PRAM: a novel pooling approach for discovering intergenic transcripts from large-scale RNA sequencing experiments

Peng Liu, Alexandra A. Soukup, Emery H. Bresnick, Colin N. Dewey, Sündüz Keleş

List of Supplementary Notes

List of Supplementary Tables

List of Supplementary Figures

Supplementary Note 1. Cufflinks-predicted models have false positive splice junctions.

Cufflinks-based '1-Step' and '2-Step' methods constructed a number of false positive splice junctions (Figure 1A), which seemed to contradict with the 'noise-free' input of the benchmark data. We found that most of these false positives had 5'- and 3'- splice sites shifted by the same number of base pairs (Supplementary Table 2) and a large fraction of them shifted by only one or two base pairs (Supplementary Figure 3 and Supplementary Figure 4). This suggests a further investigation to understand why Cufflinks built splice junctions this way.

Supplementary Note 2. Benchmark on simulated RNA-seq data containing noise.

We used RSEM's simulator to simulate RNA-seq data containing noise. RSEM's simulator can generate sequencing fragments based on pre-defined gene expression levels and on features such as noise level, fragment length distribution, read start position distribution, and sequencing error models learned from real datasets. In particular, noisy reads from RSEM simulator are randomly generated read sequences. Therefore, we used RSEM for our simulations.

Next, we describe the details of the simulations. We used the 1,256 benchmark transcripts as the full set of transcripts in the genome throughout the simulations that also included random noisy reads not originating from these transcripts. We used the quantification on the GENCODE transcripts in each of the 30 ENCODE RNA-seq datasets (https://github.com/pliu55/PRAM_paper) to extract the TPMs of benchmark transcripts and re-scaled them to one million to fulfill the assumption that benchmark transcripts were the only transcripts in the genome. We then simulated two million fragments based on the re-scaled TPMs and the noise ratios learned from each of the ENCODE datasets. The rationale for the two million reads was the observed maximum of 1.9 million total number of fragments that resided within the benchmark transcripts across the ENCODE datasets. Simulated fragments were aligned to entire genome by STAR to mimic real case applications. As a result, some fragments were not only mapped to benchmark transcripts, but also to other loci in the genome. All of the two '1-Step' methods and three '2-Step' methods were used to predict transcripts. Since the input RNA-seq datasets contained 'noise', predicted transcripts that had only one exon or genomic span shorter than 200 bp were removed by PRAM's default filtering step.

A summary of the simulation results according to a number of different criteria is now included in the manuscript. Below, we discuss the overall implications of these results. Given that our simulated RNA-seq fragments did not ensure full coverage of all the target transcripts, we first compared the methods in terms of the number of targets for which they failed to predict any model. The two '1-Step' methods, 'pooling + Cufflinks' and 'pooling + StringTie', had fewer missed targets than the three '2-Step' methods (Supplementary Table 4). Supplementary Figure 7 shows an example, where both the two '1-Step' methods predicted models for the target *AC073284.4* and all of the three '2-Step' methods missed this target transcript. Since our simulation contained 'noisy' RNA-seq fragments, which could potentially give rise to false positive transcript prediction, we next quantified the numbers of predicted models that did not overlap with any of the targets. 'Pooling + Cufflinks' had more such noisy models than 'Cufflinks + Cuffmerge' and 'Cufflinks + TACO', while 'pooling + StringTie' had more noisy models than 'StringTie + merging' (Supplementary Table 4). These noisy models mostly originated from multi-mapping RNA-seq fragments that also aligned to target transcripts (Supplementary Figure 8). Most of the noisy models had lower expression levels than those of non-noisy models (Supplementary Figure 9). Thus, these noisy models are likely to be well separated from the true models and should not constitute a significant problem in real applications. The comparison of missed targets and noisy models suggested that '1-Step' methods had increased prediction coverage at the expense of noisy models.

Unlike our 'noise-free' data-driven benchmark, simulated RNA-seq fragments in these simulations did not guarantee full coverage of all target transcripts. Therefore, we separated 'detected' target transcripts (i.e., those overlapping with a predicted model) from 'undetected' ones and focused on targets that had predicted models from all of the two '1-Step' and three '2-Step' methods. This allowed us to make a fair comparison for all five methods on simulated data in addition to the missed and noisy target quantifications in Supplementary Table 4. 942 targets out of the 1256 were predicted by all five methods. For each method, we calculated precision and recall of the predicted models that overlapped with this gold-standard set. Models that overlapped with multiple targets were excluded to avoid ambiguity. Except for 'Cufflinks + Cuffmerge', which had markedly lower precision, the methods had similar precision and recall for all the three features evaluated (Supplementary Figure 10).

Supplementary Note 3. '1-Step' methods predicted a very small number of chimeric transcripts.

We first quantified the numbers of intergenic transcripts were partially overlapping since these would be the loci to give rise to chimeric transcript predictions. Since we do not have a true set of intergenic transcripts, we used the 'newly discovered' GENCODE transcripts instead, because we showed in our manuscript that they have similar features to intergenic transcripts.

There are 1034 'newly discovered' transcripts in total. To avoid double counting, we removed those that had genomic span as a subset of another 'newly discovered' transcript's genomic span, because any transcripts that partially overlapped with them would also overlap with their 'superset' transcripts. Using the 963 remaining transcripts, we looked for pairs of transcripts that were partially overlapping on the same strand. Only 51 of the 963 transcripts (5%) belonged to this category (Supplementary Table 5). The small percentage of partially overlapping transcripts indicated that chimeric transcripts would be a small fraction of the intergenic transcripts predicted by '1-Step' methods.

To estimate the number of chimeric transcripts in real predictions, we performed a benchmark test on the 'newly discovered' transcripts. To increase the biological diversity of the input RNA-seq samples, we downloaded a new set of data from ENCODE that were derived from 19 different human tissues and five different donors (Supplementary Table 6). We selected these samples with the following criteria:

- have a state as 'released' without any warning item;
- poly(A) paired-ended strand-specific human tissue RNA-seq from ENCODE, Roadmap, or GGR;
- have BAM files containing RNA-seq fragments mapped to hg38 for public download (i.e., not restricted through dbGAP);
- if multiple replicates existed for the same sample, the one with the highest number of uniquely mapped fragments was taken

To test prediction's dependency on different number of input, we created four subset datasets by selecting 5, 10, 20, and 30 samples from the 40 total samples (Supplementary Table 6). To prepare the datasets, we first ran K-means clustering on gene expression profiles of the 40 samples to divide them into 5, 10, 20, or 30 groups. As an example, Supplementary Figure 11 shows a multidimensional scaling plot of the gene expression profiles with K-means clustering into 5 groups. Samples from similar anatomic sites, such as heart (the red cluster in the lower right of Supplementary Figure 11) or colon (the blue cluster in the upper left of Supplementary Figure 11) were clustered into one group. From each group, we randomly picked one sample as the representative and therefore obtained datasets containing 5, 10, 20, or 30 RNA-seq experiments. Together with a dataset including all 40 samples, we obtained five sets with different numbers of samples (5, 10, 20, 30, and 40). We used these five sets as the input for transcript prediction in the following analysis.

To estimate the number of chimeric transcripts that may arise in actual applications of '1-Step' and '2-Step' methods, we performed a benchmark test using the 'newly discovered' transcripts as the ground truth. We utilized these instead of the select single-transcript genes we used in our 'noise-free' benchmark because they can give rise to chimeric transcripts as discussed above. We prepared mapped RNA-seq fragments in the same way as we did in our 'noise-free' benchmark for comparing '1-Step' and '2-Step' methods with the exception of not requiring full RNA-seq fragment coverage of every 'newly discovered' transcript. We applied the two '1-Step' methods, 'pooling + Cufflinks' and 'pooling + StringTie' on the 40 input RNA-seq datasets and summarized the number of predicted chimeric transcripts. There were at most 16 chimeric transcripts by the two '1-Step' methods (Supplementary Figure 12). Not every potential chimeric transcript loci generated a chimeric transcript, most likely due to the incomplete RNA-seq fragment coverage of these loci. The fact that, at most, only 16 transcripts were chimeric among the large set of (~1000) 'newly discovered' transcripts

suggested again that chimeric transcripts are a negligible fraction of intergenic transcripts predicted by ‘1-Step’ methods.

We also evaluated the impact of the number of input RNA-seq datasets on the number of chimeric transcripts predicted. We considered the sets of 5, 10, 20, and 30 samples randomly selected from K-means clustering as described above. The number of predicted chimeric transcripts did not increased dramatically when using 10, 20, 30, or 40 samples (Supplementary Figure 12), suggesting that increasing the number of inputs would not lead to a large increase in the predicted chimeric transcripts.

Finally, we also asked whether ‘2-Step’ methods can resulted in chimeric transcripts. The ‘2-Step’ method, ‘StringTie + merging’, which had been shown to have superior performance than the other two ‘2-Step’ methods in our method-comparing benchmark, predicted 6 to 14 chimeric transcripts (Supplementary Figure 12). Although these numbers were slightly smaller than the ones from ‘1-Step’ methods, it nevertheless suggested that the ‘2-Step’ method was not immune from predicting chimeric transcripts.

Supplementary Note 4. PRAM has competitive time and memory cost for intergenic transcript discovery.

We evaluated PRAM's performance on the 30 RNA-seq datasets that were used to construct our benchmark. After PRAM extracted the RNA-seq alignments within intergenic regions (≥ 10 kb away from any GENCODE version 24 genes or pseudo-genes), the average number of uniquely mapped fragments was reduced from 71 million to 0.43 million and the average number of multi-mapping fragments decreased from 7.7 million to 0.08 million (Supplementary Table 7). Consequently, the average size of the input BAM files shrunk from 15 GB to 0.07 GB (Supplementary Table 8). The dramatic reduction in the number of alignments to be processed allowed PRAM to finish model building in under four hours using eight 2.1 GHz AMD CPUs in parallel by the default 'pooling + Cufflinks' method (Supplementary Table 9). Furthermore, 'pooling + StringTie' on the same input took only seven minutes, which is comparable to or much faster than the '2-Step' methods (Supplementary Table 9). The marked reduction of input size and competitive computing time illustrates that PRAM streamlines the process of intergenic transcript discovery via the '1-Step' approach.

To study the trade-off between accuracy and execution time, we ran the two '1-Step' methods, 'pooling + Cufflinks' and 'pooling + StringTie' on the 5, 10, 20, 30, and 40 RNA-seq samples that described above. To increase statistical power, we removed the full coverage requirement of the target transcripts and thus expanded to target set to all of the 3,643 GENCODE (version 24)'s one-transcript genes on Chromosome 1 to 22 and X. We did a benchmark evaluation in the same way as we did in our 'noise-free' benchmark. When the number of input datasets increased, the running time for 'pooling + Cufflinks' increased dramatically, whereas the running time for 'pooling + StringTie' increased slightly (Supplementary Figure 13). The memory cost remained about the same for both '1-Step' methods when input increased from 10 to 40 samples (Supplementary Figure 13).

To evaluate prediction accuracy, we first assessed the number of target transcripts without any predicted models. As expected, when using more input samples, the coverage of target transcripts increased and resulted in fewer missed target transcripts for both '1-Step' methods (Supplementary Figure 14). 'Pooling + Cufflinks' missed fewer targets than 'pooling + StringTie' under the same number of input samples (Supplementary Figure 14). Next, we evaluated prediction accuracy for the 780 target transcripts with predicted models from both methods under all the different numbers of input samples. For prediction of exon nucleotides, recall increased as the number of input samples increased, while precision remained at a nearly perfect level for both methods (Supplementary Figure 15). For prediction of individual junction and transcript structure, both precision and recall increased as the number of input increased for both methods and 'pooling + StringTie' had markedly higher precision than 'pooling + Cufflinks' at the same numbers of samples (Supplementary Figure 15). For each method, there was a very small difference of precision and recall using 30 or 40 input samples (Supplementary Figure 15). In summary, this exposition suggests that for a high prediction coverage, i.e., detection, 'pooling + Cufflinks' with a large number of input RNA-seq samples is the best strategy. In general, we expect further scrutinization on predicted transcripts in the form of TPM filtering as we have shown in the simulation setting with added noise or integrating with other genomic samples as we have shown in our later case study in mouse. For high precision on the finer structures of the transcripts, 'pooling + StringTie' with a larger number of samples provided marked improvement, but after 10 samples, the gains were negligible. We would like to also emphasize that the setting we have considered here does not address all the cases one might encounter in practice. However, it provides reassurance that even going from 5 to 30 samples across diverse tissues, the evaluation metrics indicate gain in power.

Supplementary Note 5. PRAM transcripts are unlikely to be eRNAs or uROFs.

A comparison of the PRAM transcripts to the FANTOM5 'robust set' of 38,554 predicted enhancers (referred to as "enhancers") revealed that only 2.8% (1,091) of all the enhancers overlapped with 8.8% (1,246) of our master set of transcripts. These 1,091 enhancers had markedly shorter lengths (median = 316 bp) than those of the 1,246 transcript models (median = 7,977 bp) (Supplementary Figure 21). This further confirms that our master set of transcript models are unlikely to harbor eRNAs. The genomic distance constraint to the annotated genes (10kb away) we applied during PRAM prediction excluded the possibility that PRAM were upstream open read frames (uORFs).

Supplementary Note 6. Protein-coding potential of PRAM transcripts.

We assessed the protein-coding potential of PRAM transcripts by BLAST and PhyloCSF. 31% of PRAM transcripts aligned to at least one protein sequence by BLAST against non-redundant mammalian protein sequences (Supplementary Table 16). PRAM transcripts and 'newly discovered' GENCODE transcripts shared a similar distribution of numbers of matched proteins (Supplementary Figure 26). In particular, 41 repeat-free PRAM transcripts matched to more than 100 proteins via BLAST (Supplementary Figure 26). Since only 62 of the 1,034 (6%) 'newly discovered' transcripts are protein-coding and a large fraction of BLAST-matched proteins on the 41 PRAM transcripts are uncertain (Supplementary Figure 27), we also evaluated their ORFs and PhyloCSF scores. Predicted ORF lengths of the 41 PRAM transcripts are similar to those of 'newly discovered' transcripts, but shorter than those of 'long-standing' transcripts (Supplementary Figure 28). Most of these predicted ORFs have negative PhyloCSF scores, suggesting that they are unlikely to be protein-coding (Supplementary Figure 29). Only one of the transcripts, plcf_chr2_minus.9034.2, has a predicted ORF with a positive PhyloCSF score (Supplementary Figure 29 and Supplementary Figure 30). This transcript is also the only one with an ORF that partially overlaps with a PhyloCSF-predicted coding region on the same strand and frame (Supplementary Figure 30), suggesting some protein-coding potential. We noted that, over this locus, PhyloCSF's statistical power had a maximum around 0.4, which is much lower than the maximum possible power of 1.0 (Supplementary Figure 30). Therefore, interpreting this transcript's protein-coding potential warrants some caution.

Another transcript at the same locus, plcf_chr2_minus.9034.1, is also one of the 41 PRAM transcripts. This transcript has exons overlapping with PhyloCSF-predicted coding regions on the same strand regardless of frame (Supplementary Figure 30). However, due to an unmatched frame in the middle region of its ORF (Frame 3 vs. Frame 2) and lack of an overlap of its 3' region with predicted coding regions on Frame 3, its PhyloCSF score is negative, suggesting low protein-coding potential.

Most of the BLAST-matched proteins for PRAM transcript plcf_chr2_minus.9034.1 and plcf_chr2_minus.9034.2 are hypothetical or predicted proteins (Supplementary Table 17 and Supplementary Table 18), which we classified as 'uncertain proteins' earlier. These two lists of proteins indicate again that protein-coding potential of these two transcripts' should be interpreted cautiously.

Since the number of BLAST-matched proteins may not be sufficient to support a transcript's protein-coding potential as we observed from the 41 PRAM transcripts above, we decided to expand our PhyloCSF analysis to all PRAM transcripts. A large number of PRAM transcripts do not have any overlap with PhyloCSF-predicted coding regions (Supplementary Table 19), suggesting that they are unlikely to be protein-coding. However, there are two transcripts that have relatively large overlap with PhyloCSF-predicted coding regions regardless of frame (Supplementary Table 19). These two transcripts are highlighted in Supplementary Figure 31.

Supplementary Note 7. Transcripts predicted uniquely by '1-Step' methods were supported by RAMPAGE and histone mark ChIP-seq data.

As a final assessment of PRAM transcripts, we asked whether these '1-Step'-predicted transcripts were missed by '2-Step' methods and yet had supporting promoter activities and epigenetic signals. Specifically, we compared transcripts built by the '1-Step' method ('pooling + Cufflinks') with those from '2-Step' methods ('Cufflinks + Cuffmerge' and 'Cufflinks + TACO') and exclusively focused on those that were predicted by only one method (Supplementary Table 20). We followed the above strategy of examining RAMPAGE and histone modification signals as a function of gene expression. There was only one model solely predicted by the '1-Step' method that had TPM ≥ 1 (Supplementary Table 20), whereas the '2-Step' methods had one or three such unique models depending on the category of mappability-filtering. The '1-Step' prediction had the highest promoter activity of all the uniquely predicted models (Supplementary Figure 33), suggesting that this model was most likely a true transcript. Of transcripts with TPM $\in$ [0.1, 1), there were thirteen and seventeen models predicted uniquely by the '1-Step' method in GM12878 and K562, respectively (Supplementary Table 20), compared to at most five models predicted uniquely by the '2-Step' methods. At least one of the '1-Step' predictions had higher promoter activity compared to '2-Step' models with similar expression levels (Supplementary Figure 33). Histone modification signals of all these models had distributions with higher medians than those from models with TPM < 0.1 (Supplementary Figure 34). Both the promoter activities and epigenetic signals suggested that '1-Step' method identified well-supported transcripts that were not predicted by '2-Step' methods, demonstrating again that the '1-Step' approach outperforms '2-Step' approaches.

Supplementary Note 8. '2-Step' methods and Cufflinks missed validated '1-Step' mouse models.

Four gene models (CUFFm.chr12.33668, CUFFm.chr17.20196, CUFFp.chr10.20259, and CUFFp.chr12.15498) built by '1-Step' method 'pooling + Cufflinks' had been detected by semi-qRT-PCR in G1E-ER-GATA1 cells. We asked whether '2-Step' methods or transcript reconstruction based on individual RNA-seq dataset can also predict these four 'hit' gene models. We applied 'Cufflinks + Cuffmerge' and 'Cufflinks + TACO' on the same 32 input RNA-seq datasets as well as applying Cufflinks on each of the input RNA-seq datasets (Supplementary Table 21). Since our gene models were built by 'pooling + Cufflinks', we only used Cufflinks and did not include StringTie here as to make a fair comparison. Although both of the two '2-Step' methods and 25 out of 32 Cufflinks runs built models that overlapped with our four hits, only those from 'Cufflinks + Cuffmerge' and 8 out of 32 Cufflinks runs remained after selecting by differential expression (Supplementary Table 27). Following the same selection steps as we did for our hits, none of the '2-Step' methods or Cufflinks produced gene models overlapped with CUFFm.chr17.20196 or CUFFp.chr12.15498. Only 'Cufflinks + Cuffmerge' produced a gene model overlapping with CUFFp.chr10.20259 (Supplementary Table 28). This comparison further reinforced the fact that '1-Step' approach outperformed '2-Step' approach.

Supplementary Note 9. Gene model CUFFp.chr7.6106 was not expressed in K562.

CUFFp.chr7.6106 had both TPM and expected fragment counts as zero in all RNA-seq datasets, indicating that it was not expressed at all (Supplementary Figure 42 A and B). Further investigation showed that CUFFp.chr7.6106 was built on an RNA-seq fragment from the K562 dataset ENCSR109IQO (replicate 2) with its 5'-splice site not compatible with this fragment (Supplementary Figure 43). This is attributable to Cufflinks's shift of splice sites as observed in our 'noise-free' benchmark (Supplementary Note 1; Supplementary Figure 3 and Supplementary Figure 4). As a result, no K562 RNA-seq fragment was compatible with CUFFp.chr7.6106 and thus its expected count was zero in all datasets.

Supplementary Table 1. ENCODE human RNA-seq datasets for benchmark test. Accession IDs and metadata were obtained from ENCODE website: https://www.encodeproject.org. BAM files were downloaded directly from ENCODE to build the benchmark. This dataset contains all of the strand-specific paired-end poly(A) mRNA-seq alignments for human untreated immortalized cell lines released by ENCODE as of February, 2017. RNA-seq datasets from subcellular fractions, including membrane, nucleolus, nucleus, cytosol, chromatin, or nucleoplasm, were excluded. All of the alignments were mapped to human genome hg38 annotated by GENCODE version 24.

Experiment	Cell	Biological replicate index	BAM	Mate1 FASTQ	Mate2 FASTQ
ENCSR000AED	GM12878	1	ENCFF802TLC	ENCFF001REK	ENCFF001REJ
ENCSR000AED	GM12878	2	ENCFF428VBU	ENCFF001REI	ENCFF001REH
ENCSR000AEF	GM12878	1	ENCFF547YFO	ENCFF001RDG	ENCFF001RCY
ENCSR000AEF	GM12878	2	ENCFF782IVX	ENCFF001RDF	ENCFF001RCX
ENCSR000AEM	K562	1	ENCFF912SZP	ENCFF001RED	ENCFF001RDZ
ENCSR000AEM	K562	2	ENCFF207ZSA	ENCFF001REG	ENCFF001REF
ENCSR000AEO	K562	1	ENCFF846WOV	ENCFF001RDE	ENCFF001RCW
ENCSR000AEO	K562	2	ENCFF588YLF	ENCFF001RDD	ENCFF001RCV
ENCSR000CON	A549	1	ENCFF125RAL	ENCFF000EJJ	ENCFF000EJV
ENCSR000CON	A549	2	ENCFF739OVZ	ENCFF000EJW	ENCFF000EKB
ENCSR000COQ	GM12878	1	ENCFF709IUX	ENCFF000EWJ	ENCFF000EWX
ENCSR000COQ	GM12878	2	ENCFF244ZQA	ENCFF000EWW	ENCFF000EXE
ENCSR000CPE	HepG2	1	ENCFF315VHI	ENCFF000FVT	ENCFF000FVU
ENCSR000CPE	HepG2	2	ENCFF834ITU	ENCFF000FVI	ENCFF000FVV
ENCSR000CPH	K562	1	ENCFF048ODN	ENCFF000HFF	ENCFF000HFG
ENCSR000CPH	K562	2	ENCFF381BQZ	ENCFF000HFH	ENCFF000HFY
ENCSR000CPR	HeLa-S3	1	ENCFF343WEZ	ENCFF000FOM	ENCFF000FOV
ENCSR000CPR	HeLa-S3	2	ENCFF444SCT	ENCFF000FOK	ENCFF000FOY
ENCSR000CPT	MCF-7	1	ENCFF367VEP	ENCFF000HQR	ENCFF000HQP
ENCSR000CPT	MCF-7	2	ENCFF983FHE	ENCFF000HQQ	ENCFF000HRH
ENCSR000CTT	SK-N-SH	1	ENCFF263OLY	ENCFF000IMA	ENCFF000IMR
ENCSR000CTT	SK-N-SH	2	ENCFF978ACT	ENCFF000IMC	ENCFF000IMS
ENCSR310FIS	MCF-7	1	ENCFF904OHO	ENCFF002DKR	ENCFF002DKU
ENCSR310FIS	MCF-7	2	ENCFF838JGD	ENCFF002DKX	ENCFF002DKY
ENCSR545DKY	K562	1	ENCFF044SJL	ENCFF059IUV	ENCFF104ZSG
ENCSR545DKY	K562	2	ENCFF728JKQ	ENCFF628GUZ	ENCFF695XOC
ENCSR561FEE	HepG2	1	ENCFF306YQS	ENCFF946VBP	ENCFF982FAM
ENCSR561FEE	HepG2	2	ENCFF521KYZ	ENCFF787PPA	ENCFF564BSM
ENCSR985KAT	HepG2	1	ENCFF800YJR	ENCFF002DKZ	ENCFF002DLC
ENCSR985KAT	HepG2	2	ENCFF782TAX	ENCFF002DLE	ENCFF002DLG

Supplementary Table 2. Number of false positive splice junctions by '1-Step' and '2-Step' methods. False positives were only from Cufflinks-based methods and most of them had the 5'- and 3'-splice sites shifted by the same number of base pairs.

Method	Number of false positive splice junctions	
	Total	**5'- and 3'-splice site shifted by the same number of base pairs**
pooling + Cufflinks	192	192
Cufflinks + Cuffmerge	549	544
Cufflinks + TACO	251	249
pooling + StringTie	0	0
StringTie + merging	0	0

Supplementary Table 3. Number of transcripts missed by '1-Step' and '2-Step' methods in benchmark test. Two '1-Step' methods ('pooling + Cufflinks' and 'pooling + StringTie') and three '2-Step' methods ('Cufflinks + Cuffmerge', 'Cufflinks + TACO', and 'StringTie + merging') are compared here. 'Predicted' refers to cases with recall = 1 and precision = 1, and 'missed' to cases with recall = 0. We compared the number of transcripts that had their structures predicted correctly (i.e., transcripts with recall = 1 and precision = 1) by one type of meta-assembly method, but missed (recall = 0) by the other. There were 918 transcripts constructed by both of the '1-Step' methods. Among these, eighteen were missed by all three '2-Step' methods and 28 were missed by two '2-Step' methods. In comparison, there were only 461 transcripts constructed by all three '2-Step' methods, none of which were missed by the '1-Step' methods. Similarly, of the 433 transcripts that were predicted by two '2-Step' methods, only six were missed by both of the '1-Step' methods.

		Number of '2-Step' methods that missed the transcript				
		0	**1**	**2**	**3**	**Total**
Number of '1-Step' methods that predicted the transcript	**0**	1	20	13	27	61
	1	105	82	70	20	277
	2	635	237	28	18	918

		Number of '1-Step' methods that missed the transcript			
		0	**1**	**2**	**Total**
Number of '2-Step' methods that predicted the transcript	**0**	34	33	30	97
	1	181	75	9	265
	2	364	63	6	433
	3	446	15	0	461

Supplementary Table 4. Number of atypical target transcripts and predicted models. 'Missed targets' refers to target transcripts without any predicted models overlapping with their genomic span on the same strand. 'Noisy models' are predicted models with genomic span do not overlap with any target transcripts on the same strand.

method	number of missed targets	number of noisy models
pooling + Cufflinks	30	382
pooling + StringTie	37	260
Cufflinks + Cuffmerge	254	354
Cufflinks + TACO	59	159
StringTie + merging	41	177

Supplementary Table 5. Number of 'newly discovered' transcripts with potential to give rise of predicted chimeric transcripts.

'newly discovered' transcripts	number of transcripts
total	1034
after removing transcripts that were a subset of other transcripts	963
transcripts that partially overlapped with other transcripts	51

Supplementary Table 6. ENCODE human tissue RNA-seq datasets. Datasets randomly selected as representatives for each K-means clustering were denoted as 'Y' for 5, 10, 20, or 30 clusters. They were used as the input RNA-seq datasets for later analysis.

tissue	donor	RNA-seq ID	BAM ID	N=5	N=10	N=20	N=30
adipose tissue	female adult (30 years)	ENCSR686JJB	ENCFF717MIN			Y	Y
adipose tissue	male adult (34 years)	ENCSR741QEH	ENCFF491UPQ				Y
adipose tissue	male child (3 years)	ENCSR718CDN	ENCFF668YHY				Y
adrenal gland	female adult (30 years)	ENCSR146ZKR	ENCFF181WWD				
adrenal gland	male adult (21 year)	ENCSR680AAZ	ENCFF917FLU				Y
adrenal gland	male adult (34 years)	ENCSR598KJX	ENCFF225BUX			Y	Y
aorta	female adult (30 years)	ENCSR995BHD	ENCFF864QLZ			Y	Y
aorta	male adult (34 years)	ENCSR763NOO	ENCFF081DYZ			Y	Y
esophagus	female adult (30 years)	ENCSR993QGR	ENCFF441FVJ				
esophagus	male adult (34 years)	ENCSR102TQN	ENCFF251YUZ		Y	Y	Y
heart	female adult (30 years)	ENCSR635GTY	ENCFF710AVC	Y		Y	Y
heart left ventricle	male adult (34 years)	ENCSR769LNJ	ENCFF466PKR				
heart left ventricle	male child (3 years)	ENCSR693CSQ	ENCFF127SLH				
heart right ventricle	male adult (34 years)	ENCSR433XCV	ENCFF684RZI		Y	Y	Y
heart right ventricle	male child (3 years)	ENCSR439SPU	ENCFF884HIK			Y	Y
liver	male child (3 years)	ENCSR714KDG	ENCFF131MIC	Y	Y	Y	Y
lung	female adult (30 years)	ENCSR917YHC	ENCFF024HAR				Y
lung	male child (3 years)	ENCSR278UYN	ENCFF841DQD			Y	Y
ovary	female adult (30 years)	ENCSR725TPW	ENCFF325FVQ	Y	Y		Y
ovary	female adult (47 years)	ENCSR046XHI	ENCFF448GEE			Y	Y
pancreas	female adult (30 years)	ENCSR571BML	ENCFF085TWC		Y	Y	
pancreas	male adult (34 years)	ENCSR629VMZ	ENCFF199EFU				Y
psoas muscle	female adult (30 years)	ENCSR502OTI	ENCFF677DKS				Y
psoas muscle	male adult (34 years)	ENCSR843HXR	ENCFF738TTD		Y	Y	Y
psoas muscle	male child (3 years)	ENCSR817TLH	ENCFF344BDW				Y

right cardiac atrium	male adult (34 years)	ENCSR675YAS	ENCFF704QQH				Y
sigmoid colon	female adult (30 years)	ENCSR825GWD	ENCFF588TQX			Y	Y
sigmoid colon	male adult (34 years)	ENCSR999ZCI	ENCFF050JVY	Y			Y
sigmoid colon	male child (3 years)	ENCSR396GIH	ENCFF323GDO		Y		
small intestine	female adult (30 years)	ENCSR039ICU	ENCFF584GAJ			Y	
small intestine	male adult (34 years)	ENCSR719HRO	ENCFF894PUN		Y		Y
small intestine	male child (3 years)	ENCSR618IQY	ENCFF484FAQ			Y	Y
spleen	female adult (30 years)	ENCSR510PSL	ENCFF717MVQ				
spleen	male adult (34 years)	ENCSR910QOX	ENCFF918SPI			Y	Y
spleen	male child (3 years)	ENCSR663IOE	ENCFF618CGM				
stomach	female adult (30 years)	ENCSR980UEY	ENCFF056CIU		Y		
stomach	male adult (34 years)	ENCSR721HDG	ENCFF268XDH			Y	Y
stomach	male child (3 years)	ENCSR922VBO	ENCFF525ANA	Y			Y
thymus	male child (3 years)	ENCSR775KCE	ENCFF401QEF		Y	Y	Y
urinary bladder	male child (3 years)	ENCSR448DCX	ENCFF934KPK			Y	Y

Supplementary Table 7. Number of RNA-seq alignments before and after filtering for intergenic regions. BAM accession IDs correspond to the ones in Supplementary Table 1. Uni- and multi-mapped fragments were defined by whether their BAM NH tags were equal or higher to 1.

RNA-seq BAM accession ID	Number of RNA-seq fragments (million)			
	Uniquely mapping		Multi-mapping	
	ENCODE	Intergenic	ENCODE	Intergenic
ENCFF044SJL	37.38	0.37	4.50	0.10
ENCFF048ODN	84.81	0.82	10.10	0.14
ENCFF125RAL	73.97	0.25	10.21	0.08
ENCFF207ZSA	88.73	0.63	16.02	0.10
ENCFF244ZQA	103.28	1.11	10.89	0.13
ENCFF263OLY	116.56	0.20	9.26	0.04
ENCFF306YQS	16.46	0.04	1.50	0.02
ENCFF315VHI	98.30	0.47	8.42	0.10
ENCFF343WEZ	96.53	0.77	7.45	0.11
ENCFF367VEP	95.76	0.76	11.77	0.20
ENCFF381BQZ	87.32	0.85	10.96	0.15
ENCFF428VBU	76.24	0.32	10.65	0.05
ENCFF444SCT	94.03	0.60	8.08	0.08
ENCFF521KYZ	19.45	0.04	1.77	0.02
ENCFF547YFO	35.02	0.17	2.62	0.03
ENCFF588YLF	53.76	0.37	4.22	0.05
ENCFF709IUX	88.74	1.35	9.03	0.16
ENCFF728JKQ	38.02	0.35	4.71	0.10
ENCFF739OVZ	96.35	0.31	9.28	0.09
ENCFF782IVX	103.55	0.43	8.02	0.07
ENCFF782TAX	56.15	0.12	3.67	0.03
ENCFF800YJR	12.97	0.02	0.95	0.01
ENCFF802TLC	75.75	0.30	14.73	0.05
ENCFF834ITU	97.18	0.42	8.21	0.09
ENCFF838JGD	47.79	0.22	4.12	0.04
ENCFF846WOV	39.14	0.23	3.14	0.03
ENCFF904OHO	47.93	0.17	3.47	0.03
ENCFF912SZP	69.56	0.44	14.49	0.07
ENCFF978ACT	82.19	0.19	7.14	0.04
ENCFF983FHE	99.67	0.67	12.10	0.16

Supplementary Table 8. RNA-seq BAM file sizes before and after filtering fragments for intergenic regions. BAM accession IDs corresponds to the ones in Supplementary Table 1.

RNA-seq BAM accession ID	BAM file size (GB)	
	ENCODE	Intergenic
ENCFF044SJL	4.324	0.044
ENCFF048ODN	18.469	0.140
ENCFF125RAL	17.758	0.055
ENCFF207ZSA	25.198	0.101
ENCFF244ZQA	20.418	0.174
ENCFF263OLY	20.138	0.037
ENCFF306YQS	1.707	0.005
ENCFF315VHI	17.777	0.077
ENCFF343WEZ	18.084	0.123
ENCFF367VEP	22.498	0.153
ENCFF381BQZ	19.791	0.149
ENCFF428VBU	18.016	0.050
ENCFF444SCT	18.066	0.095
ENCFF521KYZ	2.008	0.006
ENCFF547YFO	9.593	0.033
ENCFF588YLF	12.434	0.075
ENCFF709IUX	18.162	0.211
ENCFF728JKQ	4.559	0.043
ENCFF739OVZ	17.526	0.059
ENCFF782IVX	24.285	0.088
ENCFF782TAX	12.220	0.030
ENCFF800YJR	3.083	0.006
ENCFF802TLC	22.013	0.049
ENCFF834ITU	17.227	0.070
ENCFF838JGD	11.608	0.050
ENCFF846WOV	9.675	0.047
ENCFF904OHO	10.775	0.038
ENCFF912SZP	21.608	0.070
ENCFF978ACT	15.930	0.038
ENCFF983FHE	21.377	0.117

Supplementary Table 9. Computing time and memory usage for '1-Step' and '2-Step' methods. Each method was ran on 2.1 GHz AMD CPUs using eight threads.

Method	Time cost (minute)	Memory usage (MB)
pooling + Cufflinks	219	594
pooling + StringTie	7	151
Cufflinks + Cuffmerge	150	155
Cufflinks + TACO	145	162
StringTie + merging	5	156

Supplementary Table 10. TPMs of two 'eliminated' PRAM transcripts in the 30 ENCODE RNA-seq datasets. They got eliminated because they have TPM = 0 in at least one RNA-seq replicate from each of the seven cell lines.

cell line	replicate index	BAM ID	plcf_chr1_minus.82.1	plcf_chr9_plus.10254.3
A549	1	ENCFF125RAL	0.00	0.00
A549	2	ENCFF739OVZ	0.00	0.00
GM12878	1	ENCFF802TLC	0.38	0.00
GM12878	2	ENCFF428VBU	0.12	0.00
GM12878	3	ENCFF547YFO	0.11	0.00
GM12878	4	ENCFF782IVX	0.22	0.00
GM12878	5	ENCFF709IUX	0.01	0.00
GM12878	6	ENCFF244ZQA	0.00	0.00
HeLa-S3	1	ENCFF343WEZ	0.00	0.00
HeLa-S3	2	ENCFF444SCT	0.00	0.00
HepG2	1	ENCFF315VHI	0.00	0.00
HepG2	2	ENCFF834ITU	0.01	0.00
HepG2	3	ENCFF306YQS	0.00	0.00
HepG2	4	ENCFF521KYZ	0.00	0.00
HepG2	5	ENCFF800YJR	0.00	0.00
HepG2	6	ENCFF782TAX	0.00	0.00
K562	1	ENCFF912SZP	0.00	0.00
K562	2	ENCFF207ZSA	0.00	1.12
K562	3	ENCFF846WOV	0.00	0.77
K562	4	ENCFF588YLF	0.00	1.70
K562	5	ENCFF048ODN	0.00	0.00
K562	6	ENCFF381BQZ	0.00	0.00
K562	7	ENCFF044SJL	0.00	1.20
K562	8	ENCFF728JKQ	0.00	1.17
MCF-7	1	ENCFF367VEP	0.00	0.00
MCF-7	2	ENCFF983FHE	0.00	0.00
MCF-7	3	ENCFF904OHO	0.00	0.00
MCF-7	4	ENCFF838JGD	0.00	0.00
SK-N-SH	1	ENCFF263OLY	0.00	0.00
SK-N-SH	2	ENCFF978ACT	0.00	0.00

Supplementary Table 11. Number of PRAM transcripts before and after elimination. Transcripts that had inconsistent expression states and got eliminated were labelled as 'TPM=0'. For GENCODE transcripts, we first removed short ones that had a single exon or genomic span shorter than 200 bp. These short transcripts were labeled as 'short'. None of PRAM transcripts fits into this short criteria.

source	total	short	TPM=0	kept
GENCODE: newly discovered	1,034	748	178	108
GENCODE: long-standing	197,167	26,262	61,738	109,167
PRAM	14,226	0	8,837	5,389

Supplementary Table 12. Number of GENCODE and PRAM transcripts by TPM range. Transcript models were predicted based on the 30 human RNA-seq datasets in Supplementary Table 1.

category	total	TPM range	GM12878			K562		
			by TPM range*	promoter mappability ≥ 0.8†	transcript mappability ≥ 0.8‡	by TPM range*	promoter mappability ≥ 0.8†	transcript mappability ≥ 0.8‡
GENCODE: long-standing	197,167	< 0.1	88,300	76,132	74,685	84,569	73,472	72,012
		[0.1, 1)	2,062	1,882	1,872	1,973	1,796	1,792
		>= 1	19,878	18,767	18,415	22,081	20,675	20,240
		indeterminate	86,927	80,164	78,786	88,544	81,002	79,714
GENCODE: newly discovered	1,034	< 0.1	795	531	491	751	517	479
		[0.1, 1)	17	12	12	12	8	8
		>= 1	17	6	6	31	6	7
		indeterminate	205	118	122	240	136	137
pooling + Cufflinks	14,226	< 0.1	9,873	7,085	7,758	10,526	8,129	8,669
		[0.1, 1)	135	88	92	158	106	118
		>= 1	30	20	23	48	27	34
		indeterminate	4,188	3,157	3,382	3,494	2,088	2,434

* Transcripts and models were stratified by their expression levels in the six GM12878 and eight K562 RNA-seq datasets. Transcripts or models that were classified into TPM < 0.1, 0.1 ≤ TPM < 1, or TPM ≥ 1 were required to have all of their TPMs for the corresponding cell line within this range. Otherwise, they were classified as 'indeterminate'.
† A transcript or model's promoter mappability was based on the 500 bp region flanking its transcription start site, where RAMPAGE signal was calculated.
‡ A transcript or model's mappability on the region including all of its exons and introns, where histone modification ChIP-seq signal was calculated.

Supplementary Table 13. ENCODE RAMPAGE bigWig files. Accession IDs and metadata were from https://www.encodeproject.org.

Cell	Accession ID	Biological replicate index	File type
GM12878	ENCFF039WHT	1	plus strand signal of unique reads
	ENCFF143FSY	1	minus strand signal of unique reads
	ENCFF707RLJ	2	plus strand signal of unique reads
	ENCFF354OFJ	2	minus strand signal of unique reads
K562	ENCFF783EAC	1	plus strand signal of unique reads
	ENCFF518WII	1	minus strand signal of unique reads
	ENCFF663DTD	2	plus strand signal of unique reads
	ENCFF809GTW	2	minus strand signal of unique reads

Supplementary Table 14. ENCODE histone modification ChIP-seq datasets. Accession IDs and metadata were from https://www.encodeproject.org.

BAM accession ID	Cell	Histone mark	Biological replicate index
ENCFF958QVX	GM12878	H3K36me3	1
ENCFF460TXJ	GM12878	H3K36me3	2
ENCFF676NDU	GM12878	H3K79me2	1
ENCFF231YZJ	GM12878	H3K79me2	2
ENCFF639PLN	K562	H3K36me3	1
ENCFF673KBG	K562	H3K36me3	2
ENCFF947DVY	K562	H3K79me2	1
ENCFF408YHI	K562	H3K79me2	2

Supplementary Table 15. Number of conserved GENCODE and PRAM transcripts. GENCODE (version 24) and PRAM transcripts were mapped from human genome (hg38) to mouse genome (mm10) using the liftOver function from Bioconductor package rtracklayer. Human GENCODE transcripts were divided into 'long-standing' and 'newly discovered' by whether they overlapped with transcripts from the oldest available GENCODE (version 20) annotation for hg38. A transcript was considered as 'conserved' if its genomic span mapped to the same chromosome on the same strand in mouse. A 'conserved' transcript was further examined to see whether it overlapped with any mouse GENCODE (vM19) transcripts.

transcript type	human GENCODE		PRAM
	long-standing	newly discovered	
total	197,167	1,034	14,226
conserved	143,013 (72.5%)	555 (53.7%)	9,164 (64.4%)
conserved and overlapping with mouse GENCODE	127,137 (64.5%)	173 (16.7%)	1,170 (8.2%)

Supplementary Table 16. Number of BLAST-matched proteins for PRAM transcripts. All the 14,226 master set transcripts were aligned to the mammalian protein sequences (taxonomy ID: 40674) using BLAST against the non-redundant protein sequences databases (downloaded on Dec. 14th, 2018). The alignment was performed by blastx (version 2.7.1+) requiring a maximum e-value of 10^{-15} and searching in the orientation as transcript's 5'- to 3'-end. All the other options were set to default. A matched protein was required to contain ≥ 60 amino acids and ≥ 75% of its sequence was aligned. These criteria have been used previously to compile the CHESS human gene catalog.

number range of matched proteins	transcript models	
	number	percentage
0	9,823	69.05
[1, 10]	1,782	12.53
(10, 50]	1,002	7.04
(50, 100]	708	4.98
>100	911	6.40

Supplementary Table 17. PRAM transcript plcf_chr2_minus.9034.1's matched proteins by BLAST.

species	name	ID
Aotus nancymaae	LOW QUALITY PROTEIN: putative uncharacterized protein encoded by LINC00596, partial	XP_021531542
Callithrix jacchus	PREDICTED: putative uncharacterized protein encoded by LINC00269, partial	XP_017819497
Gorilla gorilla gorilla	PREDICTED: ribosome biogenesis protein BMS1 homolog	XP_018888574
Gorilla gorilla gorilla	PREDICTED: ribosome biogenesis protein BMS1 homolog isoform X1	XP_018889689
Gorilla gorilla gorilla	PREDICTED: ribosome biogenesis protein BMS1 homolog isoform X2	XP_018889690
Gorilla gorilla gorilla	PREDICTED: ribosome biogenesis protein BMS1 homolog, partial	XP_018875818
Homo sapiens	FAM175A protein	AAH16905
Homo sapiens	PRO1902	AAF22026
Homo sapiens	hCG1814039, partial	EAW68953
Homo sapiens	hCG1817437	EAW47553
Homo sapiens	hCG1818479	EAW95069
Homo sapiens	hCG1979495	EAW55411
Homo sapiens	hCG2038438, partial	EAW65538
Homo sapiens	hCG2038961, partial	EAW48306
Homo sapiens	hCG2039009, partial	EAW64637
Homo sapiens	hCG2039054, partial	EAW89122
Homo sapiens	hCG2039105, partial	EAX04768
Homo sapiens	hCG2039110, partial	EAX06591
Homo sapiens	hCG2042258, partial	EAW75601
Homo sapiens	hCG2042307	EAW98491
Homo sapiens	ribosome biogenesis protein BMS1 homolog isoform X4	XP_011516396
Homo sapiens	unnamed protein product	BAB15056
Homo sapiens	unnamed protein product	BAH12795
Homo sapiens	unnamed protein product	BAC85209
Homo sapiens	unnamed protein product	BAC04333
Macaca fascicularis	Putative BMS1-like protein ENSP00000383088, partial	EHH64667
Macaca fascicularis	hypothetical protein EGM_00005, partial	EHH62889
Macaca fascicularis	hypothetical protein EGM_00324, partial	EHH49632
Macaca fascicularis	hypothetical protein EGM_01641	EHH50766
Macaca fascicularis	hypothetical protein EGM_01642, partial	EHH50767
Macaca fascicularis	hypothetical protein EGM_01778, partial	EHH50883
Macaca fascicularis	hypothetical protein EGM_01780, partial	EHH50885
Macaca fascicularis	hypothetical protein EGM_03478, partial	EHH66476

Macaca fascicularis	hypothetical protein EGM_03798, partial	EHH66749
Macaca fascicularis	hypothetical protein EGM_04619, partial	EHH55411
Macaca fascicularis	hypothetical protein EGM_04788, partial	EHH55556
Macaca fascicularis	hypothetical protein EGM_04997, partial	EHH55734
Macaca fascicularis	hypothetical protein EGM_08759, partial	EHH58816
Macaca fascicularis	hypothetical protein EGM_08825, partial	EHH58869
Macaca fascicularis	hypothetical protein EGM_09292, partial	EHH59230
Macaca fascicularis	hypothetical protein EGM_09449, partial	EHH59362
Macaca fascicularis	hypothetical protein EGM_10210, partial	EHH59972
Macaca fascicularis	hypothetical protein EGM_11255, partial	EHH51808
Macaca fascicularis	hypothetical protein EGM_11598, partial	EHH60270
Macaca fascicularis	hypothetical protein EGM_11981, partial	EHH60591
Macaca fascicularis	hypothetical protein EGM_12341, partial	EHH51985
Macaca fascicularis	hypothetical protein EGM_12528, partial	EHH52138
Macaca fascicularis	hypothetical protein EGM_14979, partial	EHH54194
Macaca fascicularis	hypothetical protein EGM_15018, partial	EHH54230
Macaca fascicularis	hypothetical protein EGM_15176, partial	EHH54354
Macaca fascicularis	hypothetical protein EGM_15972, partial	EHH63076
Macaca fascicularis	hypothetical protein EGM_16090, partial	EHH63176
Macaca fascicularis	hypothetical protein EGM_17106, partial	EHH64004
Macaca fascicularis	hypothetical protein EGM_17177, partial	EHH64058
Macaca fascicularis	hypothetical protein EGM_17267, partial	EHH64131
Macaca fascicularis	hypothetical protein EGM_17802, partial	EHH64557
Macaca fascicularis	hypothetical protein EGM_17881	EHH64622
Macaca fascicularis	hypothetical protein EGM_18770, partial	EHH60881
Macaca fascicularis	hypothetical protein EGM_19342, partial	EHH61346
Macaca fascicularis	unnamed portein product	BAB01630
Macaca fascicularis	unnamed protein product	BAE89602
Macaca fascicularis	unnamed protein product	BAE89854
Macaca fascicularis	unnamed protein product	BAE89454
Macaca mulatta	hypothetical protein EGK_00351, partial	EHH14429
Macaca mulatta	hypothetical protein EGK_01319, partial	EHH15253
Macaca mulatta	hypothetical protein EGK_01586, partial	EHH15486
Macaca mulatta	hypothetical protein EGK_02088, partial	EHH15918
Macaca mulatta	hypothetical protein EGK_02111, partial	EHH15935
Macaca mulatta	hypothetical protein EGK_02411, partial	EHH19699
Macaca mulatta	hypothetical protein EGK_03041, partial	EHH20232
Macaca mulatta	hypothetical protein EGK_03509, partial	EHH20620

Macaca mulatta	hypothetical protein EGK_03652, partial	EHH20736
Macaca mulatta	hypothetical protein EGK_03802, partial	EHH20863
Macaca mulatta	hypothetical protein EGK_03909, partial	EHH20949
Macaca mulatta	hypothetical protein EGK_04085, partial	EHH21096
Macaca mulatta	hypothetical protein EGK_05144, partial	EHH21966
Macaca mulatta	hypothetical protein EGK_05194, partial	EHH22013
Macaca mulatta	hypothetical protein EGK_05619, partial	EHH22373
Macaca mulatta	hypothetical protein EGK_07177, partial	EHH23662
Macaca mulatta	hypothetical protein EGK_07262, partial	EHH23728
Macaca mulatta	hypothetical protein EGK_08749, partial	EHH24999
Macaca mulatta	hypothetical protein EGK_09403, partial	EHH29075
Macaca mulatta	hypothetical protein EGK_10762, partial	EHH30155
Macaca mulatta	hypothetical protein EGK_11692, partial	EHH16412
Macaca mulatta	hypothetical protein EGK_11851, partial	EHH16558
Macaca mulatta	hypothetical protein EGK_11888, partial	EHH16588
Macaca mulatta	hypothetical protein EGK_12471, partial	EHH31407
Macaca mulatta	hypothetical protein EGK_12542, partial	EHH31460
Macaca mulatta	hypothetical protein EGK_13122, partial	EHH31951
Macaca mulatta	hypothetical protein EGK_13267, partial	EHH16986
Macaca mulatta	hypothetical protein EGK_13278, partial	EHH16997
Macaca mulatta	hypothetical protein EGK_13471, partial	EHH17143
Macaca mulatta	hypothetical protein EGK_13706, partial	EHH17322
Macaca mulatta	hypothetical protein EGK_13768, partial	EHH17376
Macaca mulatta	hypothetical protein EGK_16111, partial	EHH26203
Macaca mulatta	hypothetical protein EGK_16433, partial	EHH26452
Macaca mulatta	hypothetical protein EGK_17439, partial	EHH27277
Macaca mulatta	hypothetical protein EGK_18274, partial	EHH27951
Macaca mulatta	hypothetical protein EGK_18276, partial	EHH27953
Macaca mulatta	hypothetical protein EGK_18756, partial	EHH28336
Macaca mulatta	hypothetical protein EGK_19508, partial	EHH18929
Macaca mulatta	hypothetical protein EGK_19530	EHH18945
Macaca mulatta	hypothetical protein EGK_19543, partial	EHH18952
Macaca mulatta	hypothetical protein EGK_19562, partial	EHH18962
Macaca mulatta	hypothetical protein EGK_19586, partial	EHH18977
Macaca mulatta	hypothetical protein EGK_20357, partial	EHH30617
Macaca mulatta	hypothetical protein EGK_20358, partial	EHH30618
Macaca mulatta	hypothetical protein EGK_20417, partial	EHH30664
Nomascus leucogenys	PREDICTED: LOW QUALITY PROTEIN: putative uncharacterized protein encoded by LINC00269, partial	XP_012353247

Nomascus leucogenys	PREDICTED: ribosome biogenesis protein BMS1 homolog, partial	XP_012365973
Pan troglodytes	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta	BAK62850
Pan troglodytes	ribosome biogenesis protein BMS1 homolog	XP_024208465
Papio anubis	putative uncharacterized protein encoded by LINC00269, partial	XP_009203223
Piliocolobus tephrosceles	putative uncharacterized protein encoded by LINC00596, partial	XP_026311328
Pongo abelii	LOW QUALITY PROTEIN: IDNK isoform 1	PNJ71634

Supplementary Table 18. PRAM transcript plcf_chr2_minus.9034.2's matched proteins by BLAST.

species	name	ID
Aotus nancymaae	LOW QUALITY PROTEIN: putative uncharacterized protein encoded by LINC00596, partial	XP_021531542
Callithrix jacchus	PREDICTED: putative uncharacterized protein encoded by LINC00269, partial	XP_017819497
Gorilla gorilla gorilla	PREDICTED: putative uncharacterized protein encoded by LINC00269	XP_018871999
Gorilla gorilla gorilla	PREDICTED: ribosome biogenesis protein BMS1 homolog	XP_018888574
Gorilla gorilla gorilla	PREDICTED: ribosome biogenesis protein BMS1 homolog isoform X1	XP_018889689
Gorilla gorilla gorilla	PREDICTED: ribosome biogenesis protein BMS1 homolog isoform X2	XP_018889690
Gorilla gorilla gorilla	PREDICTED: ribosome biogenesis protein BMS1 homolog, partial	XP_018875818
Homo sapiens	FAM175A protein	AAH16905
Homo sapiens	PRO1902	AAF22026
Homo sapiens	hCG1814039, partial	EAW68953
Homo sapiens	hCG1817437	EAW47553
Homo sapiens	hCG1818479	EAW95069
Homo sapiens	hCG1979495	EAW55411
Homo sapiens	hCG2038438, partial	EAW65538
Homo sapiens	hCG2038961, partial	EAW48306
Homo sapiens	hCG2039009, partial	EAW64637
Homo sapiens	hCG2039054, partial	EAW89122
Homo sapiens	hCG2039105, partial	EAX04768
Homo sapiens	hCG2039110, partial	EAX06591
Homo sapiens	hCG2042258, partial	EAW75601
Homo sapiens	hCG2042307	EAW98491
Homo sapiens	ribosome biogenesis protein BMS1 homolog isoform X4	XP_011516396
Homo sapiens	unnamed protein product	BAB15056
Homo sapiens	unnamed protein product	BAH12795
Homo sapiens	unnamed protein product	BAC85209
Homo sapiens	unnamed protein product	BAC04333
Macaca fascicularis	Putative BMS1-like protein ENSP00000383088, partial	EHH64667
Macaca fascicularis	hypothetical protein EGM_00005, partial	EHH62889
Macaca fascicularis	hypothetical protein EGM_00324, partial	EHH49632
Macaca fascicularis	hypothetical protein EGM_01641	EHH50766
Macaca fascicularis	hypothetical protein EGM_01642, partial	EHH50767
Macaca fascicularis	hypothetical protein EGM_01778, partial	EHH50883
Macaca fascicularis	hypothetical protein EGM_01780, partial	EHH50885

Macaca fascicularis	hypothetical protein EGM_03478, partial	EHH66476
Macaca fascicularis	hypothetical protein EGM_03798, partial	EHH66749
Macaca fascicularis	hypothetical protein EGM_04619, partial	EHH55411
Macaca fascicularis	hypothetical protein EGM_04788, partial	EHH55556
Macaca fascicularis	hypothetical protein EGM_04997, partial	EHH55734
Macaca fascicularis	hypothetical protein EGM_08759, partial	EHH58816
Macaca fascicularis	hypothetical protein EGM_08825, partial	EHH58869
Macaca fascicularis	hypothetical protein EGM_09292, partial	EHH59230
Macaca fascicularis	hypothetical protein EGM_09449, partial	EHH59362
Macaca fascicularis	hypothetical protein EGM_10210, partial	EHH59972
Macaca fascicularis	hypothetical protein EGM_11255, partial	EHH51808
Macaca fascicularis	hypothetical protein EGM_11598, partial	EHH60270
Macaca fascicularis	hypothetical protein EGM_11981, partial	EHH60591
Macaca fascicularis	hypothetical protein EGM_12341, partial	EHH51985
Macaca fascicularis	hypothetical protein EGM_12528, partial	EHH52138
Macaca fascicularis	hypothetical protein EGM_14979, partial	EHH54194
Macaca fascicularis	hypothetical protein EGM_15018, partial	EHH54230
Macaca fascicularis	hypothetical protein EGM_15176, partial	EHH54354
Macaca fascicularis	hypothetical protein EGM_15972, partial	EHH63076
Macaca fascicularis	hypothetical protein EGM_16090, partial	EHH63176
Macaca fascicularis	hypothetical protein EGM_17106, partial	EHH64004
Macaca fascicularis	hypothetical protein EGM_17177, partial	EHH64058
Macaca fascicularis	hypothetical protein EGM_17267, partial	EHH64131
Macaca fascicularis	hypothetical protein EGM_17802, partial	EHH64557
Macaca fascicularis	hypothetical protein EGM_17881	EHH64622
Macaca fascicularis	hypothetical protein EGM_18770, partial	EHH60881
Macaca fascicularis	hypothetical protein EGM_19342, partial	EHH61346
Macaca fascicularis	unnamed portein product	BAB01630
Macaca fascicularis	unnamed protein product	BAE89602
Macaca fascicularis	unnamed protein product	BAE89854
Macaca fascicularis	unnamed protein product	BAE89454
Macaca mulatta	hypothetical protein EGK_00351, partial	EHH14429
Macaca mulatta	hypothetical protein EGK_01319, partial	EHH15253
Macaca mulatta	hypothetical protein EGK_01586, partial	EHH15486
Macaca mulatta	hypothetical protein EGK_02088, partial	EHH15918
Macaca mulatta	hypothetical protein EGK_02111, partial	EHH15935
Macaca mulatta	hypothetical protein EGK_02411, partial	EHH19699
Macaca mulatta	hypothetical protein EGK_03041, partial	EHH20232

Macaca mulatta	hypothetical protein EGK_03509, partial	EHH20620
Macaca mulatta	hypothetical protein EGK_03652, partial	EHH20736
Macaca mulatta	hypothetical protein EGK_03802, partial	EHH20863
Macaca mulatta	hypothetical protein EGK_03909, partial	EHH20949
Macaca mulatta	hypothetical protein EGK_04085, partial	EHH21096
Macaca mulatta	hypothetical protein EGK_05144, partial	EHH21966
Macaca mulatta	hypothetical protein EGK_05194, partial	EHH22013
Macaca mulatta	hypothetical protein EGK_05619, partial	EHH22373
Macaca mulatta	hypothetical protein EGK_07177, partial	EHH23662
Macaca mulatta	hypothetical protein EGK_07262, partial	EHH23728
Macaca mulatta	hypothetical protein EGK_08749, partial	EHH24999
Macaca mulatta	hypothetical protein EGK_09403, partial	EHH29075
Macaca mulatta	hypothetical protein EGK_10762, partial	EHH30155
Macaca mulatta	hypothetical protein EGK_11692, partial	EHH16412
Macaca mulatta	hypothetical protein EGK_11851, partial	EHH16558
Macaca mulatta	hypothetical protein EGK_11888, partial	EHH16588
Macaca mulatta	hypothetical protein EGK_12471, partial	EHH31407
Macaca mulatta	hypothetical protein EGK_12542, partial	EHH31460
Macaca mulatta	hypothetical protein EGK_13122, partial	EHH31951
Macaca mulatta	hypothetical protein EGK_13267, partial	EHH16986
Macaca mulatta	hypothetical protein EGK_13278, partial	EHH16997
Macaca mulatta	hypothetical protein EGK_13471, partial	EHH17143
Macaca mulatta	hypothetical protein EGK_13706, partial	EHH17322
Macaca mulatta	hypothetical protein EGK_13768, partial	EHH17376
Macaca mulatta	hypothetical protein EGK_16111, partial	EHH26203
Macaca mulatta	hypothetical protein EGK_16433, partial	EHH26452
Macaca mulatta	hypothetical protein EGK_17439, partial	EHH27277
Macaca mulatta	hypothetical protein EGK_18274, partial	EHH27951
Macaca mulatta	hypothetical protein EGK_18276, partial	EHH27953
Macaca mulatta	hypothetical protein EGK_18756, partial	EHH28336
Macaca mulatta	hypothetical protein EGK_19508, partial	EHH18929
Macaca mulatta	hypothetical protein EGK_19530	EHH18945
Macaca mulatta	hypothetical protein EGK_19543, partial	EHH18952
Macaca mulatta	hypothetical protein EGK_19562, partial	EHH18962
Macaca mulatta	hypothetical protein EGK_19586, partial	EHH18977
Macaca mulatta	hypothetical protein EGK_20357, partial	EHH30617
Macaca mulatta	hypothetical protein EGK_20358, partial	EHH30618
Macaca mulatta	hypothetical protein EGK_20417, partial	EHH30664

Nomascus leucogenys	PREDICTED: LOW QUALITY PROTEIN: putative uncharacterized protein encoded by LINC00269, partial	XP_012353247
Nomascus leucogenys	PREDICTED: ribosome biogenesis protein BMS1 homolog, partial	XP_012365973
Pan troglodytes	LOW QUALITY PROTEIN: IDNK isoform 2	PNI62172
Pan troglodytes	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta	BAK62850
Pan troglodytes	ribosome biogenesis protein BMS1 homolog	XP_024208465
Papio anubis	putative uncharacterized protein encoded by LINC00269, partial	XP_009203223
Piliocolobus tephrosceles	putative uncharacterized protein encoded by LINC00596, partial	XP_026311328
Pongo abelii	LOW QUALITY PROTEIN: IDNK isoform 1	PNJ71634

Supplementary Table 19. Number of PRAM transcripts stratified by the fractions of their exons that overlap with PhyloCSF-predicted coding regions. Overlap is required to be on the same strand regardless of frame.

range of fraction	number of transcripts
0	13730
(0, 0.1]	432
(0.1, 0.2]	27
(0.2, 0.3]	12
(0.3, 0.4]	11
(0.4, 0.5]	3
(0.5, 0.6]	8
(0.6, 0.7]	1
(0.7, 0.8]	2
(0.8, 0.9]	0
(0.9, 1]	0

Supplementary Table 20. Number of human transcripts predicted by '1-Step' and '2-Step' methods. Models were predicted by '1-Step' method ('pooling + Cufflinks') and '2-Step' methods ('Cufflinks + Cuffmerge'; 'Cufflinks + TACO') based on the 30 human RNA-seq datasets in Supplementary Table 1. Expression levels of models were determined by the six GM12878 RNA-seq datasets and the eight K562 RNA-seq datasets listed in Supplementary Table 1. The number of models that were shown as points in Supplementary Figure 33 and Supplementary Figure 34 are highlighted. For 'pooling + Cufflinks', the highlighted thirteen models with TPMs in '[0.1, 1)' in GM12878 were not identical and differed by two in each mappability selection category. Similarly, the highlighted seventeen models with TPMs in '[0.1, 1)' K562 were not identical either and differed by one in each category. The highlighted model with TPM ≥ 1 in GM12878 is the same one in each mappability selection category.

method	master list	method specific[§]	TPM range	GM12878			K562		
				by TPM range[*]	promoter mappability ≥ 0.8[†]	transcript mappability ≥ 0.8[‡]	by TPM range[*]	promoter mappability ≥ 0.8[†]	transcript mappability ≥ 0.8[‡]
pooling + Cufflinks	14,226	3,082	< 0.1	2,160	1,100	1,359	2,175	1,244	1,463
			[0.1, 1)	28	13	13	25	17	17
			≥ 1	1	1	1	3	0	0
			indeterminate	893	510	594	879	363	487
Cufflinks + Cuffmerge	8,779	251	< 0.1	157	115	124	156	126	135
			[0.1, 1)	5	4	4	1	1	1
			≥ 1	1	1	1	0	0	0
			indeterminate	88	79	81	94	72	74
Cufflinks + TACO	10,147	476	< 0.1	297	171	200	304	199	232
			[0.1, 1)	7	3	5	4	1	1
			≥ 1	6	3	1	6	1	3
			indeterminate	166	103	114	162	79	84

[§] Models that were predicted by only one method and not by the other two. Their genomic spans did not overlap with any other model's genomic span on the same strand predicted by the other two methods.
[*] Defined in the same way as Supplementary Table 12.
[†] Defined in the same way as Supplementary Table 12.
[‡] Defined in the same way as Supplementary Table 12.

Supplementary Table 21. Hematopoietic mouse ENCODE RNA-seq datasets. Each accession contains two RNA-seq replicates and there are 32 RNA-seq datasets in total. All of the datasets are paired-end on untreated cells related to hematopoiesis. In order to include as many datasets as possible, both RNA-seq and poly(A) mRNA RNA-seq data were collected.

Accession[1]	Cell	Assay
ENCSR767VHR	CMP	RNA-seq
ENCSR826IXR	G1E	RNA-seq
ENCSR000CHV	G1E	poly(A) mRNA RNA-seq
ENCSR000CHY	G1E-ER4	poly(A) mRNA RNA-seq
ENCSR833HPM	GMP	RNA-seq
ENCSR549QME	MEP	RNA-seq
ENCSR661TLW	erythroblast	RNA-seq
ENCSR000CHS	erythroblast	poly(A) mRNA RNA-seq
ENCSR558PXY	erythroid progenitor cell	RNA-seq
ENCSR000CHU	hematopoietic multipotent progenitor cell	poly(A) mRNA RNA-seq
ENCSR236ZIE	hematopoietic stem cell	RNA-seq
ENCSR000CHT	leukemia stem cell	poly(A) mRNA RNA-seq
ENCSR340NCF	megakaryocyte	RNA-seq
ENCSR000CIC	megakaryocyte	poly(A) mRNA RNA-seq
ENCSR848LXY	megakaryocyte progenitor cell	RNA-seq
ENCSR000CIF	megakaryocyte-erythroid progenitor cell	poly(A) mRNA RNA-seq

[1]ENCODE RNA-seq experiment accession ID (https://www.encodeproject.org)

Supplementary Table 22. Mouse hematopoiesis-related RNA-seq datasets.

Name	Source	Condition A	Condition B	Accession[1]	Reference
AGM	aorta-gonad-mesonephros	wild type	*Gata2* +9.5 enhancer deletion	N/A	Gao et al. 2013
fetal livers	fetal livers	wild type	*Gata2* -77 enhancer knockout	GSE69786	Johnson et al. 2015
G1E	G1E-ER-GATA1	untreated	β-estradiol treated	GSE74371	Tanimura et al. 2016
ES	pluripotent embryonic stem cell	wild type	nuclear RNase (*Exosc10*) mutant	SRP042355	Pefanis et al. 2015

[1]Accession ID for Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) or Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra)

Supplementary Table 23. Number of selected PRAM mouse gene and transcript models.

Selection step	Number of models	
	Gene	Transcript
a transcript has ≥ 2 exons and with genomic span ≥ 200 bp	6969	8652
a gene does not overlap with any GENCODE or RefSeq gene on either strand and has mappability ≥ 0.8	2657	3189
a gene is differentially expressed in ≥ 2 hematopoiesis-related systems	10	18
a gene maps to one chromosome on one strand in hg38	7	14
a gene has all exons with mappability ≥ 0.001 and at least one exon with mappability ≥ 0.8	6	13

Supplementary Table 24. GATA2 and TAL1 mouse ChIP-seq datasets.

Accession[1]	Cell	Treatment	Antibody	Alias
GSE69776	416b	None	GATA2	416B_GATA2_Rep1
			TAL1	416B_TAL1_Rep1
			IgGR	416B_Input_Rep1
GSE22178	HPC7	None	GATA2	HPC7_GATA2_Rep1
			TAL1	HPC7_TAL1_Rep1
			IgG	HPC7_Input_Rep1
GSE31331	G1ME	None	GATA2	G1ME_GATA2_Rep1
				G1ME_GATA2_Rep2
			None	G1ME_Input_Rep1
GSE26031	Lin- bone marrow hematopoietic progenitor	None	GATA2	LIN_GATA2_Rep1
			TAL1	LIN_TAL1_Rep1
			IgG	LIN_Input_Rep1
				LIN_Input_Rep2
GSE29193	G1E	BMP	GATA2	G1Echb_GATA2_BMP_Rep1
			None	G1Echb_Input_BMP_Rep1
PRJEB2019	MEL	DMSO	TAL1	MELera_TAL1_DMSO_Rep1
			Input	MELera_Input_DMSO_Rep1
		None	TAL1	MELera_TAL1_Rep1
				MELera_TAL1_Rep2
			Input	MELera_Input_Rep1
				MELera_Input_Rep2
GSE36029	G1E	None	GATA2	G1Epsu_GATA2_Rep1
				G1Epsu_GATA2_Rep2
			TAL1	G1Epsu_TAL1_Rep1
				G1Epsu_TAL1_Rep2
			Input	G1Epsu_Input_Rep1
				G1Epsu_Input_Rep2
	G1E-ER4	diffProtD_24hr	GATA2	G1EER4psu_GATA2_Rep1
				G1EER4psu_GATA2_Rep2
			TAL1	G1EER4psu_TAL1_Rep1
				G1EER4psu_TAL1_Rep2
			Input	G1EER4psu_Input_Rep1
				G1EER4psu_Input_Rep2
	MEL	None	TAL1	MELpsu_TAL1_Rep1
				MELpsu_TAL1_Rep2
			Input	MELpsu_Input_Rep1
				MELpsu_Input_Rep2
	Erythroblast, ter119+ cells from liver	None	TAL1	ERYpsu_TAL1_Rep1
				ERYpsu_TAL1_Rep2
				ERYpsu_TAL1_Rep3
			Input	ERYpsu_Input_Rep1
				ERYpsu_Input_Rep2
	Megakaryocyte	None	TAL1	MEGpsu_TAL1_Rep1
				MEGpsu_TAL1_Rep2

				MEGpsu_TAL1_Rep3
				MEGpsu_TAL1_Rep4
			Input	MEGpsu_Input_Rep1
				MEGpsu_Input_Rep2
GSE30142	G1E	None	GATA2	G1Epsu2_GATA2_Rep1
				G1Epsu2_GATA2_Rep2
			TAL1	G1Epsu2_TAL1_Rep1
				G1Epsu2_TAL1_Rep2
			Input	G1Epsu2_Input_Rep1
	G1E-ER4	E2 24 hrs	GATA2	G1EER4psu2_GATA2_Rep1
				G1EER4psu2_GATA2_Rep2
			TAL1	G1EER4psu2_TAL1_Rep1
				G1EER4psu2_TAL1_Rep2
				G1EER4psu2_TAL1_Rep3
				G1EER4psu2_TAL1_Rep4
	Ter119+	None	TAL1	TERpsu2_TAL1_Rep1
				TERpsu2_TAL1_Rep2
			Input	TERpsu2_Input_Rep1
GSE18720	fetal liver erythroblast WT	None	TAL1	FLE_TAL1_Rep1
			Input	FLE_Input_Rep1
	fetal liver erythroblast RER mutant		TAL1	FLERER_TAL1_Rep1
			Input	FLERER_Input_Rep1

[1]Accession ID for Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) or BioProject (https://www.ncbi.nlm.nih.gov/bioproject/),

Supplementary Table 25. PCR primers and their sequences.

name	sequence
mouse	
CUFFm.chr12.32594 F	GGTGACTGTTAGGTACCATGTGGG
CUFFm.chr12.32594 R	GGAGGCTAGATTCCCATGGTAGC
CUFFm.chr12.33668 R	GAGCCTCAGCGACAAGGCC
CUFFm.chr12.33668 F	CAGCTGCCAGGACCACTCC
CUFFm.chr12 33668 1.2 F	GACCTGGGCTCTTCCACCC
CUFFm.chr12 33668 1.3 R	TGACGAGAGCCATCAGAAGC
CUFFm.chr12 33668 2.1 F 9	CTATCACACTCTTGCTGTCAATG
CUFFm.chr12 33668 2.2 F	GAGACACCTGGCAACAGTACTT
CUFFm.chr12 33668 2.2 R	GGCTACACCTAGAGCTGCTCTC
CUFFm.chr12 33668 2.3 R	CCAGACGGCAACCGTACAGT
CUFFm.chr17.20196 R	GATTCCTTCGATGGACGTGCC
CUFFm.chr17.20196 F	GACTGCCCACCCACCATTCT
CUFFp.chr10.20259 F2	TGTCCATGTCTGCATAGCGGT
CUFFp.chr10.20259 R2	GATGTCTGTGGGTATTGGCTC
CUFFp.chr12.15498 1F	GACTTCGGACCCTGCTTGTC
CUFFp.chr12.15498 2F	GCTATGCCATCCTGCCTGTC
CUFFp.chr12.15498 R	CAGAGCATGGGATGATGTCACC
CUFFm.chr10.13181 F	GGCAGCACGACCATGAGGC
CUFFm.chr10.13181 34R	CGCATCCACTTCTCGCAGTTAAC
CUFFm.chr10.13181 2R	CCACTGAGCCATCTCGCCAG
Gata1 e4 5F	GGTTCACCTGATGGAGCTTGA
Gata1 e4 5R	GGCCCAAGAAGCGAATGATT
Gata2 e6 R	GCACTTGGAGAGCTCCTCG
Gata2 e5 F	GGCACCTGTTGTGCAAATTGTCA
Pik3cg e6 F	CTGATCCCACAGTCCTATCC
Pik3cg e7 R	GGTCCAGAGATTCAGTCTCC
Prkar2b e4 F	ACCGATGATCAGAGAAACAGAT
Prkar2b e5 R	TCACCGTCATCACCTTGGTC
human	
CUFFm.chr7.6148 1.1 2.1 F1	CATCCCGTGGTGTTGAAGAG
CUFFm.chr7.6148 1.1 2.1 F2	ATACTTTCACCACAGCACACC
CUFFm.chr7.6148 1.1 2.1 R	GTGCGTCCCTGTTTGCCGC

CUFFm.chr7.6148 1.2 2.2 R	GGAGCACTCCCATGAATCCT
CUFFm.chr7.6148 1.2 F	GTAGAGACGGGGTTTCACAG
CUFFm.chr7.6148 1.3 R	CTGTCTAGTGACCTTGCAGC
CUFFm.chr7.6148 2.2 F	GCCTGCCCAGACCTTGGAC
CUFFm.chr7.6148 2.3 R	CCGTAGTCAGTCCCAGGTAC
PIK3CG e6 F	TGCCGATCCTACAGCCCTATC
PIK3CG e7 R	GATCCAAAGATTCAGTCTCCCA
PRKAR2B e4 F	ATCAAGGTGACGATGGTGACAACT
PRKAR2B e5 R	GTTCGCCGAAACTCCCACG

Supplementary Table 26. Expression levels of the six PRAM gene models. Expression levels are represented in TPM. Conditions that have a model's TPM ≥ 1 in all replicates were highlighted. Dataset names correspond to those in Supplementary Table 22.

Gene model ID	AGM						fetal livers					
	Mutant			WT			Mutant			WT		
	Rep1	Rep2	Rep3	Rep1	Rep2	Rep3	Rep1	Rep2	Rep3	Rep1	Rep2	Rep3
CUFFm.chr12.32594	0.02	0.03	0.04	3.14	3.91	1.36	0.00	0.14	0.18	1.45	0.50	0.49
CUFFm.chr12.33668	0.05	0.05	0.12	3.73	5.83	0.85	0.03	0.14	0.13	0.65	0.43	0.20
CUFFm.chr17.20196	0.20	0.24	0.11	0.38	0.50	0.31	0.14	0.40	0.35	0.56	0.33	0.41
CUFFp.chr10.20259	0.31	0.18	0.21	1.31	1.52	0.60	0.00	0.03	0.04	0.13	0.09	0.07
CUFFp.chr12.15498	0.05	0.16	0.01	0.90	1.06	0.16	0.01	0.11	0.05	0.19	0.15	0.01
CUFFm.chr10.13181	0.63	0.26	0.26	0.49	0.82	0.38	1.65	0.71	0.64	0.87	0.96	0.62

Gene model ID	G1E						ES			
	β-estradiol treated			untreated			KO		WT	
	Rep1	Rep2	Rep3	Rep1	Rep2	Rep3	Rep1	Rep2	Rep1	Rep2
CUFFm.chr12.32594	0.04	0.02	0.05	0.70	0.64	0.53	0.00	0.00	0.00	0.00
CUFFm.chr12.33668	59.31	61.94	61.20	2.73	2.57	2.44	0.00	0.00	0.00	0.00
CUFFm.chr17.20196	2.08	2.03	2.03	0.26	0.26	0.18	0.02	0.01	0.01	0.01
CUFFp.chr10.20259	25.29	25.86	26.03	10.68	10.27	10.34	0.01	0.01	0.00	0.00
CUFFp.chr12.15498	12.99	13.80	14.48	0.93	0.98	0.85	0.00	0.00	0.00	0.00
CUFFm.chr10.13181	0.11	0.00	0.16	0.36	0.83	0.95	0.54	0.46	1.28	1.15

Supplementary Table 27. Number of '2-Step' and Cufflinks models overlapping with the four validated '1-Step' models. Models were built either by '2-Step' methods ('Cufflinks + Cuffmerge' and 'Cufflinks + TACO') or by Cufflinks based on individual RNA-seq data sets (labeled by DCC accession ID and biological replicate index). The four PRAM models detected by semi-qRT-PCR are: CUFFm.chr12.33668, CUFFm.chr17.20196, CUFFp.chr10.20259, and CUFFp.chr12.15498.

Method	Number of gene models in each selection step			
	built	by DE[1]	by DE and conservation	by DE, conservation, and mappability
Cufflinks + Cuffmerge	4	3	2	2
Cufflinks + TACO	3	0	0	0
ENCSR000CHS.Rep1	3	1	1	1
ENCSR000CHS.Rep2	2	0	0	0
ENCSR000CHT.Rep1	1	1	1	1
ENCSR000CHT.Rep2	1	0	0	0
ENCSR000CHU.Rep1	4	0	0	0
ENCSR000CHU.Rep2	1	0	0	0
ENCSR000CHV.Rep1	5	0	0	0
ENCSR000CHV.Rep2	3	0	0	0
ENCSR000CHY.Rep1	5	0	0	0
ENCSR000CHY.Rep2	2	0	0	0
ENCSR000CIC.Rep1	2	0	0	0
ENCSR000CIC.Rep2	0	0	0	0
ENCSR000CIF.Rep1	2	0	0	0
ENCSR000CIF.Rep2	2	0	0	0
ENCSR236ZIE.Rep1	0	0	0	0
ENCSR236ZIE.Rep2	3	0	0	0
ENCSR340NCF.Rep1	1	0	0	0
ENCSR340NCF.Rep2	0	0	0	0
ENCSR549QME.Rep1	0	0	0	0
ENCSR549QME.Rep2	1	0	0	0
ENCSR558PXY.Rep1	1	1	1	1
ENCSR558PXY.Rep2	1	1	1	1
ENCSR661TLW.Rep1	3	1	1	1
ENCSR661TLW.Rep2	1	1	1	1
ENCSR767VHR.Rep1	0	0	0	0
ENCSR767VHR.Rep2	0	0	0	0
ENCSR826IXR.Rep1	3	1	1	1
ENCSR826IXR.Rep2	3	1	1	1
ENCSR833HPM.Rep1	0	0	0	0
ENCSR833HPM.Rep2	1	0	0	0
ENCSR848LXY.Rep1	1	0	0	0
ENCSR848LXY.Rep2	1	0	0	0

[1] Selection by DE requires that gene model is differentially expressed in ≥ 2 experiments.

Supplementary Table 28. '2-Step' methods and Cufflinks missed two of the four validated '1-Step' models. Labels for 'Method' and number of models denoted the same as Supplementary Table 27. For those methods shown in Supplementary Table 27 but not here, none of them has any gene model overlapping with the four validated gene models. None of '2-Step' or Cufflinks models overlapped with CUFFm.chr17.20196 or CUFFp.chr12.15498.

Method	CUFFm.chr12.33668	CUFFm.chr17.20196	CUFFp.chr10.20259	CUFFp.chr12.15498
Cufflinks + Cuffmerge	1	0	1	0
ENCSR000CHS.Rep1	1	0	0	0
ENCSR000CHT.Rep1	1	0	0	0
ENCSR558PXY.Rep1	1	0	0	0
ENCSR558PXY.Rep2	1	0	0	0
ENCSR661TLW.Rep1	1	0	0	0
ENCSR661TLW.Rep2	1	0	0	0
ENCSR826IXR.Rep1	1	0	0	0
ENCSR826IXR.Rep2	1	0	0	0

Supplementary Table 29. Protein-coding potential of PRAM mouse transcripts. Listed are the top ten mammalian proteins that PRAM mouse transcripts aligned to by blastx. Proteins were ranked by E-value and the fraction of aligned protein segment length over protein's total length. Blastx searches were carried out in the same way as in Supplementary Table 16. CUFFm.chr12.33668.1 had only one matched protein. CUFFm.chr13.33668.2 and CUFFp.chr12.15498.2 had 31 and 165 matched proteins, respectively.

aligned transcript				aligned protein							
ID	length	start	end	ID	name	species	E-value	fraction	length	start	end
CUFFm.chr12.33668.1	9047	4534	4737	EDL09413	mCG147326	*Mus musculus*	8.41E-21	0.821	84	16	84
CUFFm.chr12.33668.2	7944	1770	2168	EDL29766	mCG148020	*Mus musculus*	9.06E-56	0.950	140	1	133
		1854	2246	EDL14187	mCG147486	*Mus musculus*	1.34E-50	1.000	132	1	132
		1770	2105	CAA37650	ORF7	*Rattus norvegicus*	1.46E-50	1.000	112	1	112
		4230	4796	CAA29034	ORF1	*Rattus norvegicus*	1.04E-48	0.995	189	1	188
		1854	2168	BAE33613	unnamed protein product	*Mus musculus*	2.33E-48	0.827	127	1	105
		1770	2105	ACT99045	unknown	*Rattus norvegicus*	7.65E-48	1.000	112	1	112
		1770	2072	EDL11227	mCG133245, isoform CRA_a	*Mus musculus*	5.78E-37	0.990	102	1	101
		1770	2087	BAC29583	unnamed protein product	*Mus musculus*	1.55E-36	0.841	126	1	106
		3021	3431	EDK98251	mCG146853	*Mus musculus*	4.73E-35	0.915	153	14	153
		3099	3602	EFB21087	hypothetical protein PANDA_017931, partial	*Ailuropoda melanoleuca*	2.70E-34	1.000	171	1	171
CUFFp.chr12.15498.2	6380	2774	4192	EDL78838	rCG59047, partial	*Rattus norvegicus*	0	0.996	475	3	475
		2774	4192	EDL95042	rCG20251, partial	*Rattus norvegicus*	0	0.996	475	3	475
		3545	5479	CAA43592	unnamed protein product, partial	*Rattus norvegicus*	0	0.946	685	1	648
		3914	5479	CAA37646	ORF3	*Rattus norvegicus*	0	0.944	556	2	526
		3641	5479	EDM13183	rCG47246, partial	*Rattus norvegicus*	0	0.943	653	1	616
		2780	4192	CAA43595	unnamed protein product, partial	*Rattus norvegicus*	0	0.942	500	30	500
		2729	4036	CAA27363	unnamed protein product, partial	*Mus musculus*	0	0.936	466	12	447
		2777	5479	ELR58510	hypothetical protein M91_05513, partial	*Bos mutus*	0	0.772	1170	3	905
		4067	5326	EDL78640	rCG65853	*Rattus norvegicus*	0	0.766	552	130	552
		2777	5479	ELR51705	hypothetical protein M91_10420, partial	*Bos mutus*	0	0.766	1179	3	905

Supplementary Table 30. ENCODE K562 RNA-seq datasets. All the datasets are strand-specific and paired-end. To avoid bias, we required that data were from untreated K562 cells and were not from subcellular fractions, such as membrane, nucleus, cytosol, etc. We included all of the data sets labeled as RNA-seq or poly(A) mRNA RNA-seq so that we could pool a large collection of K562 samples.

Accession[1]	Assay	FASTQ ID[1]	Biological replicate index	Mate index
ENCSR000AEL	RNA-seq	ENCFF001RFF	1	1
ENCSR000AEL	RNA-seq	ENCFF001RFE	1	2
ENCSR000AEL	RNA-seq	ENCFF001RFD	2	1
ENCSR000AEL	RNA-seq	ENCFF001RFC	2	2
ENCSR000AEM	poly(A) mRNA RNA-seq	ENCFF001RED	1	1
ENCSR000AEM	poly(A) mRNA RNA-seq	ENCFF001RDZ	1	2
ENCSR000AEM	poly(A) mRNA RNA-seq	ENCFF001REG	2	1
ENCSR000AEM	poly(A) mRNA RNA-seq	ENCFF001REF	2	2
ENCSR000AEN	RNA-seq	ENCFF001RDC	1	1
ENCSR000AEN	RNA-seq	ENCFF001RCU	1	2
ENCSR000AEN	RNA-seq	ENCFF001RDB	2	1
ENCSR000AEN	RNA-seq	ENCFF001RCT	2	2
ENCSR000AEO	poly(A) mRNA RNA-seq	ENCFF001RDE	1	1
ENCSR000AEO	poly(A) mRNA RNA-seq	ENCFF001RCW	1	2
ENCSR000AEO	poly(A) mRNA RNA-seq	ENCFF001RDD	2	1
ENCSR000AEO	poly(A) mRNA RNA-seq	ENCFF001RCV	2	2
ENCSR000AEP	RNA-seq	ENCFF001RVV	1	1
ENCSR000AEP	RNA-seq	ENCFF001RWA	1	2
ENCSR000AEP	RNA-seq	ENCFF001RWD	2	1
ENCSR000AEP	RNA-seq	ENCFF001RVU	2	2
ENCSR000AEQ	poly(A) mRNA RNA-seq	ENCFF001RWF	1	1
ENCSR000AEQ	poly(A) mRNA RNA-seq	ENCFF001RWC	1	2
ENCSR000AEQ	poly(A) mRNA RNA-seq	ENCFF001RWE	2	1
ENCSR000AEQ	poly(A) mRNA RNA-seq	ENCFF001RWG	2	2
ENCSR000CPH	poly(A) mRNA RNA-seq	ENCFF000HFF	1	1
ENCSR000CPH	poly(A) mRNA RNA-seq	ENCFF000HFG	1	2
ENCSR000CPH	poly(A) mRNA RNA-seq	ENCFF000HFH	2	1
ENCSR000CPH	poly(A) mRNA RNA-seq	ENCFF000HFY	2	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DWT	1	1
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DWW	1	1
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DWV	1	1
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DWU	1	1
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DWX	1	1
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXL	1	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXM	1	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXN	1	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXP	1	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXO	1	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXC	2	1
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXF	2	1
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXD	2	1

ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXE	2	1
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXU	2	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXW	2	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXV	2	2
ENCSR000EYO	poly(A) mRNA RNA-seq	ENCFF000DXX	2	2
ENCSR109IQO	RNA-seq	ENCFF002DKA	1	1
ENCSR109IQO	RNA-seq	ENCFF002DKE	1	2
ENCSR109IQO	RNA-seq	ENCFF002DKF	2	1
ENCSR109IQO	RNA-seq	ENCFF002DKI	2	2
ENCSR545DKY	poly(A) mRNA RNA-seq	ENCFF059IUV	1	1
ENCSR545DKY	poly(A) mRNA RNA-seq	ENCFF104ZSG	1	2
ENCSR545DKY	poly(A) mRNA RNA-seq	ENCFF628GUZ	2	1
ENCSR545DKY	poly(A) mRNA RNA-seq	ENCFF695XOC	2	2
ENCSR885DVH	RNA-seq	ENCFF267RKD	1	1
ENCSR885DVH	RNA-seq	ENCFF455VYN	1	2
ENCSR885DVH	RNA-seq	ENCFF606ZTR	2	1
ENCSR885DVH	RNA-seq	ENCFF444KCV	2	2

[1]ENCODE RNA-seq experiment and file accession ID (https://www.encodeproject.org)

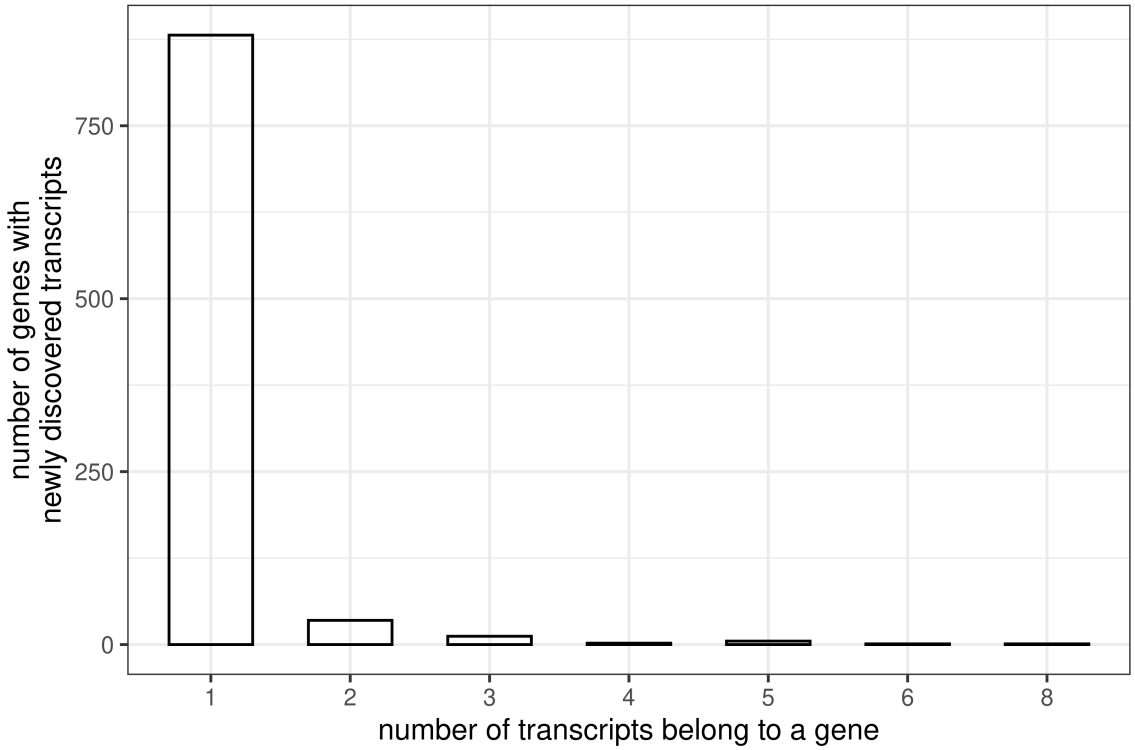
Supplementary Table 31. Protein-coding potential of PRAM mouse transcript's human counterpart. Listed are the top ten mammalian proteins that CUFFm.chr7.6148.1 aligned to by blastx. Proteins were ranked by E-value and the fraction of aligned protein segment length over protein's total length. Blastx searches were carried out in the same way as in Supplementary Table 16. CUFFm.chr7.6148.1 had sixteen matched proteins.

aligned transcript				aligned protein							
ID	**length**	**start**	**end**	**ID**	**name**	**species**	**E-value**	**fraction**	**length**	**start**	**end**
CUFFm.chr7.6148.1	6176	1006	1221	EHH59533	hypothetical protein EGM_09670, partial	*Macaca fascicularis*	3.81E-25	0.81	89	17	88
		1006	1221	EAX05977	hCG2038848, partial	*Homo sapiens*	1.14E-24	0.83	87	14	85
		972	1247	EHH64951	hypothetical protein EGM_18285, partial	*Macaca fascicularis*	1.23E-21	0.92	93	8	93
		5734	5949	EHH14529	hypothetical protein EGK_00471, partial	*Macaca mulatta*	3.13E-20	0.95	77	4	76
		5737	5952	EHH22132	hypothetical protein EGK_05340, partial	*Macaca mulatta*	2.43E-18	0.76	93	1	71
		990	1247	EHH22132	hypothetical protein EGK_05340, partial	*Macaca mulatta*	3.66E-18	0.97	93	4	93
		5737	5952	EHH21769	hypothetical protein EGK_04905, partial	*Macaca mulatta*	4.35E-18	0.80	88	1	70
		5737	5952	EHH59445	hypothetical protein EGM_09562, partial	*Macaca fascicularis*	1.77E-17	0.82	87	1	71
		5734	5952	EHH64647	hypothetical protein EGM_17921, partial	*Macaca fascicularis*	3.89E-17	0.81	88	6	76
		5737	5952	EHH29987	hypothetical protein EGK_10551, partial	*Macaca mulatta*	4.01E-17	0.93	75	2	71

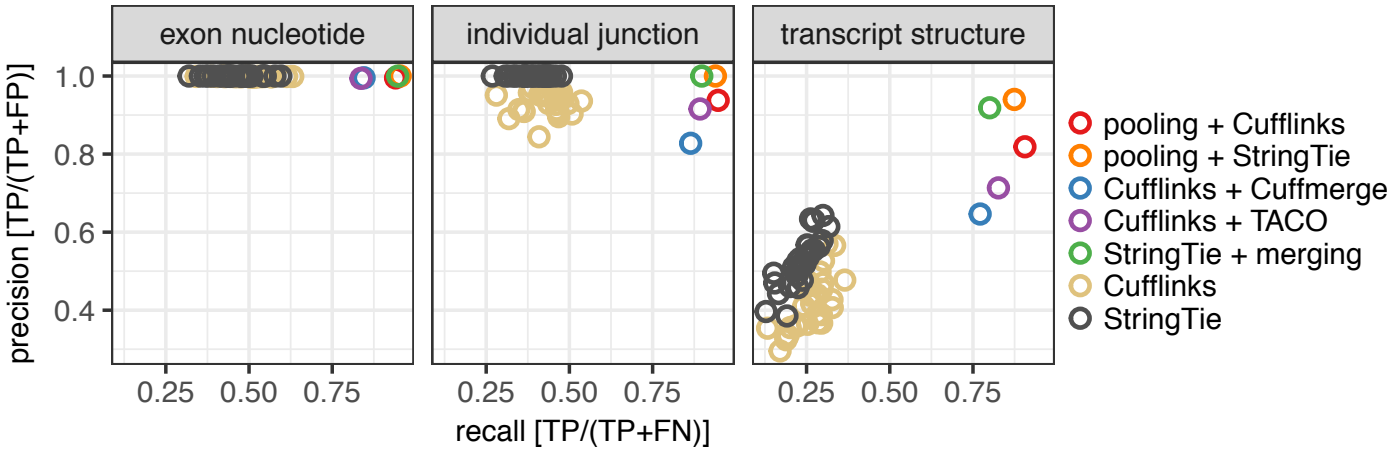
Supplementary Table 32. GATA2 and TAL1 human ChIP-seq datasets.

Accession[1]	Cell	Treatment	Antibody	Alias
GSE60792	Erythroid progenitors derived from human CD34+ bone marrow cells	DMSO	GATA2	CD34ace_GATA2_DMSO_Rep1
		ACY-957		CD34ace_GATA2_ACY957_Rep1
GSE45144	CD34+ Human Blood Stem/Progenitor Cells	None	SCL	CD34uns_TAL1_Rep1
			GATA2	CD34uns_GATA2_Rep1
			IgG	CD34uns_Input_Rep1
GSE29194	CD34+ progenitors	BMP	GATA2	CD34chb_GATA2_BMP_Rep1
			WCE	CD34chb_Input_BMP_Rep1
GSE31477	HUVEC	None	GATA2	HUVEC_GATA2_Rep1
				HUVEC_GATA2_Rep2
			Input	HUVEC_Input_Rep1
	K562	None	GATA2	K562usc_GATA2_Rep1
				K562usc_GATA2_Rep2
			Input	K562usc_Input_Rep1
	SH-SY5Y	None	GATA2	SHSY5Y_GATA2_Rep1
				SHSY5Y_GATA2_Rep2
			Input	SHSY5Y_Input_Rep1
	K562	None	TAL1	K562sta_TAL1_Rep1
				K562sta_TAL1_Rep2
			Input	K562sta_Input_Rep1
				K562sta_Input_Rep2
GSE31363	K562	None	GATA2	K562uch_GATA2_Rep1
				K562uch_GATA2_Rep2
			Input	K562uch_Input_Rep1
GSE32465	K562	None	GATA2	K562hai_GATA2_Rep1
				K562hai_GATA2_Rep2
			Input	K562hai_Input_Rep1
				K562hai_Input_Rep2
				K562hai_Input_Rep3
				K562hai_Input_Rep4

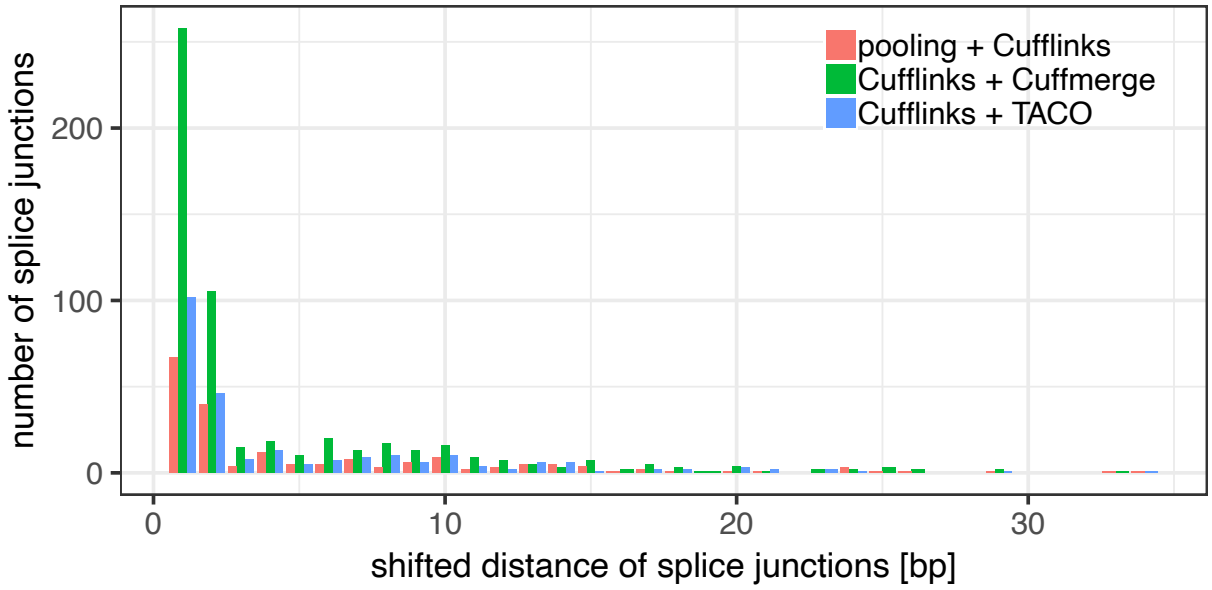
[1]Accession ID for Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/)



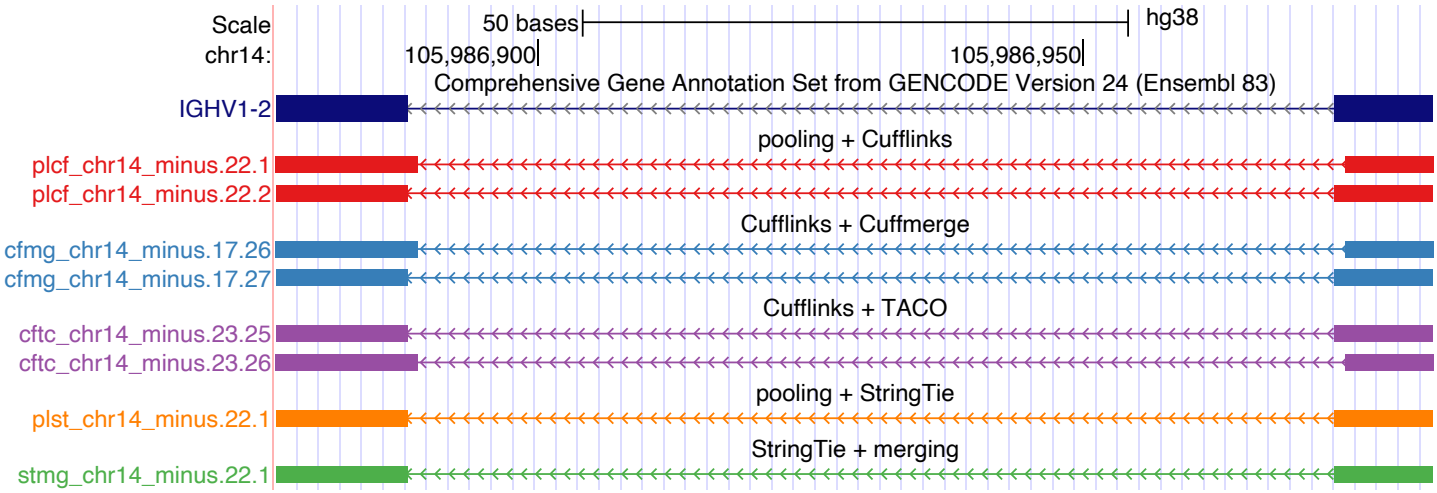
Supplementary Figure 1. Distribution of genes with 'newly discovered' transcripts. Genes were stratified by their number of 'newly discovered' transcripts. 'Newly discovered' genes were those that existed in GENCODE version 24, but not in GENCODE version 20. As described later in the manuscript, 'newly discovered' genes share similar features with intergenic transcripts. 94% of the genes (881 of 937) contain only one newly discovered transcripts. This high percentage suggested that the single-transcript genes we used in the benchmark dataset are representative for the intergenic transcripts we aimed to predict.



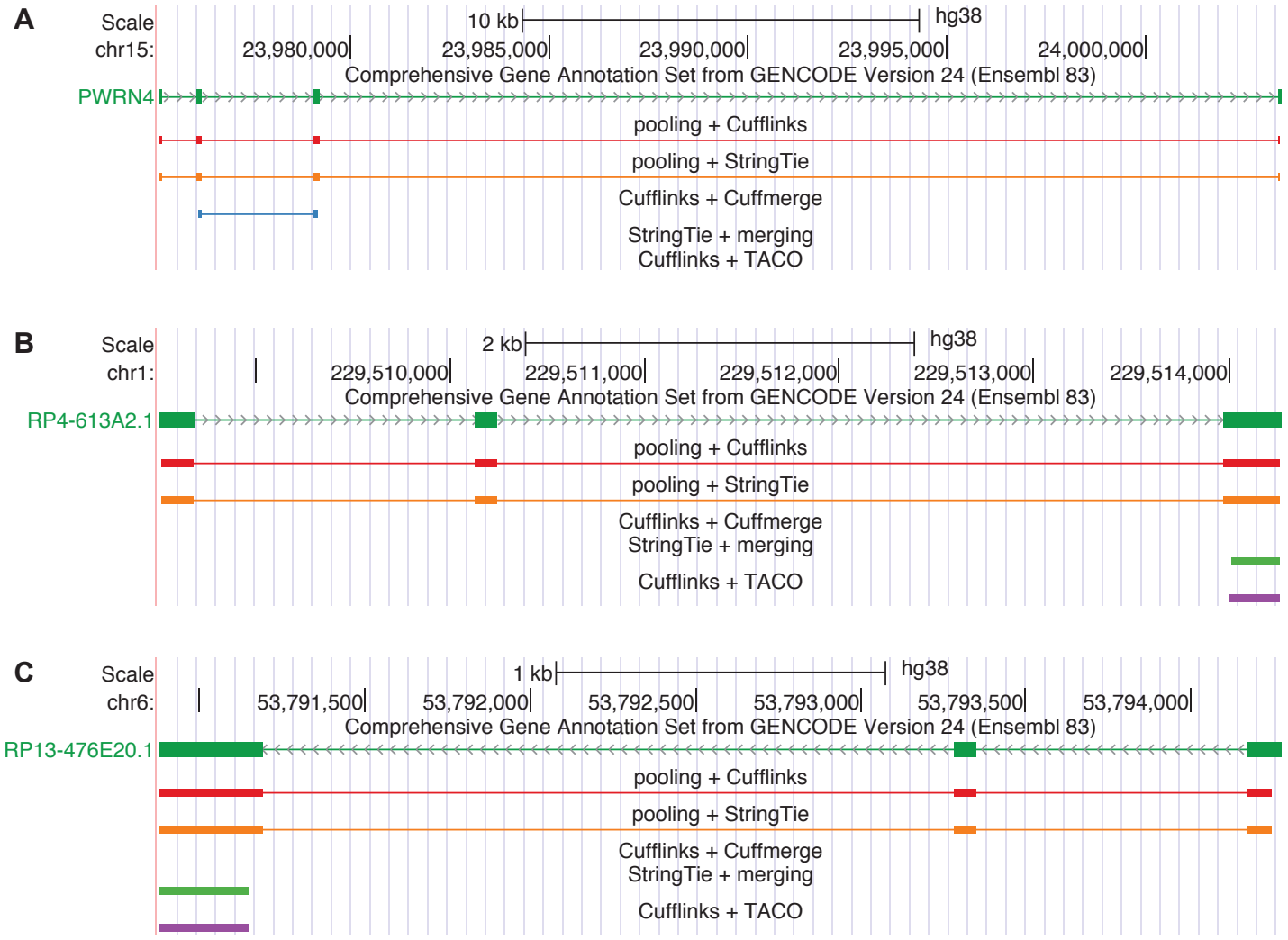
Supplementary Figure 2. Benchmark results of Cufflinks, StringTie, '1-Step' and '2-Step' methods.



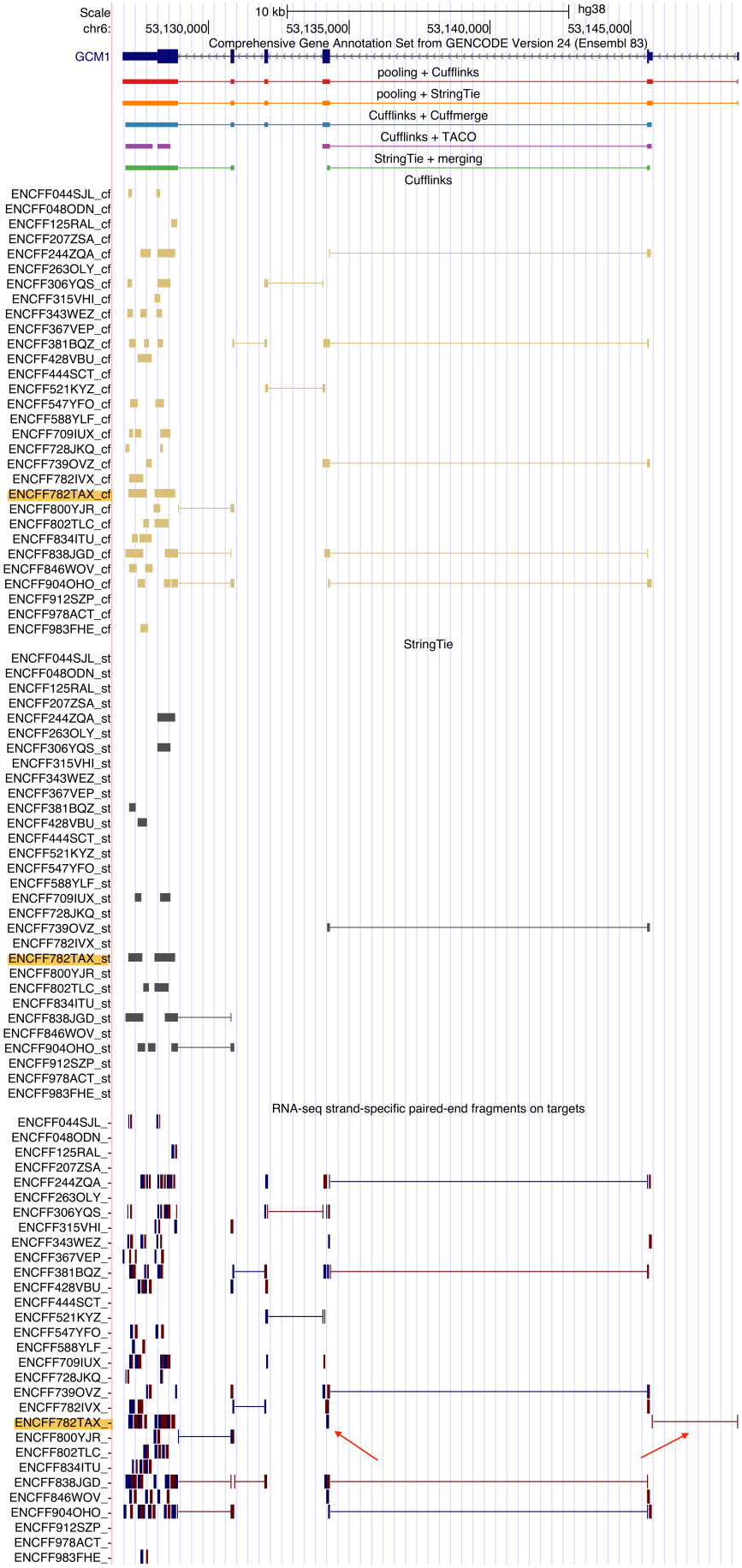
Supplementary Figure 3. Distribution of shift for false positive junctions by Cufflinks-based methods. False positive junctions shown here are those with both 5'- and 3'-splice sites shifted by the same number of base pairs compared to the benchmark transcripts.



Supplementary Figure 4. An example of shifted 5'- and 3'-splice sites by Cufflinks-based methods. For reconstructing transcript *IGHV1-2*, 'pooling+ Cufflinks', 'Cufflinks + Cuffmerge', and 'Cufflinks + TACO' built models (plcf_chr14_minus.22.1, cfmg_chr14_minus.17.26, cftc_chr14_minus.23.26) containing false positive splice junctions, where 5'- and 3'-splice sites were shifted by one base pair.

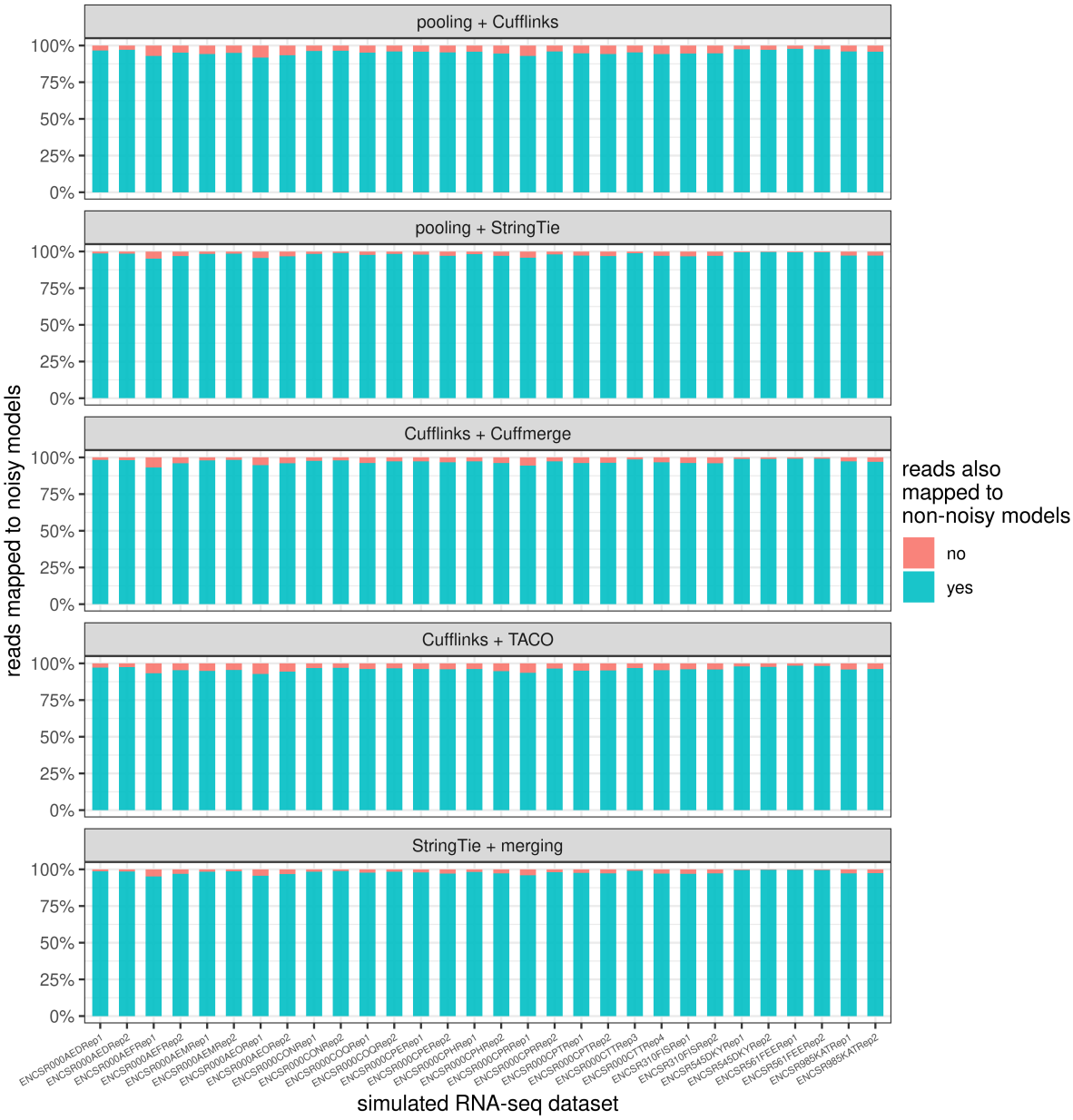


Supplementary Figure 5. Transcript structures missed by '2-Step', but predicted by '1-Step' methods. The definition of 'predicted' and 'missed' transcript structures are the same as Supplementary Table 3.

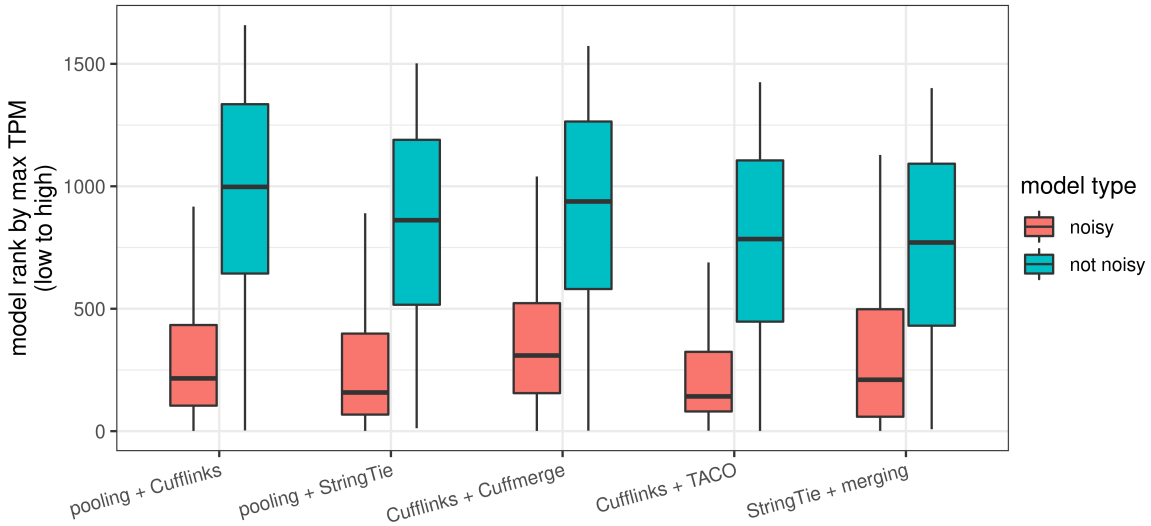


Supplementary Figure 6. Input alignments from the 30 RNA-seq datasets for *GCM1*. ENCFF782TAX was the only RNA-seq dataset that contained a fragment for *GCM1*'s first splice junction. The two mates of this fragment are labelled by red arrows. Track names for transcript models built by Cufflinks and StringTie based on ENCFF782TAX and track name for ENCFF782TAX RNA-seq alignments are highlighted in yellow.

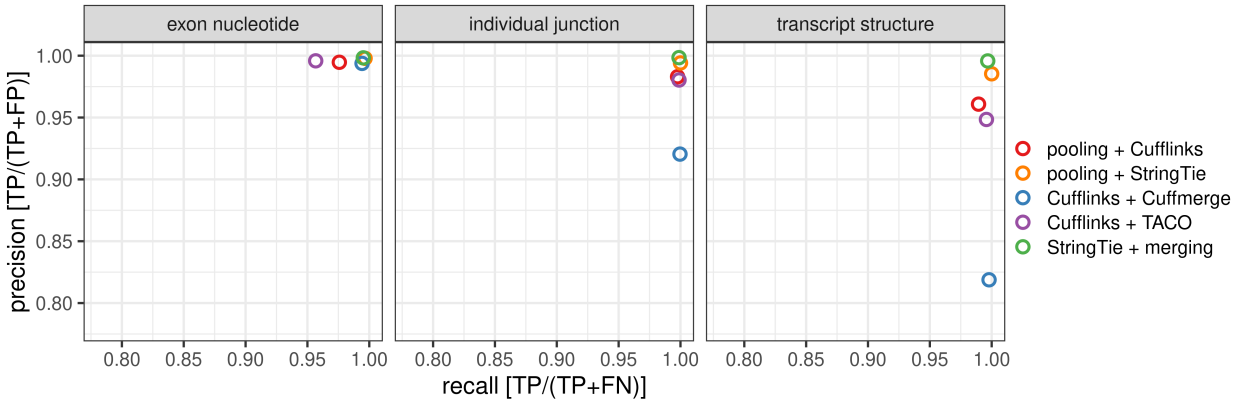
Supplementary Figure 7. UCSC Genome Browser screenshot of a benchmark transcript that had transcript structure predicted by both '1-Step' methods and missed by all three '2-Step' methods on simulated RNA-seq fragments. Shown are GENCODE annotation of transcript *AC073284.4*, predictions from '1-Step' and '2-Step' methods, and simulated RNA-seq fragments that served as the input.



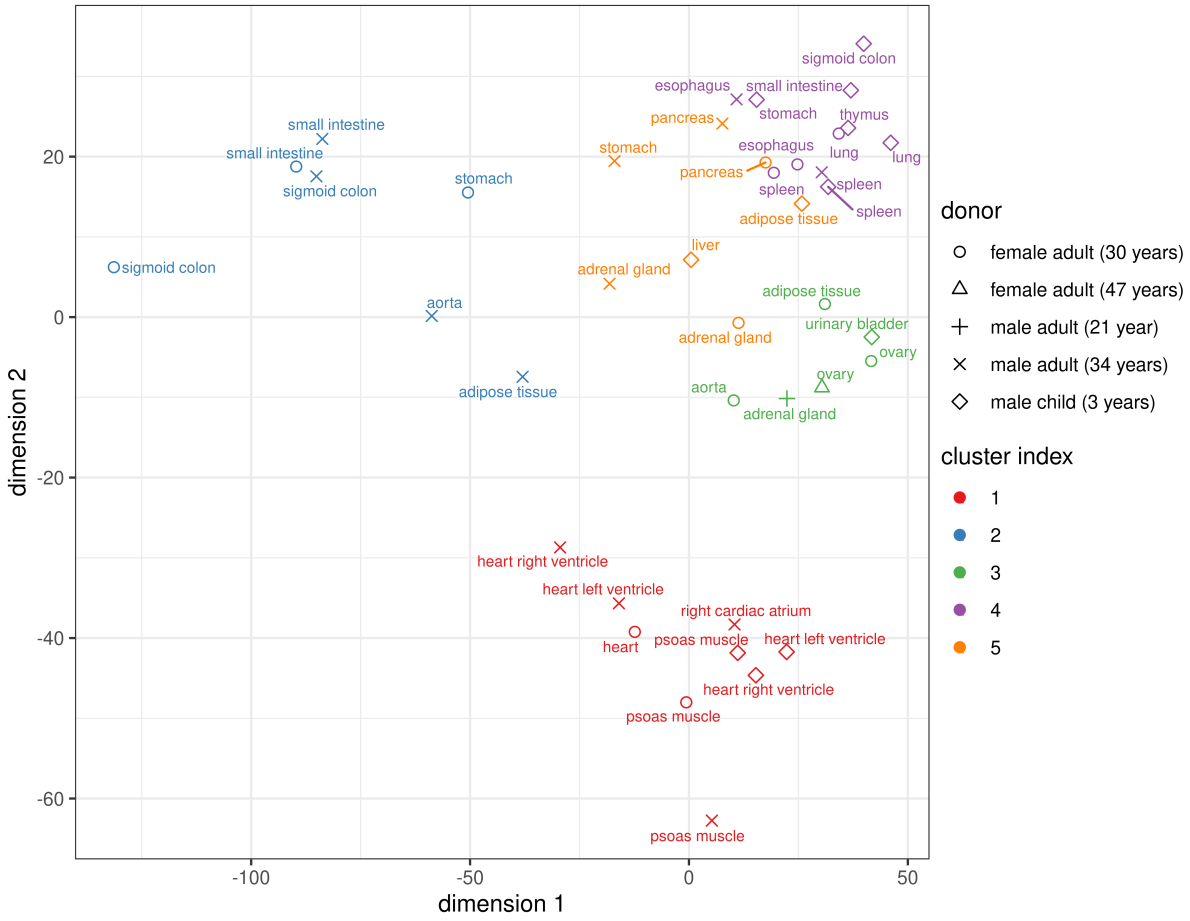
Supplementary Figure 8. Percentage of simulated RNA-seq reads that mapped to noisy models and were shared by 'non-noisy' models. Cyan bars depicts the percentage of reads that mapped to noisy models and were also shared by the 'non-noisy' models.



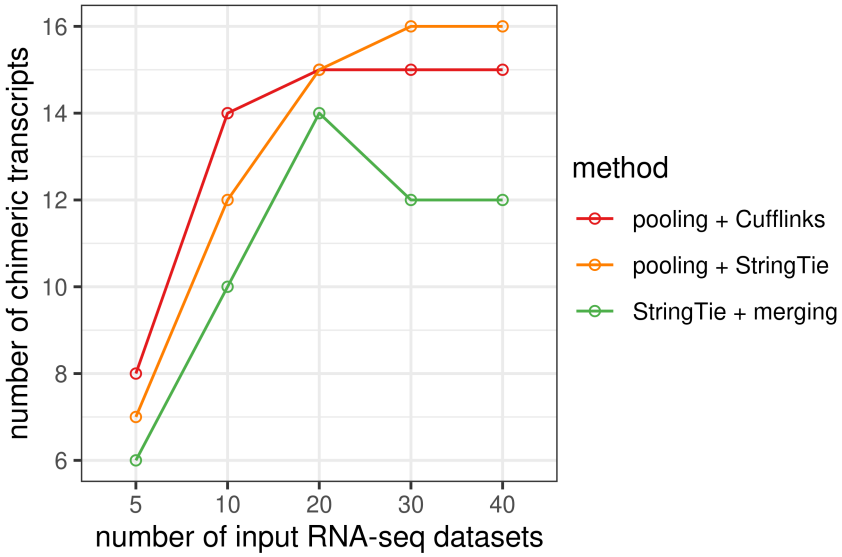
Supplementary Figure 9. Comparison of expression levels for noisy and correctly detected transcript models. Expression level was defined as the maximum TPM from the 30 simulated RNA-seq datasets. Models were ranked by their maximum TPM from low to high. A smaller ranking number indicates a lower expression level.



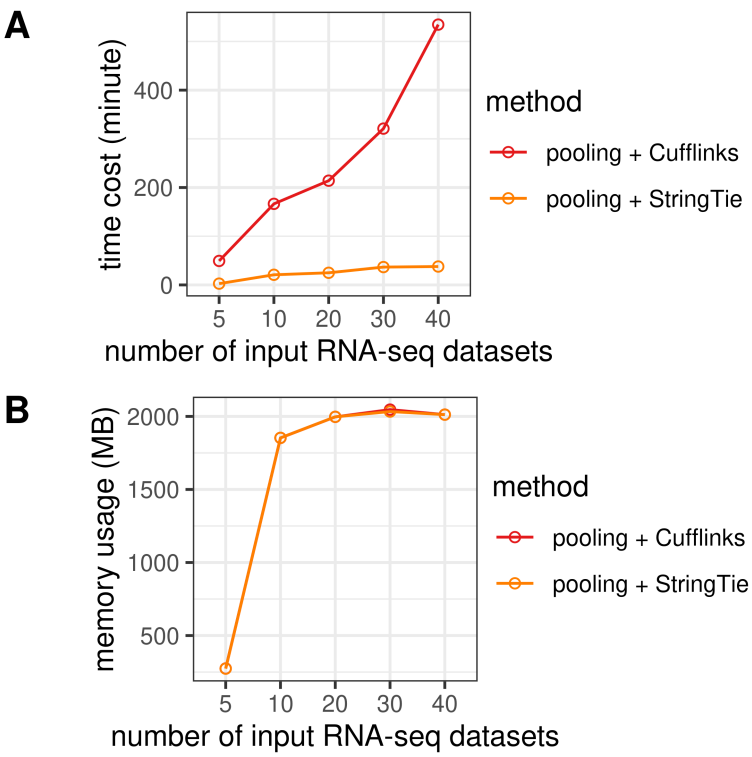
Supplementary Figure 10. Comparison of 1-Step and 2-Step methods on simulated RNA-seq fragments based on parameters learned from the 30 ENCODE datasets. Target transcripts with predicted models from all of the five methods are set as the gold standard. Target transcripts that shared an overlapping predicted model with another target transcript were excluded from evaluation. Predicted models with a single exon, a genomic span < 200bp, or not overlapping any target transcript were excluded from evaluation.



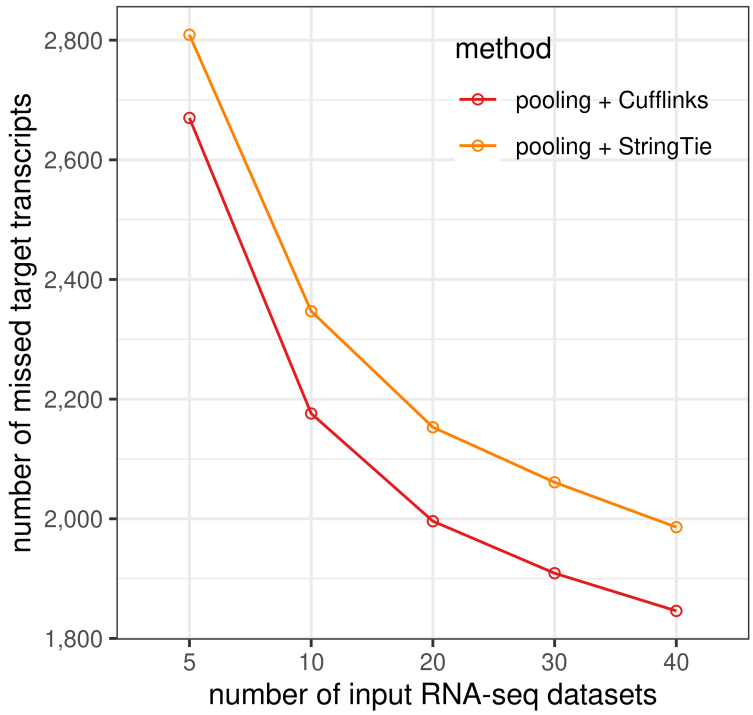
Supplementary Figure 11. 40 ENCODE RNA-seq datasets were grouped into five clusters by K-means clustering. Tissues were clustered based on expression profiles of protein-coding genes and projected on to two dimensions by multidimensional scaling of their $\boldsymbol{log_{10}(FPKM)}$. FPKM values of the genes were added by a constant $\mathbf{10^{-3}}$ to avoid taking logarithm of zero.



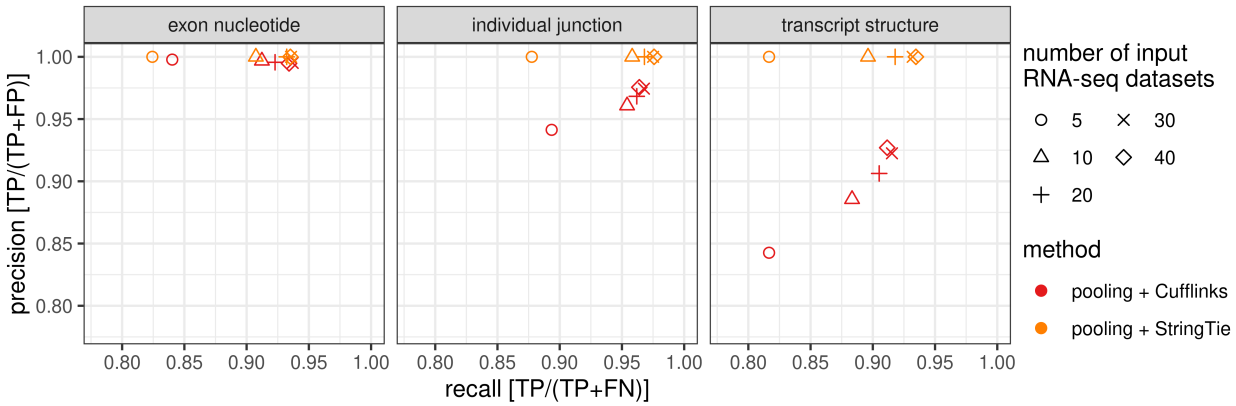
Supplementary Figure 12. Number of predicted chimeric transcripts by '1-Step' and '2-Step' methods using different number of input RNA-seq datasets. The dashed line indicates the number of loci that could give rise to chimeric transcripts.



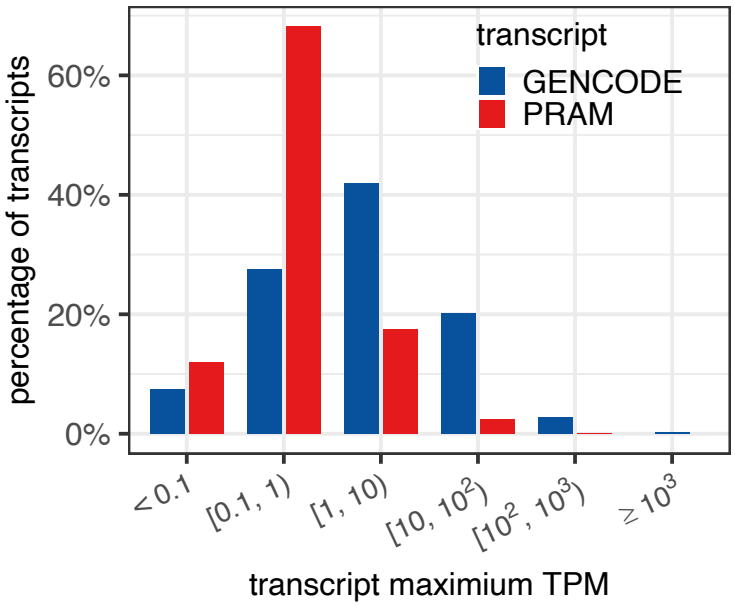
Supplementary Figure 13. Computing time and memory usage for '1-Step' method predictions on different number of input RNA-seq datasets. Predictions were made on one-transcript genes ran on 2.1 GHz AMD CPUs using eight threads.



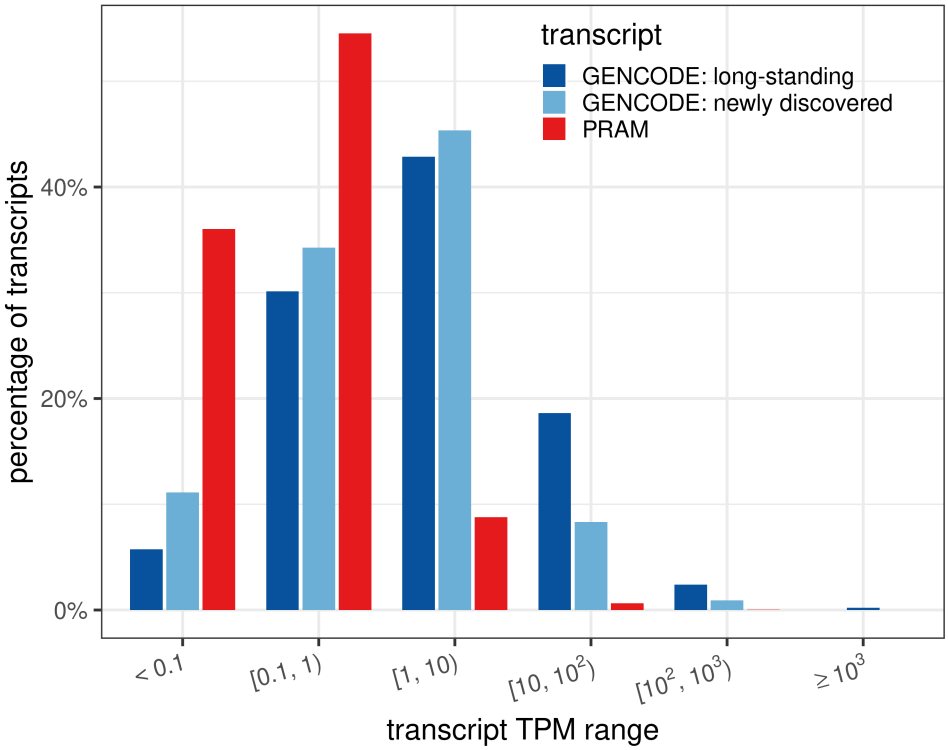
Supplementary Figure 14. Number of target transcripts not detected by '1-Step' methods under different number of input RNA-seq datasets. Whether or not a target transcript was detected by a given methods is based on the whether or not any of the predicted transcript models overlapped the genomic span of the target transcript.



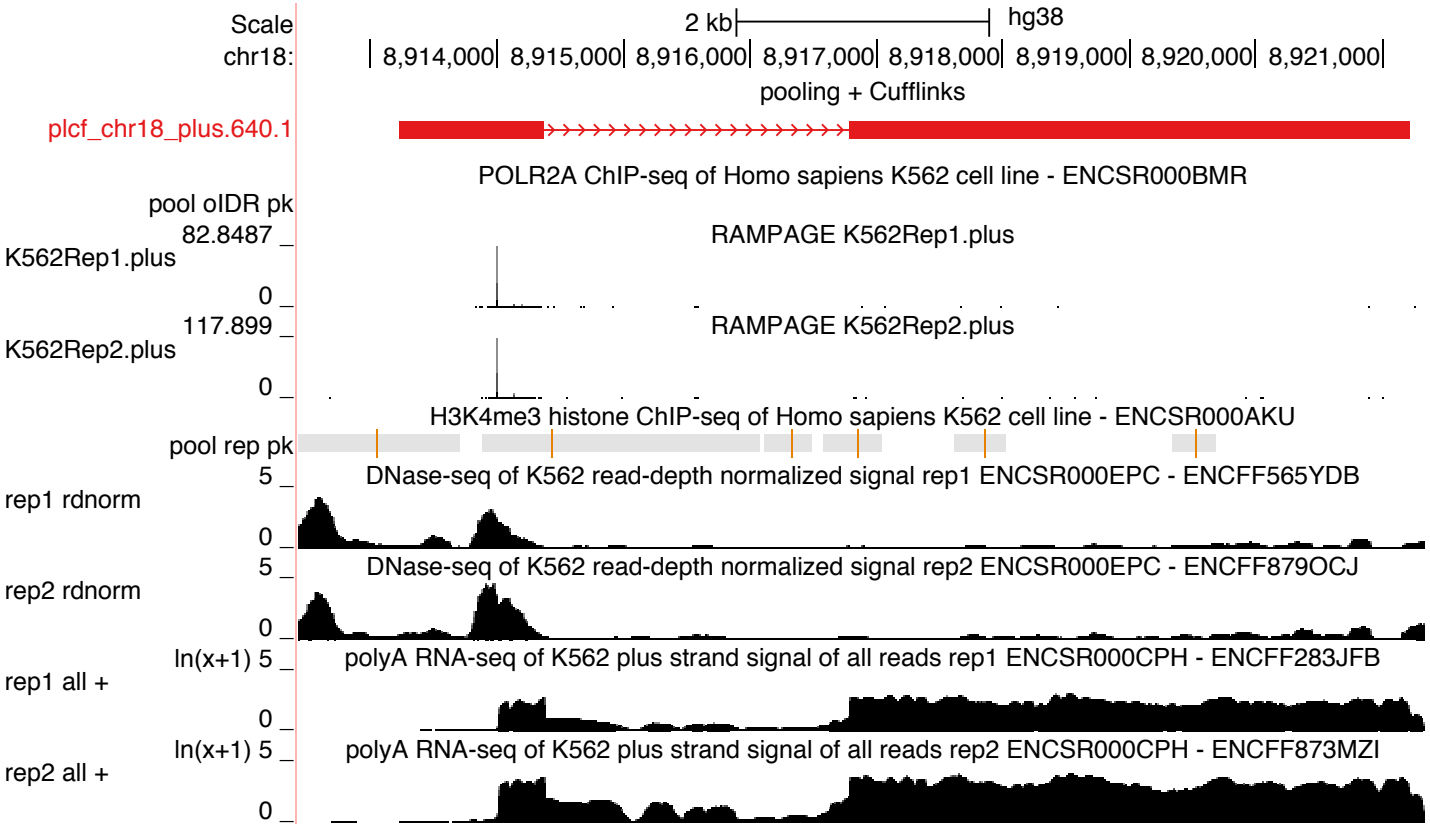
Supplementary Figure 15. Precision and recall on the 780 target transcripts with predicted models under all combinations of the five different numbers of input RNA-seq datasets and two '1-Step' methods.



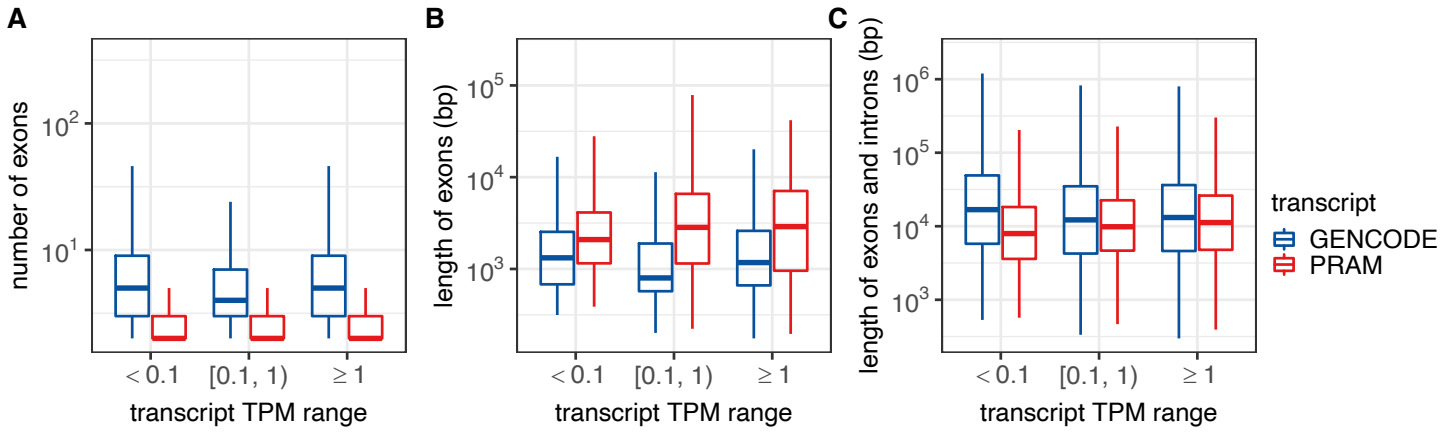
Supplementary Figure 16. Distribution of GENCODE and PRAM transcripts by their maximum TPMs. A transcript's final TPM was defined as its maximum TPM across all the 30 RNA-seq datasets (Supplementary Table 1). Transcripts with single exon, genomic span < 200 bp, or maximum TPM as 0 were excluded.



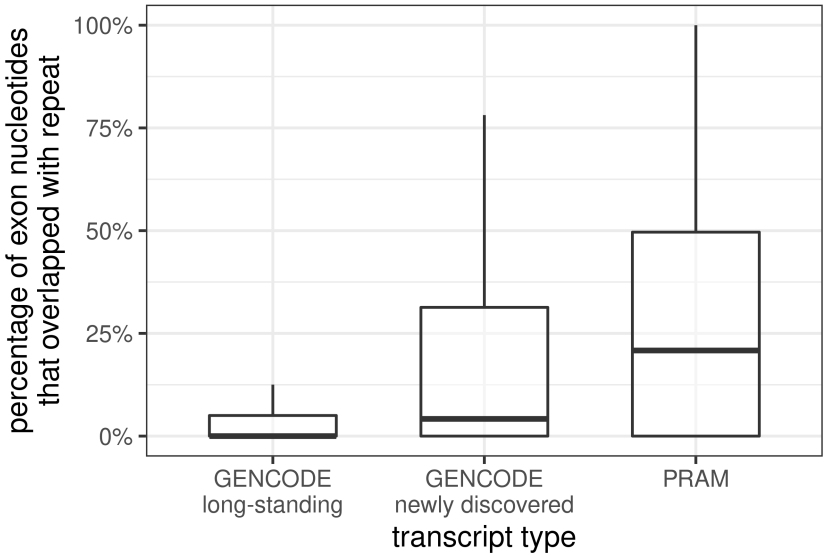
Supplementary Figure 17. Distribution of GENCODE and PRAM transcripts stratified by their average expression levels in the seven cell lines. Transcripts were from the 'kept' category in Supplementary Table 11.



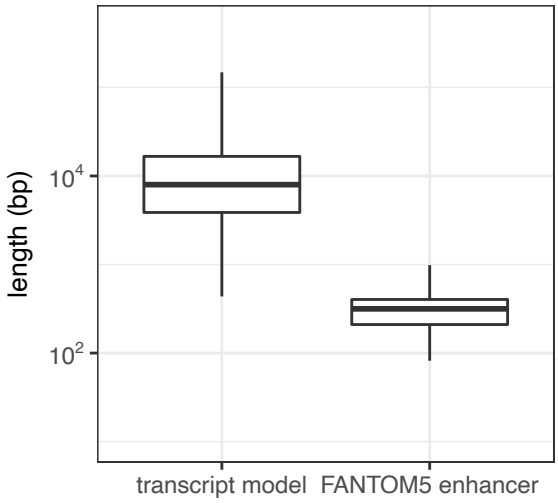
Supplementary Figure 18. The second highly expressed PRAM transcript with supported genomic features. All the genomic datasets were from ENCODE (https://www.encodeproject.org) with their accession IDs listed above each track. Accession IDs for RAMPAGE datasets are listed in Supplementary Table 13. The model 'plcf_chr18_plus.640.1' had an average TPM of 124 in K562 cells. It had high DNase-seq signals around its 5'-exon suggesting high chromatin accessibility and had multiple H3K4me3 ChIP-seq peaks suggesting active transcription. Moreover, 'plcf_chr18_plus.640.1' had strong RAMPAGE signals in close proximity to its transcription start site. All of these external genomic data supported the existence of this highly-expressed PRAM transcript.



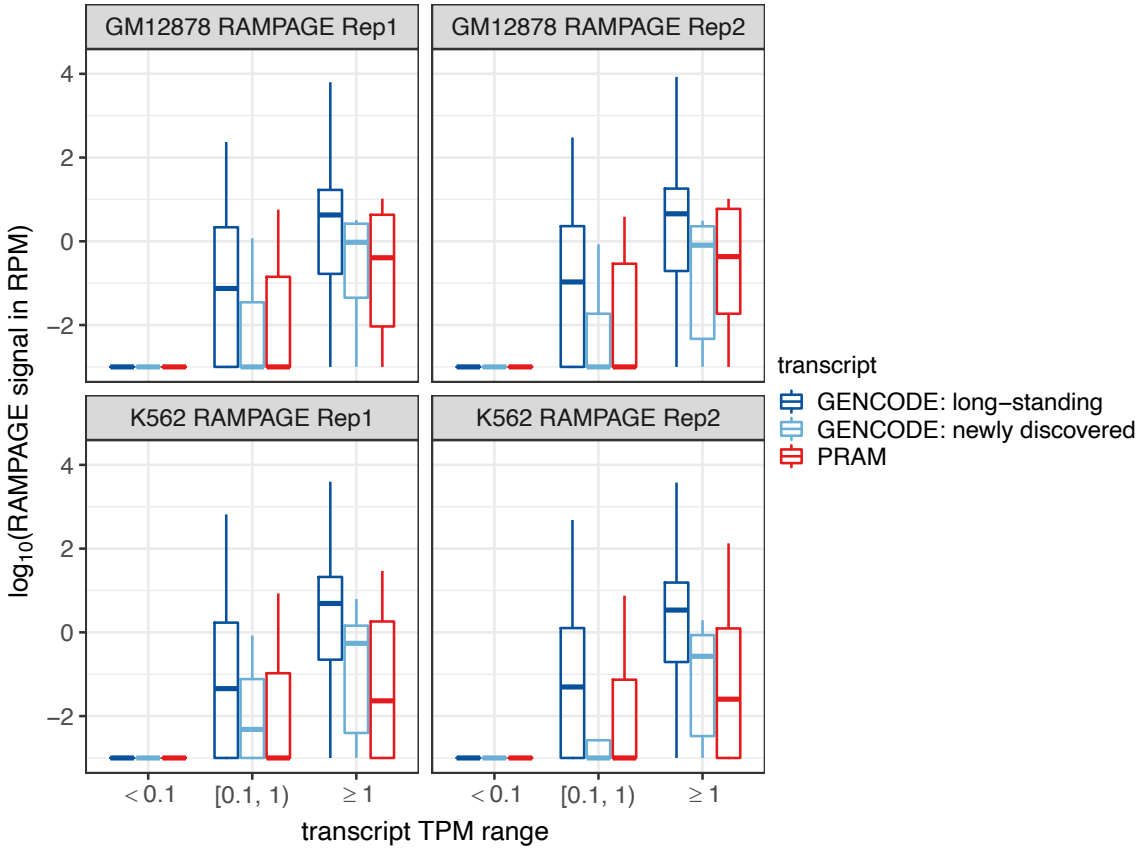
Supplementary Figure 19. Numbers and lengths of GENCODE and PRAM transcript exon and introns. Selection of transcripts were the same as in Figure 2B.



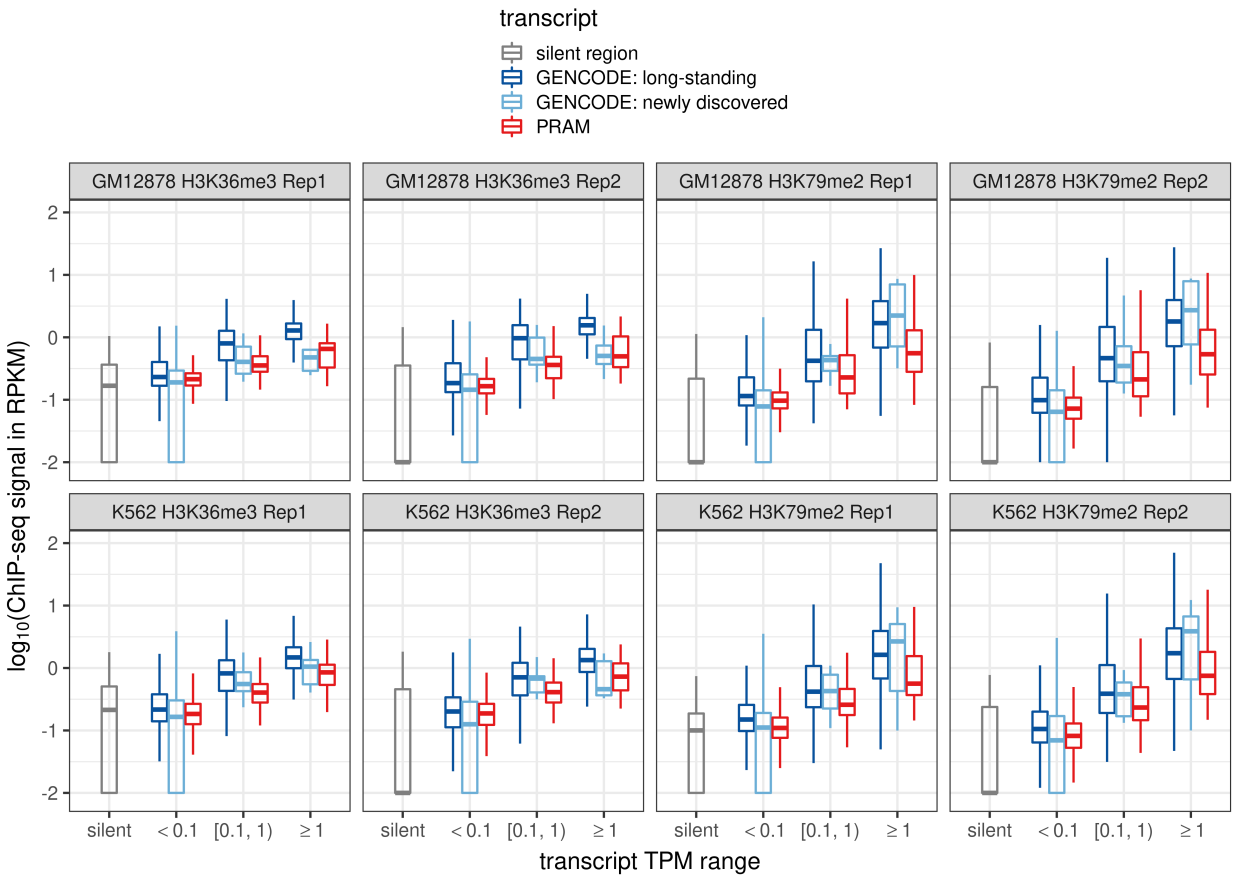
Supplementary Figure 20. Fraction of exon nucleotides that overlapped with repeats. 14,226 PRAM transcripts were compared with 1,034 GENCODE 'newly discovered' transcripts and 197,167 GENCODE 'long-standing' transcripts. Repeats were downloaded from UCSC Genome Browser's RepeatMasker track for hg38 (https://genome.ucsc.edu/cgi-bin/hgTables). We quantified overlap of exon nucleotides of PRAM transcripts to those of repeats from RepeatMasker and observed that about half of PRAM transcripts had less than 25% of their exon nucleotides overlapping with repeats and about three quarters of PRAM transcripts had less than 50% of their exon nucleotides overlapping with repeats. As positive controls, we repeated the same analysis for 'newly discovered' and 'long-standing' GENCODE transcripts. They have lower overlap compared to PRAM transcripts. In particular, 'long-standing' transcripts have median fraction at 0 and the 3rd quartile at about 5%. However, PRAM transcripts did not completely or largely correspond to repeats. 26% (3,686 of 14,226) of the PRAM transcripts and 45% (469 of 1,034) 'newly discovered' GENCODE transcripts did not have their exons overlapping with any repeat at all.



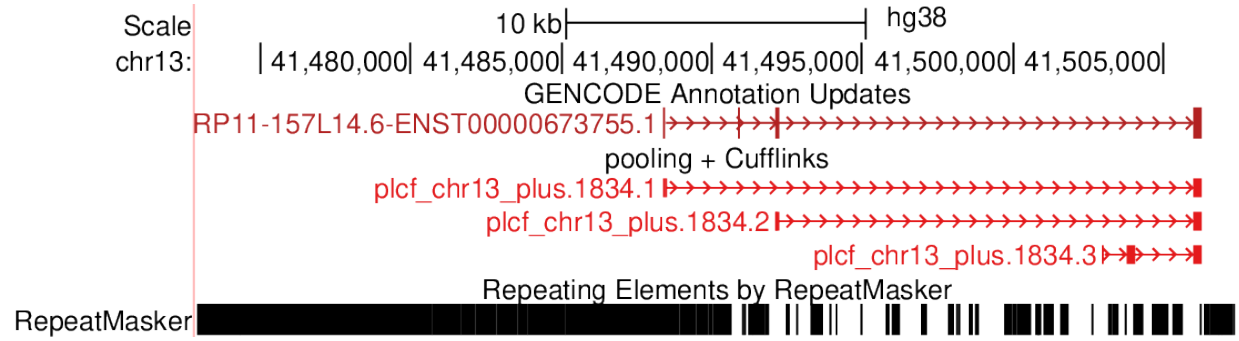
Supplementary Figure 21. Lengths of PRAM transcripts and FANTOM5 enhancers. Fantom5 enhancers were from the 'robust set' downloaded from http://enhancer.binf.ku.dk/presets/robust_enhancers.bed and lifted from human genome hg19 to hg38 for comparison with PRAM transcripts.



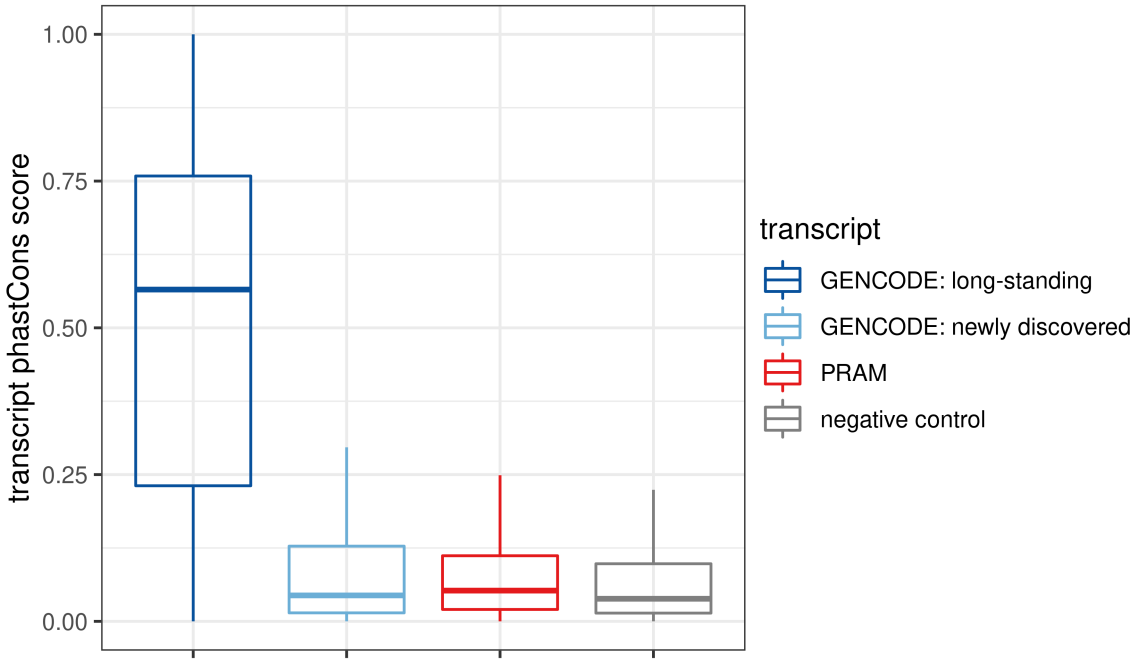
Supplementary Figure 22. RAMPAGE signals of human GENCODE and PRAM transcripts. Box plots are based on transcripts listed as 'promoter mappability ≥ 0.8' in Supplementary Table 12. RAMPAGE signals are from the two GM12878 replicates and the two K562 replicates listed in Supplementary Table 13 and are displayed as panel strip titles. RAMPAGE signals were calculated as read per millions (RPM) with an added factor of 10^{-3} (maximum non-zero RPM is 0.0176) to avoid logarithm of zero.



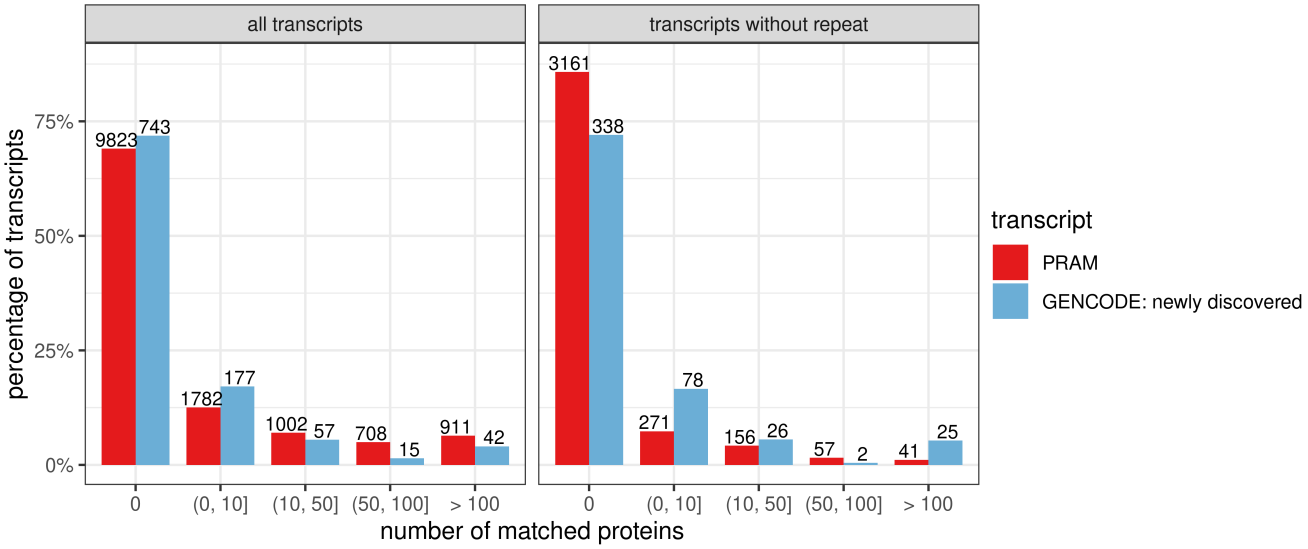
Supplementary Figure 23. Transcription-associated epigenetic signals of human GENCODE and PRAM transcripts as well as silent regions as negative control. Box plots were based on transcripts listed as 'transcript mappability ≥ 0.8' in Supplementary Table 12. ChIP-seq signals were from the datasets listed in Supplementary Table 14 and are displayed as panel strip titles. ChIP-seq signals were calculated as read per kilobase millions (RPKM) with an added factor of $\mathbf{10^{-2}}$ to avoid taking logarithm of zero. We used the transcriptional repressive histone mark H3K27me3 (ChIP-seq peaks called by ENCODE: ENCFF512TQI for HeLa-S3; ENCFF337XQQ for HepG2; ENCFF140SFK for MCF-7; and ENCFF277NRX for SK-N-SH) to define a set of silent regions as negative controls. Since we used histone marks from GM12878 and K562 cell lines to validate transcripts, we avoided internal correlation of histone marks within the same cell line by utilizing H3K27me3 peaks from the other five cell lines that were used to predict PRAM transcripts. ENCODE does not have H3K27me3 ChIP-seq data in un-treated A549 cells (there are two datasets on A549 with treatments); therefore, we only had H3K27me3 peaks from four cell lines. We defined 'silent regions' as regions: (i) overlapping with H3K27me3 peaks across all four cell lines; (ii) with a minimal width of 200 bp; (iii) on Chromosomes 1-22 or X, where PRAM transcripts were derived. There were 163 such silent regions in total. We quantified their signals for marks H3K36me3 and H3K79me2 that associate with transcription, and compared them with those of GENCODE and PRAM transcripts. Overall, silent regions had lower transcription associated histone mark signals than GENCODE and PRAM transcripts in all three TPM ranges (except for GENCODE newly discovered transcripts in the lowest TPM category). This suggests that GENCODE and PRAM predictions have higher evidence of transcription as measured by these two histone modifications even at lower TPM settings.



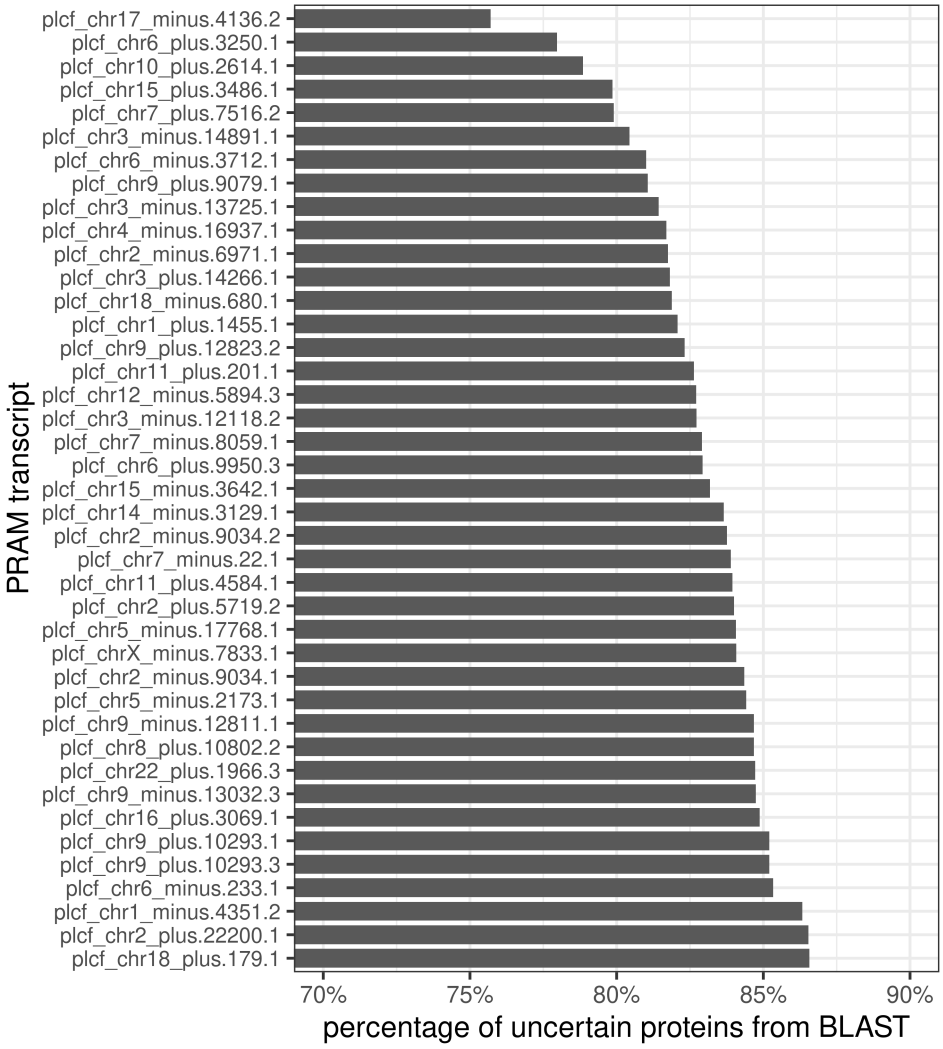
Supplementary Figure 24. UCSC Genome Browser screenshot of a recently updated GENCODE transcript overlapped with PRAM transcripts. We downloaded GENCODE's regularly updated data after the most recent release from http://ftp.ebi.ac.uk/pub/databases/gencode/update_trackhub/data/hg38.bed.gz on Nov 26, 2019. Of all the 2,003 transcripts, 23 resided within the intergenic regions on Chromosome 1-22 and X, which we used to predict PRAM transcripts. 48% (11 out of 23) transcripts overlapped with PRAM transcripts. While not perfect, the high percentage overlap illustrated PRAM's predicted accuracy.



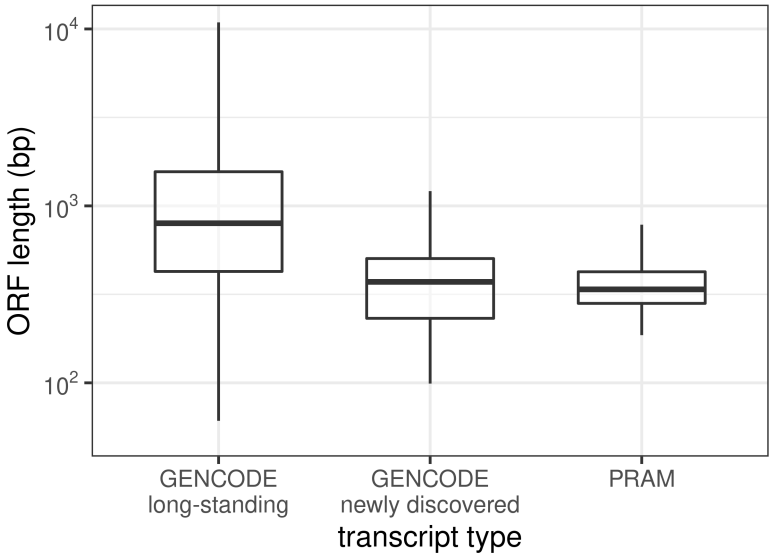
Supplementary Figure 25. The phastCons scores of GENCODE and PRAM human transcripts. GENCODE transcripts were divided into 'long-standing' and 'newly discovered'. Randomly selected 1,000 non-transcript genomic regions with a width of 1kb were used as negative controls. The phastCons scores estimates the probability that each genomic nucleotide belongs to a conserved element. It takes on values between 0 and 1, where higher value indicates higher probability of conservation. We downloaded phastCons scores (hg38, based multiple alignments of 100 vertebrate species) from UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/phastCons100way/hg38.phastCons100way.bw). We calculated the phastCons scores for all the 197,167 long-standing and 1,034 newly discovered GENCODE transcripts, and 14,226 PRAM transcripts. A transcript's phastCons score was defined as the average phastCons scores of all of its exons. We also generated a negative control, by randomly sampling 1,000 genomic regions with width of 1kb from the genomic locations that did not overlap with any GENCODE or PRAM transcripts. Next, we compared the distribution of phastCons scores for these four categories. The phastCons scores of newly discovered and PRAM transcripts were significantly differently from long-standing transcripts (Wilcoxon rank-sum test $p < 10^{-16}$ for both newly discovered and PRAM transcripts) and the control regions (Wilcoxon rank-sum test $p=5.4 \times 10^{-4}$ for newly discovered transcript; $p=1.2 \times 10^{-9}$ for PRAM transcripts), whereas the difference between newly discovered and PRAM transcripts were not significant (Wilcoxon rank-sum test p=0.25) under p-value cutoff of 0.01. This comparison suggested that newly discovered and PRAM transcripts shared similar degree of conservation across vertebrate. Their conservation is significantly higher than expected by chance and significantly lower than those of long-standing transcripts.



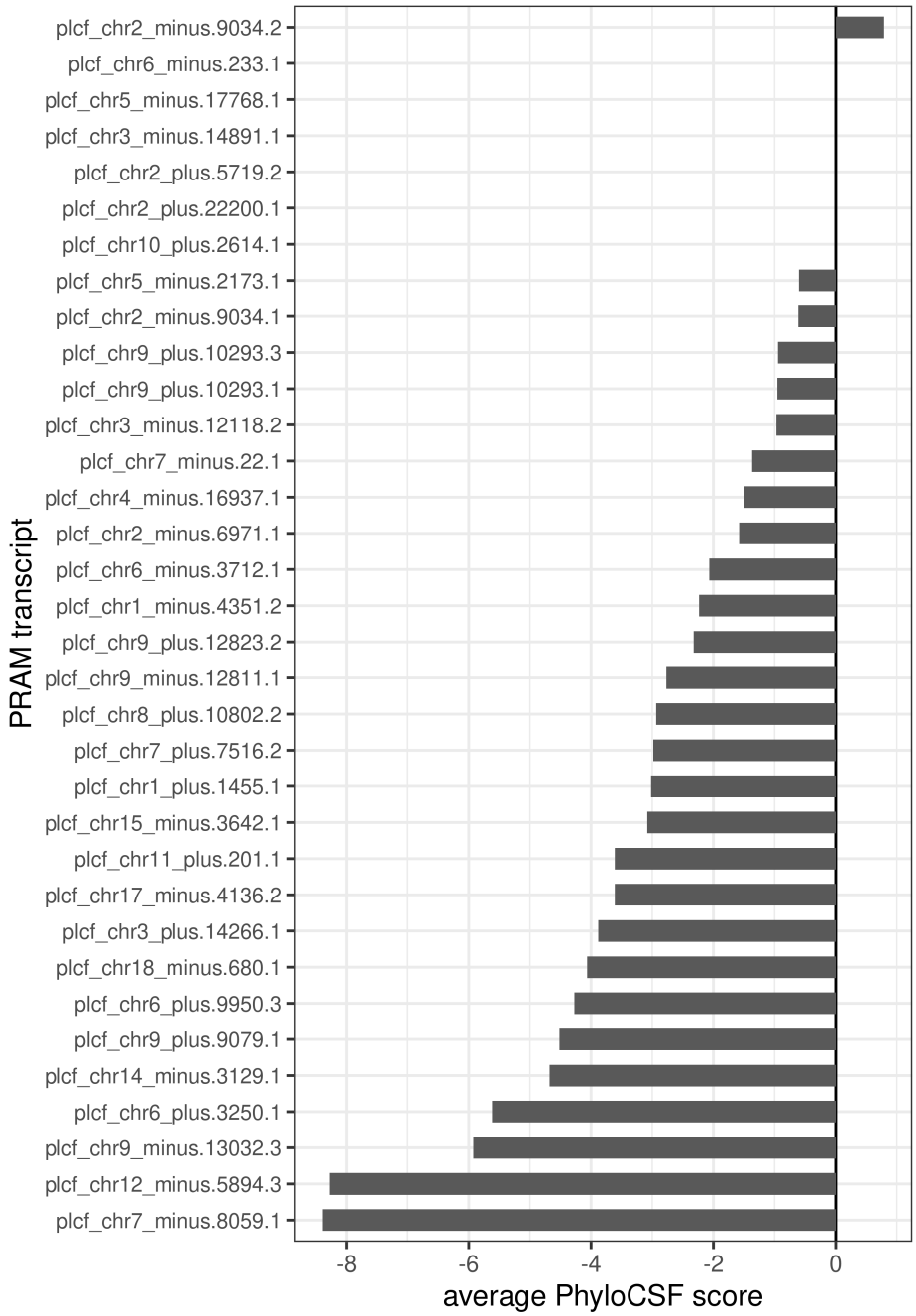
Supplementary Figure 26. Percentage of PRAM and GENCODE 'newly discovered' transcripts stratified by the number of BLAST-matched proteins. All the transcripts (left panel) and transcripts without any exon overlap with RepeatMasker repeats (right panel) were studied. The number of transcripts within each category are reported above the corresponding bar. We carried out the same BLAST analysis on the 1,034 GENCODE newly discovered transcripts. The percentages of matches to proteins stratified by number of proteins had a similar distribution to those of PRAM transcripts, which again suggested shared features between PRAM and GENCODE newly discovered transcripts. To rule out transcripts that mapped to retrotransposon sequences, we selected the 3,686 PRAM and 469 GENCODE newly discovered transcripts that did not have any exon overlap with RepeatMasker repeats. Their percentages of matched proteins showed a similar distribution to those of all the transcripts, where > 70% transcripts did not match to any proteins at all. We also noticed that there were 41 PRAM transcripts with matches to > 100 proteins.



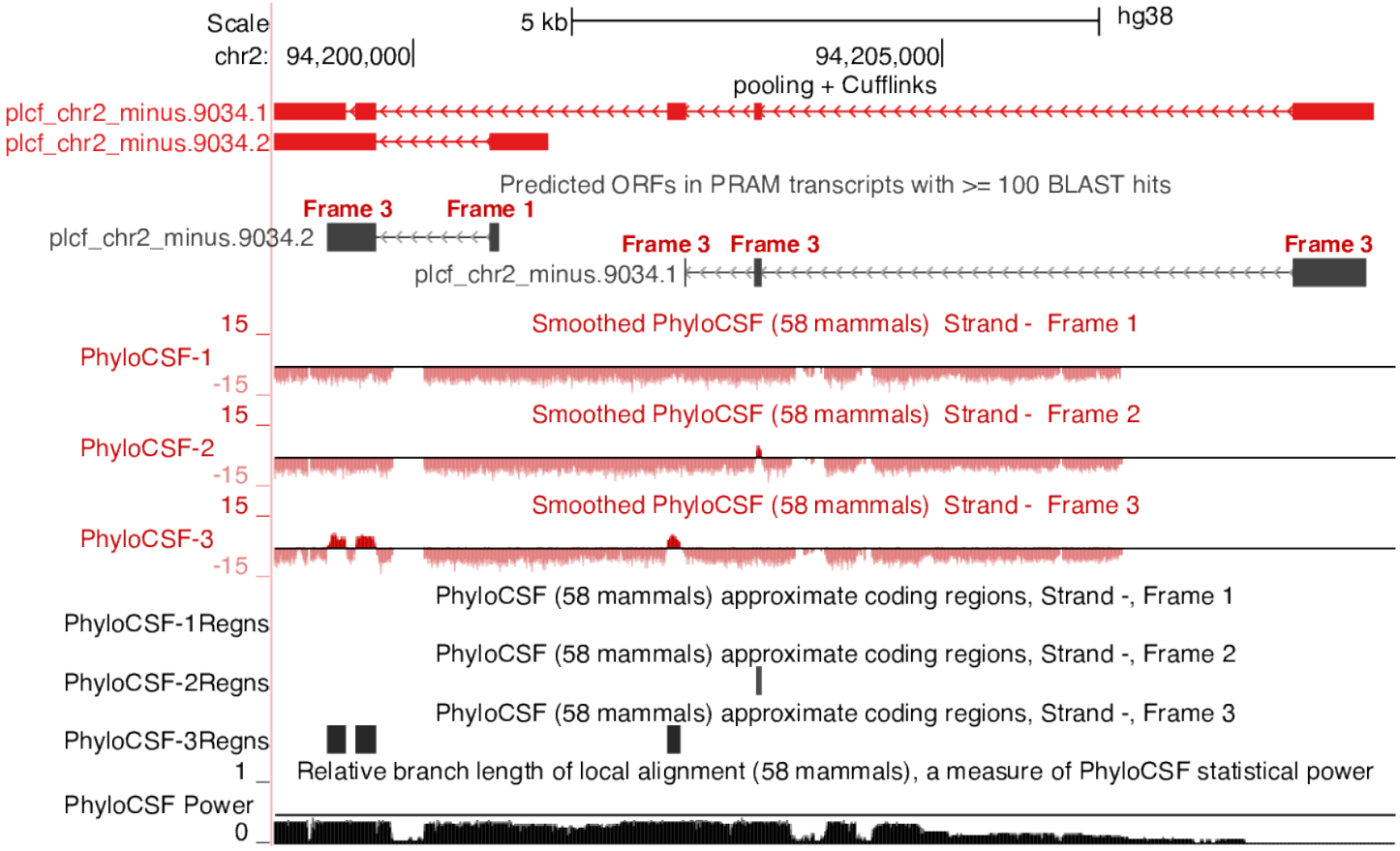
Supplementary Figure 27. Percentage of uncertain BLAST-matched proteins for the 41 PRAM transcripts. A protein is considered as uncertain if its name contains the following word: hypothetical, predicted, putative, uncharacterized, unknown, or unnamed. The 41 PRAM transcripts are those without any repeat and matched to more than 100 proteins by BLAST (Supplementary Figure 26).



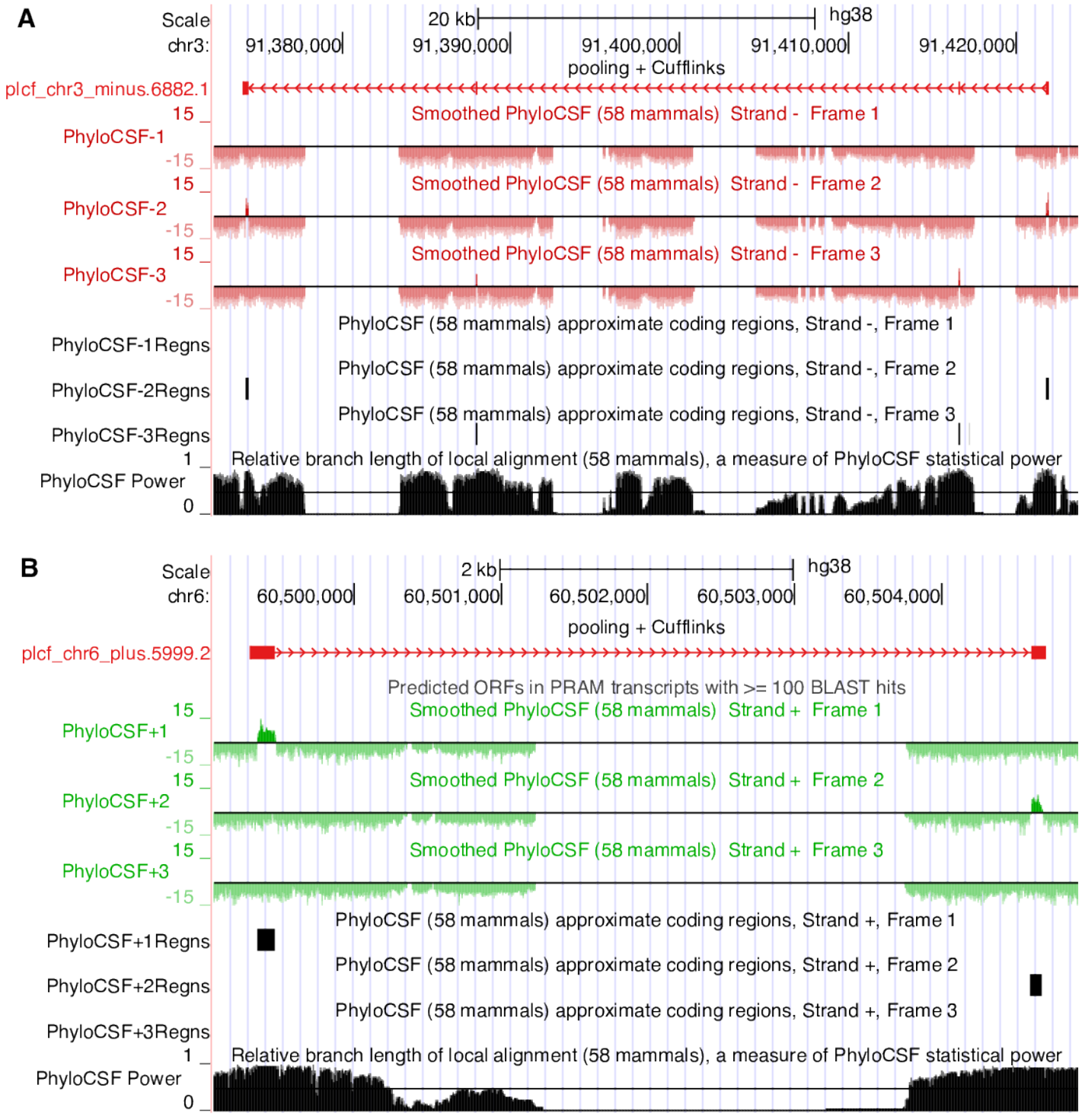
Supplementary Figure 28. Comparison of ORF lengths between GENCODE and the 41 PRAM transcripts. GENCODE transcript ORFs were defined by GENCODE version 24 and were required to have feature column as 'CDS', gene type as 'protein_coding', and transcript type as 'protein_coding'. There are 79,654 such transcripts with ORFs from Chromosome 1 to 22 and X. 62 of them are 'newly discovered' GENCODE transcripts and the others are 'long-stranding' transcripts. ORFs of PRAM transcripts were predicted by ORFfinder (version 0.4.3, https://www.ncbi.nlm.nih.gov/orffinder/). ORF for each PRAM transcript was defined as the longest ORF among all predicted ones. The 41 PRAM transcripts are those without any repeat and matched to more than 100 proteins by BLAST (Supplementary Figure 26). 34 of the 41 transcripts have at least one ORF predicted.



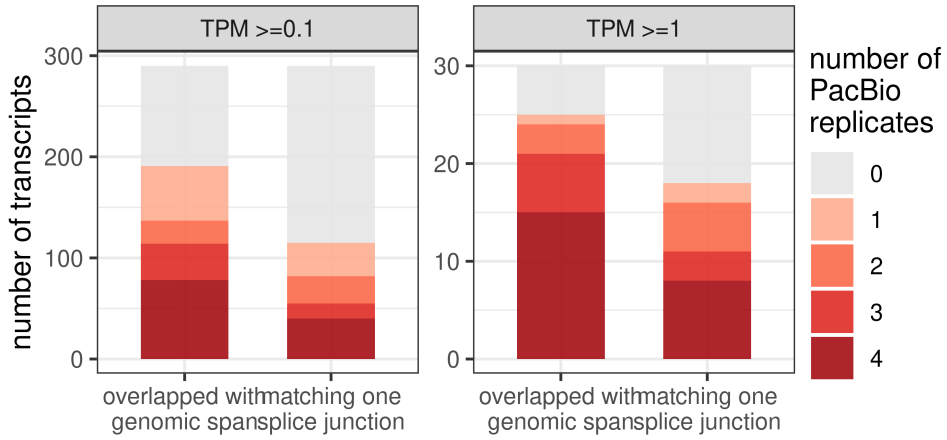
Supplementary Figure 29. PhyloCSF scores of PRAM transcripts' predicted ORFs. ORF's score was calculated using strand- and frame-matched smoothed PhyloCSF scores derived from 58-mammal alignments (https://data.broadinstitute.org/compbio1/PhyloCSFtracks/). A positive score indicates protein-coding potential. Shown are the 34 of 41 PRAM transcripts with at least one ORF predicted (Supplementary Figure 28).



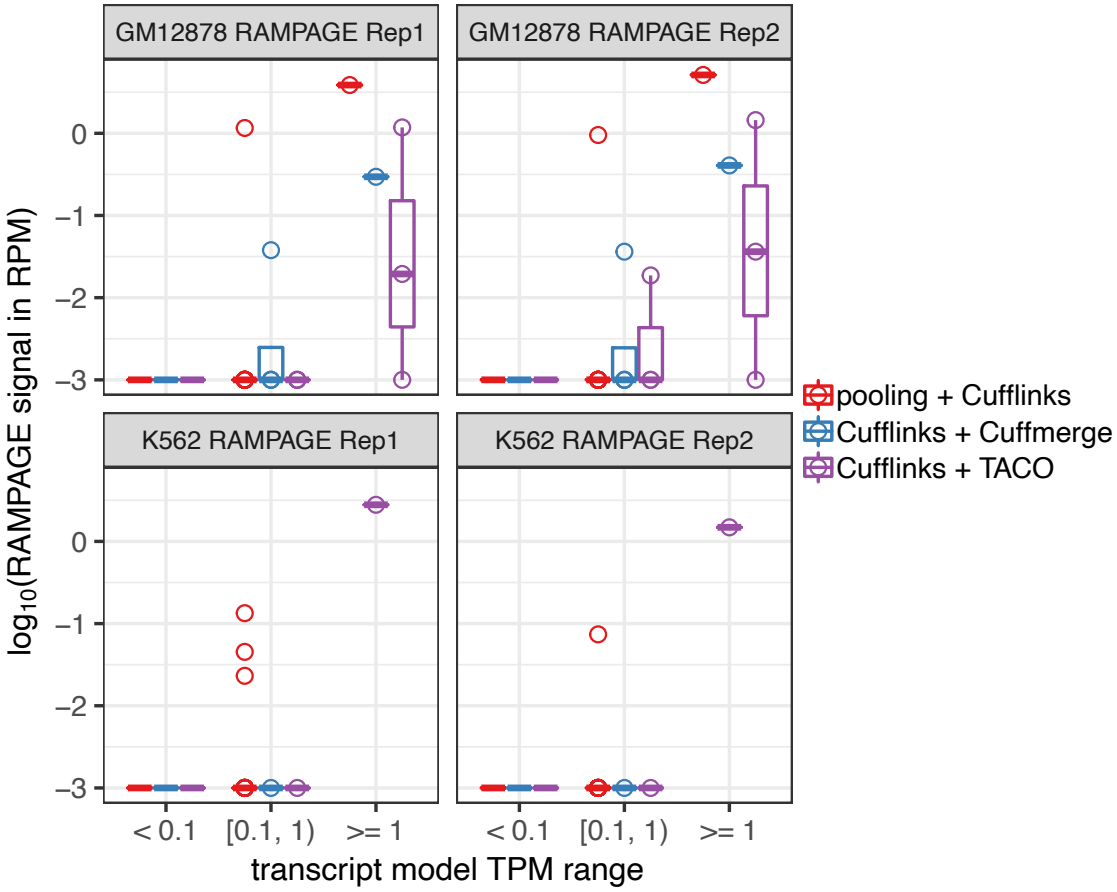
Supplementary Figure 30. UCSC Genome Browser screenshot of two of the 41 PRAM transcripts, their predicted ORFs and PhyloCSF scores. ORF's matched PhyloCSF frame was labeled in red bold text above each ORF.



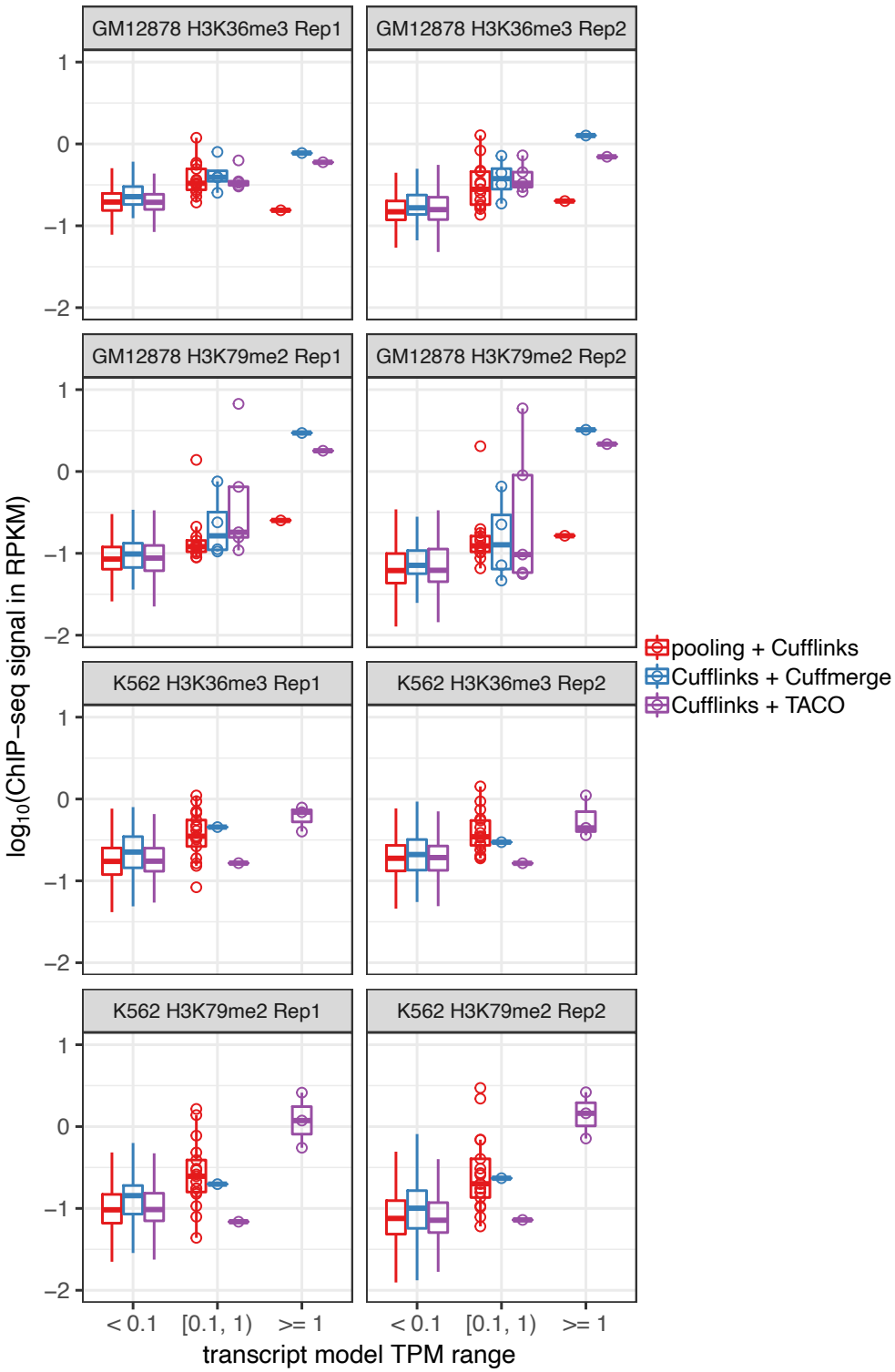
Supplementary Figure 31. UCSC Genome Browser screenshot of the two PRAM transcripts that have >70% of their exons overlapped with PhyloCSF-predicted coding regions. Overlap was defined as on the same strand regardless of frame.



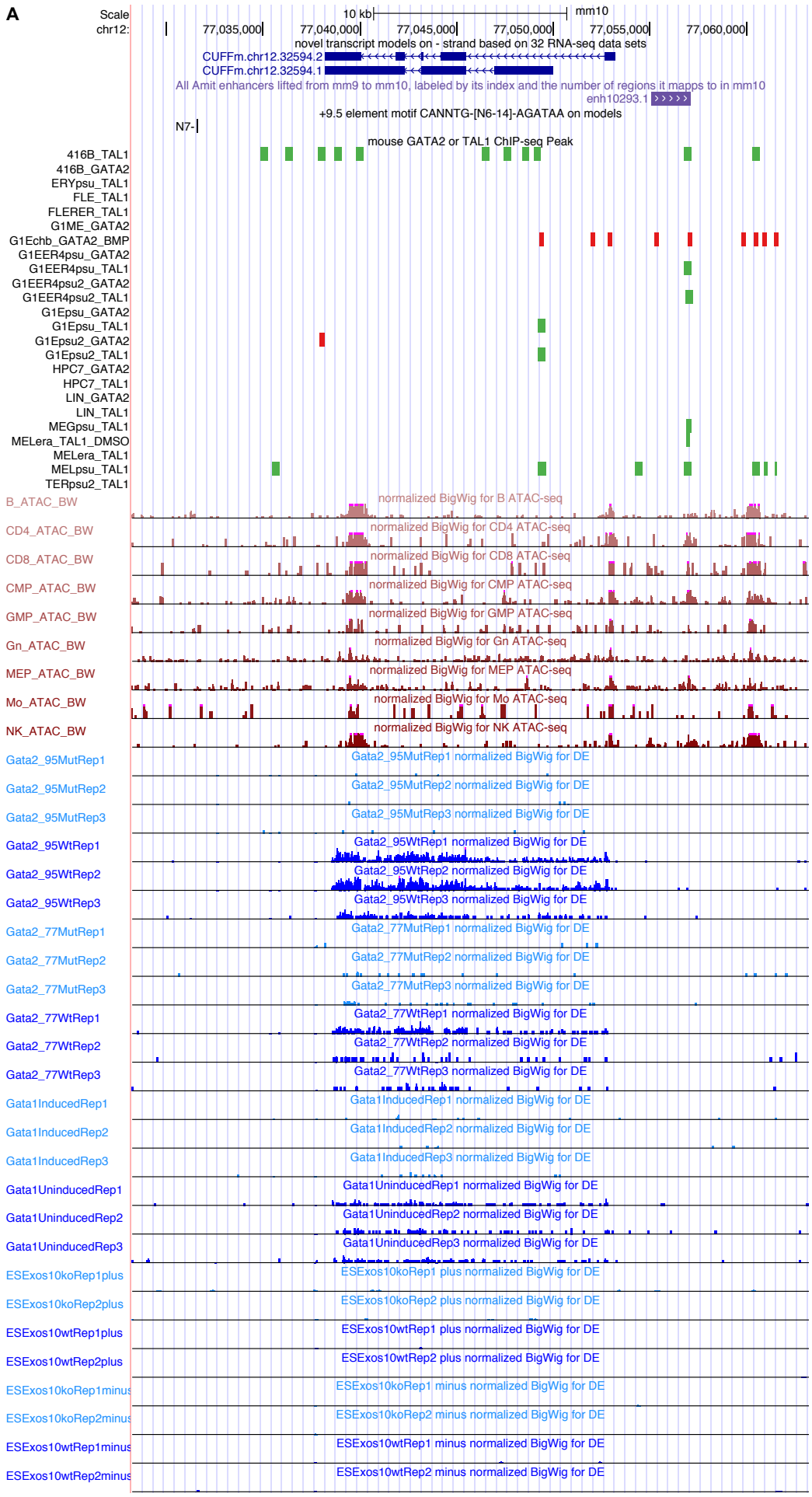
Supplementary Figure 32. Comparison of PRAM transcripts with ENCODE GM12878 PacBio long reads. ENCODE has PacBio data on GM12878 (https://www.encodeproject.org/experiments/ENCSR706ANY/) in the format of FASTQ files from four replicates (as of Nov. 14, 2019). We followed ENCODE's long read RNA-seq analysis protocol (https://www.encodeproject.org/documents/7ec9d66a-3b7e-4183-8677-e1df14770b44/@@download/attachment/ENCODE%20Long%20Read%20RNA-Seq%20Analysis%20Pipeline%20%28Human%29.pdf) to align reads to human genome hg38. The comparison of PacBio reads and PRAM transcripts were evaluated by two features: (i) overlap with genomic span: whether a PRAM transcript had a PacBio read overlapping with its genomic span (exons and introns); (ii) match to one splice junction: whether at least one of the splice junctions of a PRAM transcript matched exactly to a PacBio read splice junction. We made comparisons on two sets of PRAM transcripts: (i) $\boldsymbol{TPM \geq 0.1}$: transcripts with $\boldsymbol{TPM \geq 0.1}$ in all of the six GM12878 ENCODE short-read RNA-seq datasets. This resulted in 290 transcripts. (ii) $\boldsymbol{TPM \geq 1}$: transcripts with $\boldsymbol{TPM \geq 1}$ in all of the six GM12878 ENCODE short-read RNA-seq datasets. This resulted in 30 transcripts that were among the 290 transcripts identified above. In the '$\boldsymbol{TPM \geq 0.1}$' set (left panel), 66% (191 out of 290) transcripts had a PacBio read overlapping with their genomic span from at least one PacBio replicate. 27% (78 out of 290) transcripts had a PacBio read overlapping with their genomic span from all four PacBio replicates. The fraction of splice junction match was relatively smaller. 40% (115 out of 290) of PRAM transcripts had at least one of its splice junctions matched by a long read from at least one PacBio replicate. In the '$\boldsymbol{TPM \geq 1}$' set (right panel), 83% (25 out of 30) of PRAM transcripts had long-read overlap from at least one PacBio replicate and half (15 out of 30) of PRAM transcripts had long-read overlap from all four PacBio replicates. For splice junction match, the corresponding percentages were 60% (18 out of 30) and 27% (8 out of 30) from at least one PacBio replicate and all four replicates, respectively. In conclusion, a substantial fraction of PRAM transcripts overlapped with PacBio long reads and had matching splice junctions as well, providing further support for PRAM transcripts.

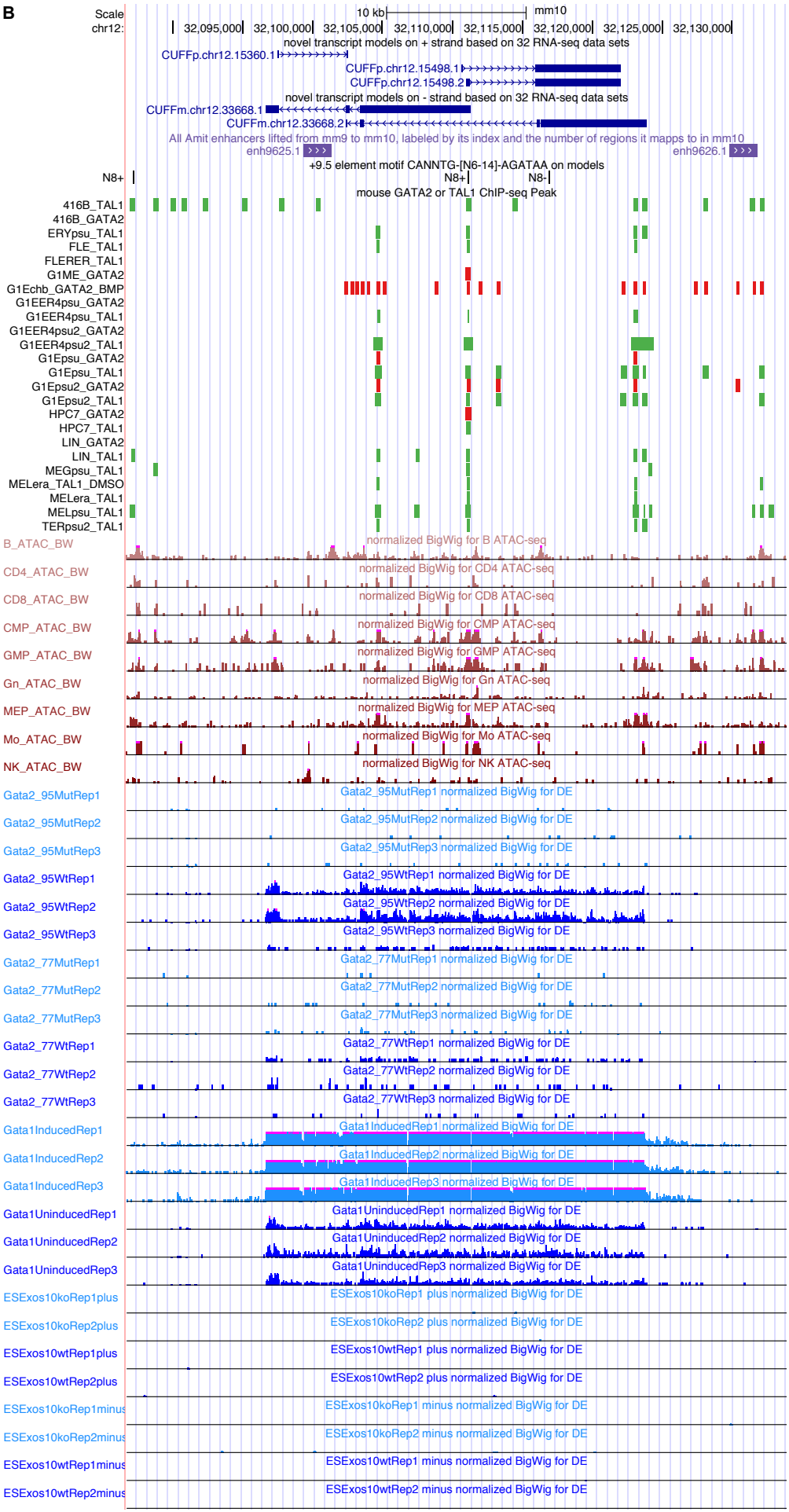


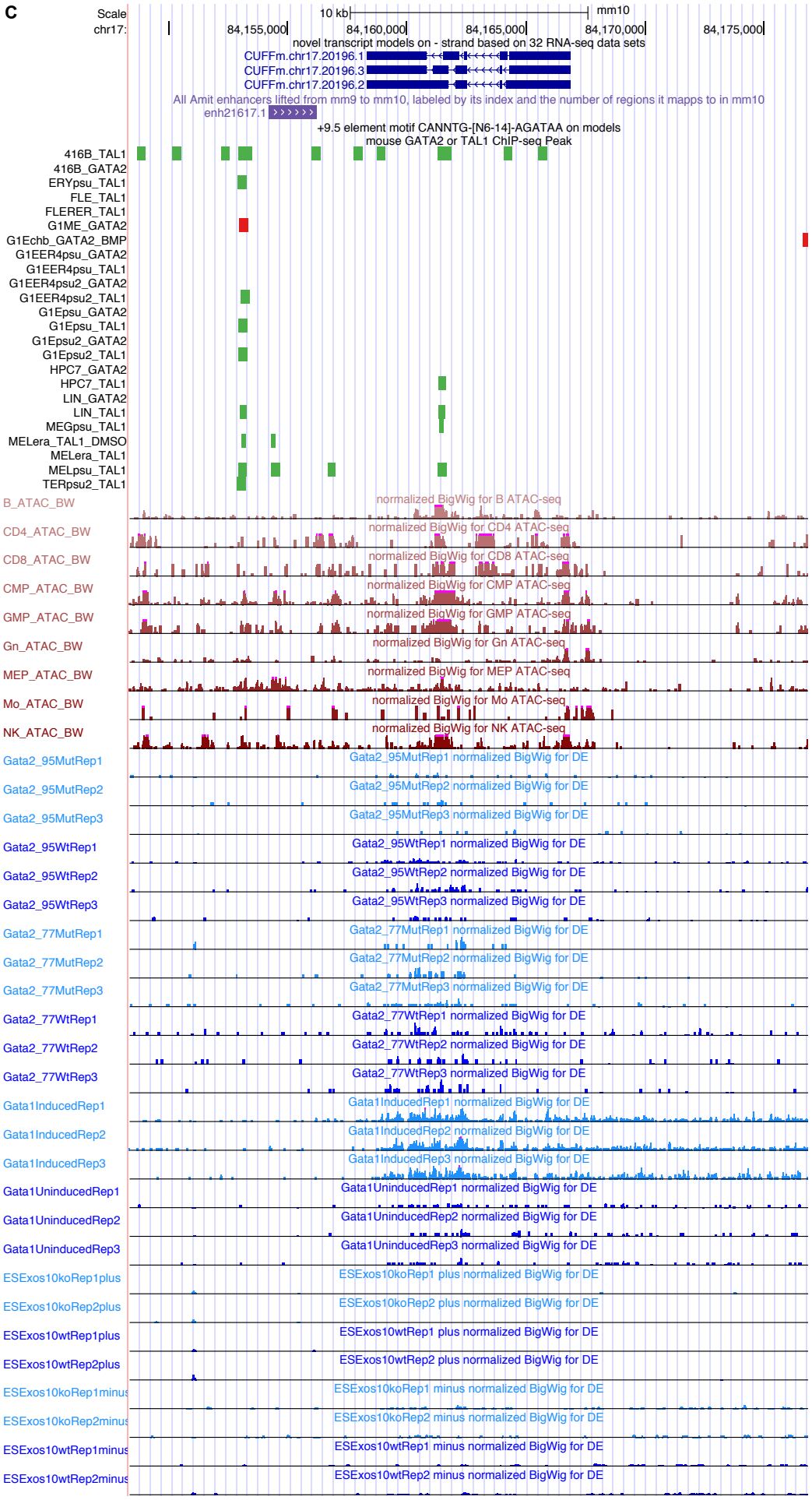
Supplementary Figure 33. RAMPAGE signals of '1-Step' and '2-Step' specific human transcripts. Box plots are based on models listed as 'promoter mappability ≥ 0.8' in Supplementary Table 20. Models with TPM range of [0.1, 1] and '>= 1' are also displayed as points. 'pooling + Cufflinks' and 'Cufflinks + Cuffmerge' did not have any model with TPM ≥ 1 in K562. RAMPAGE signals were based on the two GM12878 replicates and the two K562 replicates listed in Supplementary Table 13 and displayed as panel strip titles. RAMPAGE signals were calculated as read per millions (RPM) with an added factor of 10^{-3} (maximum non-zero RPM is 0.0176) to avoid logarithm of zero.

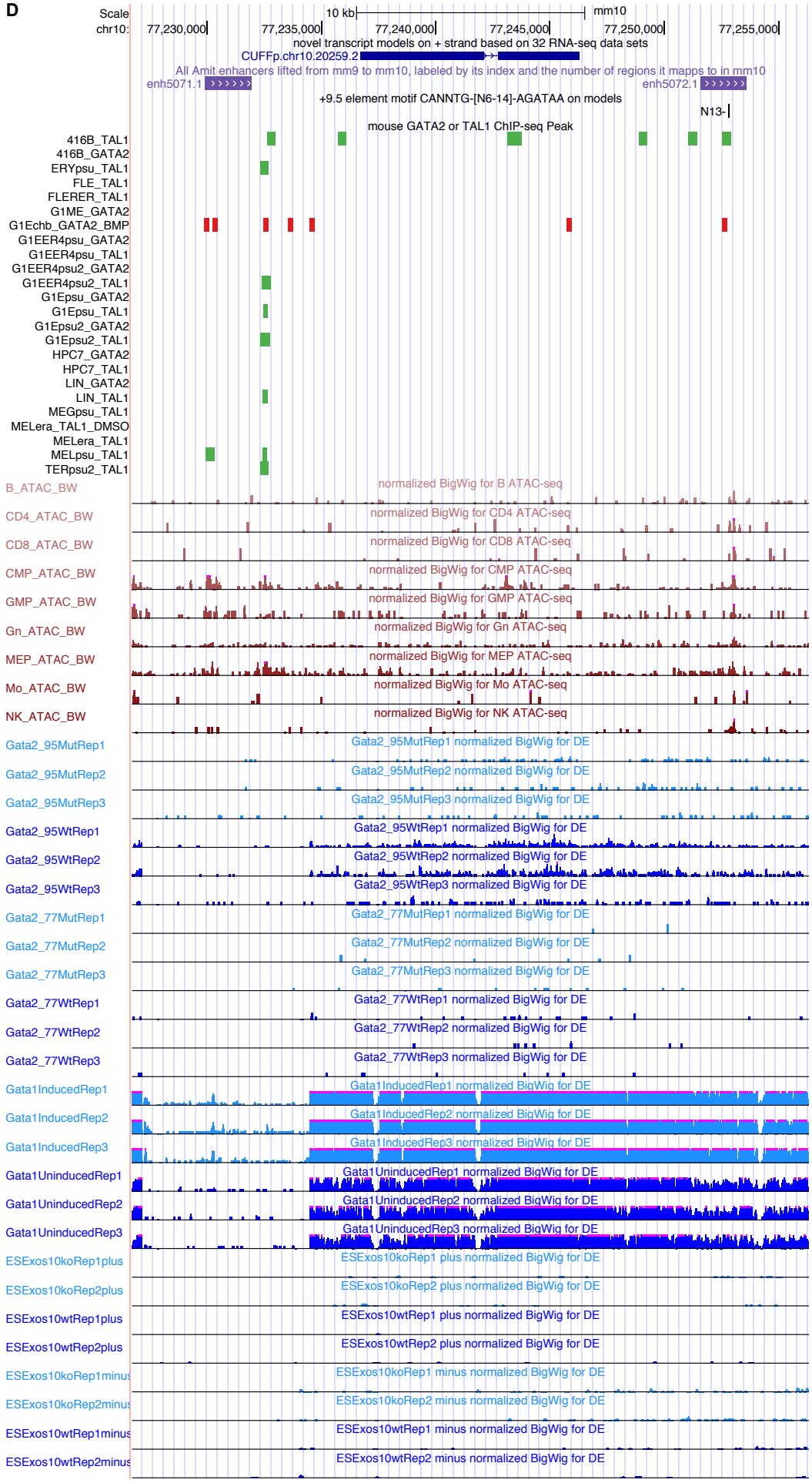


Supplementary Figure 34. Epigenetic signals of '1-Step' and '2-Step' specific human transcripts. Box plots are based on models listed as 'transcript mappability ≥ 0.8' in Supplementary Table 20. Models with TPM range of '[0.1, 1)' and '>= 1' are also displayed as points. 'pooling + Cufflinks' and 'Cufflinks + Cuffmerge' did not have any model with TPM ≥ 1 in K562. ChIP-seq signals are from the datasets listed in Supplementary Table 14 and displayed as panel strip titles. ChIP-seq signals were calculated as read per kilobase millions (RPKM) with an added factor of 10^{-5} to avoid logarithm of zero.

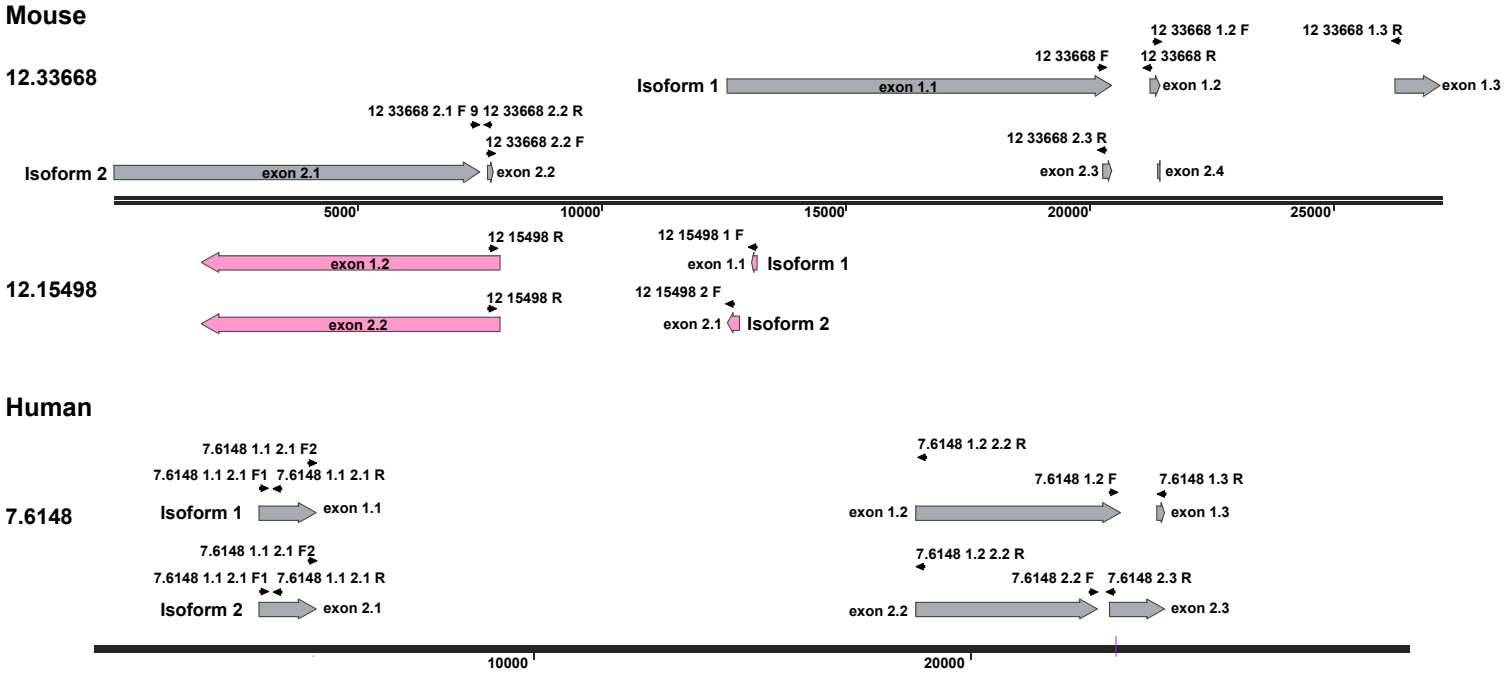

A
Scale
chr12:
10 kb
mm10
77,035,000
77,040,000
77,045,000
77,050,000
77,055,000
77,060,000
novel transcript models on - strand based on 32 RNA-seq data sets
CUFFm.chr12.32594.2
CUFFm.chr12.32594.1
All Amit enhancers lifted from mm9 to mm10, labeled by its index and the number of regions it mapps to in mm10
enh10293.1
+9.5 element motif CANNTG-[N6-14]-AGATAA on models
N7-
mouse GATA2 or TAL1 ChIP-seq Peak
416B_TAL1
416B_GATA2
ERYpsu_TAL1
FLE_TAL1
FLERER_TAL1
G1ME_GATA2
G1Echb_GATA2_BMP
G1EER4psu_GATA2
G1EER4psu_TAL1
G1EER4psu2_GATA2
G1EER4psu2_TAL1
G1Epsu_GATA2
G1Epsu_TAL1
G1Epsu2_GATA2
G1Epsu2_TAL1
HPC7_GATA2
HPC7_TAL1
LIN_GATA2
LIN_TAL1
MEGpsu_TAL1
MELera_TAL1_DMSO
MELera_TAL1
MELpsu_TAL1
TERpsu2_TAL1
B_ATAC_BW
normalized BigWig for B ATAC-seq
CD4_ATAC_BW
normalized BigWig for CD4 ATAC-seq
CD8_ATAC_BW
normalized BigWig for CD8 ATAC-seq
CMP_ATAC_BW
normalized BigWig for CMP ATAC-seq
GMP_ATAC_BW
normalized BigWig for GMP ATAC-seq
Gn_ATAC_BW
normalized BigWig for Gn ATAC-seq
MEP_ATAC_BW
normalized BigWig for MEP ATAC-seq
Mo_ATAC_BW
normalized BigWig for Mo ATAC-seq
NK_ATAC_BW
normalized BigWig for NK ATAC-seq
Gata2_95MutRep1
Gata2_95MutRep1 normalized BigWig for DE
Gata2_95MutRep2
Gata2_95MutRep2 normalized BigWig for DE
Gata2_95MutRep3
Gata2_95MutRep3 normalized BigWig for DE
Gata2_95WtRep1
Gata2_95WtRep1 normalized BigWig for DE
Gata2_95WtRep2
Gata2_95WtRep2 normalized BigWig for DE
Gata2_95WtRep3
Gata2_95WtRep3 normalized BigWig for DE
Gata2_77MutRep1
Gata2_77MutRep1 normalized BigWig for DE
Gata2_77MutRep2
Gata2_77MutRep2 normalized BigWig for DE
Gata2_77MutRep3
Gata2_77MutRep3 normalized BigWig for DE
Gata2_77WtRep1
Gata2_77WtRep1 normalized BigWig for DE
Gata2_77WtRep2
Gata2_77WtRep2 normalized BigWig for DE
Gata2_77WtRep3
Gata2_77WtRep3 normalized BigWig for DE
Gata1InducedRep1
Gata1InducedRep1 normalized BigWig for DE
Gata1InducedRep2
Gata1InducedRep2 normalized BigWig for DE
Gata1InducedRep3
Gata1InducedRep3 normalized BigWig for DE
Gata1UninducedRep1
Gata1UninducedRep1 normalized BigWig for DE
Gata1UninducedRep2
Gata1UninducedRep2 normalized BigWig for DE
Gata1UninducedRep3
Gata1UninducedRep3 normalized BigWig for DE
ESExos10koRep1plus
ESExos10koRep1 plus normalized BigWig for DE
ESExos10koRep2plus
ESExos10koRep2 plus normalized BigWig for DE
ESExos10wtRep1plus
ESExos10wtRep1 plus normalized BigWig for DE
ESExos10wtRep2plus
ESExos10wtRep2 plus normalized BigWig for DE
ESExos10koRep1minus
ESExos10koRep1 minus normalized BigWig for DE
ESExos10koRep2minus
ESExos10koRep2 minus normalized BigWig for DE
ESExos10wtRep1minus
ESExos10wtRep1 minus normalized BigWig for DE
ESExos10wtRep2minus
ESExos10wtRep2 minus normalized BigWig for DE


B
Scale
chr12:
10 kb
mm10
32,095,000
32,100,000
32,105,000
32,110,000
32,115,000
32,120,000
32,125,000
32,130,000
novel transcript models on + strand based on 32 RNA-seq data sets
CUFFp.chr12.15360.1
CUFFp.chr12.15498.1
CUFFp.chr12.15498.2
novel transcript models on - strand based on 32 RNA-seq data sets
CUFFm.chr12.33668.1
CUFFm.chr12.33668.2
All Amit enhancers lifted from mm9 to mm10, labeled by its index and the number of regions it mapps to in mm10
enh9625.1
enh9626.1
+9.5 element motif CANNTG-[N6-14]-AGATAA on models
N8+
N8+
N8-
mouse GATA2 or TAL1 ChIP-seq Peak
416B_TAL1
416B_GATA2
ERYpsu_TAL1
FLE_TAL1
FLERER_TAL1
G1ME_GATA2
G1Echb_GATA2_BMP
G1EER4psu_GATA2
G1EER4psu_TAL1
G1EER4psu2_GATA2
G1EER4psu2_TAL1
G1Epsu_GATA2
G1Epsu_TAL1
G1Epsu2_GATA2
G1Epsu2_TAL1
HPC7_GATA2
HPC7_TAL1
LIN_GATA2
LIN_TAL1
MEGpsu_TAL1
MELera_TAL1_DMSO
MELera_TAL1
MELpsu_TAL1
TERpsu2_TAL1
B_ATAC_BW
normalized BigWig for B ATAC-seq
CD4_ATAC_BW
normalized BigWig for CD4 ATAC-seq
CD8_ATAC_BW
normalized BigWig for CD8 ATAC-seq
CMP_ATAC_BW
normalized BigWig for CMP ATAC-seq
GMP_ATAC_BW
normalized BigWig for GMP ATAC-seq
Gn_ATAC_BW
normalized BigWig for Gn ATAC-seq
MEP_ATAC_BW
normalized BigWig for MEP ATAC-seq
Mo_ATAC_BW
normalized BigWig for Mo ATAC-seq
NK_ATAC_BW
normalized BigWig for NK ATAC-seq
Gata2_95MutRep1
Gata2_95MutRep1 normalized BigWig for DE
Gata2_95MutRep2
Gata2_95MutRep2 normalized BigWig for DE
Gata2_95MutRep3
Gata2_95MutRep3 normalized BigWig for DE
Gata2_95WtRep1
Gata2_95WtRep1 normalized BigWig for DE
Gata2_95WtRep2
Gata2_95WtRep2 normalized BigWig for DE
Gata2_95WtRep3
Gata2_95WtRep3 normalized BigWig for DE
Gata2_77MutRep1
Gata2_77MutRep1 normalized BigWig for DE
Gata2_77MutRep2
Gata2_77MutRep2 normalized BigWig for DE
Gata2_77MutRep3
Gata2_77MutRep3 normalized BigWig for DE
Gata2_77WtRep1
Gata2_77WtRep1 normalized BigWig for DE
Gata2_77WtRep2
Gata2_77WtRep2 normalized BigWig for DE
Gata2_77WtRep3
Gata2_77WtRep3 normalized BigWig for DE
Gata1InducedRep1
Gata1InducedRep1 normalized BigWig for DE
Gata1InducedRep2
Gata1InducedRep2 normalized BigWig for DE
Gata1InducedRep3
Gata1InducedRep3 normalized BigWig for DE
Gata1UninducedRep1
Gata1UninducedRep1 normalized BigWig for DE
Gata1UninducedRep2
Gata1UninducedRep2 normalized BigWig for DE
Gata1UninducedRep3
Gata1UninducedRep3 normalized BigWig for DE
ESExos10koRep1plus
ESExos10koRep1 plus normalized BigWig for DE
ESExos10koRep2plus
ESExos10koRep2 plus normalized BigWig for DE
ESExos10wtRep1plus
ESExos10wtRep1 plus normalized BigWig for DE
ESExos10wtRep2plus
ESExos10wtRep2 plus normalized BigWig for DE
ESExos10koRep1minus
ESExos10koRep1 minus normalized BigWig for DE
ESExos10koRep2minus
ESExos10koRep2 minus normalized BigWig for DE
ESExos10wtRep1minus
ESExos10wtRep1 minus normalized BigWig for DE
ESExos10wtRep2minus
ESExos10wtRep2 minus normalized BigWig for DE

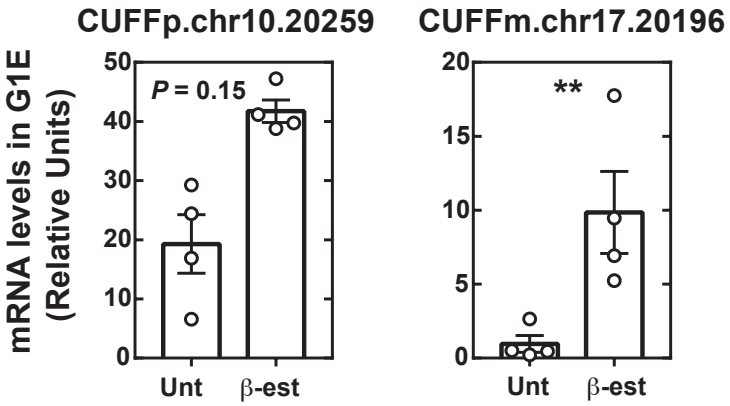

C
Scale
chr17:
10 kb
mm10
84,155,000
84,160,000
84,165,000
84,170,000
84,175,000
novel transcript models on - strand based on 32 RNA-seq data sets
CUFFm.chr17.20196.1
CUFFm.chr17.20196.3
CUFFm.chr17.20196.2
All Amit enhancers lifted from mm9 to mm10, labeled by its index and the number of regions it mapps to in mm10
enh21617.1
+9.5 element motif CANNTG-[N6-14]-AGATAA on models
mouse GATA2 or TAL1 ChIP-seq Peak
416B_TAL1
416B_GATA2
ERYpsu_TAL1
FLE_TAL1
FLERER_TAL1
G1ME_GATA2
G1Echb_GATA2_BMP
G1EER4psu_GATA2
G1EER4psu_TAL1
G1EER4psu2_GATA2
G1EER4psu2_TAL1
G1Epsu_GATA2
G1Epsu_TAL1
G1Epsu2_GATA2
G1Epsu2_TAL1
HPC7_GATA2
HPC7_TAL1
LIN_GATA2
LIN_TAL1
MEGpsu_TAL1
MELera_TAL1_DMSO
MELera_TAL1
MELpsu_TAL1
TERpsu2_TAL1
B_ATAC_BW
normalized BigWig for B ATAC-seq
CD4_ATAC_BW
normalized BigWig for CD4 ATAC-seq
CD8_ATAC_BW
normalized BigWig for CD8 ATAC-seq
CMP_ATAC_BW
normalized BigWig for CMP ATAC-seq
GMP_ATAC_BW
normalized BigWig for GMP ATAC-seq
Gn_ATAC_BW
normalized BigWig for Gn ATAC-seq
MEP_ATAC_BW
normalized BigWig for MEP ATAC-seq
Mo_ATAC_BW
normalized BigWig for Mo ATAC-seq
NK_ATAC_BW
normalized BigWig for NK ATAC-seq
Gata2_95MutRep1
Gata2_95MutRep1 normalized BigWig for DE
Gata2_95MutRep2
Gata2_95MutRep2 normalized BigWig for DE
Gata2_95MutRep3
Gata2_95MutRep3 normalized BigWig for DE
Gata2_95WtRep1
Gata2_95WtRep1 normalized BigWig for DE
Gata2_95WtRep2
Gata2_95WtRep2 normalized BigWig for DE
Gata2_95WtRep3
Gata2_95WtRep3 normalized BigWig for DE
Gata2_77MutRep1
Gata2_77MutRep1 normalized BigWig for DE
Gata2_77MutRep2
Gata2_77MutRep2 normalized BigWig for DE
Gata2_77MutRep3
Gata2_77MutRep3 normalized BigWig for DE
Gata2_77WtRep1
Gata2_77WtRep1 normalized BigWig for DE
Gata2_77WtRep2
Gata2_77WtRep2 normalized BigWig for DE
Gata2_77WtRep3
Gata2_77WtRep3 normalized BigWig for DE
Gata1InducedRep1
Gata1InducedRep1 normalized BigWig for DE
Gata1InducedRep2
Gata1InducedRep2 normalized BigWig for DE
Gata1InducedRep3
Gata1InducedRep3 normalized BigWig for DE
Gata1UninducedRep1
Gata1UninducedRep1 normalized BigWig for DE
Gata1UninducedRep2
Gata1UninducedRep2 normalized BigWig for DE
Gata1UninducedRep3
Gata1UninducedRep3 normalized BigWig for DE
ESExos10koRep1plus
ESExos10koRep1 plus normalized BigWig for DE
ESExos10koRep2plus
ESExos10koRep2 plus normalized BigWig for DE
ESExos10wtRep1plus
ESExos10wtRep1 plus normalized BigWig for DE
ESExos10wtRep2plus
ESExos10wtRep2 plus normalized BigWig for DE
ESExos10koRep1minus
ESExos10koRep1 minus normalized BigWig for DE
ESExos10koRep2minus
ESExos10koRep2 minus normalized BigWig for DE
ESExos10wtRep1minus
ESExos10wtRep1 minus normalized BigWig for DE
ESExos10wtRep2minus
ESExos10wtRep2 minus normalized BigWig for DE


D
Scale
chr10:
10 kb
mm10
77,230,000
77,235,000
77,240,000
77,245,000
77,250,000
77,255,000
novel transcript models on + strand based on 32 RNA-seq data sets
CUFFp.chr10.20259.2
All Amit enhancers lifted from mm9 to mm10, labeled by its index and the number of regions it mapps to in mm10
enh5071.1
enh5072.1
+9.5 element motif CANNTG-[N6-14]-AGATAA on models
N13-
mouse GATA2 or TAL1 ChIP-seq Peak
416B_TAL1
416B_GATA2
ERYpsu_TAL1
FLE_TAL1
FLERER_TAL1
G1ME_GATA2
G1Echb_GATA2_BMP
G1EER4psu_GATA2
G1EER4psu_TAL1
G1EER4psu2_GATA2
G1EER4psu2_TAL1
G1Epsu_GATA2
G1Epsu_TAL1
G1Epsu2_GATA2
G1Epsu2_TAL1
HPC7_GATA2
HPC7_TAL1
LIN_GATA2
LIN_TAL1
MEGpsu_TAL1
MELera_TAL1_DMSO
MELera_TAL1
MELpsu_TAL1
TERpsu2_TAL1
B_ATAC_BW
normalized BigWig for B ATAC-seq
CD4_ATAC_BW
normalized BigWig for CD4 ATAC-seq
CD8_ATAC_BW
normalized BigWig for CD8 ATAC-seq
CMP_ATAC_BW
normalized BigWig for CMP ATAC-seq
GMP_ATAC_BW
normalized BigWig for GMP ATAC-seq
Gn_ATAC_BW
normalized BigWig for Gn ATAC-seq
MEP_ATAC_BW
normalized BigWig for MEP ATAC-seq
Mo_ATAC_BW
normalized BigWig for Mo ATAC-seq
NK_ATAC_BW
normalized BigWig for NK ATAC-seq
Gata2_95MutRep1
Gata2_95MutRep1 normalized BigWig for DE
Gata2_95MutRep2
Gata2_95MutRep2 normalized BigWig for DE
Gata2_95MutRep3
Gata2_95MutRep3 normalized BigWig for DE
Gata2_95WtRep1
Gata2_95WtRep1 normalized BigWig for DE
Gata2_95WtRep2
Gata2_95WtRep2 normalized BigWig for DE
Gata2_95WtRep3
Gata2_95WtRep3 normalized BigWig for DE
Gata2_77MutRep1
Gata2_77MutRep1 normalized BigWig for DE
Gata2_77MutRep2
Gata2_77MutRep2 normalized BigWig for DE
Gata2_77MutRep3
Gata2_77MutRep3 normalized BigWig for DE
Gata2_77WtRep1
Gata2_77WtRep1 normalized BigWig for DE
Gata2_77WtRep2
Gata2_77WtRep2 normalized BigWig for DE
Gata2_77WtRep3
Gata2_77WtRep3 normalized BigWig for DE
Gata1InducedRep1
Gata1InducedRep1 normalized BigWig for DE
Gata1InducedRep2
Gata1InducedRep2 normalized BigWig for DE
Gata1InducedRep3
Gata1InducedRep3 normalized BigWig for DE
Gata1UninducedRep1
Gata1UninducedRep1 normalized BigWig for DE
Gata1UninducedRep2
Gata1UninducedRep2 normalized BigWig for DE
Gata1UninducedRep3
Gata1UninducedRep3 normalized BigWig for DE
ESExos10koRep1plus
ESExos10koRep1 plus normalized BigWig for DE
ESExos10koRep2plus
ESExos10koRep2 plus normalized BigWig for DE
ESExos10wtRep1plus
ESExos10wtRep1 plus normalized BigWig for DE
ESExos10wtRep2plus
ESExos10wtRep2 plus normalized BigWig for DE
ESExos10koRep1minus
ESExos10koRep1 minus normalized BigWig for DE
ESExos10koRep2minus
ESExos10koRep2 minus normalized BigWig for DE
ESExos10wtRep1minus
ESExos10wtRep1 minus normalized BigWig for DE
ESExos10wtRep2minus
ESExos10wtRep2 minus normalized BigWig for DE

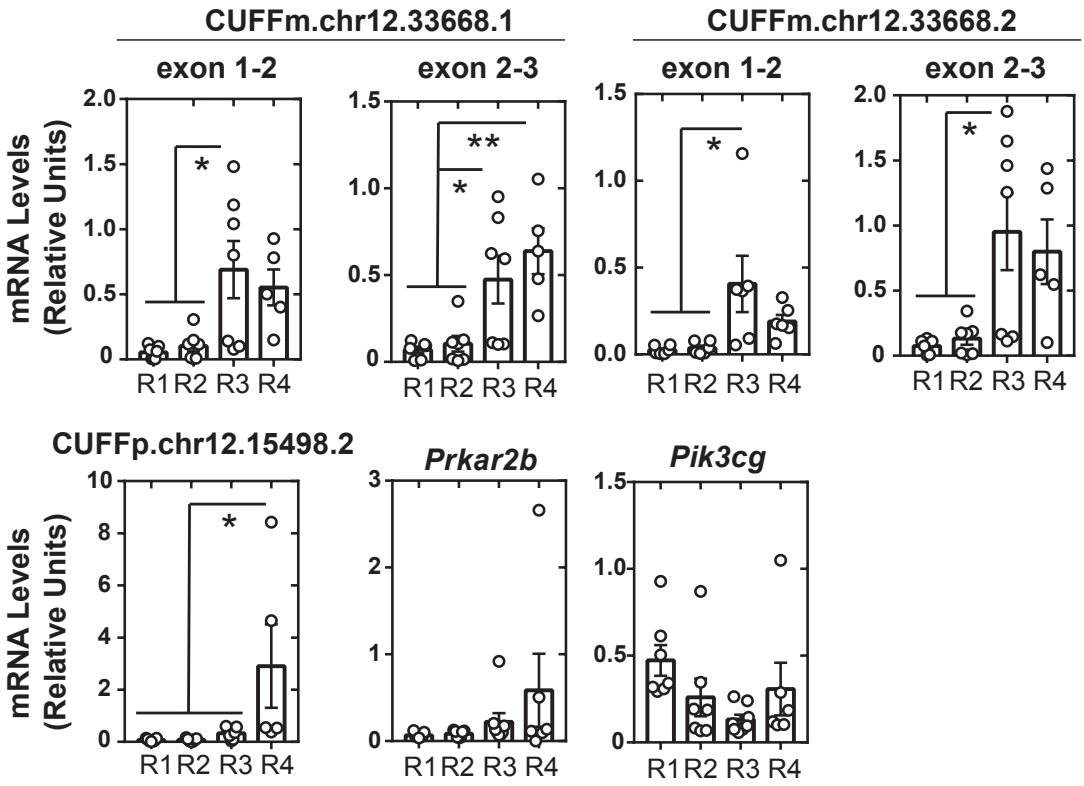
Supplementary Figure 35. The six PRAM mouse gene models and their genomic features. (**A**) CUFFm.chr12.32594; (**B**) CUFFm.chr12.33668 and CUFFp.chr12.15498; (**C**) CUFFm.chr17.20196; (**D**) CUFFp.chr10.20259; (**E**) CUFFm.chr10.13181.



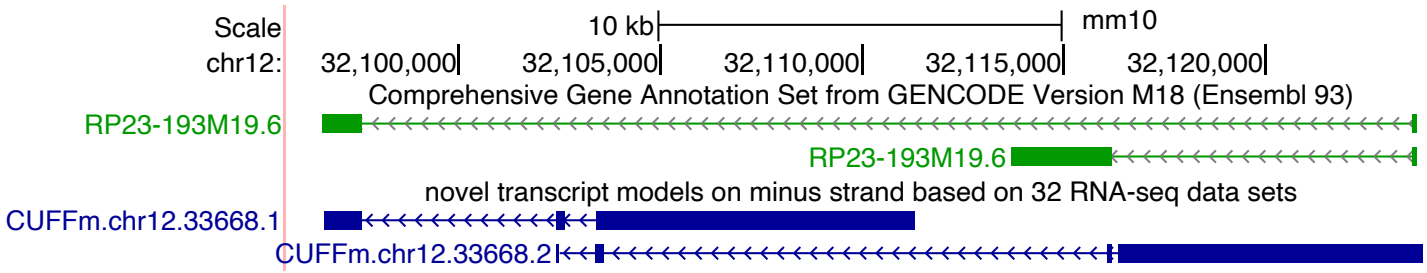
Supplementary Figure 36. Primer diagrams for PRAM mouse and human transcripts. Forward (F) and reverse (R) primers were denoted for PRAM mouse transcripts of CUFFm.chr12.33668 and CUFFp.chr12.15498, human K562 transcripts of CUFFm.chr7.6148. Primer sequences were listed in Supplementary Table 25. Prefixes of model names were removed for brevity.



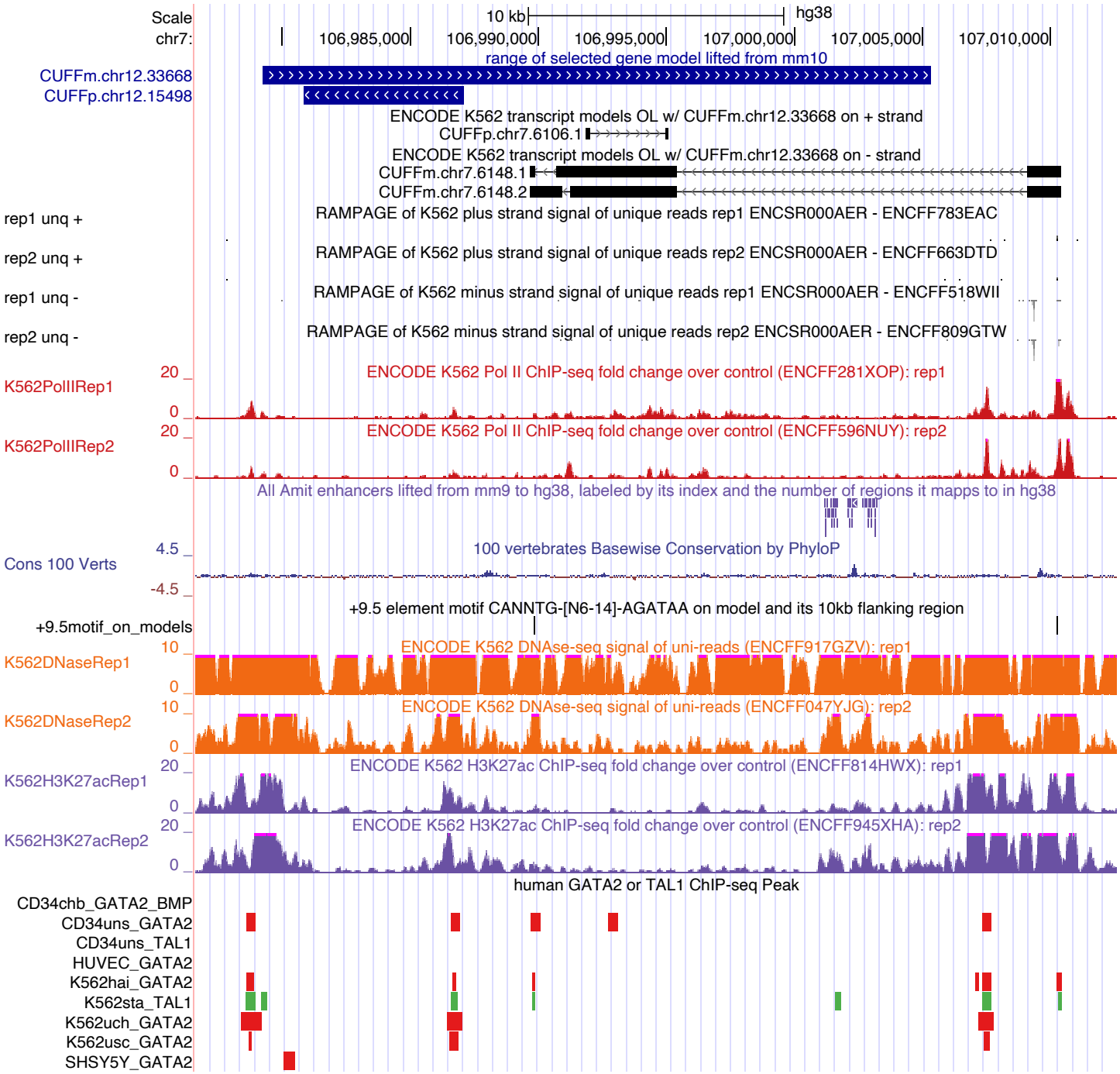
Supplementary Figure 37. CUFFp.chr10.20259 and CUFFm.chr17.20196 expression levels in G1E ER-GATA1 by qRT-PCR. Measurements were performed in untreated (Unt) and ß-estradiol-treated (ß-est) G1E-ER-GATA1 cells for 48 hours. P values were calculated by two-tailed Student's *t*-test (** for $p < 0.01$).



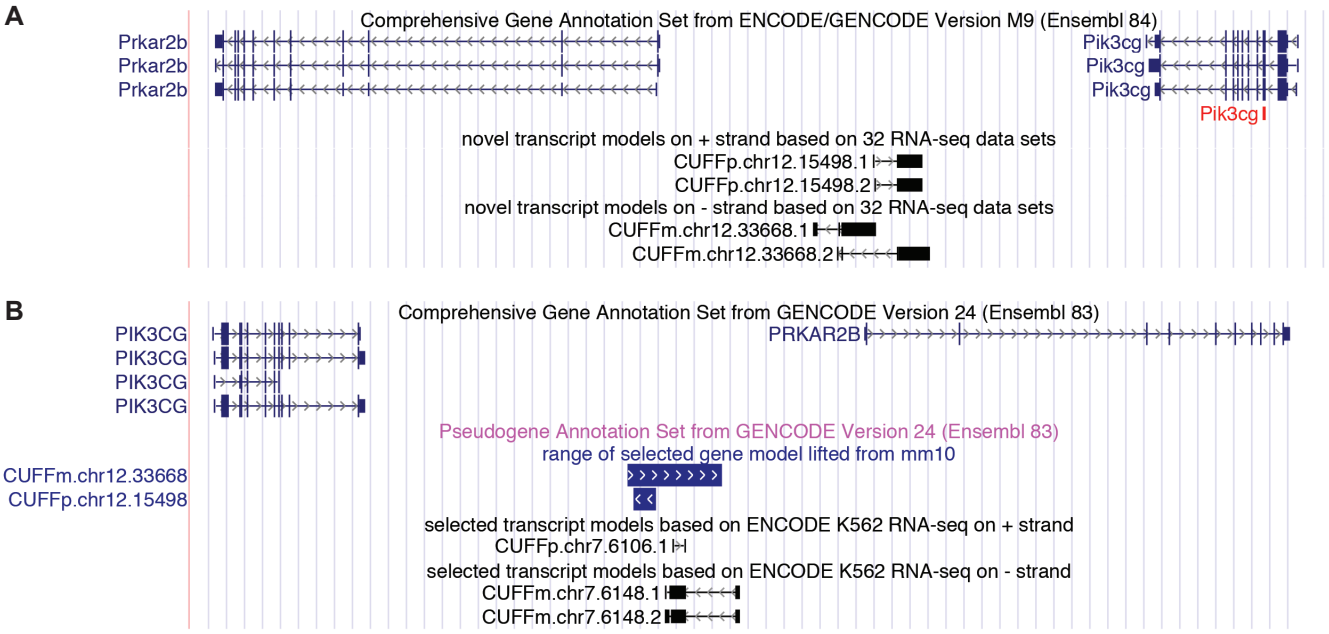
Supplementary Figure 38. Expression levels of PRAM models and their neighboring genes in sorted fetal liver cells by qRT-PCR. Expression levels of two mouse PRAM gene models CUFFm.chr12.33668 and CUFFp.chr12.15498 and their upstream and downstream neighbors *Prkar2b* and *Pik3cg* were measured by qRT-PCR during erythroid maturation (R1 to R4) of fetal liver cells. P values were calculated by two tailed Student's *t*-test (* for $p < 0.05$, ** for $p < 0.01$).



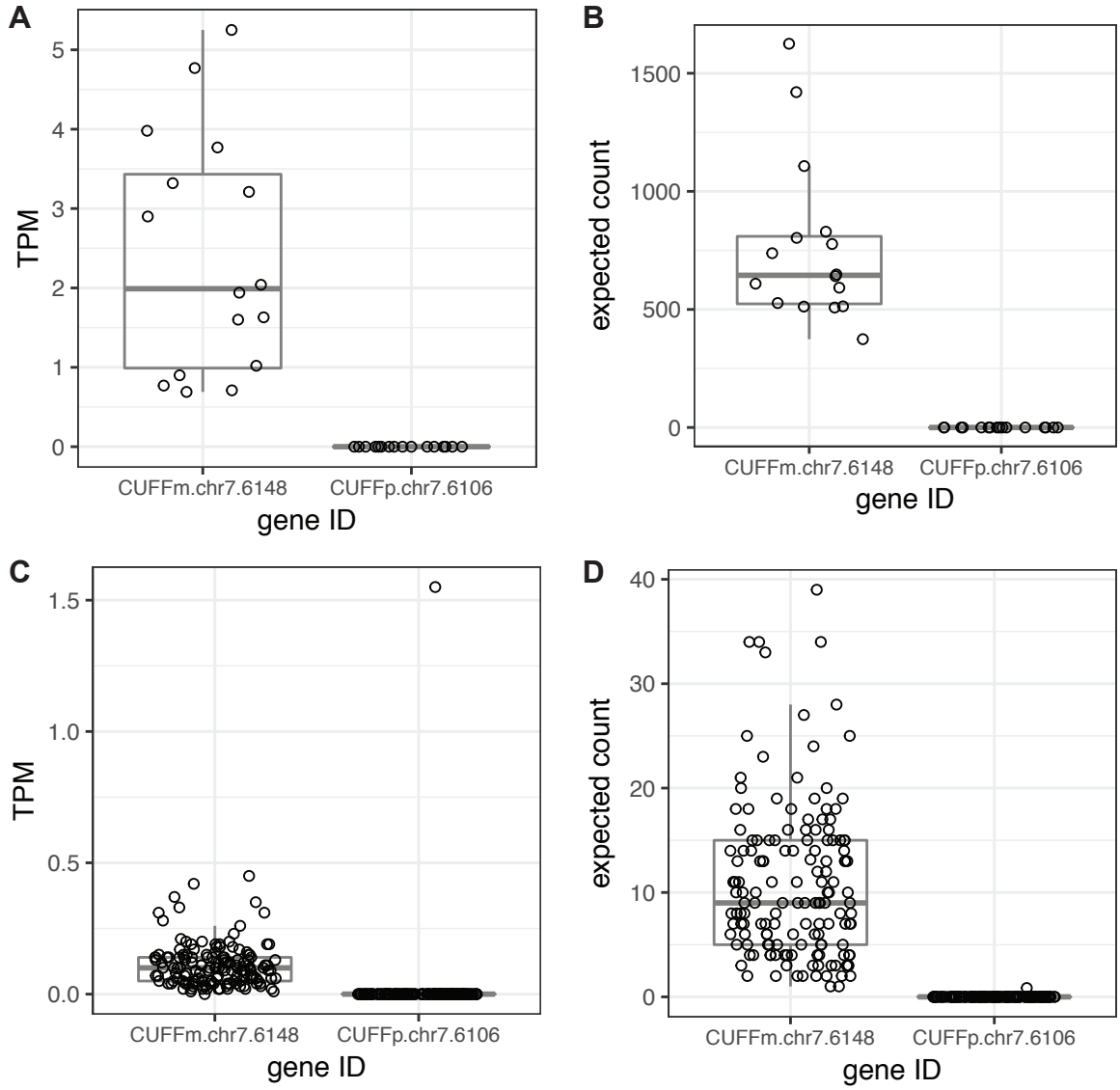
Supplementary Figure 39. PRAM mouse transcripts overlapped with newly annotated GENCODE transcripts. UCSC Genome Browser screenshot of PRAM mouse transcript CUFFm.chr12.33668.1 and CUFFm.chr12.33668.2 with transcripts from a recent mouse GENCODE annotation (vM18).



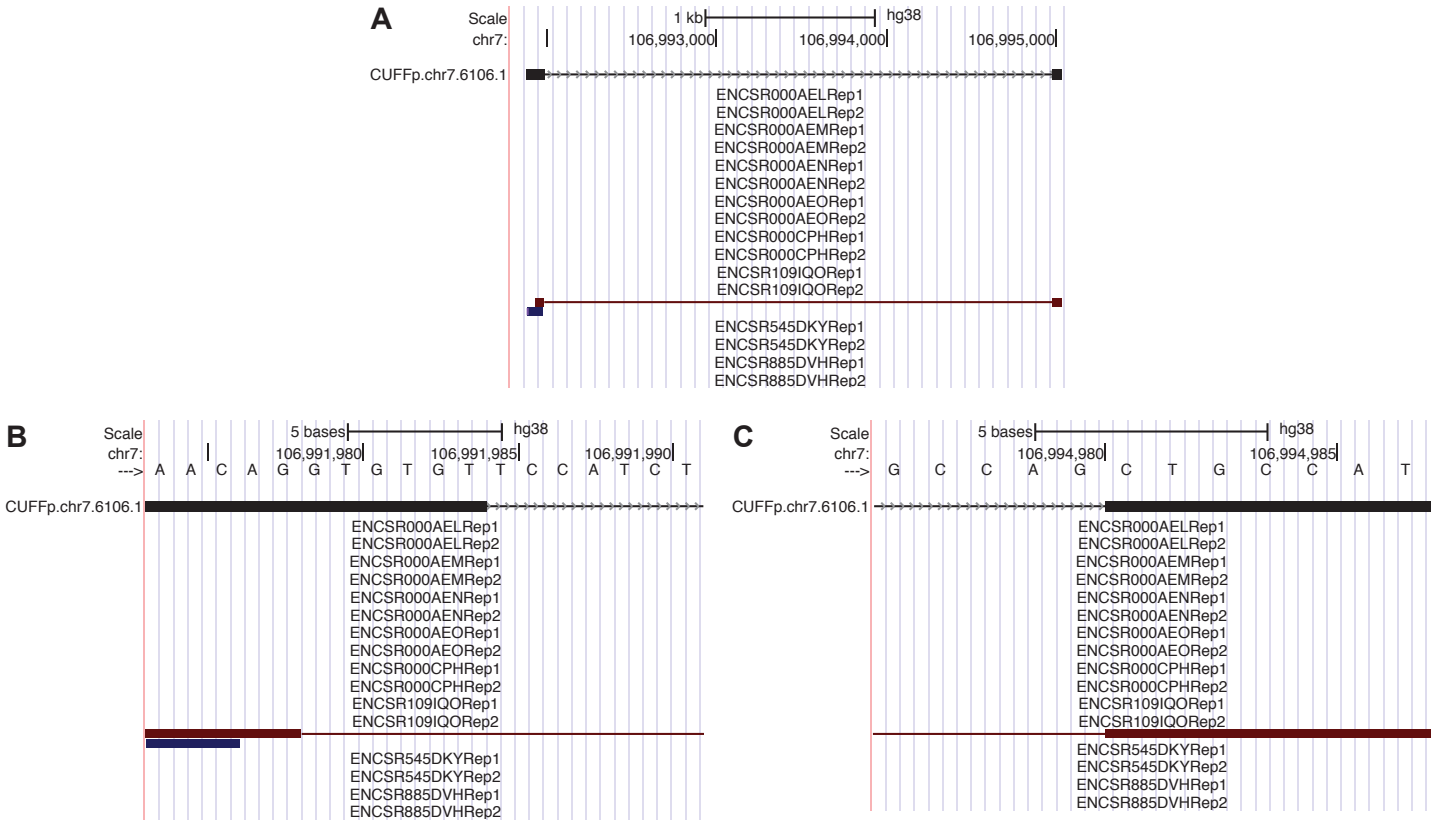
Supplementary Figure 40. PRAM K562 transcripts and their genomic features. Transcripts were built from K562 RNA-seq datasets (Supplementary Table 30) and resided in the lifted genomic range of experimentally validated PRAM mouse models CUFFm.chr12.33668. No model was found overlapping with lifted genomic range of CUFFp.chr12.15498.



Supplementary Figure 41. PRAM mouse and K562 transcripts and their neighboring genes. Synteny was maintained between mouse (**A**) and human (**B**).



Supplementary Figure 42. Estimated expression levels and fragment counts for PRAM K562 transcripts. (**A** & **C**) Estimated expression levels in K562 (**A**) and TCGA-LAML patients (**C**); (**B** & **D**) RNA-seq fragment counts in K562 (**B**) and TCGA-LAML patients (**D**).



Supplementary Figure 43. Splice sites and input RNA-seq fragments of CUFFp.chr7.6106. (**A**) Full structure of CUFFp.chr7.6106.1 and the paired-end RNA-seq fragment (mate1 in blue and mate2 in red) from ENCSR109IQO's replicate 2. This fragment is the only one from all the sixteen K562 RNA-seq datasets (Supplementary Table 30) that has a splice junction within the range of CUFFp.chr7.6106.1. Therefore, it should be the fragment that CUFFp.chr7.6106.1 was built on. (**B**) CUFFp.chr7.6106.1's 5'-splice site, which did not fit the RNA-seq fragment and was shifted by six bp, most likely due to Cufflinks's adjustment. (**C**) CUFFp.chr7.6106.1's 3'-splice site, which agreed with the input RNA-seq fragment.