Supplemental Material:

Integration of high-resolution promoter profiling assays reveals novel, cell type-specific transcription start sites across 115 human cell and tissue types

Jill E. Moore[1], Xiao-Ou Zhang[1], Shaimae I. Elhajjajy[1], Kaili Fan[1], Henry E. Pratt[1], Fairlie Reese[2], Ali Mortazavi[2], and Zhiping Weng*[1]

1 Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA..

2 Department of Developmental and Cell Biology, University of California Irvine, Irvine, CA, USA

* Correspondence should be addressed to ZW (zhiping.weng@umassmed.edu)

Table of Contents

Supplemental Tables (*external files*)

Supplemental Methods

Generating a collection of RAMPAGE rPeaks

Curating RAMPAGE experiments

As of September 1, 2020 there were 155 ENCODE3 RAMPAGE experiments at the ENCODE portal (https://www.encodeproject.org/search/?type=Experiment&status=released&perturbed=false&assay_title=RAMPAGE&award.rfa=ENCODE3&perturbed=true). From the portal, we downloaded RAMPAGE BAM alignment files, which contained reads mapped to the GRCh38/hg38 reference genome by the ENCODE Data Coordination Center using STAR (Dobin et al. 2013) (https://www.encodeproject.org/data-standards/rampage/). We then removed redundant reads as described in (Zhang et al. 2019), briefly summarized as follows. First, we properly aligned read pairs (R1 and R2 denote each mate of a read pair) with uniquely aligned R2 reads were collapsed with the same alignment coordinates and the identical 15-bp barcode at the 5´-end of R2 reads to remove PCR duplicates. We then pooled read pairs from biological replicates together after the PCR duplicate removal and created signal bigWig files of the 5' ends of R1 reads that we used for all subsequent signal quantifications (available for download on our companion site). Finally, we excluded all experiments with a non-redundancy fraction less than 0.25, which resulted in a final collection of 115 high quality RAMPAGE experiments (**Supplemental Table S1**).

Relevant script: rm_pcr.py

Calling RAMPAGE peaks

We called RAMPAGE peaks as described in (Zhang et al. 2019). Briefly, RAMPAGE peaks were clustered with the 5´-most base of aligned R1 reads using F-seq (Boyle et al. 2008) (parameter: feature length = 30 and fragment size = 0). For each peak, we identified a high-density region, which contained 80% of the reads in each original peak, and a summit, which was the genomic position with the highest number of R1 5' read ends.

Relevant script: call_peak.py

Filtering RAMPAGE peaks

When combining annotations from many experiments to build a TSS catalog, it is very important to start off with a set of high quality annotations; otherwise, noisy annotations will compound as more datasets are added and the quality of the catalog would suffer. For each RAMPAGE experiment, the Gingeras lab also performed a matching total RNA-seq experiment on the same biosample, which we used to filter RAMPAGE peaks. Using bigWigAverageOverBed, we calculated the total RNA-seq and RAMPAGE signals (column four of the resulting file, sum) across each RAMPAGE peak. We excluded peaks whose RNA-seq signals were greater than their RAMPAGE signals (i.e., peaks that fell below the x=y line, **Supplemental Fig. S1**). These filtered-out peaks predominantly overlapped annotated exons and may be due to cytosolic recapping. Finally, to further select for high-quality annotations, we only retained peaks with RPM (reads per million) > 2, which resulted in a set of 1,147,456 peaks across all experiments with an average of 9,978 per experiment (**Supplemental Table S1**).

Relevant script: 1_Peak-Filtering.sh

Generating representative RAMPAGE peaks

To generate representative RAMPAGE peaks (RAMPAGE rPeaks), we adapted the representative DNase I Hypersensitivity Site (rDHS) pipeline as described by the ENCODE Project Consortium (The ENCODE Project Consortium et al. 2020). First, to retain strand-specific information, we separated peaks based on DNA strand, and then clustered the strand-specific peaks across all 115 experiments using BEDtools *merge* (Quinlan and Hall 2010). For each cluster, we selected the peak with the highest RPKM (reads per kilobase per million) signal as the rPeak. All peaks that overlapped this rPeak—as defined by using BEDtools *intersect* with default parameters—were then removed. We iteratively repeated this process until all 1.1 M RAMPAGE peaks were represented by a collection of 80,157 non-overlapping rPeaks. To reduce false positives, we discarded all singleton rPeaks (i.e., rPeaks that represented only one experiment) unless they had an RPM > 5, resulting in a final set of 52,546 rPeaks. It is important to note that the 1.1 M individual peaks are a pooled set of peaks across the 115 experiments; many of these peaks directly overlap each other. For example, a TSS that is expressed in every experiment would contribute 115 peaks to the pooled set. Individual experiments have between 6.2k and 17k peaks. Our iterative process of consolidation resulted in a non-redundant set of rPeaks that serves as the anchor for systematic comparison between experiments, and we keep the peaks with sufficiently high signals in just one biosample.

Relevant script: 2_rPeak-Annotation.sh

Genomic context and enrichment

Determining genomic context

We used the following hierarchical approach to assign genomic contexts to annotations (including RAMPAGE rPeaks and FANTOM CAGE peaks). We used BEDtools *intersect* to determine overlapping features with overlap requirements as described below:

1) *TSS-overlapping*: rPeak overlapped an annotated TSS from GENCODEv31 basic annotations. Use default parameters for BEDtools *intersect*.
2) *TSS-Proximal*: rPeak fell within ± 500 bp of an annotated TSS from GENCODEv31 basic. Required at least 50% of RAMPAGE rPeak to overlap region (-f 0.5).
3) *Exon*: rPeak overlapped "exon" annotation from GENCODEv31 basic which include coding exons (CDS), exons of non-coding genes, and untranslated regions (UTRs). Required at least 50% of RAMPAGE rPeak to overlap exon (-f 0.5).
4) *Intron:* rPeak overlapped an annotated gene from GENCODEv31 basic but not an exon. Required at least 50% of RAMPAGE rPeak to overlap gene (-f 0.5).
5) *Intergenic:* all remaining rPeaks

Relevant script: Determine-Genomic-Context.sh

Assigning strand

1) *TSS-overlapping*: assign strand of the transcript the RAMPAGE rPeak overlaps. If the RAMPAGE rPeak overlaps TSSs on both strands, the strand matching the rPeak is assigned.
2) *Proximal*: assign strand of the transcript RAMPAGE rPeak falling within 500 bp. If the RAMPAGE rPeak overlaps TSSs on both strands, the strand matching the rPeak is assigned.
3) *Exon*: assign strand of the transcript containing the exon the rPeak overlaps. If the RAMPAGE rPeak overlaps exons on both strands, the strand matching the rPeak is assigned.
4) *Intron:* assign strand of the transcript the rPeak overlaps. If the RAMPAGE rPeak overlaps transcripts on both strands, the strand matching the rPeak is assigned.
5) *Intergenic:* assign strand of the closest transcript as determined by BEDtools *closest* using the TSS basic annotations. If the RAMPAGE rPeak is equally close to transcripts on both strands, the strand matching the rPeak is assigned.

Relevant script: Determine-Genomic-Context.sh

Determining genomic context enrichment

To determine the genomic background, we calculated the percentage of the GRCh38 genome comprising each of the annotations: (TSS: 0.0004%; TSS-proximal: 2.2%; Exon: 3.7%; Intron: 45.5%; Intergenic: 48.6%). We then determined the percentage of total rPeaks falling in each annotation and calculated fold enrichment.

Relevant script: Determine-Genomic-Context.sh

Boundary and summit analysis

For each rPeak, we calculated the median peak boundary, high-density boundary and summit variation for each peak that was represented. We did not include peaks that were selected as the rPeaks in this analysis.

Relevant script: Compare-Boundary-Variation.sh

UMAP

We performed two separate UMAP analyses: one using all 115 biosamples (**Supplemental Fig. S1F**) and one using the subset of all 87 tissue samples (**Fig. 1G**). For each biosample, we calculated the RPKM (reads per kilobase per million) at each rPeak. We then combined these results to create two input matrices, 52,547 by 115 and 52,547 by 87, respectively, where each row is a RAMPAGE rPeak and each column is a biosample. For each entry of the matrix we took the log10 of each entry and normalized each row using sklearnStandardScaler. We then implemented the UMAP algorithm using the Python UMAP-learn package with n_neighbors = 10 and default values for the remaining parameters.

Relevant script: UMAP-Analysis.sh

Comparisons with other transcription annotations

Comparison with CAGE peaks

We downloaded CAGE peaks and quantifications from the FANTOM consortium:

- Peaks: https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/CAGE_peaks/hg38_fair+new_CAGE_peaks_phase1and2.bed.gz
- Quantifications: https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/CAGE_peaks_expression/hg38_fair+new_CAGE_peaks_phase1and2_tpm.osc.txt.gz

To compare the overall concordance of peak collections, we intersected the entire collection of CAGE peaks with the entire collection of RAMPAGE peaks using BEDtools *intersect* with the requirement that at least 25% of the CAGE peak overlapped the RAMPAGE peak and the peaks fell on the same strand.

Relevant script: RAMPAGE-CAGE-All-Peak-Comparison.sh

To extract peaks active in K562 and GM12878, we selected all peaks with an average TPM (transcripts per million) > 2 across the three surveyed replicates (columns 563-565 for K562 and columns 171-173 for GM12878). We compared these peaks with RAMPAGE rPeaks with RPM > 2 in K562 and GM12878, respectively, using BEDtools *intersect*, requiring overlapping peaks to be on the same strand and overlap a minimum of 25% of the CAGE peak.

Relevant script: Extract-CAGE-Peaks.sh

Comparison with CAGE and NET-CAGE enhancers

We downloaded CAGE & NET-CAGE enhancers from (Hirabayashi et al. 2019) (https://fantom.gsc.riken.jp/5/suppl/Hirabayashi_et_al_2019/data/Supplementary_Data_1_Human_FANTOM-NET_enhancers.bed.gz). We lifted the annotations from the hg19 to the hg38 genome using the UCSC liftOver tool. We intersected the enhancers with RAMPAGE rPeaks using default BEDtools *intersect* parameters We then stratified the enhancer annotations as to whether they were detected by CAGE (N=65,423) or only NET-CAGE (N=20,363) and calculated the total percent overlap with the RAMPAGE rPeaks.

Relevant script: Compare-FANTOM-Enhancers.sh

Comparison with PacBio long-read RNA-seq data

We downloaded the following BAM files from the ENCODE project data portal: ENCFF546DOT and ENCFF709YES for K562 and ENCFF247TLH, ENCFF431IOE, ENCFF520MMC, and ENCFF626GWM for GM12878. We merged and sorted BAM files for each cell type, split reads by genomic strand, and used BEDtools *bamtobed* to extract the 5' ends of reads.

Relevant script: Format-PacBio-Data.sh

We used BEDtools *intersect* with default parameters to intersect PacBio 5' read ends with RAMPAGE and CAGE peaks. To only count strand matching intersections, RAMPAGE and CAGE peaks were first split by strand and then intersected with 5' ends on the same strand.

Relevant script: Compare-TSS-Annotations.sh

Comparison with GRO-cap signal

We downloaded the following GRO-cap signal files from GEO under accessions GSM1480321 and GSM1480323 for K562 and GM12878, respectively:

- GSM1480321_K562_GROcap_wTAP_plus.bigWig
- GSM1480321_K562_GROcap_wTAP_minus.bigWig
- GSM1480323_GM12878_GROcap_wTAP_plus.bigWig
- GSM1480323_GM12878_GROcap_wTAP_minus.bigWig

To calculate average signal at RAMPAGE rPeaks, CAGE peaks, and PacBio 5' ends, we lifted down the 1 bp summits or read ends to the hg19 genome using the UCSC liftOver tool (Kuhn et al. 2013) with default parameters. We then set region width to a uniform 50 bp centered on the peak summits or 5' ends and, using the UCSC bigWigAverageOverBed function (Kuhn et al. 2013), calculated the average signal across each region.

To determine a signal threshold for high GRO-cap signal, we first randomly selected 500k 50 bp genomic regions and calculated their average GRO-cap signal. We then selected the 99.5th percentile as the threshold for *high signal* which was 0.06 in K562 and 0.08 in GM12878, respectively.

Relevant script: Calculate-GROcap-Signal.sh

Comparison with GRO-cap peaks

We downloaded the supplemental data file from (Core et al. 2014) which contained GRO-cap peaks calls for K562 and GM12878:

- https://static-content.springer.com/esm/art%3A10.1038%2Fng.3142/MediaObjects/41588_2014_BFng3142_MOESM78_ESM.zip
    - tss_all_k562.bed
    - tss_all_gm12878.bed

We intersected GRO-cap peaks with RAMPAGE rPeaks, CAGE peaks, and RAMPAGE PacBio reads using default BEDtools *intersect* parameters and requiring annotations to be on the same strand (-s flag).

Relevant script: Compare-TSS-Annotations.sh

From the same supplemental data file we also obtained sets of paired GRO-cap peaks in GM12878 and K562 that were classified by stability (tss_SS_gm12878_plus.bed, tss_SU_gm12878_plus.bed, tss_US_gm12878_plus.bed, tss_UU_gm12878_plus.bed, etc). We lifted peaks from the hg19 to the hg38 genome using the UCSC liftOver tool. We then intersected the hg38 peak sets with RAMPAGE rPeaks with these peaks using default BEDtools *intersect* parameters and requiring annotations to be on the same strand (-s flag). We then calculated the overall percentage of each category that overlapped the rPeaks.

Relevant script: Compare-Stability-Overlap.sh

Comparison of GENCODE covered genes

We first set peak width to a uniform 100 bp centered around each peak summit or 5' read end and then intersected these regions with annotated TSSs of GENCODE 31 genes using default BEDtools *intersect* and requiring annotations to be on the same strand (-s flag). We performed Gene Ontology analysis using PantherDB's online database (Mi et al. 2017). We first performed this analysis for the entire sets of RAMPAGE and CAGE peaks, then for peaks and PacBio 5' read ends in K562 and GM12878 cells.

Relevant scripts: RAMPAGE-CAGE-Gene-Comparison.sh, RAMPAGE-CAGE-PacBio-Gene-Comparison.sh

Aggregate transcriptomic signals at RAMPAGE rPeaks

Using 1 bp bins, we calculated the average CAGE, PacBio, and GRO-cap signals, along a 4 kb window centered across the summits of RAMPAGE rPeaks active in either K562 or GM12878 cells. For CAGE we converted 5' end *ctss.bed files on the RIKEN portal to bigWigs (https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/basic/human.cell_line.hCAGE/). Similarly we selected the 5' end of PacBio reads and converted this to bigWig signal. For GRO-cap, we lifted down the rPeaks 1 bp summits to the hg19 genome using the UCSC liftOver tool (Kuhn et al. 2013) with default parameters before calculating the average signal. In all three assays, we calculated strand specific signal for each rPeaks which was then properly orientated and averaged.

Relevant scripts: Run-Aggregate-DirSignal.sh

Assigning RAMPAGE rPeaks to Genes

Curating verified GENCODE TSSs, verified unannotated TSSs, unannotated transcript TSSs and local transcription rPeaks

We developed the following computational workflow to link RAMPAGE rPeaks with genes, which is detailed in **Supplemental Fig. S3A**. Briefly, based on the genetic context of the rPeak and the location of its supporting 3' reads, we assigned the rPeak into one of six categories:

1) *Verified GENCODE TSS*: rPeak overlaps an annotated GENCODE TSS and its 3' read ends overlap a downstream exon.
2) *Verified unannotated TSS*: rPeak does not overlap an annotated GENCODE TSS (i.e., rPeak is either TSS-proximal, exonic, intronic, or intergenic) and its 3' read ends overlap a downstream exon.
3) *Candidate GENCODE TSS*: rPeak overlaps a TSS, first exon or is TSS-proximal to either a single exon transcript, or a transcript with a first exon greater than 500 nt.
4) *Unannotated transcript TSS*: rPeak is supported by reads with 3' ends that do not overlap an annotated GENCODE exon.
5) *Local transcription*: rPeak is supported by reads that span less than 1 kb or map to the first exon of the transcript.
6) *Discard*: We discarded all rPeaks that overlapped exons that were not the first exon of a transcript or only supported by reads that spanned more than 500 kb.

Relevant scripts: Analyze-RAMPAGE-Read-Mates.sh
Assign-Gene-Links.sh

Overlap of novel transcripts with lncRNAs

We downloaded lncRNA annotations (lncRNA_LncBook_GRCh38_9.28.gtf) from lncBook (Ma et al. 2019) (https://bigd.big.ac.cn/lncbook/index) and extracted annotated TSSs. Then, we intersected RAMPAGE rPeaks using default BEDtools *intersect* parameters and requiring annotations to be on the same strand (-s flag). We also calculated the overlap of lncBook TSSs with 500k 100 bp random genomic regions generated using BEDtools *random*.

Relevant script: Intersect-lncBook-TSSs.sh

Scanning transcripts for open reading frames

We intersected our RAMPAGE rPeaks with PacBio reads to delineate produced transcripts and then scanned these transcripts using NCBI's ORFfinder tool (Wheeler et al. 2003). Stratifying by our rPeak TSS assignment, we calculated the number of uniquely identified ORFs for each rPeak

Relevant script: Scan-For-ORFs.sh

Characterizing biosample profiles of RAMPAGE TSSs

We selected all GENCODE genes with at least one linked RAMPAGE rPeak (either verified GENCODE or verified unannotated). For each gene, we calculated two metrics:

1) The total number of biosamples in which the gene was expressed. This was a union of all the tissues for which any linked RAMPAGE rPeak was expressed with a maximum value of 115
2) Total biosample space which was a concatenated list of all biosamples for which any linked RAMPAGE rPeak was expressed. The maximum value would be the number of linked rPeaks x 115.

To evaluate the cell type-specificity of gene and transcript expression, we compared the number of active biosamples (RPM > 2) for each RAMPAGE rPeak and its linked gene.

To determine whether the transcripts resulting from rPeak TSSs correspond to major or minor isoforms, we calculated the total number of biosamples for which the rPeak has an RPM > 2 and then divided this by the total biosamples space (2) of its linked gene.

Relevant script: Calculate-Percentage-Tissue-Space.sh

Comparison of GENCODE and verified TSSs

Generating sets of matched GENCODE TSSs

We first selected all GENCODE genes that did not have a single annotated TSS overlapping a RAMPAGE rPeak. Of these, we then selected all genes with a RAMPAGE-verified TSS. Because of the no overlapping requirement, these RAMPAGE-verified TSSs were either TSS-proximal, exonic, intronic, or intergenic. The GENCODE-annotated TSSs of these genes served as the matched GENCODE TSS set. In total, we curated 8,391 GENCODE-matched TSSs to compare with 6,243 RAMPAGE-verified TSSs. We also curated K562-specific annotations by selecting all RAMPAGE-verified TSSs with an RPM > 2 in K562 and their matched GENCODE TSSs, resulting in a set of 1,768 GENCODE-matched TSSs to compare with 966 RAMPAGE-verified TSSs in K562 cells.

Unlike the RAMPAGE-verified TSSs, GENCODE TSSs were only 1 bp in width; therefore, to eliminate biases due to region width, we generated uniform 100 bp regions centered on either RAMPAGE -verified TSS summits or GENCODE TSSs, respectively.

Relevant script: Compare-Verified-TSS-GENCODE.sh

Overlap of RAMPAGE-verified and matched GENCODE TSSs with ENCODE cCREs

We downloaded cell type-agnostic cCREs and K562-specific cCREs from the ENCODE SCREEN database (screen.encodeproject.org). For the K562 cCREs, we filtered out "Low-DNase" cCREs, which are regulatory regions deemed inactive in the cell type. Using BEDtools *intersect* with default parameters and the -u flag to count unique elements, we intersected the uniform 100 bp sized TSS regions (as described above) with the cell type-agnostic cCREs. We repeated this analysis using the K562 cCREs and the uniform 100 bp sized K562 regions.

Relevant script: Compare-Verified-TSS-GENCODE.sh

Overlap of RAMPAGE-verified and matched GENCODE TSSs with GTEx eQTLs

We downloaded eQTLs from the GTEx database (GTEx_Analysis_v8_eQTL.tar), aggregated across all *signif_variant_gene_pairs.txt files, and reformatted the results into BED format. Using BEDtools *intersect* with default parameters and the -u flag to count unique elements, we intersected the uniform 100 bp sized TSS regions (as described above) with the eQTL BED file.

Relevant script: Compare-Verified-TSS-GENCODE.sh

Overlap of K562 RAMPAGE-verified and matched GENCODE TSSs with SuRE peaks

We downloaded SuRE peaks from the Supplementary Data section of van Arensbergen *et al. (van Arensbergen et al. 2017)* (SuRE-peaks_K562.45.55_raw_sep_globalLambda.annotated_LP160616.txt).

We reformatted this file into BED format, and lifted the regions from the hg19 genome up to the hg38 genome using UCSC's liftOver tool with default parameters and the hg19ToHg38.over.chain chain file. We then intersected the uniform 100 bp sized TSS regions from K562 (as described above) with hg38 K562 SuRE peaks using BEDtools *intersect* with default parameters and the -u flag to count the number of unique regions that overlapped.

Relevant script: Compare-Verified-TSS-GENCODE.sh

Aggregate epigenomic signals at RAMPAGE-verified and matched GENCODE TSSs

Using 1 bp bins, we calculated the average DNase I-seq, and H3K4me3, H3K27ac and Pol II ChIP-seq signals, along a 4 kb window centered across the RAMPAGE-verified rPeak summit or matched GENCODE TSS, respectively, accounting for strand orientation. We used the following uniformly processed bigWig files from the ENCODE portal:

Signal	**Experiment**	**BigWig Accession**
DNase I-seq	ENCSR921NMD	ENCFF971AHO
H3K4me3	ENCSR000DWD	ENCFF847JMY
H3K27ac	ENCSR000AKP	ENCFF779QTH
Pol II	ENCSR388QZF	ENCFF321FZQ

Relevant script: Run-Aggregate-Signal.sh

Overlap of RAMPAGE-verified and matched GENCODE TSSs with PacBio 5' read ends

We intersected the uniform 100 bp sized TSS regions from K562 (as described above) with K562 PacBio 5' read ends (generated from Format-PacBio-Data.sh, see above for more details) using BEDtools *intersect* with the -c flag—which counts the number of overlapping entries—and requiring genomic strands to match.

Relevant script: Compare-Verified-TSS-GENCODE.sh

Conservation of RAMPAGE-verified and matched GENCODE TSSs

We calculated the average phastCons conservation (100way vertebrate) across the uniform 100 bp regions (as described above) using UCSC's bigWigAverageOverBed (Kent et al. 2010).

We lifted the uniform 100 bp sized TSS regions (as described above) over to the mm10 genome using UCSC's liftOver tool (Hinrichs et al. 2006) with a minMatch = 0.5 and the hg38ToMm10.over.chain chain file. We then calculated the percentage of total regions that successfully lifted over. We also compared the liftOver rates of ENCODE cCREs-dELS—extracted from the cell type-agnostic set of cCREs—and 500k random regions of the genome generated from BEDtools *random*. For comparison, both these sets of regions were resized to 100 bp around the region center.

Relevant script: Compare-Verified-TSS-GENCODE.sh

Interpreting GWAS variants with the RAMPAGE rPeak catalog

Overlap of GWAS variants

We curated SNPs reported by the NHGRI-EBI GWAS catalog as of January 2019 and using population specific linkage disequilibrium, incorporating all SNPs in high LD ($r^2 > 0.7$) with this collection, as described in (The ENCODE Project Consortium et al. 2020). We created a master BED file with these annotations and intersected them with our RAMPAGE rPeak catalog using default BEDtools parameters. To compare gene assignments, we extracted reported and mapped genes from the original studies (columns 14 and 15 from the downloaded NHGRI-EBI GWAS catalog file) and determined if our rPeak linked genes (from read pair analysis) were represented on the list.

Relevant script: Overlap-GWAS-SNPs.sh

Comparison with eQTLs

We downloaded eQTLs from the GTEx database (GTEx_Analysis_v8_eQTL.tar), aggregated across all *signif_variant_gene_pairs.txt files, and reformatted the results into BED format. We then compared overlap between GWAS SNPs and matched controls as defined in (The ENCODE Project Consortium et al. 2020) and calculated the number of SNPs in each group that was linked to the same gene by both RAMPAGE reads and expression changes (eQTL).

Relevant script: Compare-eQTL-Links.sh

Cell type enrichment

We tested whether sets of GWAS SNPs were enriched in RAMPAGE rPeaks activity in specific biosamples using the same GWAS enrichment pipeline as described in (The ENCODE Project Consortium et al. 2020). Because RAMPAGE rPeaks have a much smaller genomic footprint than other collections of genomic regions (e.g., cCREs), we only included studies for which at least 15 LD blocks contained a SNP that overlapped a RAMPAGE rPeak (67 out of 397 initially tested GWAS studies). We reported all enrichments with an FDR corrected *p*-value less than 0.05 (**Supplemental Table S5B**).

Relevant repository: https://github.com/weng-lab/VIPER

3D chromatin interactions between ZH38T0028803 and KCNH7

We downloaded the cardiomyocyte promoter capture Hi-C data (Montefiori et al. 2018) from ArrayExpression under the accession E-MTAB-6014:

- E-MTAB-6014.processed.1.zip —> capt-CM-replicated-interactions-1kb.bedpe

and iPSC neuron promoter capture Hi-C data (Song et al. 2019) from GEO under the accession GSE113481:

- GSM3106832_cortical.cutoff.5.washU.txt.gz
- GSM3598046_hippocampal.cutoff.5.washU.txt.gz
- GSM3598048_motor.cutoff.5.washU.txt.gz

We also requested iPSC neuron Hi-C loop calls directly from (Rajarajan et al. 2018), who generously provided these annotations.

Using BEDtools *intersect* with default parameters, we intersected links with the *KCNH7* locus, requiring one of the *KCNH7* GENCODE TSSs to overlap one anchor and ZH38T0028803 to overlap the other anchor. Because the cardiomyocyte data was mapped to the hg19 genome, we lifted down the *KCNH7* TSSs and ZH38T0028803 coordinates to hg19 using UCSC liftOver with default parameters.

Relevant script: Compare-3D-Chromatin-Links.sh

Supplemental References

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**: 2537–2538.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**: 1311–1320.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: 699–710.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**: D590–8.

Hirabayashi S, Bhagat S, Matsuki Y, Takegami Y, Uehata T, Kanemaru A, Itoh M, Shirakawa K, Takaori-Kondo A, Takeuchi O, et al. 2019. NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat Genet* **51**: 1369–1379.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204–2207.

Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC genome browser and associated tools. *Brief Bioinform* **14**: 144–161.

Ma L, Cao J, Liu L, Du Q, Li Z, Zou D, Bajic VB, Zhang Z. 2019. LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res* **47**: 2699.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* **45**: D183–D189.

Montefiori LE, Sobreira DR, Sakabe NJ, Aneas I, Joslin AC, Hansen GT, Bozek G, Moskowitz IP, McNally EM, Nóbrega MA. 2018. A promoter interaction map for cardiovascular disease genetics. *Elife* **7**. http://dx.doi.org/10.7554/eLife.35788.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

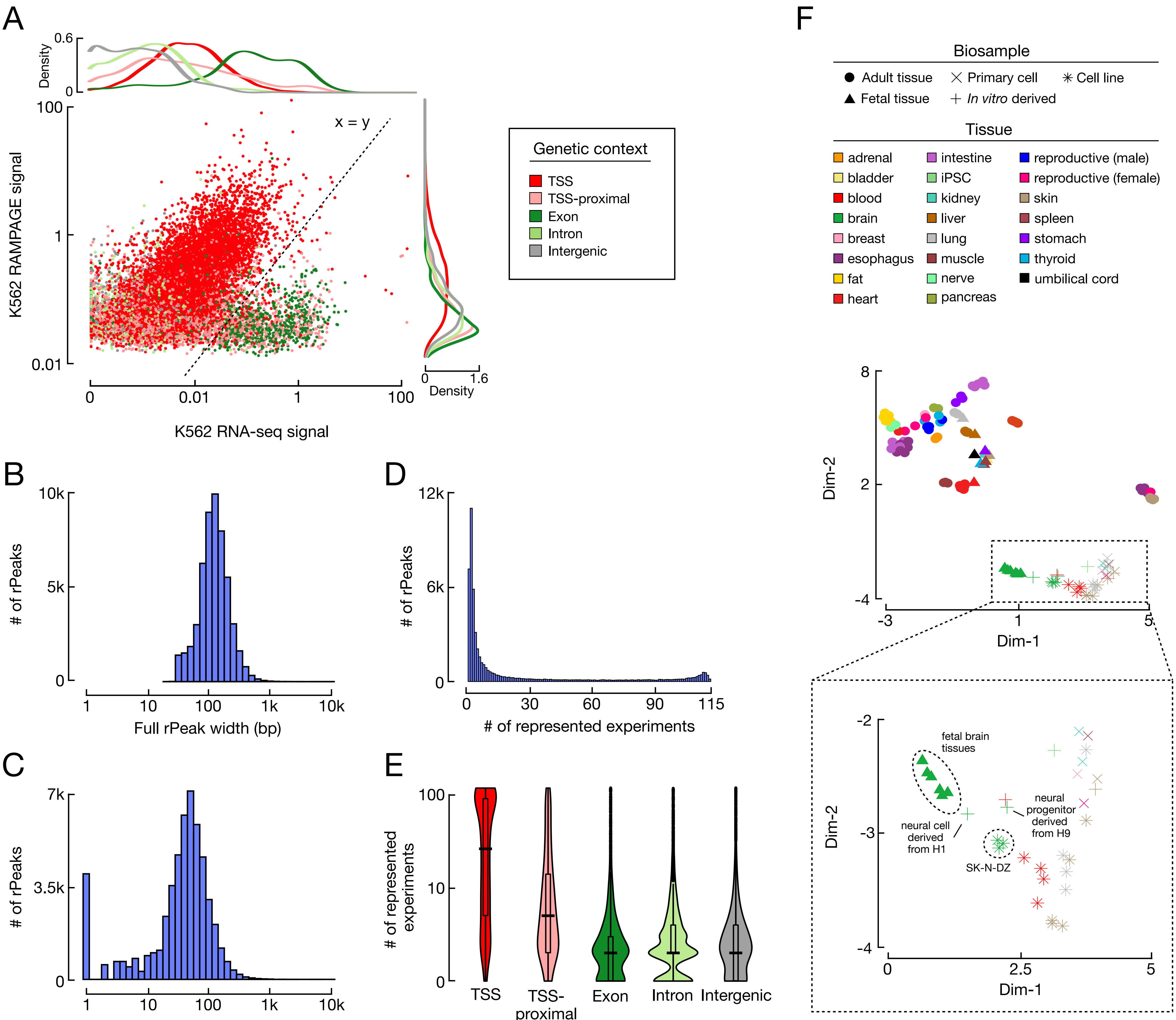
Rajarajan P, Borrman T, Liao W, Schrode N, Flaherty E, Casiño C, Powell S, Yashaswini C, LaMarca EA, Kassim B, et al. 2018. Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* **362**. http://dx.doi.org/10.1126/science.aat4311.

Song M, Yang X, Ren X, Maliskova L, Li B, Jones IR, Wang C, Jacob F, Wu K, Traglia M, et al. 2019. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat Genet* **51**: 1252–1262.

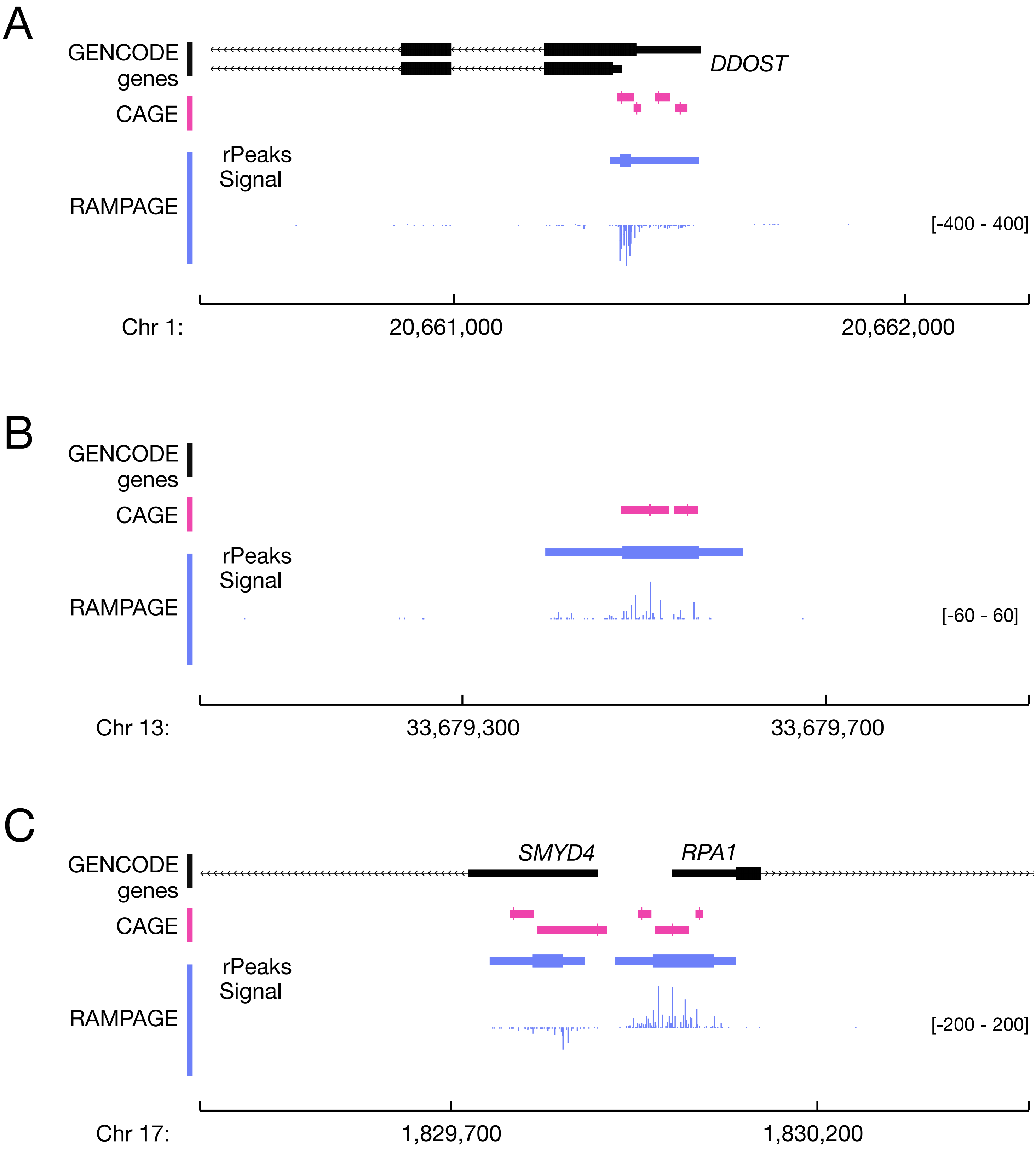
van Arensbergen J, FitzPatrick VD, de Haas M, Pagie L, Sluimer J, Bussemaker HJ, van Steensel B. 2017. Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* **35**: 145–153.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, et al. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31**: 28–33.

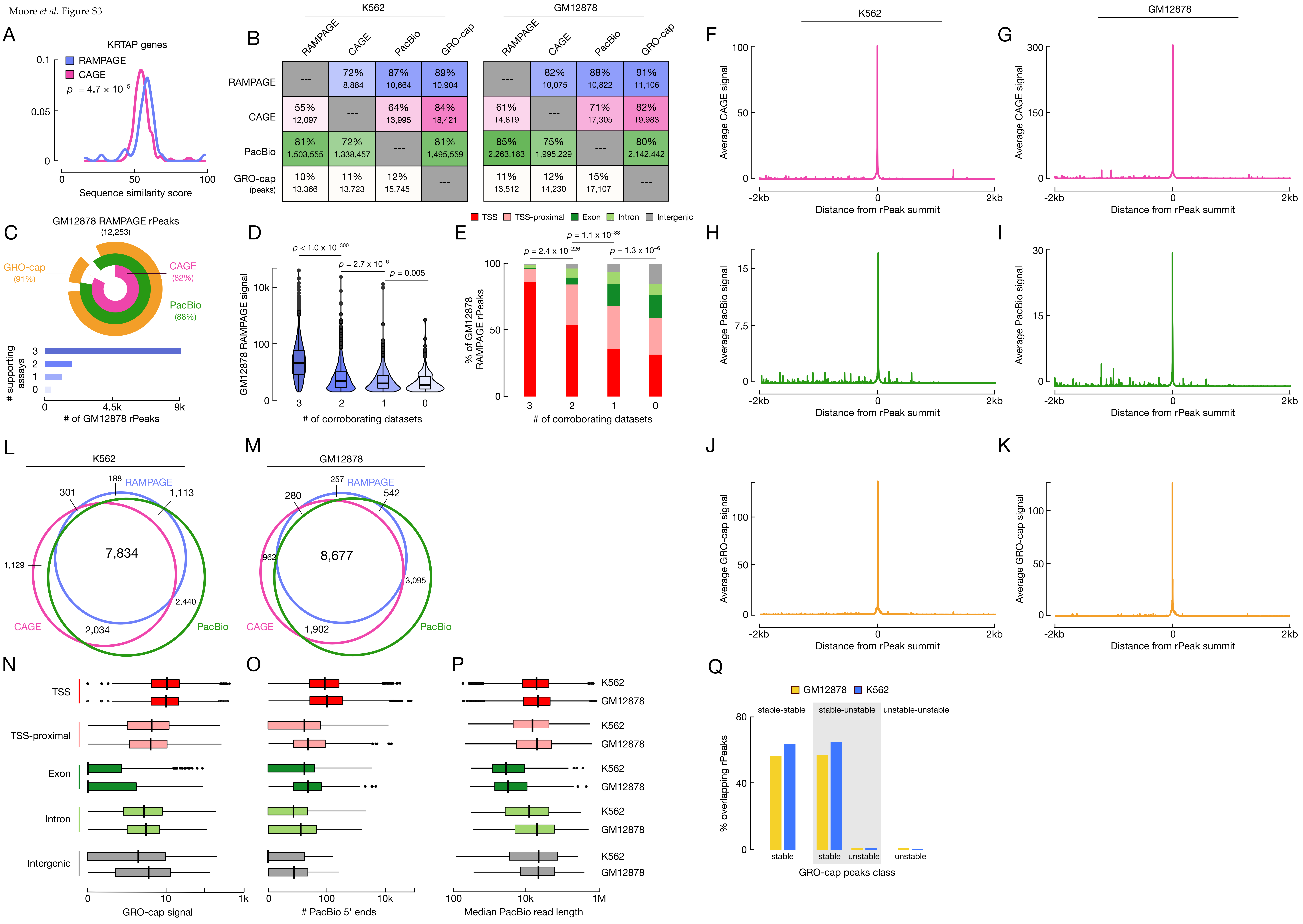
Zhang XO, Gingeras TR, Weng Z. 2019. Genome-wide analysis of polymerase III–transcribed Alu elements suggests cell-type–specific enhancer function. *Genome Res*. https://genome.cshlp.org/content/29/9/1402.short.



Supplemental Figure S1. General properties of RAMPAGE rPeaks. (*A*) A scatterplot comparing the RNA-seq signals (x-axis) and RAMPAGE signals (y-axis) in K562 cells at RAMPAGE peaks identified in K562. Points are colored by genomic context: TSS in red, TSS-proximal in pink, exon in dark green, intron in light green and intergenic in grey. Density plots along the x- and y-axes show the distributions of signal stratified by genomic context for RNA-seq and RAMPAGE signals, respectively. The dashed line represents the x=y line used to filter RAMPAGE peaks prior to the rPeak pipeline; peaks falling below this line were excluded. (*B*) A histogram depicting the distribution of rPeak full-peak widths. (*C*) A histogram depicting the distribution of rPeak high-density region widths. (*D*) A histogram depicting the number of RAMPAGE experiments represented by each rPeak. (*E*) A violin-boxplot depicting the number of experiments represented by each rPeak stratified by genomic context as in *A*. (*F*) Scatterplot displaying a two-dimensional Uniform Manifold Approximation and Projection (UMAP) embedding of 115 biosamples using RAMPAGE signal across all rPeaks as input features. Markers are shaped by biosample category and colored by tissue of origin as defined in the legend.



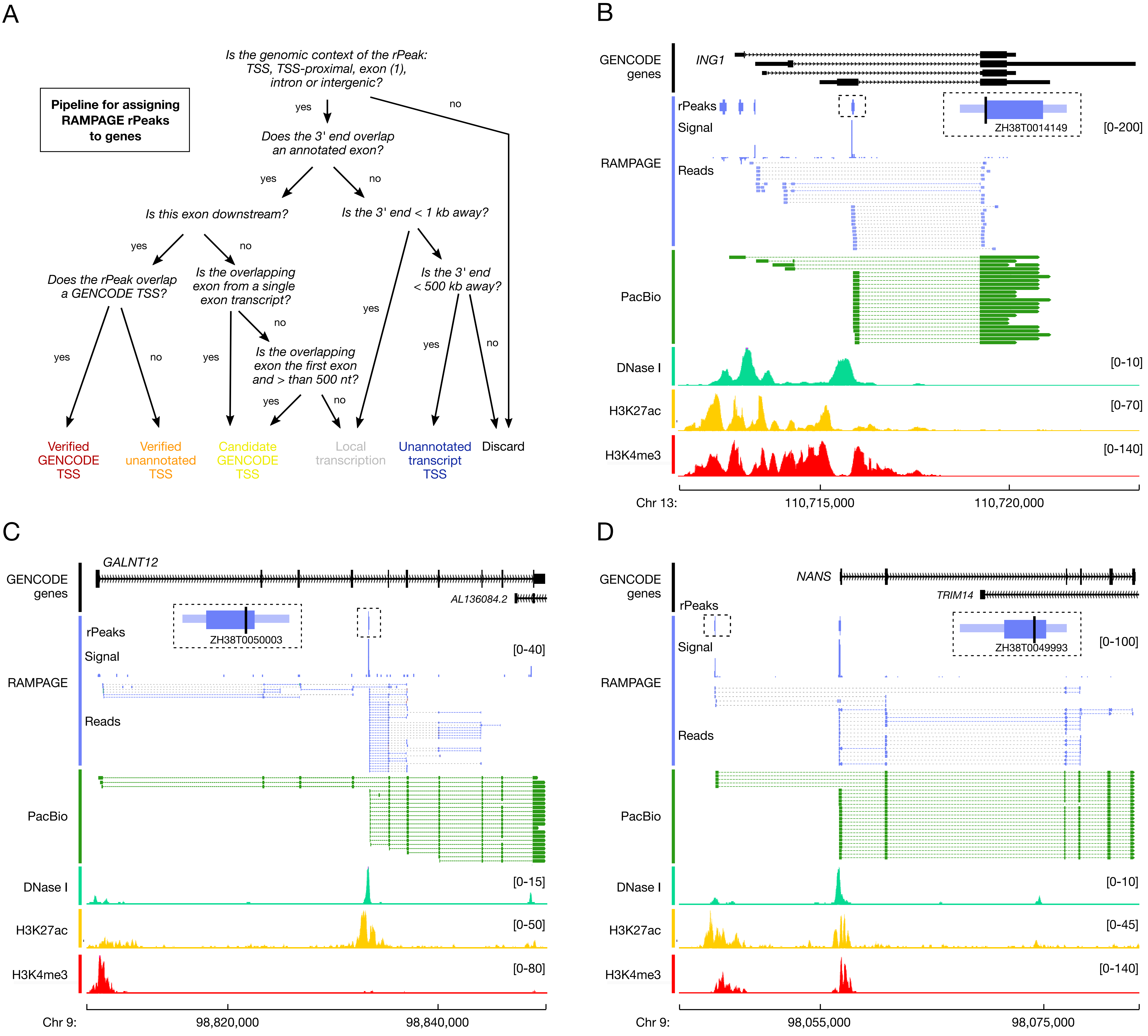
Supplemental Figure S2. RAMPAGE rPeaks that overlap multiple CAGE peaks are sites of dispersed transcription. (*A-C*) Three examples of RAMPAGE rPeaks (purple) that overlap multiple CAGE peaks (pink). GM12878 RAMPAGE signal (purple) reveals that these are sites of dispersed transcription.



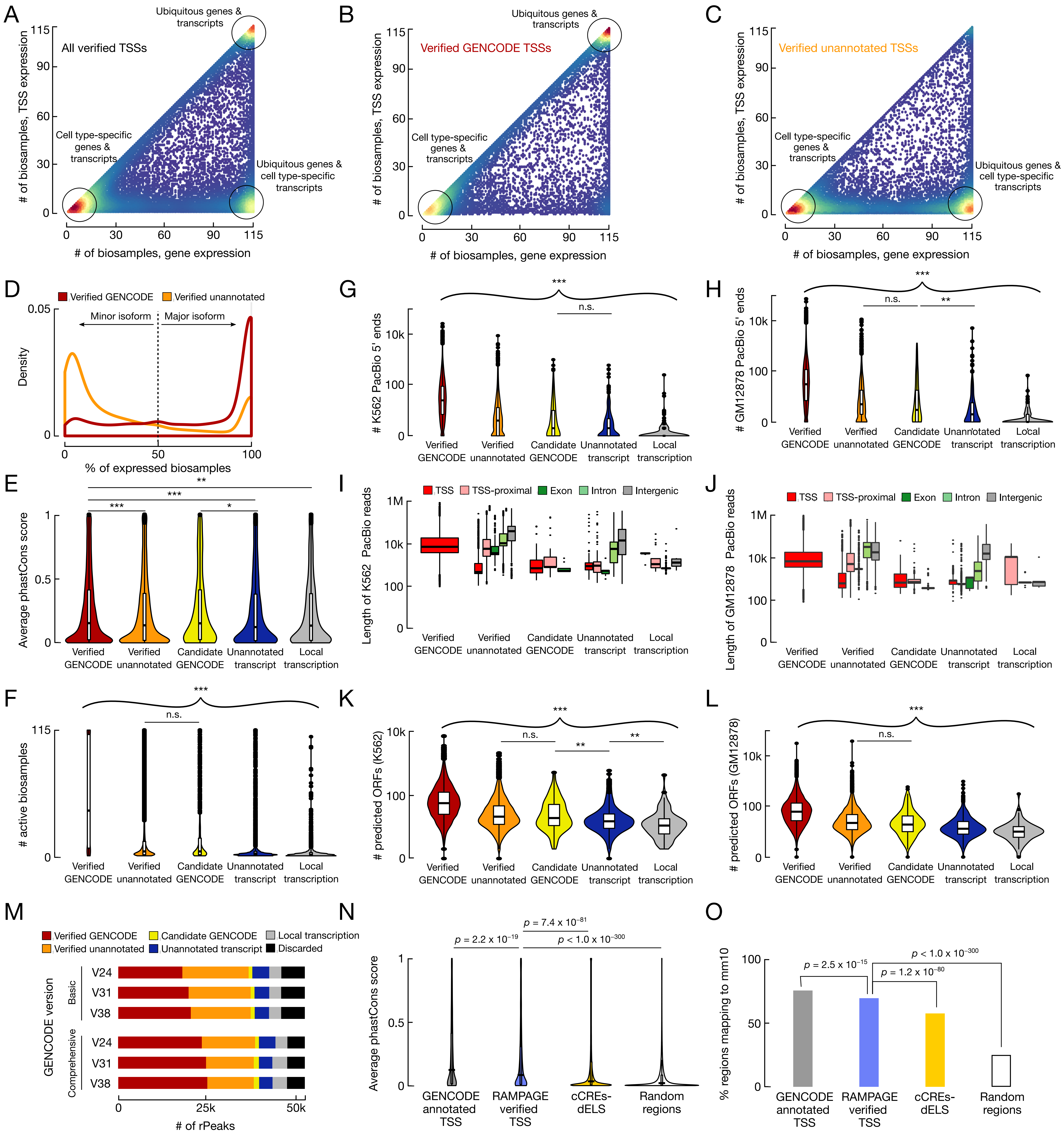
K562	RAMPAGE	CAGE	PacBio	GRO-cap
RAMPAGE	---	72% 8,884	87% 10,664	89% 10,904
CAGE	55% 12,097	---	64% 13,995	84% 18,421
PacBio	81% 1,503,555	72% 1,338,457	---	81% 1,495,559
GRO-cap (peaks)	10% 13,366	11% 13,723	12% 15,745	---

GM12878	RAMPAGE	CAGE	PacBio	GRO-cap
RAMPAGE	---	82% 10,075	88% 10,822	91% 11,106
CAGE	61% 14,819	---	71% 17,305	82% 19,983
PacBio	85% 2,263,183	75% 1,995,229	---	80% 2,142,442
GRO-cap (peaks)	11% 13,512	12% 14,230	15% 17,107	---

Supplemental Figure S3. Comparison of RAMPAGE rPeaks with individual transcriptome annotations. (*A*) A density plot showing the distributions of the similarity scores for sequences surrounding the TSSs of RAMPAGE-only (purple) and CAGE-only (pink) KRTAP genes. Sequence similarity is calculated by taking the maximum score of all pairwise local alignments. *P*-value corresponds to a two-sided Wilcoxon test. (*B*) A heatmap displaying the percentage and number of of RAMPAGE rPeaks (purple), CAGE peaks (pink), PacBio 5' read ends (green) and GRO-cap peaks (orange) that overlap each other and have high GRO-cap signal in K562 (left) and GM12878 (right) cells. (*C*) (top) VennPie diagram displaying the percentage of GM12878 RAMPAGE rPeaks that overlap K562 CAGE peaks (pink) or PacBio 5' ends (green), or have high GRO-seq signals (orange). Concentric circles show the percentages that are similarly supported between the three assays. (bottom) Bar plot with the number of GM12878 rPeaks stratified by the number of supporting transcriptomic assays as described in the above VennPie. (*D*) Violin-boxplot showing the distributions of the average GM12878 RAMPAGE signal across rPeaks stratified by the number of supporting assays as defined in *C*. *P*-values correspond to two-sided pairwise Wilcoxon tests with FDR correction. (*E*), Stacked bar graphs showing the percentage of GM12878 rPeaks belonging to each genomic context (TSS: red, TSS-proximal: pink, exon: dark green, intron: light green, intergenic: gray) stratified by the number of supporting assays as defined in *C*. *P*-values correspond to Chi-square tests. (*F*) Aggregate K562 CAGE 5' end signal at K562 RAMPAGE rPeaks centered on rPeak summit. (*G*) Aggregate GM12878 CAGE 5' end signal at GM12878 RAMPAGE rPeaks centered on rPeak summit. (*H*) Aggregate K562 PacBio 5' end signal at K562 RAMPAGE rPeaks centered on rPeak summit. (*I*) Aggregate GM12878 PacBio 5' end signal at GM12878 RAMPAGE rPeaks centered on rPeak summit. (*J*) Aggregate K562 GRO-cap 5' end signal at K562 RAMPAGE rPeaks centered on rPeak summit. (*K*) Aggregate GM12878 GRO-cap 5' end signal at GM12878 RAMPAGE rPeaks centered on rPeak summit. (*L*) Venn diagram showing the overlap of genes with TSSs that overlap either RAMPAGE rPeaks (purple), CAGE peaks (pink) or PacBio 5' ends (green) in K562. (*M*) Venn diagram as described in I for annotations in GM12878. (*N*) Boxplots showing the GRO-cap signal at RAMPAGE rPeaks stratified by genomic context in K562 (top) and GM12878 cells (bottom). (*O*) Boxplots showing the number of PacBio 5' ends that overlap RAMPAGE rPeaks stratified by genomic context in K562 (top) and GM12878 cells (bottom). (*P*) Boxplots showing the median lengths of the PacBio reads with 5' ends that overlap each RAMPAGE rPeak stratified by genomic context in K562 (top) and GM12878 cells (bottom). (*Q*) Overlap of GRO-cap peaks in GM12878 (yellow) and K562 (blue) classified by stability (Core et al. 2014) with RAMPAGE rPeaks in the respective cell types. Bidirectional GRO-cap peaks were classified into three groups based on comparisons with CAGE data: both peaks stable (left), one peak stable and one peak unstable (center), and both peaks unstable (right).



Supplemental Figure S4. Examples of exonic, intronic, and intergenic rPeaks in K562 cells. (*A*) Decision tree depicting the computational workflow for assigning RAMPAGE rPeaks to genes. (*B*) Genome browser view of the *ING1* locus. Exonic RAMPAGE rPeak ZH38T0014149 (in a small dashed box, with a magnified version shown to its right) overlaps an annotated exon of *ING1*, but is also a novel TSS for this gene. This annotation is supported not only by RAMPAGE signal and reads (purple), but PacBio reads (green) and epigenomic signals (DNase I: teal; H3K27ac: yellow; H3K4me3: red) in K562. Read pairs are denoted by dashed gray lines and split reads (i.e., single reads that span splice junctions) are denoted by colored lines (purple for RAMPAGE and green for PacBio). (*C*) Genome browser view of the *GALNT12* locus. Intronic RAMPAGE rPeak ZH38T0050003 (the smaller dashed box, with a magnified version shown to its left), overlaps an annotated intron of *GALNT12*, but is also a novel TSS for this gene. This annotation is supported not only by RAMPAGE signal and reads (purple), but PacBio reads (green) and epigenomic signals (colored as in *B*) in K562. (*D*) Genome browser view of the *NANS* locus. Intergenic RAMPAGE rPeak ZH38T0049993 (the smaller dashed box, with a magnified version shown to the right), lies upstream of *NANS*, but is also a novel TSS for this gene. This annotation is supported not only by RAMPAGE signal and reads (purple), but PacBio reads (green) and epigenomic signals (colored as in *A*) in K562.



Supplemental Figure S5. Features of RAMPAGE rPeaks stratified by TSS class. (*A*) Heatscatter plot displaying the number of biosamples in which each rPeak is expressed and the number of biosamples in which its linked gene is expressed. Each point is a single verified GENCODE TSS or verified annotated TSS. Color denotes density of points with purple corresponding to low density and red high density. (*B*) Heatscatter plot as defined in *A* with only verified GENCODE TSSs. (*C*) Heatscatter plot as defined in *A* with only verified unannotated TSSs. (*D*) Density plots depicting the percent of expressed biosamples of each gene accounted for by individual TSSs, separated by verified GENCODE TSSs (red) and verified annotated TSSs (orange). *P*-value corresponds to a Wilcoxon test. (*E*) Nested violin-boxplots displaying the average phastCons conservation score for RAMPAGE rPeaks stratified by TSS class (verified GENCODE: dark red, verified unannotated: orange, candidate GENCODE: yellow, unannotated transcript: dark blue, local transcription: gray). *P*-values corresponding to two-sided pairwise Wilcoxon tests with FDR correction are available in **Supplemental Table S4A**. Stars denote pairs with statistically significant differences (* $p < 0.05$, ** $p < 0.01$; *** $p < 0.001$) (*F*) Nested violin-boxplots displaying the number of active biosamples (RPM > 2) of RAMPAGE rPeaks stratified by gene assignment category (as colored in *E*). *P*-values corresponding to two-sided pairwise Wilcoxon tests with FDR correction are available in **Supplemental Table S4C**. All pairs are statistically significant (*** $p < 0.001$) except where noted. (*G*) Nested violin-boxplots displaying the number of overlapping K562 PacBio 5' ends for K562 RAMPAGE rPeaks stratified by gene assignment category (as colored in *E*). *P*-values corresponding to two-sided pairwise Wilcoxon tests with FDR correction are available in **Supplemental Table S4C**. All pairs are statistically significant (*** $p < 0.001$) except where noted. (*H*) Nested violin-boxplots displaying the number of overlapping GM12878 PacBio 5' ends for GM12878 RAMPAGE rPeaks stratified by gene assignment category (as colored in *E*). *P*-values corresponding to two-sided pairwise Wilcoxon tests with FDR correction are available in **Supplemental Table S4D**. (*I*) Boxplots displaying the length of K562 PacBio reads with overlapping 5' ends for K562 RAMPAGE rPeaks stratified by gene assignment category and colored by genetic context (TSS: red, TSS-proximal: pink, exon: dark green, intron: light green, intergenic: gray). *P*-values corresponding to pairwise Fisher's exact tests with FDR correction are available in **Supplemental Table S4E**. (*J*) Boxplots displaying the length of GM12878 PacBio reads with overlapping 5' ends for GM12878 RAMPAGE rPeaks stratified by gene assignment category and colored by genetic context (as defined in *I*). *P*-values corresponding to pairwise Fisher's exact tests with FDR correction are available in **Supplemental Table S4F**. (*K*) Nested violin-boxplots displaying the number of predicted ORFs in resulting transcripts in K562 for RAMPAGE rPeaks stratified by gene assignment category (as colored in *E*). *P*-values corresponding to two-sided pairwise Wilcoxon tests with FDR correction are available in **Supplemental Table S4G**. (*L*) Nested violin-boxplots displaying the number of predicted ORFs in resulting transcripts in GM12878 for RAMPAGE rPeaks stratified by gene assignment category (as colored in *E*). *P*-values corresponding to two-sided pairwise Wilcoxon tests with FDR correction are available in **Supplemental Table S4H**. (*M*) Bar plots depicting the number of RAMPAGE rPeaks assigned to each TSS category (as colored in E, discarded peaks in black) when different GENCODE versions were used in the assignment pipeline. (*N*) Nested violin-boxplots displaying the average phastCons conservation score for RAMPAGE-verified TSSs (purple), matched GENCODE-annotated TSSs (gray), cCREs-dELS (yellow), and random genomic regions (white). P-values correspond to two-sided pairwise Wilcoxon tests with FDR correction. (*O*) Bar plots displaying the percentage of RAMPAGE-verified TSSs (purple), matched GENCODE-annotated TSSs (gray), cCREs-dELS (yellow), and random genomic regions (white) that liftOver to the mm10 genome. *P*-values correspond to pairwise Fisher's exact tests with FDR correction.