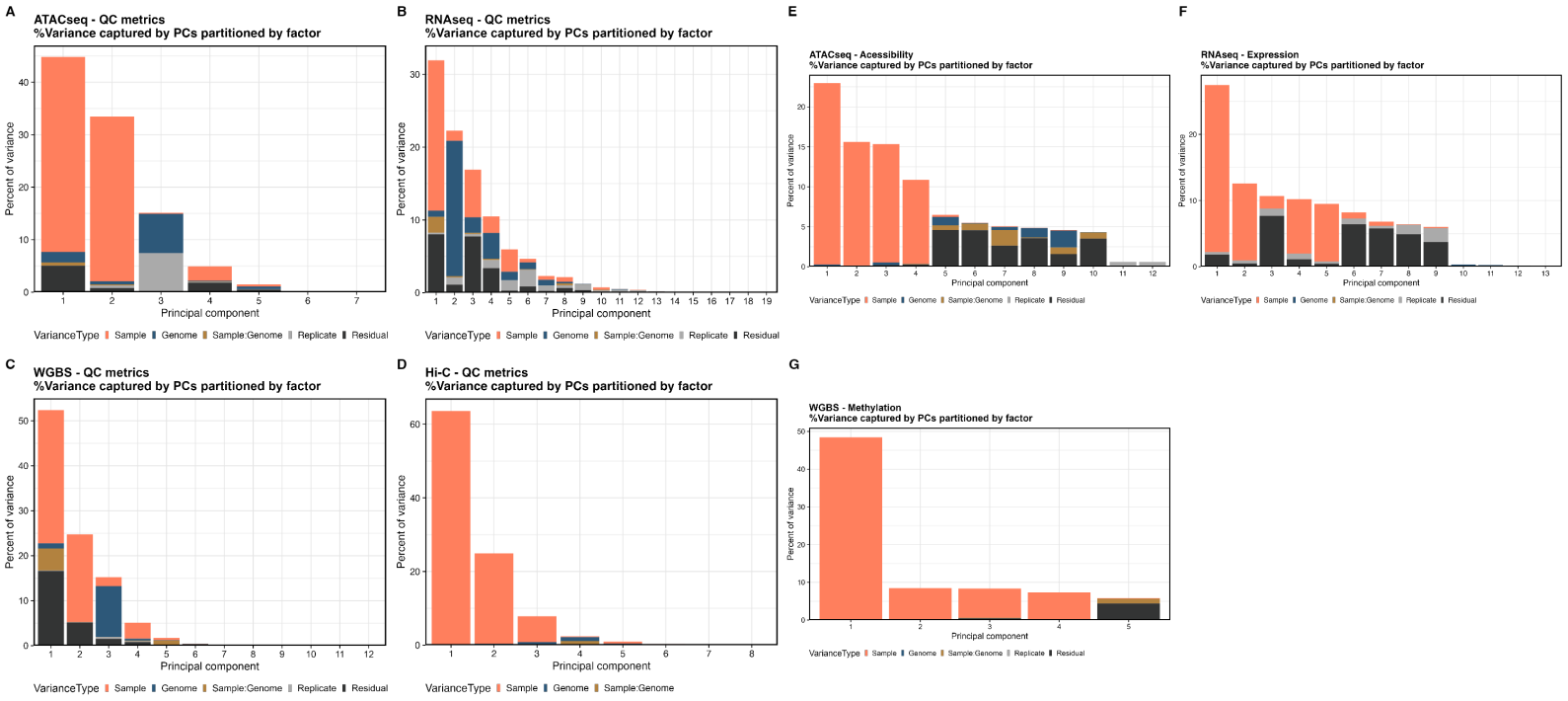
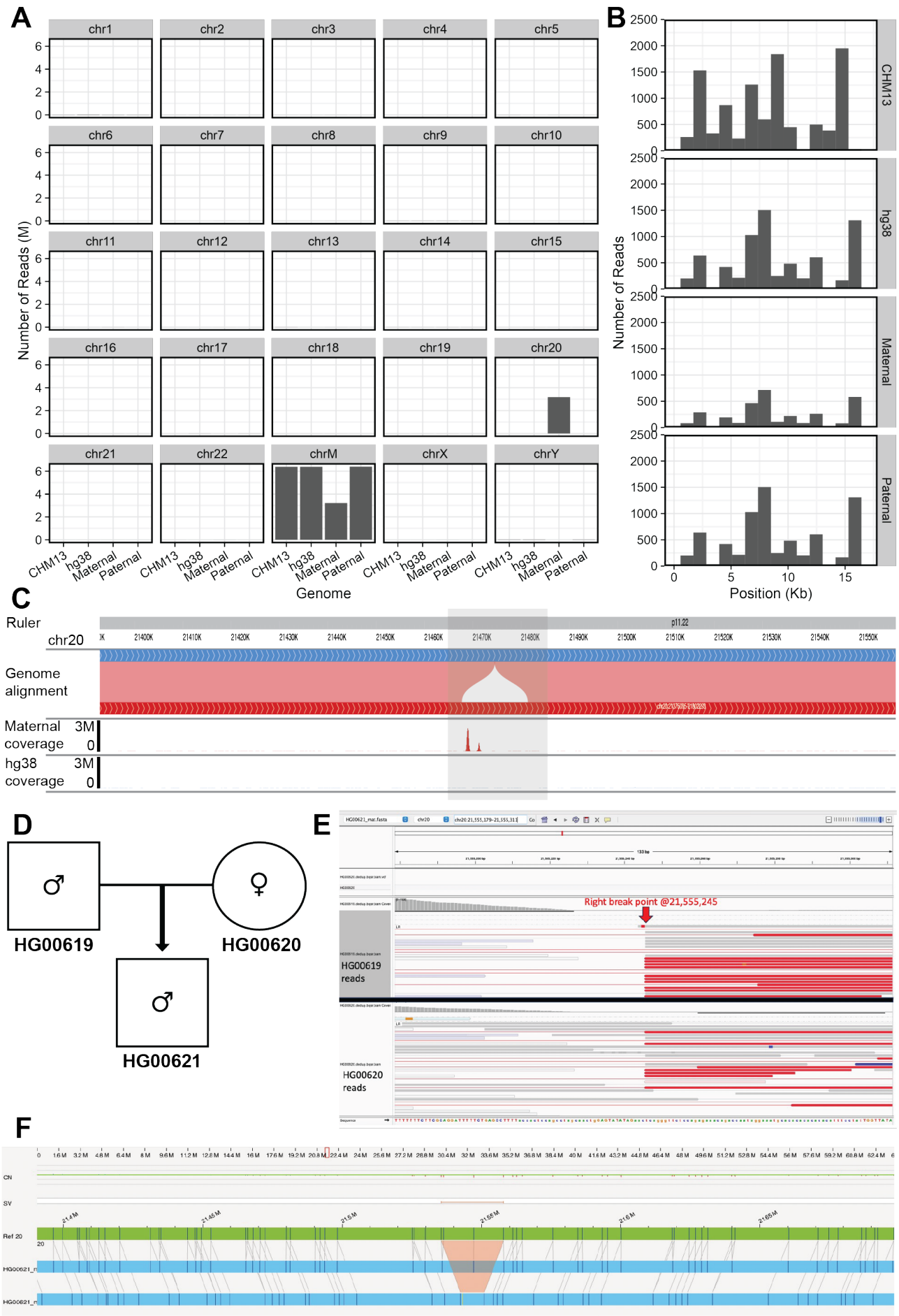
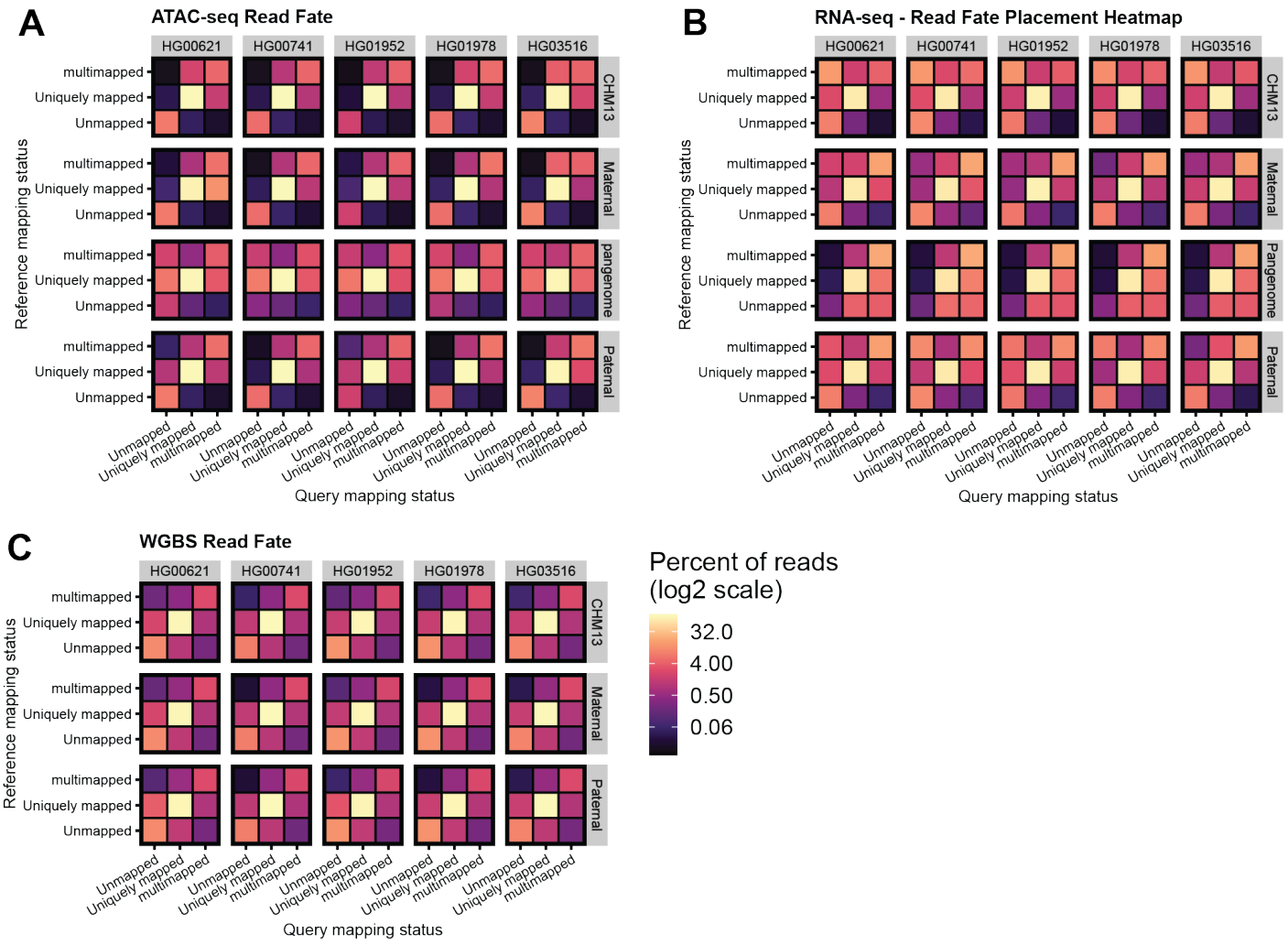


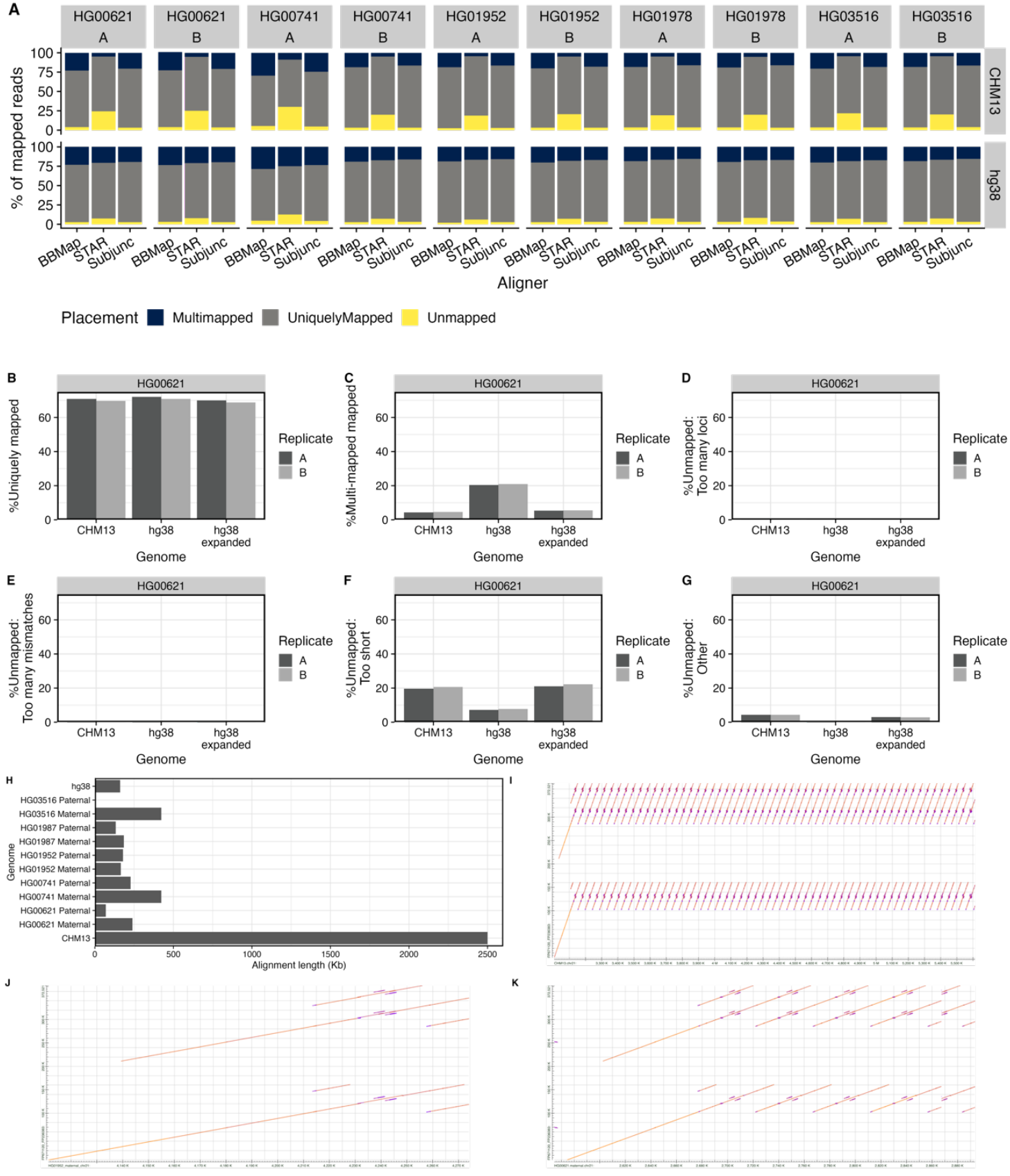
Supplementary Figure 1: Sample identity accounts for the largest fraction of accountable variation. Principal components analysis was performed for QC/alignment metrics (A-D) and Functional outputs (E-G) for ATAC-seq (A&E), RNA-seq (B&F), WGBS (C&G), and Hi-C (D) assays of genomic function. Partial eta squared estimates were calculated by MANOVA for each term along each principal component to partition the variances attributable to each term. Factors evaluated were Sample, Genome, Sample:Genome, and Replicate. The percent of variation attributable to each term for each principal component is shown.


A
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrM
chrX
chrY
Number of Reads (M)
CHM13
hg38
Maternal
Paternal
Genome
B
Number of Reads
Position (Kb)
C
Ruler
chr20
Genome alignment
Maternal coverage
hg38 coverage
D
HG00619
HG00620
HG00621
E
Right break point @21,555,245
HG00619 reads
HG00620 reads
F
CN
SV
Ref 20

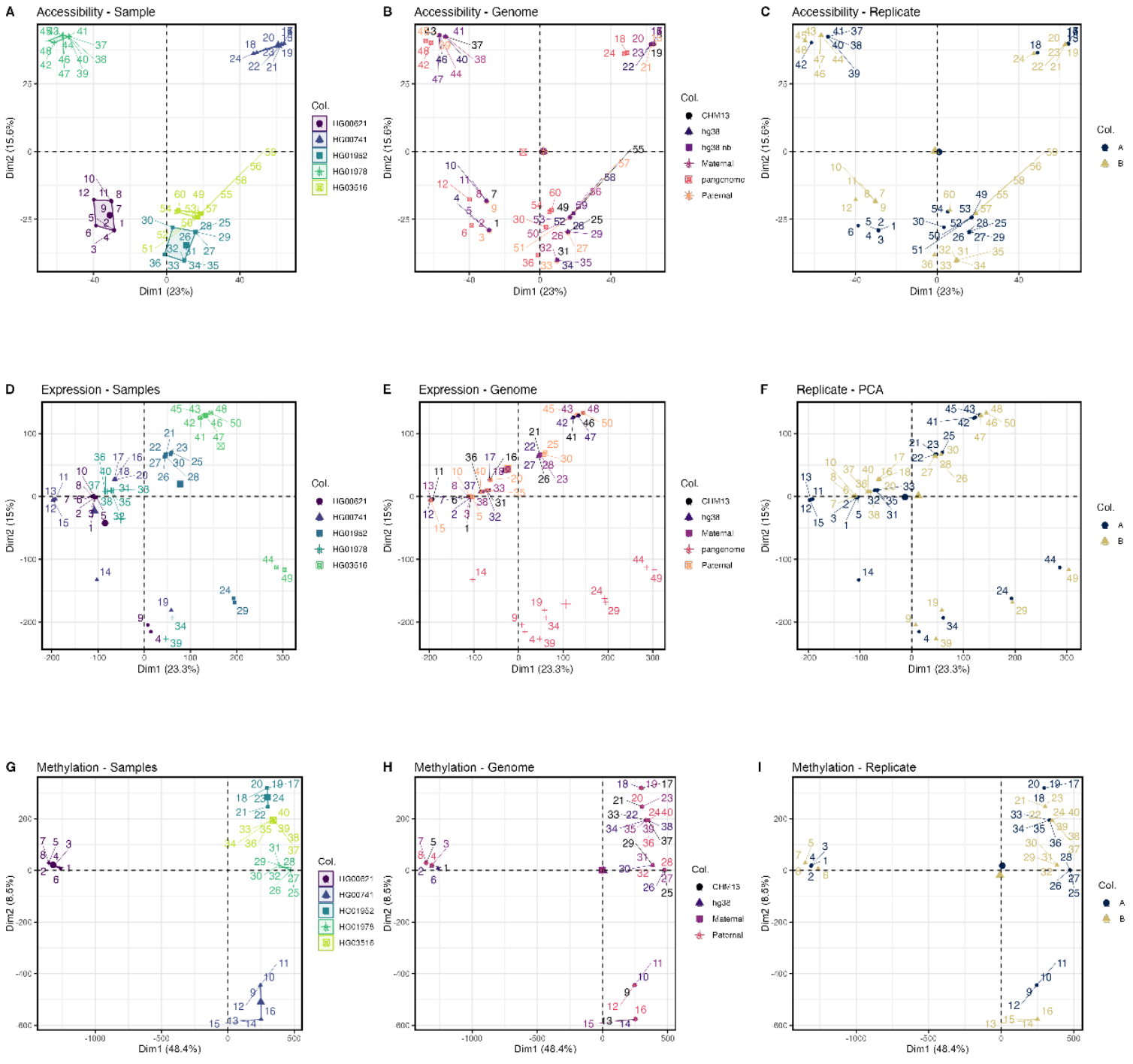
Supplementary Figure 2. A nuclear mitochondrial transfer (NuMT) into chromosome 20 alters the ATAC-seq %ChrM metric for HG00621 when the maternal assembly is used. Use of the maternal assembly decreases the fraction of uniquely mapped mitochondrial reads and increases the fraction of multi-mapped reads. A) Chromosomal distribution of reads that are uniquely mapped in hg38 but multi-mapped in the HG00621 maternal assembly. B) Read density across chrM for hg38, CHM13, the maternal assembly, and the paternal assembly. C) WashU Epigenome Browser view of the HG00621 maternal chr20 insertion site. Tracks show hg38 coordinates, the alignment-based relationship between hg38 and the HG00621 maternal assembly, ATAC-seq read density after alignment to the maternal assembly, and ATAC-seq read density after alignment to hg38. In the genome comparison track, blue denotes hg38 and red denotes the maternal assembly. BLAST identified the inserted sequence as a NuMT. D) Pedigree schematic showing HG00621 and the sequenced parents HG00619 and HG00620 used for validation. E) IGV view of parental whole-genome sequencing reads aligned to the HG00621 maternal chr20 assembly at the NuMT breakpoint. Reads from HG00620 span the breakpoint, whereas reads from HG00619 do not, indicating that the NuMT is maternally derived. F) Bionano optical map confirmation of the same region. The CN track denotes predicted copy number variation, the SV track denotes predicted structural variation, the green track shows the HG00621 maternal assembly, and the two lower tracks show the maternal and paternal Bionano haplotype maps.



Supplementary Figure 3. Breakdown per sample and replicate of read fates compared. For every read its fate is assigned as being uniquely mapped, multi-mapped, or unmapped. Its fate is compared for ATAC-seq (A), RNA-seq (B), and WGBS (C) assays. The comparisons are between each query genome {Maternal, Paternal, (where applicable) the draft pangenome, CHM13} relative to the reference genome {hg38}. The percent of reads with each pair of fates between query and reference are shown as the color value.

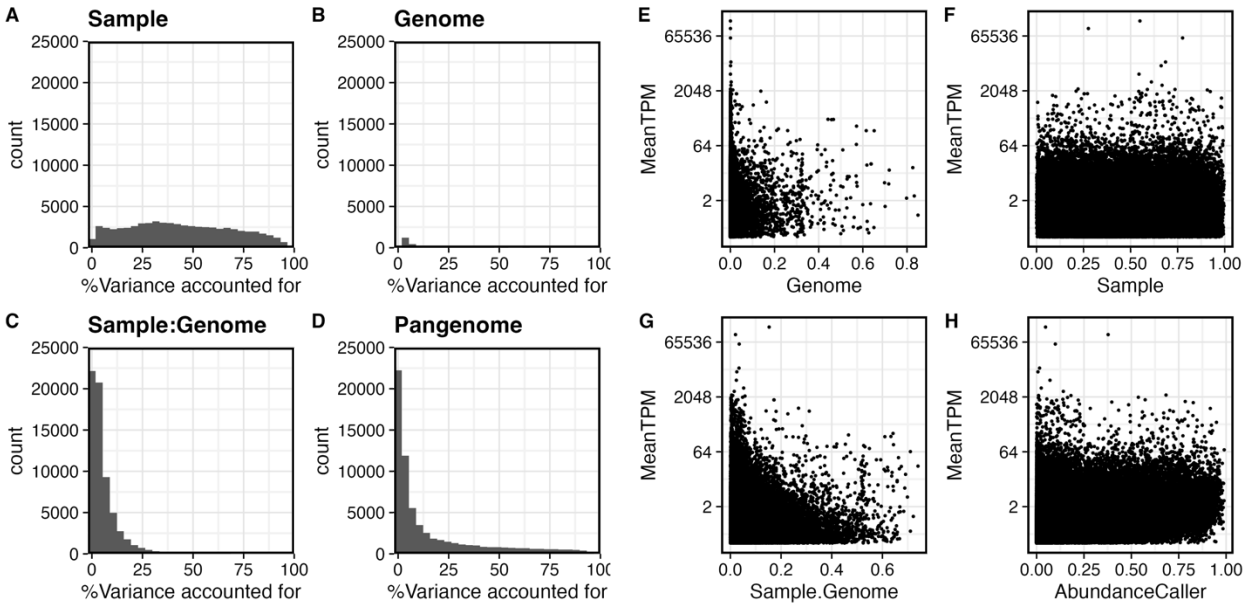

A
HG00621 A
HG00621 B
HG00741 A
HG00741 B
HG01952 A
HG01952 B
HG01978 A
HG01978 B
HG03516 A
HG03516 B
% of mapped reads
CHM13
hg38
BBMap
STAR
Subjunc
Aligner
Placement
Multimapped
UniquelyMapped
Unmapped
B
%Uniquely mapped
C
%Multi-mapped mapped
D
%Unmapped: Too many loci
E
%Unmapped: Too many mismatches
F
%Unmapped: Too short
G
%Unmapped: Other
Replicate
A
B
Genome
CHM13
hg38
hg38 expanded
H
hg38
HG03516 Paternal
HG03516 Maternal
HG01987 Paternal
HG01987 Maternal
HG01952 Paternal
HG01952 Maternal
HG00741 Paternal
HG00741 Maternal
HG00621 Paternal
HG00621 Maternal
CHM13
Alignment length (Kb)
I
J
K

Supplementary Figure 4. Evaluation of how fully resolved rDNA arrays of CHM13 affect mapping of RNA-seq data using seed-and-extend strategy (STAR) compared to seed-and-vote strategy (A). Samples aligned against CHM13 have a substantial increase in the number of unmapped reads which are otherwise multi-mapped in other assemblies (B-C). To isolate the effect of the rDNA expansion, the syntenic rDNA element in hg38 is expanded to match the number of copies in CHM13. STAR breaks down reason for failed alignment (D-G), showing the flagged reasons which increase with expansion of the rDNA array are in the “too short” and “other” class. Length of chr21 rDNA element alignments for each assembly against to hg38 (H). Number of repeats visualized by dot-matrix plot for CHM13 vs hg38 (I), HG01952 Maternal vs hg38 (J), and HG00621 Maternal (K).

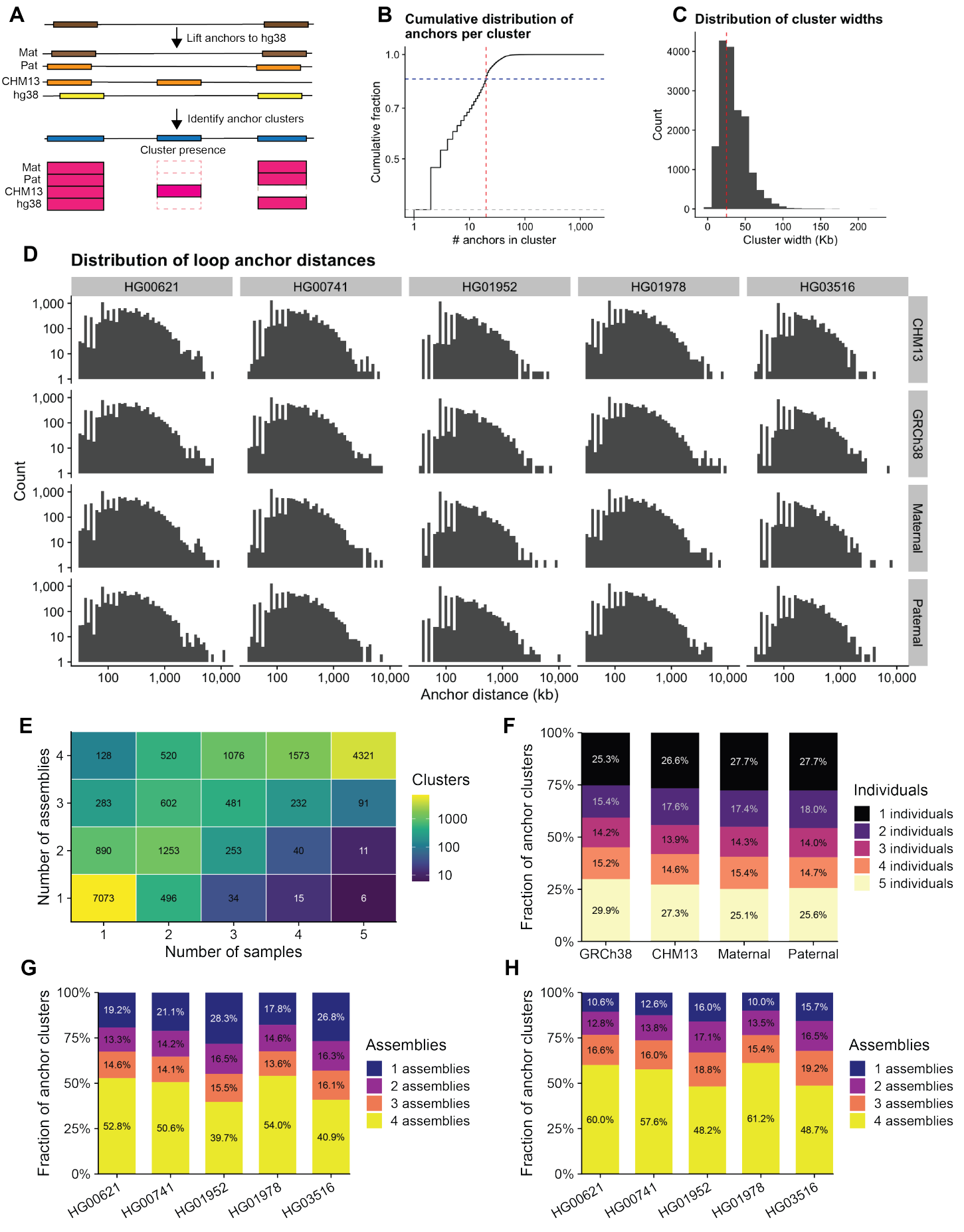


Supplementary Figure 5. Effects of genome choice on functional outputs. To evaluate genome-wide effects of genome choice on functional outputs, principal components

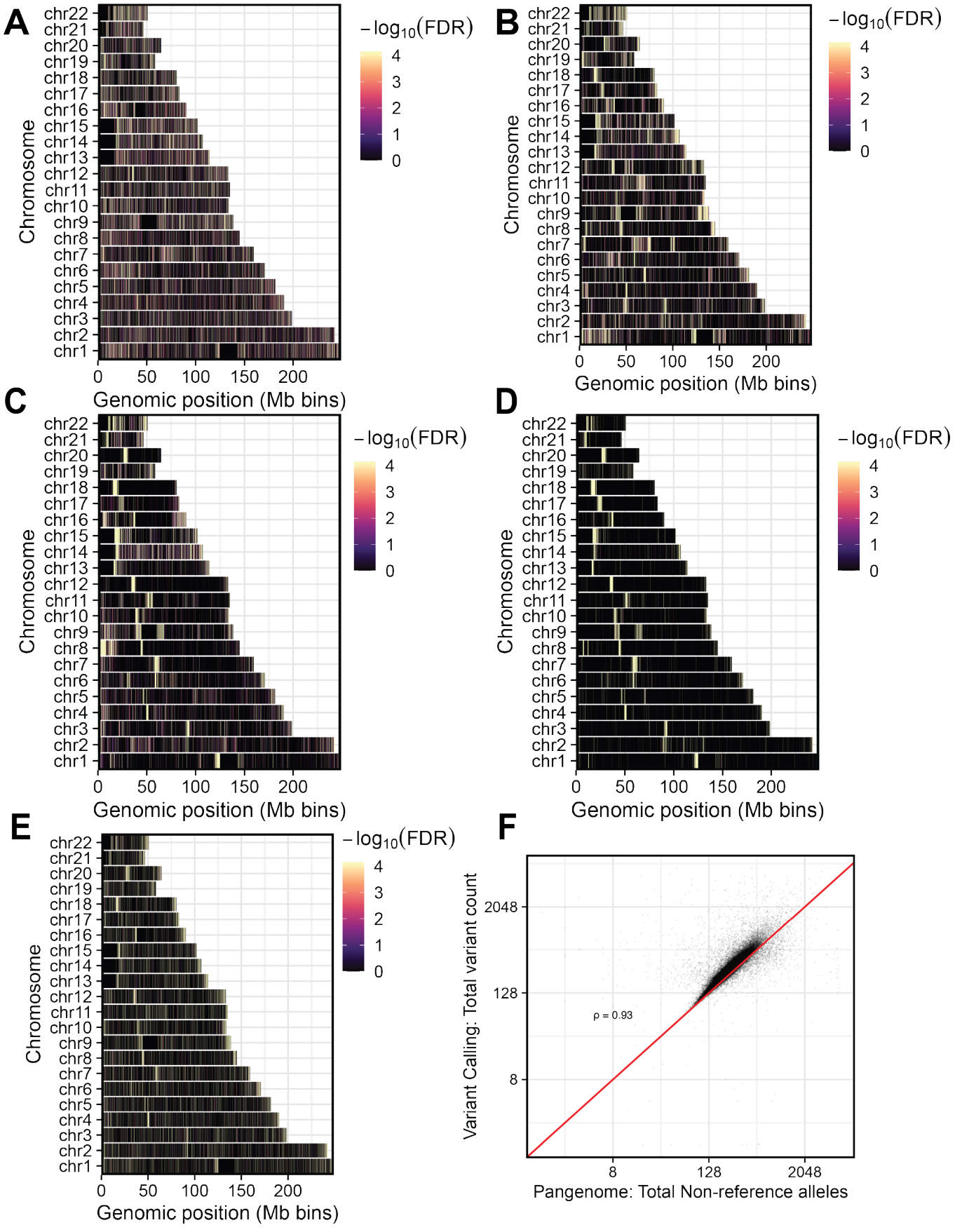
analysis was performed to dimensionally reduce functional outputs. This was done for ATAC-seq (A-C), RNA-seq (D-F), and WGBS (G-I) assays. Runs are grouped (color-coded) by sample identity (A, D, G), replicate (B, E, H), and genome choice (C, F, I). Individual runs are numbered and represented as smaller shapes; group means are represented as unnumbered large shapes. Sample identity accounts for the largest amount of genome-wide variation in functional outputs.



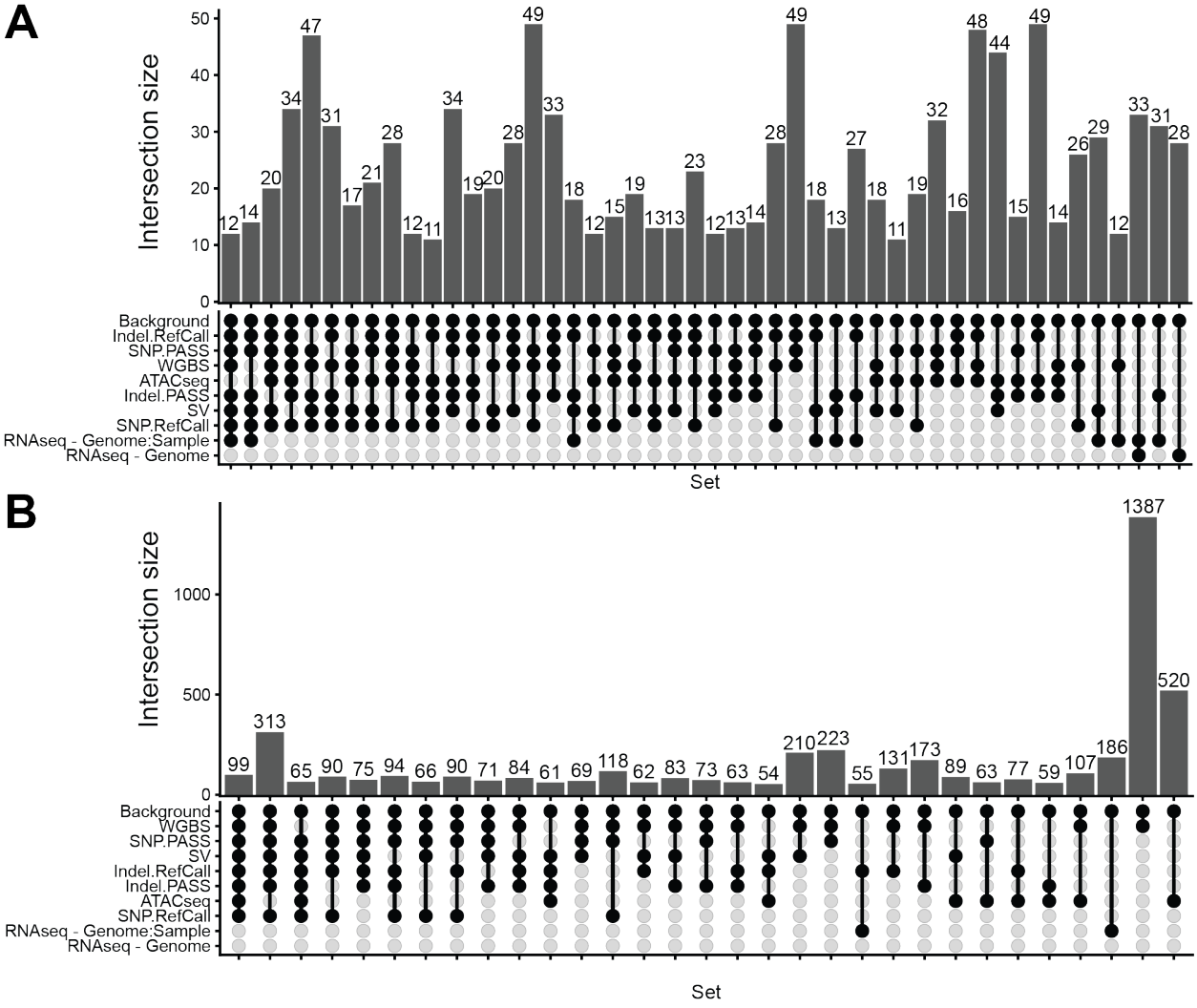
Supplementary Figure 6. Distribution and association with average expression of percent of variance in abundance estimates. For each transcript the analysis of variance is performed to estimate partial eta squared for each predictor term. The distribution of variance attributable to Sample (A), Genome (B), Sample:Genome (C), and Pangenome (D) terms. The average abundance (TPM) estimate for each transcript is calculated and compared to the fraction of variance attributable to predictor terms (E-H). Showing a trend towards lowly expressed genes being more susceptible to effects of genome choice.


A
Lift anchors to hg38
Mat
Pat
CHM13
hg38
Identify anchor clusters
Cluster presence
B
Cumulative distribution of anchors per cluster
Cumulative fraction
anchors in cluster
C
Distribution of cluster widths
Count
Cluster width (Kb)
D
Distribution of loop anchor distances
HG00621
HG00741
HG01952
HG01978
HG03516
CHM13
GRCh38
Maternal
Paternal
Count
Anchor distance (kb)
E
Number of assemblies
Number of samples
Clusters
F
Fraction of anchor clusters
Individuals
1 individuals
2 individuals
3 individuals
4 individuals
5 individuals
GRCh38
CHM13
Maternal
Paternal
G
Fraction of anchor clusters
Assemblies
1 assemblies
2 assemblies
3 assemblies
4 assemblies
H
Fraction of anchor clusters
Assemblies
1 assemblies
2 assemblies
3 assemblies
4 assemblies

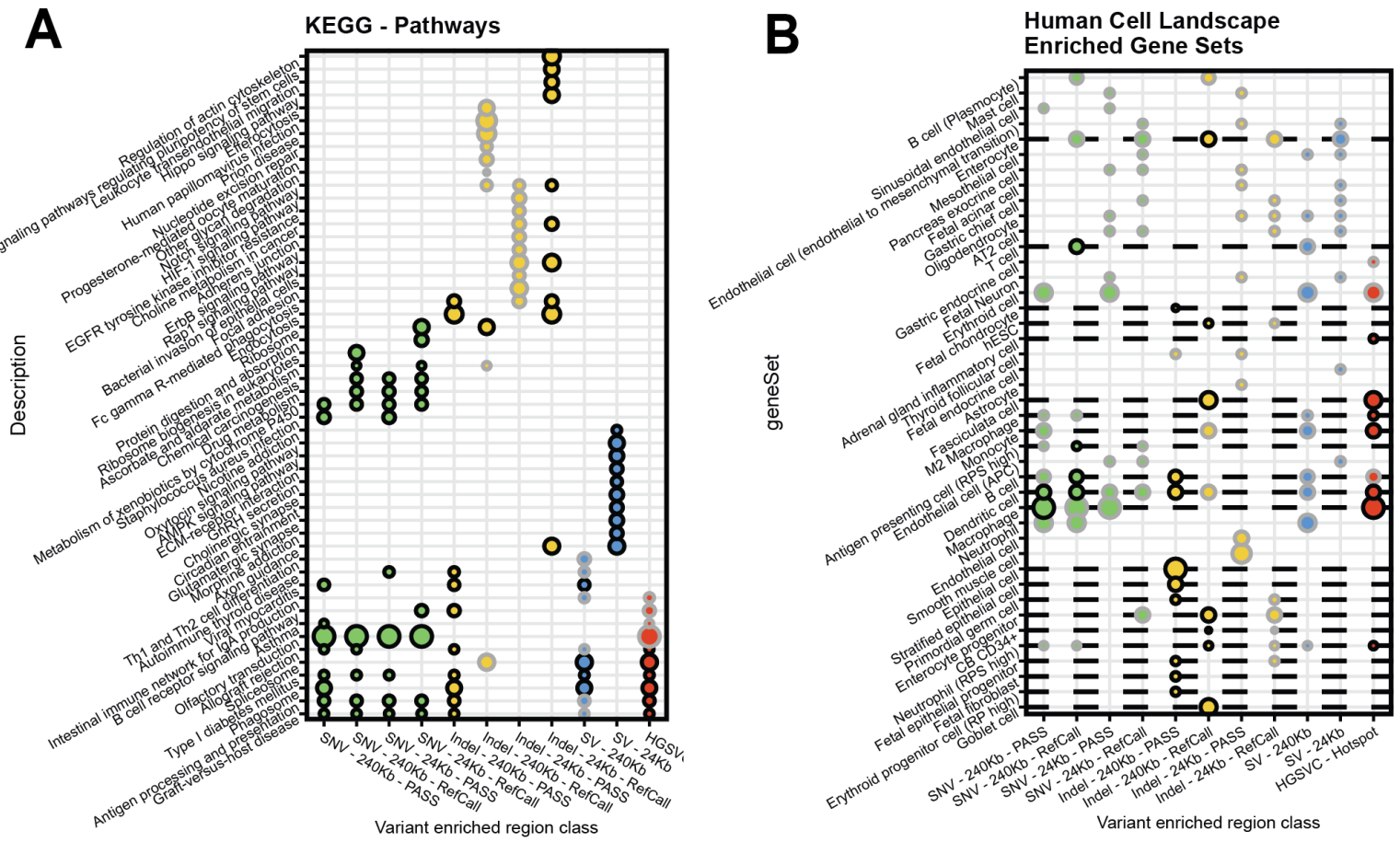
Supplementary Figure 7. Hi-C loop anchor clustering across assemblies and individuals. A) Workflow schematic of the analysis. Loop anchors were lifted to hg38 coordinates and grouped into anchor clusters defined by overlapping lifted anchor intervals on the same chromosome after merging with bedtools merge. Each merged region was assigned a unique cluster ID. Clusters were compared across four genome assembly classes (GRCh38, CHM13, maternal, paternal) and five individuals (HG00621, HG00741, HG01952, HG01978, HG03516). B) Cumulative distribution of anchors per cluster. The cumulative fraction of clusters is plotted against the number of anchors per cluster on log-log axes. The red dashed line marks 20 anchors, the threshold used to flag highly complex clusters; the blue dashed line marks the 85th percentile; and the grey dashed line marks the proportion of single-anchor clusters. C) Distribution of cluster widths. The x-axis shows cluster width in kilobases and the y-axis shows cluster count. The red dashed line at 25 kb marks the loop-anchor width before liftover. D) Distribution of genomic distances between paired loop anchors after liftover. Panels are faceted by genome assembly class (rows) and individual (columns). E) Relationship between the number of individuals and the number of assemblies represented within each anchor cluster. Tile color indicates the number of clusters, with counts shown in the tiles. Clusters containing a single anchor were excluded. F) Cluster sharing between individuals within each assembly class. G) Cluster sharing across assemblies within each individual. Clusters detected in only one assembly range from 17.8 to 28.3%. H) Cluster sharing across assemblies within each individual, restricted to clusters supported by at least one parental assembly. Clusters detected in only one assembly range from 10.0 to 16.0%.


A
B
C
D
E
F
Chromosome
chr22
chr21
chr20
chr19
chr18
chr17
chr16
chr15
chr14
chr13
chr12
chr11
chr10
chr9
chr8
chr7
chr6
chr5
chr4
chr3
chr2
chr1
0
50
100
150
200
Genomic position (Mb bins)
−log10(FDR)
4
3
2
1
0
Variant Calling: Total variant count
Pangenome: Total Non-reference alleles
2048
128
8
ρ = 0.93

Supplementary Figure 8. Identification of regions with elevated rates of variation relative to chromosome-constrained randomization null distributions of per-interval variant counts. PacBio HiFi long-read variant calls from 47 individuals in the draft human pangenome were used to quantify per-interval counts of indels passing filters (A), indels not passing filters (B), single-nucleotide variants passing filters (C), single-nucleotide variants not passing filters (D), and structural variants (E). The genome was partitioned into non-overlapping 24 kb intervals. For each variant class, enrichment of observed interval counts was assessed using one-tailed randomization tests based on 1,000 within-chromosome shuffles of variant positions, and p-values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure. Intervals were classified as enriched if they had false discovery rate (FDR) ≤ 0.005 and observed counts in the top 10% genome-wide for that variant class. In A-E, the x-axis shows genomic position in megabases and the y-axis shows chromosome. Colored intervals indicate enriched regions, with color intensity proportional to $-\log10(\mathrm{FDR})$ as shown in the scale bars; black intervals did not meet enrichment criteria. F) Relationship between the summed per-interval variant counts across classes and the corresponding number of non-reference alleles from the decomposed graph VCF. Points represent 24 kb intervals, and the red line indicates the identity line. Pearson's correlation coefficient is shown in the panel. Exact nominal and Benjamini-Hochberg-adjusted p-values for A-E, and the exact p-value for the correlation in F, are provided in the supplementary data file 8.

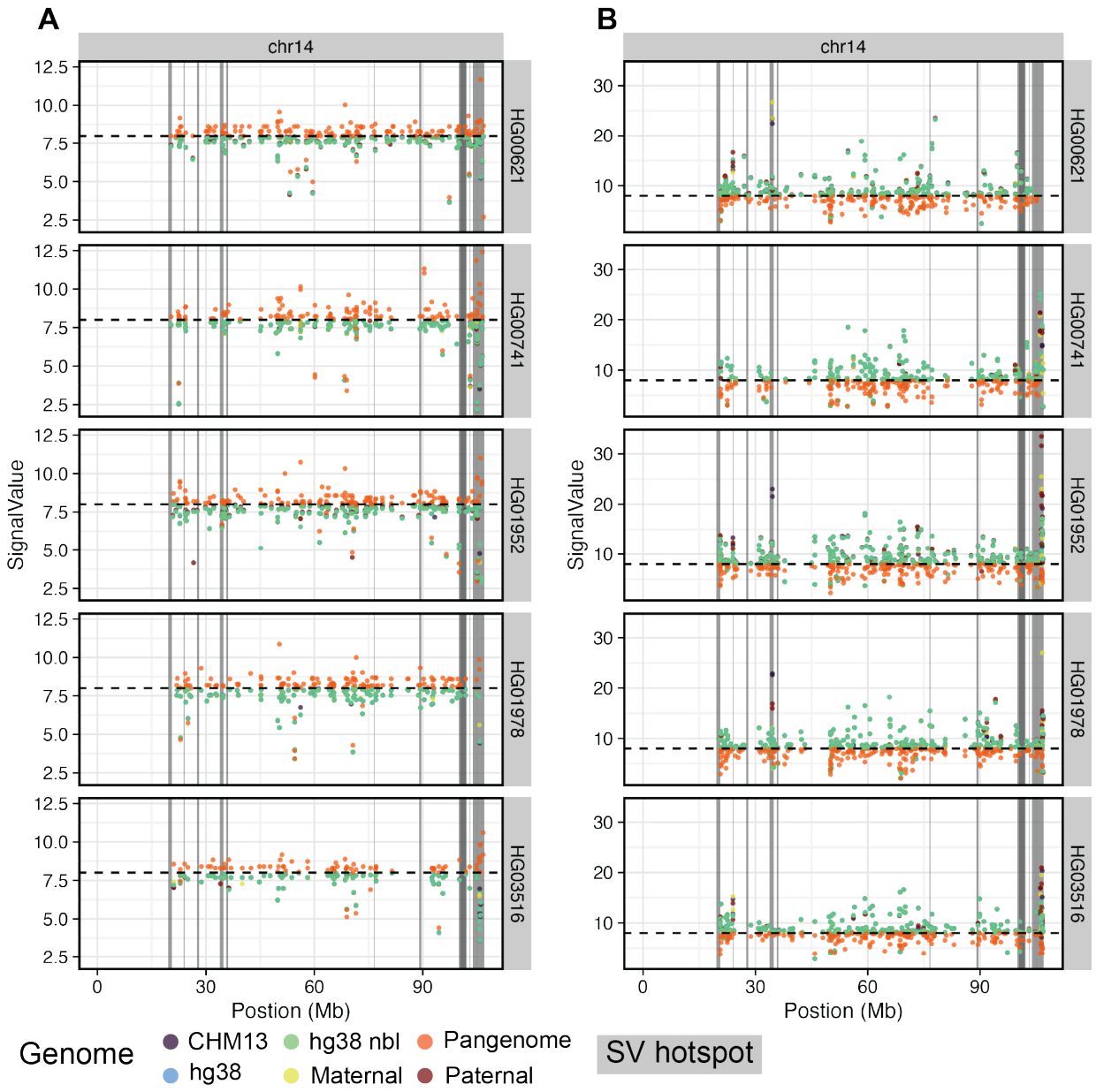


Supplementary Figure 9. Colocalization of genome choice effects and enrichment in regions of elevated variation. Features showing a genome choice effect on their functional estimate were assigned to the nearest gene, or in the case of transcripts, to their corresponding gene. Intersections among these genome choice-proximal genes across assay types are shown, together with their overlap within regions of elevated variation stratified by variant class. Intersections are grouped by size to aid interpretability. Panel A shows intersections containing more than 10 genes, highlighting moderately shared genome choice effects across assays. Panel B shows intersections containing more than 50 genes, representing highly shared, core genome choice-associated gene sets.



Supplementary Figure 10. Gene set enrichment of genes located in regions with elevated rates of genetic variation. To assess biological processes and cellular contexts that may be especially susceptible to genome choice effects beyond the five benchmark individuals analyzed here, overrepresentation analyses were performed in WebGestalt using genes within 24 kb and 240 kb intervals with elevated rates of variation as test gene sets and all annotated genes in the genome as the background gene set. Five databases were queried: GO Biological Process, GO Cellular Component, GO Molecular Function,

KEGG Pathway, and Human Phenotype Ontology. A) KEGG enrichment results for genes in regions with elevated variation. B) Human Cell Landscape enrichment results for the same gene sets. Points are organized by variant class and interval size as labeled on the x-axis. Point size is proportional to relative gene set size. Overrepresentation analysis was performed in WebGestalt using one-sided hypergeometric tests with false-discovery-rate correction. Exact nominal and FDR-adjusted p-values for all displayed terms are provided in the Source Data file.



Supplementary Figure 11. Analysis of ATAC-seq peak calling between linear and pangenomic methods on chr14. Peak calls using alignments surjected from the pangenome and alignments from linear assemblies were used to identify peaks of open chromatin. A substantial number of peaks after signal value thresholding were specific to (A) or specifically missing from (B) the analysis when using the draft pangenome. The location (x-axis) of all peaks within cross-assembly peak clusters for which the

pangenome does not call or exclusively call peaks for an individual are plotted per-chromosome against the pre-thresholding signal value (y-axis) of all peaks within said peak-cluster. Genome-wide these instances show that the peaks are call across assemblies for these sites, but there is a shift in the signal value across assemblies. Where the pangenome boosts signal in some places and dampens it in others.