High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing (CLS)

Authors

Julien Lagarde*[1,2], Barbara Uszczynska-Ratajczak*[1,2,6], Silvia Carbonell[3], Carrie Davis[4], Thomas R. Gingeras[4], Adam Frankish[5], Jennifer Harrow[5,7], Roderic Guigo#[1,2], Rory Johnson#[1,2,8]

* Equal contribution

Corresponding authors: rory.johnson@dkf.unibe.ch, roderic.guigo@crg.cat

Author affiliations

[1] Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain.

[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain.

[3] R&D Department, Quantitative Genomic Medicine Laboratories (qGenomics), Barcelona, Spain.

[4] Functional Genomics Group, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.

[5] Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK CB10 1HH.

[6] Present address: International Institute of Molecular and Cell Biology, Ks. Trojdena 4, 02-109 Warsaw, Poland

[7] Present address: Illumina, Cambridge, UK.

[8] Present address: Department of Clinical Research, University of Bern, Murtenstrasse 35, 3010 Bern, Switzerland.

Keywords

Long non-coding RNA; lncRNA; lincRNA; RNA sequencing; transcriptomics; GENCODE; annotation; CaptureSeq; third generation sequencing; long read sequencing; PacBio; KANTR.

Abbreviations

bp: base pair

FL: full length

nt: nucleotide

ROI: read of insert, *i.e.* PacBio reads

SJ: splice junction

SMRT: single-molecule real-time

TM: transcript model

Abstract

Accurate annotations of genes and their transcripts is a foundation of genomics, but no annotation technique presently combines throughput and accuracy. As a result, current reference gene collections remain far from complete: many genes models are fragmentary, while thousands more remain uncatalogued—particularly for long non coding RNAs (lncRNAs). To accelerate lncRNA annotation, the GENCODE consortium has developed RNA Capture Long Seq (CLS), combining targeted RNA capture with third generation long-read sequencing. We present an experimental re-annotation of the entire GENCODE intergenic lncRNA population in matched human and mouse tissues. CLS approximately doubles the annotated complexity of targeted loci, in terms of validated splice junctions and transcript models. The full-length transcript models produced by CLS enable us to definitively characterize the genomic features of lncRNAs, including promoter- and gene-structure, and protein-coding potential. Thus CLS removes a longstanding bottleneck of transcriptome annotation, generating manual-quality full-length transcript models at high-throughput scales.

Introduction

Long noncoding RNAs (lncRNAs) represent a vast and largely unexplored component of the mammalian genome. Efforts to assign lncRNA functions rest on the availability of high-quality transcriptome annotations. At present such annotations are still rudimentary: we have little idea of the total lncRNA count, and for those that have been identified, transcript structures remain largely incomplete.

The number and size of available lncRNA annotations have grown rapidly thanks to projects using diverse approaches. Early gene sets, deriving from a mixture of FANTOM cDNA sequencing efforts and public databases (1,2) were joined by the "lincRNA" (long intergenic non-coding RNA) sets, discovered through analysis of chromatin signatures (3). More recently, studies have applied *de novo* transcript-reconstruction software, such as *Cufflinks* (4) and *Scripture* (5), to identify novel genes in short-read RNA sequencing (RNAseq) datasets (6-10). However the reference for lncRNAs, as for protein-coding genes, has become the regularly-updated, manual annotations from GENCODE, which are based on curation of cDNAs and ESTs by human annotators (11). GENCODE has been adopted by most international genomics consortia (12-18).

At present, annotation efforts are caught in a trade-off between throughput and quality. *De novo* methods deliver large annotations with low hands-on time and financial investment. In contrast, manual annotation is relatively slow and requires long-term funding. However the quality of *de novo* annotations is often doubtful, due to the inherent difficulty of reconstructing transcript structures from much shorter sequence reads. Such structures tend to be incomplete, often lacking terminal exons or omitting splice junctions between adjacent exons (19). This particularly affects lncRNAs, whose low expression results in low read coverage (12). The outcome is a growing divergence between automated annotations of large size but uncertain quality (*e.g.* 90,062 for NONCODE (19)), and smaller but highly-curated "conservative" annotations of GENCODE (15,767 for version 25) (12).

Annotation incompleteness takes two forms. First, genes may be entirely missing from the annotation. Many genomic regions are suspected to transcribe RNA but presently contain no annotation, including "orphan" small RNAs with presumed long precursors (20), enhancers (21)

and ultraconserved elements (22,23). Similarly, thousands of single-exon predicted transcripts may be valid, but are generally excluded owing to doubts over their origin (12). The second form of incompleteness refers to missing or partial gene structures in already-annotated lncRNAs. Start and end sites frequently lack independent supporting evidence (12), and lncRNAs as annotated have shorter spliced lengths and fewer exons than mRNAs (8,12,24). Recently, RACE-Seq was developed to complete lncRNA annotations, but at relatively low throughput (24).

One of the principal impediments to lncRNA annotation arises from their low steady-state levels (3,12). To overcome this, targeted transcriptomics, or "RNA Capture Sequencing" (CaptureSeq) (25) is used to boost the concentration of known or suspected low-abundance transcripts in cDNA libraries. These studies have relied on Illumina short read sequencing and *de novo* transcript reconstruction (25-27), with accompanying doubts over transcript structure quality. Thus, while CaptureSeq achieves high throughput, its transcript structures lack the confidence required for inclusion in GENCODE.

In order to harness the power of CaptureSeq while eliminating *de novo* transcript assembly, we have developed RNA Capture Long Seq (CLS). CLS couples targeted RNA capture with third generation long-read cDNA sequencing. We used CLS to interrogate the GENCODE catalogue of intergenic lncRNAs, together with thousands of suspected novel loci, in six tissues each of human and mouse. CLS dramatically extends known annotations with high-quality novel structures. These data can be combined with 5' and 3' information to yield full-length transcript models in an automated way, allowing us to enumerate fundamental lncRNA promoter and gene structure properties for the first time. Thus CLS represents a significant advance in transcriptome annotation, and the dataset produced here advances our understanding of lncRNA's basic properties.

Results

Capture Long Seq approach to extend the GENCODE lncRNA annotation

Our aim was to develop an experimental approach to improve and extend reference transcript annotations, while minimizing human intervention and avoiding de novo transcript assembly. We designed a method, Capture Long Seq (CLS), which couples targeted RNA capture to Pacific Biosciences ("PacBio") Third Generation long-read sequencing (Figure 1A). The novelty of CLS is that it captures full-length, unfragmented cDNAs: this enables the targeted sequencing of low-abundance transcripts, while avoiding the uncertainty of assembled transcript structures from short-read sequencing.

CLS may be applied to two distinct objectives: to improve existing gene models, or to identify novel loci (blue and orange in Figure 1A, respectively). Although the present study focuses mainly on the former objective, we demonstrate also that novel loci can be captured and sequenced. Here we employed CLS with the goal of improving lncRNA annotations. With this in mind, we created a comprehensive capture library targeting the set of intergenic GENCODE lncRNAs in human and mouse. It should be noted that annotations for human are presently more complete than for mouse, and this accounts for the differences in the annotation sizes throughout (9,090 vs 6,615 genes, respectively). To these we added tiled probes targeting loci that may produce lncRNAs: small RNA genes (28), enhancers (29) and ultraconserved elements (30). For mouse we also added orthologue predictions of human lncRNAs from PipeR (31). Numerous control probes were added, including a series targeting half of the ERCC synthetic spike-ins (32). Together, these sequences were used to design capture libraries of temperature-matched and non-repetitive oligonucleotide probes (Figure 1B).

To access the maximal breadth of lncRNA diversity, we chose a set of transcriptionally-complex and biomedically-relevant organs from mouse and human: whole brain, heart, liver and testis (Figure 1C). To these we added two deeply-studied ENCODE human cell lines, HeLa and K562 (33), and two mouse embryonic time-points (E7 and E15).

We designed a protocol to capture full-length, oligo-dT-primed cDNAs (full details can be found in Materials and Methods). Barcoded, unfragmented cDNAs were pooled and captured.

Preliminary tests using quantitative PCR indicated strong and specific enrichment for targeted regions (Supplementary Figure S1). PacBio sequencing tends to favour shorter templates in a mixture (34). Therefore pooled, captured cDNA was size-selected into three ranges (1-1.5kb, 1.5-2.5kb, >2.5kb) (Supplementary Figure S2), and used to construct sequencing libraries for PacBio SMRT (single-molecular real-time) technology (35).

CLS yields an enriched long-read transcriptome

Samples were sequenced on altogether 130 SMRT cells, yielding ~2 million reads in total in each species (Figure 2A). PacBio sequence reads, or "reads of insert" (ROIs) were demultiplexed to retrieve their tissue of origin and mapped to the genome. We observed high mapping rates (>99% in both cases), of which 86% and 88% were unique, in human and mouse, respectively (Supplementary Figure S3). For brevity, all data are henceforth quoted in order of human then mouse. The use of short barcodes meant that, for ~30% of reads, the tissue of origin could not be retrieved (Supplementary Figure S4). This may be remedied in future by the use of longer barcode sequences. Representation was evenly distributed across tissues, with the exception of testis (Supplementary Figure S5). The ROIs had a median length of 1 - 1.5 kb (Figure 2B) consistent with previous reports (34) and longer than typical lncRNA annotation of ~0.5 kb (12).

Capture performance is assessed in two ways: by "on-target" rate – the proportion of reads originating from probed regions – and by enrichment, or increase of on-target rate following capture (36). To estimate this, we sequenced pre-capture libraries using MiSeq. CLS achieved on-target rates of 26.6% / 16.6%, representing 16- / 11-fold increase over pre-capture RNA (Figure 2C, D and Supplementary Figure S6). The majority of off-target signal arises from non-targeted, annotated protein-coding genes (dark blue in Figure 2C).

CLS on-target rates were lower than previous studies using fragmented cDNA (36). Side-by-side comparisons showed that this is likely due to the lower efficiency of capturing long cDNA fragments (Supplementary Figure S7), as observed by others (27), and thus representing a future target for protocol optimization.

Synthetic spike-in sequences at known concentrations were used to assess CLS sensitivity and quantitativeness. We compared the relationship of sequence reads to starting concentration for the 42 probed (green) and 50 non-probed (violet) synthetic ERCC sequences in pre- and post-capture samples (Figure 2E, top and bottom rows). Given the low sequencing depth, CLS is surprisingly sensitive, extending detection sensitivity by two orders of magnitude, and capable of detecting molecules at approximately 0.005 copies per cell (Materials and Methods). As expected, it is less quantitative than conventional CaptureSeq (27), particularly at higher concentrations where the slope falls below unity. This suggests saturation of probes by cDNA molecules during hybridisation. A degree of noise, as inferred by the coefficient of determination (R^2) between read counts and templates, is introduced by the capture process (R^2 of 0.63 / 0.87 in human post-capture and pre-capture, respectively).

CLS expands the complexity of known and novel lncRNAs.

CLS discovers a wealth of novel transcript structures within annotated lncRNA loci. A good example is the *SAMMSON* oncogene (*LINC01212*) (37), where we discover a variety of new exons, splice sites, and transcription termination sites that are not present in existing annotations (Figure 3A, more examples in Supplementary Figures S8, S9, S10).

Gathering the non-redundant union of all ROIs, we measured the amount of new complexity discovered in targeted lncRNA loci. CLS detected 58% and 45% of targeted lncRNA nucleotides, and extended these annotations by 6.3 / 1.6 Mb nucleotides (86% / 64% increase compared to existing annotations) (Supplementary Figure S11). CLS discovered 45,673 and 11,038 distinct splice junctions (SJs), of which 36,839 and 26,715 are novel (Figure 3B and Supplementary Figure S12, left bars). For independent validation, we deeply sequenced captured cDNA by Illumina HiSeq at an average depth of 35 million / 26 million pair-end reads per tissue sample. Split reads from this data exactly matched 78% / 75% SJs from CLS. These "high-confidence" SJs alone represent a 160% / 111% increase over the existing, probed annotations. Novel high-confidence lncRNA SJs are rather tissue-specific, with greatest numbers observed in testis (Supplementary Figure S13), and were also discovered across other classes of targeted and non-targeted loci (Supplementary Figure S14).

To further evaluate novel lncRNA SJs, we computed their strength using standard position weight matrix models from donor and acceptor sites (38) (Figure 3C, Supplementary Figure S15). High-confidence novel SJs from lncRNAs (orange, upper panel) far exceed the predicted strength of background SJ-like dinucleotides (bottom panels), and are essentially indistinguishable from annotated SJs in protein-coding and lncRNA loci (pink, upper and middle panels). Even unsupported, novel SJs (black) tend to have high scores in excess of background, although with a significant low-scoring tail. Novel SJs also display weak but non-random evidence of selected function between human and mouse (Supplementary Figure S16).

We estimated how close these sequencing data are to saturation of true gene structures, that is, to reaching a definitive lncRNA annotation. In each tissue sample, we tested the rate of novel splice junction and transcript model discovery as a function of increasing depth of randomly-sampled ROIs (Figure 3D, Supplementary Figures S17, S18). We observed an ongoing gain of novelty with increasing depth, up to that presented here. Thus, considerable additional sequencing is required to fully define the complexity of annotated GENCODE lncRNAs.

Beyond lncRNA characterization, CLS can be of utility to characterize many other types of transcriptional units. As an illustration, we searched for precursor transcripts of small RNAs (microRNAs, snoRNAs and snRNAs), whose annotation remains poor (20). We probed 1 kb windows around all "orphan" small RNAs, *i.e.* those with no annotated overlapping transcript. We identified more than one hundred likely exonic primary transcripts, and hundreds more potential precursors harbouring small RNAs within their introns (Figure 3E). One intriguing example was the cardiac-enriched hsa-mir-143 (Supplementary Figure S19). We previously identified a standalone lncRNA in the same locus, *CARMEN1*, which is necessary for cardiac precursor cell differentiation (39). CLS identifies a new isoform of *CARMEN1* that overlaps hsa-mir-143, suggesting it is a bifunctional lncRNA directing a complex auto-regulatory feedback loop in cardiogenesis.

Assembling a full-length lncRNA annotation.

A unique benefit of the CLS approach is the ability to identify full-length transcript models with confident 5' and 3' termini. ROIs of oligo-dT-primed cDNAs carry a fragment of

the poly(A) tail, which can identify the transcription termination site (TTS) with basepair precision (34). Using conservative filters, 73 / 64% of ROIs had identifiable TTS sites (Supplementary Table S1) representing 16,961 / 12,894 novel TTS when compared to end positions of all GENCODE annotations. Both known and novel TTS were accompanied by canonical polyadenylation motifs (Supplementary Figure S20). Similarly, the 5' completeness of ROIs was confirmed by proximity to methyl-guanosine caps identified by CAGE (Cap Analysis of Gene Expression) (18) (Supplementary Figure S21). Together, TSS and TTS sites were used to define the 5' / 3' completeness of all ROIs (Figure 4A).

We developed a pipeline to merge ROIs into a non-redundant collection of transcript models (TMs). In contrast to previous approaches (4), our "anchored merging" method respects confirmed internal TSS or TTS sites (Figure 4B). Applying this to captured ROIs results in a greater number of unique TMs than would be identified otherwise (Figure 4C, Supplementary Figure S22). Specifically, we identified 179,993/129,556 transcript models across all biotypes (Supplementary Table S2). The *CCAT1* locus is an example where several novel transcripts are identified, each with CAGE and polyA support of 5' and 3' termini, respectively (Figure 4D). CLS here suggests that adjacent *CCAT1* and *CASC19* gene models are in fact fragments of the same underlying gene (40).

Merged TMs can be defined by their end support: full length (5' and 3' supported), 5' only, 3' only, or unsupported (Figure 4B, E, F). We identified a total of 65,736 / 44,673 full length (FL) transcript models (Figure 4E, F, left panels): 47,672 (73%) / 37,244 (83%) arise from protein coding genes, and 13,071 (20%) / 5,329 (12%) from lncRNAs (Supplementary Table S2). An additional 3,742 (6%) / 1,258 (3%) represent FL models that span loci of different biotypes (listed in Figure 1B), usually including one protein-coding gene ("Multi-Biotype"). Of the remaining non-coding FL transcript models, 295 / 434 are novel, arising from unannotated gene loci. Altogether, 11,429 / 4,350 full-length structures arise from probed lncRNA loci, of which 8,494 / 3,168 (74% / 73%) are novel (Supplementary Table S2). We identified at least one FL read for 19% / 17% of the originally-probed lncRNA annotation (Figure 4G, Supplementary Figure S23).

In addition to probed lncRNA loci, CLS also discovered several thousand novel TMs originating from unannotated regions, mapping to probed (blue in Figure 1B) or unprobed

regions (Supplementary Figures S24, S25). These TMs tended to have lower detection rates (Supplementary Figure S26) and lower rates of 5' and 3' support than probed lncRNAs, although a small number are full length ("other" in Figure 4E, F right panels). A recent CaptureSeq study probed the human GENCODE v12 annotation and sequenced fragmented cDNA using Illumina short read technology (27). The study captured the unfiltered lncRNA annotation (both intergenic and overlapping protein-coding genes) in a panel of 20 human tissues and four cell lines sequenced on five lanes of HiSeq 2000. Bearing in mind the differences in capture design and sequencing depth, we compared the annotations resulting from that study and the CLS approach using an identical pipeline. CLS yields a higher proportion of novel TMs, counting either all TMs or FL models alone (Supplementary Figure S27). More importantly, FL TMs are almost twice as long as *de novo* models (Figure 4H). We estimate that this was achieved by 1630-fold less nucleotides sequenced (Supplementary Figure S28).

Together, these findings show that CLS is effective in creating large numbers of full-length transcript annotations for probed gene loci, in a highly scalable way.

Re-defining lncRNA promoter and gene characteristics with full-length annotations

With a full-length lncRNA catalogue, we could revisit the question of fundamental differences of lncRNA and protein-coding genes.

Existing lncRNA transcripts, as annotated, are significantly shorter and have less exons than mRNAs (6,12). However it has remained unresolved whether this is a genuine biological trend, or simply the result of annotation incompleteness (24). Considering FL TMs, we find that the median lncRNA transcript to be 1150 / 1138 nt, similar to mRNAs mapped by the same criteria (1236 / 1317 nt) (Figure 5A, Supplementary Figure S29). This is still shorter than the median length of annotated protein-coding transcripts (1,543 nt in GENCODE v20), but much larger than the median of annotated lncRNAs (668 nt). Because of the length limitation of PacBio reads, we cannot conclude that lncRNAs are the same length as mRNAs; however, we do not find evidence here that they are shorter either.

We previously observed a striking enrichment for two-exon genes in lncRNAs, which was not observed in mRNAs (12). However, this is clearly an artefact arising from annotation incompleteness: the mean number of exons for lncRNAs in the FL models is 4.27, compared to 6.69 for mRNAs (Figure 5B, Supplementary Figure S29). This difference is explained by lncRNAs' longer exons, although with a peak of approximately 150 bp, or one nucleosomal turn (Supplementary Figure S30).

The usefulness of TSS annotation used here is demonstrated by the fact that FL transcripts' TSS are, on average, closer than existing annotations to expected promoter features, including promoters and enhancers predicted by genome segmentations (41) and CpG islands, although not evolutionarily-conserved elements or phenotypic GWAS sites (42) (Figure 5C). More accurate mapping of lncRNA promoters in this way may provide new hypotheses for the latter's' mechanism of action. For example, an improved 5' annotation strengthens the link between GWAS SNP rs246185, correlating with QT-interval and lying in the promoter of heart- and muscle-expressed RP11-65J2 (ENSG00000262454), for which it is an expression quantitative trait locus (eQTL) (Supplementary Figure S31) (43).

The improved 5' definition provided by CLS transcript models also enables us to compare lncRNA and mRNA promoters. Recent studies, based on the start position of gene annotations, have pointed to strong apparent differences across a range of features (44,45). To make fair comparisons between gene sets, we created an expression-matched set of mRNAs in HeLa and K562 cells, and removed bidirectional promoters. These were compared across a variety of datasets from ENCODE (15) (Supplementary Figures S32, S33).

We observe a series of similar and divergent features of lncRNAs' and mRNAs' promoters. For example, activating promoter histone modifications such as H3K4me3 (Figure 5D) and H3K9ac (Figure 5E), are essentially indistinguishable between full-length lncRNAs (dark blue) and protein-coding genes (red), suggesting that, when accounting for expression differences, active promoter architecture of lncRNAs is not unique. The contrast of these findings with previous reports, suggest that the latter's reliance on annotations alone led to inaccurate promoter identification (44,45).

On the other hand, and as observed previously, lncRNA promoters are distinguished by elevated levels of repressive chromatin marks, such as H3K9me3 (Figure 5F) and H3K27me3

(Supplementary Figures S32, S33). This may be the consequence of elevated recruitment to lncRNAs of Polycomb Repressive Complex, as evidenced by its subunit Ezh2 (Figure 5G). Surprisingly, we also observed that the promoters of lncRNAs are distinguished from those of protein-coding genes by a localised peak of insulator protein CTCF binding (Figure 5H). Finally, there is a clear region of evolutionary conservation at lncRNA promoters, although lower than for protein-coding genes (Figure 5G).

Two conclusions are drawn. First, that CLS-inferred TSS have greater density of expected promoter features, compared to corresponding GENCODE annotations, implying that CLS improves TSS annotation. And second, that when adjusting for expression, lncRNA have comparable activating histone modifications, but distinct repressive modifications, compared to protein-coding genes.

Discovery of new potential open reading frames.

Recently a number of studies have suggested that many lncRNA loci encode peptide sequences through unannotated open reading frames (ORFs) (46,47). We searched for signals of protein-coding potential in FL models using two complementary methods, based on evolutionary conservation and intrinsic sequence features (Figure 6A, Materials and Methods, Supplementary Data File 1) (48,49). This analysis finds evidence for protein-coding potential in a small fraction of lncRNA FL TMs (109/1271=8.6%), with a similar number of protein-coding FL TMs displaying no evidence of encoding protein (2900/42,758=6.8%).

CLS FL models may lead to reclassification of protein-coding potential for seven cases in five distinct gene loci (Figure 6C, Supplementary Figure S34, Supplementary Data File 2). A good example is the *KANTR* lncRNA, where CLS identifies an unannotated exon harbouring a placental mammal-conserved 76aa ORF with no detectable protein orthologue, composed of two sequential transmembrane domains (Figure 6D, Supplementary Figure S35) (50). This region derives from the antisense strand of a LINE1 transposable element. Another case is *LINC01138*, linked with prostate cancer, where a potential 42aa ORF is found in the extended transcript (51). Again, this ORF has no identifiable domains or orthologues. We could not find peptide evidence for either ORF's translation (see Materials and Methods). Whole-cell expression, as well as

cytoplasmic-to-nuclear distributions, also showed that potentially protein-coding lncRNAs' behaviour is consistently more similar to annotated lncRNAs than to mRNAs (Supplementary Figures S36, S37, S38). Together, these findings demonstrate the utility of CLS in improving the biotype annotation of the small minority of lncRNAs that may encode proteins.

Discussion

We have introduced an annotation methodology that resolves the competing needs of quality and throughput. Capture Long Read Sequencing produces transcript models with quality approaching that of human annotators, yet with throughput comparable to *de novo* transcriptome assembly. In fact, by incorporating 5' and 3' mapping, CLS advances beyond all contemporary annotation methods by providing full-length transcript models.

In the context of GENCODE, CLS will be used to accelerate annotation pipelines. Transcript models, accompanied by meta-data describing 5', 3' and splice junction support, will be stratified by confidence level. These will receive attention from human annotators as a function of their incompleteness, with FL TMs passed directly to published annotations. Future workflows will utilise d*e novo* models from short read data from diverse cell types and developmental time points to perform new rounds of CLS. This approach lays the path towards a truly comprehensive human transcriptome annotation.

CLS is appropriate for virtually any class of RNA transcript. CLS' versatility and throughput makes it suited to rapid, low-cost transcriptome annotation in non-model organisms. Preliminary bioinformatic homology screens for potential genes (including protein-coding, lncRNAs, microRNAs etc.), in newly-sequenced genomes, or first-pass short read RNA-Seq, could be used to design capture libraries. Resulting annotations would be substantially more accurate than those produced by current pipelines based on homology and short-read data.

In economic terms, CLS is also competitive. Using conservative estimates, with 2016 prices ($1700 for 1 lane of PE100bp HiSeq, $500 for 1 SMRT), and including the cost of sequencing alone, we estimate that CLS yielded one novel lncRNA structure for every $0.70 spent, compared to $0.40 for CaptureSeq. This difference is almost entirely accounted for by the lower on-target rate of CLS, which is likely to improve in the near future.

CLS could also be applied to personal genomics studies. Targeted sequencing of gene panels, perhaps those with medical relevance, could examine the little-studied question of alternative transcript variability across individuals—i.e. whether there exist isoforms that are private to given individuals or populations.

Full-length annotations have provided the first confident view of lncRNA gene properties. These are more similar to mRNAs than previously thought, in terms of spliced length and exon count (12,52). A similar trend is seen for promoters: when lncRNA promoters are accurately mapped by CLS and compared to matched protein-coding genes, we find them to be surprisingly similar for activating modifications. This suggests that previous studies, which placed confidence in annotations of TSS, should be reassessed (44,45). On the other hand lncRNA promoters do have unique properties, including elevated levels of repressive histone modification, recruitment of Polycomb group proteins, and interaction with the insulator protein CTCF. To our knowledge, this is the first report to suggest a relationship between lncRNAs and insulator elements. Overall, these results suggest that that lncRNA gene features *per se* are generally comparable to mRNAs, after normalising for their differences in overall expression. Finally, extended TMs do not yield evidence for widespread protein-coding capacity encoded in lncRNAs.

In summary, by resolving a longstanding roadblock in lncRNA transcript annotation, the CLS approach promises to dramatically accelerate our progress towards an eventual "complete" mammalian transcriptome annotation. These updated lncRNA catalogues represent a valuable resource to the genomic and biomedical communities, and address fundamental issues of lncRNA biology.

Figure legends

Figure 1: Capture Long Seq approach to extend the GENCODE lncRNA annotation.

(A) Strategy for automated, high-quality transcriptome annotation. CLS may be used to complete existing annotations (blue), or to map novel transcript structures in suspected loci (orange). Capture oligonucleotides (black bars) are designed to tile across targeted regions. PacBio libraries are prepared for from the captured molecules. Illumina HiSeq short-read sequencing can be performed for independent validation of predicted splice junctions. Predicted transcription start sites can be confirmed by CAGE clusters (green), and transcription termination sites by non-genomically encoded polyA sequences in PacBio reads. Novel exons are denoted by lighter coloured rectangles.

(B) Summary of human and mouse capture library designs. Shown are the number of individual gene loci that were probed. "PipeR pred.": orthologue predictions in mouse genome of human lncRNAs, made by PipeR (31); "UCE": ultraconserved elements; "Prot. coding": expression-matched, randomly-selected protein-coding genes; "ERCC": spike-in sequences; "Ecoli": randomly-selected *E. coli* genomic regions. Enhancers and UCEs are probed on both strands, and these are counted separately. "Total nts": sum of targeted nucleotides.

(C) RNA samples used.

Figure 2: CLS yields an enriched, long-read transcriptome.

(A) Summary statistics for long-read sequencing. ROI = "Read Of Insert", or PacBio reads.

(B) Length distributions of ROIs. Sequencing libraries were prepared from three size-selected cDNA fractions (see Supplementary Figure S2).

(C) Breakdown of detected regions by gene biotype, pre- and post-capture. *Y*-axis: Number of read overlaps before (left, "Pre-capture") and after (right, "Post-capture") capture. Illumina MiSeq was used to sequence Pre-capture samples. Colours denote the biotype of the genomic feature, targeted or not, from which the reads originate, notably: Dark blue:

"off-target genic" reads, mainly representing highly-expressed protein-coding genes; light blue: lncRNAs targeted in the capture design; pink: targeted protein-coding genes. Note that when a given read overlapped more than one targeted class of regions, it was counted in each of these classes separately. As a consequence, the maximum of each bar plot exceeds the total number of sequenced reads.

(D) Summary of capture performance. The *y*-axis shows the percent of all mapped ROIs originating from a targeted region ("on-target"). Enrichment is defined as the ratio of this value in Post- and Pre-capture samples. Note that Pre- and Post-capture on-target rates were calculated using MiSeq and PacBio reads, respectively, although similar results were obtained when using MiSeq also for the Post-capture samples.

(E) Response of read counts in captured cDNA to input RNA concentration. Upper panels: Pre-capture; lower panels: Post-capture. Left: human; right: mouse. Note the log scales for each axis. Each point represents one of 92 spiked-in synthetic ERCC RNA sequences. 42 were probed in the capture design (green), while the remaining 50 were not (violet). Lines represent linear fits to each dataset, whose parameters are shown above. Given the log-log representation, a linear response of read counts to template concentrate should yield an equation of type $y = c + mx$, where m is 1.

Figure 3: Extending known lncRNA gene structures

(A) Novel transcript structures from the *SAMMSON* (*LINC01212*) locus. Annotation as present in GENCODE v20 is shown in green, capture probes in grey, CLS reads in black (confirming known structure) and red (novel structures).

(B) Novel splice junction (SJ) discovery. The *y*-axis denotes counts of unique SJs for human (equivalent mouse data in Supplementary Figure S12). Only "on-target" junctions originating from probed lncRNA loci are considered. Grey represents annotated SJs that are not detected. Dark green represents GENCODE-annotated SJs that are detected by CLS. Light green represent novel SJs that are identified by CLS but not annotated. The left column represents all SJs, and the right column represents only high-confidence SJs, supported by at least one split-read from Illumina short read sequencing.

(C) Splice junction (SJ) motif strength. Panels plot the distribution of predicted SJ strength, for acceptors (left) and donors (right). Data shown are for human, equivalent analysis for mouse may be found in Supplementary Figure S15. The strength of the splice sites were computed using standard position weight matrices used by GeneID (38). Data are shown for non-redundant SJs from CLS transcript models from targeted lncRNAs (top), all annotated protein-coding genes (middle), or a background distribution sampled from randomly-selected AG (acceptor-like) and GT (donor-like) dinucleotides.

(D) Novel splice junction discovery as a function of sequencing depth in human. Each panel represents the number of novel splice junctions (SJs) discovered (*y*-axis) in simulated analysis where increasing numbers of ROIs (*x*-axis) were randomly sampled from the experiment. The randomizations were repeated a hundred times, and a boxplot summarizes the results at each simulated depth. The highest *y* value represents the actual number of novel SJs discovered. Equivalent data for mouse can be found in Supplementary Figure S17, and for rates of novel transcript model discovery in Supplementary Figure S18.

(E) Identification of putative precursor transcripts of small RNA genes. For each gene biotype, the figures show the count of unique genes in each group. "Orphans" are those with no annotated same-strand overlapping transcript in GENCODE, and were used for capture probe design in this project. "Pot. Precursors" (potential precursors) represent orphan small RNAs that reside in the intron of and on the same strand as a novel transcript identified by CLS; "Precursors" represent those that reside in the exon of a novel transcript.

Figure 4: Full-length transcript annotation.

(A) The 5' (transcription start site, TSS) and 3' (transcription termination site, TTS) termini of new transcript models can be inferred using CAGE clusters and sequenced polyA tails, respectively. The latter correspond to polyA fragments identified at ROI 3' ends that are not genomically-encoded.

(B) "Anchored" merging of ROIs to create transcript models, while respecting their TSS and TTS. In conventional merging (left), transcripts' TSS and TTS are lost when they overlap

exons of other transcripts. Anchored merging (right) respects and does not collapse TSS and TTS that fall within exons of other transcripts.

(C) Anchored merging yields more distinct transcript models. The *y*-axis represents total counts of ROIs (pink), anchor-merged transcript models (brown) and conventionally-merged transcript models (turquoise). Transcript models were merged irrespective of tissue-origin.

(D) Example of full-length, TSS- and TTS-mapped transcript models at the human *CCAT1* / *CASC19* locus. GENCODE v20 annotation is shown in green, novel full-length CLS models in red. Note the presence of a CAGE-supported TSS (green star) and multiple distinct polyA-mapped TTS sites (red stars).

(E) The total numbers of anchor-merged transcript models identified by CLS for human. The *y*-axis of each panel shows unique transcript model (TM) counts. Left panel: All merged TMs, coloured by end-support. Middle panel: Full length (FL) TMs, broken down by novelty with respect to existing GENCODE annotations. Green areas are novel and multi-exonic: "overlap" intersect an annotation on the same strand, but do not respect all its splice junctions; "intergenic" overlap no annotation on the same strand; "extension" respect all of an annotation's splice junctions, and add novel ones. Right panel: Novel FL TMs, coloured by their biotype. "Other" refers to transcripts not mapping to any GENCODE protein-coding or lncRNA annotation. Note that the majority of "multi-biotype" models link a protein-coding gene to another locus.

(F) As (E), but for mouse.

(G) The number of unique full-length transcript models identified per probed lncRNA locus in human. The total number of probed loci is shown above. Note that probed loci giving rise to more than 20 full-length transcripts were aggregated into a single bin, labelled ">20" on the *x*-axis. Equivalent mouse data can be found in Supplementary Figure S23.

(H) Comparing the length of annotations from CLS full-length models and Cufflinks models of a recent CaptureSeq project using short-read sequencing in human (27). Note that only models mapping to GENCODE v20 lncRNAs were considered for each. Note that discrepancies between TM numbers shown here and elsewhere, result from the method used to make a fair comparison of CLS and short-read data. Here, the standard *Cuffcompare* approach was used to count transcripts mapping to GENCODE annotation.

Figure 5: Discovery of novel lncRNA transcripts.

(A) The mature, spliced transcript length of: CLS full-length transcript models from targeted lncRNA loci (dark blue); transcript models from the targeted and detected GENCODE lncRNA loci (light blue); CLS full-length transcript models from protein-coding loci (red).

(B) The numbers of exons per full length transcript model, from the same groups as in (A). Dotted lines represent medians.

(C) Distance of annotated transcription start sites (TSS) to genomic features. Each cell displays the mean distance to nearest neighbouring feature for each TSS. TSS sets correspond to the classes from (A). "Shuffled" represent FL lincRNA TSS randomly placed throughout genome.

(D) – (G) Comparing promoter profiles across gene sets. The aggregate density of various features is shown across the TSS of indicated gene classes. Note that overlapping TSS were merged within classes, and TSSs belonging to bi-directional promoters were discarded (see Methods). The *y*-axis denotes the mean signal per TSS, and grey fringes represent the standard error of the mean. Gene sets are: Dark blue, full length lncRNA models from CLS; Light blue, the GENCODE annotation models from which the latter were derived; Red, a subset of protein-coding genes with similar expression in HeLa as the CLS lncRNAs.

Figure 6: Properties of full-length lncRNAs.

(A) The predicted protein-coding potential of all full-length transcript models mapped to lncRNA (left) or protein-coding loci (right). Each point represents a single full length (FL) transcript model (TM). The *y*-axis displays the coding likelihood according to *PhyloCSF*, based on multiple genome alignments, while the *x*-axis displays that calculated by CPAT, an alignment free method. Red lines indicate score thresholds, above which transcript models are considered protein-coding. Models mapping to multiple different biotypes were not considered.

(B) The numbers of classified transcript models (TMs) from (A).

(C) Discovery of new protein-coding transcripts as a result of full-length CLS reads, using PhyloCSF. For each probed lncRNA locus, we calculated the transcript isoform with highest scoring ORF (*x*-axis). From each locus, we identified the full-length transcript model with high scoring ORF (*y*-axis). LncRNA loci from existing GENCODE v20 annotation predicted to encode proteins are highlighted in yellow. LncRNA loci where new ORFs are discovered as a result of CLS transcript models are highlighted in red.

(D) *KANTR*, an example of an annotated lncRNA locus where CLS discovers novel protein-coding sequence. The upper panel shows the structure of the lncRNA and the associated ORF (highlighted region) falling within two novel full-length CLS transcripts (red). Note how this ORF lies outside existing GENCODE annotation (green), and its overlap with a highly-conserved region (see green PhyloCSF conservation track, below). The lower panel, generated by *CodAlignView* (53), reveals conservative substitutions in the predicted ORF of 76 aa consistent with a functional peptide product. High-confidence predicted SMART (54) domains are shown as coloured bars below. The entire ORF lies within and antisense to a L1 transposable element (grey bar).

Author contributions

RJ, RG, JH, AF, BU-R and JL designed the experiment. SC generated cDNA libraries and performed the Capture. CD and TRG performed the PacBio sequencing of Capture libraries. JL and BU-R analysed the data under the supervision of RG and RJ. RJ wrote the manuscript, with contributions from JL, BU-R and RG.

Competing financial interests

The author declare no competing financial interest.

Acknowledgements

We thank members of the Guigó laboratory for their valuable input and help when handling samples, analysing data and writing the manuscript, including Emilio Palumbo, Ferran Reverter, Amaya Abad, Dmitri Pervouchine, Carme Arnan and Francisco Camara. We wish to thank Lluis Armengol (qGenomics) for advice on RNA capture, Diego Garrido (CRG) for help with eQTL analysis, Sarah Bonnin (CRG) for help with data manipulation in R, Irwin Jungreis (MIT) for advice on PhyloCSF. James Wright and Jyoti Choudhary (Sanger Institute) helped in searching for peptide hits to putative coding regions. This work and publication were supported by the National Human Genome Research Institute of the National Institutes of Health (grant numbers U41HG007234, U41HG007000 and U54HG007004) and the Wellcome Trust (grant number WT098051). RJ was supported by Ramón y Cajal RYC-2011-08851. Work in laboratory of RG was supported by Awards Number U54HG0070, R01MH101814 and U41HG007234 from the National Human Genome Research Institute. This research was partly supported by the NCCR RNA & Disease funded by the Swiss National Science Foundation (to RJ). We thank Romina Garrido (CRG) for administrative support. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208.

Data availability

Raw and processed data is deposited in the Gene Expression Omnibus under accession GSE93848. Genome-aligned data were assembled into a public Track Hub, which can be loaded into the UCSC Genome Browser (pre-loaded URL: http://genome-euro.ucsc.edu/cgi-bin/hgTracks?hubUrl=http://public_docs.crg.es/rguigo/CLS/data/trackHub//hub.txt). In addition, a supplementary data portal is available on the web at https://public_docs.crg.es/rguigo/CLS/ .

References

1. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559-1563.

2. Jia, H., Osak, M., Bogu, G.K., Stanton, L.W., Johnson, R. and Lipovich, L. (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA*, **16**, 1478-1487.
3. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223-227.
4. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, **7**, 562-578.
5. Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, **28**, 503-510.
6. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, **25**, 1915-1927.
7. Hangauer, M.J., Vaughn, I.W. and McManus, M.T. (2013) Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS genetics*, **9**, e1003569.
8. Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics*.
9. Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic acids research*, **44**, D203-208.
10. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, **28**, 511-515.
11. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, **22**, 1760-1774.
12. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, **22**, 1775-1789.
13. Mele, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660-665.
14. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, **45**, 1113-1120.
15. Consortium, E. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
16. Chen, L., Kostadima, M., Martens, J.H., Canu, G., Garcia, S.P., Turro, E., Downes, K., Macaulay, I.C., Bielczyk-Maczynska, E., Coe, S. *et al.* (2014) Transcriptional diversity during lineage commitment of human blood progenitors. *Science*, **345**, 1251033.

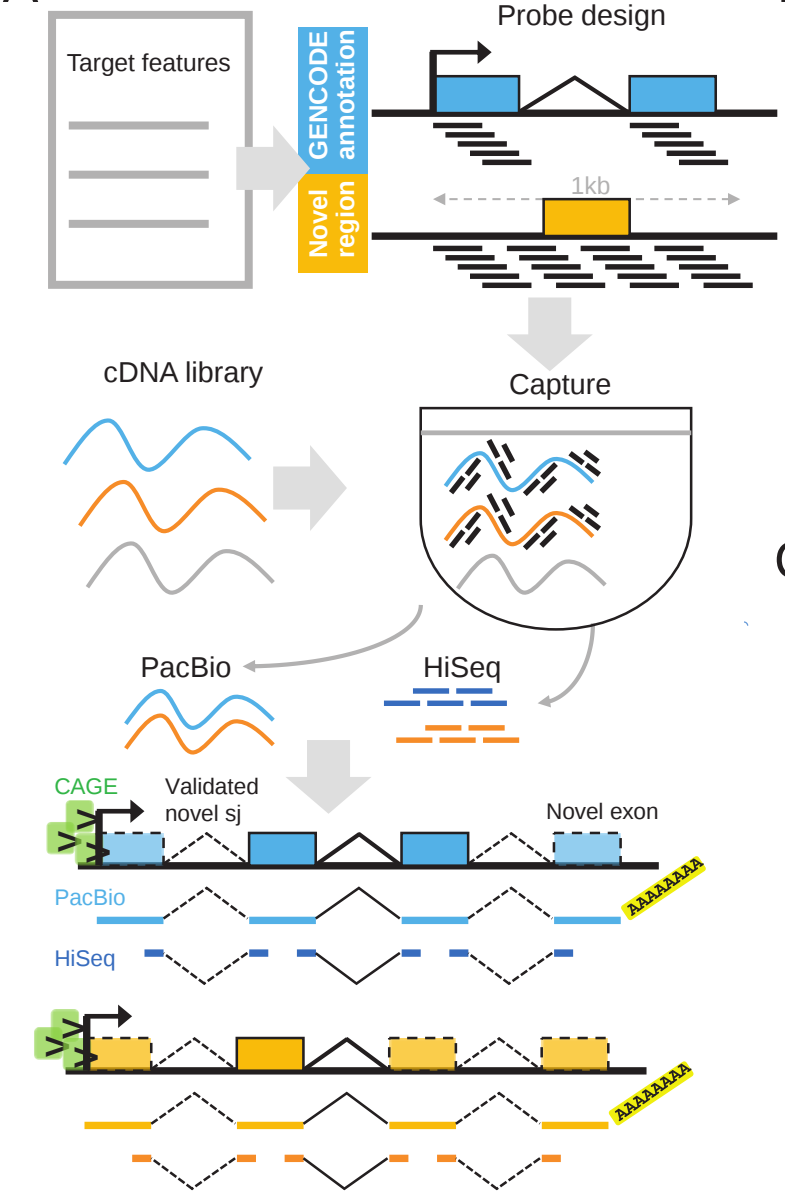
17. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317-330.
18. Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462-470.
19. Steijger, T., Abril, J.F., Engstrom, P.G., Kokocinski, F., Hubbard, T.J., Guigo, R., Harrow, J. and Bertone, P. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nature methods*, **10**, 1177-1184.
20. Georgakilas, G., Vlachos, I.S., Paraskevopoulou, M.D., Yang, P., Zhang, Y., Economides, A.N. and Hatzigeorgiou, A.G. (2014) microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nature communications*, **5**, 5700.
21. Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q. *et al.* (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**, 46-58.
22. Ferdin, J., Nishida, N., Wu, X., Nicoloso, M.S., Shah, M.Y., Devlin, C., Ling, H., Shimizu, M., Kumar, K., Cortez, M.A. *et al.* (2013) HINCUTs in cancer: hypoxia-induced noncoding ultraconserved transcripts. *Cell death and differentiation*, **20**, 1675-1687.
23. Calin, G.A., Liu, C.G., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E.J., Wojcik, S.E. *et al.* (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer cell*, **12**, 215-229.
24. Lagarde, J., Uszczynska-Ratajczak, B., Santoyo-Lopez, J., Gonzalez, J.M., Tapanari, E., Mudge, J.M., Steward, C.A., Wilming, L., Tanzer, A., Howald, C. *et al.* (2016) Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nature communications*, **7**, 12339.
25. Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S. and Rinn, J.L. (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature biotechnology*, **30**, 99-104.
26. Bussotti, G., Leonardi, T., Clark, M.B., Mercer, T.R., Crawford, J., Malquori, L., Notredame, C., Dinger, M.E., Mattick, J.S. and Enright, A.J. (2016) Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome research*, **26**, 705-716.
27. Clark, M.B., Mercer, T.R., Bussotti, G., Leonardi, T., Haynes, K.R., Crawford, J., Brunck, M.E., Cao, K.A., Thomas, G.P., Chen, W.Y. *et al.* (2015) Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nature methods*, **12**, 339-342.
28. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, **42**, D68-73.
29. Visel, A., Minovitsky, S., Dubchak, I. and Pennacchio, L.A. (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic acids research*, **35**, D88-92.
30. Dimitrieva, S. and Bucher, P. (2013) UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic acids research*, **41**, D101-109.
31. Bussotti, G., Raineri, E., Erb, I., Zytnicki, M., Wilm, A., Beaudoing, E., Bucher, P. and Notredame, C. (2011) BlastR--fast and accurate database searches for non-coding RNAs. *Nucleic acids research*, **39**, 6886-6895.
32. Kralj, J.G. and Salit, M.L. (2013) Characterization of in vitro transcription amplification linearity and variability in the low copy number regime using External RNA Control Consortium (ERCC) spike-ins. *Analytical and bioanalytical chemistry*, **405**, 315-320.

33. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101-108.
34. Sharon, D., Tilgner, H., Grubert, F. and Snyder, M. (2013) A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*, **31**, 1009-1014.
35. Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, **13**, 341.
36. Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E. and Mattick, J.S. (2014) Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nature protocols*, **9**, 989-1009.
37. Leucci, E., Vendramin, R., Spinazzi, M., Laurette, P., Fiers, M., Wouters, J., Radaelli, E., Eyckerman, S., Leonelli, C., Vanderheyden, K. *et al.* (2016) Melanoma addiction to the long non-coding RNA SAMMSON. *Nature*, **531**, 518-522.
38. Blanco, E., Parra, G. and Guigo, R. (2007) Using geneid to identify genes. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, **Chapter 4**, Unit 4 3.
39. Ounzain, S., Micheletti, R., Arnan, C., Plaisance, I., Cecchi, D., Schroen, B., Reverter, F., Alexanian, M., Gonzales, C., Ng, S.Y. *et al.* (2015) CARMEN, a human super enhancer-associated long noncoding RNA controlling cardiac specification, differentiation and homeostasis. *Journal of molecular and cellular cardiology*, **89**, 98-112.
40. Nissan, A., Stojadinovic, A., Mitrani-Rosenbaum, S., Halle, D., Grinbaum, R., Roistacher, M., Bochem, A., Dayanc, B.E., Ritter, G., Gomceli, I. *et al.* (2012) Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. *International journal of cancer*, **130**, 1598-1606.
41. Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R. and Ponting, C.P. (2013) Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome biology*, **14**, R131.
42. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, **42**, D1001-1006.
43. Arking, D.E., Pulit, S.L., Crotti, L., van der Harst, P., Munroe, P.B., Koopmann, T.T., Sotoodehnia, N., Rossin, E.J., Morley, M., Wang, X. *et al.* (2014) Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nature genetics*, **46**, 826-836.
44. Alam, T., Medvedeva, Y.A., Jia, H., Brown, J.B., Lipovich, L. and Bajic, V.B. (2014) Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PloS one*, **9**, e109443.
45. Mele, M., Mattioli, K., Mallard, W., Shechner, D.M., Gerhardinger, C. and Rinn, J.L. (2016) Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome research*.
46. Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M. *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome biology*, **16**, 179.
47. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal*, **33**, 981-993.

48. Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P. and Li, W. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research*, **41**, e74.
49. Lin, M.F., Jungreis, I. and Kellis, M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275-282.
50. Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M. *et al.* (2013) Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife*, **2**, e01749.
51. Wan, X., Huang, W., Yang, S., Zhang, Y., Pu, H., Fu, F., Huang, Y., Wu, H., Li, T. and Li, Y. (2016) Identification of androgen-responsive lncRNAs as diagnostic and prognostic markers for prostate cancer. *Oncotarget*.
52. Ruiz-Orera, J., Messeguer, X., Subirana, J.A. and Alba, M.M. (2014) Long non-coding RNAs as a source of new peptides. *eLife*, **3**, e03523.
53. I Jungreis, M Lin and Kellis, M. CodAlignView: a tool for visualizing protein-coding constraint. *In Preparation*.
54. Letunic, I., Doerks, T. and Bork, P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic acids research*, **43**, D257-260.

Figure 1

A



B

Feature	Human	Mouse
lincRNAs	5,953	1,920
PipeR pred.	0	2,271
MicroRNAs	786	494
snRNAs	838	721
snoRNAs	401	850
Enhancers	954	203
UCE	158	156
TOTAL	**9,090**	**6,615**
Prot. coding	124	112
ERCC	42	42
Ecoli	100	100
TOTAL nts	**15,541,153**	**8,307,739**

Annotated

Unannotated

Controls

C

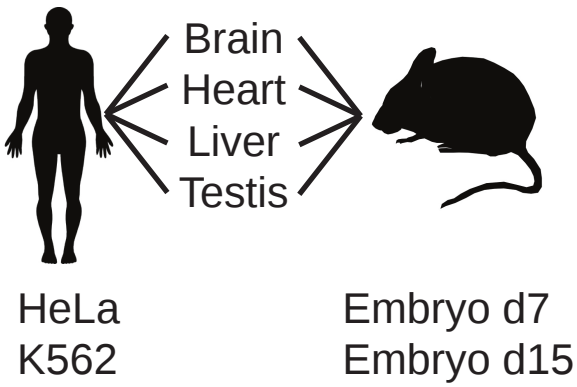
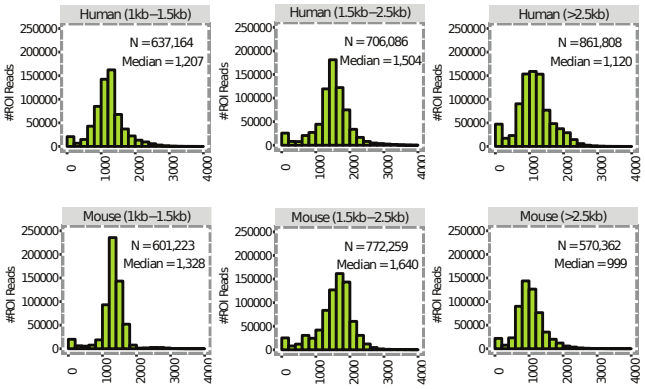


Figure 2

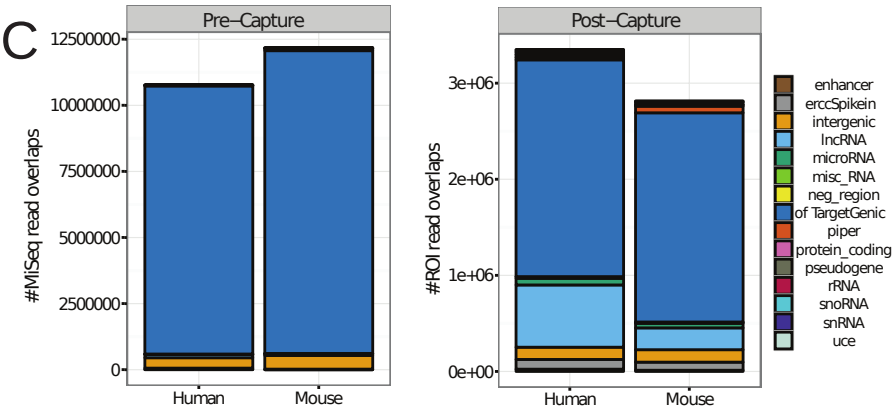
A

Human		Mouse
64	SMRT Cells	66
2,397,192	ROI Reads	2,137,176
2,383,470	Mapped	2,132,978
2,053,424	Uniquely mapped	1,870,681
1,627,322	Demultiplexed	1,509,374

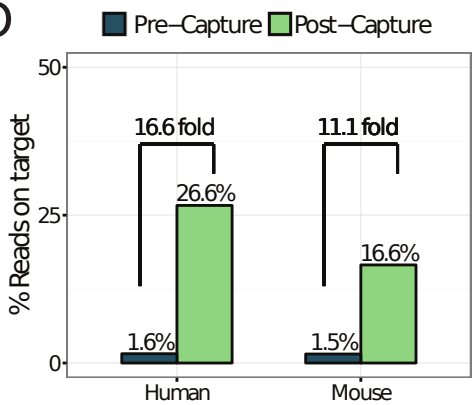
B



C



D



E

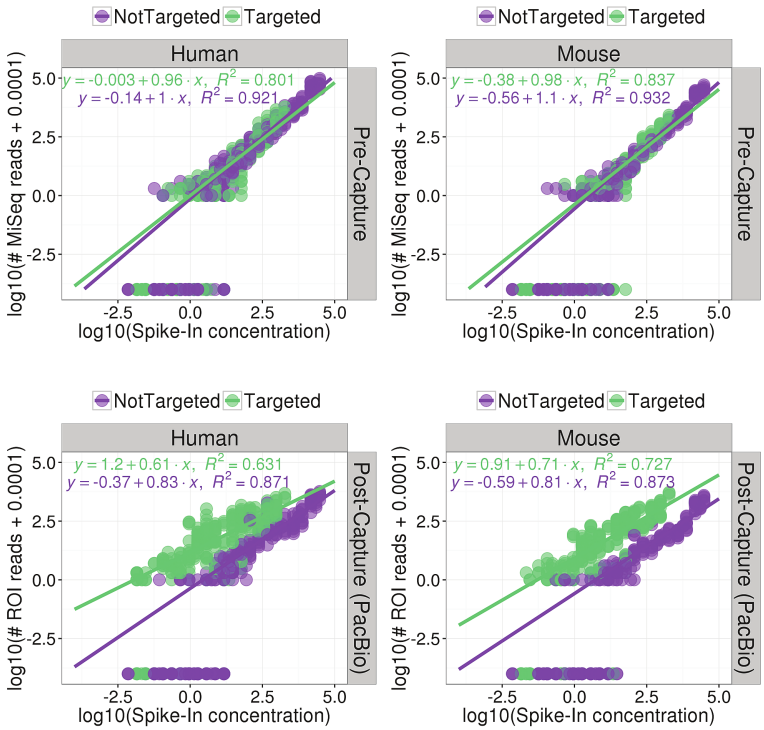


Figure 3

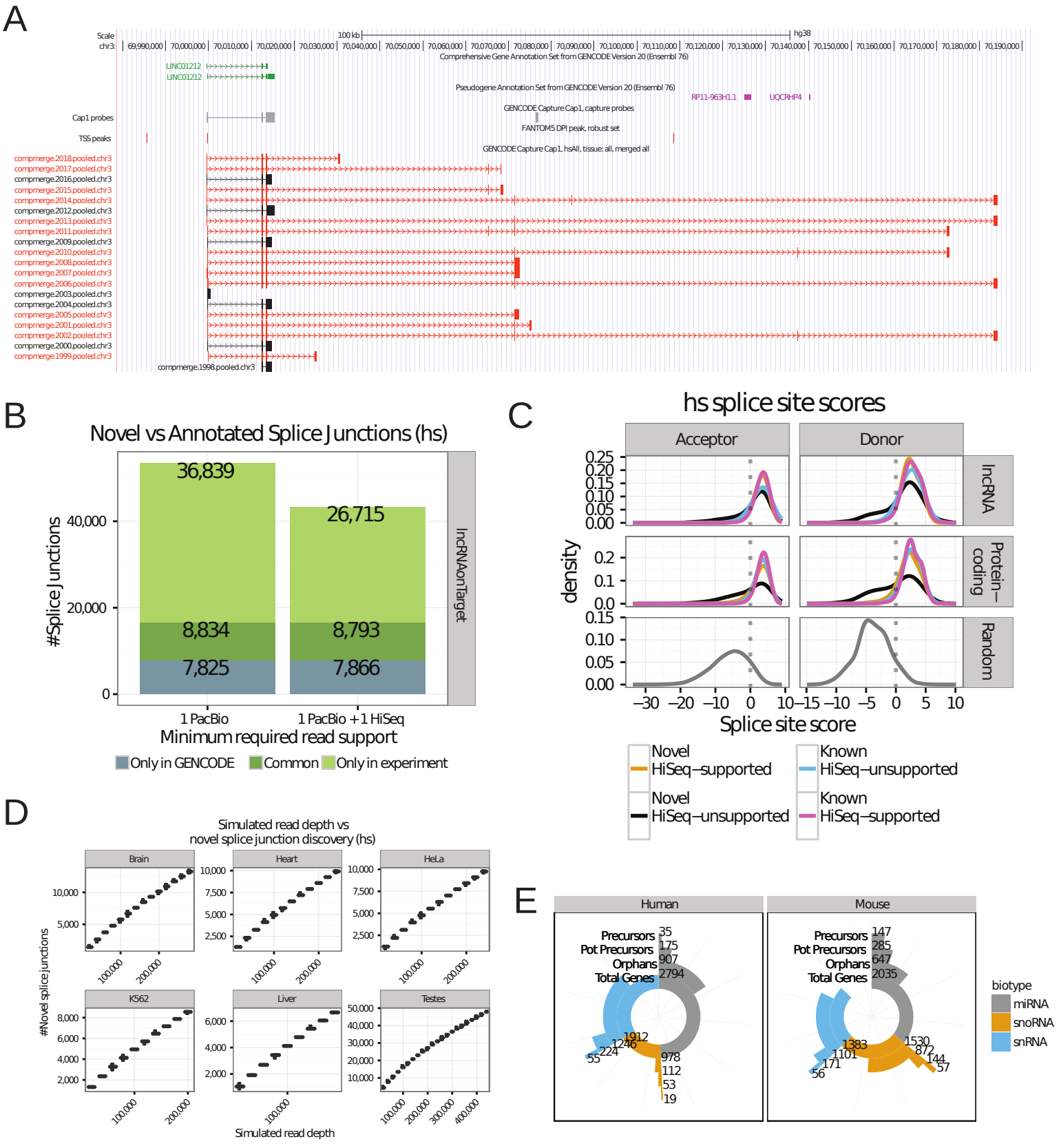

A
Scale
chr3:
100 kb
hg38
Comprehensive Gene Annotation Set from GENCODE Version 20 (Ensembl 76)
LINC01212
LINC01212
Pseudogene Annotation Set from GENCODE Version 20 (Ensembl 76)
RP11-963H1.1
UQCRHP4
GENCODE Capture Cap1, capture probes
Cap1 probes
FANTOM5 DPI peak, robust set
TSS peaks
GENCODE Capture Cap1, hsAll, tissue: all, merged all
compmerge.2018.pooled.chr3
compmerge.2017.pooled.chr3
compmerge.2016.pooled.chr3
compmerge.2015.pooled.chr3
compmerge.2014.pooled.chr3
compmerge.2012.pooled.chr3
compmerge.2013.pooled.chr3
compmerge.2011.pooled.chr3
compmerge.2009.pooled.chr3
compmerge.2010.pooled.chr3
compmerge.2008.pooled.chr3
compmerge.2007.pooled.chr3
compmerge.2006.pooled.chr3
compmerge.2003.pooled.chr3
compmerge.2004.pooled.chr3
compmerge.2005.pooled.chr3
compmerge.2001.pooled.chr3
compmerge.2002.pooled.chr3
compmerge.2000.pooled.chr3
compmerge.1999.pooled.chr3
compmerge.1998.pooled.chr3
B
Novel vs Annotated Splice Junctions (hs)
#Splice Junctions
36,839
26,715
8,834
8,793
7,825
7,866
lncRNAonTarget
1 PacBio
1 PacBio +1 HiSeq
Minimum required read support
Only in GENCODE
Common
Only in experiment
C
hs splice site scores
Acceptor
Donor
lncRNA
Protein–coding
Random
density
Splice site score
Novel HiSeq–supported
Known HiSeq–unsupported
Novel HiSeq–unsupported
Known HiSeq–supported
D
Simulated read depth vs novel splice junction discovery (hs)
Brain
Heart
HeLa
K562
Liver
Testes
#Novel splice junctions
Simulated read depth
E
Human
Mouse
Precursors
Pot Precursors
Orphans
Total Genes
biotype
miRNA
snoRNA
snRNA

Figure 4

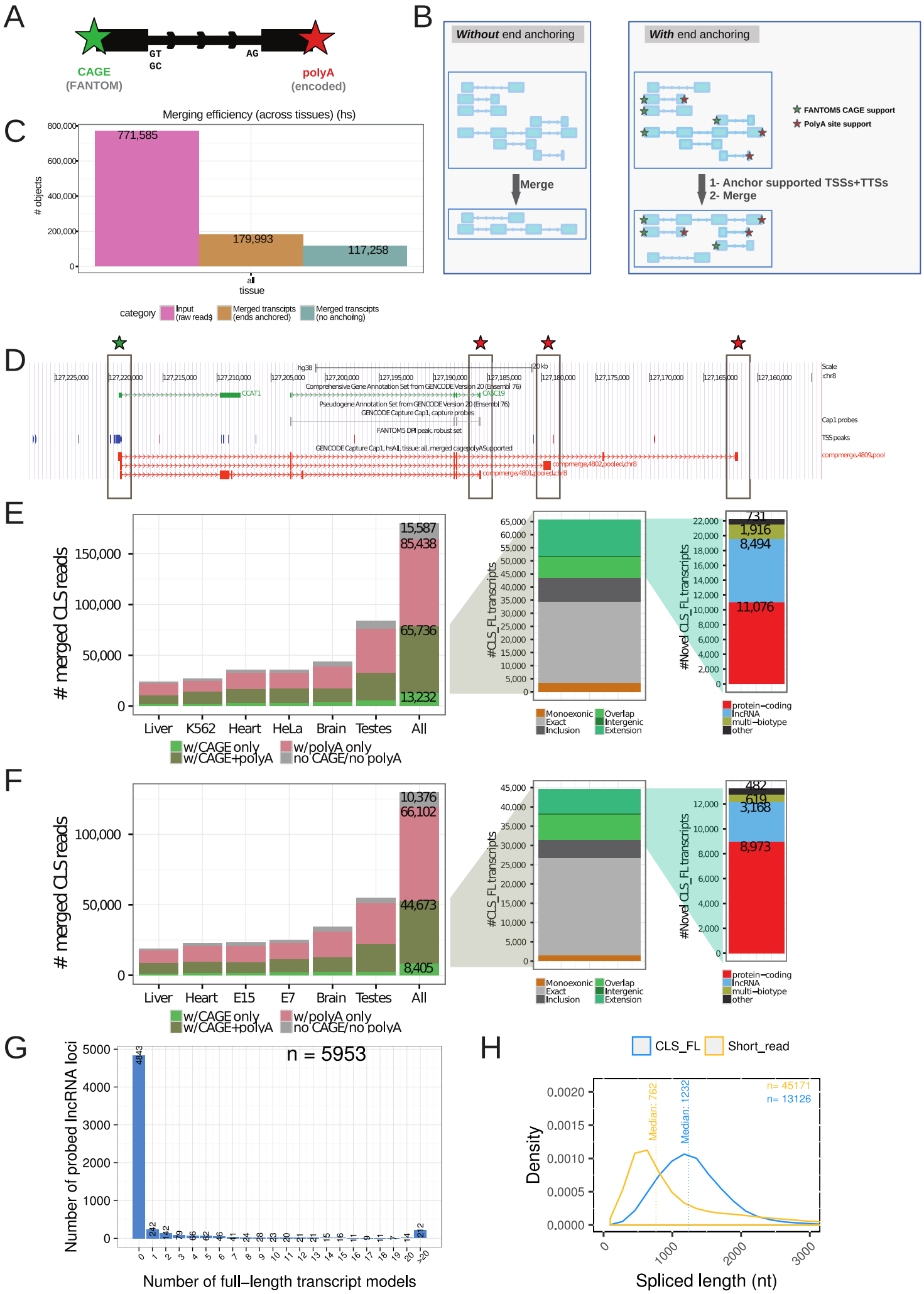

A
GT
GC
AG
CAGE
(FANTOM)
polyA
(encoded)
B
Without end anchoring
Merge
With end anchoring
FANTOM5 CAGE support
PolyA site support
1- Anchor supported TSSs+TTSs
2- Merge
C
Merging efficiency (across tissues) (hs)
objects
771,585
179,993
117,258
all
tissue
category
Input (raw reads)
Merged transcipts (ends anchored)
Merged transcipts (no anchoring)
D
hg38
Comprehensive Gene Annotation Set from GENCODE Version 20 (Ensembl 76)
Pseudogene Annotation Set from GENCODE Version 20 (Ensembl 76)
GENCODE Capture Cap1, capture probes
FANTOM5 DPI peak, robust set
GENCODE Capture Cap1, hsAll, tissue: all, merged cagepolyASupported
Cap1 probes
TSS peaks
E
merged CLS reads
15,587
85,438
65,736
13,232
Liver
K562
Heart
HeLa
Brain
Testes
All
w/CAGE only
w/CAGE+polyA
w/polyA only
no CAGE/no polyA
#CLS_FL transcripts
Monoexonic
Exact
Inclusion
Overlap
Intergenic
Extension
#Novel CLS_FL transcripts
731
1,916
8,494
11,076
protein−coding
lncRNA
multi−biotype
other
F
merged CLS reads
10,376
66,102
44,673
8,405
Liver
Heart
E15
E7
Brain
Testes
All
w/CAGE only
w/CAGE+polyA
w/polyA only
no CAGE/no polyA
#CLS_FL transcripts
Monoexonic
Exact
Inclusion
Overlap
Intergenic
Extension
#Novel CLS_FL transcripts
482
619
3,168
8,973
protein−coding
lncRNA
multi−biotype
other
G
n = 5953
Number of probed lncRNA loci
Number of full−length transcript models
H
CLS_FL
Short_read
n= 45171
n= 13126
Median: 762
Median: 1232
Density
Spliced length (nt)

Figure 5

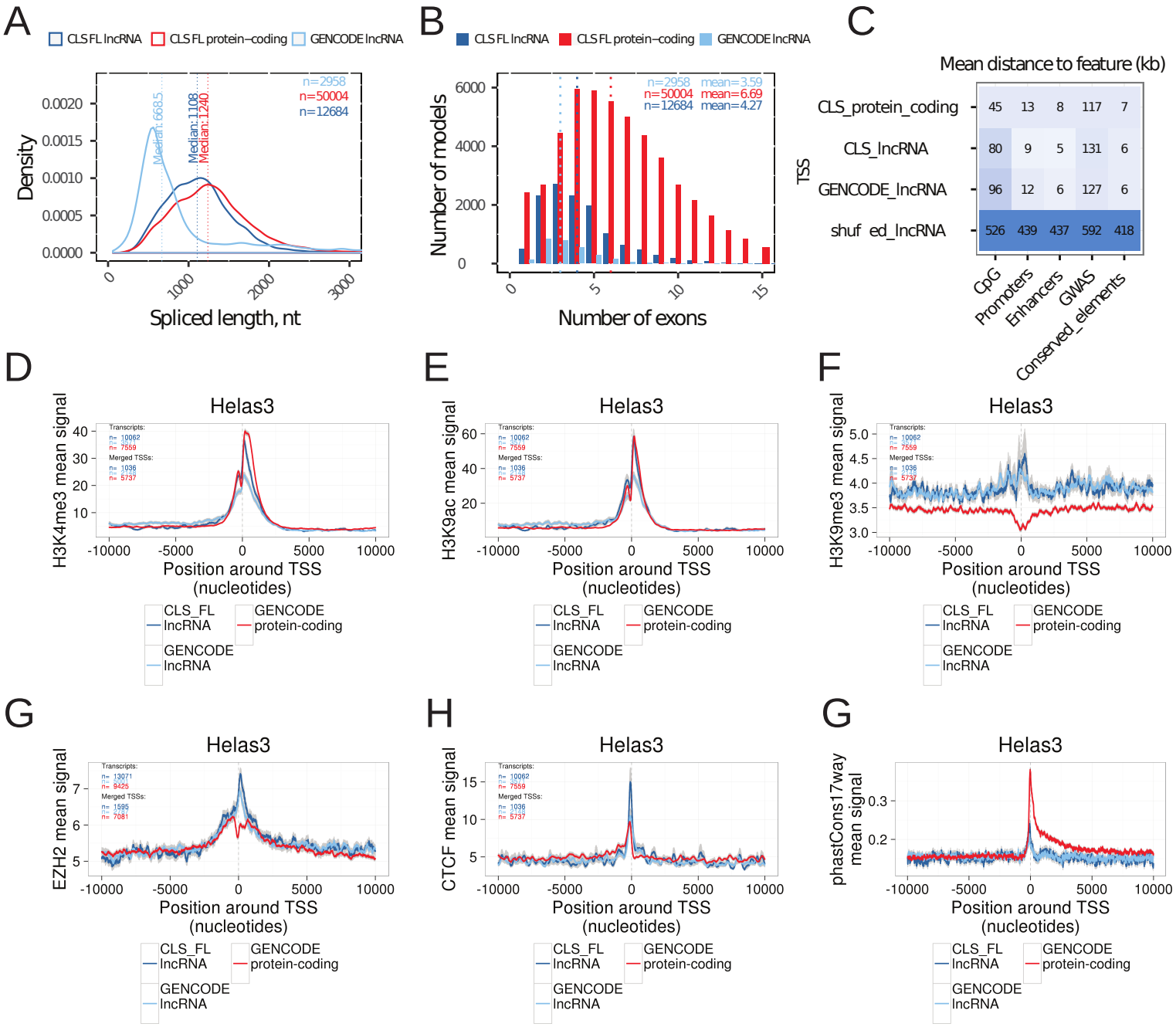
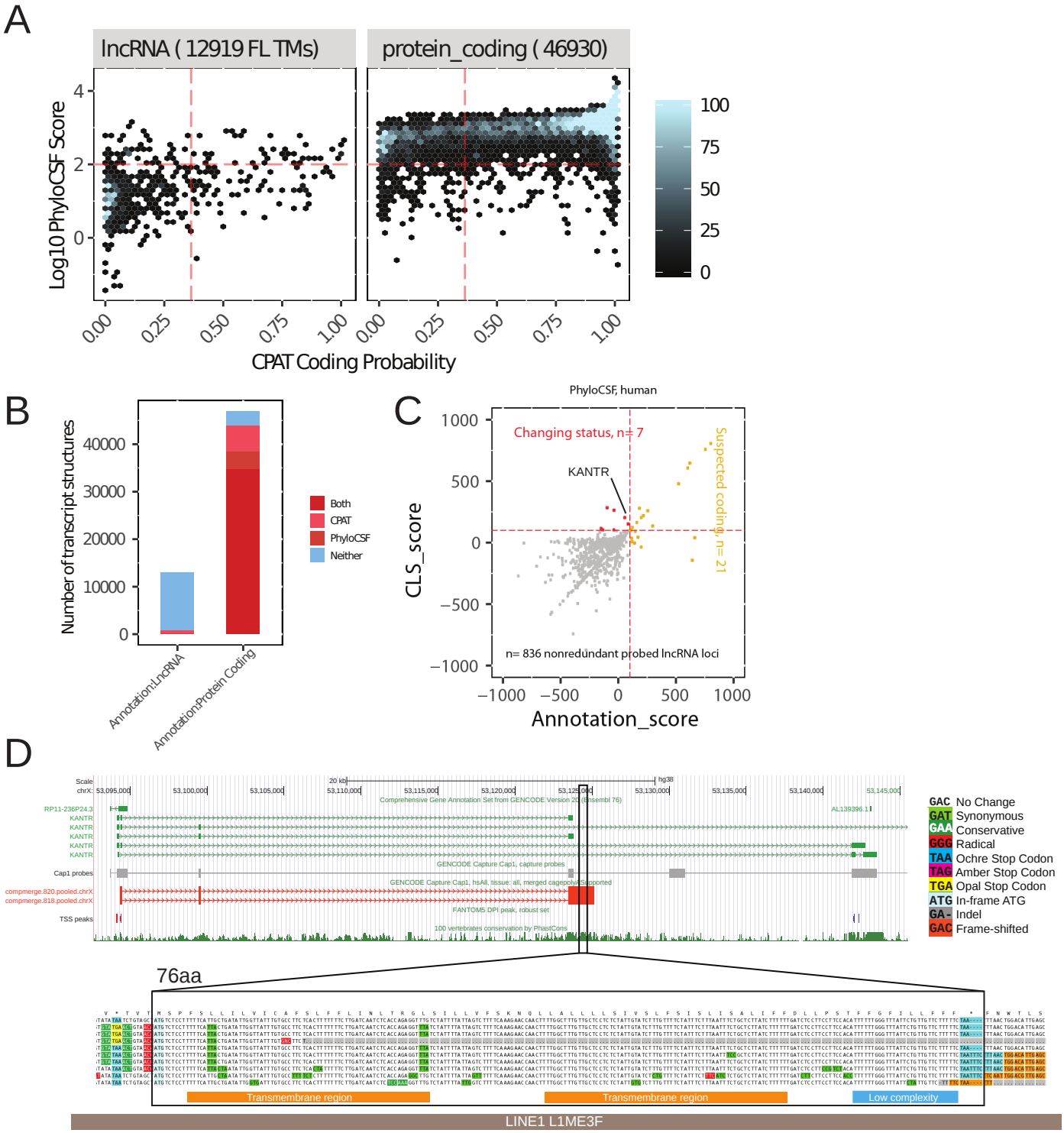
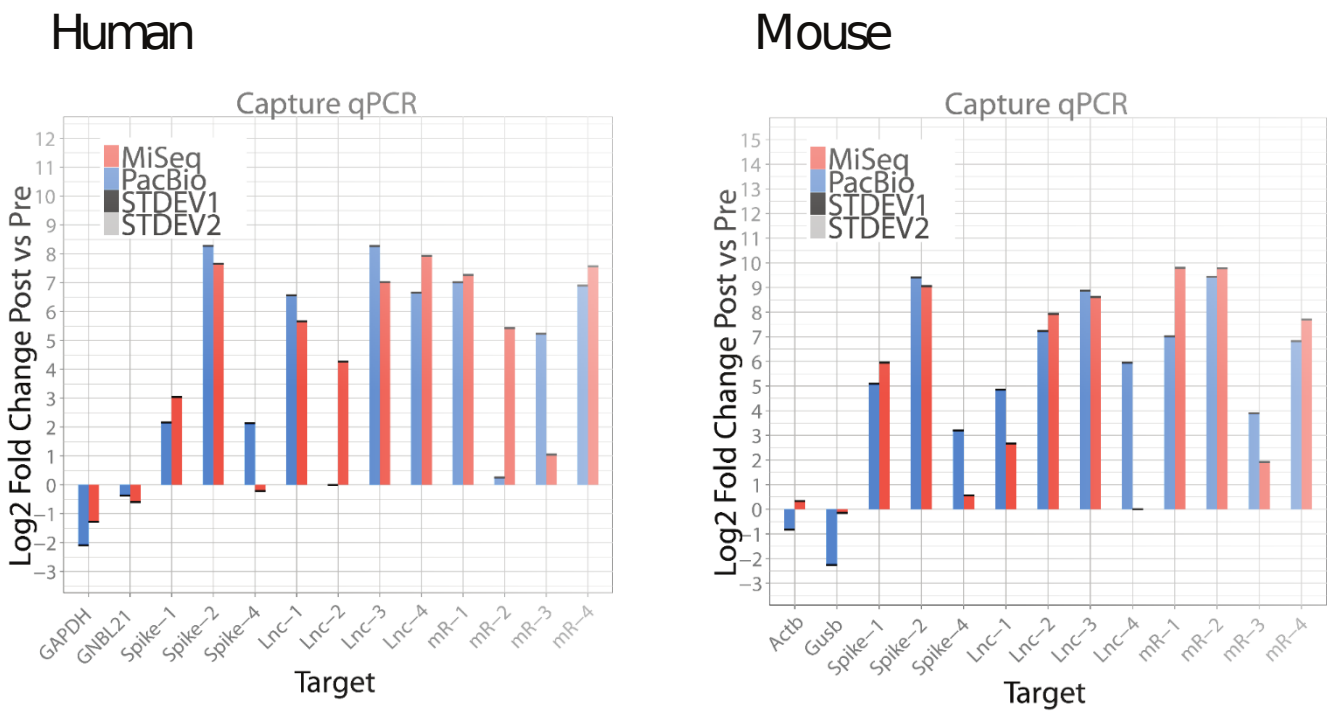

A
CLS FL lncRNA
CLS FL protein−coding
GENCODE lncRNA
Density
Spliced length, nt
n=2958
n=50004
n=12684
Median: 668.5
Median: 1108
Median: 1240
B
Number of models
Number of exons
n=2958 mean=3.59
n=50004 mean=6.69
n=12684 mean=4.27
C
Mean distance to feature (kb)
TSS
CLS_protein_coding 45 13 8 117 7
CLS_lncRNA 80 9 5 131 6
GENCODE_lncRNA 96 12 6 127 6
shuf ed_lncRNA 526 439 437 592 418
CpG
Promoters
Enhancers
GWAS
Conserved_elements
D
E
F
G
H
G
Helas3
H3K4me3 mean signal
H3K9ac mean signal
H3K9me3 mean signal
EZH2 mean signal
CTCF mean signal
phastCons17way mean signal
Position around TSS (nucleotides)
CLS_FL lncRNA
GENCODE protein-coding
GENCODE lncRNA

Figure 6


A
lncRNA (12919 FL TMs)
protein_coding (46930)
Log10 PhyloCSF Score
CPAT Coding Probability
B
Number of transcript structures
Annotation:LncRNA
Annotation:Protein Coding
Both
CPAT
PhyloCSF
Neither
C
PhyloCSF, human
Changing status, n= 7
KANTR
Suspected coding, n= 21
n= 836 nonredundant probed lncRNA loci
CLS_score
Annotation_score
D
Comprehensive Gene Annotation Set from GENCODE Version 20 (Ensembl 76)
GENCODE Capture Cap1, capture probes
FANTOM5 DPI peak, robust set
100 vertebrates conservation by PhastCons
76aa
Transmembrane region
Transmembrane region
Low complexity
LINE1 L1ME3F
GAC No Change
GAT Synonymous
GAA Conservative
GGG Radical
TAA Ochre Stop Codon
TAG Amber Stop Codon
TGA Opal Stop Codon
ATG In-frame ATG
GA- Indel
GAC Frame-shifted

Supplementary Figure S1

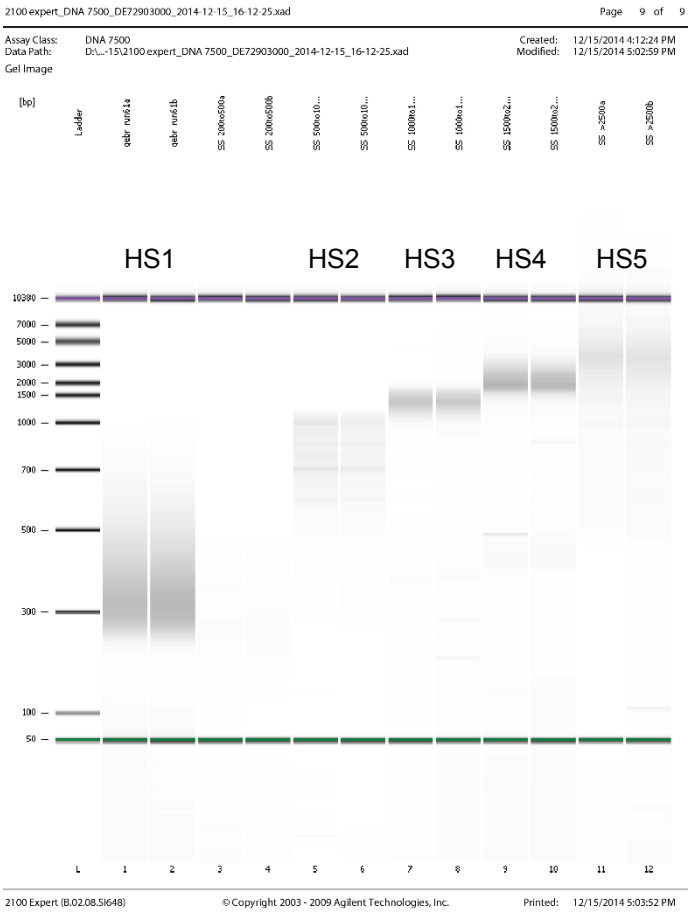


Supplementary Figure S2

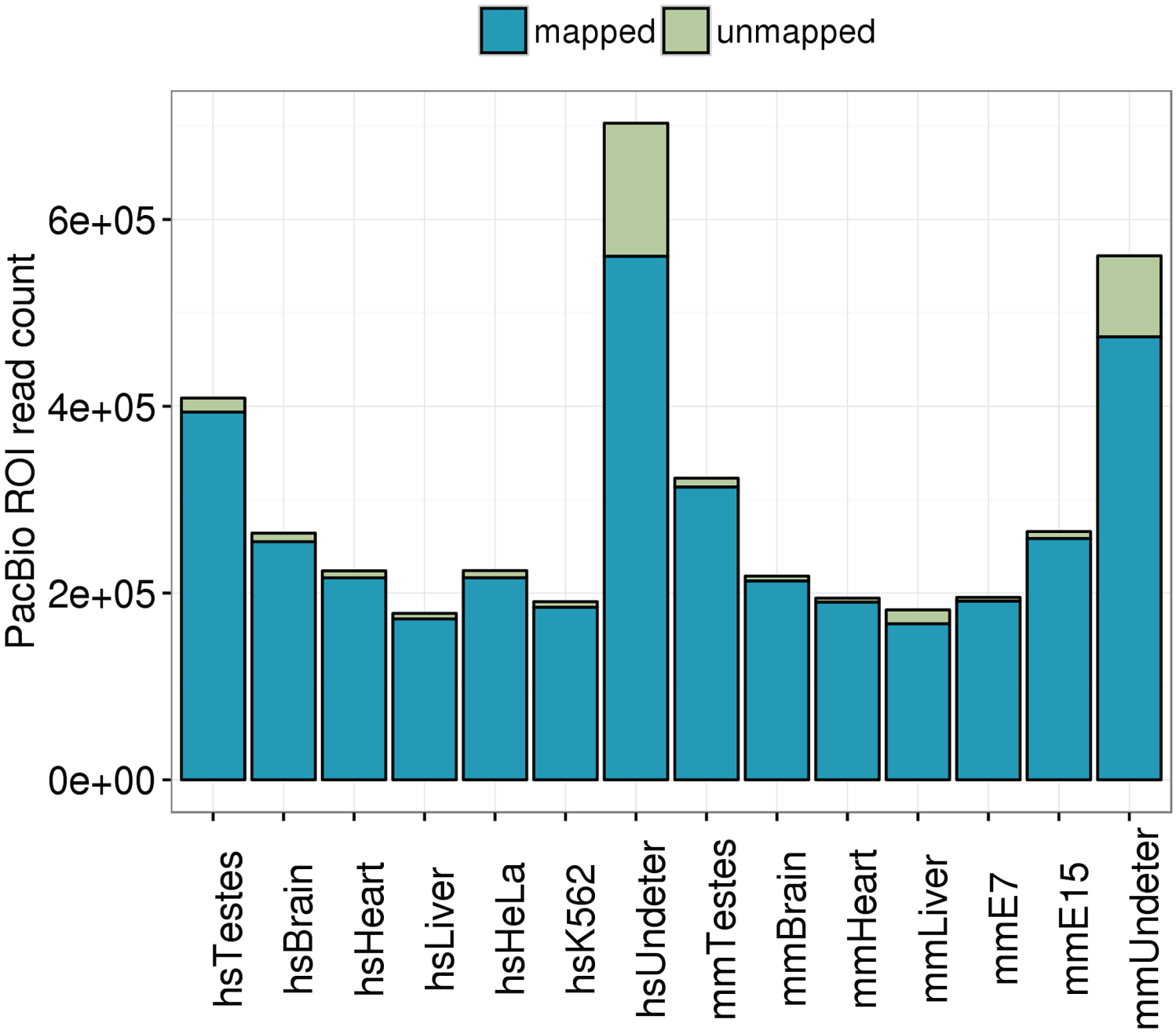
A

	Range (nt)	Human
1	200-500	HS1
2	500-1000	HS2
3	1000-1500	HS3
4	1500-2500	HS4
5	>2500	HS5

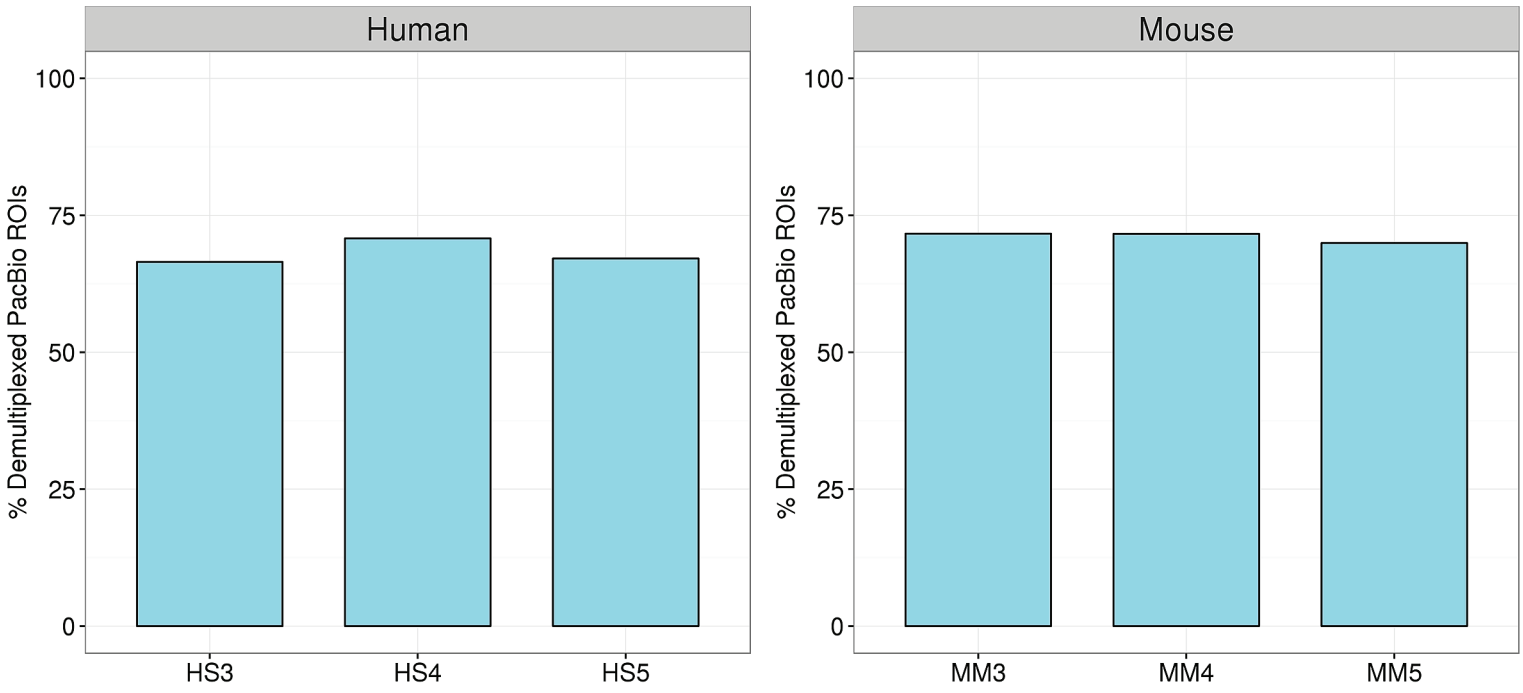
B



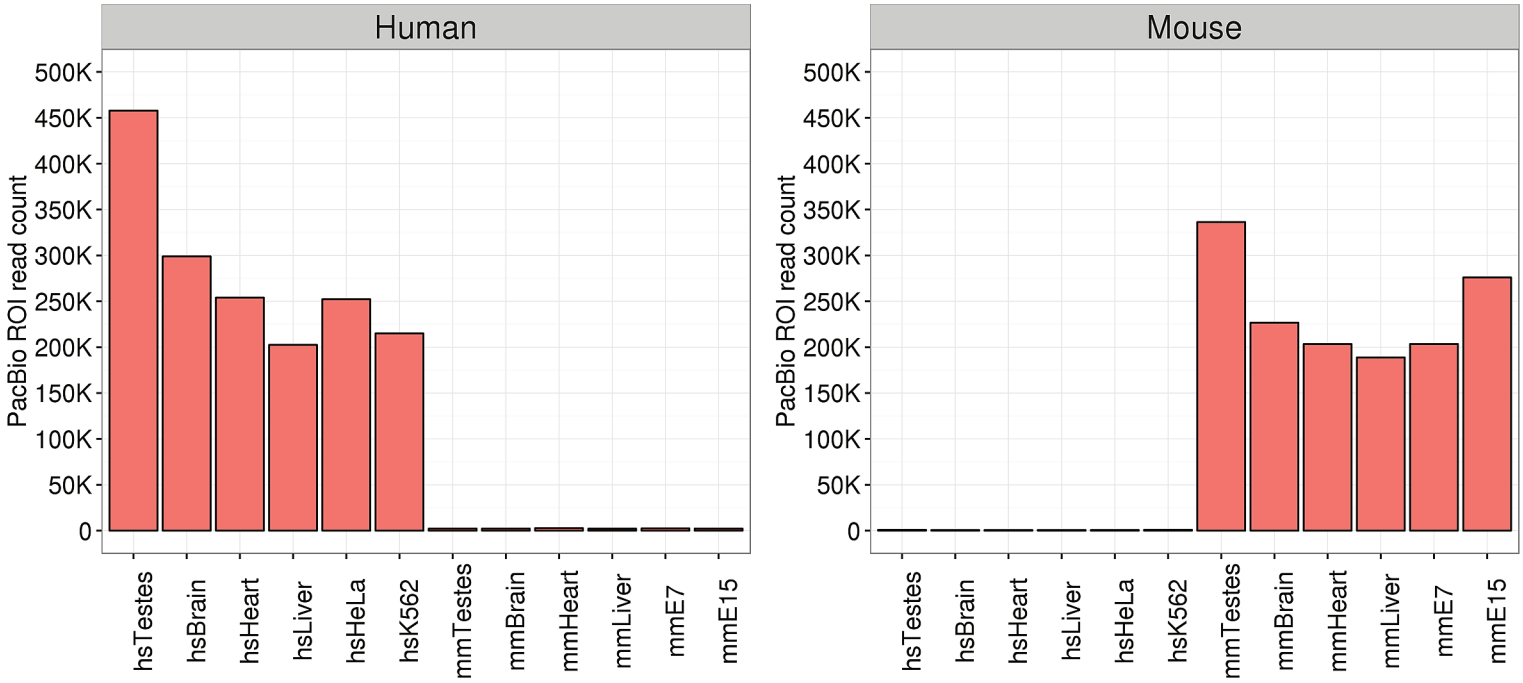
Supplementary Figure S3



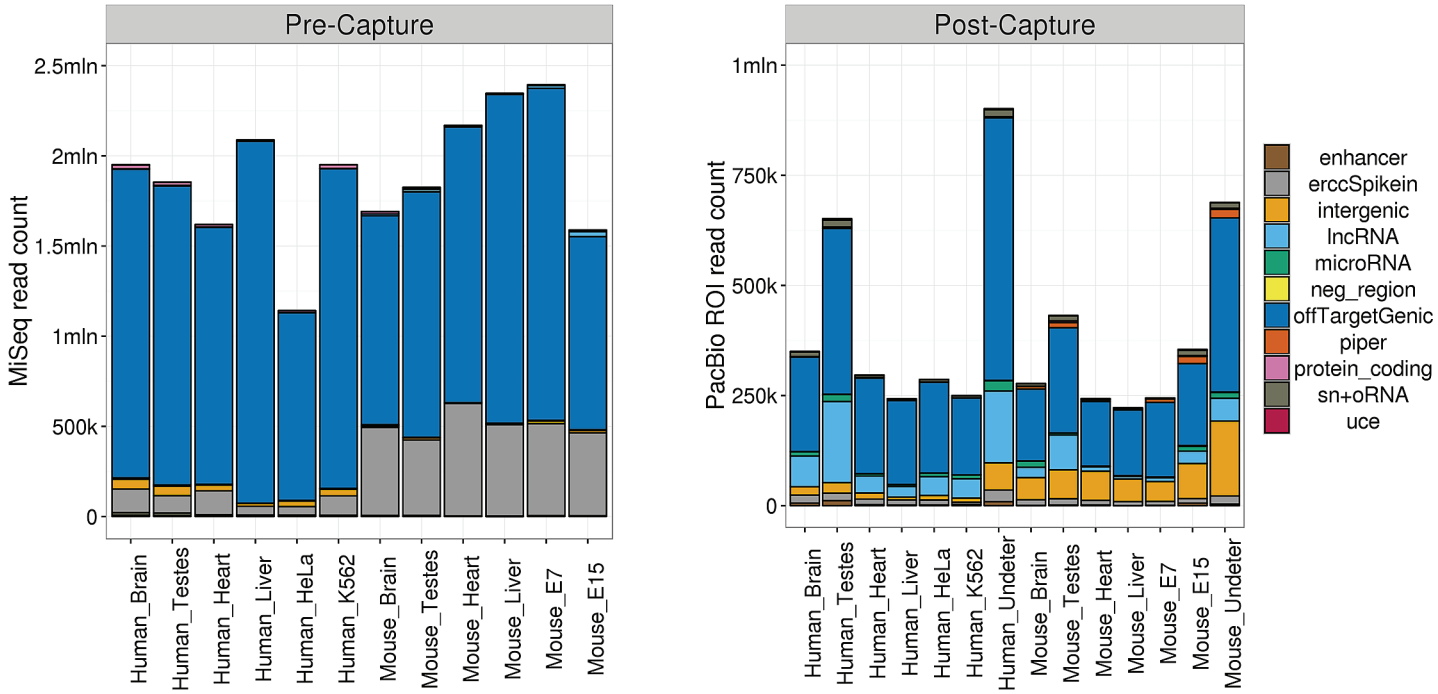
Supplementary Figure S4



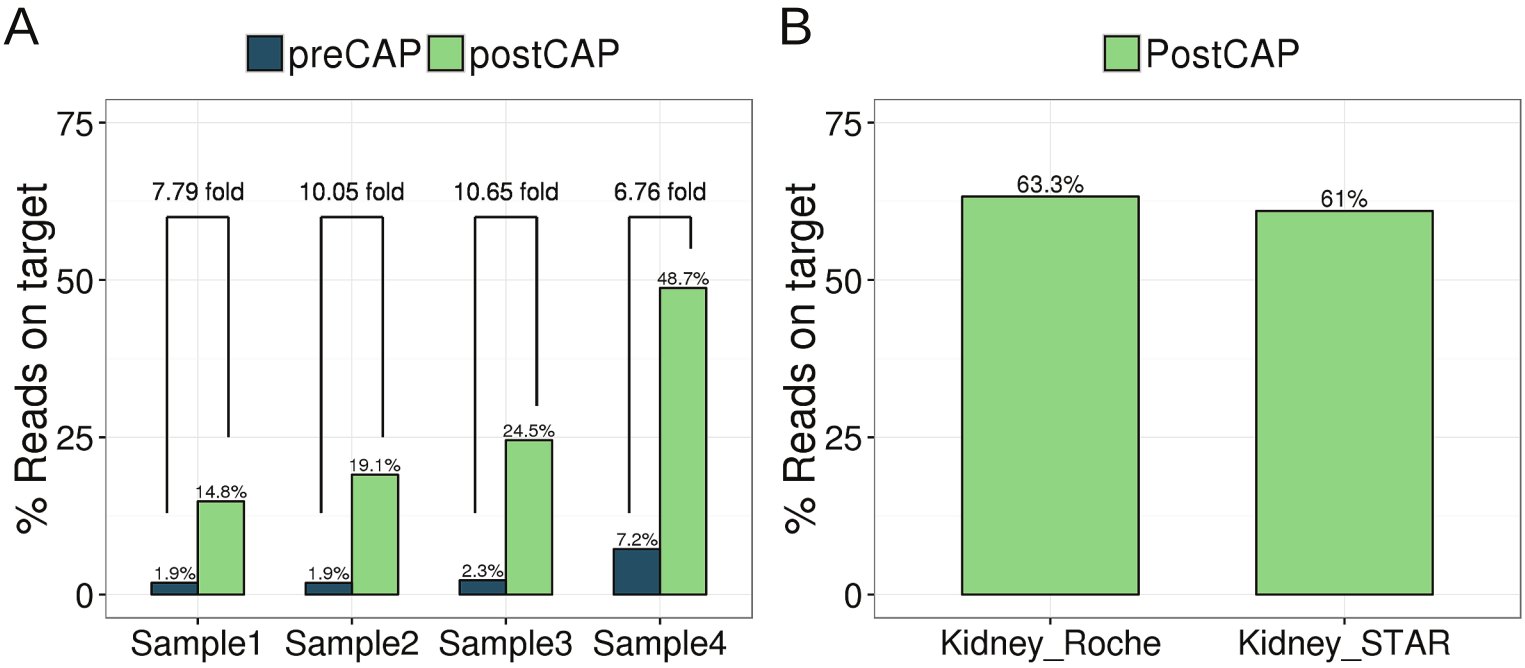
Supplementary Figure S5



Supplementary Figure S6

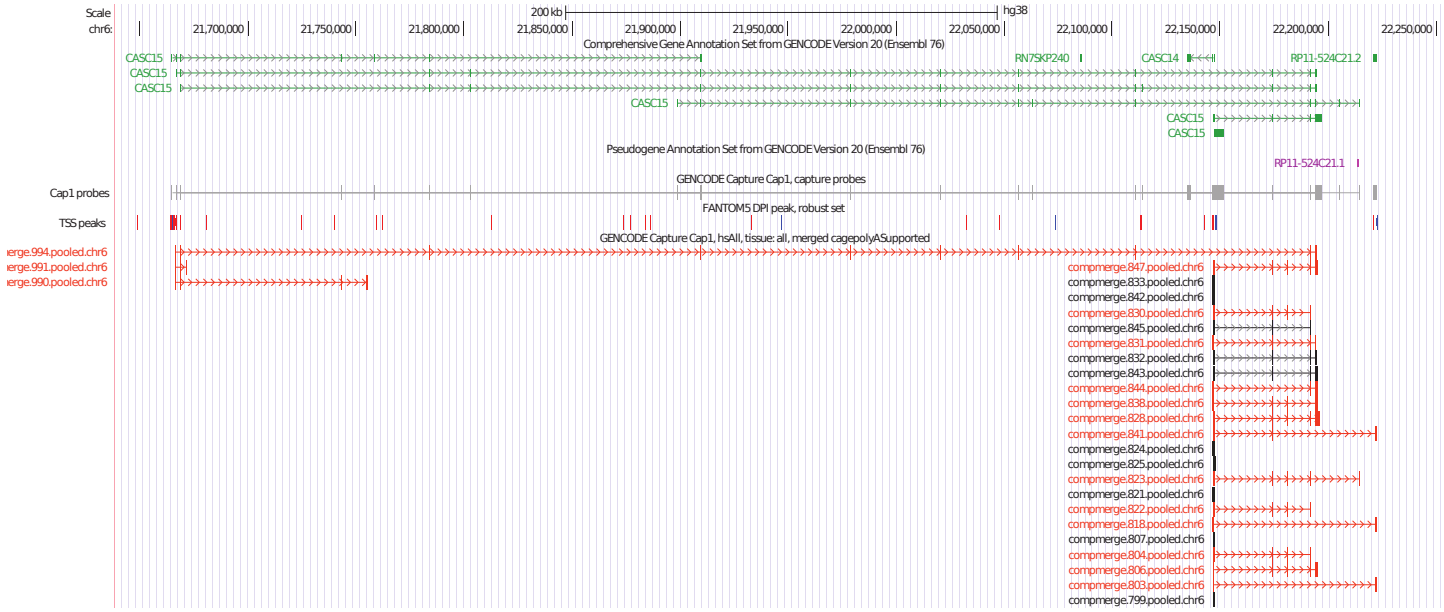


Supplementary Figure S7

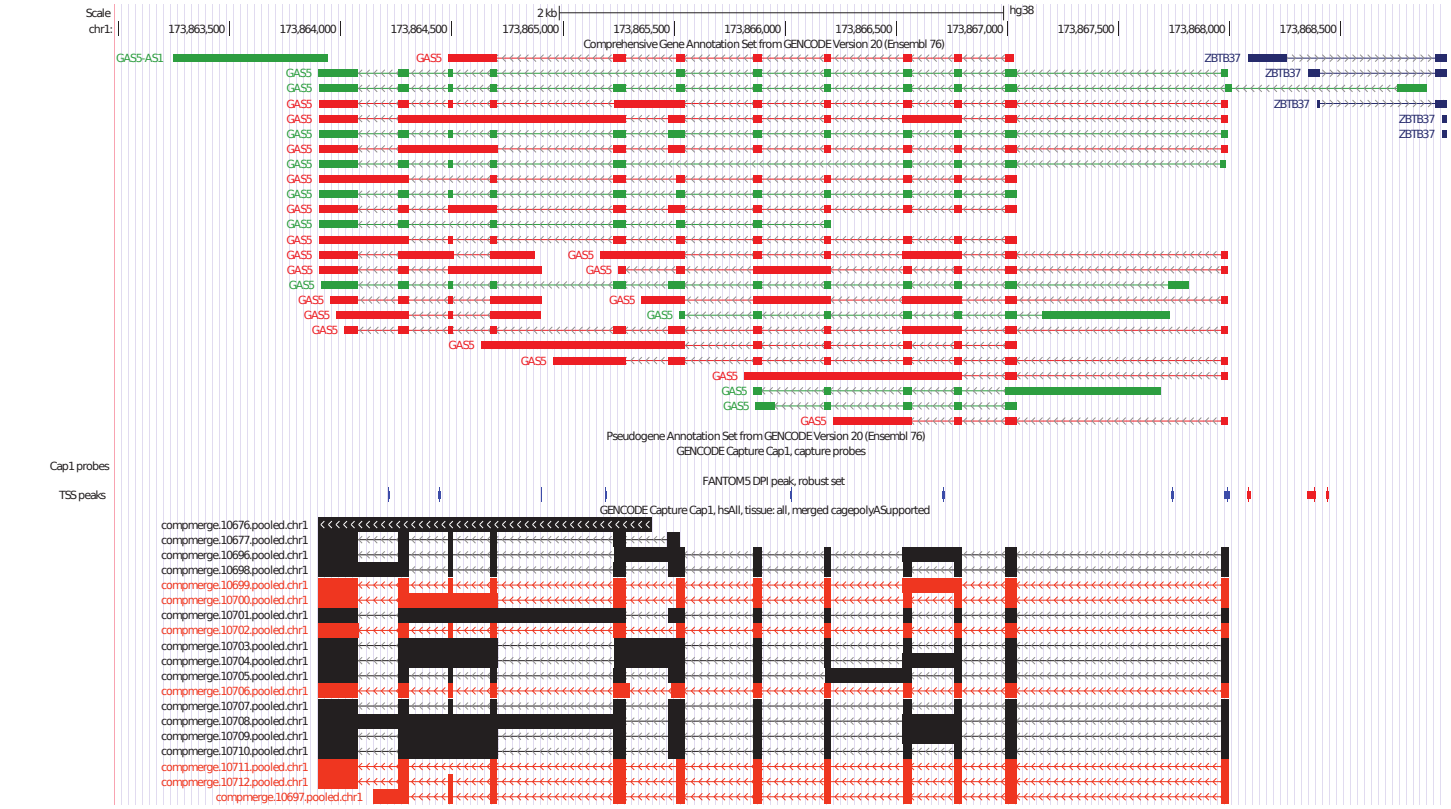


Supplementary Figure S8

CASC15 / ENSG00000272168

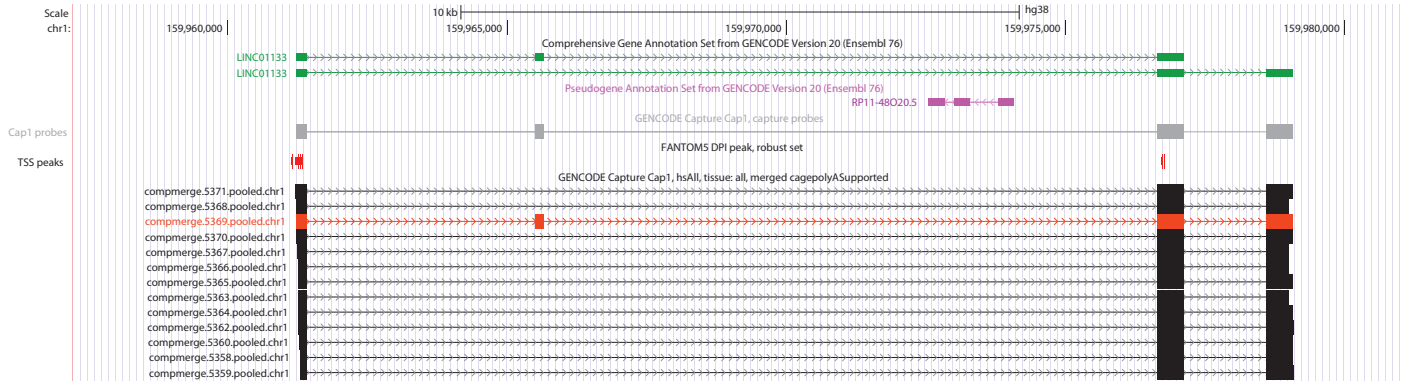


GAS5 / ENSG00000234741



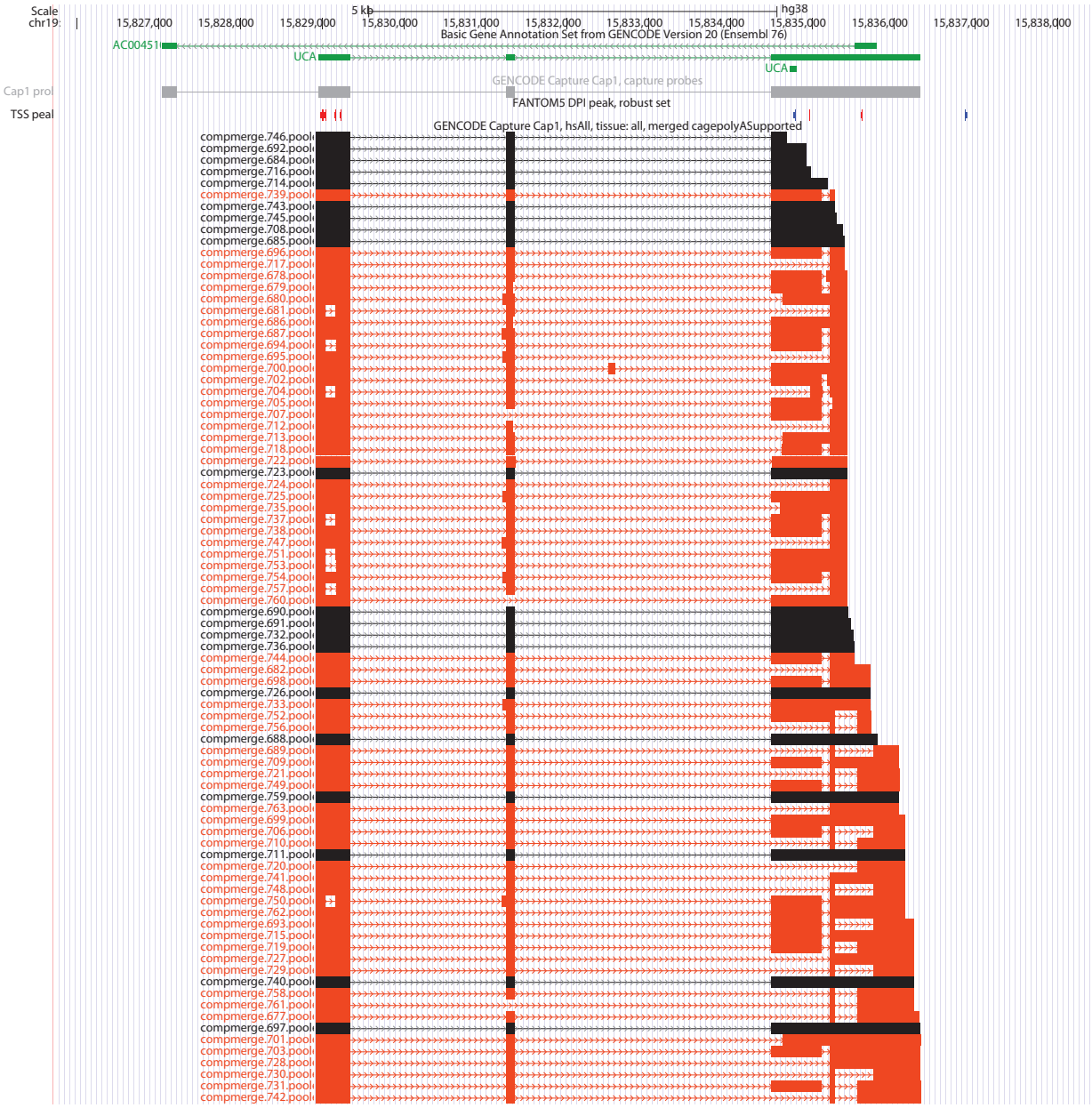
Supplementary Figure S9

LINC01133 / ENSG00000224259

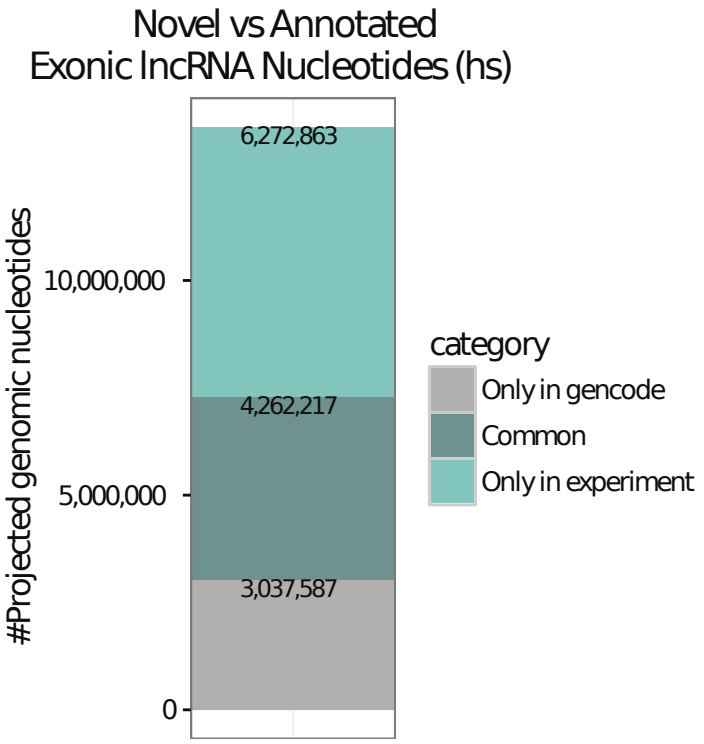


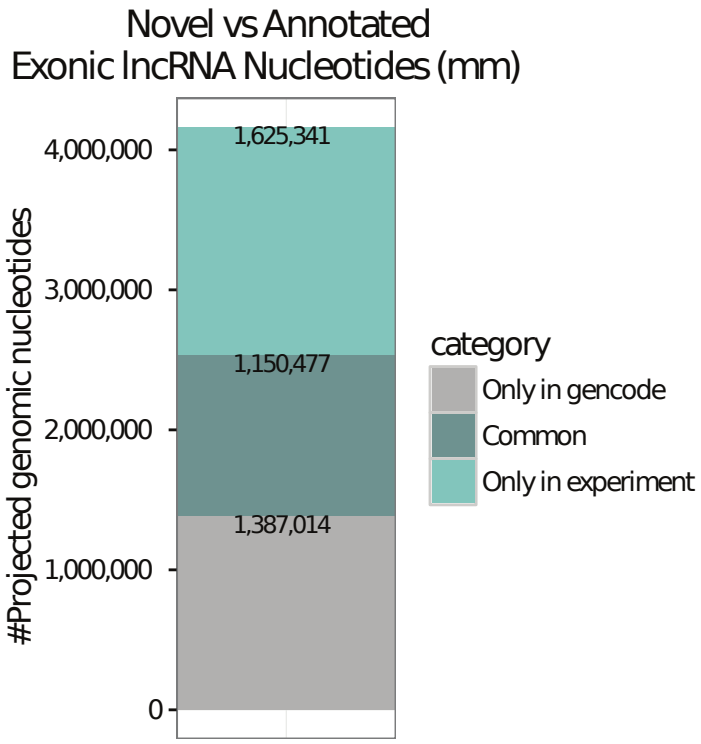
Supplementary Figure S10

UCA1 / ENSG00000214049

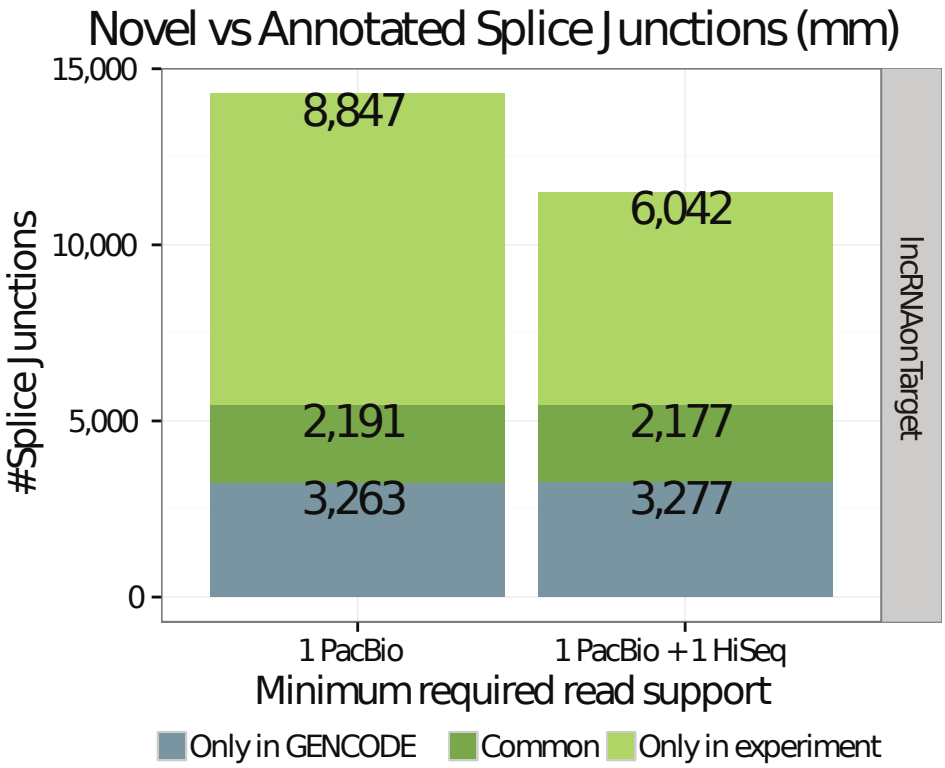


Supplementary Figure S11

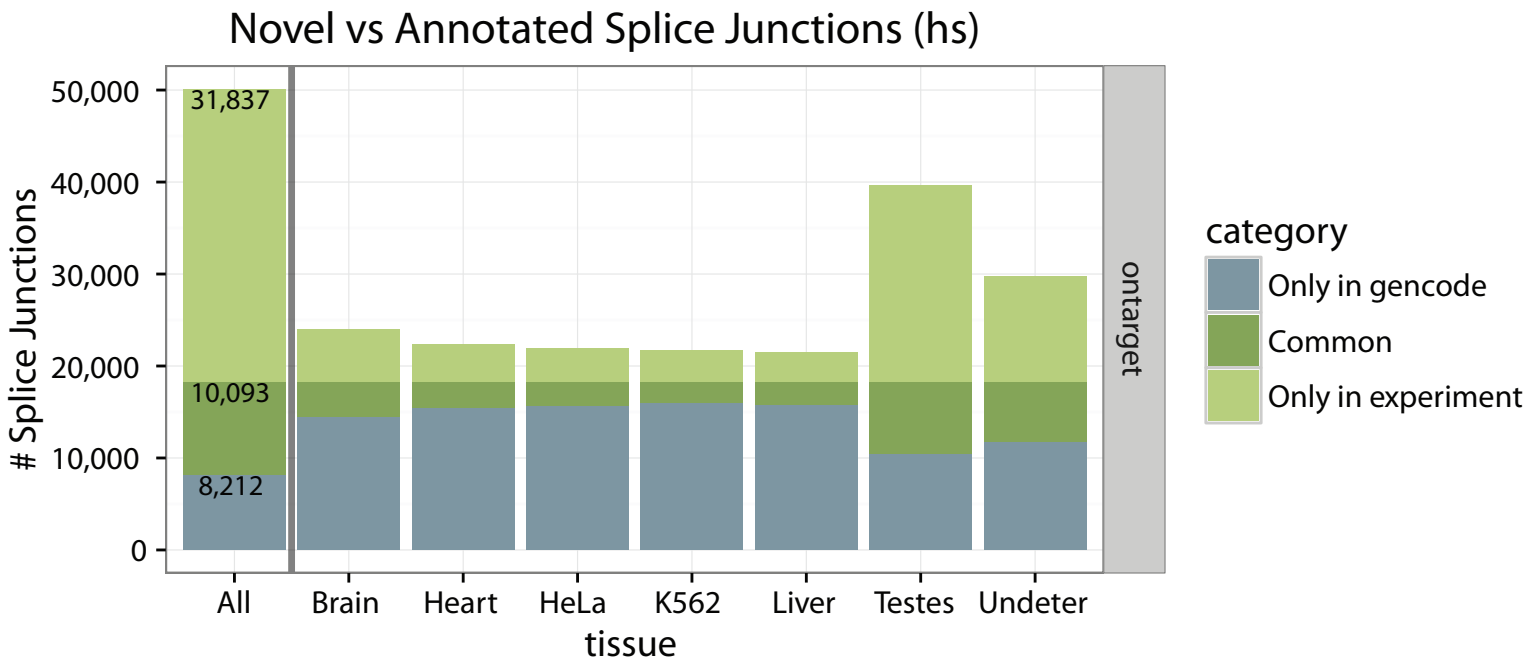


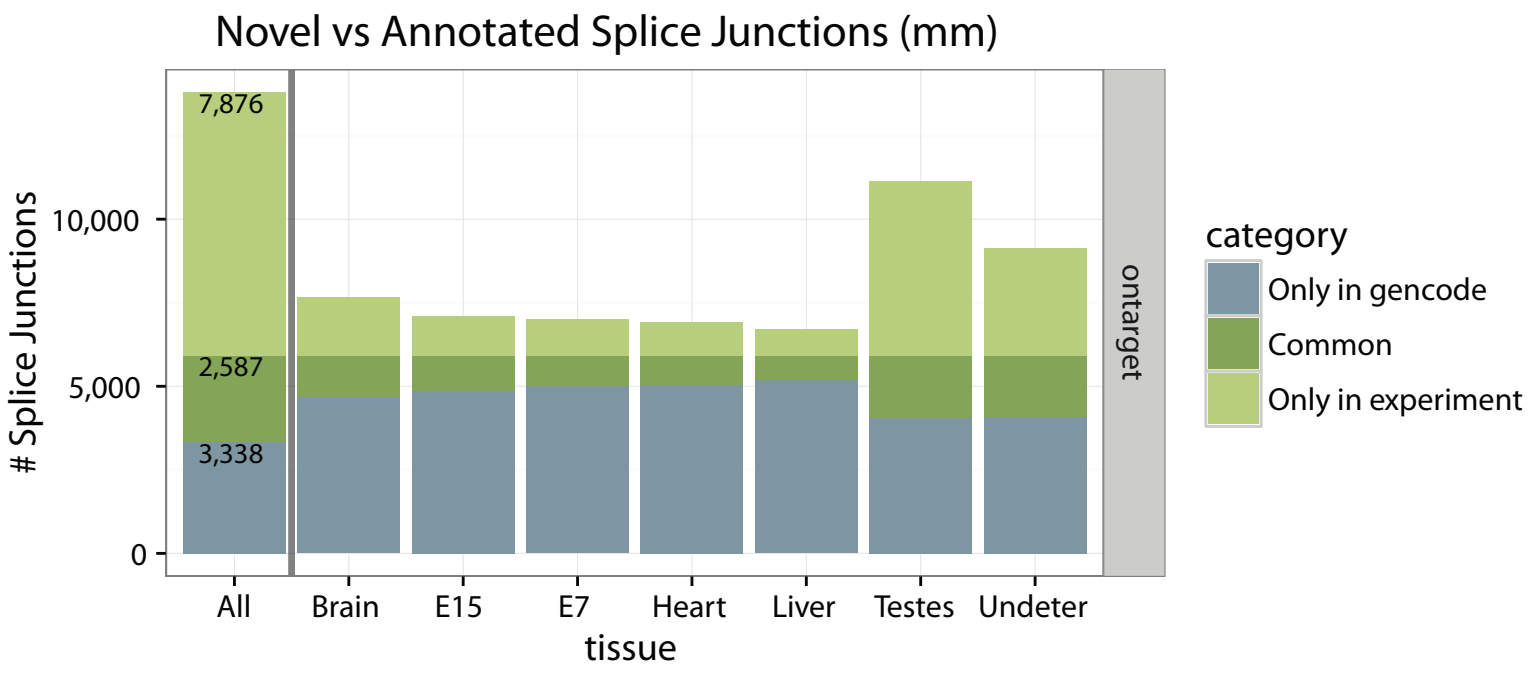


Supplementary Figure S12

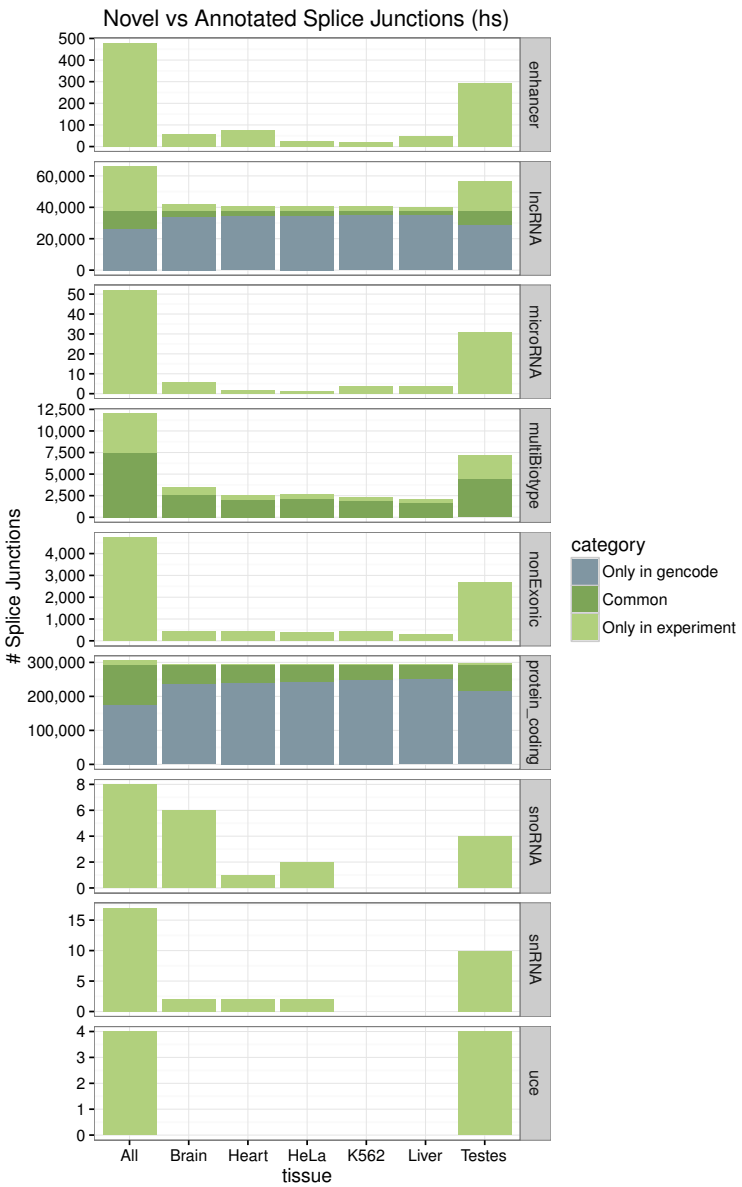


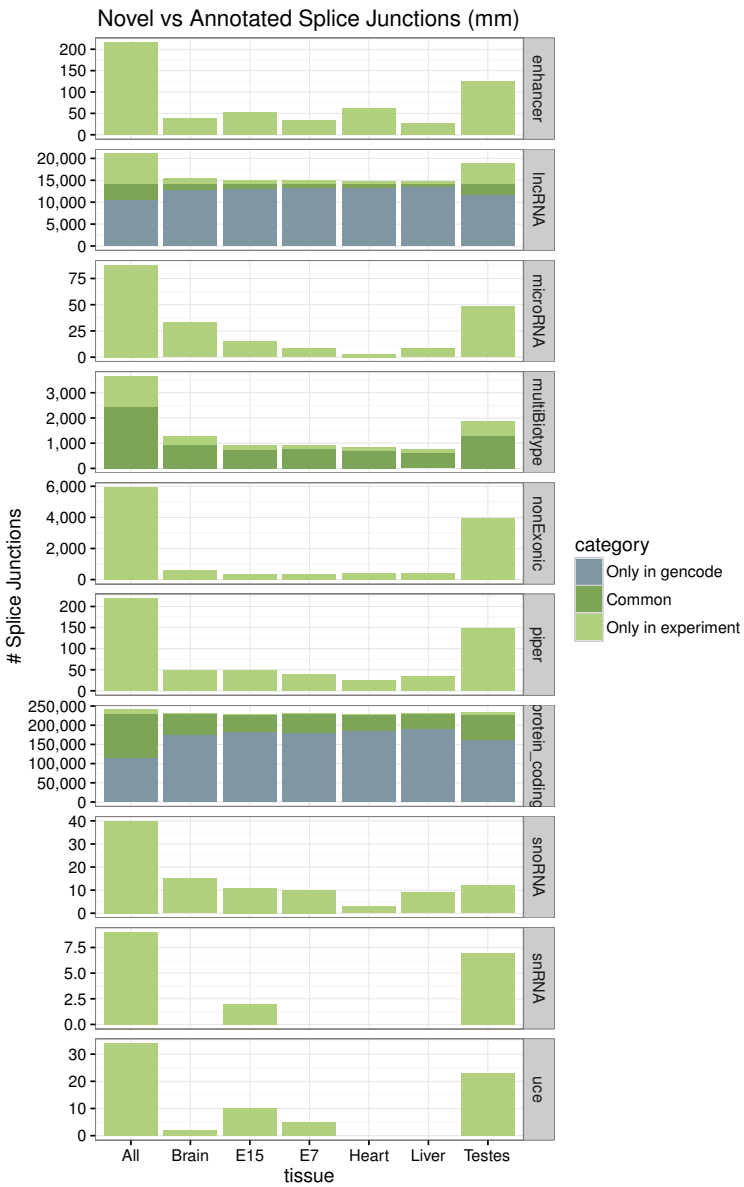
Supplementary Figure S13



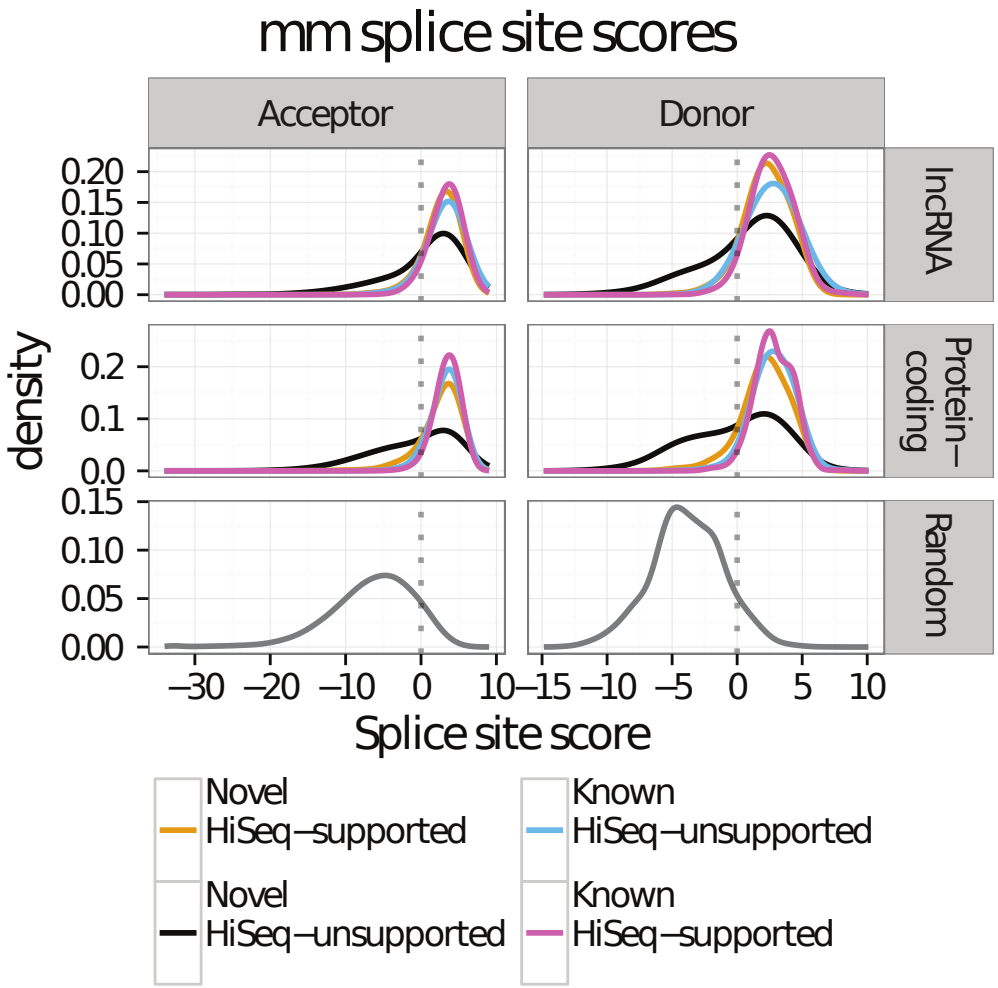


Supplementary Figure S14

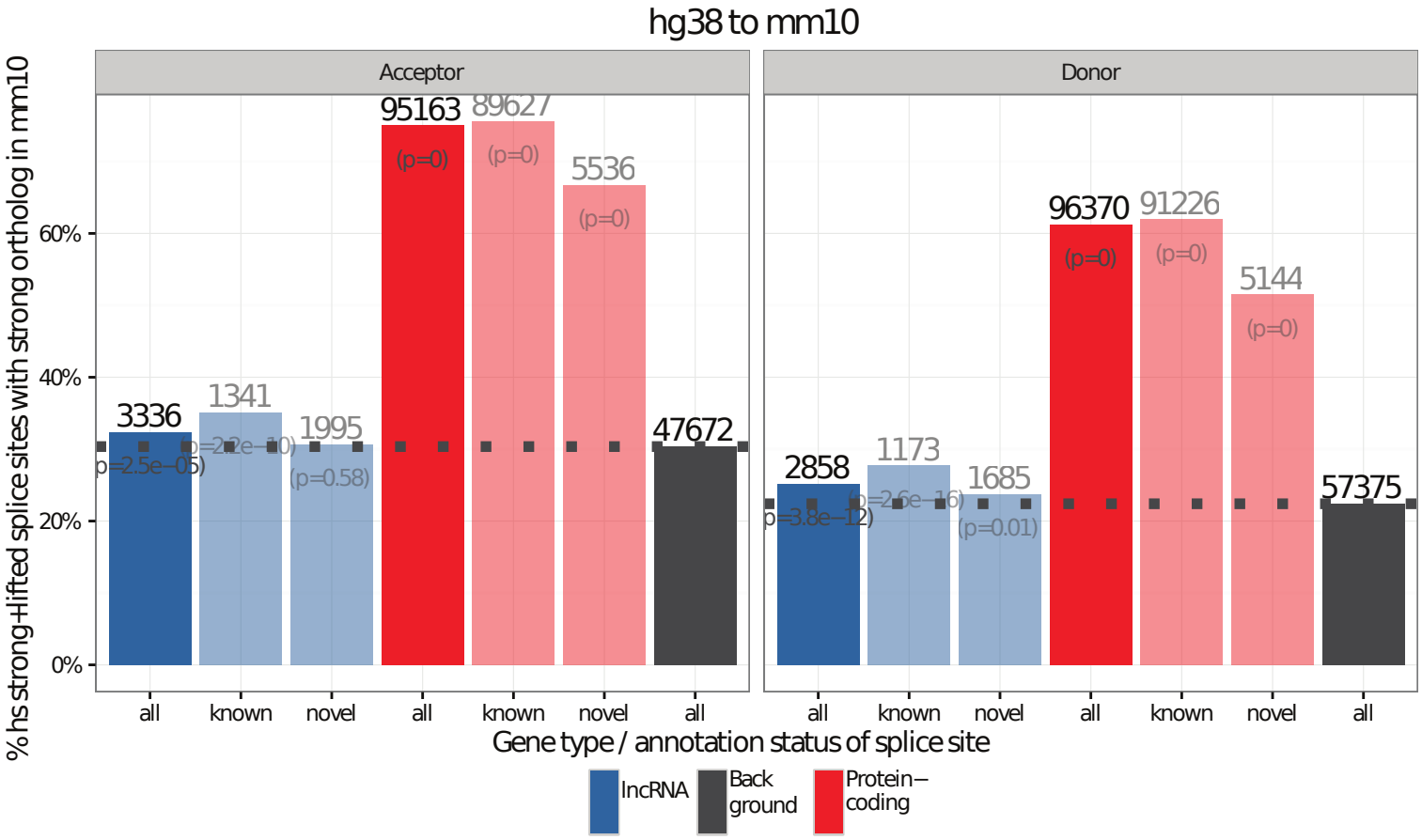


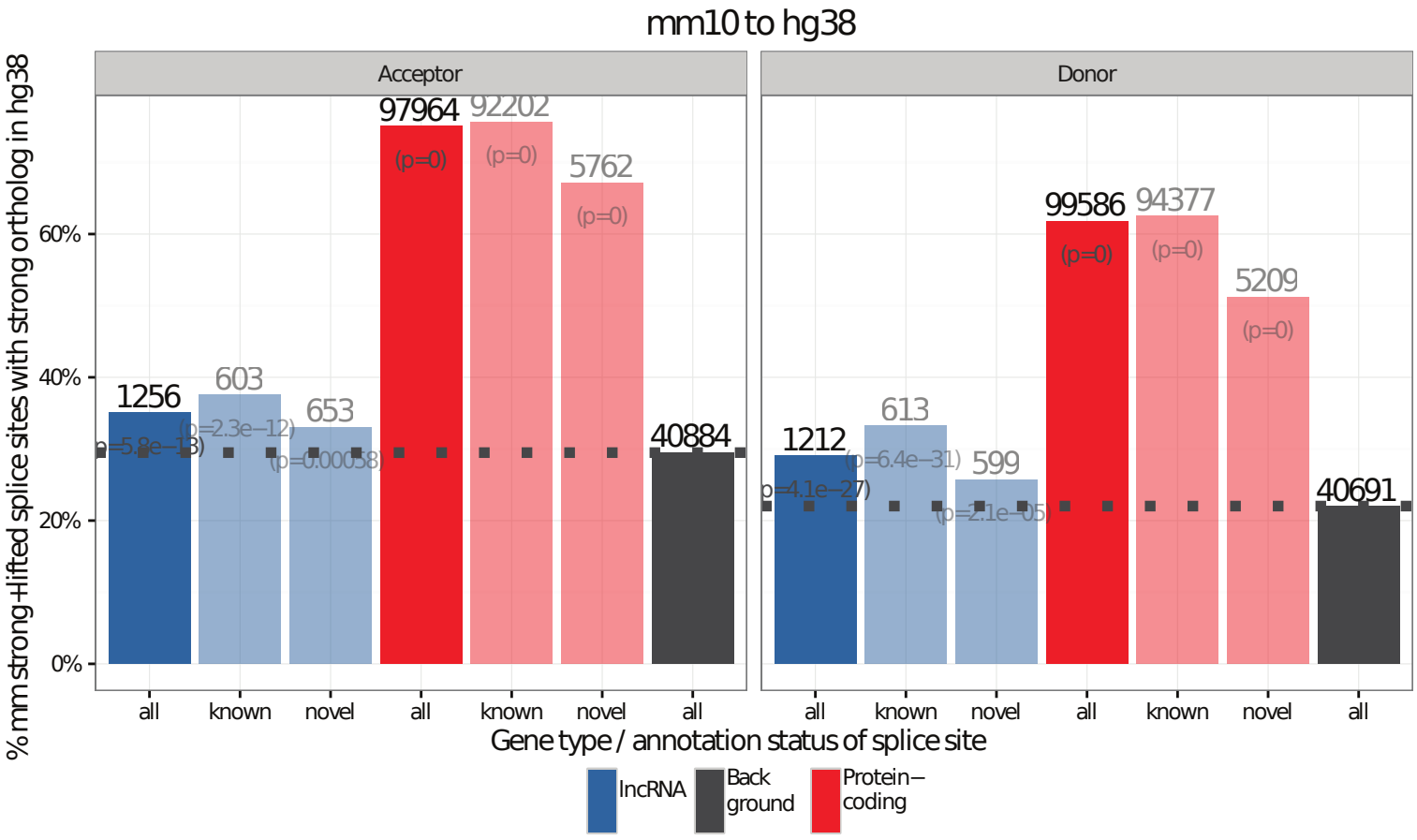


Supplementary Figure S15

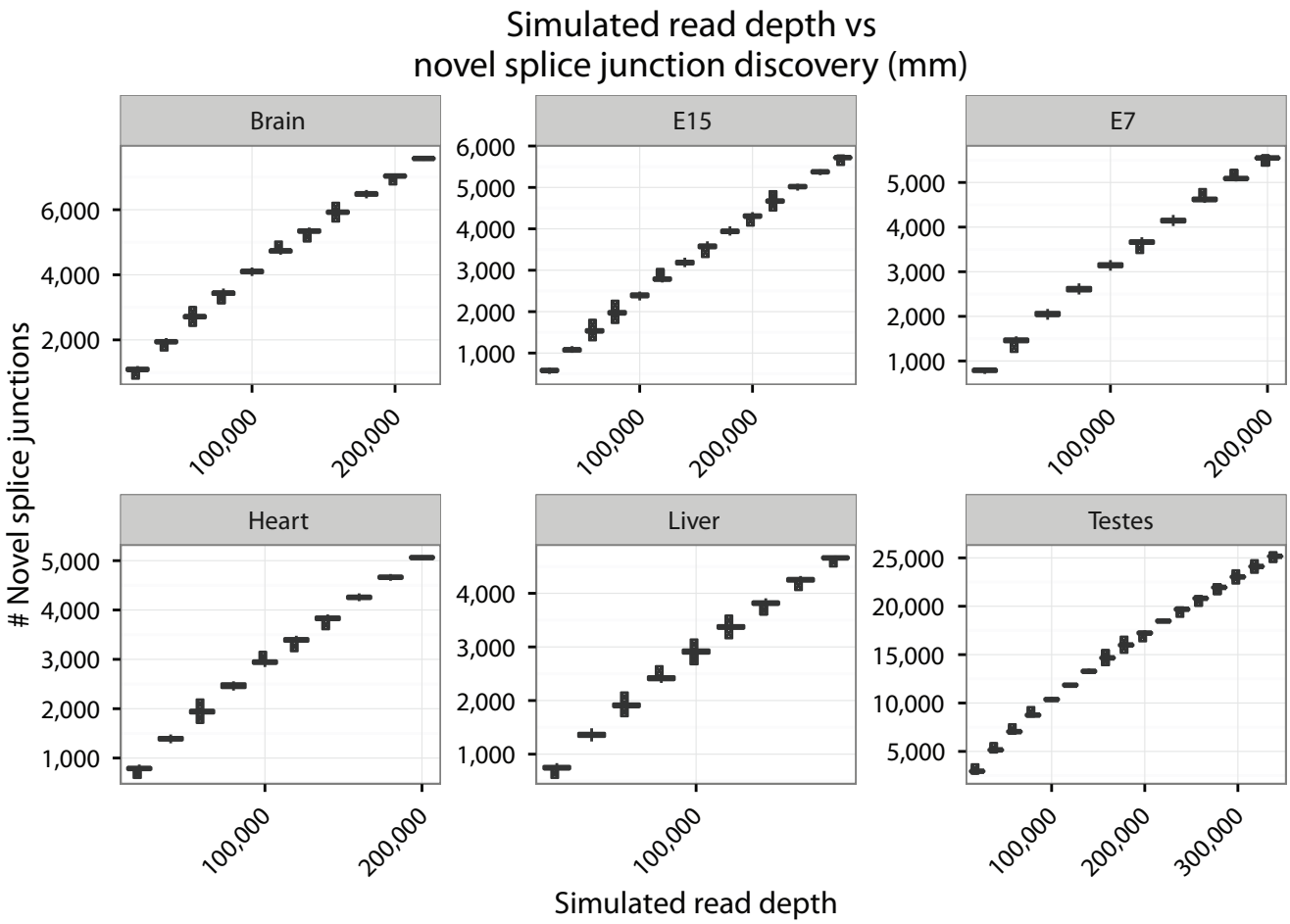


Supplementary Figure S16

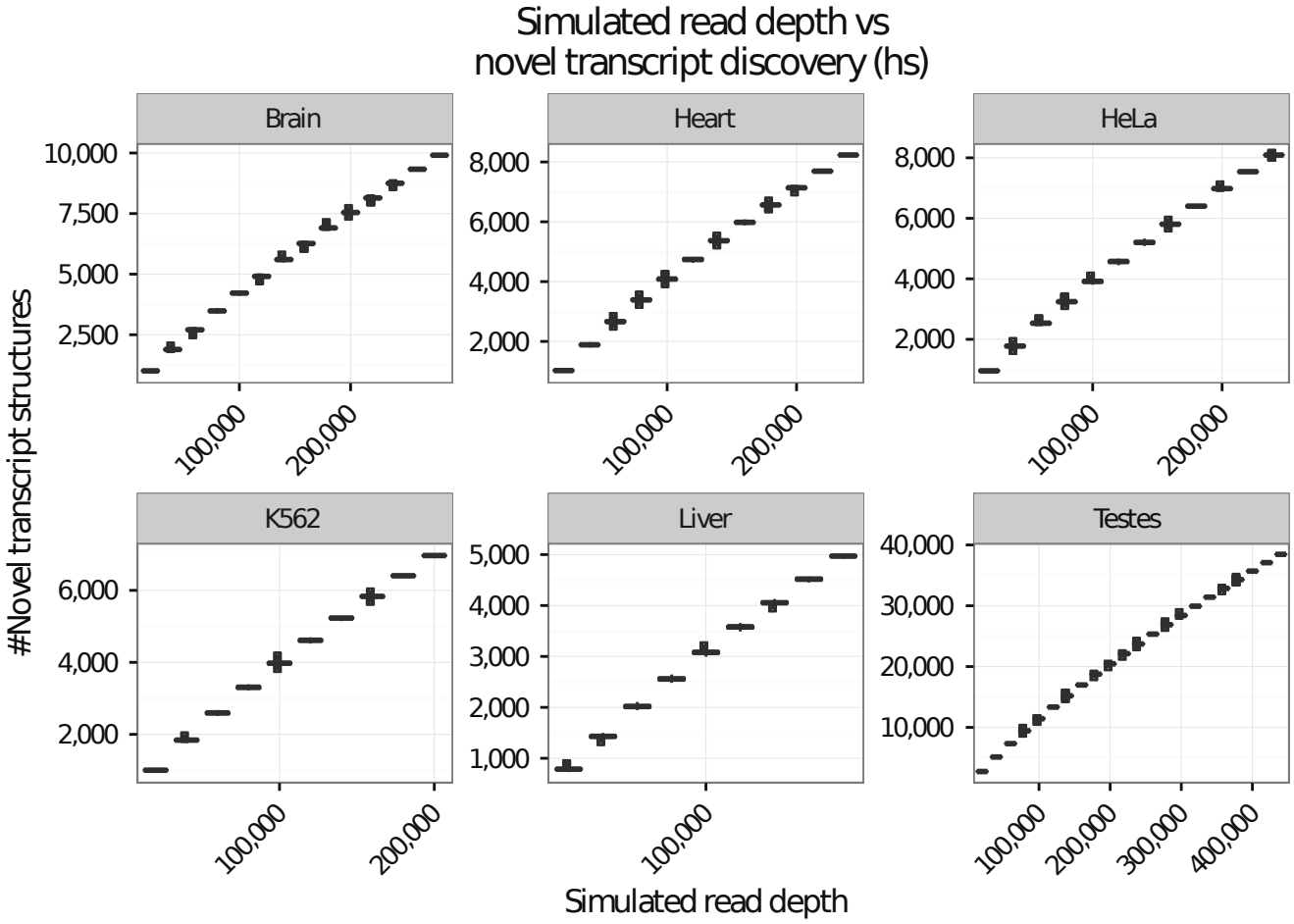


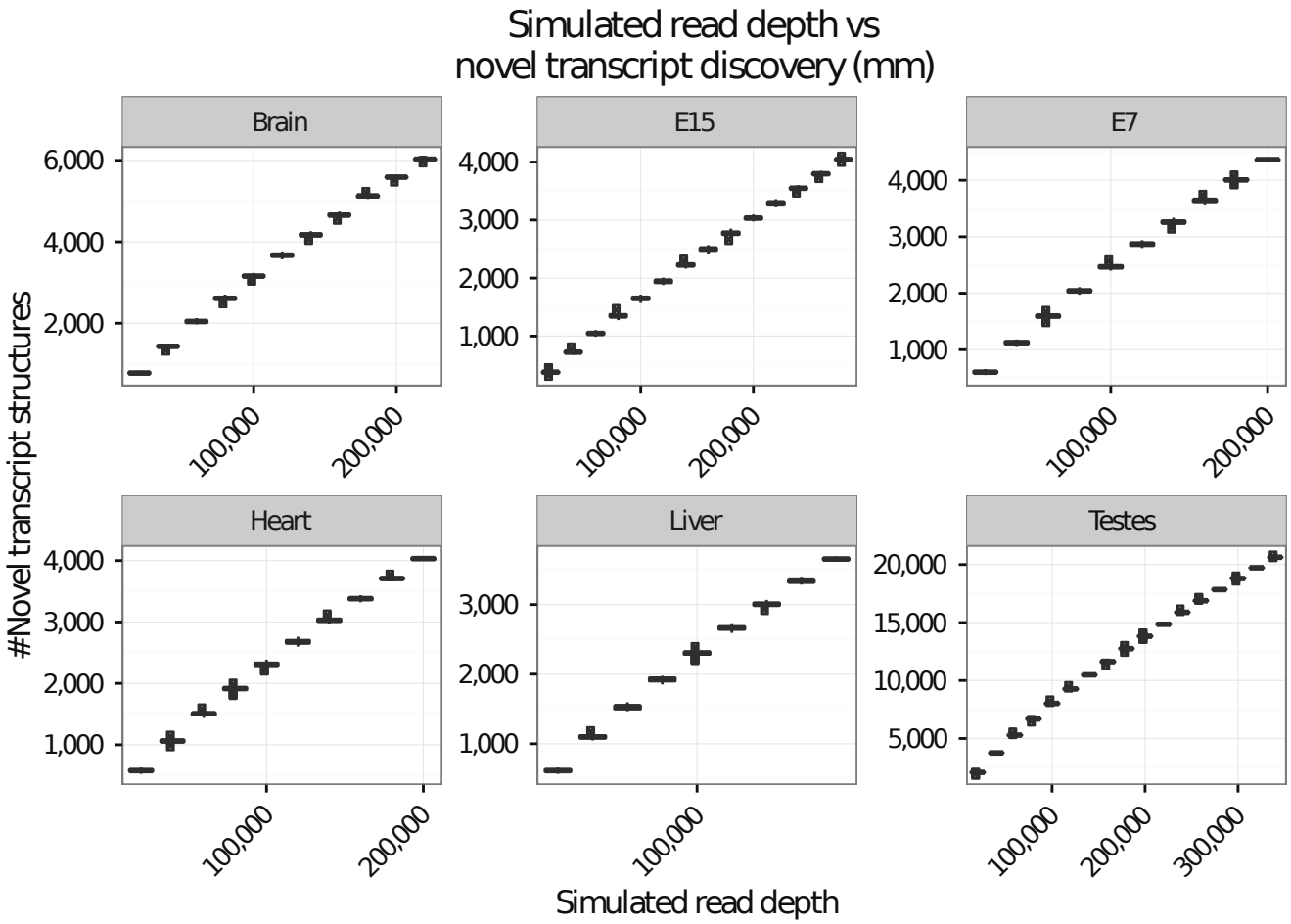


Supplementary Figure S17

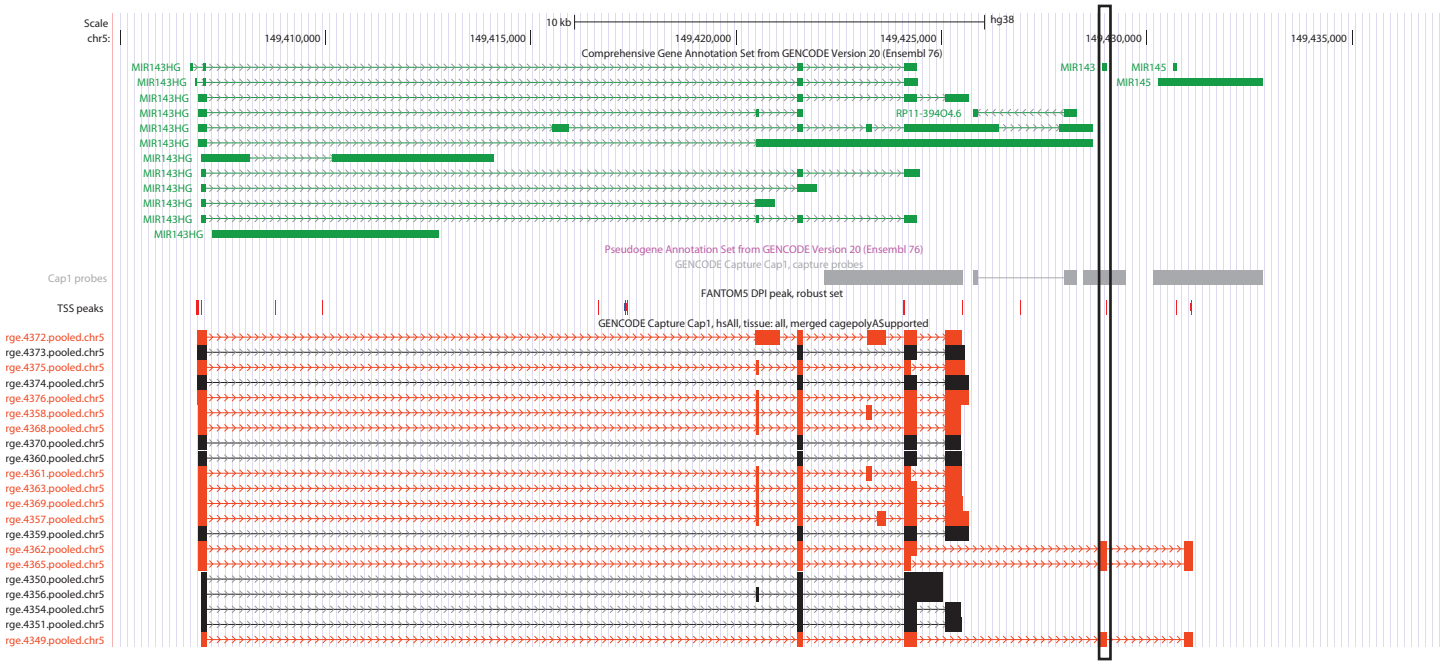


Supplementary Figure S18





Supplementary Figure S19



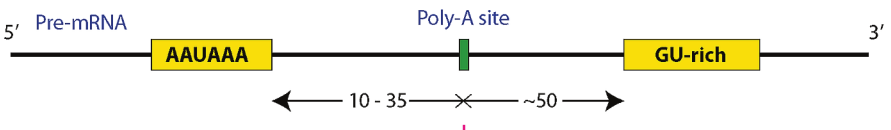
Supplementary Figure S20

A

Canonical polyA signals:

AACAAA
AATAAA
AATACA
AATAGA
AATATA
AATGAA
ACTAAA
AGTAAA
ATTAAA
CATAAA
GATAAA
TATAAA
TTTAAA

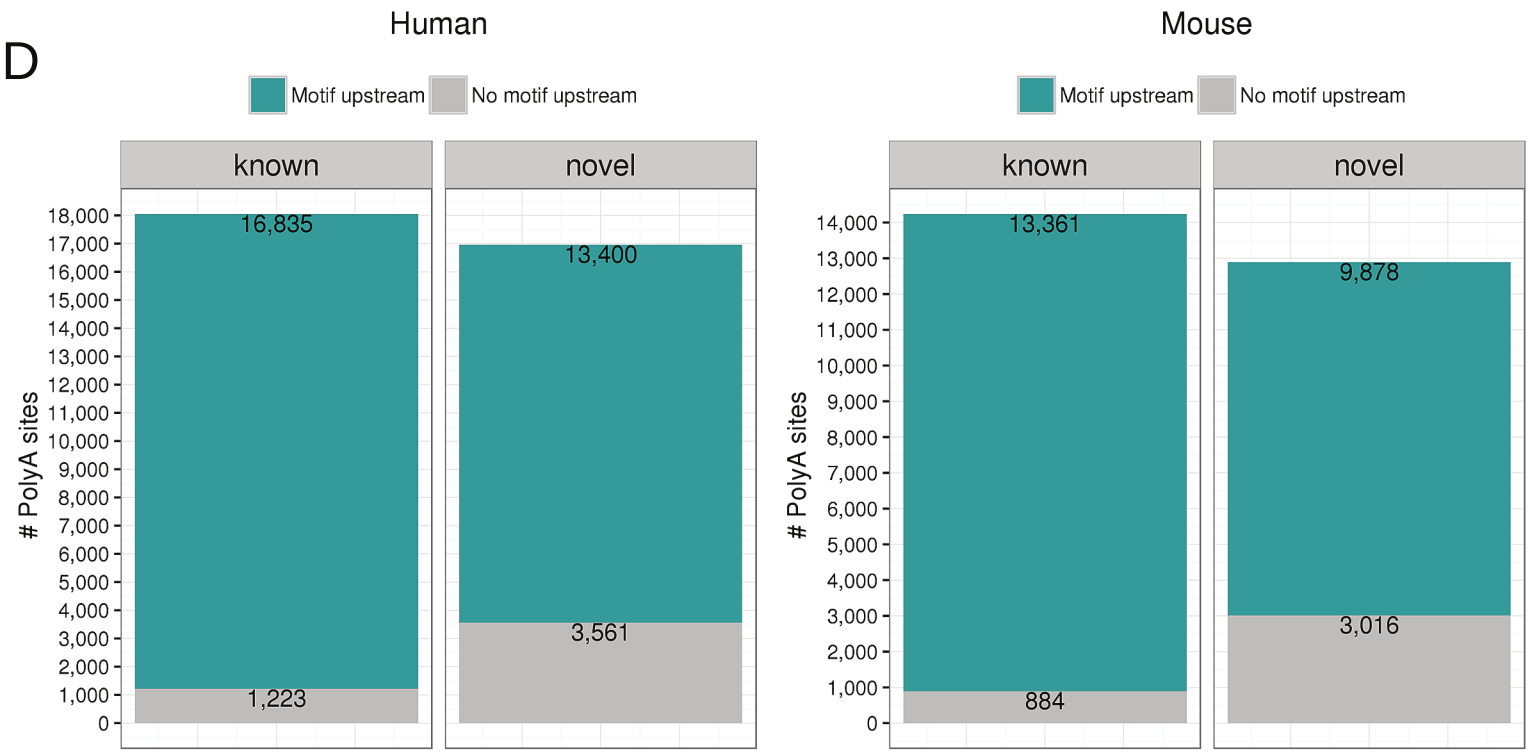
B



C

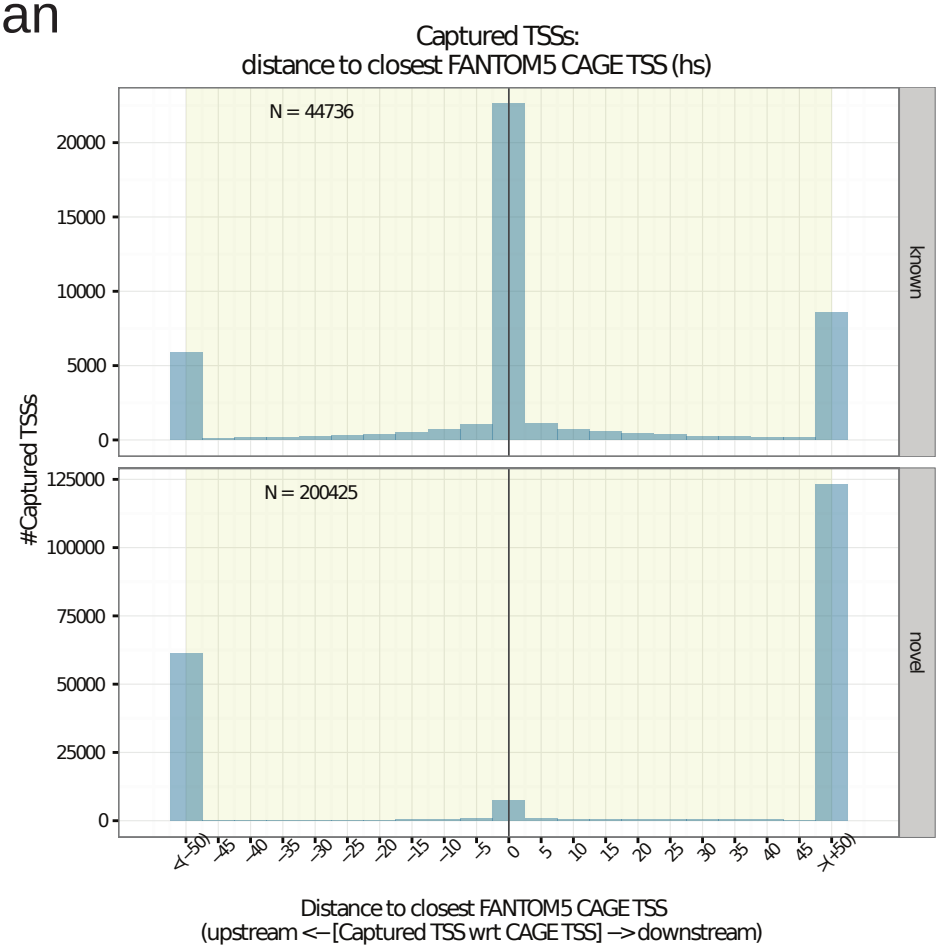
Species	Set	Total # sites	# Sites with Poly-A signal upstream (-10 to -50 nts)	% Sites with Poly-A signal upstream (-10 to -50 nts)
Human	**Poly-A sites**	35,092	30,301	**86%**
	Control	203,303	30,519	15%
Mouse	**Poly-A sites**	27,152	23,249	**86%**
	Control	175,240	20,423	12%

D

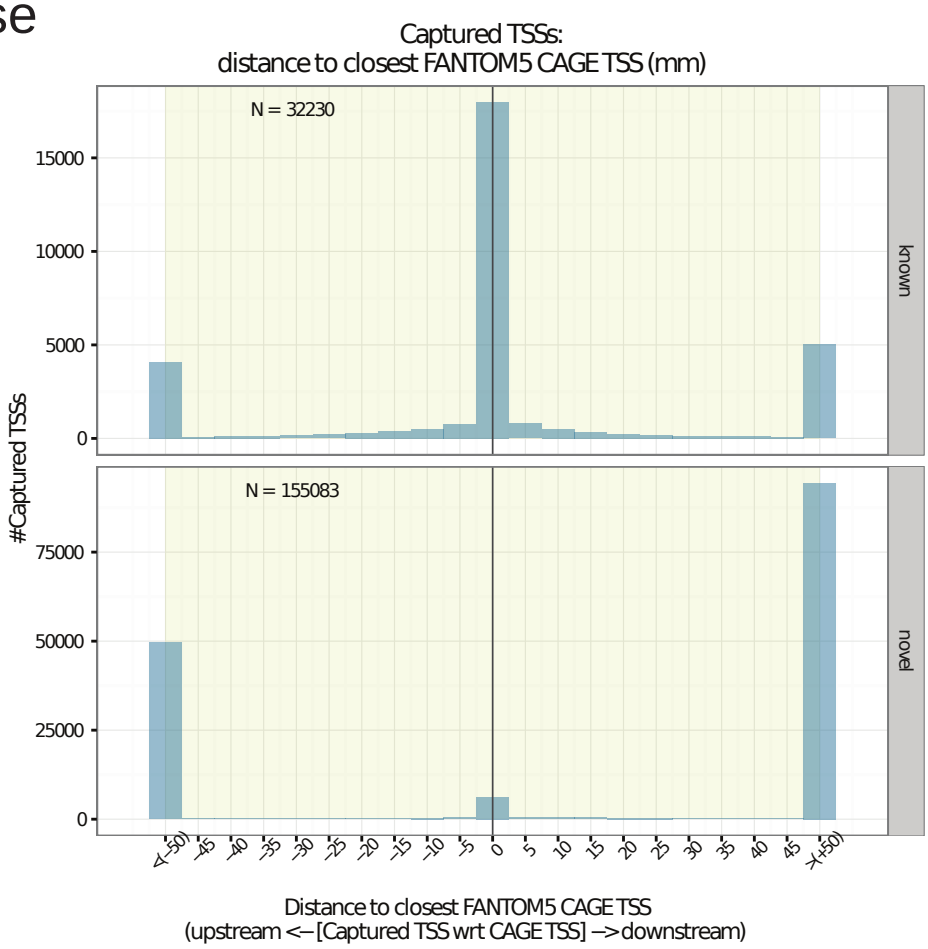


Supplementary Figure S21

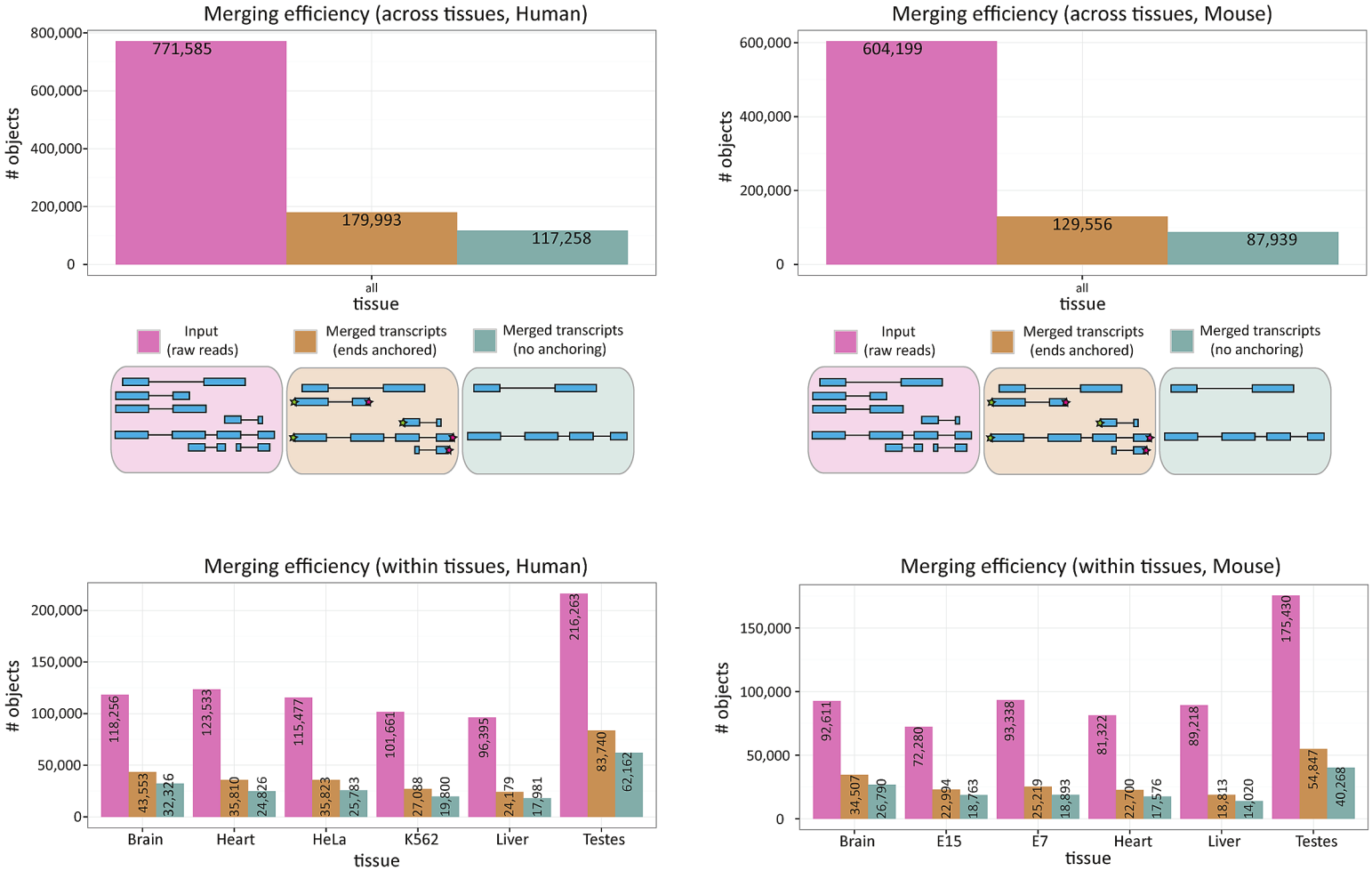
Human



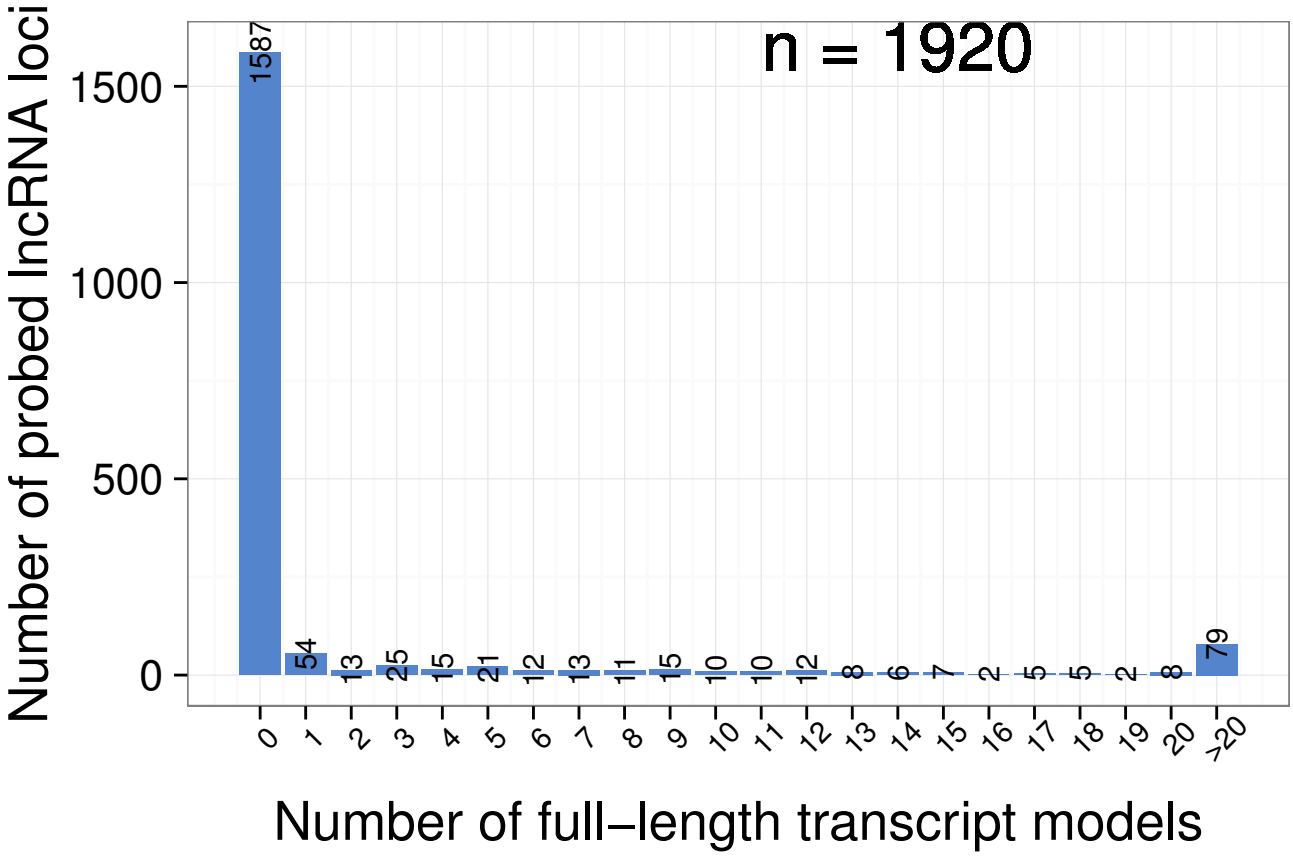
Mouse



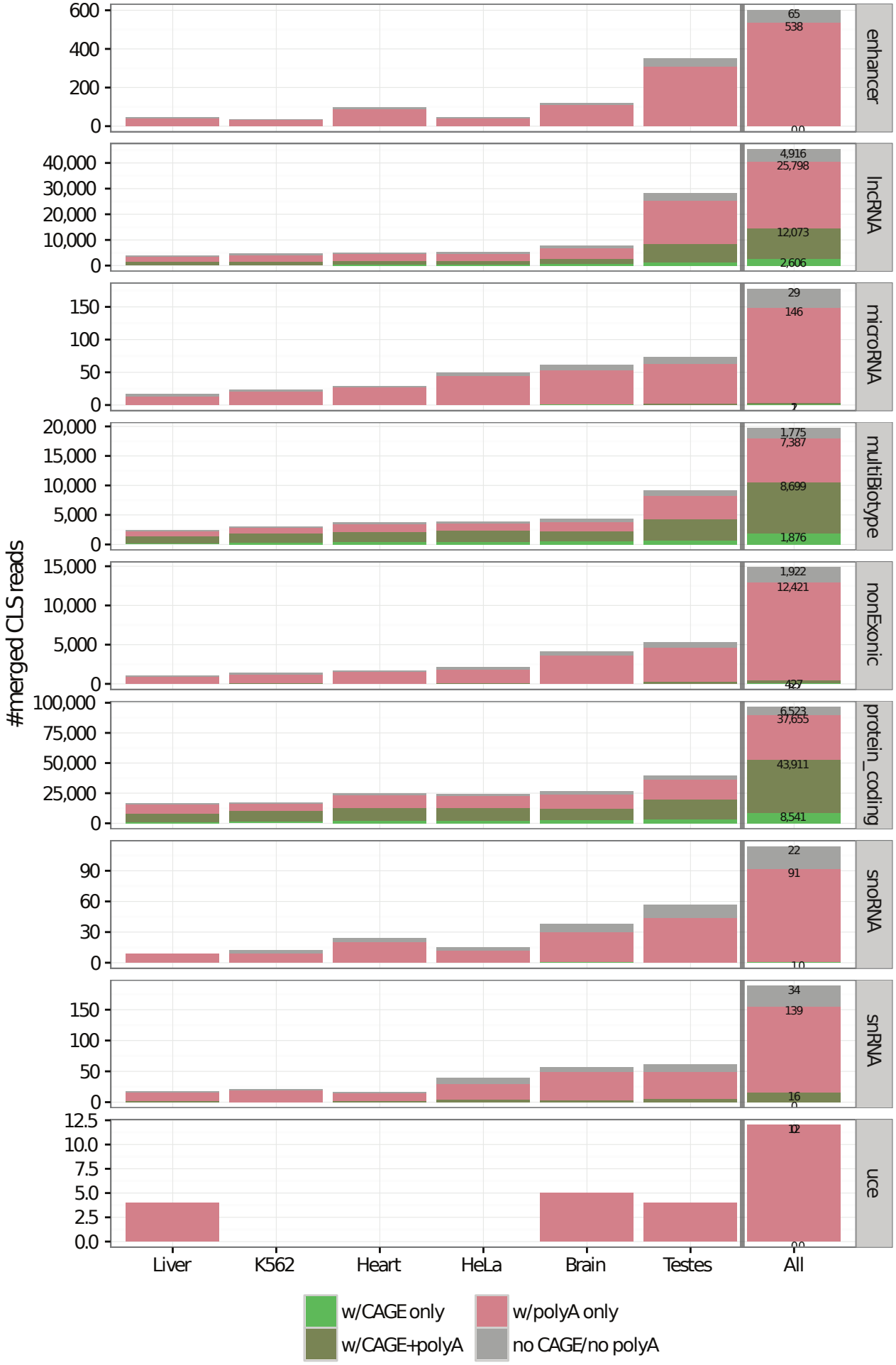
Supplementary Figure S22



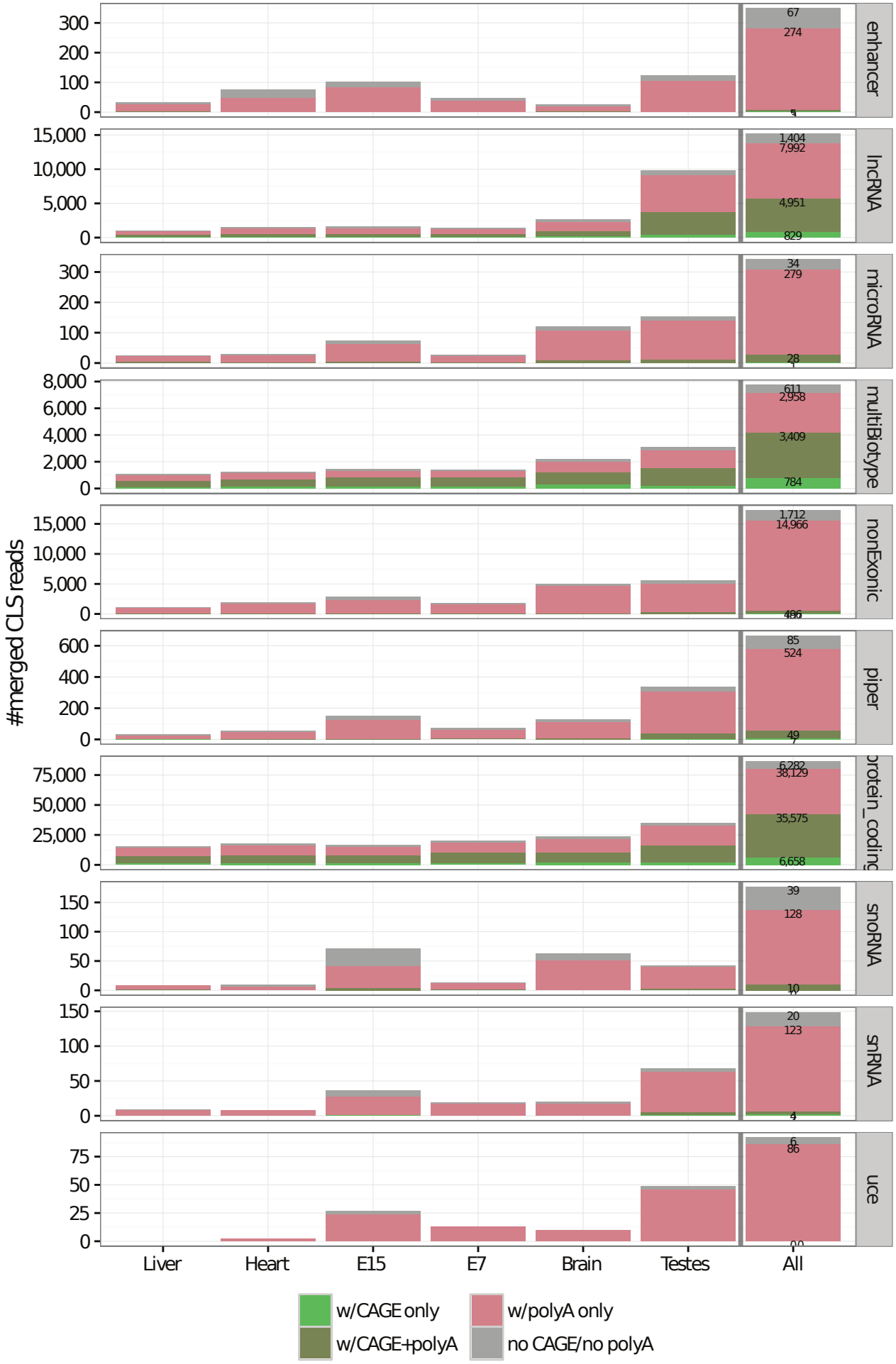
Supplementary Figure S23



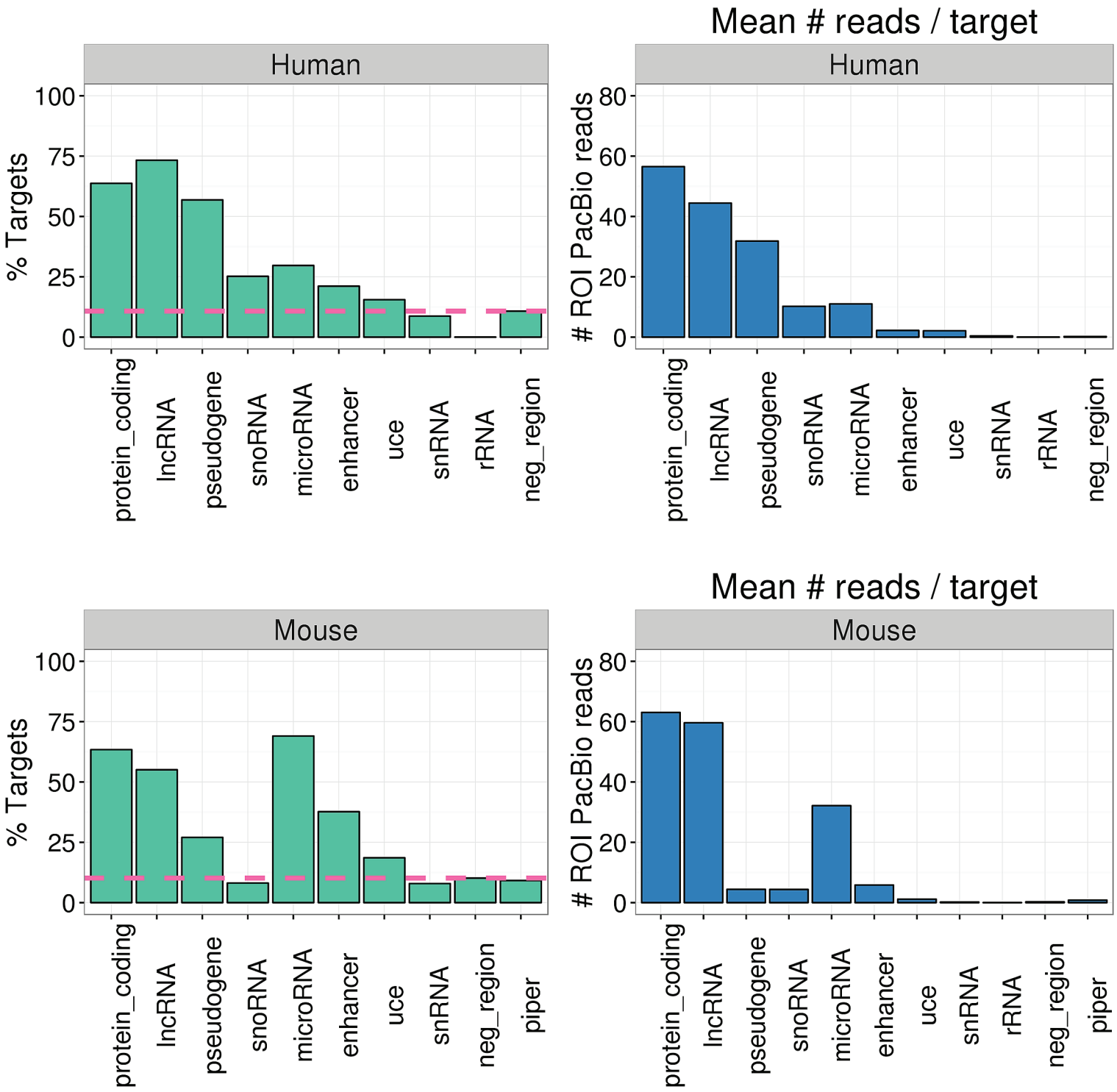
Supplementary Figure S24



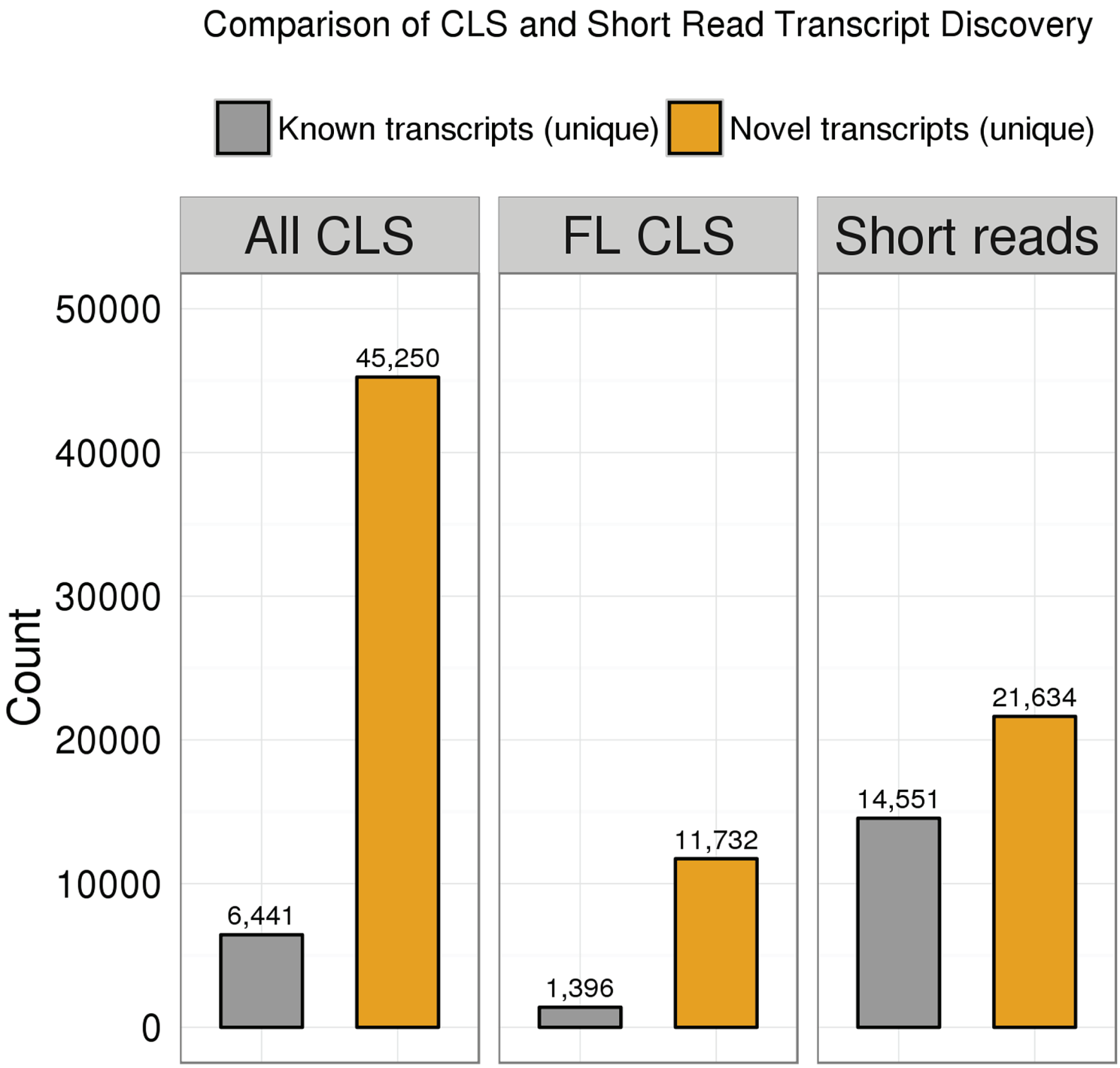
Supplementary Figure S25



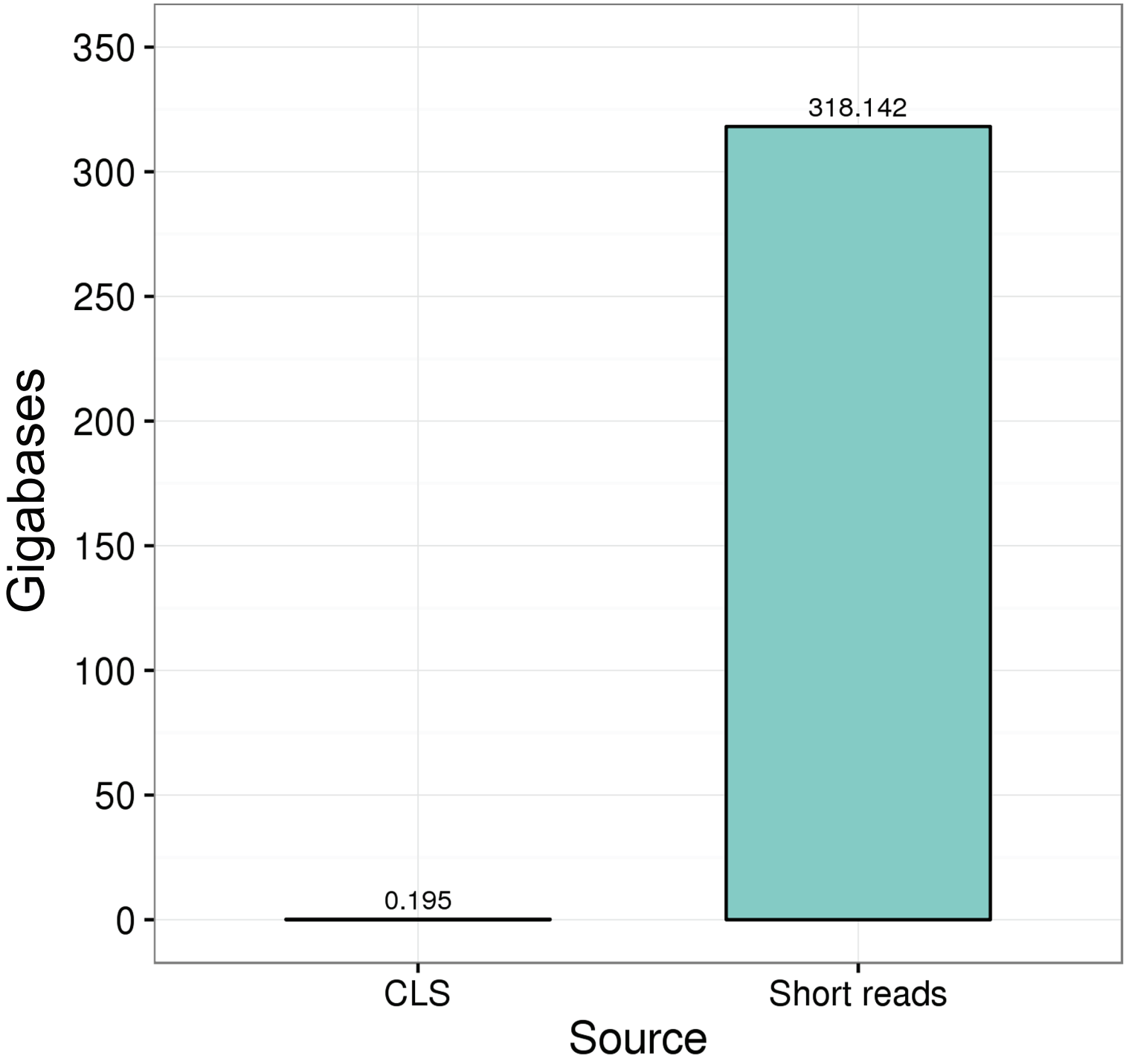
Supplementary Figure S26



Supplementary Figure S27

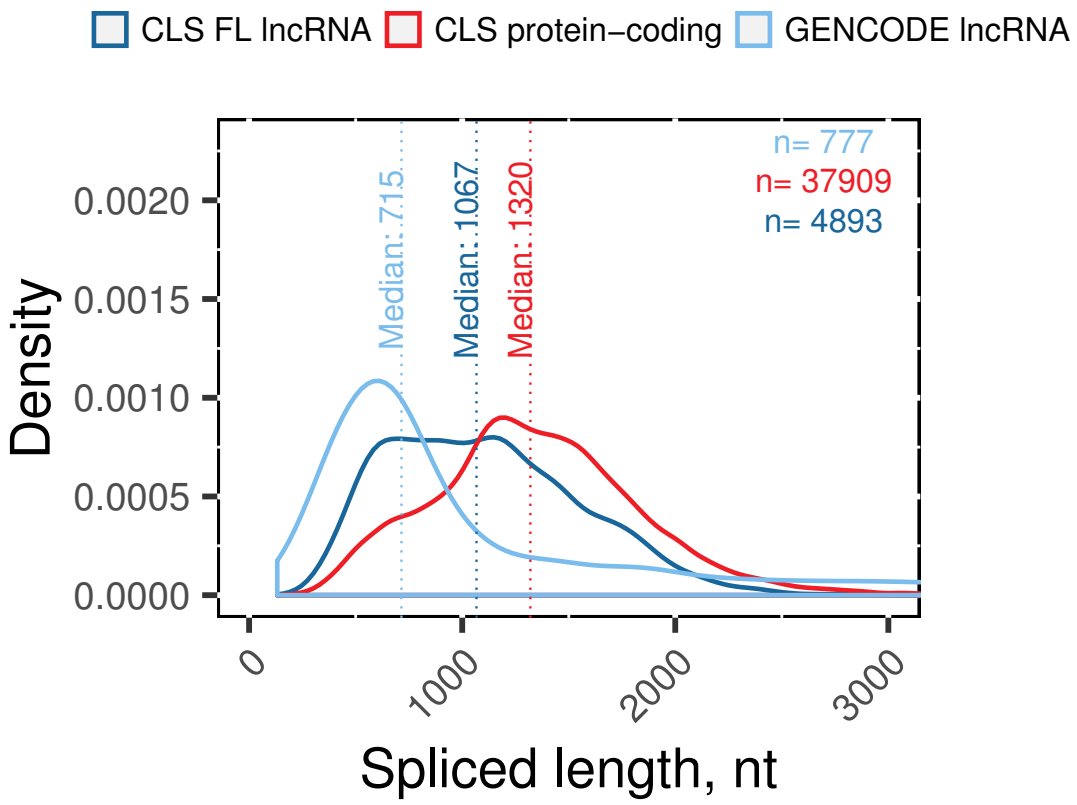


Supplementary Figure S28

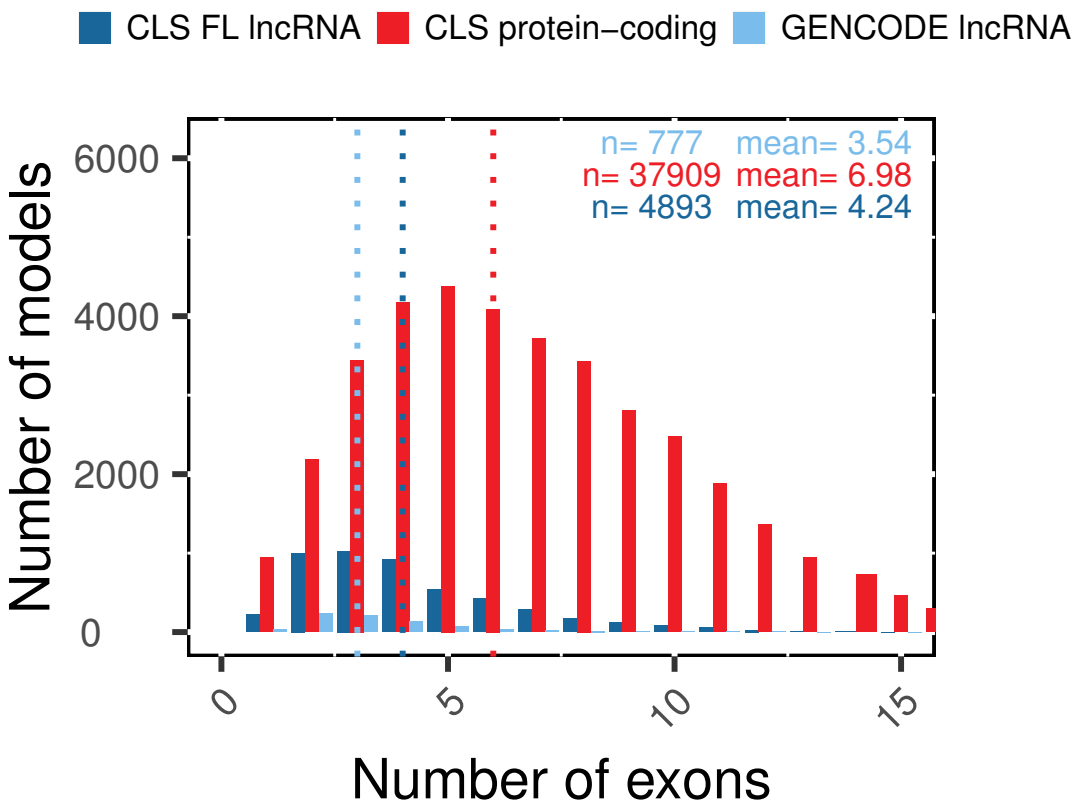


Supplementary Figure S29

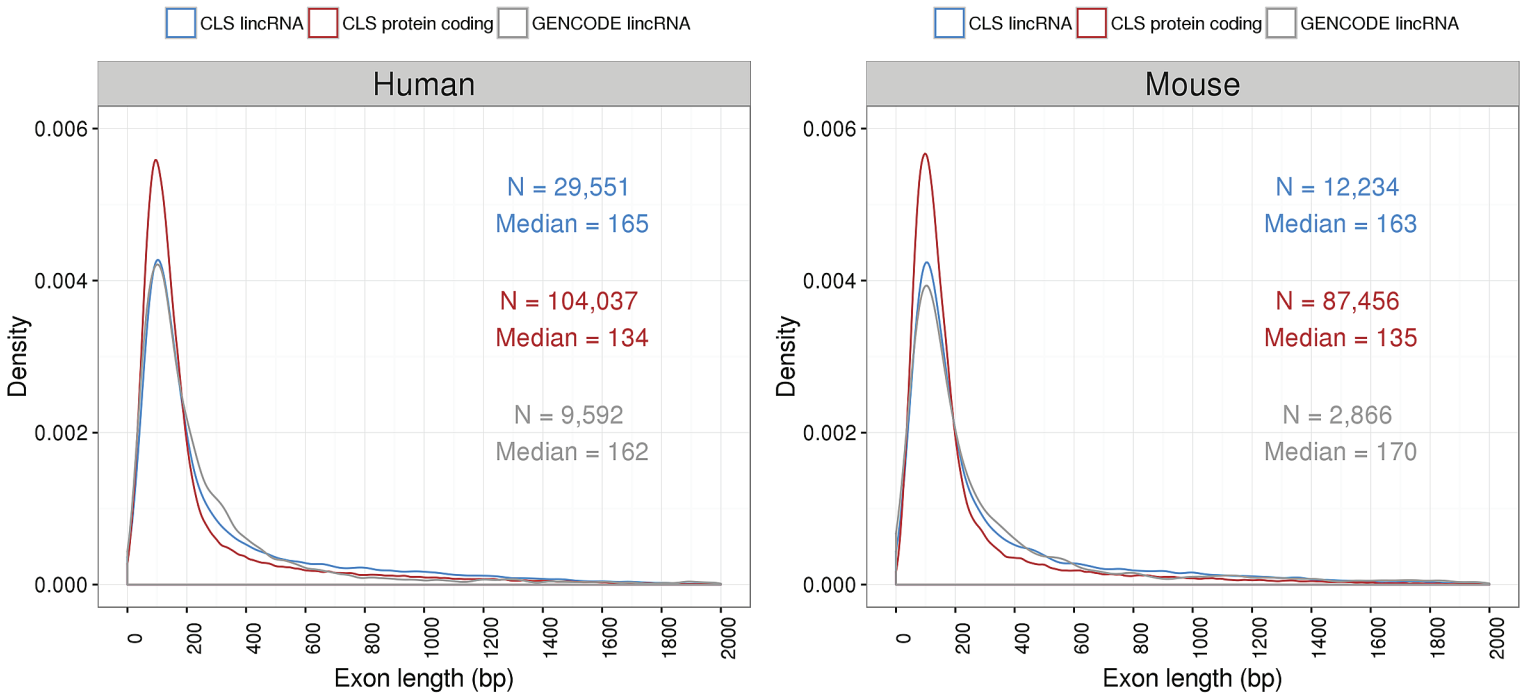
A



B

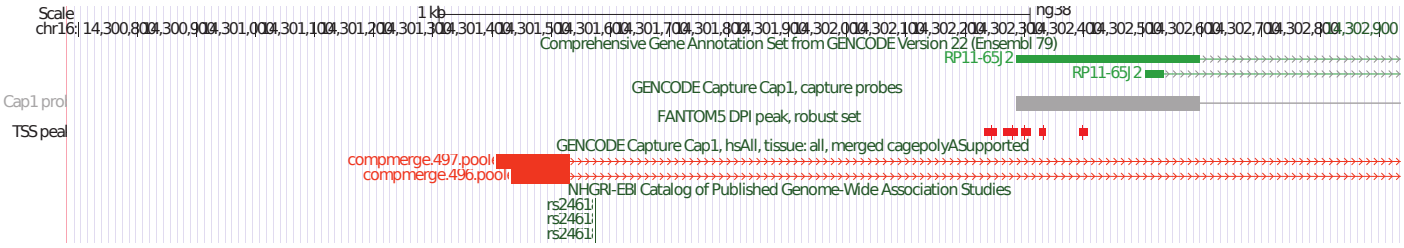


Supplementary Figure S30

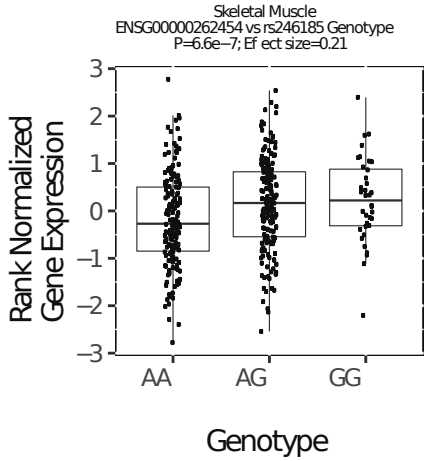


Supplementary Figure S31

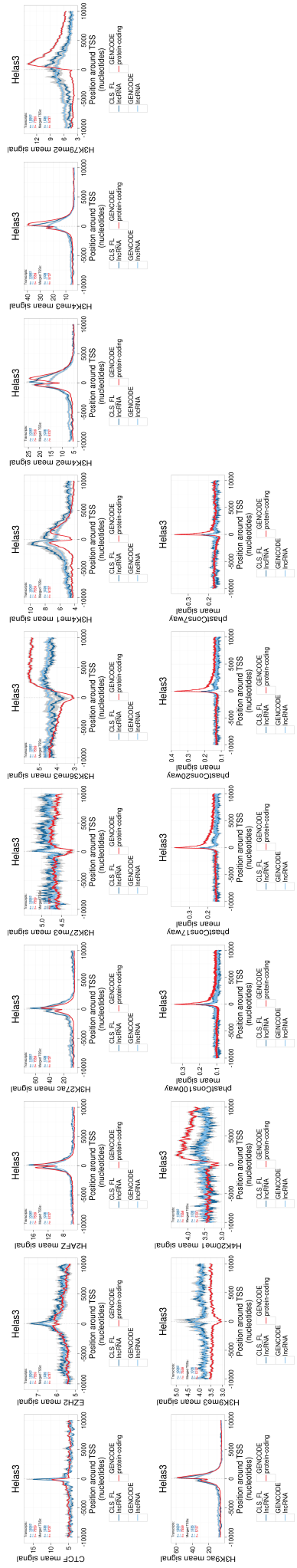
A



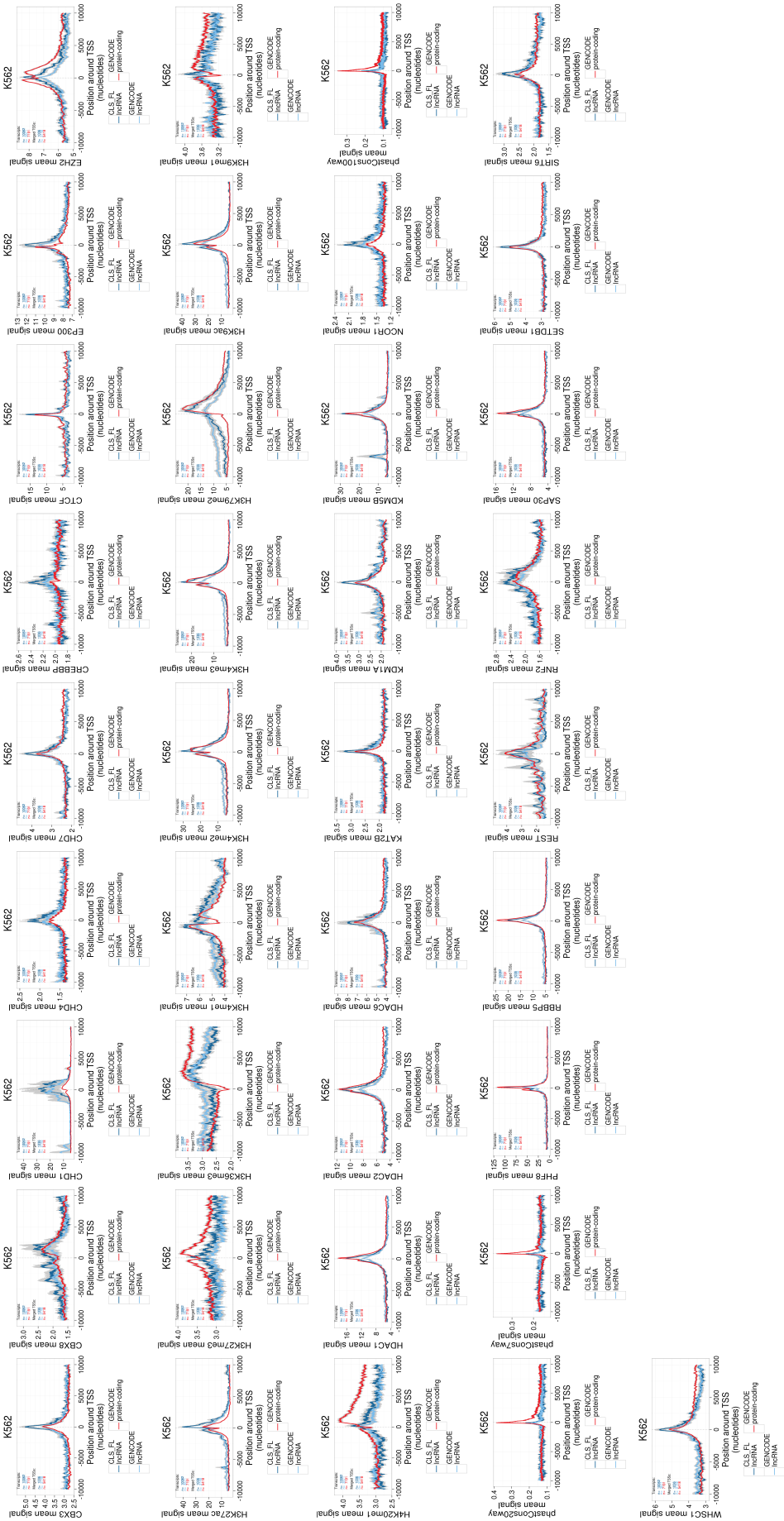
B



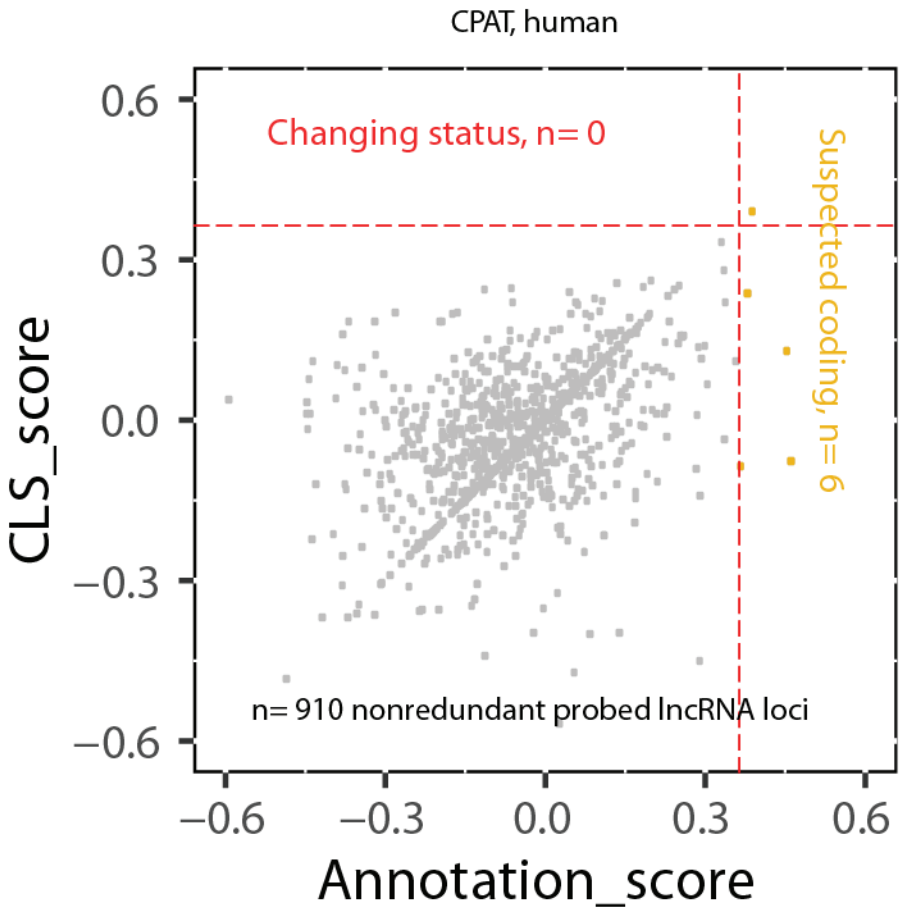
Supplementary Figure S32

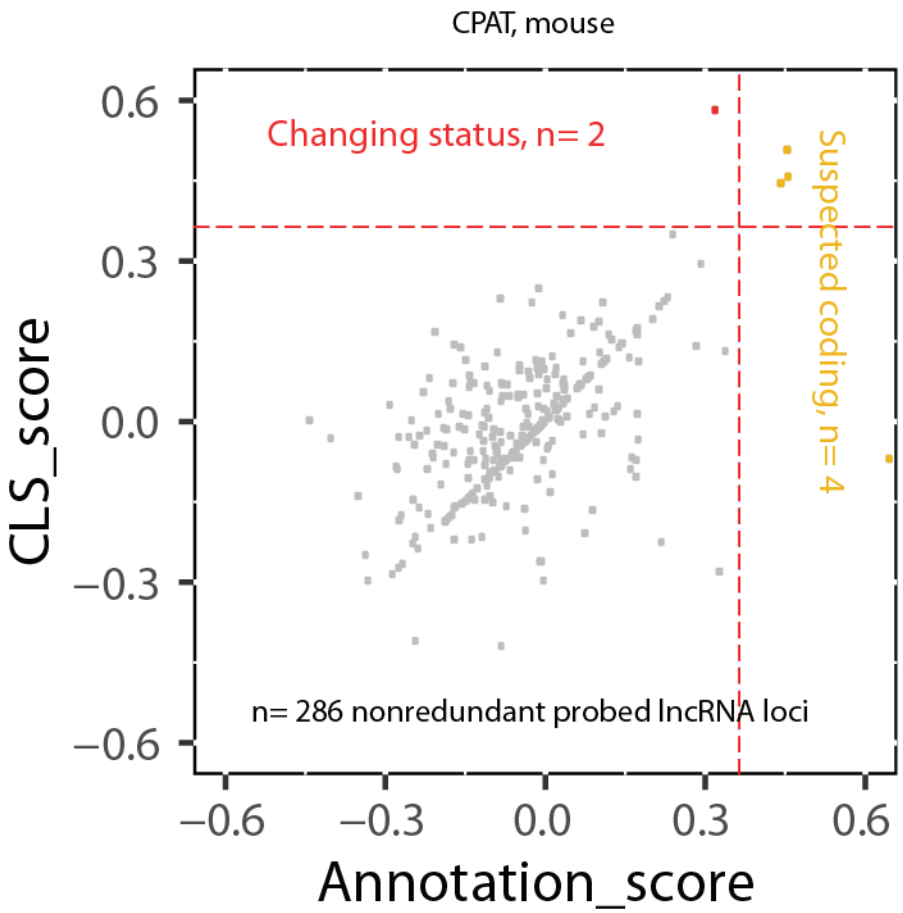


Supplementary Figure S33

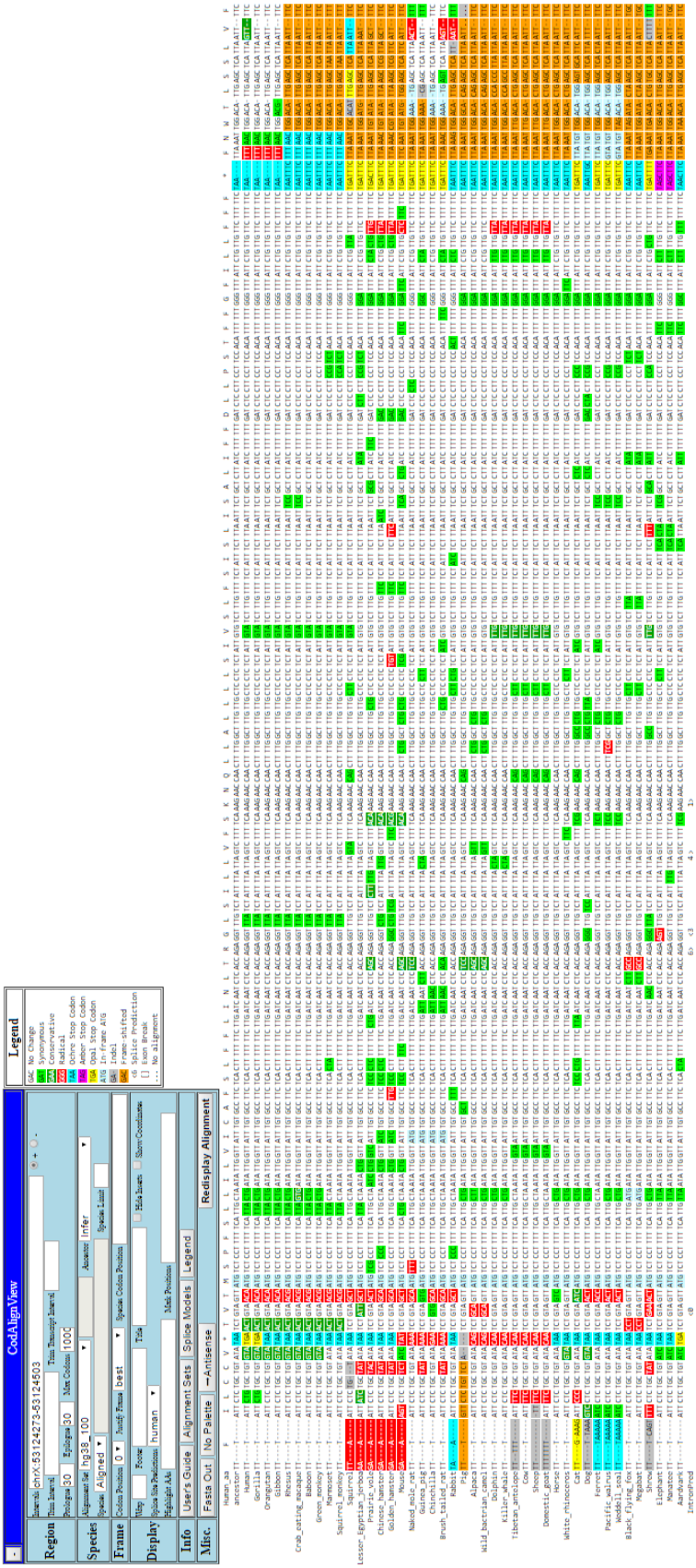


Supplementary Figure S34

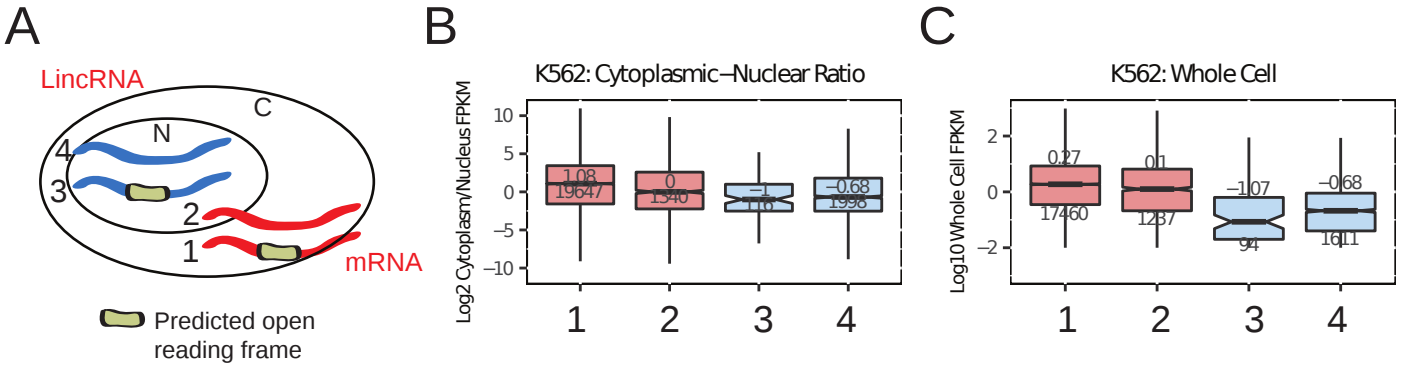




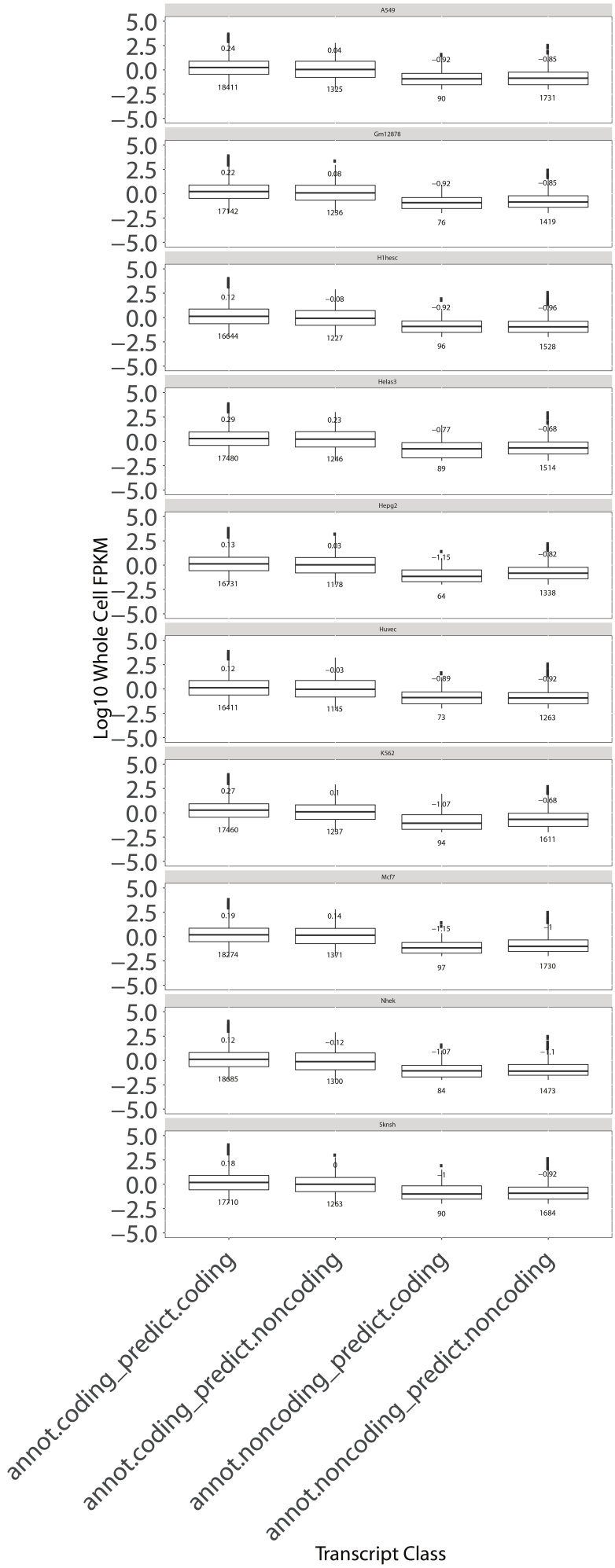
Supplementary Figure S35



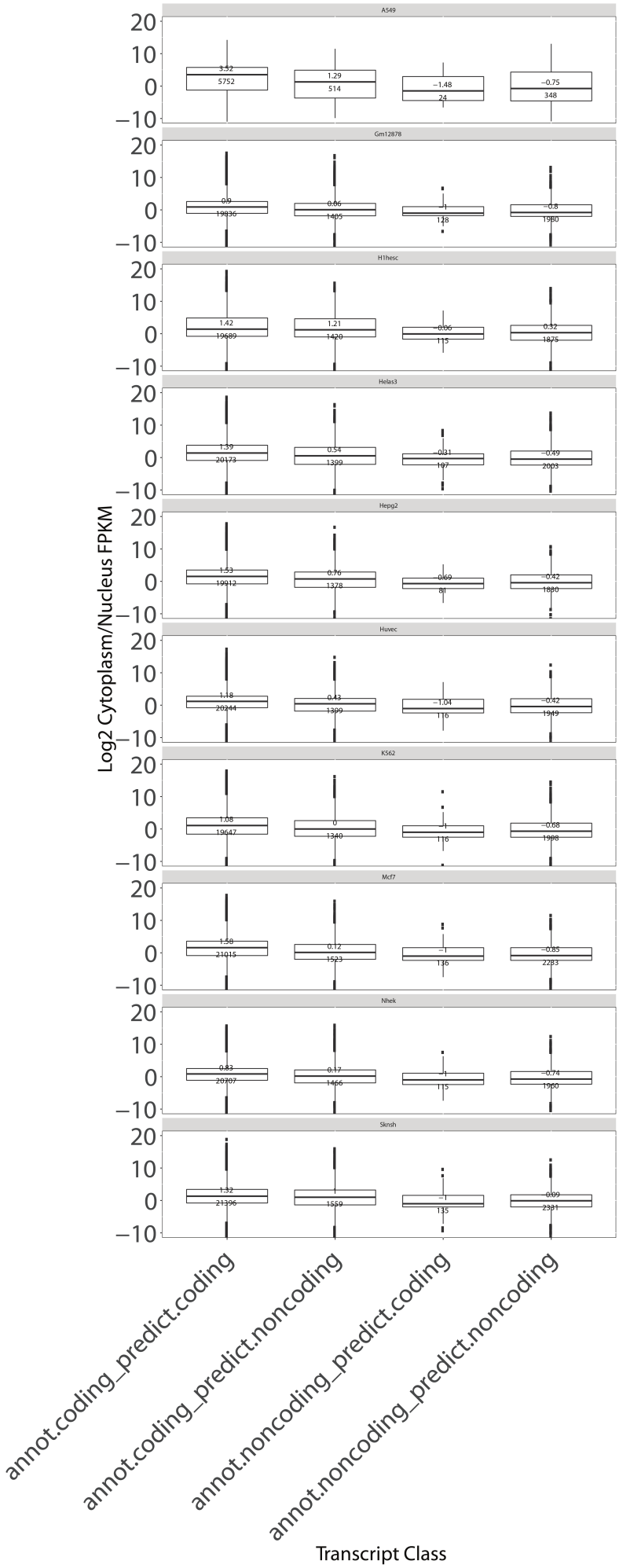
Supplementary Figure S36



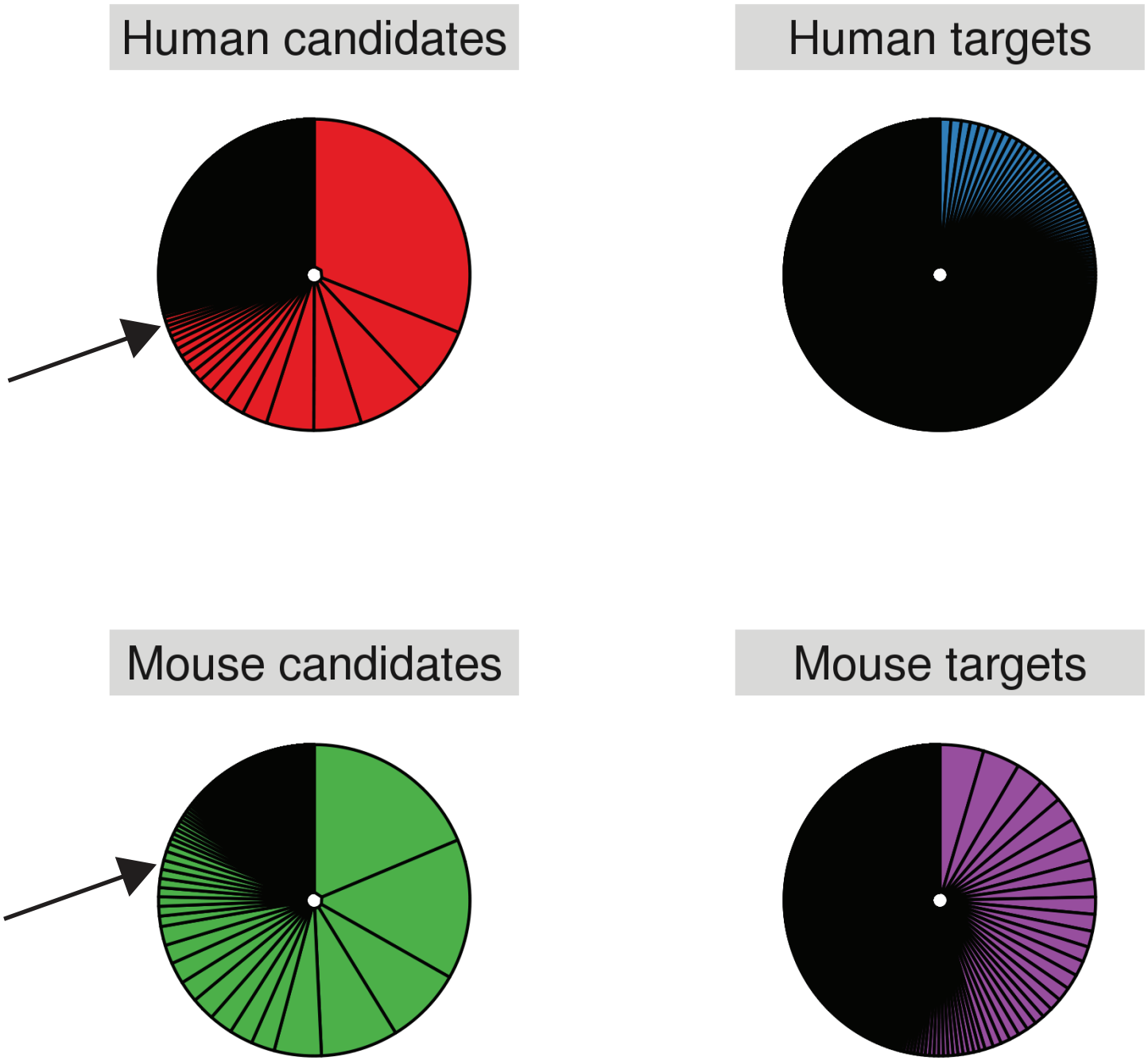
Supplementary Figure S37



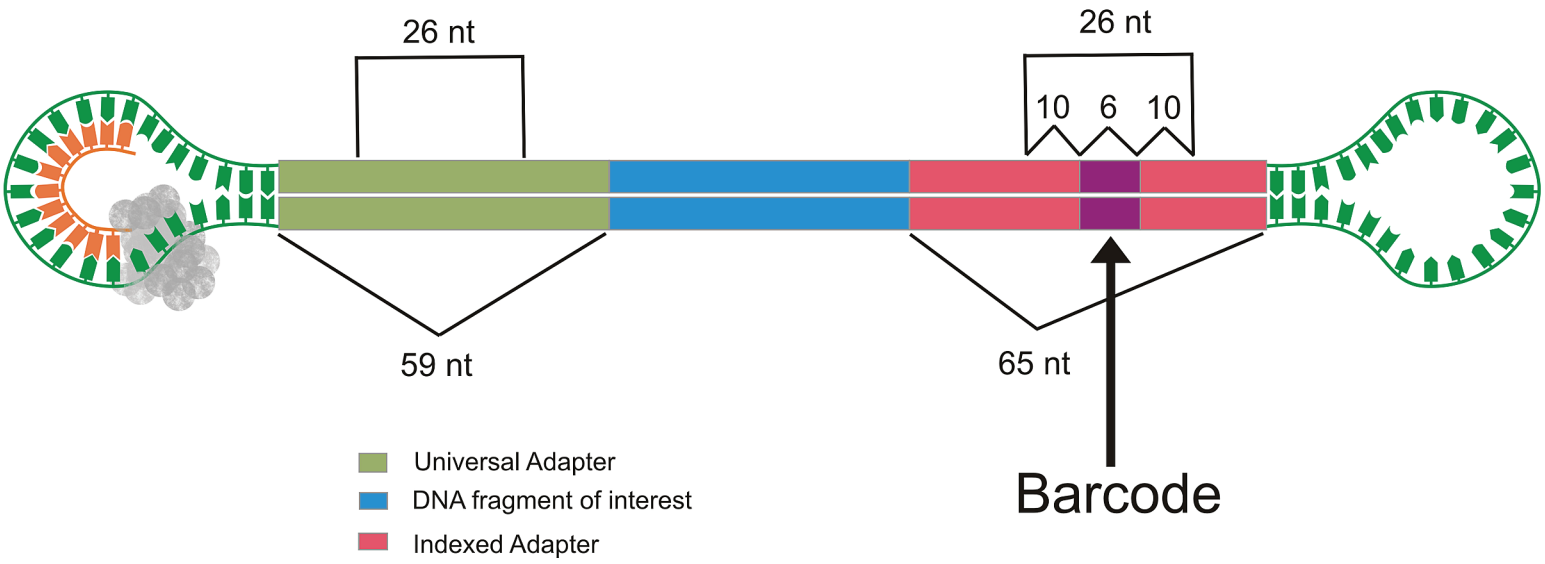
Supplementary Figure S38



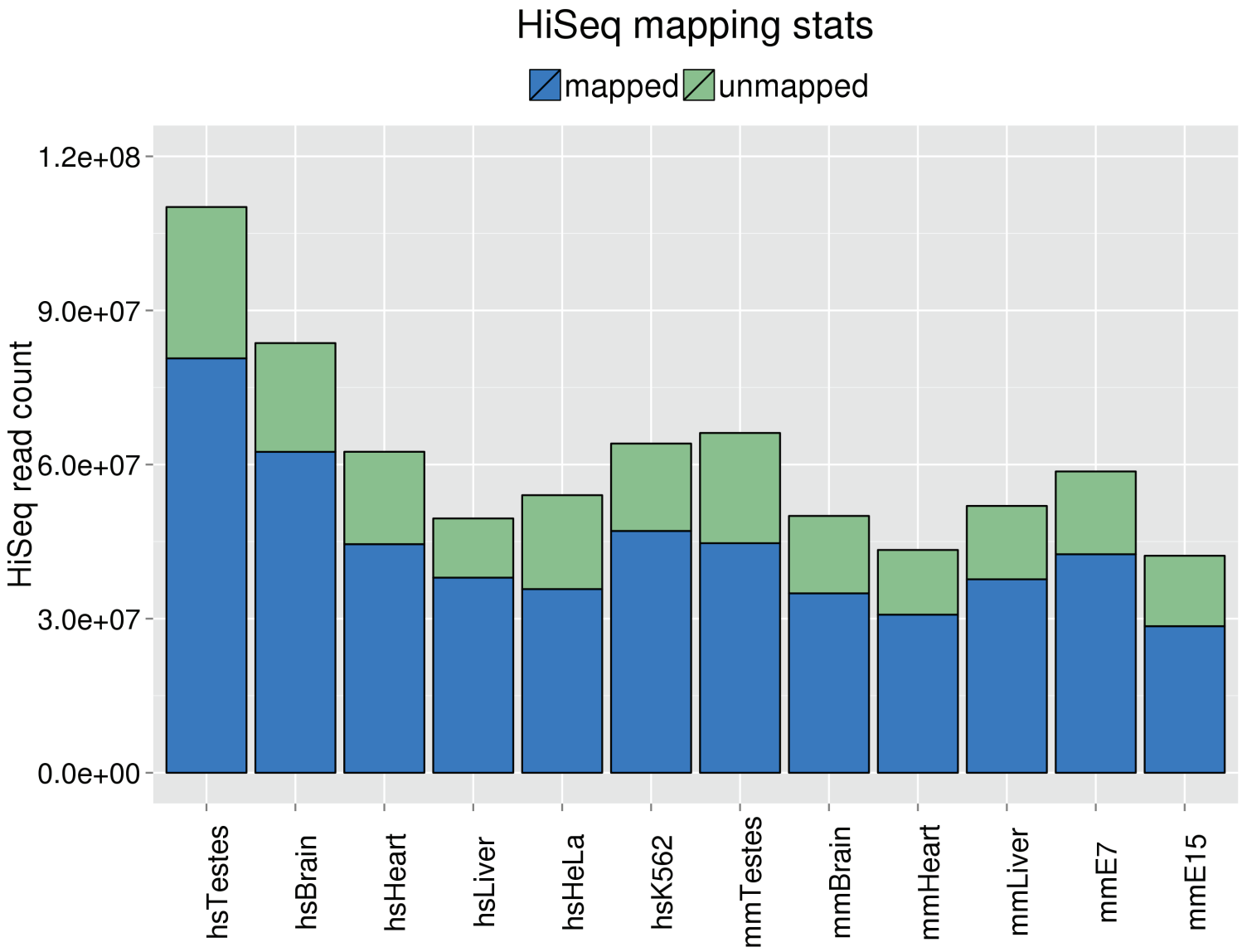
Supplementary Figure S39



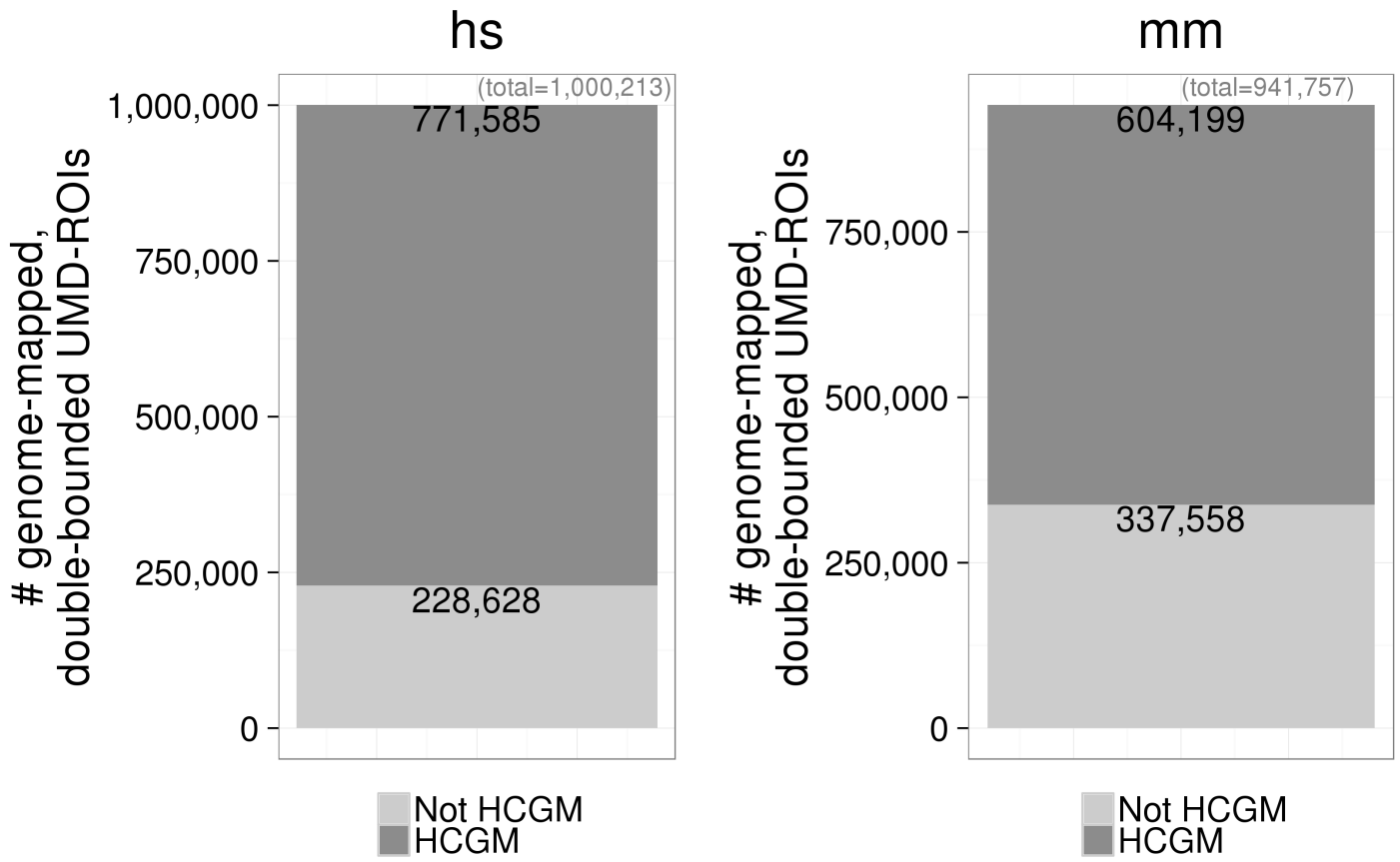
Supplementary Figure S40



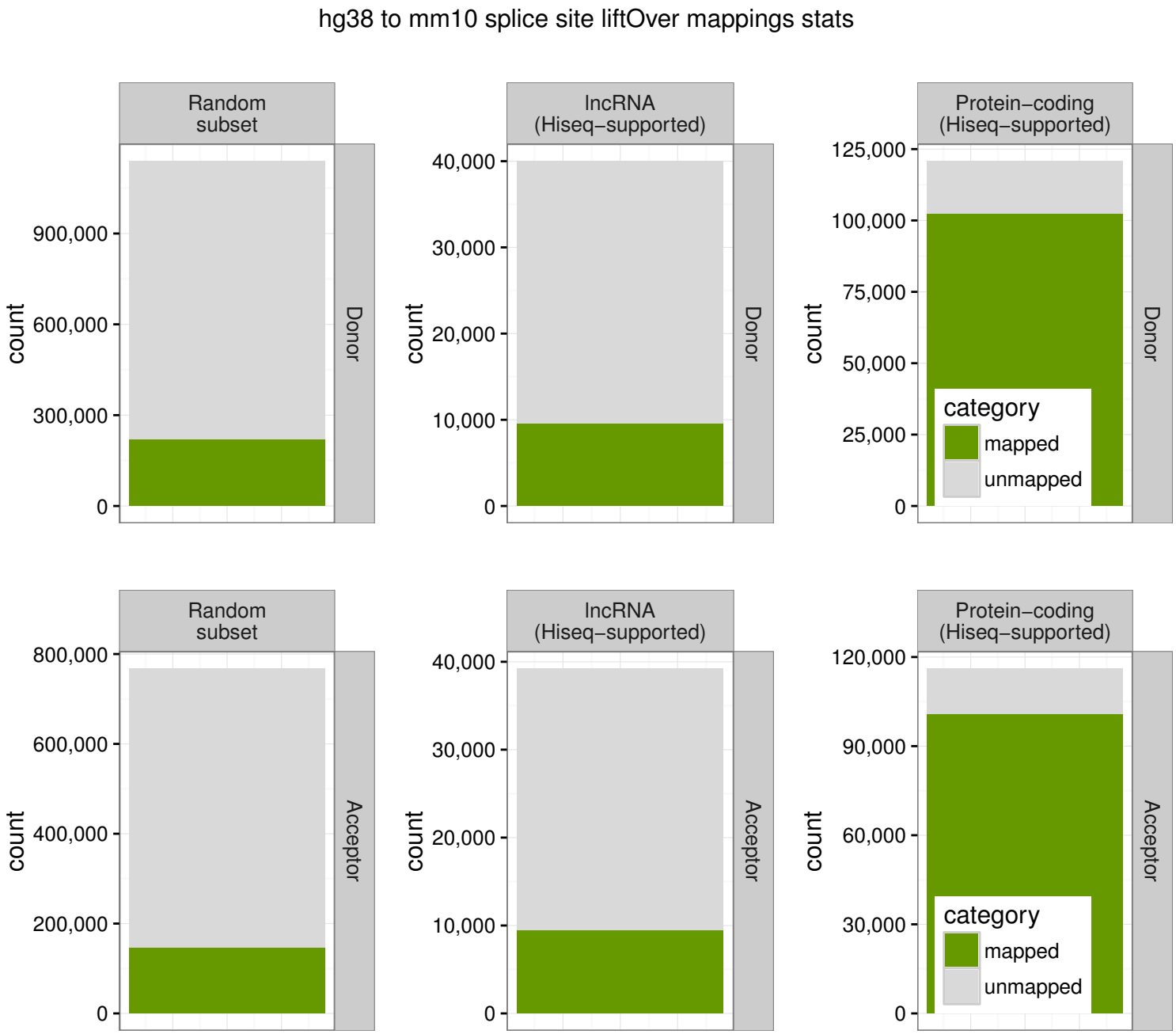
Supplementary Figure S41



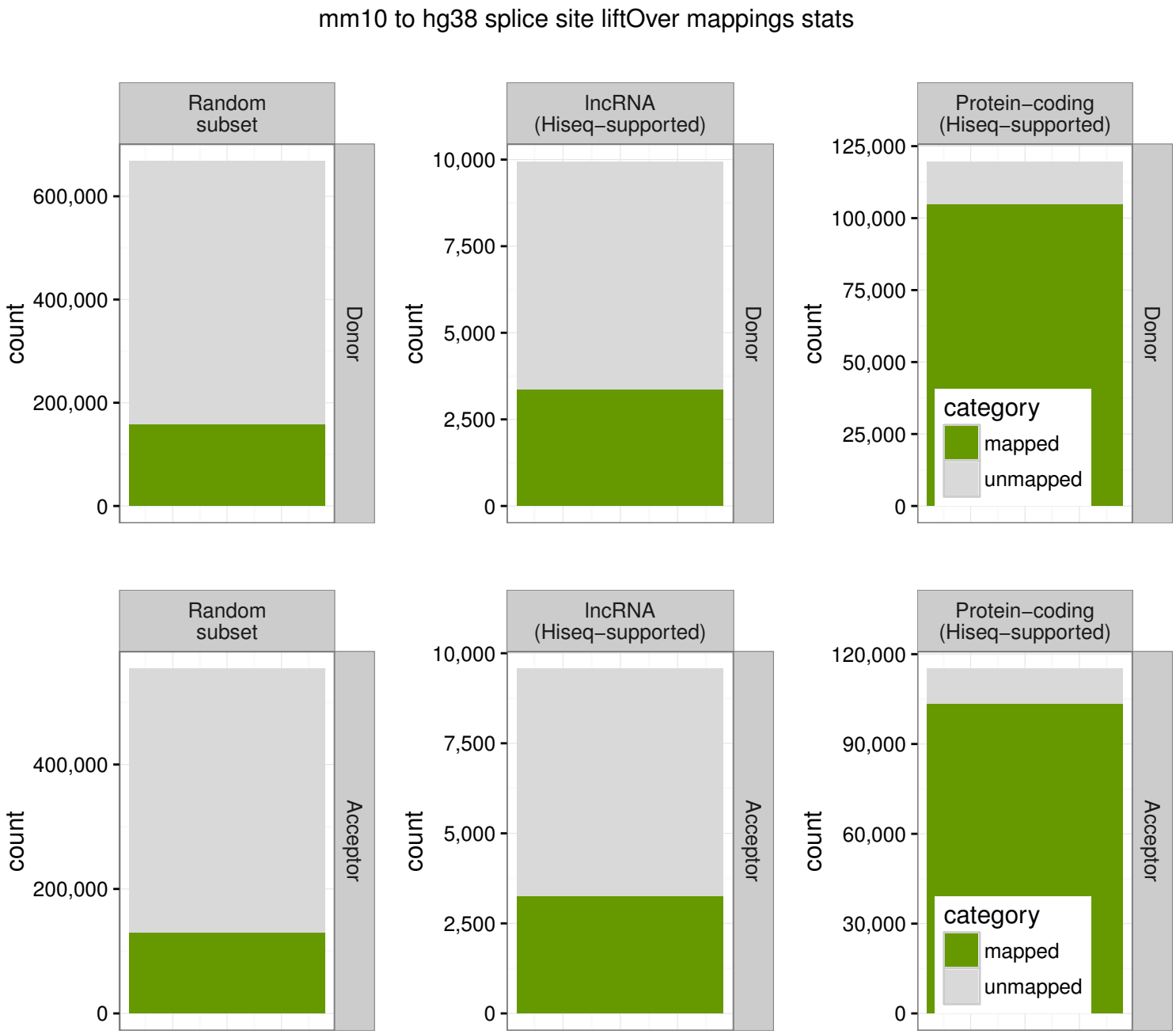
Supplementary Figure S42



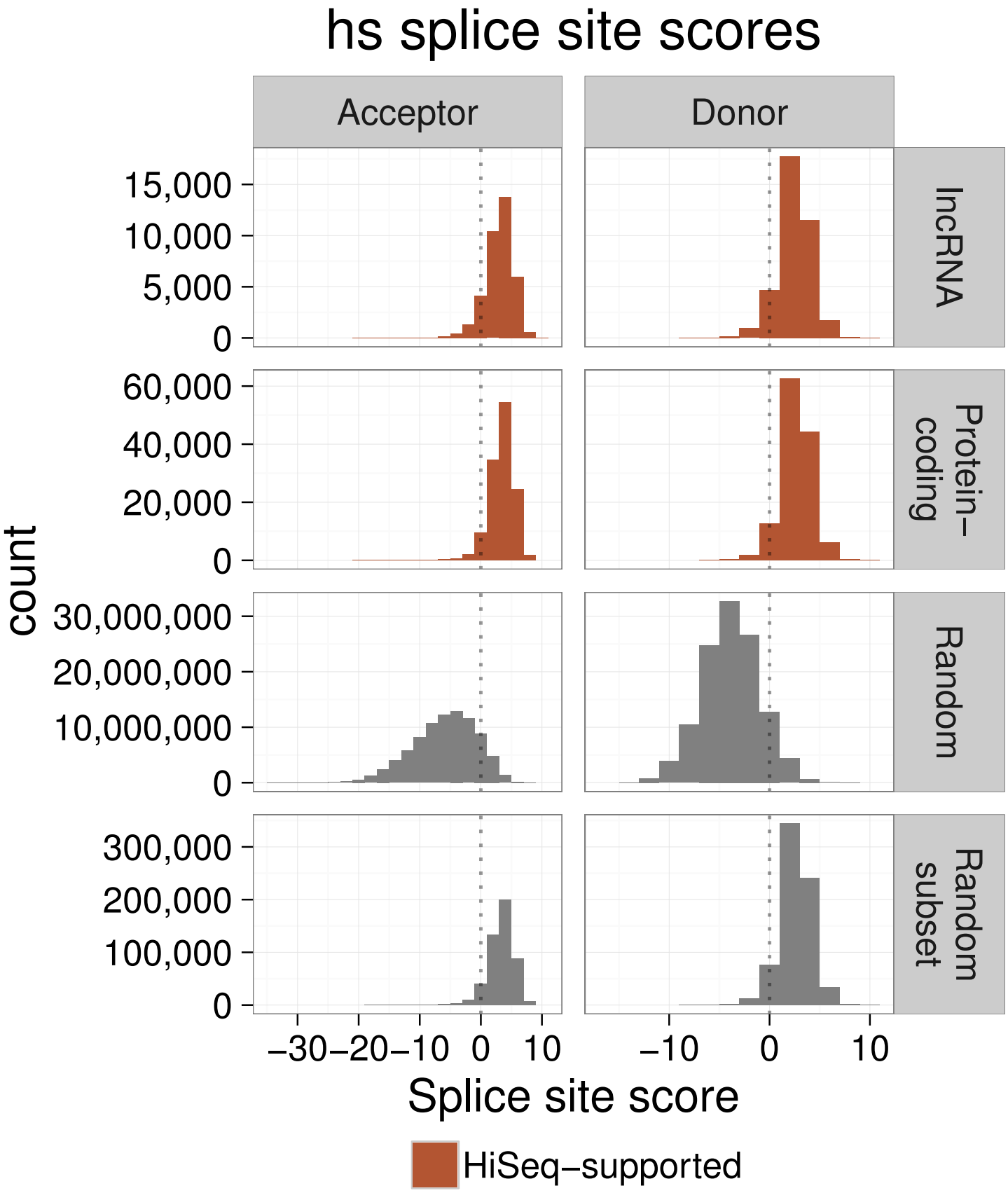
Supplementary Figure S43



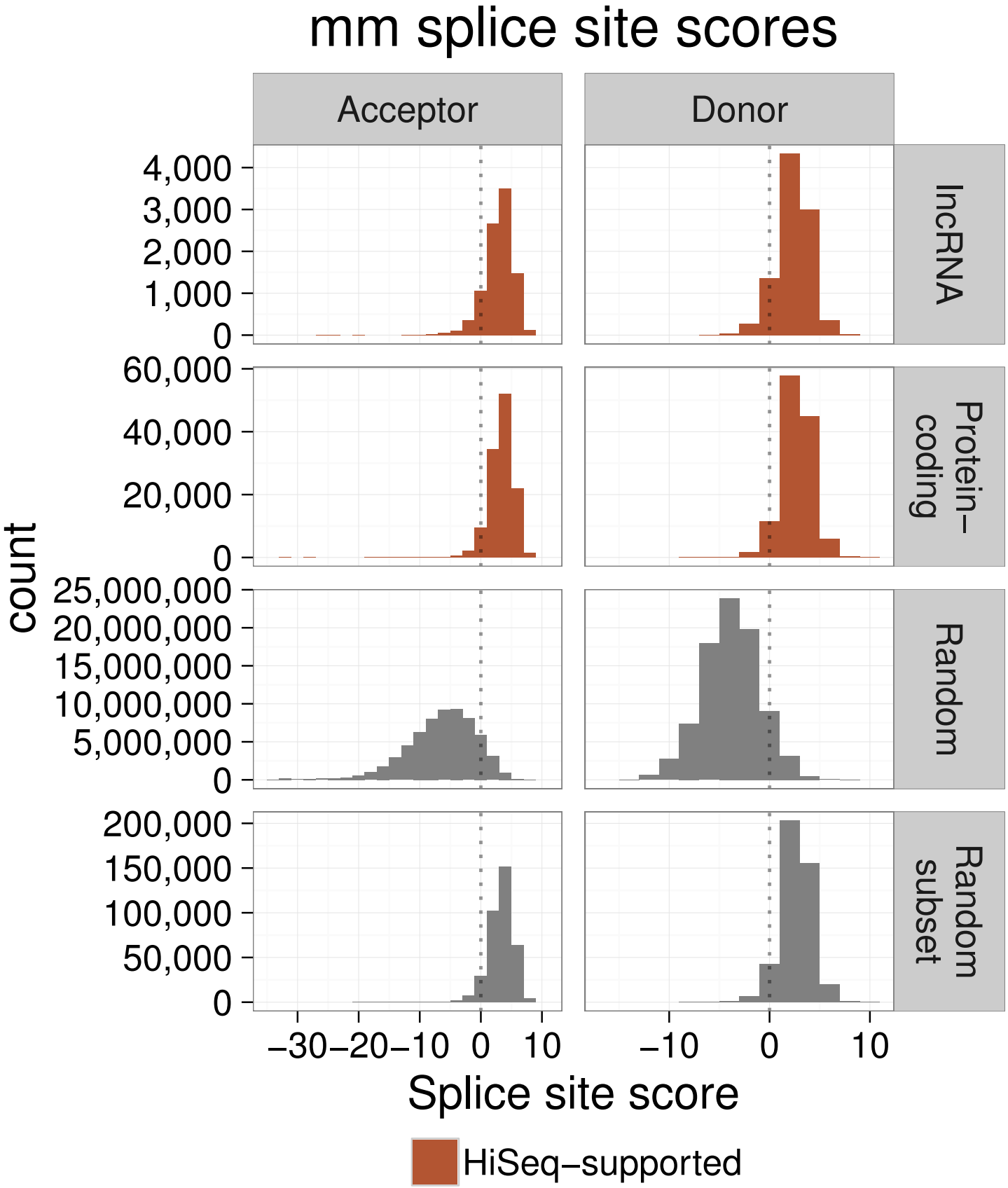
Supplementary Figure S44



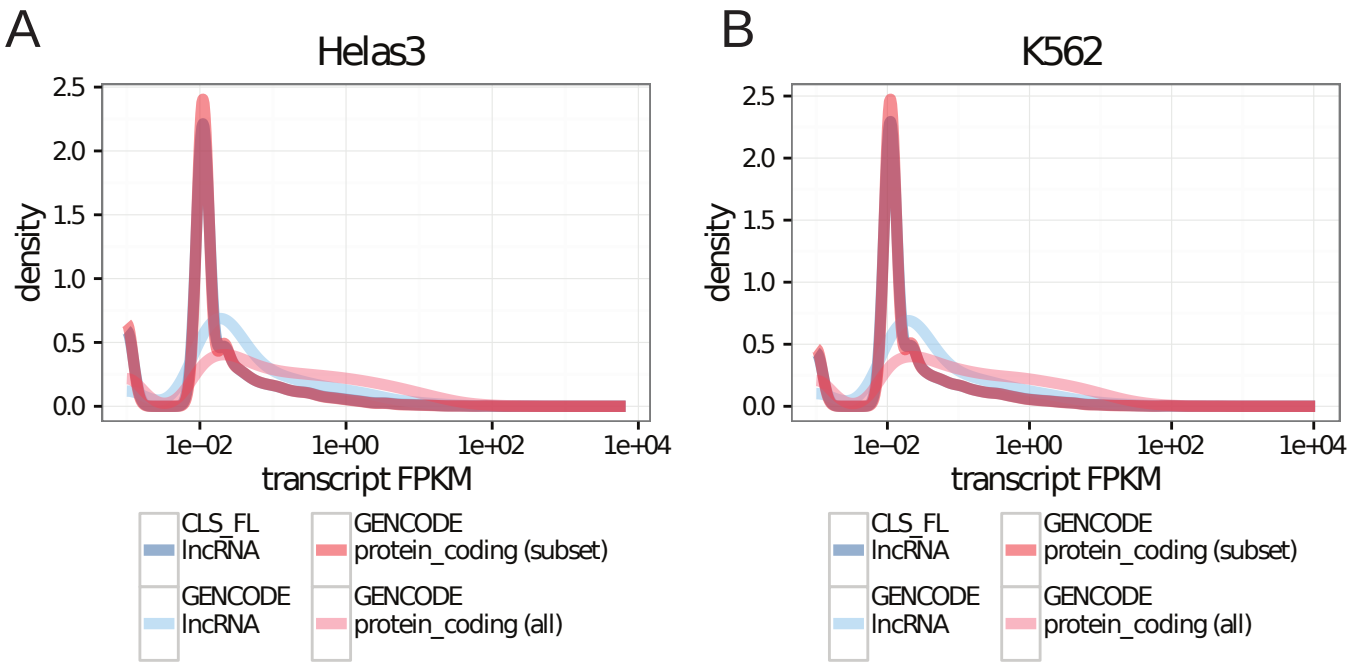
Supplementary Figure S45



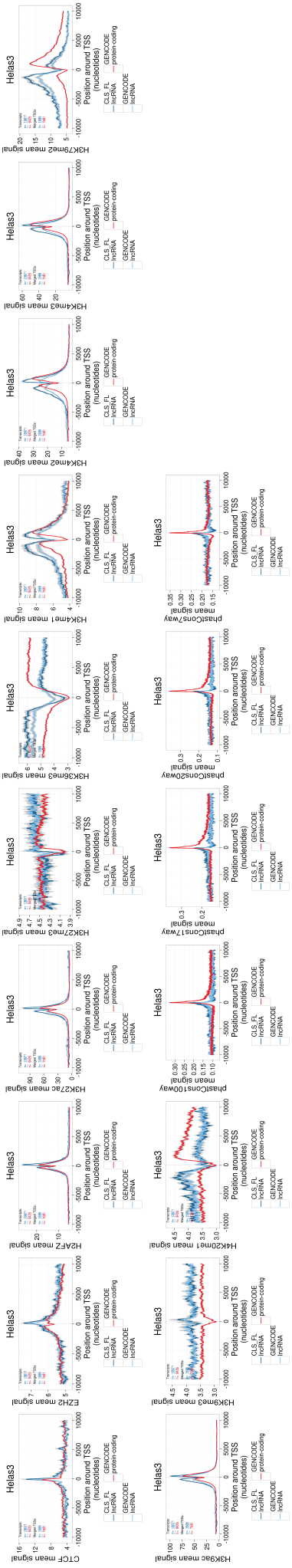
Supplementary Figure S46



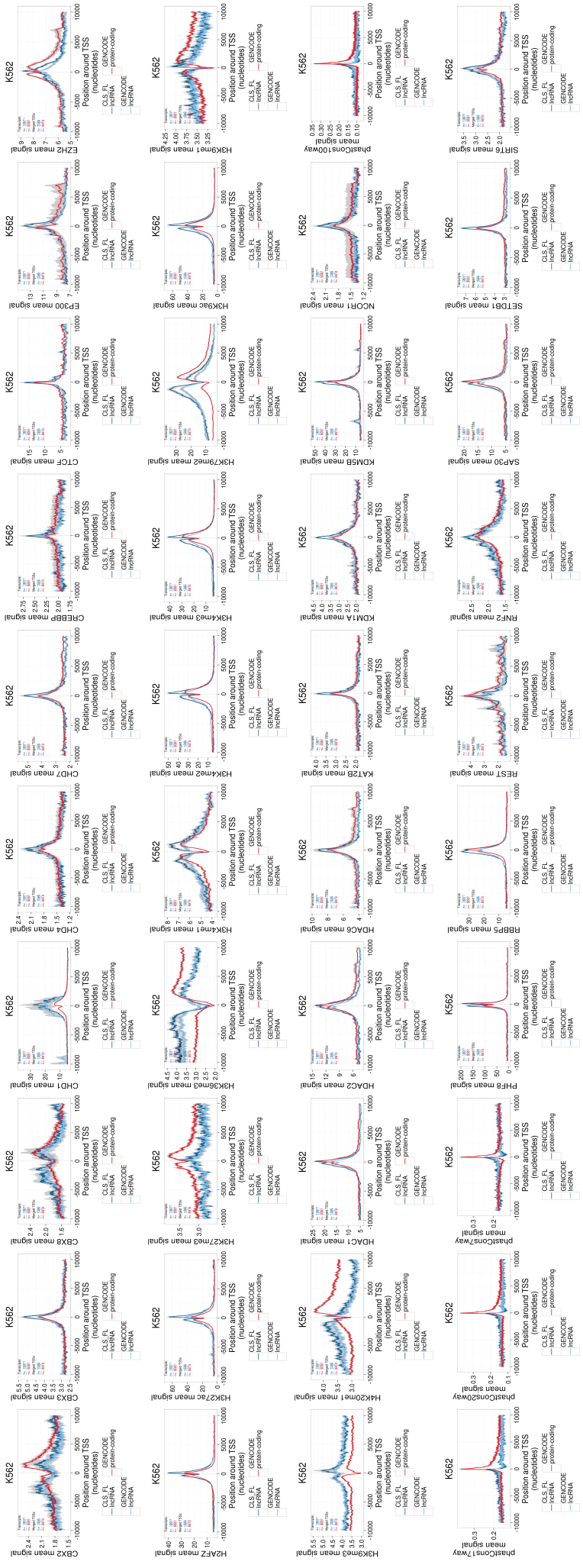
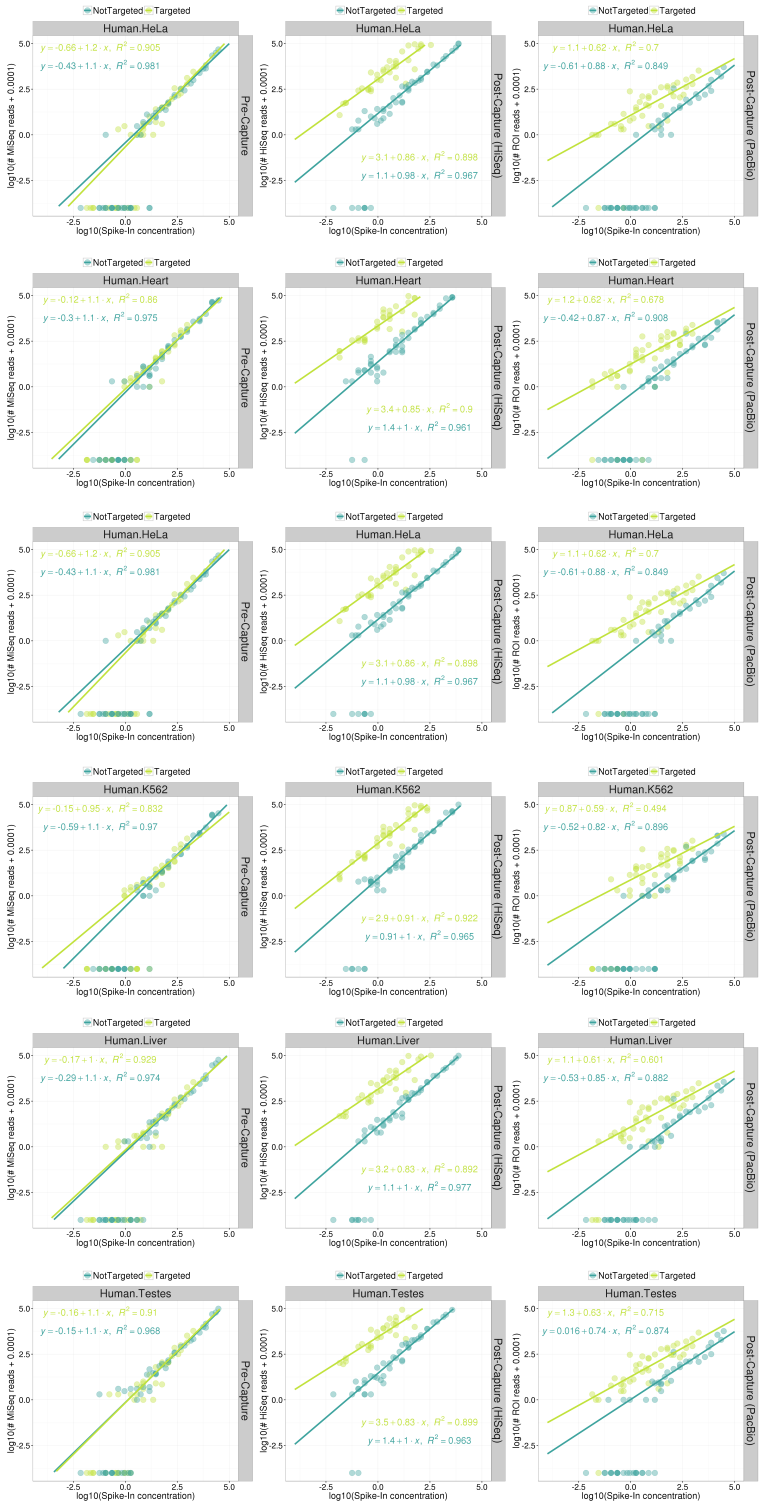
Supplementary Figure S47



Supplementary Figure S48



Supplementary Figure S49



Supplementary Figure S50