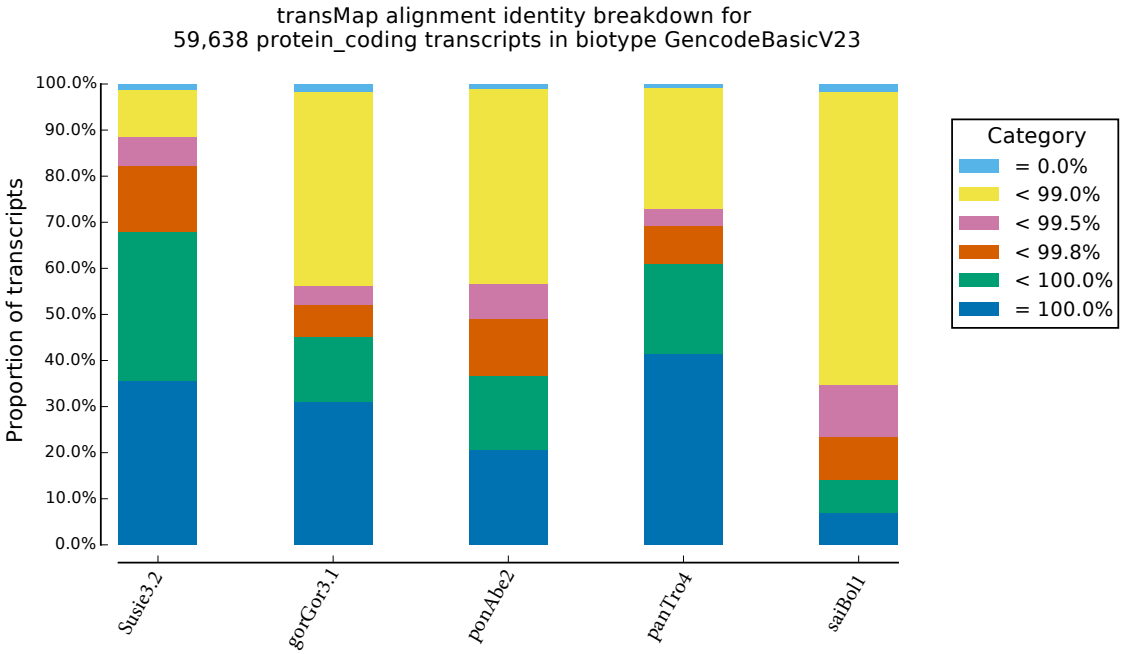

Figure S18. Proportion of GENCODE transcripts aligned at different identity thresholds, including 100% (dark blue), <100% and >=99.8% (green), <99.8% and >=99.5% (orange), <99.5% and >=99.0% (pink), <99.0% (yellow), or not aligned at all (light blue) to Susie3.2 (gorilla) and four reference assemblies including gorGor3 (gorilla), ponAbe2 (orangutan), panTro4 (chimpanzee), and saiBol1 (squirrel monkey).

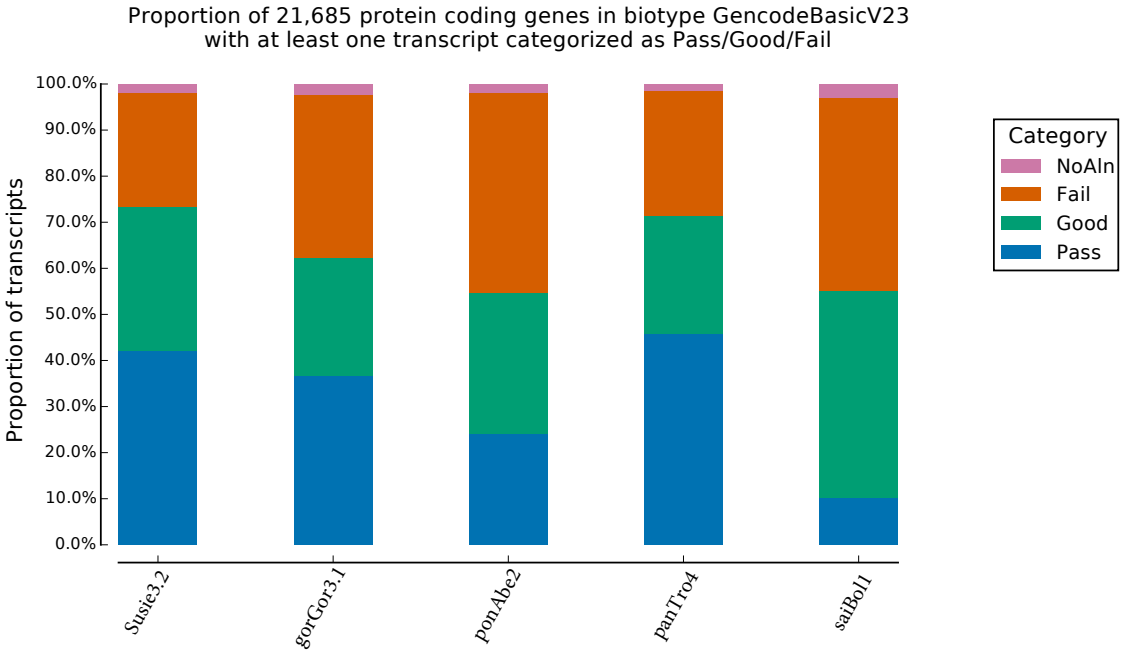

Figure S19. Proportion of genes whose transcripts passed binary classifiers (good or pass), failed, or did not have a TransMap alignment based on initial alignments to the gorilla assembly, Susie3.2, and four reference assemblies including gorGor3 (gorilla), ponAbe2 (orangutan), panTro4 (chimpanzee), and saiBol1 (squirrel monkey).

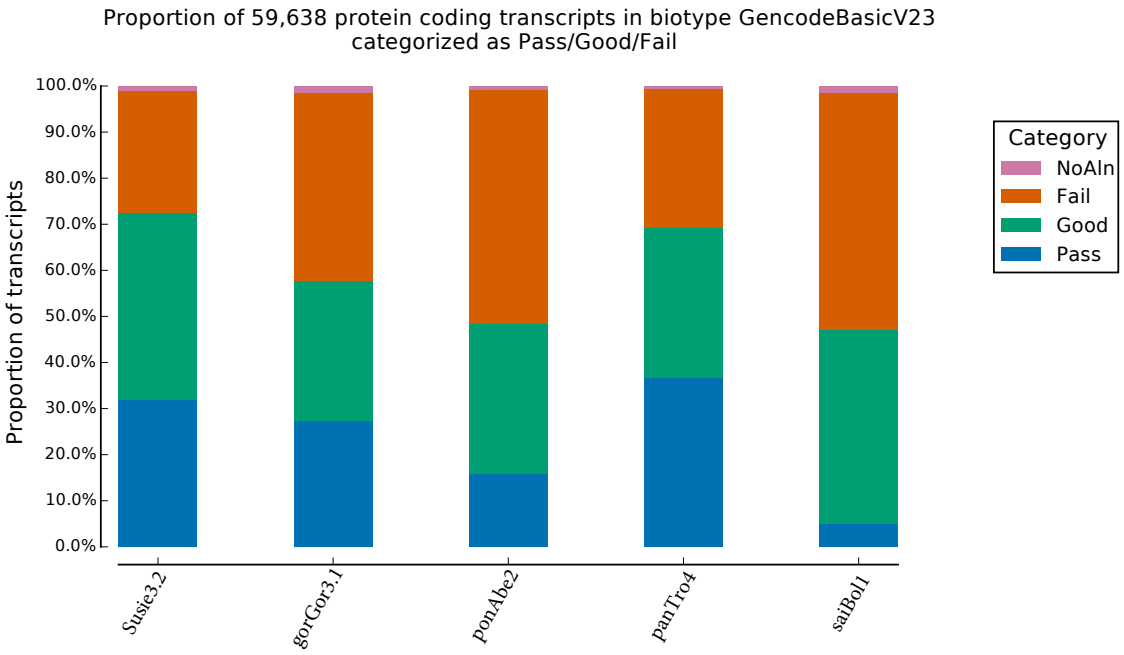

Figure S20. Proportion of transcripts that passed binary classifiers (good or pass), failed, or did not have a TransMap alignment based on initial alignments to the gorilla assembly, Susie3.2, and four reference assemblies including gorGor3 (gorilla), ponAbe2 (orangutan), panTro4 (chimpanzee), and saiBol1 (squirrel monkey).

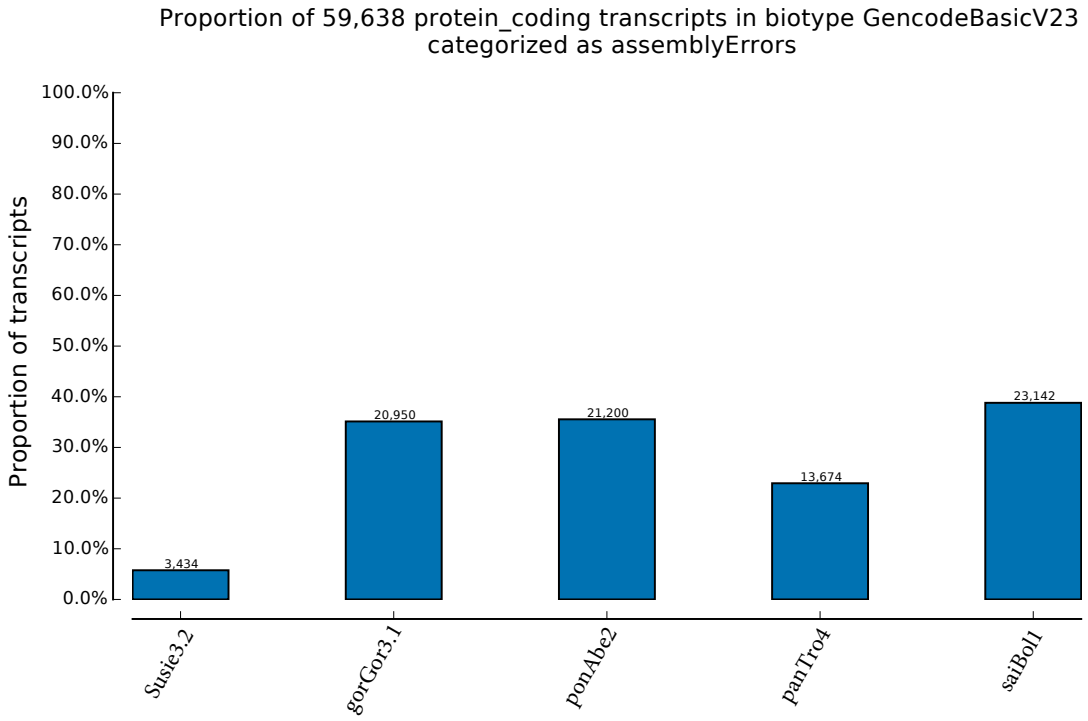

Figure S21. Proportion of transcripts with assembly errors when aligned to gorilla assembly Susie3.2 and four reference assemblies including gorGor3 (gorilla), ponAbe2 (orangutan), panTro4 (chimpanzee), and saiBol1 (squirrel monkey).

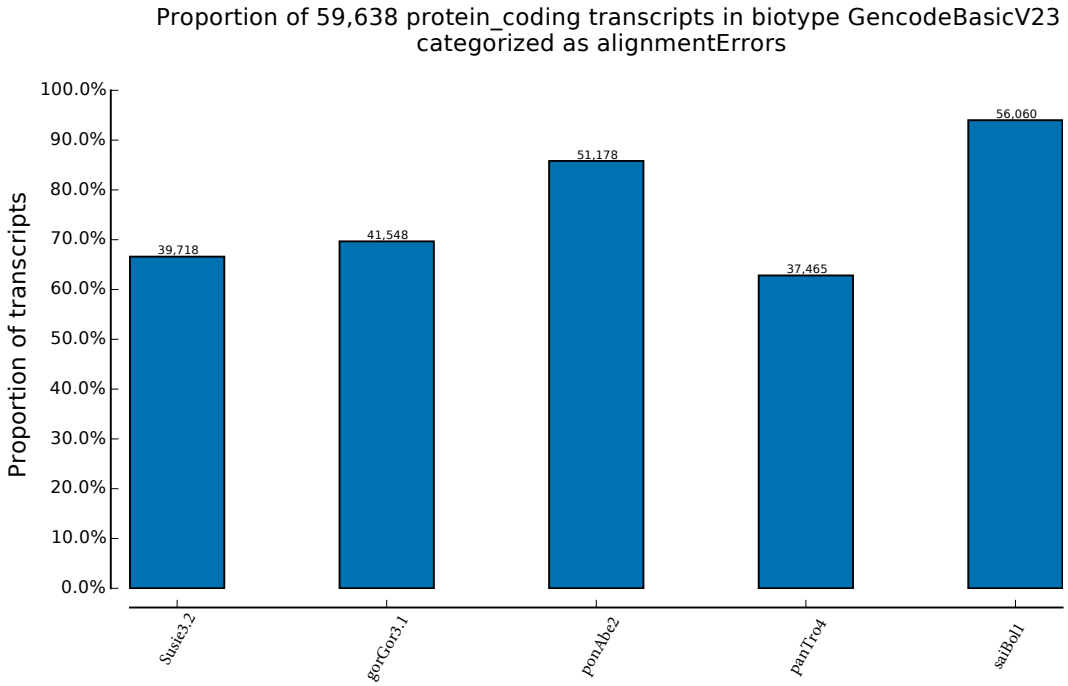

Figure S22. Proportion of transcripts with alignment errors when aligned to gorilla assembly Susie3.1 and four reference assemblies including gorGor3 (gorilla), ponAbe2 (orangutan), panTro4 (chimpanzee), and saiBol1 (squirrel monkey).

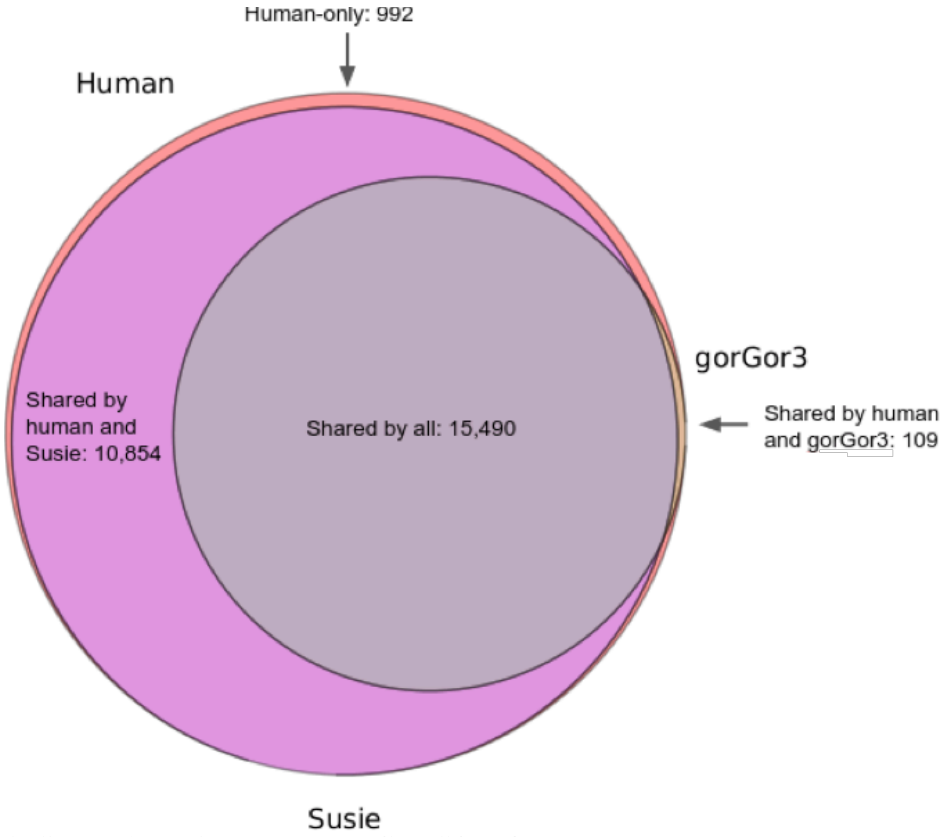

Figure S23. Comparison of human genes (GENCODE/Ensembl sourced) annotated to the human reference assembly (GRCh38) and the two gorilla assemblies (Susie and gorGor3).

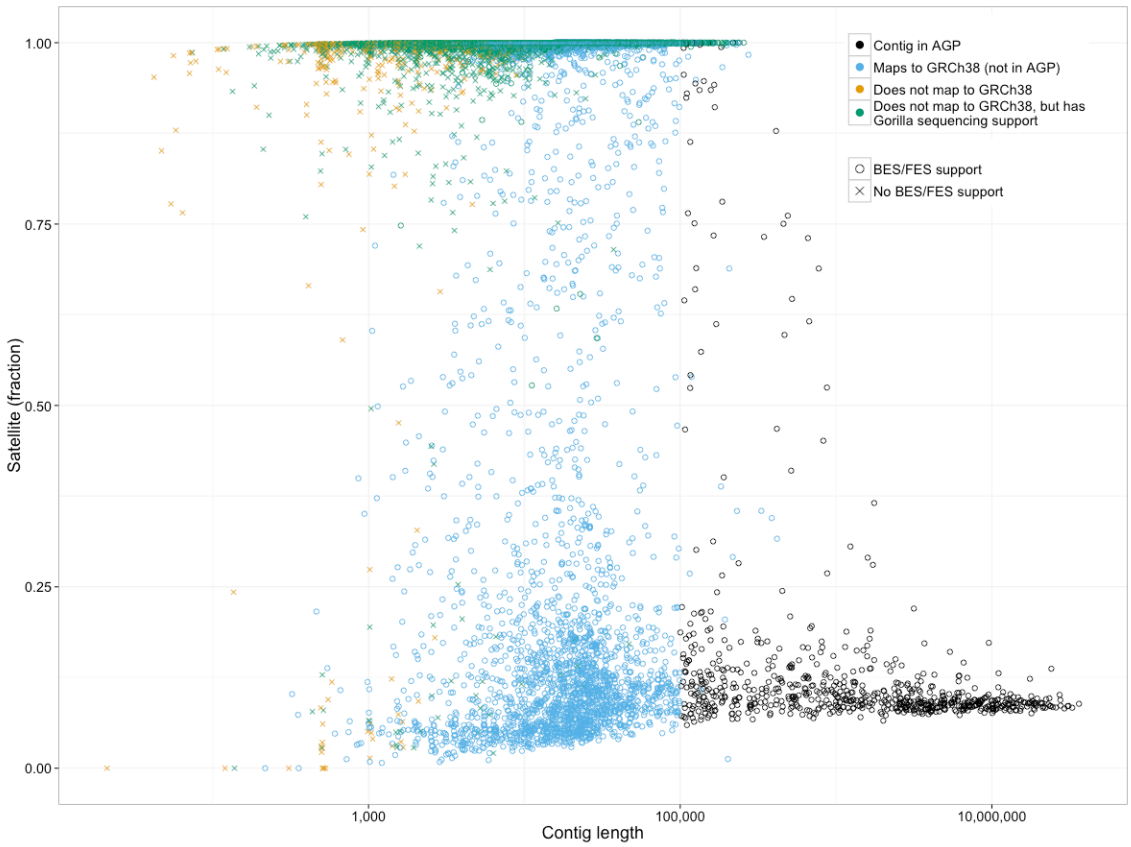

Figure S24. Satellite content of mapped and unmapped Susie3 contigs. Satellite sequence was marked using RepeatMasker and Tandem Repeats Finder. Contigs were aligned to GRCh38 using BLASR (whole-genome alignment option). Contigs colored in black constitute our AGP. Contigs not found in our AGP, but still map to GRCh38, are colored blue. The remaining unmappable contigs (colored orange and green) contain a high fraction of satellite sequence and represent heterochromatic sequences not properly assembled.

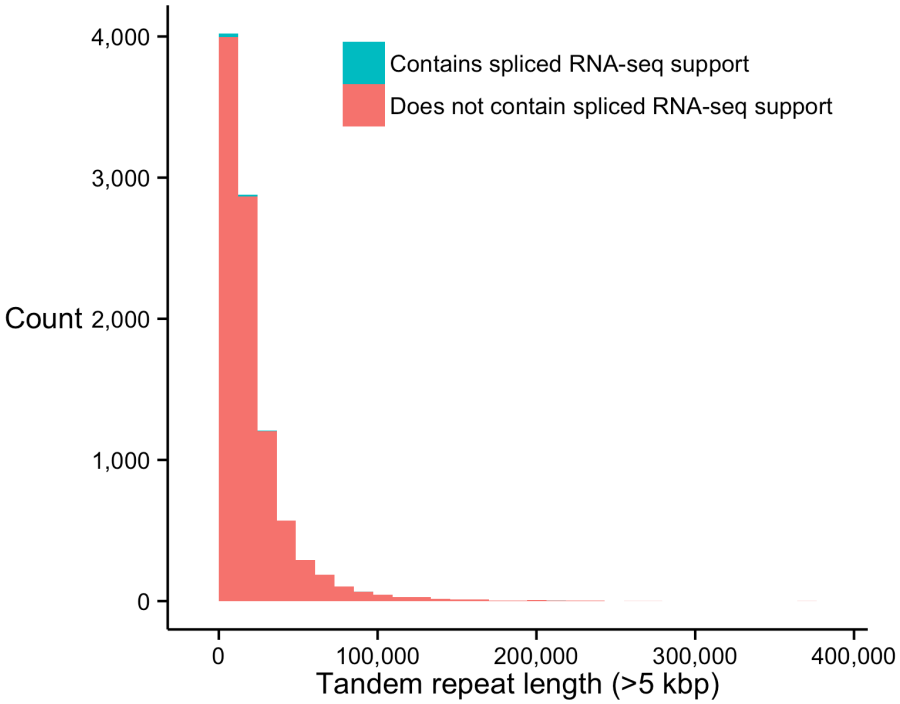

Figure S25. Length distribution of tandem repeats greater than 5 kbp. RNA-seq data map to a small fraction of these identifying those that are potentially transcribed.

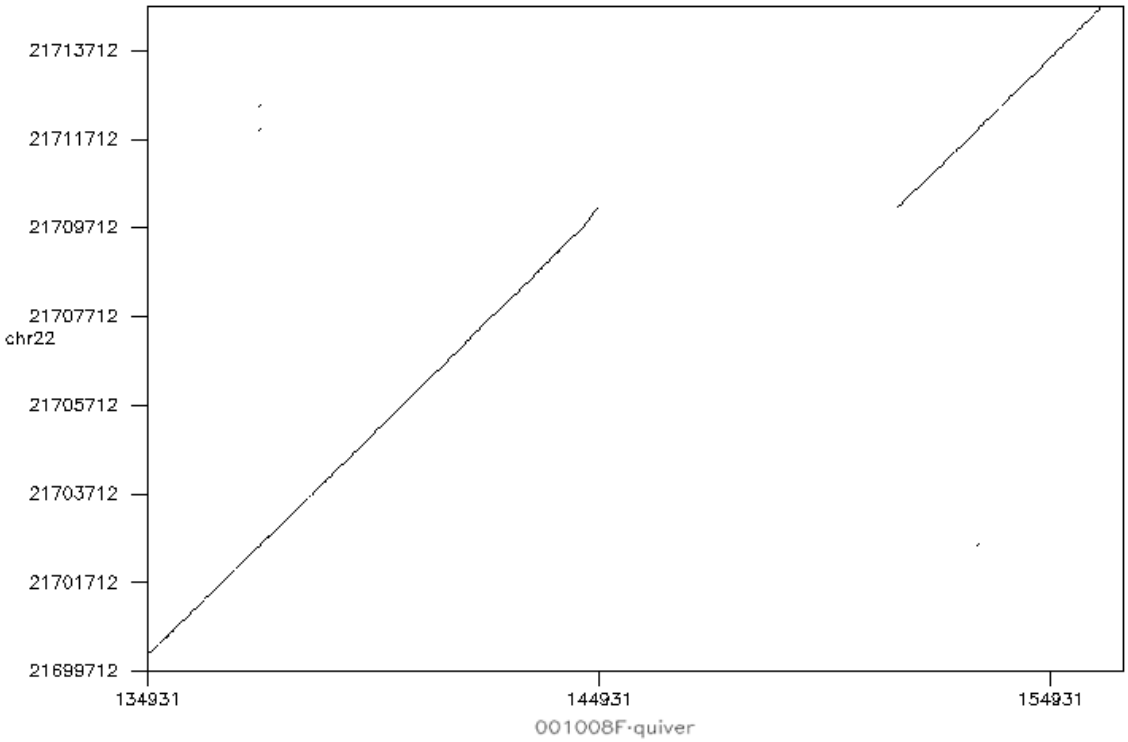

Figure S26. Novel macrosatellite found in Susie3 (001008F_quiver) that is not found in GRCh38. The location of the macrosatellite intersects with *YPEL1*.

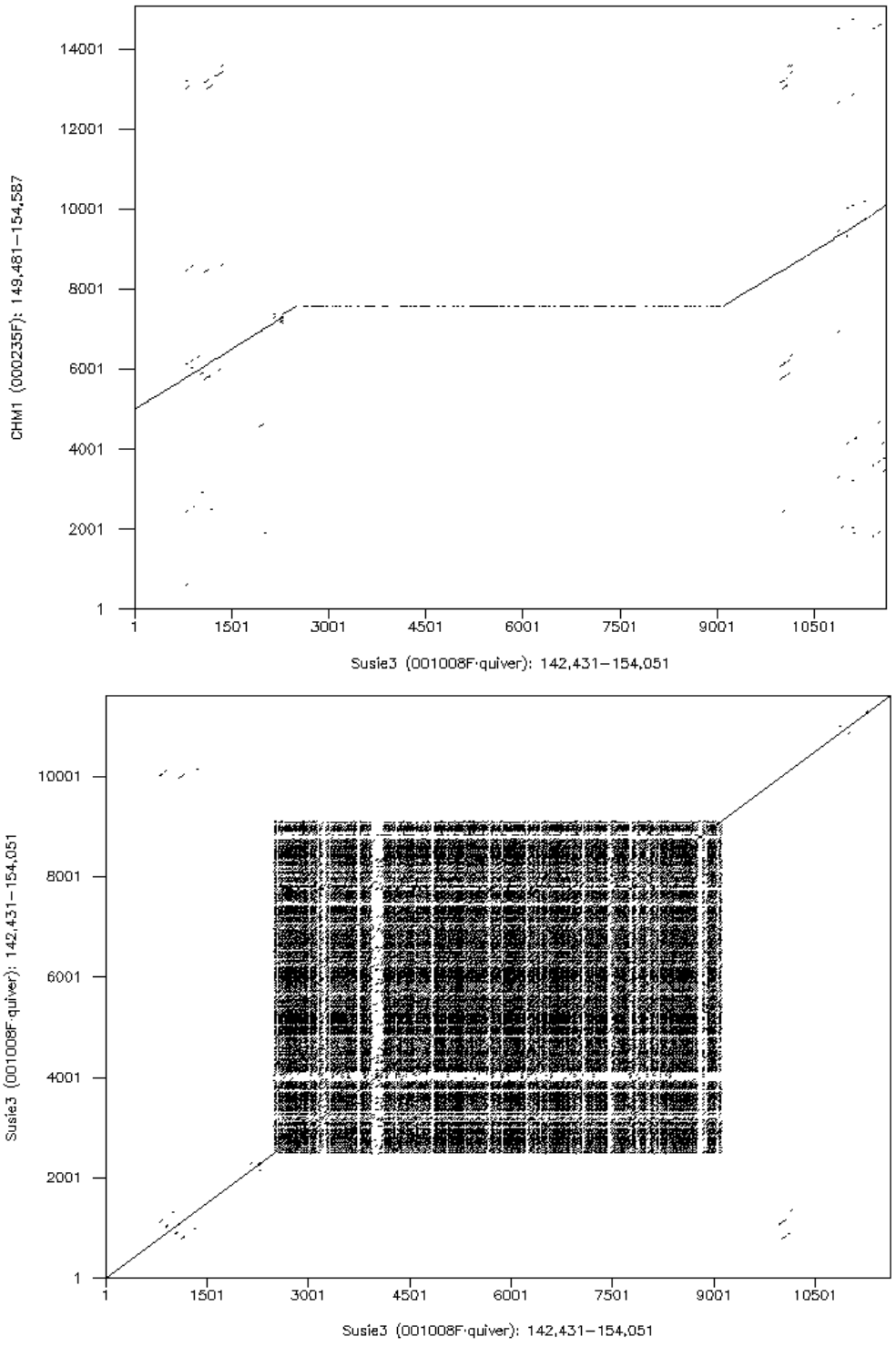

Figure S27. Novel macrosatellite found in Susie3 (001008F_quiver) that is not found in CHM1 (top). The location of the macrosatellite intersects with *YPEL1*. Susie3 aligned to itself to show the repeat structure of the macrosatellite (bottom).

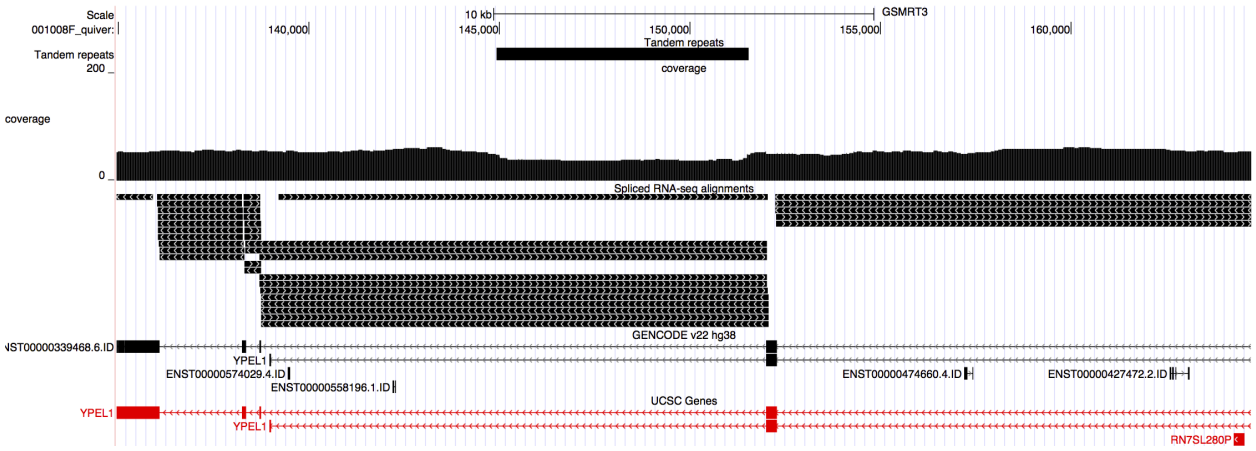

Figure S28. A genome browser shot showing the position of macrosatellite in relation to *YPEL1*.

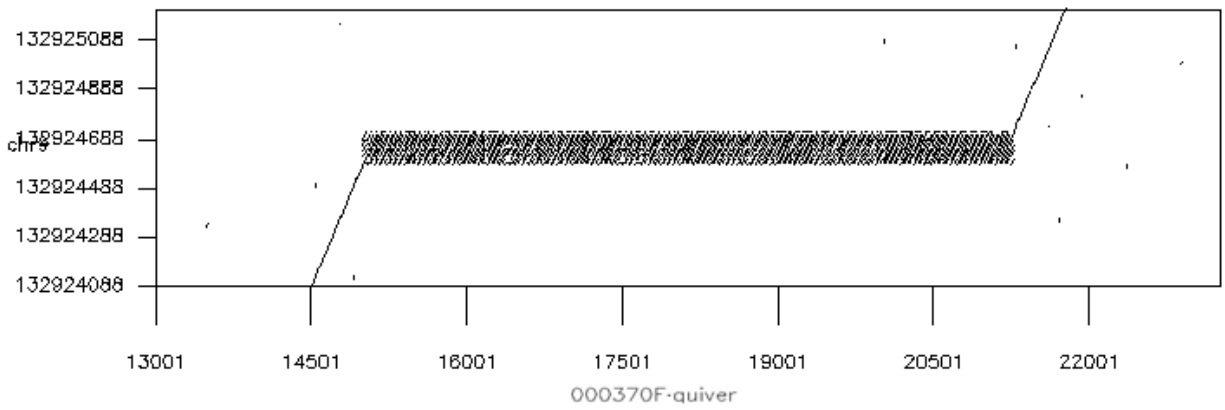

Figure S29. An expanded macrosatellite in Susie3 (00370F_quiver) found in GRCh38 chr9.

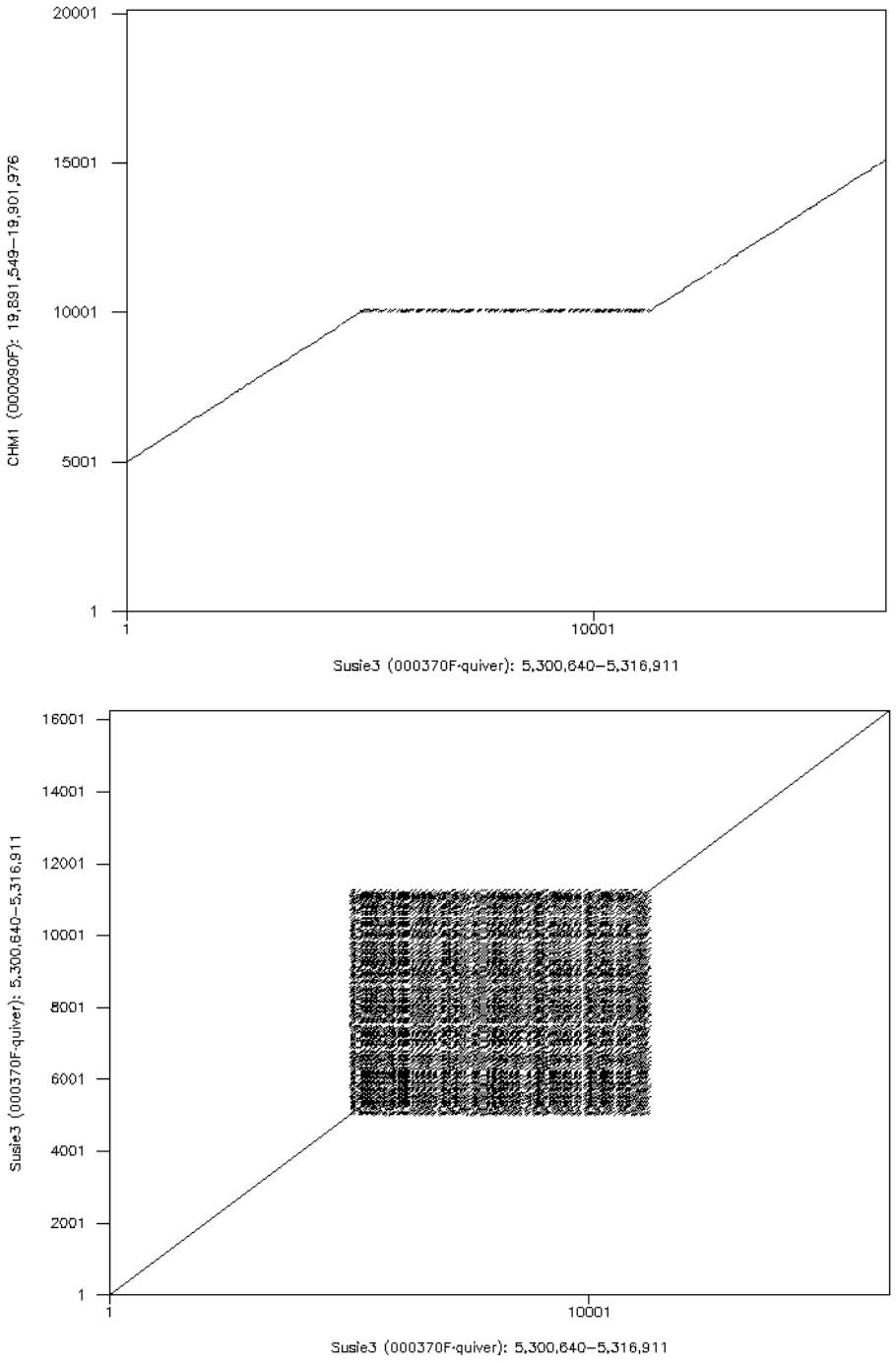

Figure S30. An alignment between Susie3 and CHM1 showing an expanded macrosatellite found in Susie3 (00370F_quiver) (top). Susie3 aligned to itself to show the repeat structure of the macrosatellite (bottom).

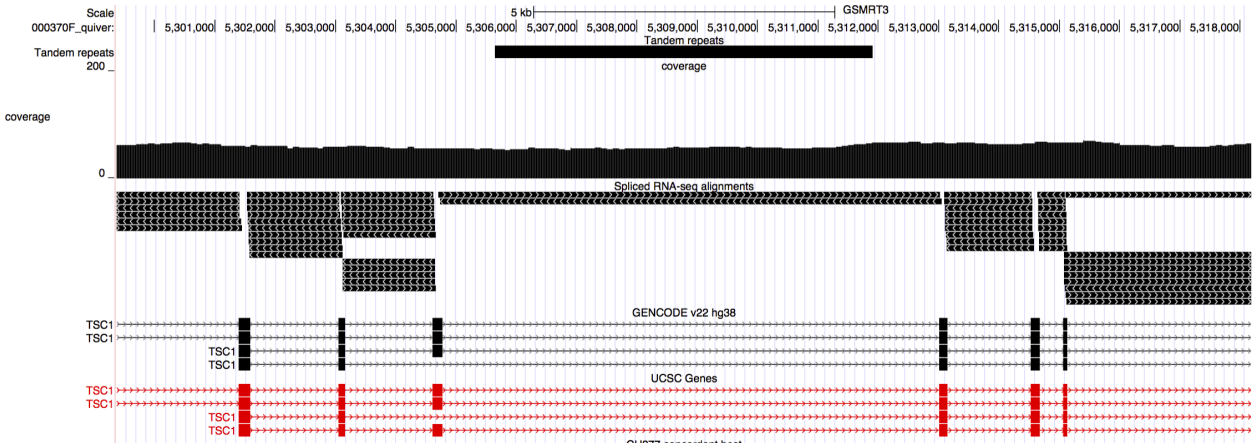

Figure S31. The expanded microsatellite in Susie3 occurs within the introns of gene *TSC1*.

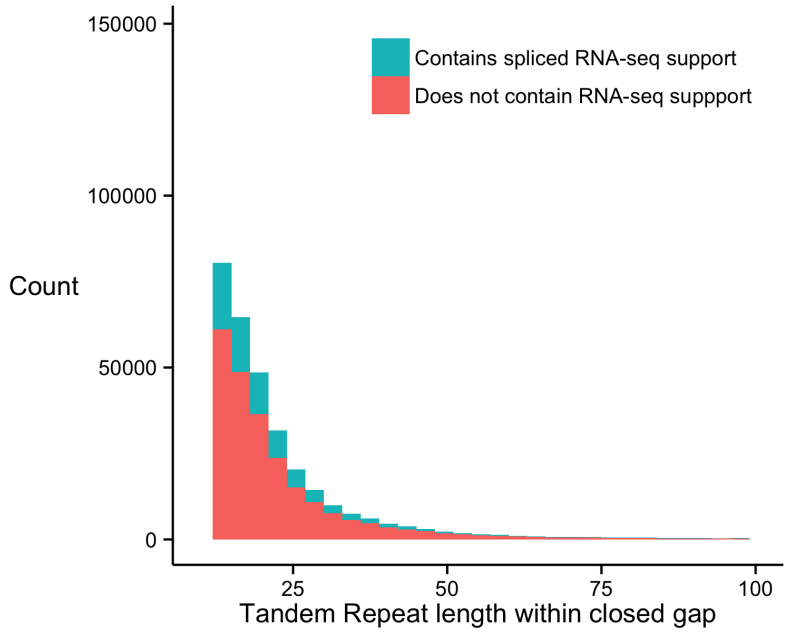

Figure S32. Length distribution of tandem repeats found within closed gaps. Tandem repeats within a closed gap and overlapping a spliced RNA-seq transcript are colored in blue.

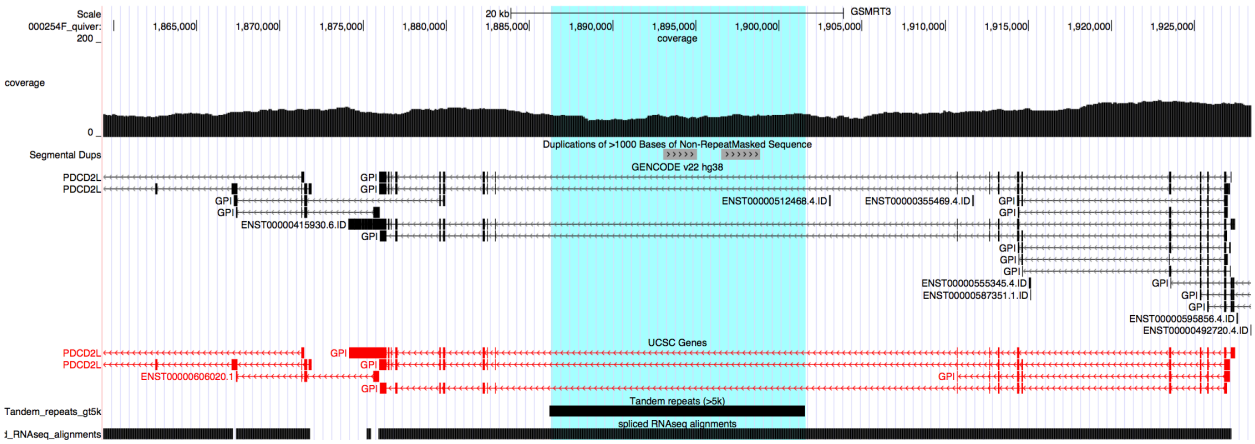

Figure S33. The expanded microsatellite from a closed gap in Susie3.

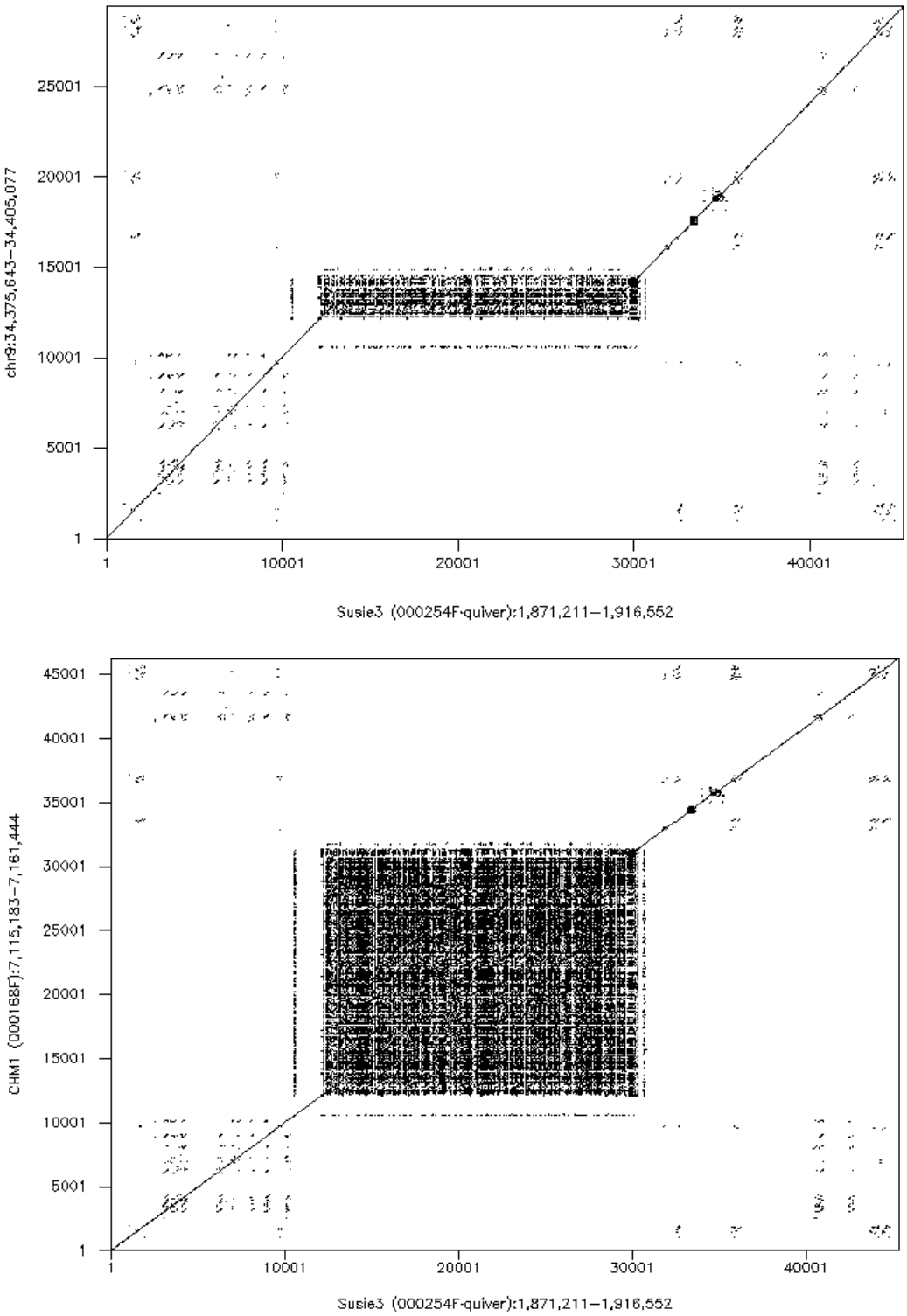

Figure S34. An expanded macrosatellite within a closed gap in Susie3 (000254F_quiver) compared to GRCh38 chromosome 9 (top). CHM1 shows the same expansion of the satellite as Susie3 (bottom).

Figure S35. A 12.3 kbp tandem repeat spans a closed gap in Susie3.

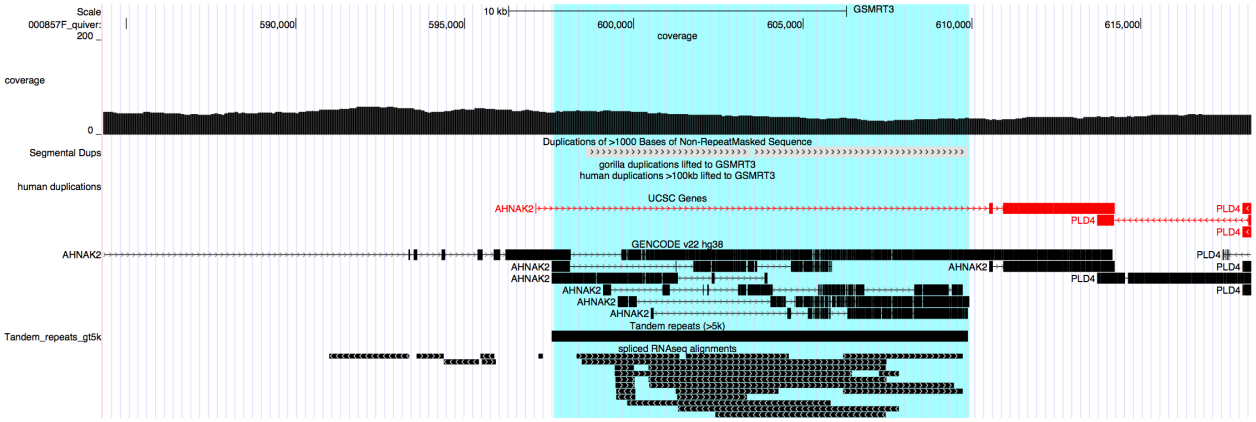

Figure S36. A 12.3 kbp tandem repeat spans a closed gap in Susie3. This tandem repeat lies within the intronic region of AHNAK2.

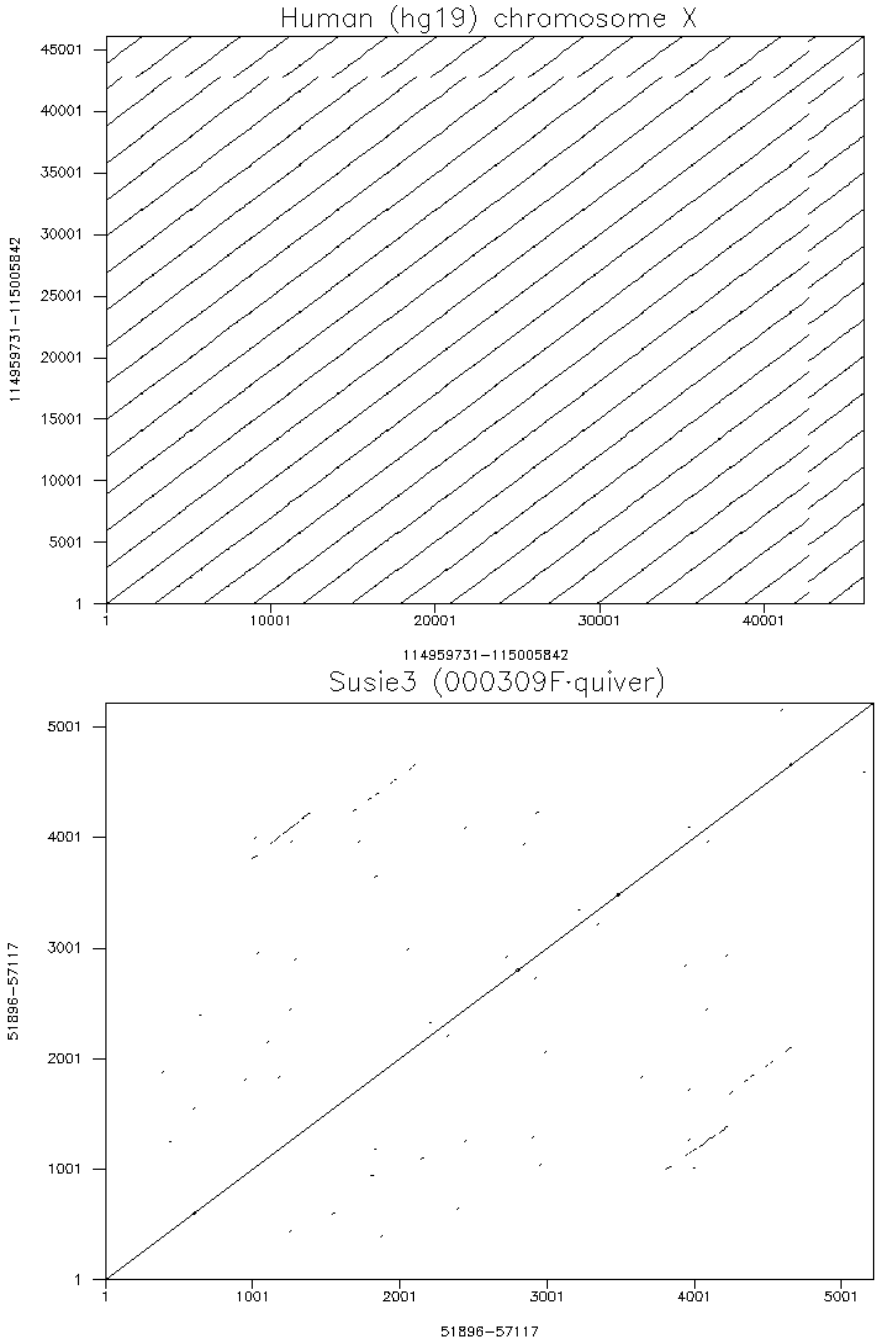

Figure S37. Dot plot of the DXZ4 macrosatellite found on hg19 chrX:114,959,731–115,005,842 (top) (*49*)**.** DXZ4 is compressed into a single 3 kbp monomer in a Susie3 contig (bottom). An additional 1 kbp of flanking sequence was added to the plot.

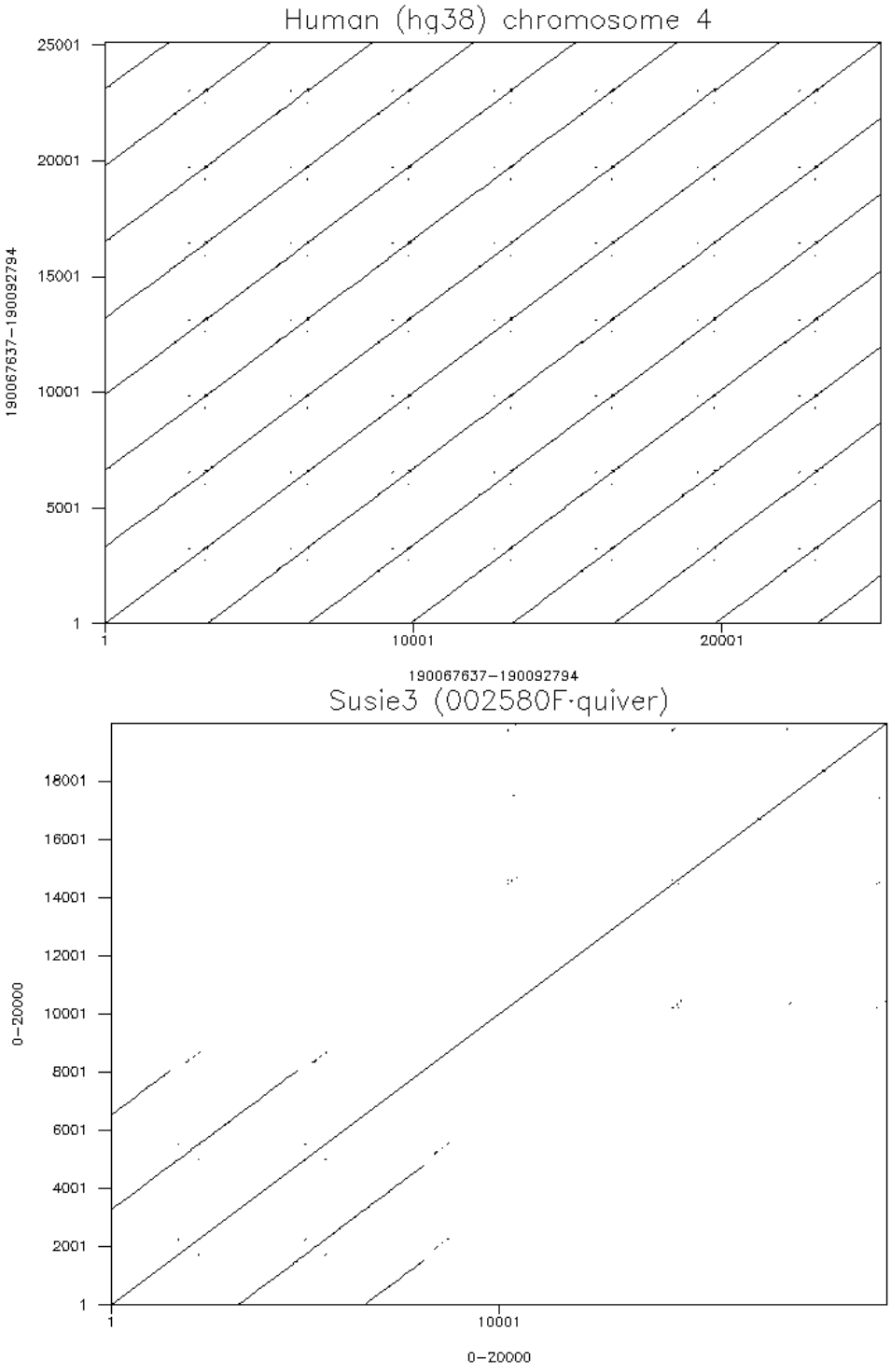

Figure S38. Dot plot of the D4Z4 macrosatellite found on hg38 chr4:190067637-190092794 (top); one end point of D4Z4 found at the end of a Susie3 contig (bottom).

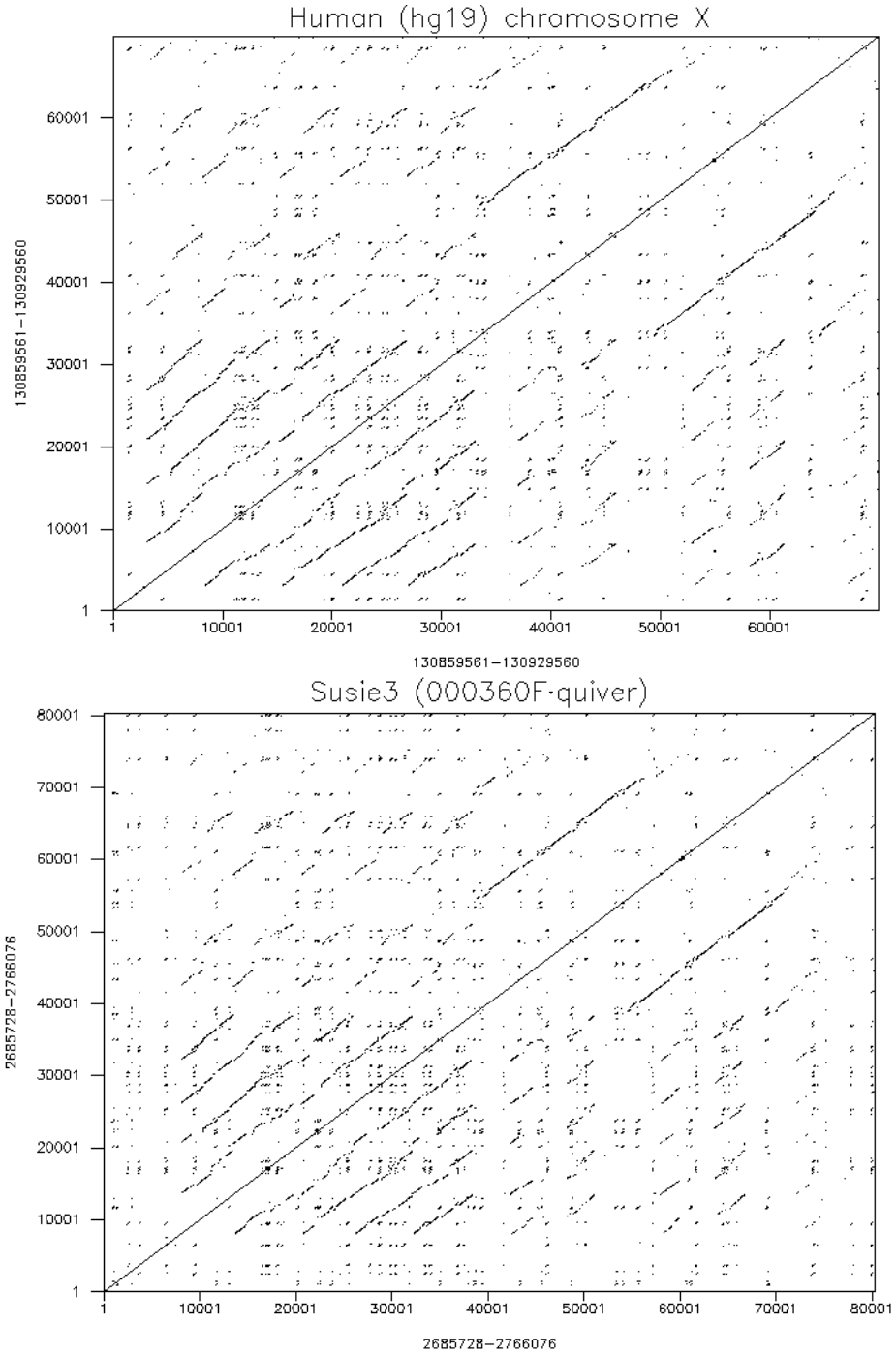

Figure S39. Dot plot of the X130 macrosatellite found on hg19 chrX:130,859,561–130,929,560 (top) (*49*)**.** The entire ~70 kbp macrosatellite is contained within a single Susie contig (bottom).

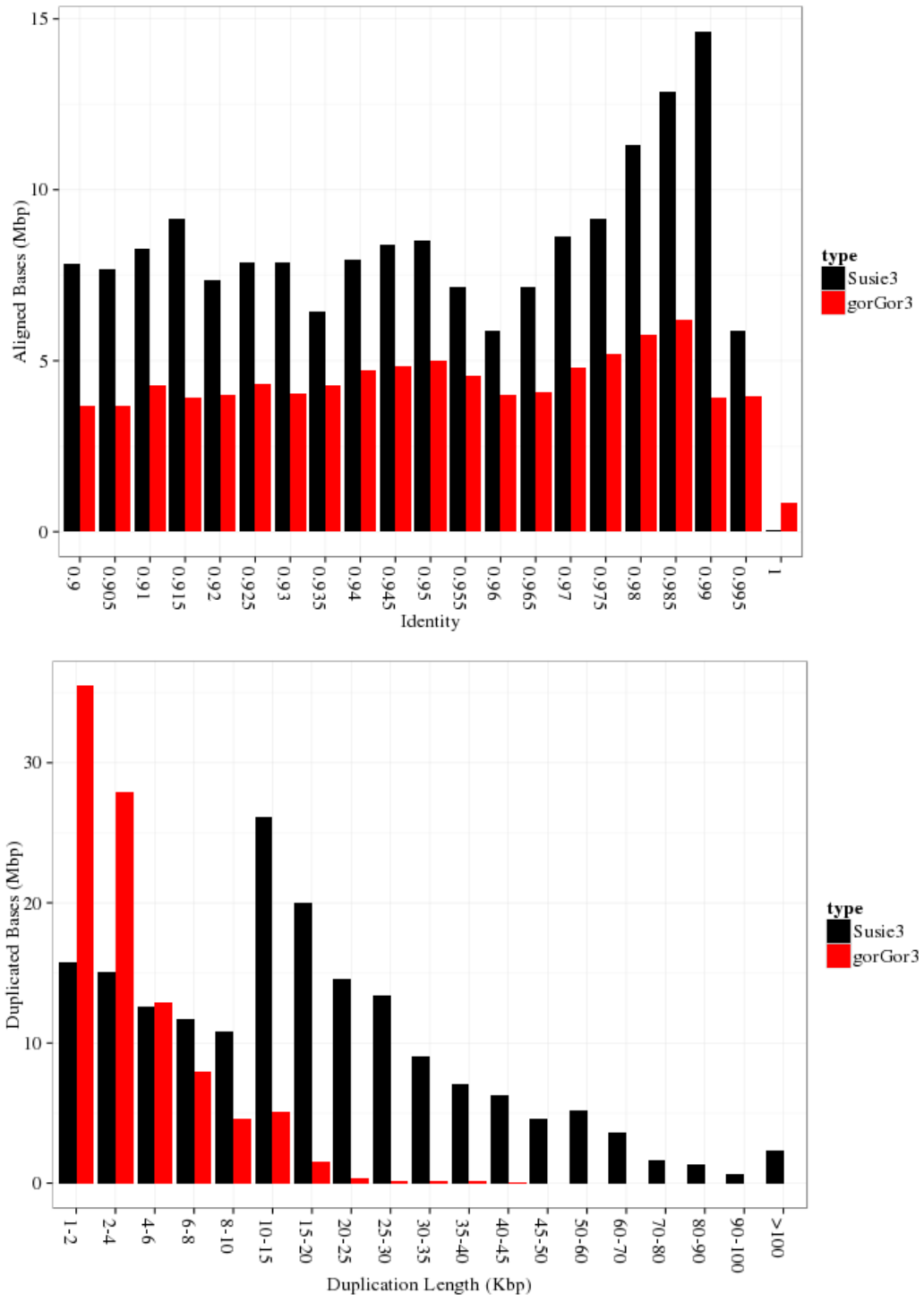

Figure S40. a) % Identity and b) length distribution of segmental duplications detected in Susie3 vs. gorGor3 genome assemblies and validated by read depth.

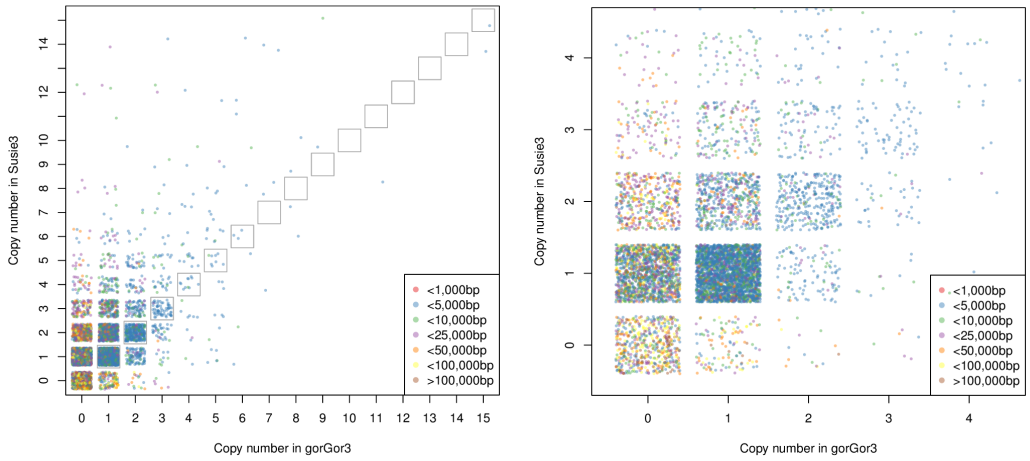

Figure S41. The copy number of copy number variable sequences detected by (*17*) **using read depth in gorGor3 and Susie3 for all sequences (left), and the subset of sequences with copy number of at most five in either assembly (right).** Copy number is integral but point location is spread for visualization. The size of the expanded regions is indicated by point color. The grid elements outlined in gray represent equal copy number, and above the gray diagonal Susie3 is increased in copy number, and below, gorGor3.

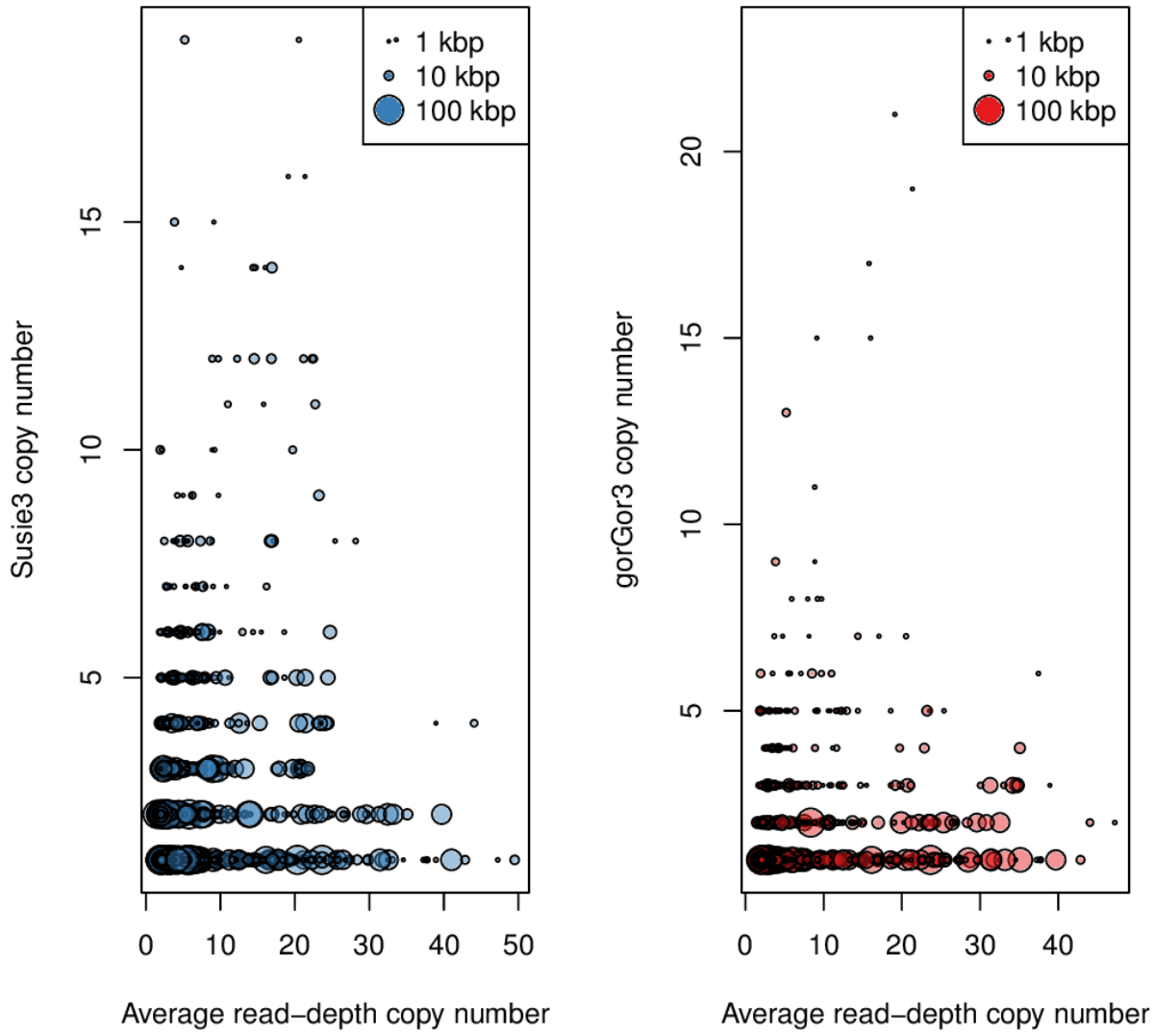

Figure S42. The copy number of sequences in Susie3 (left) and gorGor3 (right) versus the average copy number estimated through read depth by (*17*)**.**

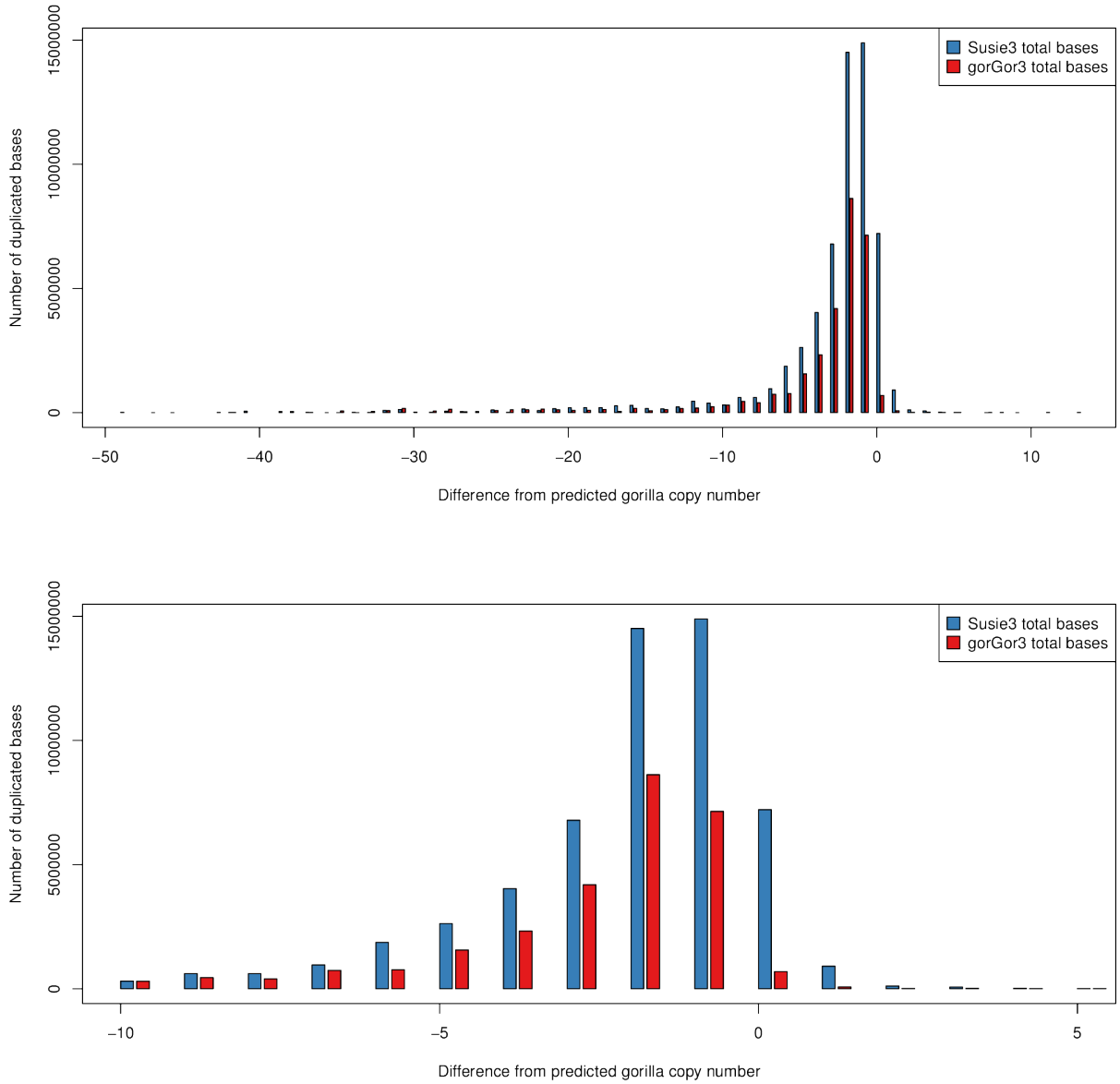

Figure S43. The difference in copy number between assemblies and read depth, counted by total bases (top) and number of loci (bottom). The copy number for read depth is the average of 28 gorilla genomes sequenced by Illumina. A sequence is counted as present in an assembly if it has at least 80% of the sequence length and 80% identity of the duplicated sequences mapped to human.

Figure S44. The intersection of genes detected as expanded in Susie3 and gorGor3, considering annotation by HUGO gene name.

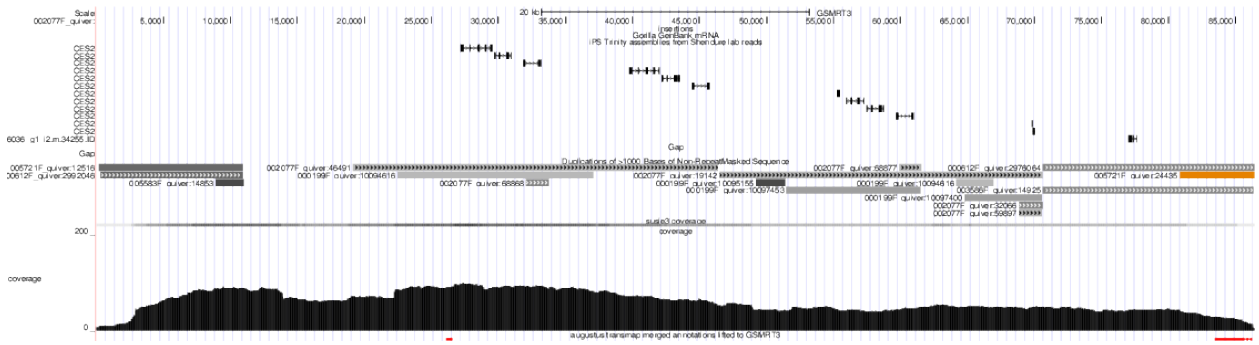

Figure S45. An example of a short (85 kbp) contig that contains multiple copies of the CES gene family. We estimate there are four copies of this gene family based on sequence read depth; however, only three are represented in the Susie3 assembly. The uneven sequence coverage profile is indicative of an assembly collapse and unresolved copy.

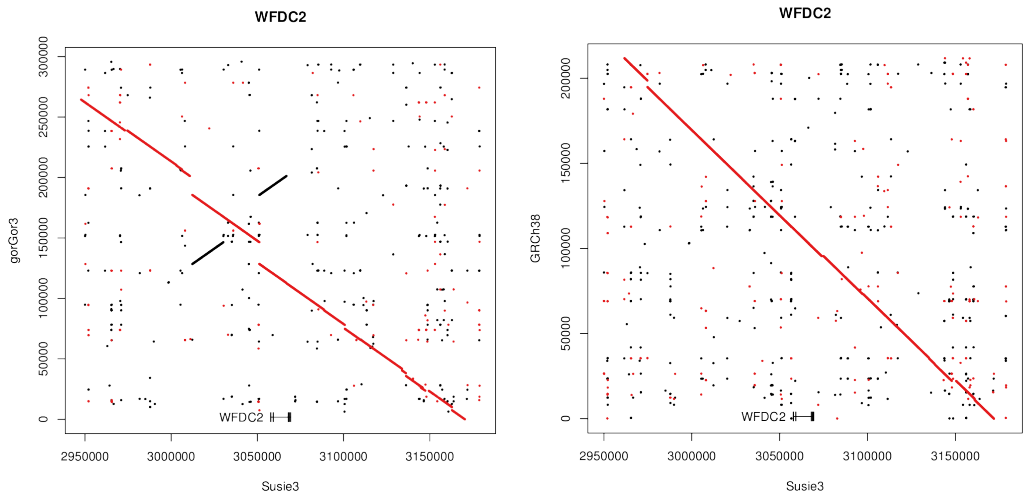

Figure S46. *WFDC2* is present in two copies in gorGor3 as an inverted duplication but found only once in human (GRCh38) and Susie3 gorilla genome assemblies. Read depth supports a single copy in Susie and Kamilah (gorGor3), indicating that the duplication of this gene in gorGor3 is likely an artifact of assembling this gene family locus.

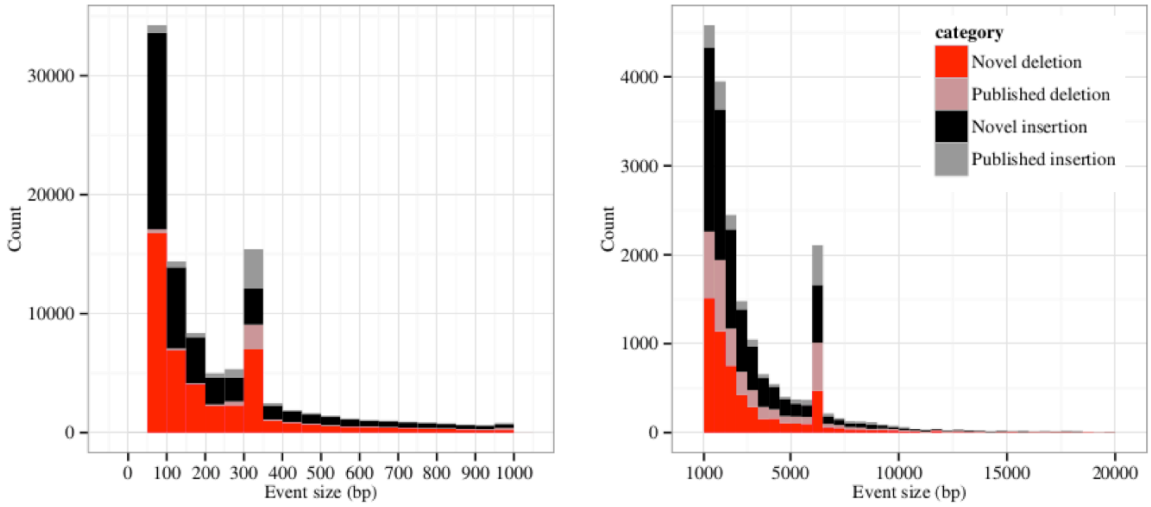

Figure S47. Distribution of insertion and deletion lengths for Susie3. Previously published deletions and insertions are shown in light red and gray, respectively.

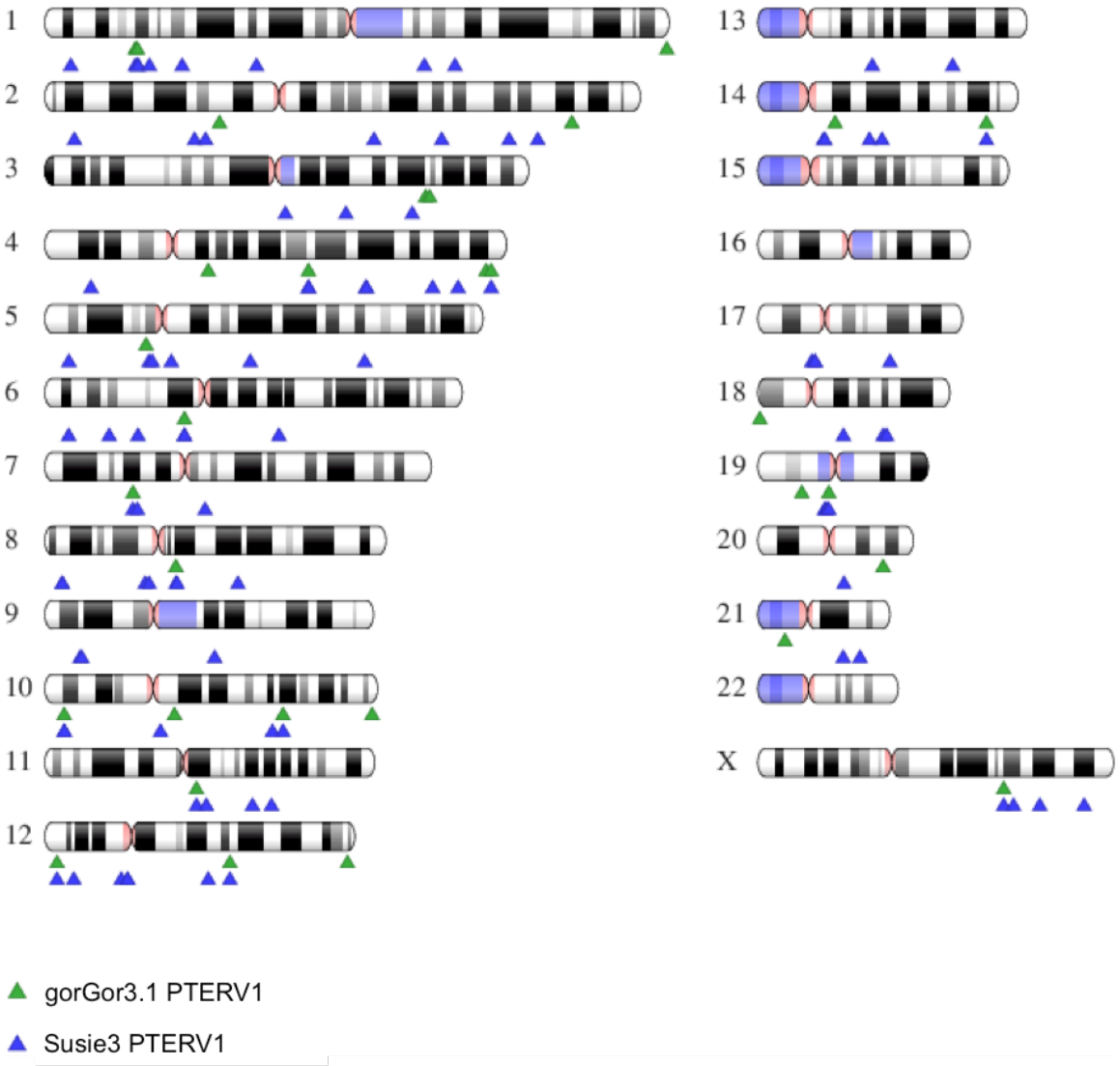

Figure S48. The locations of PTERV elements over 6 kbp for gorGor3 (blue) and GSMRT3.1 (green), mapped to GRCh38.

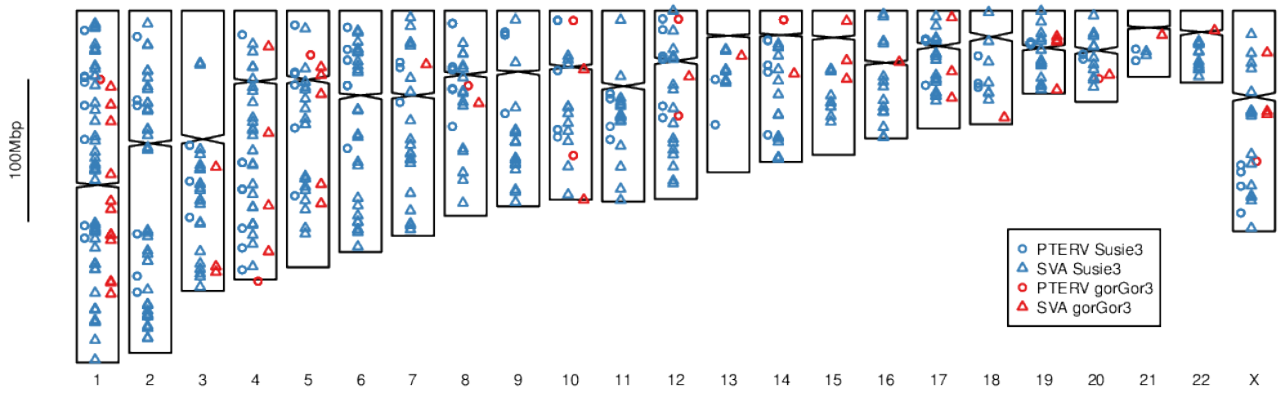

Figure S49. The locations of full-length PTERV (>6 kbp) and SVA (>2 kbp) insertions in Susie3 and gorGor3 are shown projected onto human.

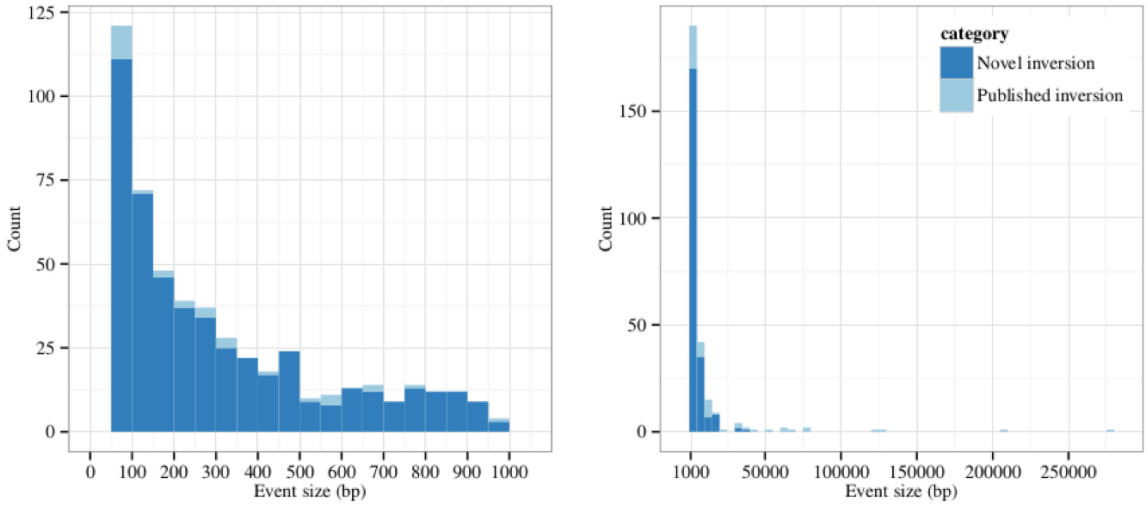

Figure S50. Distribution of inversion lengths for Susie3. Previously published inversions are shown in light blue. Of the 697 variants shown, 615 (88%) were not observed in previous studies (*3*, *16*) with a mean length of 1,584 bp and a maximum length of 35,019 bp.

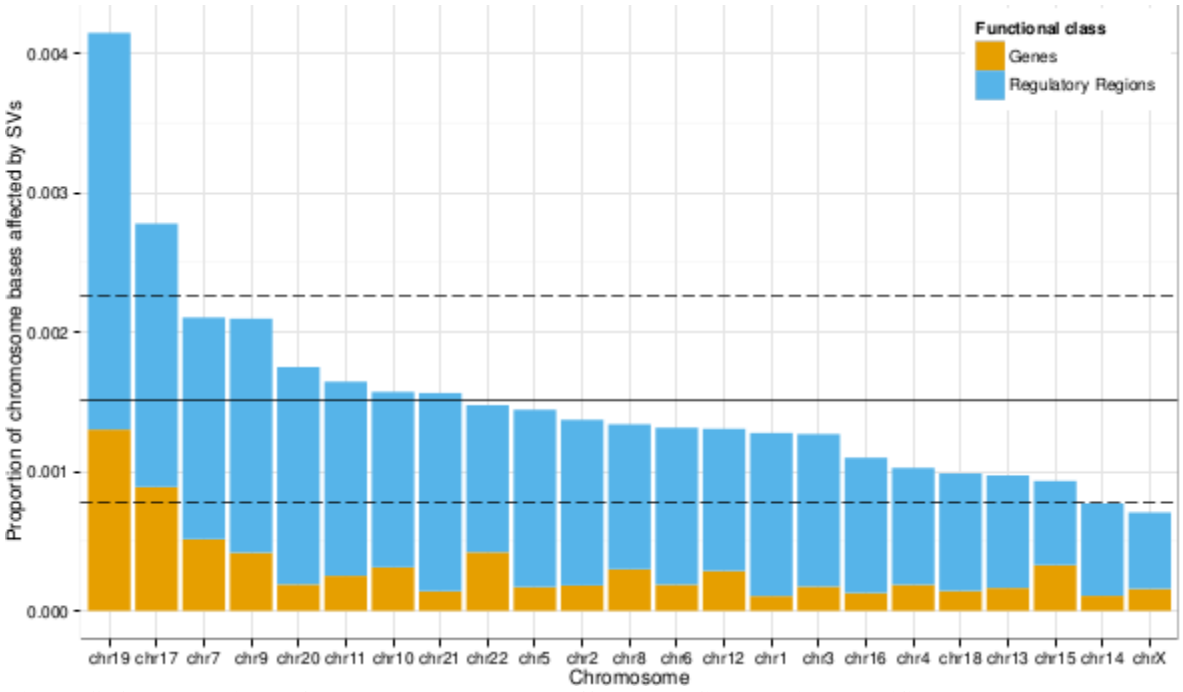

Figure S51. Proportions of human chromosomes affected by fixed gorilla-specific structural variants (GSVs) intersecting functional regions, including genes (coding and noncoding exons) in orange and regulatory regions (DNase, H3K4me3, and H3K27Ac clusters) in blue. The mean proportion of bases affected across all chromosomes is shown by the solid horizontal line and the mean +/- one standard deviation is shown by the two dashed lines.

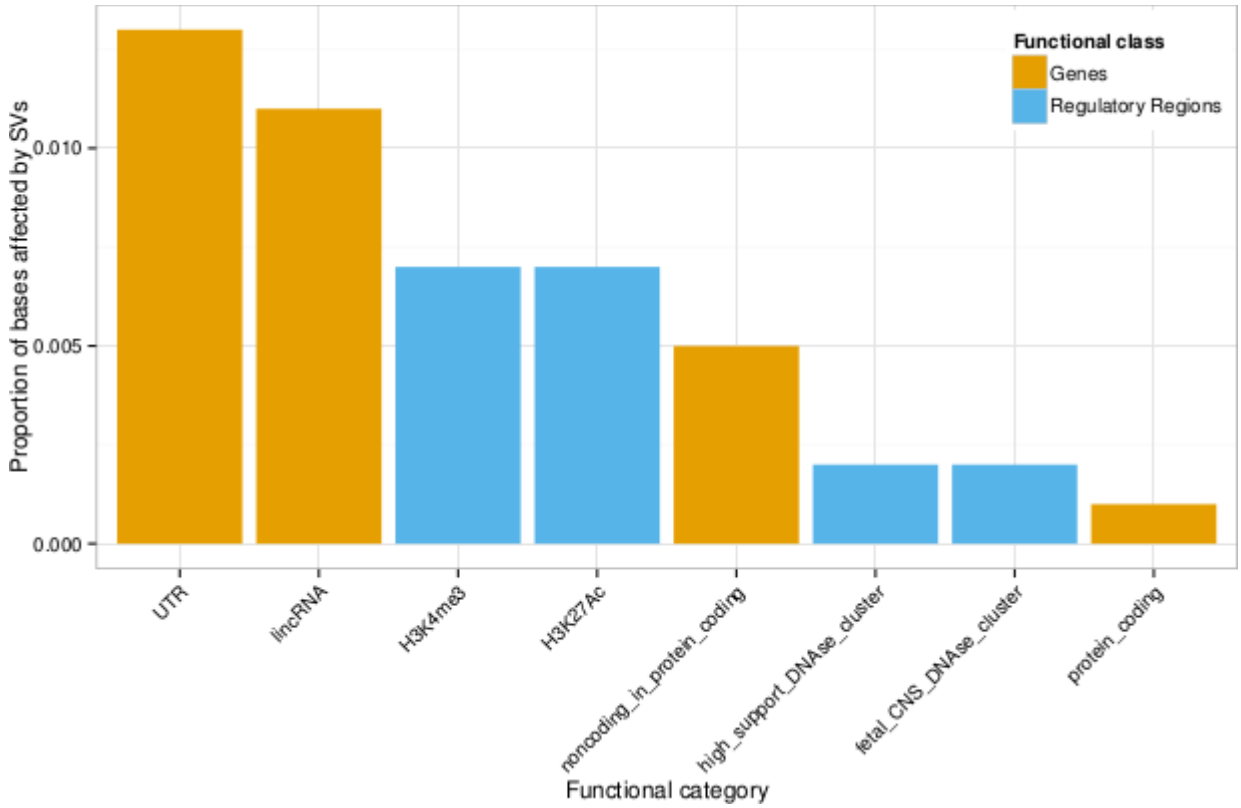

Figure S52. Proportion of functional categories affected by fixed GSVs. Proportions are calculated by the nonredundant exonic or regulatory bases of each category affected by SVs divided by the total genomic bases per category.

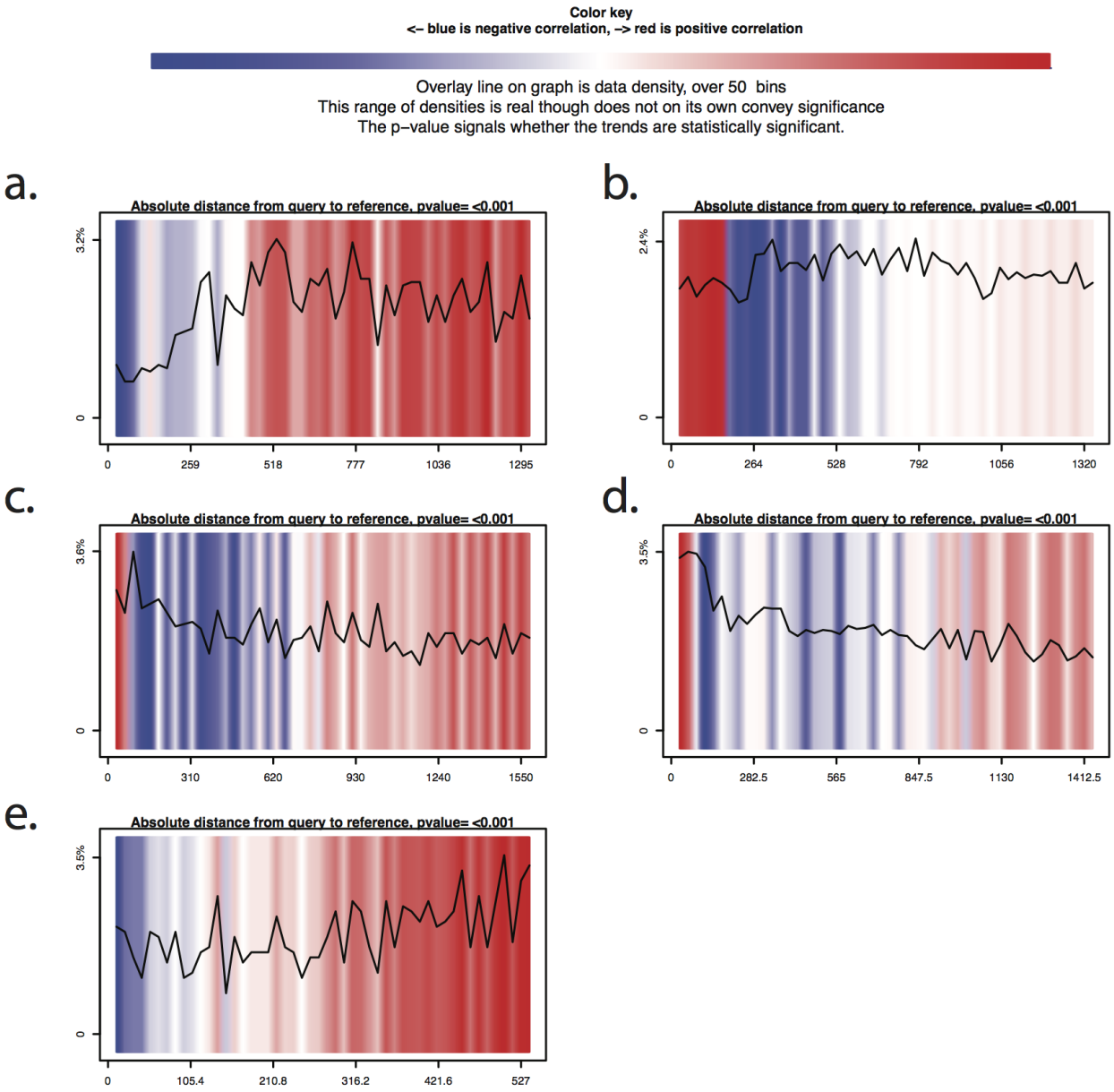

Figure S53. Spatial relationship between GSVs and protein-coding genes (A), DNase clusters from fetal central nervous system (B), H3K4me3 methylation marks (C), H3K27AC acetylation marks (D), and lincRNAs (E). Each panel shows the distance of a GSV to the closest feature (black line). The distances are binned and the percent of data is shown on the y-axis. The background shading denotes if a bin has more (red) or less (blue) GSVs than expected. For example, GSVs are often not found near protein-coding genes (A) or lincRNAs (E). While the GSVs near fetal DNase clusters appear enriched, it is not statistically significant (projection test p-value = 4.3e-10). The increased number of GSVs near H3K27AC and H3K4me3 marks is statistically significant (see text).

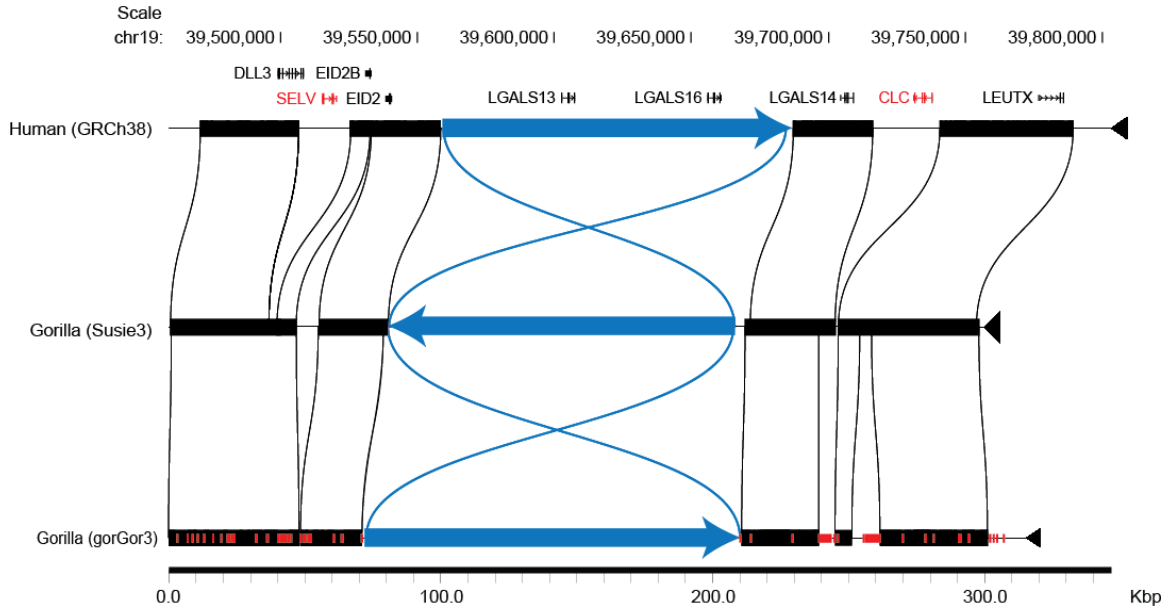

Figure S54. A ~140 kbp gorilla inversion in LGALS13/16 locus and deletion of SELV and CLC between human and gorilla assemblies. Both gorilla assemblies confirm the absence of SELV and CLC in gorilla. However, the inversion is detected only in the Susie3 assembly and not in the gorGor3 assembly most likely due to the presence of gaps adjacent to the inverted sequence in gorGor3. This locus in gorGor3 (chr19:36878605-37193959) contains 58 gaps with a mean size of 350 bp and a maximum size of 5,226 bp.

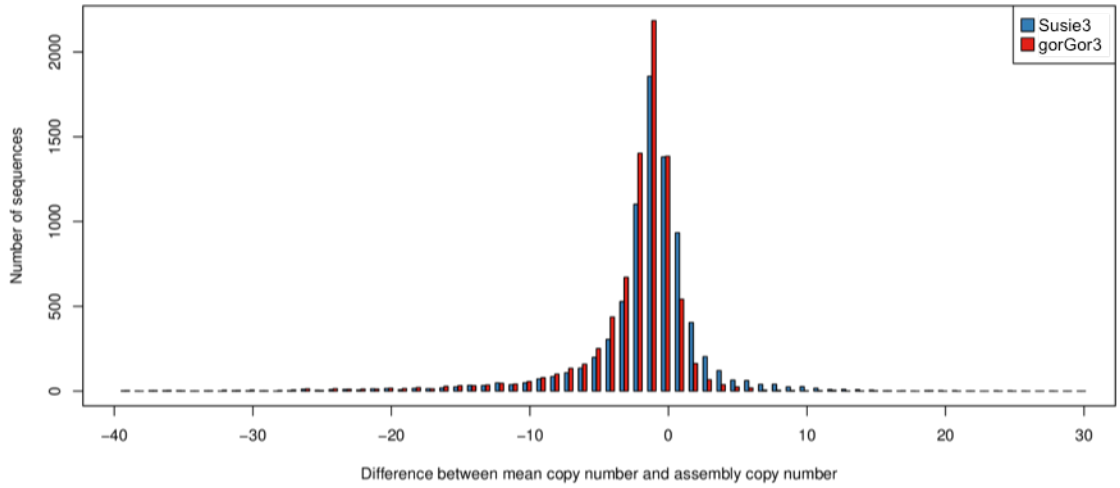

Figure S55. The difference between the copy number in Susie3 and the average copy number of CNV sequences from (*17*) (blue) and the equivalent counts for gorGor3 (red).

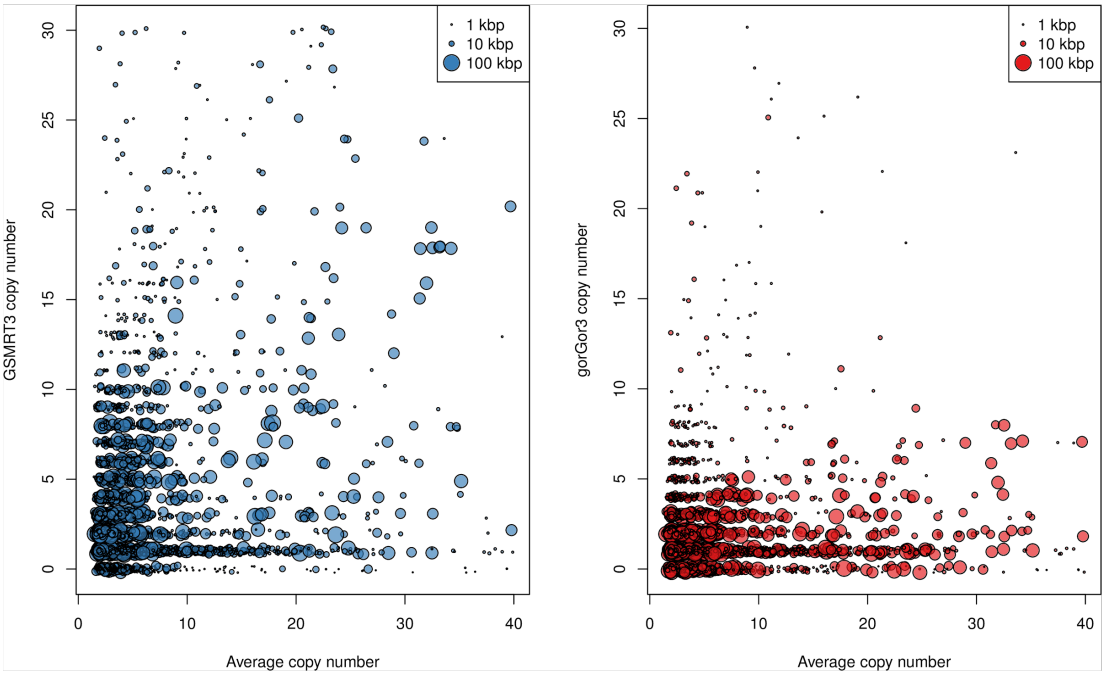

Figure S56. The copy number of CNV sequences in (*17*) are shown versus the average copy number across 28 Western lowland gorillas for Susie3 (left) and gorGor3 (right). The length of the copy number region is proportional to the area of the circle.

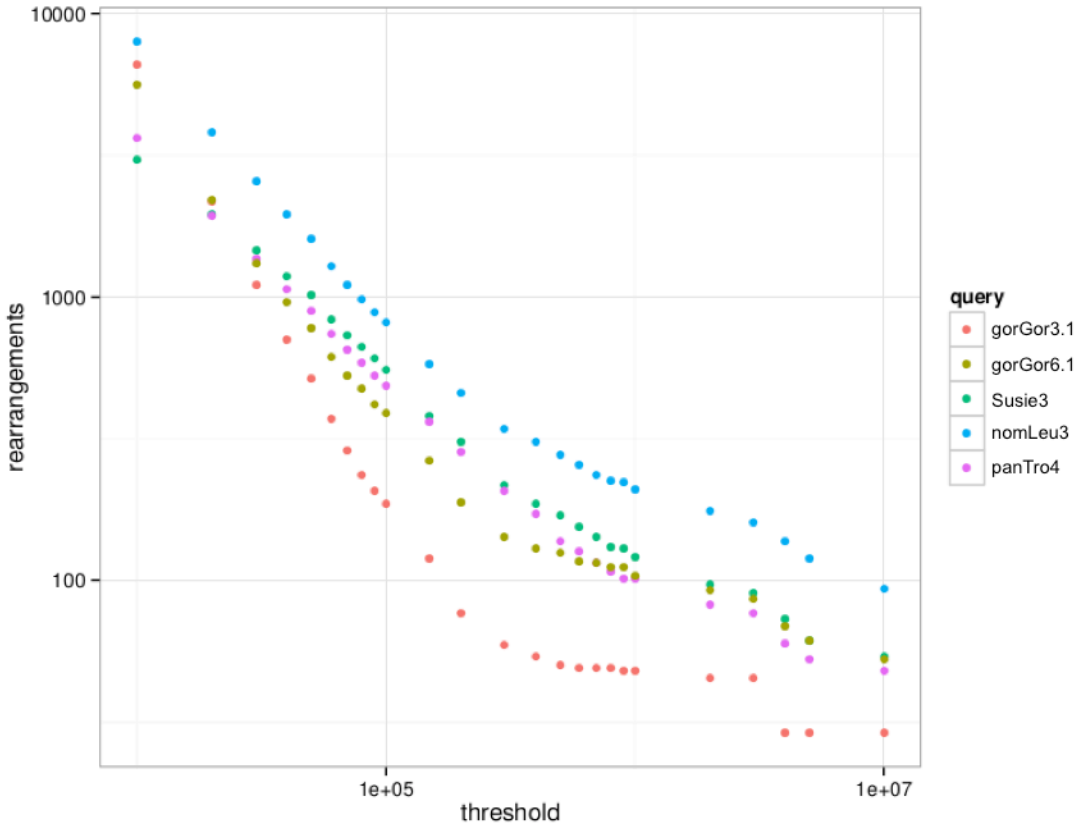

Figure S57. Number of genomic rearrangements by size threshold and reference assembly. Rearrangements are estimated by whole-genome alignment of autosomal chromosomes and chrX for each assembly against the human reference (GRCh37) with LASTZ, chaining of alignments, netting of alignment chains, and counts of total nets at each threshold as previously described (*5*).

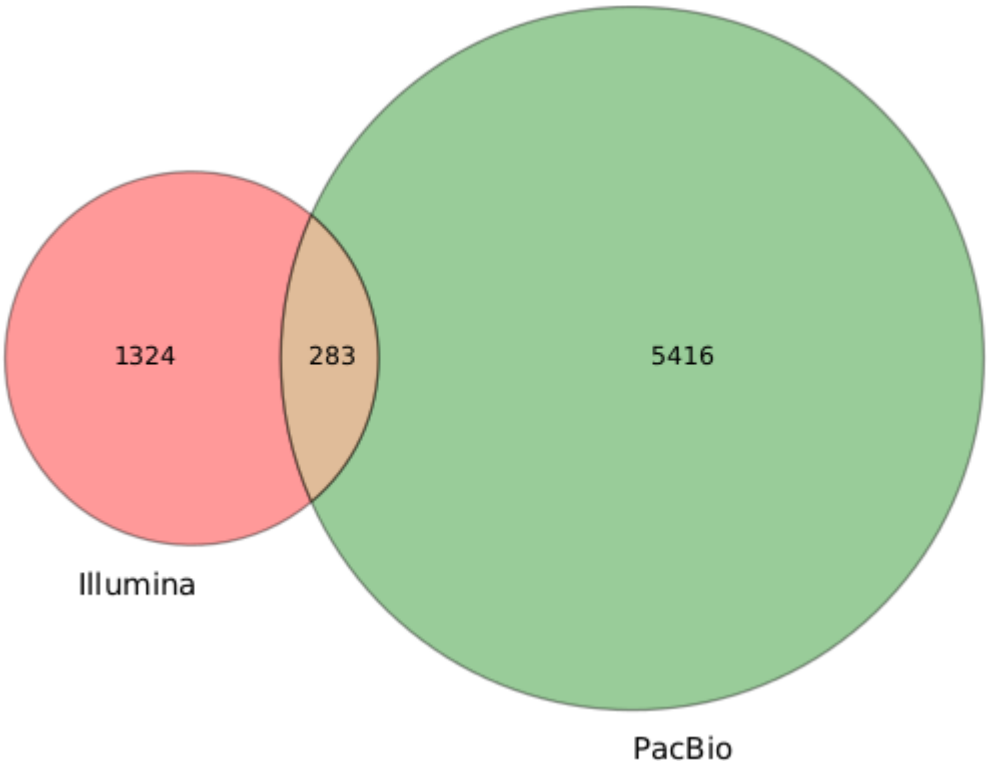

Figure S58. Genes affected by SVs or indels in the gorilla lineage relative to other great apes based on deleterious variants identified with Illumina short reads in previously published studies and variants identified with *de novo* assembly of PacBio long reads in Susie.

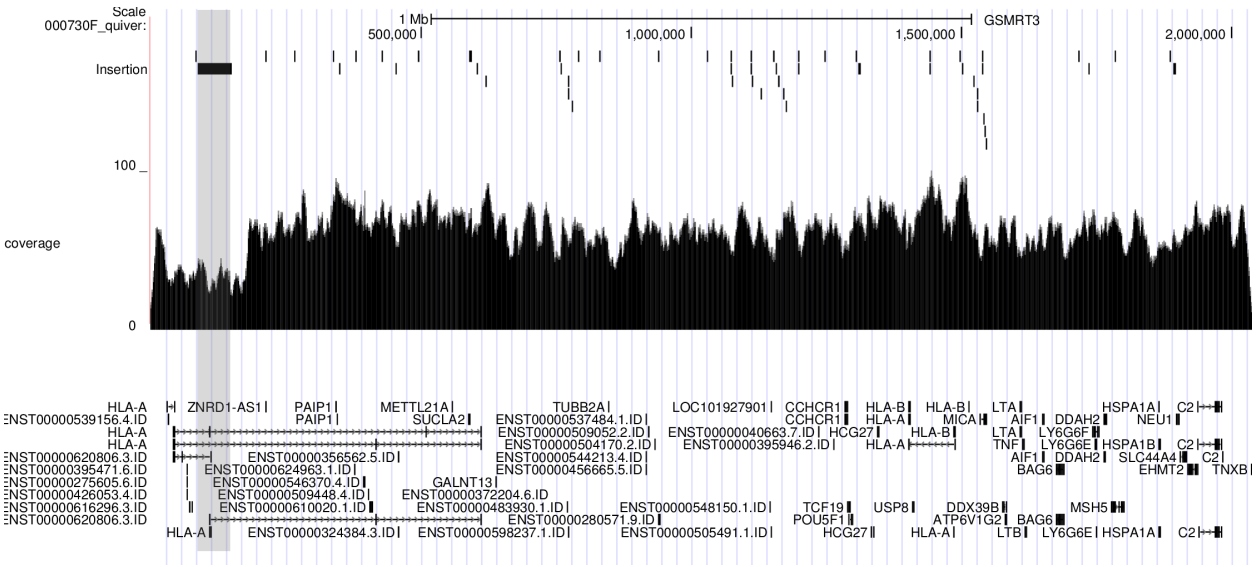

Figure S59. Susie3 contig with an inserted copy of a gene similar to *HLA-A*. The near-half drop in coverage from the flanking regions suggests that two alternate *HLA-A* haplotypes have been assembled.

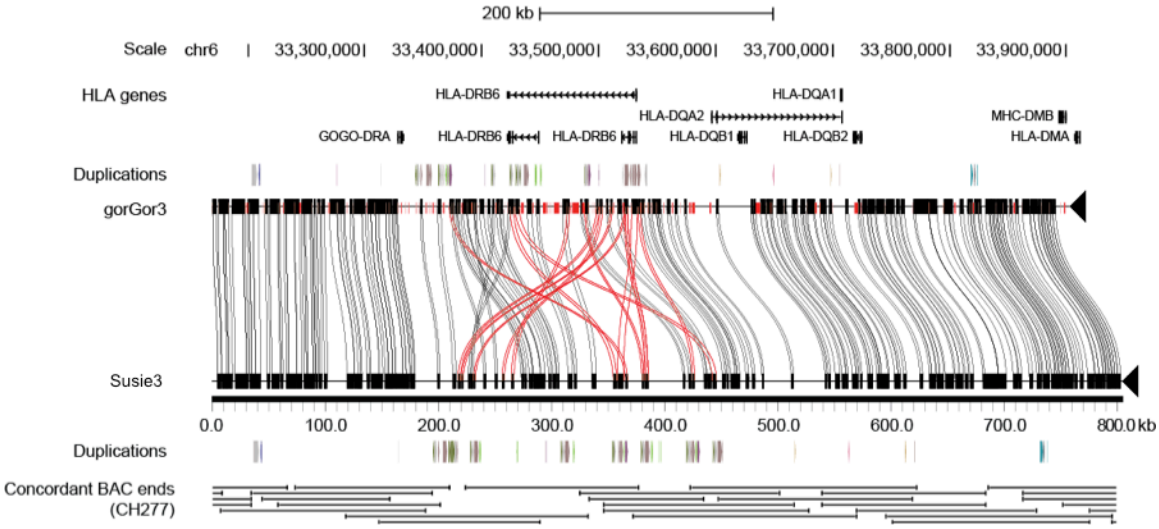

Figure S60. Alignment of MHC Class II locus in Susie3 against gorGor3 with Miropeats (*31*)**.** Sequences that disrupt the collinearity of the alignment are highlighted in red and indicate potential misassemblies in gorGor3 associated with 168 annotated gaps at this locus. Support for the proper organization of the Susie3 sequence is shown by the tiling path of concordant BAC-end sequences from the Kamilah BAC library (CHORI-277). We subsequently sequenced and confirmed the organization and sequence of Susie3 compared to gorGor3.

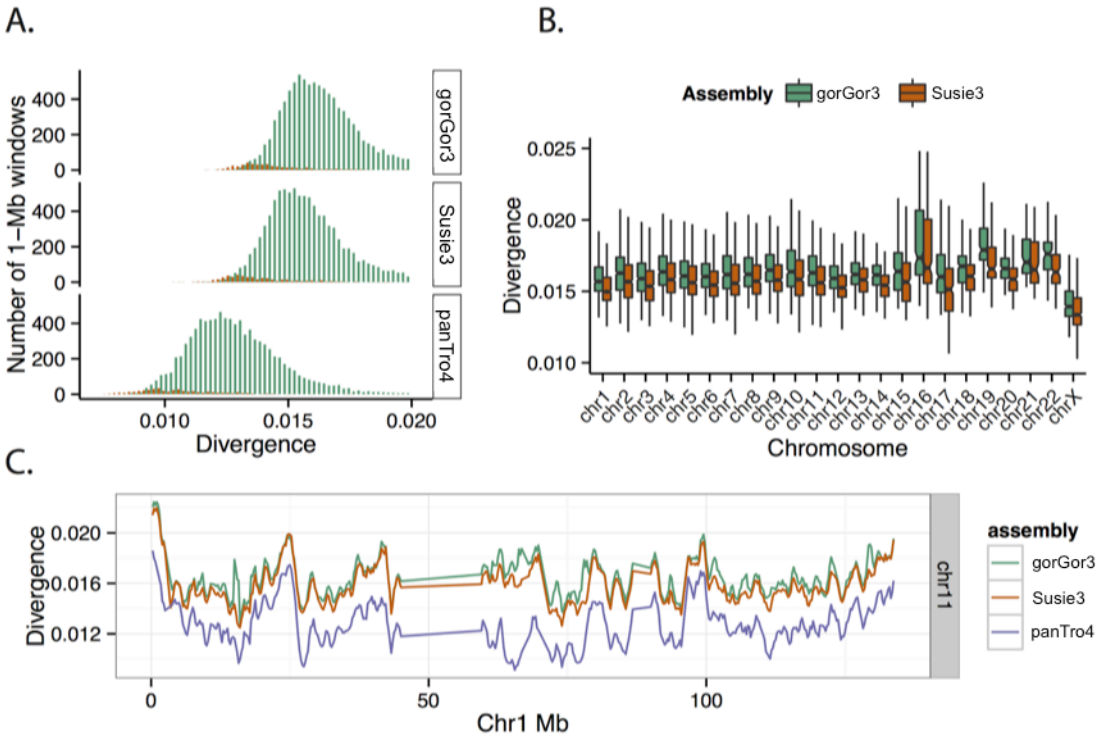

Figure S61. A) Average divergence within 1 Mbp windows between human (GRch38) and chimpanzee (panTro4), Susie3 or gorGor3. Autosomes are shown in green and the X chromosome is shown in orange (autosome means: gorGor3: 1.65%, Susie3: 1.6%, panTro4: 1.3%) (chrX means: gorGor3: 1.4%, Susie3: 1.4%, panTro4: 1.0%). All pairwise comparisons of the mean are significant (ANOVA: divergence ~ assembly; Tukey multiple comparison of mean, 95% family-wise confidence level; p <= 0.001). B) Correlation between gorGor3 and Susie3 chromosomal divergence. The empirical probability, from a permutation test, that Susie's divergence is higher than gorGor3 is 14/1e3, prob = 0.014. C) An example of divergence across GRCh38 Chr1. As previously noted, gorilla- and chimpanzee-human divergence was negatively correlated with the distance from the centromere (*4*, *67*).

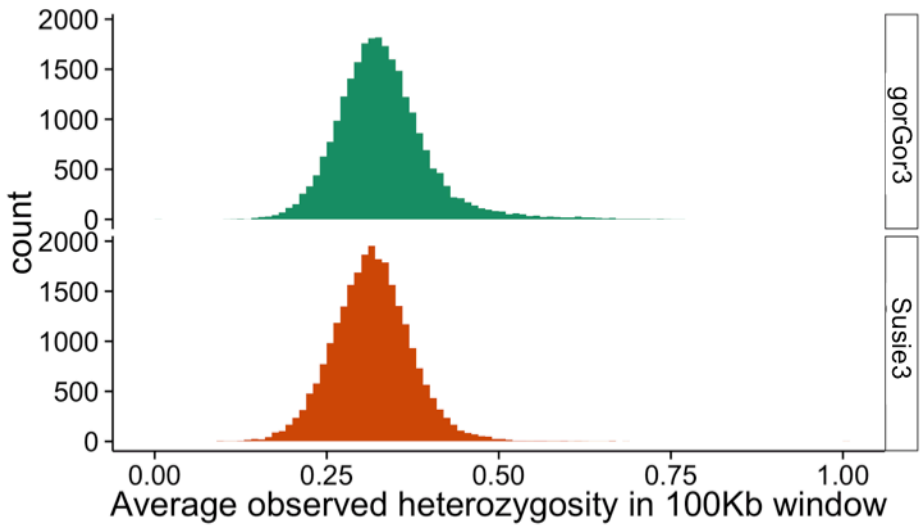

Figure S62. The distribution of autosomal heterozygosity within 20 kbp for the seven Western lowland gorillas. The means are 0.33 and 0.32 for gorGor3 and Susie3, respectively.

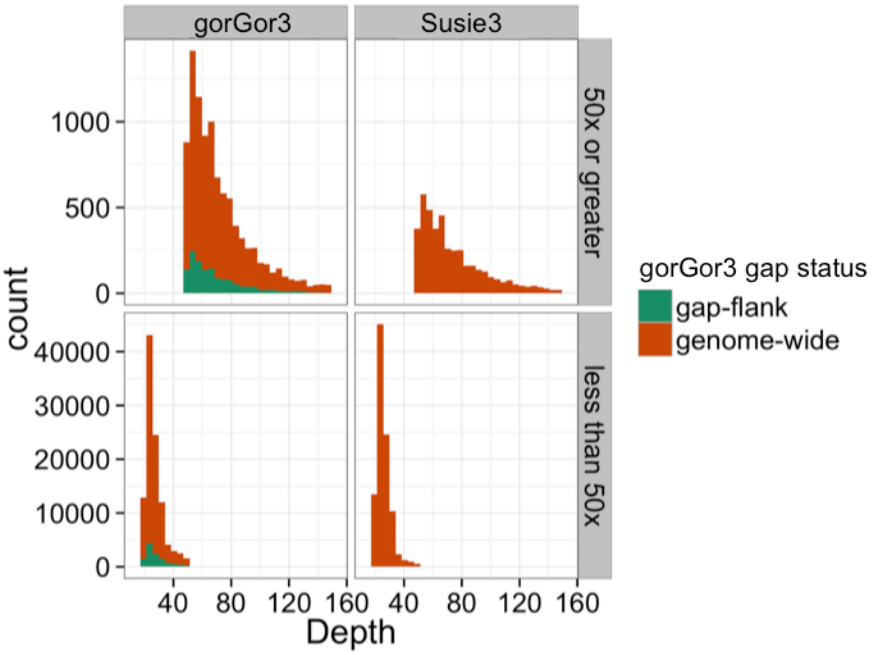

Figure S63. Depth of heterozygous sites after remapping reads containing Coco gorGor3 heterozygous sites. There are more high-depth heterozygous calls against gorGor3 compared to Susie3. Datasets were subsampled to have the same number of data points.

12. Supplementary Tables

Table S1. Western lowland gorillas targeted for genome sequence and assembly.

	Susie	**Kamilah***
Species	*G. gorilla gorilla* (Western lowland)	*G. gorilla gorilla* (Western lowland)
Sex	Female	Female
Date of birth	Sept. 27 2004	Dec. 5 1977
Location	Columbus Zoo and Aquarium, Powell, Ohio USA	San Diego Wild Animal Park, San Diego, California USA
Stud number	T1193	661
Parents	Jojo, Bahati	Pete, Nina
Sequencing	PacBio SMRT P6C4, Illumina MiSeq, HiSeq, TruSeq	Sanger BAC/fosmid-end sequencing

*Kamilah BAC/fosmid-end sequences were used to scaffold the Susie3 assembly from Susie.

Table S4. Assembly summary statistics.

	Susie3	gorGor3	gorGor4	panTro3	panTro4
Individual	Susie	Kamilah	Kamilah	Clint	Clint
Assembly size	3,080,414,926	3,035,660,144	3,063,362,754	3,307,943,878	3,323,267,922
Total coverage	74.8X	37.1X	101.1X	6X	6X
Sequencing technology	PacBio (74.8X P6C4)	WGS (2.1X) and Solexa* (35X)	WGS (2.1X), Solexa* (35X), Illumina* (64X)	WGS capillary sequencing (6X)	WGS capillary sequencing (6X)
NCBI/ENA accession	PRJEB10880	GCA_000151905.1	GCA_000151905.3	GCA_000001515.3	GCA_000001515.5
Scaffold N50	23,141,960	913,458	81,227,029	9,211,238	8,925,874
Contig N50	9,558,608	11,661	52,934	50,595	50,656
Total assembly gap length	11,793,321	206,771,311	145,977,342	407,399,385	420,897,672
Number of contigs	16,073	464,874	170,105	183,905	183,860
Number of scaffolds	554	57,196	40,730	26,994	27,005
Assembler	Falcon v.0.3.0	ABySS and Phusion	N/A	PCAP	PCAP

Comparison of assembly statistics between the gorilla SMRT genome assembly of Susie (Susie3) compared to two gorilla assemblies based on Illumina–Sanger WGS of Kamilah. Most comparisons were made between Susie3 and gorGor3 because they represent initial genome assemblies and the details for assembly of gorGor3 have been published (*4*). A summary of all primate assemblies used in this study is provided in table S5.

* Solexa was eventually acquired by Illumina, but the Solexa reference in the table refers to sequencing technology that produced average read lengths of 35 bp, while the Illumina reference refers to sequencing technology that produced an average of 110 bp.

Table S5. Mammalian genome assemblies used in the study.

Assembly name	Species	Common name	Accession
Susie3	*Gorilla gorilla gorilla*	Western lowland gorilla (Susie)	PRJEB10880
gorGor3	*Gorilla gorilla gorilla*	Western lowland gorilla (Kamilah)	GCA_000151905.1
gorGor4	*Gorilla gorilla gorilla*	Western lowland gorilla (Kamilah)	GCA_000151905.3
GRCh38	*Homo sapiens*	Human (GRC build 38)	GCA_000001405.15
ponAbe2	*Pongo pygmaeus abelii*	Sumatran orangutan (Susie)	GCA_000001545.3
panTro4	*Pan troglodytes*	Chimpanzee (Clint)	GCA_000001515.4
saiBol1	*Saimiri boliviensis boliviensis*	Bolivian squirrel monkey	GCA_000235385.1

Table S6. Gorilla clone insert sequences used for accuracy estimate.

Clone name	Accession
CH277-204E16	AC239281.3
CH277-145B18	AC239356.2
CH277-1A9	AC239360.3
CH277-217K13	AC239362.3
CH277-123C22	AC239379.3
CH277-235C21	AC239380.3
CH277-159N16	AC240953.2
CH277-205P14	AC240968.2
CH277-460D6	AC241242.3
CH277-120A17	AC241257.3
CH277-211C13	AC241451.3
CH277-485J9	AC241471.3
CH277-103D21	AC241522.3
CH277-325C3	AC241523.3
CH277-4O6	AC241525.3
CH277-481C13	AC242627.3
CH277-545D7	AC243178.3
CH277-45G13	AC254968.1
CH277-80C4	AC254993.1

Accessions of finished gorilla BACs from the Kamilah library (CHORI-277) used to assess assembly accuracy.

Table S10. BAC-end sequence accuracy estimate.

Type	Susie3	Susie3 (>=200 kbp contigs)	Susie3.2	gorGor3	gorGor4
Transition	74,408	74,374	69,825	41,997	44,521
Transversion	33,802	33,779	31,552	20,464	21,334
Deletion (1 bp)	69,950	69,885	39,389	17,558	18,788
Deletion (>1 bp)	18,045	18,027	16,204	7,661	8,859
Insertion (1 bp)	12,771	12,760	7,622	4,670	5,522
Insertion (>1 bp)	13,159	13,153	12,649	7,390	7,462
Total differences	222,135	221,978	177,241	99,740	106,486
High-quality bases	68,489,853	68,463,496	68,887,922	63,716,903	68,442,595
Expected Sanger errors	25,341	25,331	25,489	23,575	25,324
PacBio - Sanger errors	196,794	196,647	81,163	76,165	81,162
Accuracy	0.997127	0.997128	0.998822	0.998805	0.998814
Ti/Tv	2.20	2.20	2.21	2.05	2.09

Accuracy of Susie3 sequence based on alignment of BAC-end sequences from the Kamilah BAC library (CHORI-277).
Sanger-end sequences were aligned to each assembly using ALIGN (Smith-Waterman) and considered only high-quality base pairs (PHRED QV > 35).

Table S11. Repeat content of closed gaps in gorGor3.

	Total bases			Percent of bases in:		
Repeat class	Susie3	gorGor3 true gaps	gorGor3 false gaps*	Susie3	gorGor3 true gaps	gorGor3 false gaps*†
Low complexity	18,596,936	1,345,475	1,412	0.6	1.6	1.7
LTR elements	265,228,481	4,518,283	5,319	8.8	5.5	6.5
Small RNA	1,139,346	32,973	101	0.0	0.0	0.2
SINEs	383,774,077	31,457,417	17,061	12.8	38.0	20.8
LINEs	538,288,766	13,628,364	25,000	17.9	16.5	30.6
Unclassified	5,482,941	1,266,755	388	0.2	1.5	0.5
Satellites	297,678,078	4,954,047	567	9.9	6.0	0.6
Simple repeats	36,862,794	2,722,987	1,921	1.2	3.3	2.3
DNA elements	101,708,248	1,664,859	2,022	3.4	2.0	2.5
Total	1,648,759,667	61,591,160	53,791	55.0	74.5	65.9

Common repeat content (RepeatMasker (*28*)) in gaps of gorGor3 closed by Susie3 compared to the full genome content (Susie3 and GRCh38). Repeat content of true gaps is for gaps >= 3,500 bp.
*A false gap is one where there are Ns (and often some additional sequence) in gorGor3 that should not be there. In some cases the additional sequence is artifactually duplicated nearby sequence.
†This percentage is based on 81,656 negative gaps bases (not "N") in the size range -1000 to 0.

Table S12. Gap length versus segmental duplication and gene content.

		Open gaps				Closed gaps				Genes			Exons	
Gap Length	Total gaps	Total open	With segdups	With exons	With gorilla dups	Total closed	With segdups	With exons	With gorilla dups	Closed	Closed and open	Open	Closed	Open
>0	233,718	2,233	1602 (0.01)	252 (0.00)	147 (0.00)	231,485	8053 (0.03)	5834 (0.02)	774 (0.00)	3,473	52	171	11,105	1,649
>1,000	28,501	841	781 (0.03)	123 (0.00)	36 (0.00)	27,660	2666 (0.09)	1862 (0.07)	116 (0.00)	1,535	13	115	3,730	752
>2,000	10,959	662	629 (0.06)	111 (0.01)	24 (0.00)	10,297	1669 (0.15)	913 (0.08)	61 (0.01)	837	11	104	2,549	729
>3,000	5,624	533	509 (0.09)	94 (0.02)	13 (0.00)	5,091	1118 (0.20)	527 (0.09)	42 (0.01)	518	9	99	1,788	702
>4,000	3,399	437	421 (0.12)	85 (0.03)	9 (0.00)	2,962	816 (0.24)	346 (0.10)	25 (0.01)	358	6	95	1,391	673
>5,000	2,271	376	362 (0.16)	79 (0.03)	7 (0.00)	1,895	622 (0.27)	264 (0.12)	22 (0.01)	267	6	89	1,097	609
>6,000	1,667	328	318 (0.19)	76 (0.05)	6 (0.00)	1,339	481 (0.29)	211 (0.13)	17 (0.01)	215	4	90	960	590
>7,000	1,320	285	276 (0.21)	70 (0.05)	6 (0.00)	1,035	374 (0.28)	173 (0.13)	15 (0.01)	180	4	87	838	573
>8,000	1,117	257	248 (0.22)	66 (0.06)	6 (0.01)	860	302 (0.27)	147 (0.13)	9 (0.01)	156	4	84	754	549
>9,000	953	224	215 (0.23)	58 (0.06)	6 (0.01)	729	244 (0.26)	127 (0.13)	7 (0.01)	139	3	74	658	505
>10,000	802	185	177 (0.22)	49 (0.06)	6 (0.01)	617	193 (0.24)	108 (0.13)	6 (0.01)	119	3	66	570	469

Sequence content of gaps in gorGor3 by gap size and Susie3 status (open or closed) estimated by liftover of gap-flanking sequence in gorGor3 to GRCh38.
Segmental duplication as defined by human GRCh38 (WGAC) and gorilla sequence read depth (*17*).
Exons defined by human RefSeq annotation.

Table S13. Summary of gorilla transcript and RNA-seq alignments.

	RNA-seq Reads (%) Aligned		Transcripts (%) Aligned*	
Data Type	**gorGor3**	**Susie3**	**gorGor3**	**Susie3**
ESTs lymphoblastoid cell lines (13,951,363)	9,423,184 (67.5)	10,280,665 (73.7)	NA	NA
iPS RNA-seq/transcripts (32,637,775/30,511)	23,503,226 (72.0)	24,615,781 (75.4)	21,672 (71.0)	25,042 (82.1)
lncRNA (48,350,271)	37,766,213 (78.1)	38,294,411 (79.2)	NA	NA
Gorilla GenBank mRNA (294)	NA	NA	229 (77.8)	270 (91.8)
Human GENCODE mRNA v22 (93,526)	NA	NA	71,349 (76.3)	89,234 (95.4)

Summary of alignment results for ESTs, Trinity-assembled RNA-seq from iPS cells, lncRNA, mRNA from GenBank, and human GENCODE (v23 Basic) transcripts.
ESTs: SRA ERR218142, ERR21813
iPS transcripts assembled using Trinity assembler from: SRA SRR976177, SRR976178, SRR976179, SRR976180, SRR976181, SRR976182
lncRNA: SRA SRR649365
*Aligned = >90% of length

Table S15. Gorilla satellite repeat content.

	Contigs		Reads			
Satellite Type	**Count**	**Proportion**	**Count**	**Proportion**	**Reads/ contigs**	**Proportion of aligned sequence bases (224.4 Gbp)**
pCht7*	151,162,172	0.507485	8,996,019,042	0.399857	59.51	0.0400928
Sat-1_TS	85,252,475	0.286212	6,539,002,079	0.290647	76.70	0.0291426
ALRY-MAJOR_PT	49,767,226	0.167080	3,569,946,748	0.158678	71.73	0.0159103
BSR/Beta	3,858,704	0.012955	518,010,568	0.023025	134.24	0.0023086
SAR	2,957,219	0.009928	740,597,983	0.032918	250.44	0.0033006
HSATII	917,792	0.003081	38,003,444	0.001689	41.41	0.0001694
ALR/Alpha	842,973	0.002830	1,518,093,631	0.067476	1,800.88	0.0067657
SATR1	694,655	0.002332	48,404,792	0.002152	69.68	0.0002157
SATR2	298,757	0.001003	19,180,627	0.000853	64.20	0.0000855
(GAATG)n	295,423	0.000992	9,696,209	0.000431	32.82	0.0000432
(CATTC)n	288,614	0.000969	11,384,931	0.000506	39.45	0.0000507
SST1	219,671	0.000737	82,467,316	0.003666	375.41	0.0003675
HSAT4	188,371	0.000632	11,803,382	0.000525	62.66	0.0000526
REP522	146,988	0.000493	15,814,757	0.000703	107.59	0.0000705
GSATX	143,201	0.000481	12,045,742	0.000535	84.12	0.0000537
HSATI	123,231	0.000414	35,544,551	0.001580	288.44	0.0001584
CER	118,260	0.000397	8,478,664	0.000377	71.70	0.0000378
GSATII	106,164	0.000356	10,355,042	0.000460	97.54	0.0000461
GSAT	100,566	0.000338	5,942,029	0.000264	59.09	0.0000265
ALRY-MINOR_PT	86,602	0.000291	252,530,762	0.011225	2,915.99	0.0011255
MSR1	86,097	0.000289	4,158,263	0.000185	48.30	0.0000185
D20S16	75,433	0.000253	3,499,008	0.000156	46.39	0.0000156
ACRO1	62,021	0.000208	39,837,620	0.001771	642.32	0.0001775
TAR1	31,759	0.000107	3,517,736	0.000156	110.76	0.0000157
HSAT5	22,862	0.000077	1,396,290	0.000062	61.07	0.0000062
LSAU	12,447	0.000042	1,879,780	0.000084	151.02	0.0000084
HSAT6	3,986	0.000013	344,978	0.000015	86.55	0.0000015
SUBTEL_sa	1,361	0.000005	158,013	0.000007	116.10	0.0000007
Total	297,865,030	1.000000	22,498,113,987	1.000000		0.1002680
Median					80.41	

Satellite content as defined by Tandem Repeats Finder (*29*) and RepeatMasker (*28*) of assembled contigs and underlying reads.

Reads/contigs indicates the read depth of particular satellite within assembled contigs, which was used to estimate total satellite proportion in genome.

*Listed in RepeatMasker database as PTPCHT7.

Table S21. Number of inversions for which the same repeat content was present at both breakpoints.

Repeat	Observed	Expected
DNA	6	24
LINE	158	150
LTR	48	62
Satellite	2	17
Simple_repeat	22	8
SINE	122	93

Table S26. Total counts and bases of gorilla SVs for population genetic and functional categories.

Category	Variants	Bases
All variants	117,410	86,512,881
Fixed variants (n=6 gorillas supporting variant)	84,804	71,760,348
Variants absent from human (n=1)	60,329	34,999,675
Variants absent from chimp (n=1)	63,254	35,775,578
Gorilla-specific variants	44,146	23,646,417
Fixed and gorilla-specific variants	23,083	15,316,860
Variants affecting genes	2,205	2,471,973
Variants affecting regulatory regions	10,466	12,263,519
Variants affecting genes or regulatory regions	12,196	13,539,040
lincRNA exons	66	353,692
UTRs	184	290,581
noncoding exons in protein-coding genes	96	119,600
Promoters (H3K4me3)	572	1,154,800
Enhancers (H3K27Ac)	1,236	1,834,152
Protein-coding exons	46	203,433
DNase clusters with high support	250	683,894
Fetal DNase clusters	862	2,101,632
Fixed gorilla-specific variants affecting genes	392	801,751
Fixed gorilla-specific variants affecting regulatory regions	2,151	3,518,968
Fixed gorilla-specific variants affecting functional regions	2,450	3,853,719

Table S27. Enrichment of functional types affected by fixed GSVs, including genes (coding and noncoding) and regulatory regions (DNase clusters, promoters defined by H3K4me3 signals, and enhancers defined by H3K27Ac signals).

Category	Genomic bases	Genomic events	SV events	SV-affected bases	Proportion of bases affected	Enrichment[a]
lincRNA	5,997,599	16,754	66	66,735	0.011	5.57&
UTR	45,169,980	93,519	184	568,753	0.013	0.84*
noncoding_in_protein_coding[b]	29,589,040	90,960	96	141,595	0.005	0.49*
H3K4me3	73,263,202	156,254	572	519,176	0.007	0.29*
H3K27Ac	135,946,456	399,661	1,236	926,265	0.007	0.15*
protein_coding	31,282,047	191,170	46	46,407	0.001	0.14*
high_support_DNase_cluster	50,300,310	129,821	250	110,440	0.002	0.13*
fetal_CNS_DNase_cluster	152,318,161	655,462	862	366,826	0.002	0.05*

Enrichment is calculated as the proportion of all functional bases affected by SVs divided by the proportion of the genome that is functional. Values greater than 1 indicate an enrichment relative to the genomic density of a category while values less than 1 indicate a depletion. Regulatory category names are indicated in italics. Events represent exons for genes and high-quality peaks of regulatory elements with the entire event and its bases reported if a variant intersects any part of the event.

a) Enrichment of functional bases affected by SVs calculated as the proportion of genomic bases affected by SVs divided by the proportion of bases in the genome annotated for each functional category. A denominator of 3.0 Gbp was used to calculate this latter genomic proportion.

b) GENCODE annotation of a noncoding transcript for an otherwise protein coding gene.

&) $p = 0.93155$

*) $p < 10\text{-}6$

Table S33. Coco SNV calls after reads overlapping heterozygous positions (4,761,776) gorGor3 were remapped to both assemblies. Over 500K heterozygous positions in gorGor3 were lost in Susie3.

Assembly	Number of heterozygous SNVs	Number of homozygous non-reference SNVs
gorGor3	4,118,705	781,431
Susie3	3,571,820	770,623

References and Notes

1. H. Y. K. Lam, M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, M. Snyder, Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2012). Medline doi:10.1038/nbt.2065

2. J. Rogers, R. A. Gibbs, Comparative primate genomics: Emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* **15**, 347–359 (2014). Medline doi:10.1038/nrg3707

3. M. J. P. Chaisson, R. K. Wilson, E. E. Eichler, Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015). Medline doi:10.1038/nrg3933

4. A. Scally, J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, S. McCarthy, S. H. Montgomery, P. C. Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, Q. Ayub, E. V. Ball, K. Beal, B. J. Bradley, Y. Chen, C. M. Clee, S. Fitzgerald, T. A. Graves, Y. Gu, P. Heath, A. Heger, E. Karakoc, A. Kolb-Kokocinski, G. K. Laird, G. Lunter, S. Meader, M. Mort, J. C. Mullikin, K. Munch, T. D. O'Connor, A. D. Phillips, J. Prado-Martinez, A. S. Rogers, S. Sajjadian, D. Schmidt, K. Shaw, J. T. Simpson, P. D. Stenson, D. J. Turner, L. Vigilant, A. J. Vilella, W. Whitener, B. Zhu, D. N. Cooper, P. de Jong, E. T. Dermitzakis, E. E. Eichler, P. Flicek, N. Goldman, N. I. Mundy, Z. Ning, D. T. Odom, C. P. Ponting, M. A. Quail, O. A. Ryder, S. M. Searle, W. C. Warren, R. K. Wilson, M. H. Schierup, J. Rogers, C. Tyler-Smith, R. Durbin, Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012). Medline doi:10.1038/nature10842

5. L. Carbone, R. Alan Harris, S. Gnerre, K. R. Veeramah, B. Lorente-Galdos, J. Huddleston, T. J. Meyer, J. Herrero, C. Roos, B. Aken, F. Anaclerio, N. Archidiacono, C. Baker, D. Barrell, M. A. Batzer, K. Beal, A. Blancher, C. L. Bohrson, M. Brameier, M. S. Campbell, O. Capozzi, C. Casola, G. Chiatante, A. Cree, A. Damert, P. J. de Jong, L. Dumas, M. Fernandez-Callejo, P. Flicek, N. V. Fuchs, I. Gut, M. Gut, M. W. Hahn, J. Hernandez-Rodriguez, L. D. W. Hillier, R. Hubley, B. Ianc, Z. Izsvák, N. G. Jablonski, L. M. Johnstone, A. Karimpour-Fard, M. K. Konkel, D. Kostka, N. H. Lazar, S. L. Lee, L. R. Lewis, Y. Liu, D. P. Locke, S. Mallick, F. L. Mendez, M. Muffato, L. V. Nazareth, K. A. Nevonen, M. O'Bleness, C. Ochis, D. T. Odom, K. S. Pollard, J. Quilez, D. Reich, M. Rocchi, G. G. Schumann, S. Searle, J. M. Sikela, G. Skollar, A. Smit, K. Sonmez, B. Hallers, E. Terhune, G. W. C. Thomas, B. Ullmer, M. Ventura, J. A. Walker, J. D. Wall, L. Walter, M. C. Ward, S. J. Wheelan, C. W. Whelan, S. White, L. J. Wilhelm, A. E. Woerner, M. Yandell, B. Zhu, M. F. Hammer, T. Marques-Bonet, E. E. Eichler, L. Fulton, C. Fronick, D. M. Muzny, W. C. Warren, K. C. Worley, J. Rogers, R. K. Wilson, R. A. Gibbs, Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**, 195–201 (2014). Medline doi:10.1038/nature13679

6. K. Berlin, S. Koren, C. S. Chin, J. P. Drake, J. M. Landolin, A. M. Phillippy, Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015). Medline doi:10.1038/nbt.3238

7. M. Pendleton, R. Sebra, A. W. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stütz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M. H. Fritz, H. Cao, A. Cohain, G.

Deikus, R. E. Durrett, S. C. Blanchard, R. Altman, C. S. Chin, Y. Guo, E. E. Paxinos, J. O. Korbel, R. B. Darnell, W. R. McCombie, P. Y. Kwok, C. E. Mason, E. E. Schadt, A. Bashir, Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015). Medline doi:10.1038/nmeth.3454

8. C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, J. Korlach, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013). Medline doi:10.1038/nmeth.2474

9. N. J. Royle, D. M. Baird, A. J. Jeffreys, A subterminal satellite located adjacent to telomeres in chimpanzees is absent from the human genome. *Nat. Genet.* **6**, 52–56 (1994). Medline doi:10.1038/ng0194-52

10. J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, A. Cagan, C. Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F. Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A. Tishkoff, J. C. Mullikin, R. K. Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H. Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup, C. Hvilsom, A. M. Andrés, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E. Eichler, T. Marques-Bonet, Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013). Medline doi:10.1038/nature12228

11. N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, K. D. Pruitt, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44** (D1), D733–D745 (2016). Medline doi:10.1093/nar/gkv1189

12. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference

human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). Medline doi:10.1101/gr.135350.111

13. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008). Medline doi:10.1093/bioinformatics/btn013

14. A. Siepel, M. Diekhans, B. Brejová, L. Langton, M. Stevens, C. L. Comstock, C. Davis, B. Ewing, S. Oommen, C. Lau, H. C. Yu, J. Li, B. A. Roe, P. Green, D. S. Gerhard, G. Temple, D. Haussler, M. R. Brent, Targeted discovery of novel human exons by comparative genomics. *Genome Res.* **17**, 1763–1773 (2007). Medline doi:10.1101/gr.7128207

15. J. Zhu, J. Z. Sanborn, M. Diekhans, C. B. Lowe, T. H. Pringle, D. Haussler, Comparative genomics search for losses of long-established genes on the human lineage. *PLOS Comput. Biol.* **3**, e247 (2007). Medline doi:10.1371/journal.pcbi.0030247

16. M. Ventura, C. R. Catacchio, C. Alkan, T. Marques-Bonet, S. Sajjadian, T. A. Graves, F. Hormozdiari, A. Navarro, M. Malig, C. Baker, C. Lee, E. H. Turner, L. Chen, J. M. Kidd, N. Archidiacono, J. Shendure, R. K. Wilson, E. E. Eichler, Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res.* **21**, 1640–1649 (2011). Medline doi:10.1101/gr.124461.111

17. P. H. Sudmant, J. Huddleston, C. R. Catacchio, M. Malig, L. W. Hillier, C. Baker, K. Mohajeri, I. Kondova, R. E. Bontrop, S. Persengiev, F. Antonacci, M. Ventura, J. Prado-Martinez, T. Marques-Bonet, E. E. Eichler, Great Ape Genome Project, Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373–1382 (2013). Medline doi:10.1101/gr.158543.113

18. C. T. Yohn, Z. Jiang, S. D. McGrath, K. E. Hayden, P. Khaitovich, M. E. Johnson, M. Y. Eichler, J. D. McPherson, S. Zhao, S. Pääbo, E. E. Eichler, Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLOS Biol.* **3**, e110 (2005). Medline doi:10.1371/journal.pbio.0030110

19. N. Polavarapu, N. J. Bowen, J. F. McDonald, Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biol.* **7**, R51 (2006). Medline doi:10.1186/gb-2006-7-6-r51

20. Y. Xue, J. Prado-Martinez, P. H. Sudmant, V. Narasimhan, Q. Ayub, M. Szpak, P. Frandsen, Y. Chen, B. Yngvadottir, D. N. Cooper, M. de Manuel, J. Hernandez-Rodriguez, I. Lobon, H. R. Siegismund, L. Pagani, M. A. Quail, C. Hvilsom, A. Mudakikwa, E. E. Eichler, M. R. Cranfield, T. Marques-Bonet, C. Tyler-Smith, A. Scally, Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**, 242–245 (2015). Medline doi:10.1126/science.aaa3952

21. S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, D. B. Goldstein, Genic intolerance to functional variation and the interpretation of personal genomes. *PLOS Genet.* **9**, e1003709 (2013). Medline doi:10.1371/journal.pgen.1003709

22. A. M. Little, P. Parham, Polymorphism and evolution of HLA class I and II genes and molecules. *Rev. Immunogenet.* **1**, 105–123 (1999). Medline

23. E. W. Myers, Efficient local alignment discovery amongst noisy long reads. *Lect. Notes Comput. Sci.* **8701**, 52–67 (2014). doi:10.1007/978-3-662-44753-6_5

24. E. W. Myers, The fragment assembly string graph. *Bioinformatics* **21** (Suppl 2), ii79–ii85 (2005). Medline doi:10.1093/bioinformatics/bti1114

25. P. A. Pevzner, H. Tang, G. Tesler, De novo repeat classification and fragment assembly. *Genome Res.* **14**, 1786–1796 (2004). Medline doi:10.1101/gr.2395204

26. D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grützner, S. Bergmann, R. Nielsen, S. Pääbo, H. Kaessmann, The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011). Medline doi:10.1038/nature10532

27. A. Necsulea, M. Soumillon, M. Warnefors, A. Liechti, T. Daish, U. Zeller, J. C. Baker, F. Grützner, H. Kaessmann, The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014). Medline doi:10.1038/nature12943

28. A. Smit, R. Hubley, P. Green, RepeatMasker Open-3.0 (1996); available at www.repeatmasker.org.

29. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999). Medline doi:10.1093/nar/27.2.573

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

31. J. D. Parsons, Miropeats: Graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995). Medline

32. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, http://arxiv.org/abs/1303.3997 (2013).

33. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing, http://arxiv.org/abs/1207.3907 (2012).

34. B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, D. Haussler, Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011). Medline doi:10.1101/gr.123356.111

35. J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, E. E. Eichler, Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002). Medline doi:10.1126/science.1072047

36. A. M. Phillippy, M. C. Schatz, M. Pop, Genome assembly forensics: Finding the elusive mis-assembly. *Genome Biol.* **9**, R55 (2008). Medline doi:10.1186/gb-2008-9-3-r55

37. M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011). Medline doi:10.1093/bioinformatics/btq683

38. D. Gordon, P. Green, Consed: A graphical editor for next-generation sequencing. *Bioinformatics* **29**, 2936–2937 (2013). Medline doi:10.1093/bioinformatics/btt515

39. A. L. Delcher, A. Phillippy, J. Carlton, S. L. Salzberg, Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002). Medline doi:10.1093/nar/30.11.2478

40. J. J. Yunis, O. Prakash, The origin of man: A chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982). Medline doi:10.1126/science.7063861

41. R. Stanyon, M. Rocchi, O. Capozzi, R. Roberto, D. Misceo, M. Ventura, M. F. Cardone, F. Bigoni, N. Archidiacono, Primate chromosome evolution: Ancestral karyotypes, marker order and neocentromeres. *Chromosome Res.* **16**, 17–39 (2008). Medline doi:10.1007/s10577-007-1209-z

42. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011). Medline doi:10.1038/nbt.1883

43. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). Medline doi:10.1093/bioinformatics/bts635

44. W. J. Kent, BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002). Medline doi:10.1101/gr.229202. Article published online before March 2002

45. T. Y. Tan, C. T. Gordon, K. A. Miller, D. J. Amor, P. G. Farlie, YPEL1 overexpression in early avian craniofacial mesenchyme causes mandibular dysmorphogenesis by up-regulating apoptosis. *Dev. Dyn.* **244**, 1022–1030 (2015). Medline doi:10.1002/dvdy.24299

46. M. J. P. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J. M. Landolin, J. A. Stamatoyannopoulos, M. W. Hunkapiller, J. Korlach, E. E. Eichler, Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015). Medline doi:10.1038/nature13907

47. W. Xu, E. Beutler, The characterization of gene mutations for human glucose phosphate isomerase deficiency associated with chronic hemolytic anemia. *J. Clin. Invest.* **94**, 2326–2329 (1994). Medline doi:10.1172/JCI117597

48. A. Kirov, D. Kacer, B. A. Conley, C. P. H. Vary, I. Prudovsky, AHNAK2 Participates in the Stress-Induced Nonclassical FGF1 Secretion Pathway. *J. Cell. Biochem.* **116**, 1522–1531 (2015). Medline doi:10.1002/jcb.25047

49. A. H. Horakova, S. C. Moseley, C. R. McLaughlin, D. C. Tremblay, B. P. Chadwick, The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. *Hum. Mol. Genet.* **21**, 4367–4377 (2012). Medline doi:10.1093/hmg/dds270

50. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001). Medline doi:10.1101/gr.GR-1871R

51. P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, E. E. Eichler; 1000 Genomes Project, Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010). Medline doi:10.1126/science.1197005

52. F. Hormozdiari, M. K. Konkel, J. Prado-Martinez, G. Chiatante, I. H. Herraez, J. A. Walker, B. Nelson, C. Alkan, P. H. Sudmant, J. Huddleston, C. R. Catacchio, A. Ko, M. Malig, C. Baker, T. Marques-Bonet, M. Ventura, M. A. Batzer, E. E. Eichler, Great Ape Genome Project, Rates and patterns of great ape retrotransposition. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13457–13462 (2013). Medline doi:10.1073/pnas.1310914110

53. M. J. Chaisson, B. J. Raphael, P. A. Pevzner, Microinversions in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19824–19829 (2006). Medline doi:10.1073/pnas.0603984103

54. J. Lee, K. Han, T. J. Meyer, H.-S. Kim, M. A. Batzer, Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLOS ONE* **3**, e4047 (2008). Medline doi:10.1371/journal.pone.0004047

55. H.-H. Chou, H. Takematsu, S. Diaz, J. Iber, E. Nickerson, K. L. Wright, E. A. Muchmore, D. L. Nelson, S. T. Warren, A. Varki, A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11751–11756 (1998). Medline doi:10.1073/pnas.95.20.11751

56. T. N. Turner, F. Hormozdiari, M. H. Duyzend, S. A. McClymont, P. W. Hook, I. Iossifov, A. Raja, C. Baker, K. Hoekzema, H. A. Stessman, M. C. Zody, B. J. Nelson, J. Huddleston, R. Sandstrom, J. D. Smith, D. Hanna, J. M. Swanson, E. M. Faustman, M. J. Bamshad, J. Stamatoyannopoulos, D. A. Nickerson, A. S. McCallion, R. Darnell, E. E. Eichler, Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016). Medline doi:10.1016/j.ajhg.2015.11.023

57. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009). Medline doi:10.1093/nar/gkn923

58. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009). Medline doi:10.1038/nprot.2008.211

59. A. Favorov, L. Mularoni, L. M. Cope, Y. Medvedeva, A. A. Mironov, V. J. Makeev, S. J. Wheelan, Exploring massive, genome scale datasets with the GenometriCorr package. *PLOS Comput. Biol.* **8**, e1002529 (2012). Medline doi:10.1371/journal.pcbi.1002529

60. R. S. Harris, thesis, ProQuest (2007).

61. C. Bromberg, Sequencher: Version 4.1. 2. Gene Codes Corporation. (1995).

62. H. Li, Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014). Medline doi:10.1093/bioinformatics/btu356

63. K. F. McManus, J. L. Kelley, S. Song, K. R. Veeramah, A. E. Woerner, L. S. Stevison, O. A. Ryder, G. Ape Genome Project, J. M. Kidd, J. D. Wall, C. D. Bustamante, M. F. Hammer, Inference of gorilla demographic and selective history from whole-genome sequence data. *Mol. Biol. Evol.* **32**, 600–612 (2015). Medline doi:10.1093/molbev/msu394

64. F. H. Leendertz, S. Yumlu, G. Pauli, C. Boesch, E. Couacy-Hymann, L. Vigilant, S. Junglen, S. Schenk, H. Ellerbrok, A new Bacillus anthracis found in wild chimpanzees and a gorilla from West and Central Africa. *PLOS Pathog.* **2**, e8 (2006). Medline doi:10.1371/journal.ppat.0020008

65. P. J. Le Gouar, D. Vallet, L. David, M. Bermejo, S. Gatti, F. Levréro, E. J. Petit, N. Ménard, How Ebola impacts genetics of Western lowland gorilla populations. *PLOS ONE* **4**, e8375 (2009). Medline doi:10.1371/journal.pone.0008375

66. O. Thalmann, A. Fischer, F. Lankester, S. Pääbo, L. Vigilant, The complex evolutionary history of gorillas: Insights from genomic data. *Mol. Biol. Evol.* **24**, 146–158 (2007). Medline doi:10.1093/molbev/msl160

67. T. C. S. and Analysis Consortium; Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005). Medline doi:10.1038/nature04072