Supplementary Note

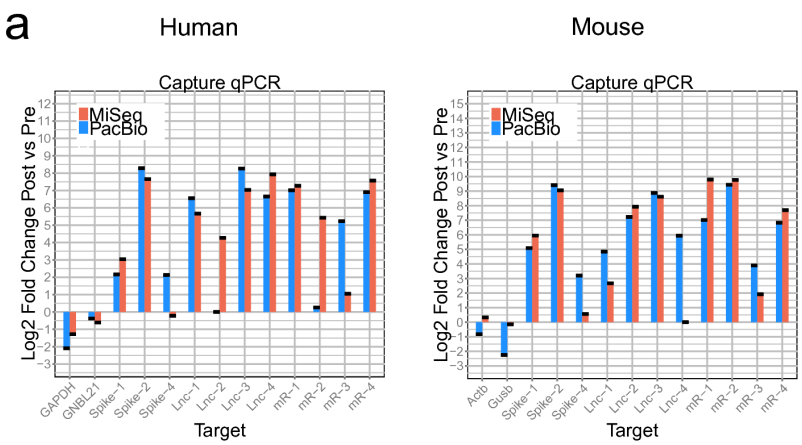
From Lagarde *et al.*, **High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing** (DOI: https://doi.org/10.1101/105064)

Contents

Supplementary Figures

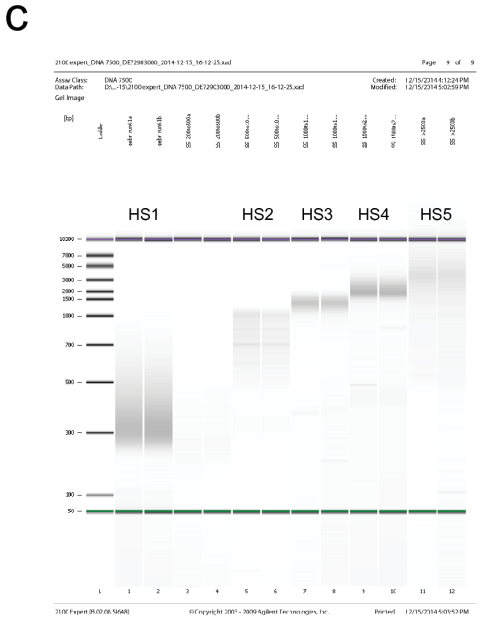
Note

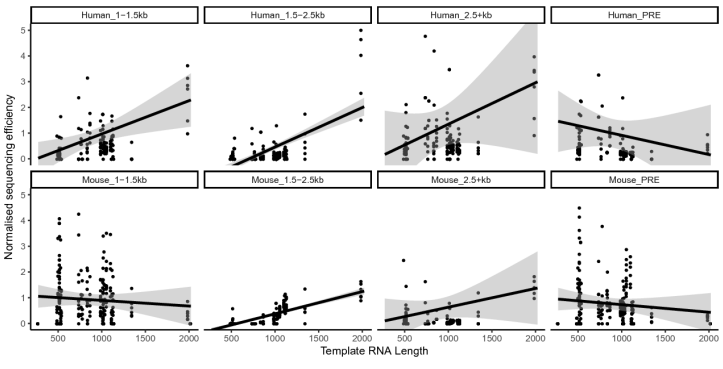
Each supplementary figure is followed by its title and caption on the following page. Links and page numbers in the Table of Contents refer to figure captions.

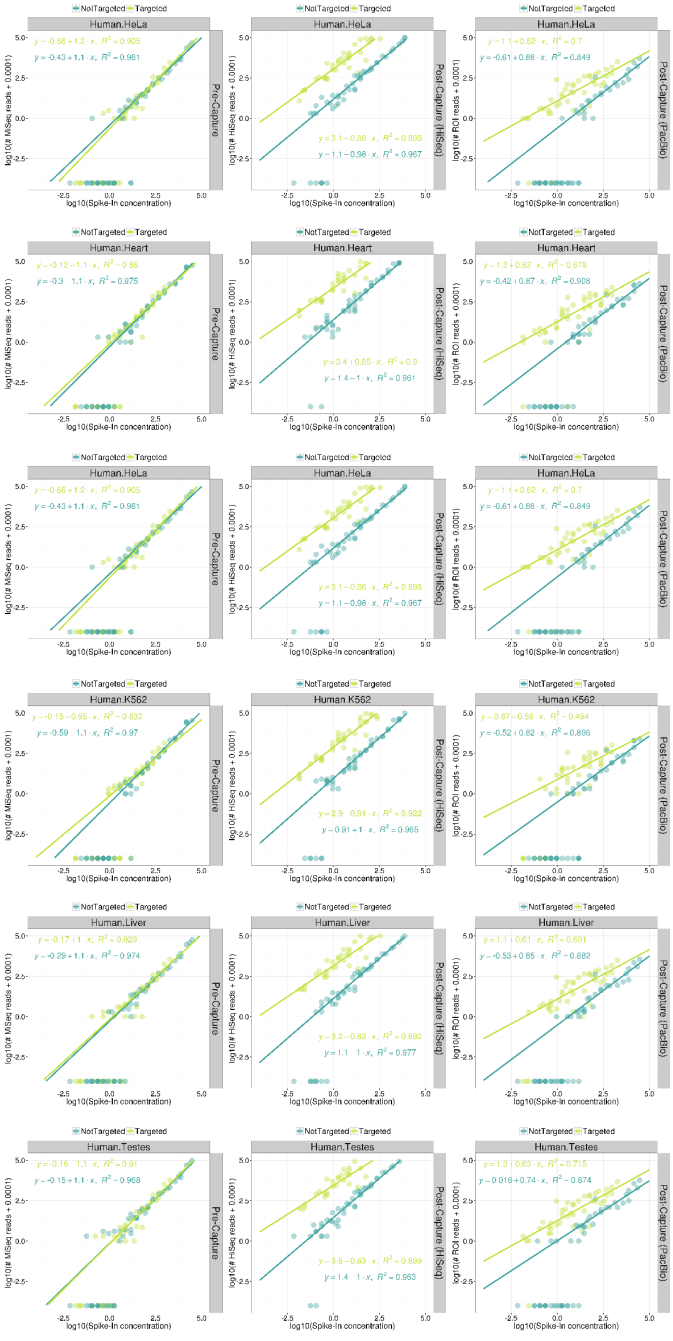


b

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







Supplementary Figure 1

Supplementary Figure 1: RNA Capture enrichment, PacBio cDNA Size Fractionation and Sequencing

(a) qPCR validation of enrichment

Quantitative PCR was performed to assess capture performance. Templates were pooled cDNA, before and after capture. Separate amplifications were performed on cDNA prepared for MiSeq (fragmented) or PacBio (full-length). Primers were designed to a selection of target sequences: GAPDH/ GNBL21/ Actb/ Gusb are housekeeping mRNAs; Spike-in 1&2 were targeted in the capture library; Spike-in 4 was present but not targeted. Lnc 1-4 and mR 1-4 refer to randomly selected, targeted lncRNAs and mRNAs respectively. Note that spike-ins are common to human and mouse experiments. y-axis shows the value of (Ct-POST - Ct-PRE), where Ct refers to the PCR threshold cycle. PCRs were carried out in technical triplicate and the mean is shown. Also shown are error bars denoting the standard deviation.

(b) Size fractionation of captured cDNA

cDNAs were size selected into five ranges. The last three were selected for subsequent sequencing.

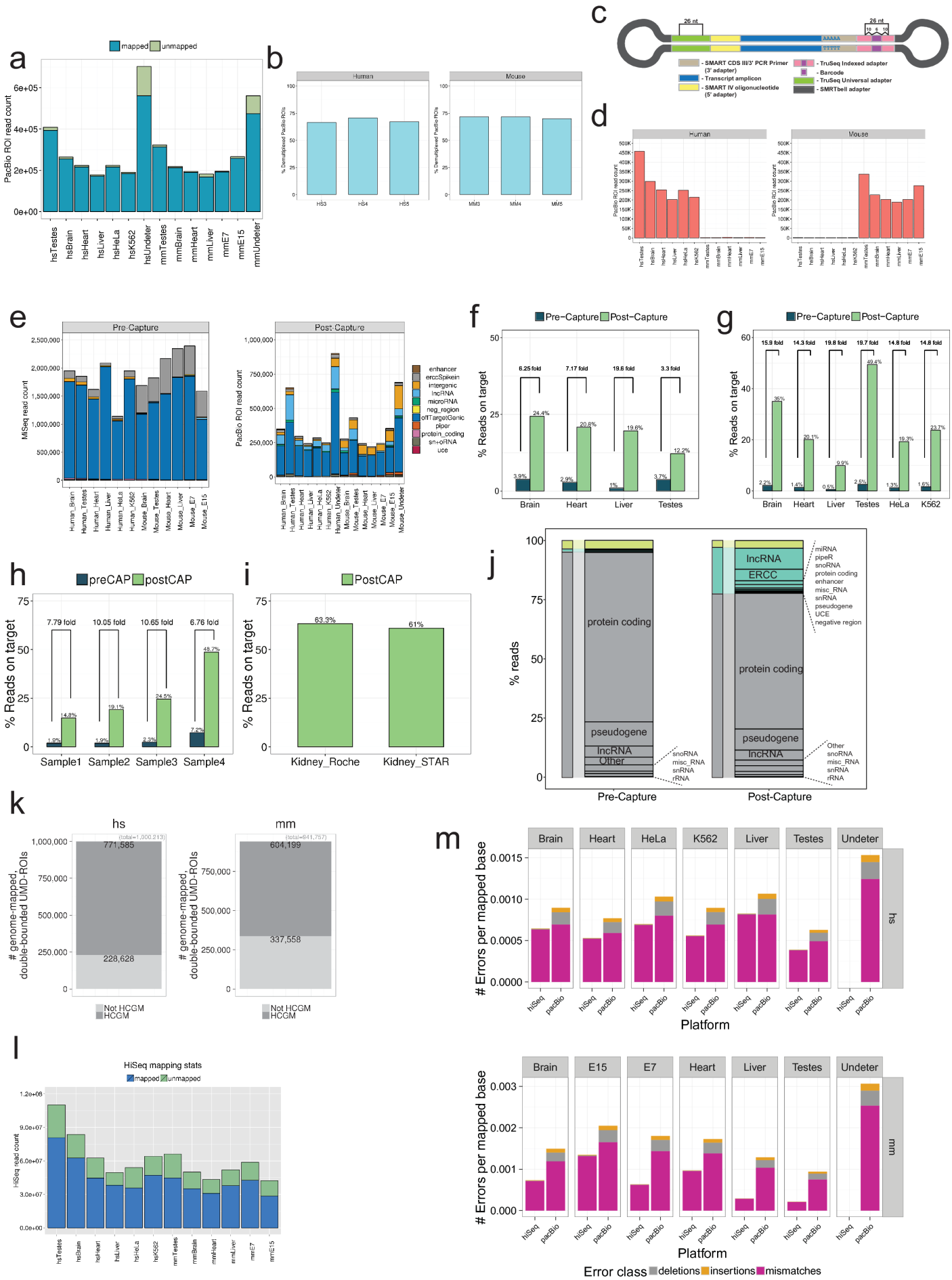
(c) Agarose gel electrophoresis of size-selected post-capture cDNA samples

(d) Template length-dependence of PacBio sequencing

Each panel shows data from a different sequencing library. Top row: human; bottom row: mouse. The first three panels of each row show post-capture PacBio data for indicated size-selected fractions. The fourth panels of each row show similar data for pre-capture MiSeq data, which did not undergo size selection. Every point represents one of the synthetic ERCC spike-in RNA sequences added to samples prior to library preparation. x-axes show the length of these sequences. y-axes show the normalised sequencing efficiency: the sequencing reads per molecule, normalised to length and sequencing depth. Details may be found in the Methods. Lines show the best linear fit, and shading indicates the 95% confidence interval.

(e) Spike-in detection curves for individual human tissues

Data are analogous to Figure 2e, but broken down by tissues (rows). First column: pre-capture samples, with MiSeq sequencing; Second column: post-capture samples, with HiSeq sequencing; Third column: post-capture samples, with PacBio sequencing. 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 (light green), while the remaining 50 were not (dark green). 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.



Supplementary Figure 2

Supplementary Figure 2: Sequencing library structure and statistics

(a) Read mapping statistics

Shown are the numbers of reads, broken down by originating sample, which could be mapped to the genome. "Undeter" refers to reads from which the barcode could not be confidently identified. "hs": human; "mm": mouse.

(b) ROI demultiplexing efficiency

The *y*-axis indicates the fraction of reads in each pooled sample whose sample of origin could be inferred based on hexamer barcodes. *x*-axis columns indicate the three size-selected fractions of each species.

(c) Schematic structure of PacBio reads and library adapters

Indexed adapters carry a unique 6-nt barcode, specific to the originating sample. The adapter and barcode sequences are available in Supplementary Data 3.

(d) Demultiplexed ROIs by sample of origin

Undetermined reads are not shown. "hs": human; "mm": mouse.

(e) Capture enrichment by tissue

Pre-Capture data was generated using MiSeq reads of pooled cDNA prior to capture, while PostCapture data represents PacBio ROI reads. "Undeter" refers to reads from which the barcode could not be confidently identified, and hence from an undetermined sample. Colours refer to the biotype of the feature to which the reads map. These feature classes are composed of targeted features (Figure 1b) or off-target features (either genic in dark blue, or intergenic in orange). Most off-target genic features are protein-coding genes. Notice the increase in representation of targeted features (mainly lncRNAs, light blue) in Post-capture compared to Pre-capture samples.

(f-g) Capture performance in individual tissues for Capture Short-Seq (CSS, data from Clark *et al.*, 2015) (f) and CLS (g)

The *y*-axis shows the percent of all mapped ROIs originating from targeted regions. Enrichment is defined as the ratio of this value in Post- and Pre-capture samples. Note that pre-capture rates in (f) were estimated using pre-capture MiSeq libraries generated in the present study.

(h-i) Comparing capture protocols shows that long cDNA targets yield lower capture efficiency

Sample1: Original CLS protocol (as used and described here), PolyA-selected, unfragmented. Sample2: Improved CLS protocol (see Methods), PolyA-selected, unfragmented. Sample 3: Improved CLS protocol, Total RNA, unfragmented. Sample 4: Roche SeqCap RNA protocol, Total RNA, fragmented. **(h)** Performance statistics for the four captures. Compare the on-target rates for Post-capture ("postCAP") material in Sample2/Sample4 and Sample3/Sample4. **(i)** Performance statistics for in-house data provided by Roche, for SeqCap capture of fragmented kidney cDNA.

(j) Breakdown of sequenced reads by gene biotype, pre- (left) and post-capture (right), for mouse

Colours denote the on/off-target status of the genomic region from which the reads originate, namely: Grey: reads originating from annotated but not targeted features; green: reads from targeted features, including lncRNAs; yellow: reads from unannotated, non-targeted regions. The ERCC class comprises only those ERCC spike-ins that were probed in this experiment. Note that when a given read overlapped more than one targeted class of regions, it was counted in each of these classes separately. Equivalent human data are found in Figure 2c.

(k) Results of the selection of High-Confidence Genome Mappings (HCGMs)

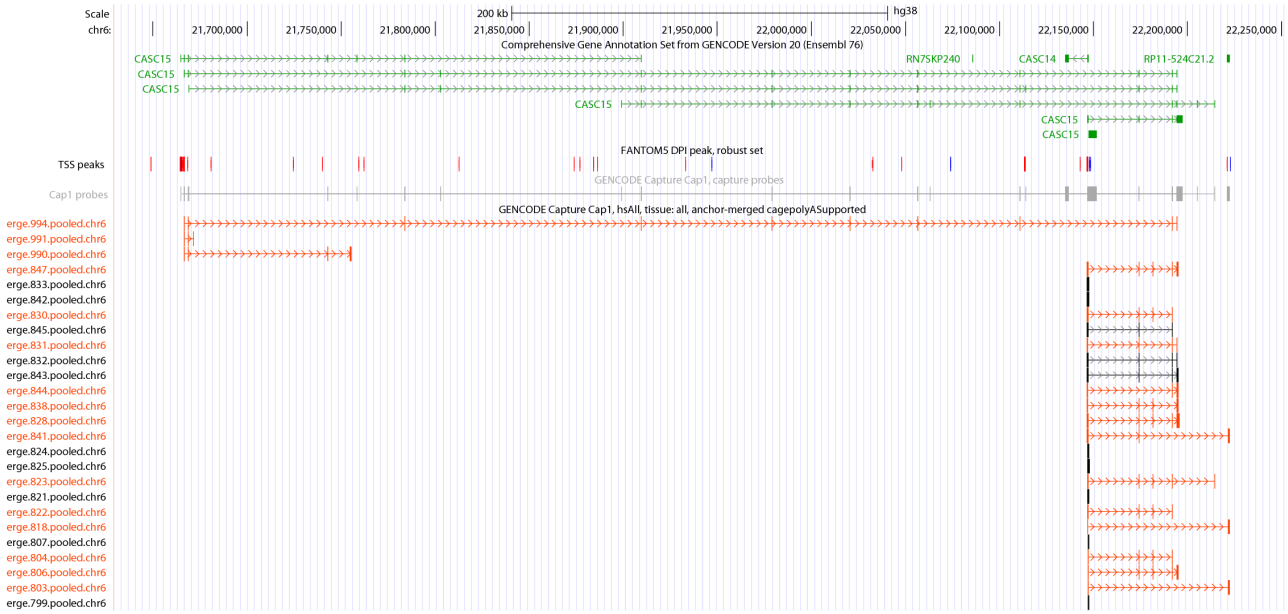
The number of double-bounded ROIs with and without HCGM (see definition in Methods) is represented in human ("hs", left panel) and mouse ("mm", right panel). The total number of double-bounded ROIs is reported at the top of each bar, in light grey.

(l) Post-Capture HiSeq read mapping statistics

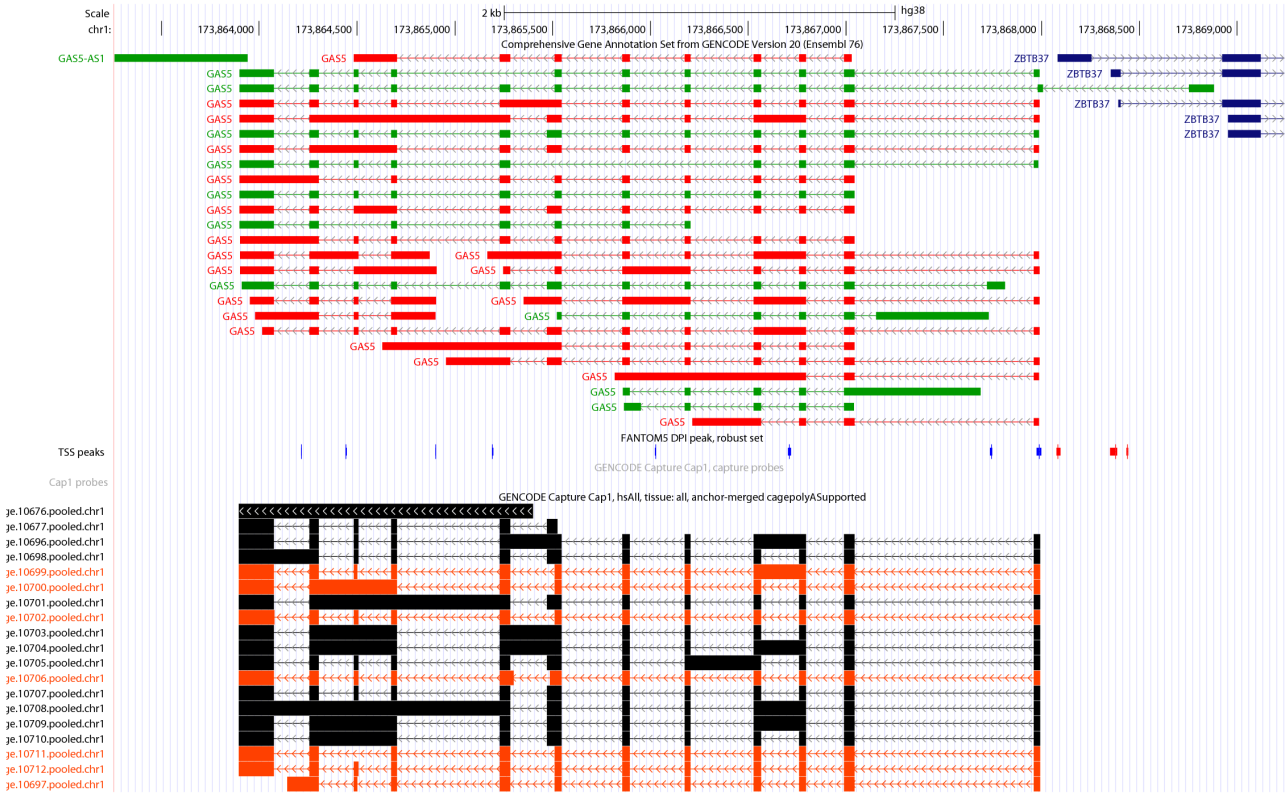
(m) Sequencing error rates in human (top) and mouse (bottom) samples

The rate of sequencing error per sample and sequencing platform is represented on the *y*-axis. Sequencing errors are subdivided into mismatches (magenta), deletions (grey) and insertions (orange) with respect to the genome reference. The top of each bar corresponds to the global error rate in each library. "Undeter": undetermined reads, *i.e.*, non-demultiplexed (not available for HiSeq libraries).

CASC15 / ENSG00000272168



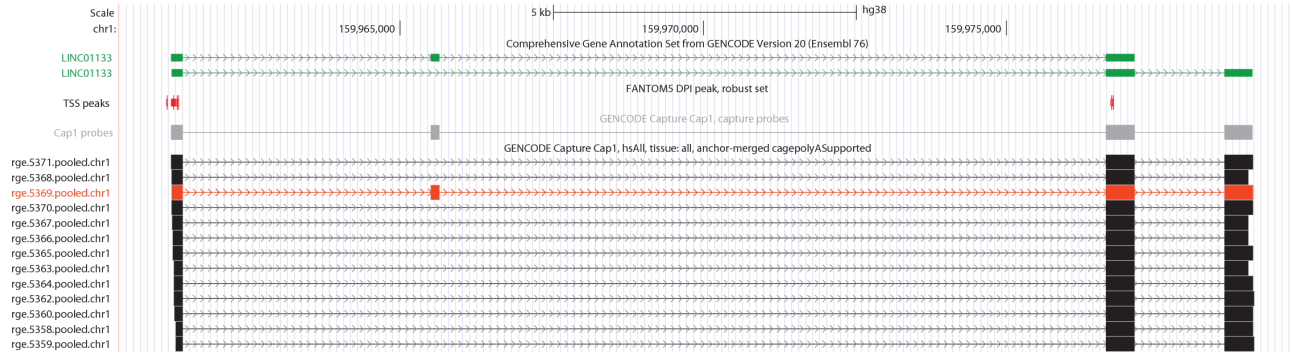
GAS5 / ENSG00000234741



Supplementary Figure 3

Supplementary Figure 3: Examples of known lncRNAs with changes in their annotated structures

LINC01133 / ENSG00000224259



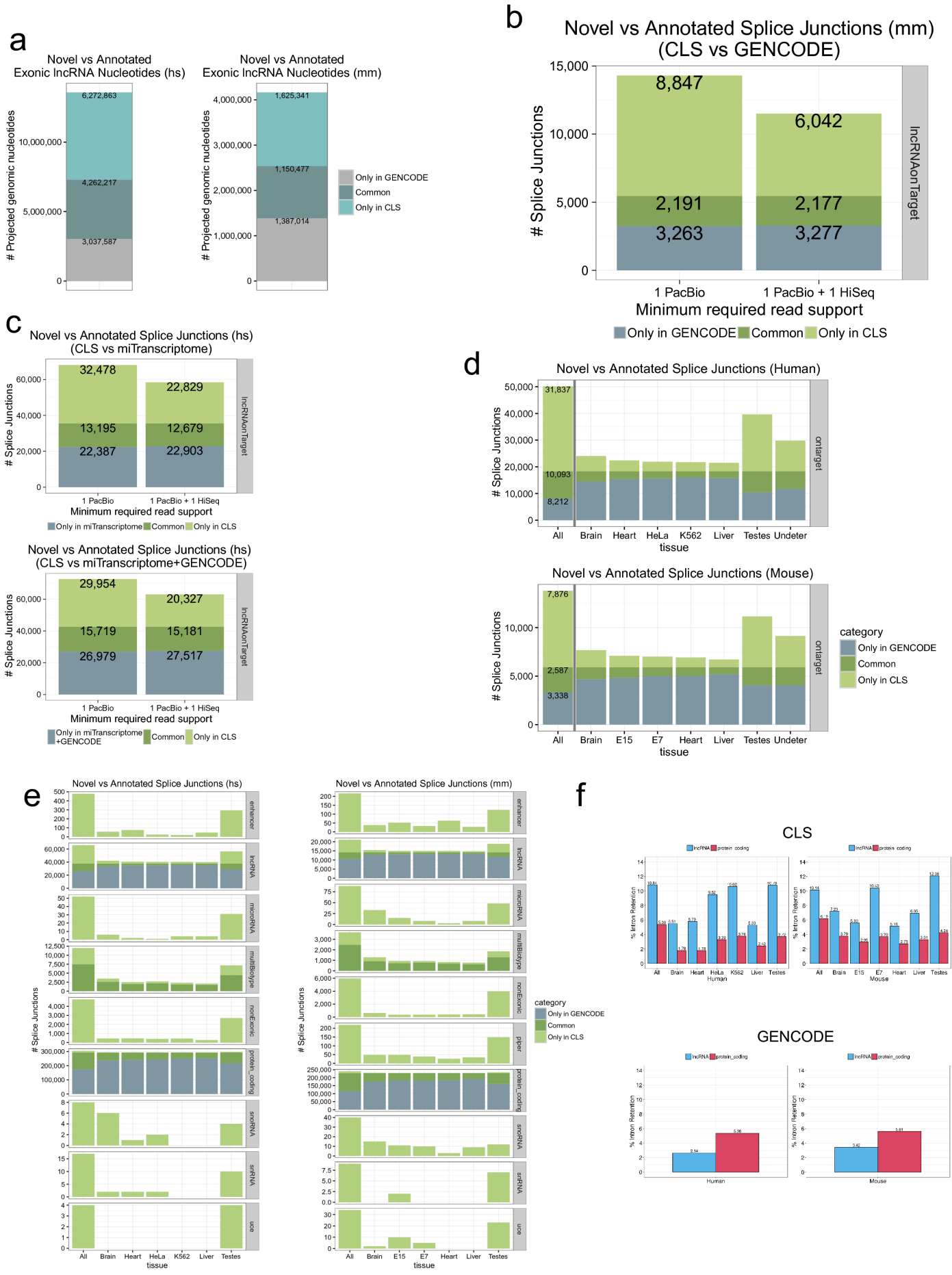
Supplementary Figure 4

Supplementary Figure 4: Examples of known lncRNAs with almost no change in their annotated structures

UCA1 / ENSG00000214049

Supplementary Figure 5

Supplementary Figure 5: Examples of known lncRNAs with changes in their annotated structures (II)



Supplementary Figure 6

Supplementary Figure 6: Exon and Intron Discovery by CLS

(a) Novel exonic bases discovered by CLS

Figures show the number of nucleotides that (i) are annotated in the targeted GENCODE lncRNA annotation but not detected ("Only in GENCODE"), (ii) are annotated and detected by CLS ("Common") or (iii) detected nucleotides that are not present in GENCODE and hence novel ("Only in CLS"). Left: human; Right: mouse. Note that nucleotide counts are from collapsed (merged annotations) and hence are non-redundant. Data on novel nucleotides only refer to ROIs that map to targeted lncRNA loci.

(b) Discovery of splice junctions (SJs) in targeted lncRNAs for mouse

GENCODE v.M3 is used as a reference. The y-axis denotes counts of unique SJs. Only "on-target" junctions originating from probed lncRNA loci are considered. Grey represents annotated SJs that are not detected. Dark green represents annotated SJs that are detected by CLS. Light green represent novel SJs that are identified by CLS but not present in the annotation. 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. Equivalent human data is in Figure 3b.

(c) Discovery of splice junctions (SJs) in targeted lncRNAs for human (comparison with *miTranscriptome*)

Novel splice junction discovery with respect to *miTranscriptome* (top panel) and the union of GENCODE and *miTranscriptome* (bottom panel) SJ sets. The figure layout and color legend is analogous to (b).

(d) Novel splice junctions by tissue in targeted loci

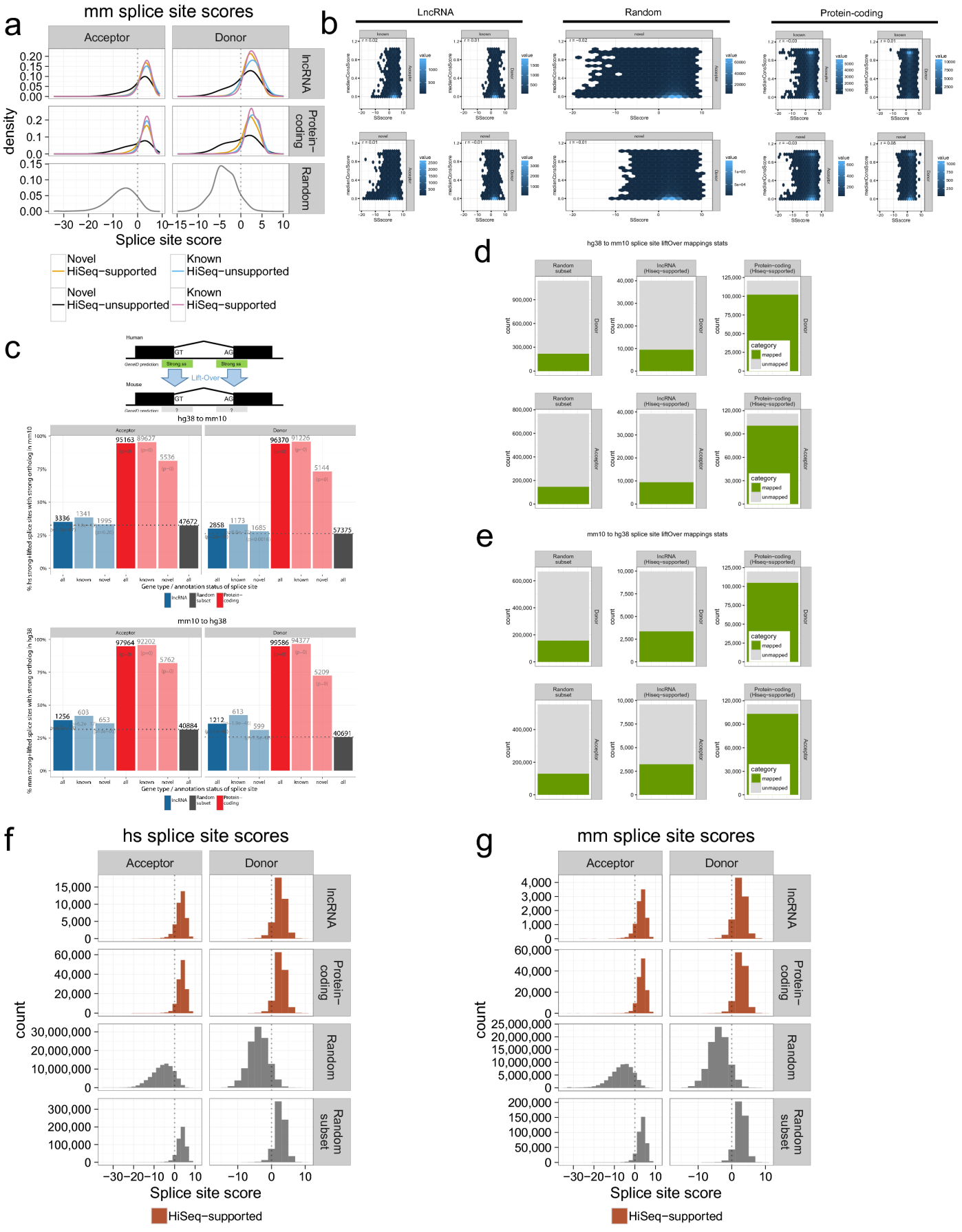
Figures display the number of splice sites discovered by CLS and compared to GENCODE annotations, broken down by tissue. Only high-confidence, HiSeq short read-supported CLS junctions are considered.

(e) Splice junction discovery statistics by tissue and biotype

Figures display the number of splice junctions discovered by CLS and compared to GENCODE annotations in human (left panel) and mouse (right panel), broken down by tissue and ROI biotype. Only high-confidence, HiSeq short read-supported CLS junctions are considered.

(f) Analysis of intron retention (IR) rates

Top panel: Proportion of transcripts with at least one retained intron in lncRNA and protein-coding CLS transcripts in human (left) and mouse (right). Bottom panel: Proportion of GENCODE lncRNA and protein-coding transcripts with at least one retained intron in human and mouse samples. Red indicates IR rate in lncRNAs, while blue indicates IR rate in protein coding transcripts.



Supplementary Figure 7

Supplementary Figure 7: Splice site (SS) motif and evolutionary analysis

(a) Splice site motif quality in mouse

Panels plot the distribution of predicted SJ strength, for acceptors (left) and donors (right). Splice site strength was computed using position weight matrices from *geneid*. Data are shown for non-redundant SJs from CLS transcript models from targeted lncRNAs (top), protein-coding genes (middle), or background distribution sampled from randomly-selected AG (acceptor-like) and GT (donor-like) dinucleotides (bottom). Analogous human data can be found in Figure 3c.

(b) Evolutionary conservation of known and novel splice sites

Panels show the distribution of splice sites (broken down by donor and acceptor sites) as a function of base-level nucleotide conservation ("medianConsScore", as calculated by PhastCons 100 vertebrate alignments) (y-axis) and predicted splice site strength ("SSscore", as determined by the *geneid* software) (x-axis).

(c) Evolutionary conservation of splice sites

The figures show the rate of conservation of different classes of splice sites (SSs). Conservation is defined by having a high-strength predicted SS at the orthologous site in the other genome. Orthologous regions were obtained from whole-genome alignments. Percentages only relate to those SSs for which an alignment exists (see (d) and (e)), HiSeq-supported, and deemed "strong" (*i.e.*, with a positive *geneid* score) in both the original and target genomes. Upper panel: conservation of human SSs in mouse; Lower panel: conservation of mouse SSs in human. Dark shades: all sites; Light shades: known/novel subsets of SSs. Background sites (referred to as "Random subset") are nearby putative SSs with no evidence of splicing, but with similar *geneid* scores, see (f) and (g). The actual SS counts are specified above each bar. Statistical significance for each set of SSs was estimated using Chi-square test of conserved/non-conserved sites, compared to background sites, and the obtained p-values are reported on each bar.

(d) human (hg38) to mouse (mm10) liftOver mapping statistics

Depicted in green are the number of hg38 strong SSs of each category for which an orthologous site could be found in mm10 using whole-genome alignments.

(e) mouse (mm10) to human (hg38) liftOver mapping statistics

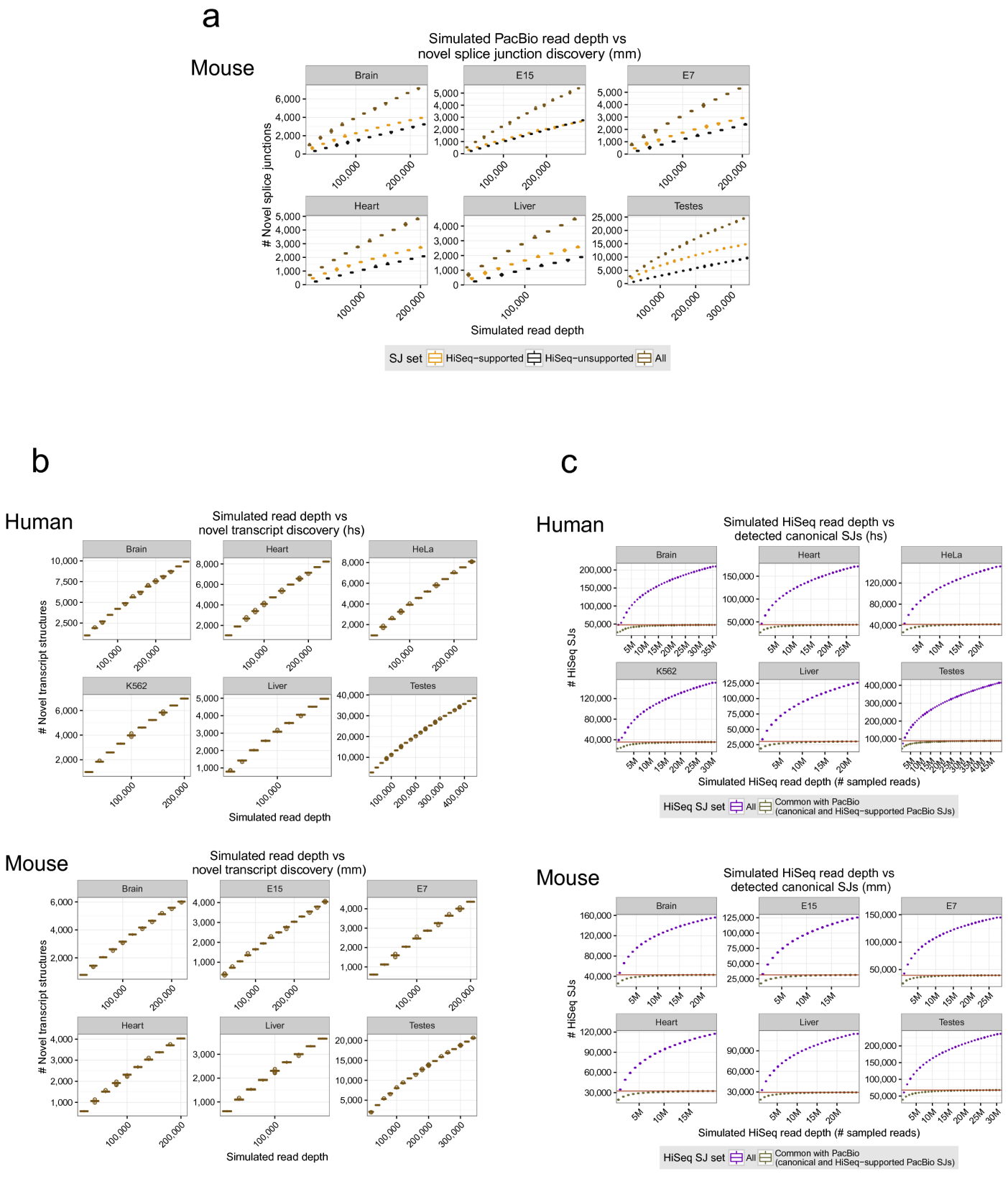
Depicted in green are the number of mm10 strong SSs of each category for which an orthologous site could be found in hg38 using whole-genome alignments.

(f) *geneid* score distribution of human splice sites used in the conservation analysis

The "Random subset" category corresponds to splice sites sampled from the "Random" set, such that its overall score distribution mimics that of lncRNA and protein-coding sites, depicted in the two upper panels.

(g) *geneid* score distribution of mouse splice sites used in the conservation analysis

Legend: see (f).



Supplementary Figure 8

Supplementary Figure 8: Discovery/Saturation analysis

(a) Novel splice junction discovery as a function of sequencing depth in mouse

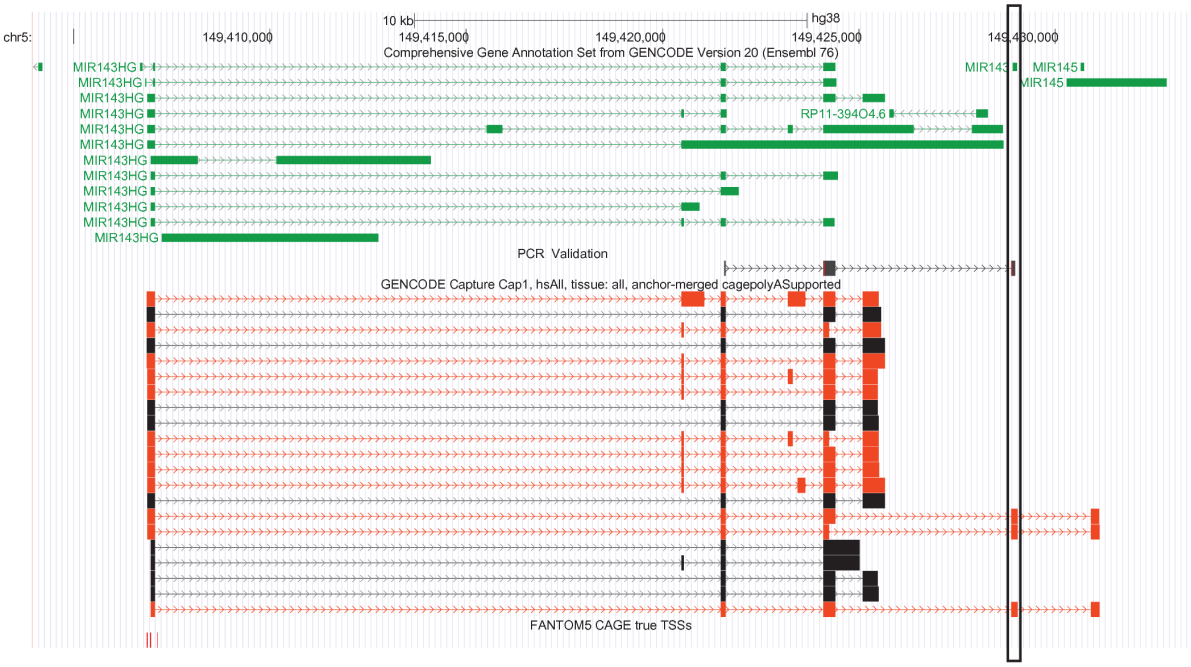
Each panel represents the number of novel splice junctions (SJs) discovered (y-axis) in a simulated analysis where increasing numbers of mapped ROIs (x-axis) were randomly sampled from the experiment. The SJs retrieved at each read depth were further stratified by level of sequencing support (Dark brown: all PacBio SJs; Orange: HiSeq-supported PacBio SJs; Black: HiSeq-unsupported PacBio SJs). Each randomization was repeated fifty times, and a boxplot summarizes the results at each simulated depth. The highest y value represents the actual number of novel SJs discovered. Analogous data for human is to be found in Figure 3d.

(b) Novel transcript discovery simulations for human (upper section) and mouse (lower section)

Each panel represents the number of novel transcript models (TMs) discovered (y-axis) in simulated analysis where increasing numbers of mapped 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 TMs discovered.

(c) Splice junction discovery simulations for human (upper section) and mouse (lower section) using captured HiSeq reads

Each panel represents the number of splice junctions (SJs) discovered (y-axis) in simulated analysis where increasing numbers of HiSeq reads (x-axis) were randomly sampled from the experiment. The randomizations were repeated five times, and a boxplot summarizes the results at each simulated depth (purple: all HiSeq-derived SJs; brown: HiSeq-derived SJs also detected in PacBio matched samples). The highest y value represents the actual number of novel SJs discovered in each sample using HiSeq. The horizontal red line marks the number of HiSeq-supported PacBio SJs detected in the corresponding sample.



Supplementary Figure 9

Supplementary Figure 9: Evidence that CARMEN1 isoforms are precursors of hsa-mir-143

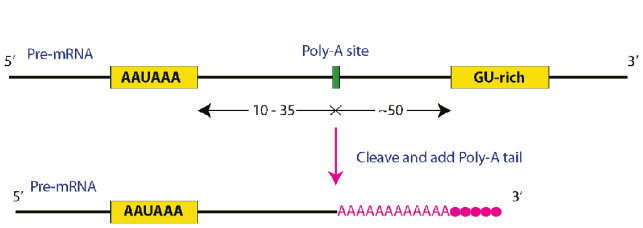
Shown is the CARMEN1 locus (chr5:149,402,925-149,452,858, hg38). GENCODE v20 annotation is green, capture probe targets in grey, full length CLS transcript models in black (known) and red (novel). Also visible are tracks for CAGE peaks from FANTOM and polyA sites from this study. Note the existence of novel isoforms directly overlapping on the same strand the mature hsa-mir-143 (boxed), for which no precursor annotation exists in GENCODE. Also shown is the sequence obtained by RT-PCR and Sanger sequencing (black).

a

Canonical polyA signals:

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

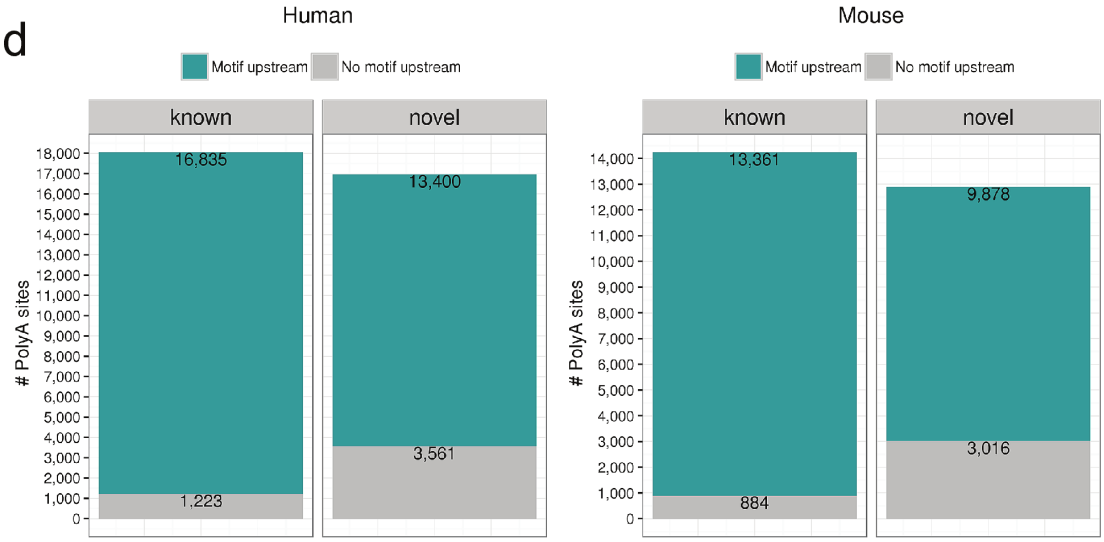
b



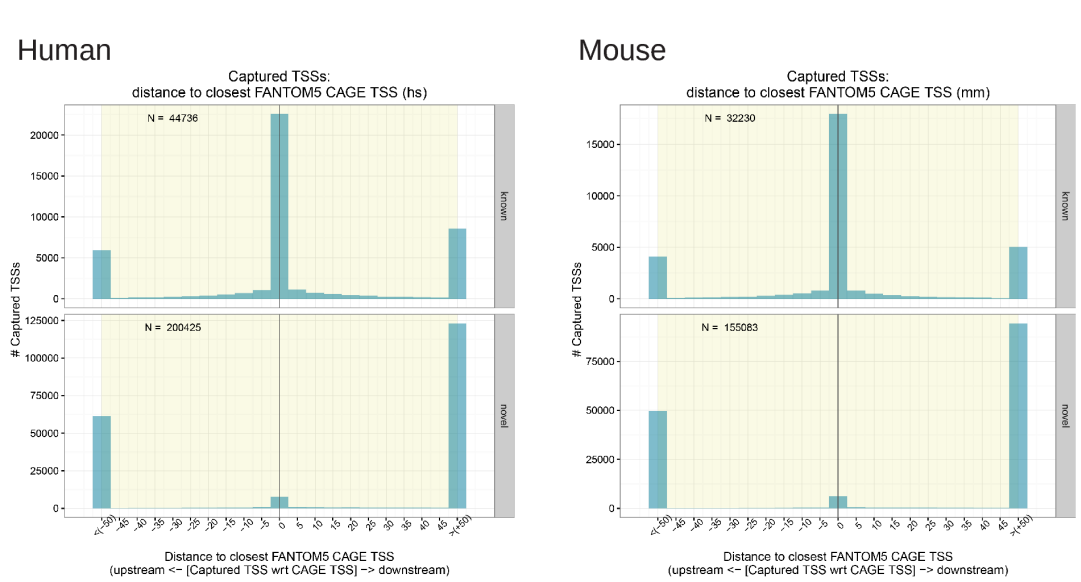
c

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



e



Supplementary Figure 10

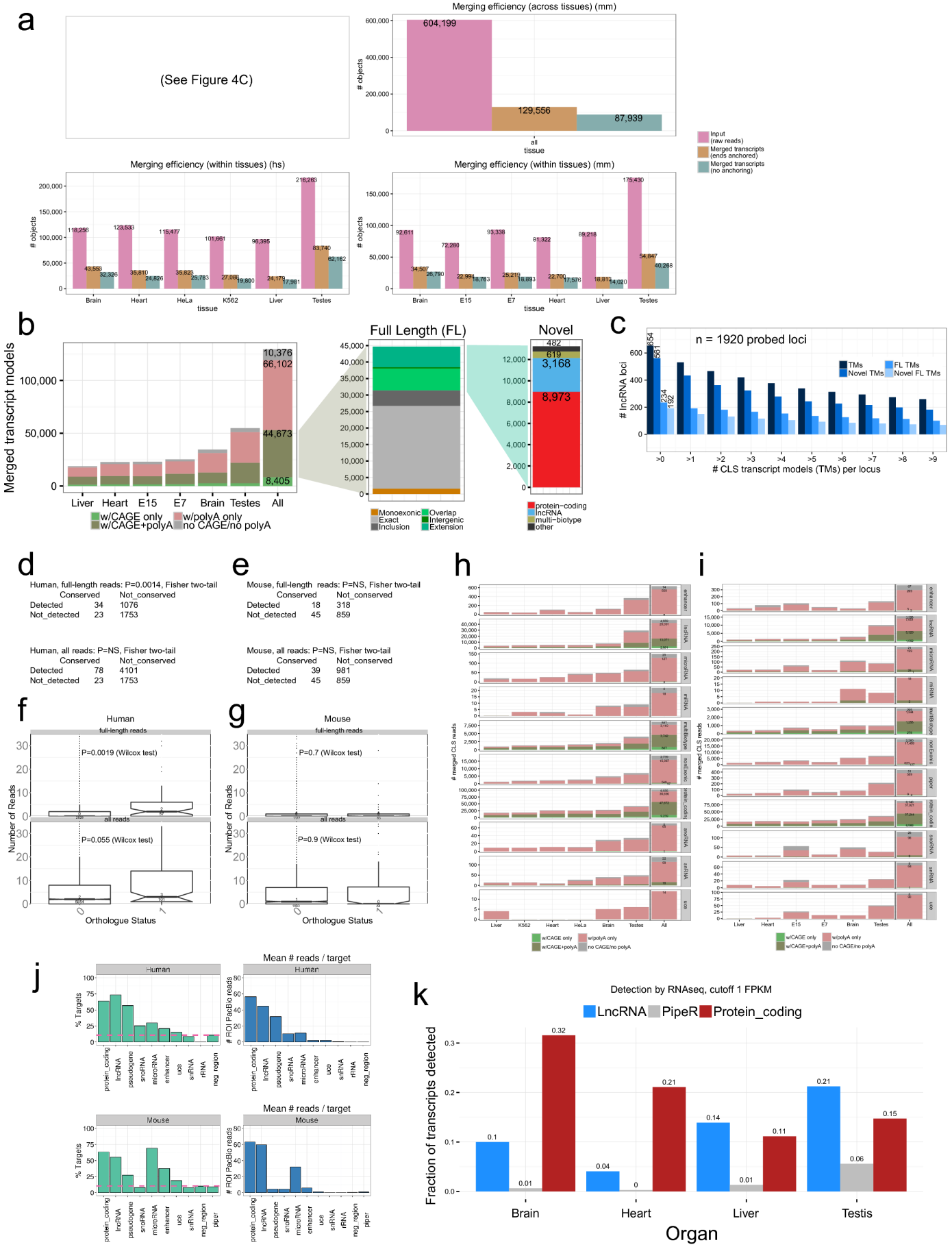
Supplementary Figure 10: Analysis of transcript ends

(a-d) Analysis of polyadenylation motifs around known and novel transcript 3′ ends

(a) list of the polyA motif considered here to be canonical. **(b)** Overview of the pre-mRNA termination and polyadenylation process. PolyA tails are generally added between 10 and 35 nt downstream of the polyA motif. **(c)** The rate at which CLS-discovered 3′ ends contain a canonical polyA signal. Control sites were generated by selecting the middle of non-terminal captured exons more than 100nt distal from the nearest captured polyA site. **(d)** Comparison of polyA motif frequency between known and novel 3′ ends.

(e) Captured TSS distance to closest CAGE cluster

Left: human; right: mouse; top: known TSSs; bottom: novel TSSs (w.r.t. GENCODE). Each plot is a histogram of the distance of the 5′ end of the start of a transcript model annotation to the FANTOM5 CAGE TSS. To the left are cases where captured TSS is upstream of the nearest CAGE TSS. The two extreme bins ("<(-50)" and ">(+50)") contain all cases where the closest CAGE cluster lies more than 50 bases away. The population size of both sets is reported in the top left corner of each plot.



Supplementary Figure 11

Supplementary Figure 11: Transcript merging, end support and detection in CLS

(a) Performance of anchored transcript model merging compared to a conventional approach

Charts indicate the number of unique transcript models created in each case (human on left, mouse on right). Upper panels show all reads, lower panel show reads broken down by sample origin.

(b) Anchor-merged transcript models identified by CLS in mouse

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. Equivalent data for mouse are found in Figure 4e.

(c) Probed lncRNA loci vs CLS transcript isoforms in mouse

The total numbers of probed lncRNA loci giving rise to CLS transcript models (TMs), novel TMs, full-length CLS TMs (FL TMs) and novel FL TMs in mouse, at increasing minimum cutoffs for each category.

(d-g) Detection rates of lncRNAs with evolutionary orthologues

(d-e): Contingency tables show the numbers of detected gene annotations in each category: "Detected" is defined as having one or more mapping reads, either of any type (upper row) or only full-length (lower row). None were significant by Fisher′s test (two-tailed). **(f-g)**: Boxplots show the same data as above, but broken down by the numbers of reads per gene. Numbers above boxes show the median (upper number) and number of data points (lower number). Orthologue status "0" and "1" indicate lncRNAs without / with identified orthologues, respectively. Further details may be found in the Methods.

(h-i) Transcript completeness by biotype and tissue source, in human (h) and mouse (i)

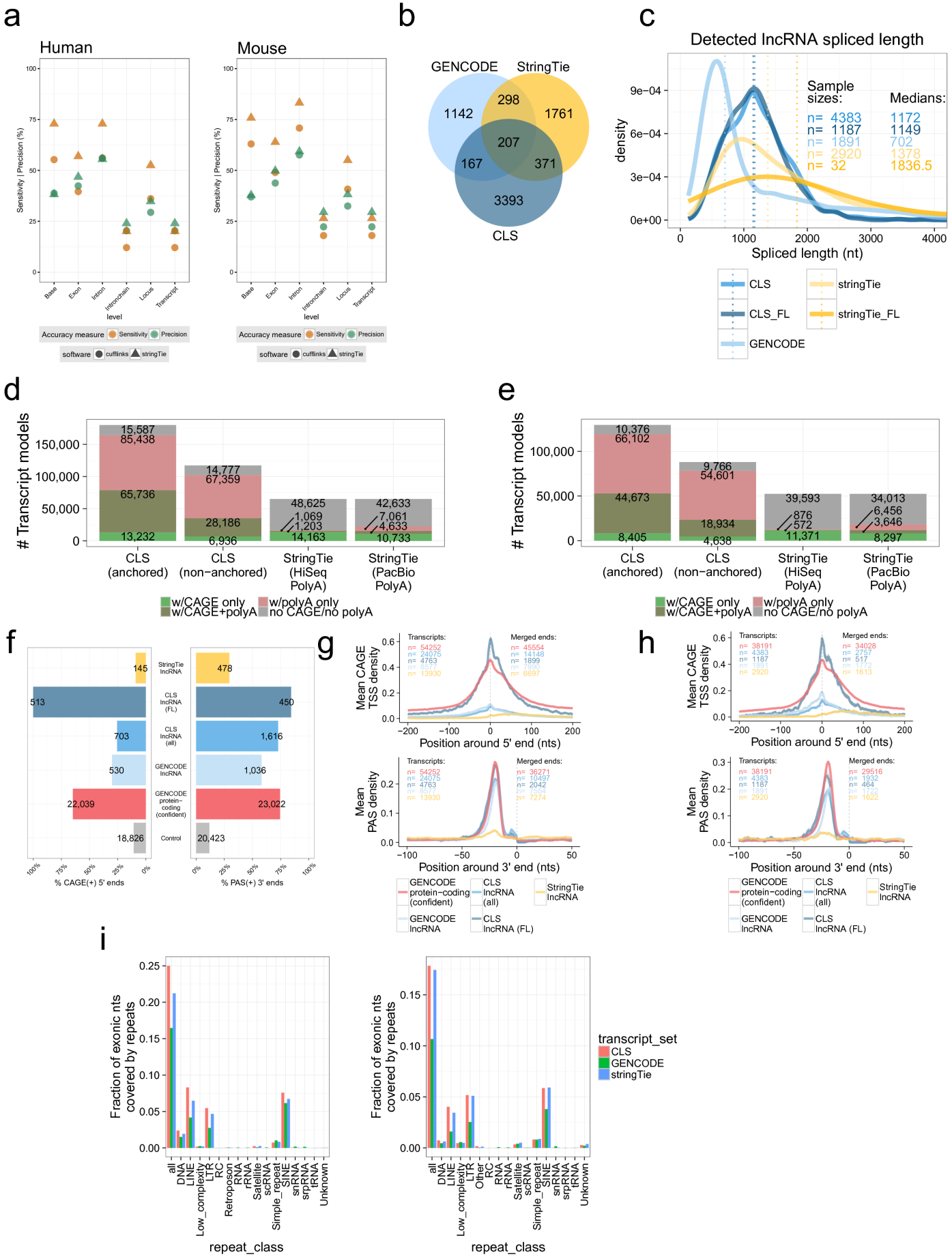
Figures show the number of unique merged transcript structures. Transcripts are coloured by 5′ or 3′ completeness.

(j) Validation rates across target categories

Left panels show the percentage of probed targets detected by at least one ROI in human (top) and mouse (bottom). The rate of detection of negative regions is indicated with a pink dashed line. Right panels show the average number of ROIs detected per target class.

(k) Expression of PipeR lncRNA predictions in mouse tissues

Shown is the fraction of detected transcript models in each class, as measured by HiSeq in pre-captured samples and using a detection cut-off of >1 FPKM. Numbers of analysed transcripts: lncRNA - 8170, PipeR - 2469, protein-coding - 77,499.



Supplementary Figure 12

Supplementary Figure 12: Comparison of CLS with short-read transcript reconstruction methods

(a) Benchmarking of the *StringTie* and *Cufflinks* transcript assembly methods using PacBio evidence as a reference

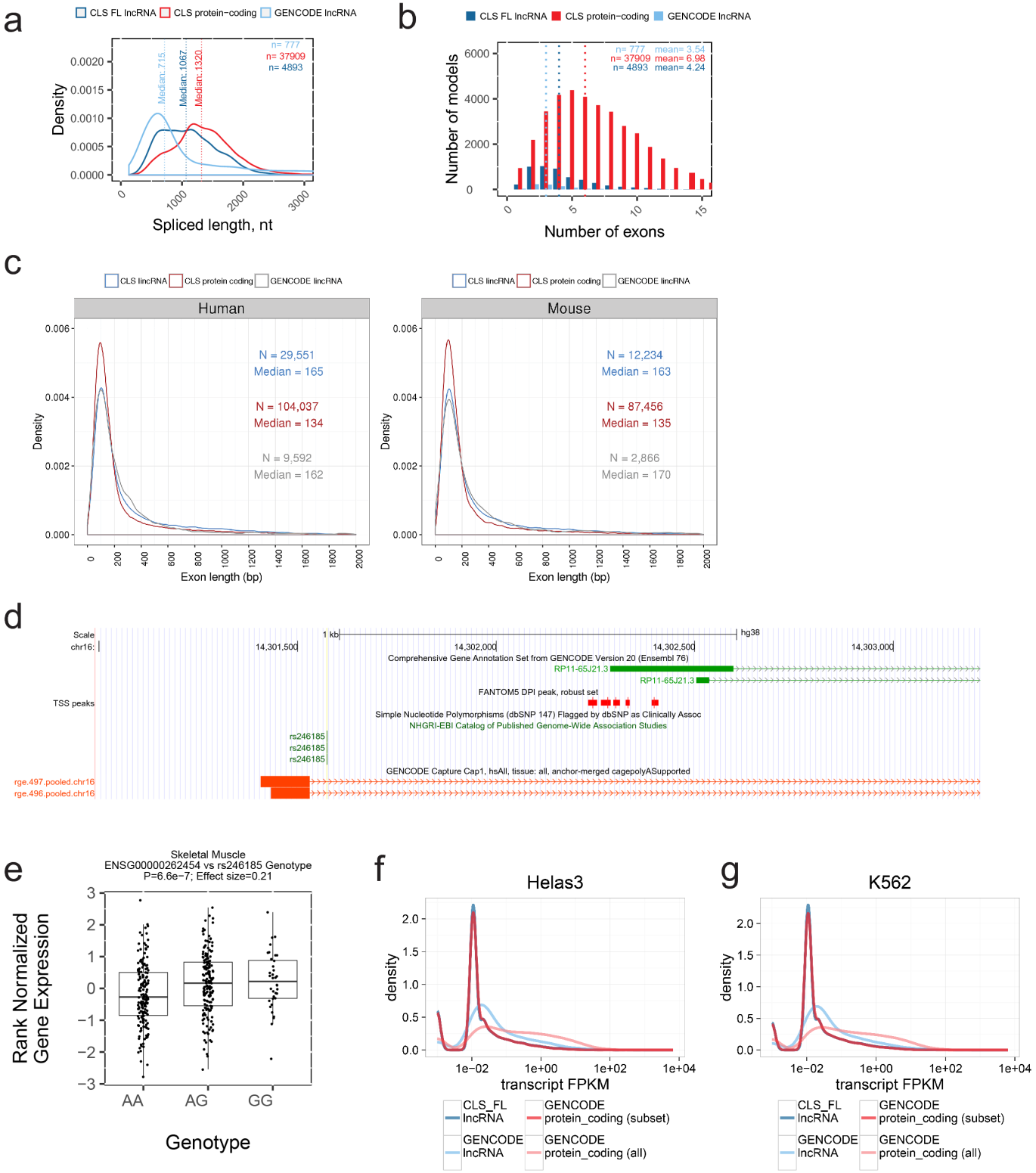
The *y*-axis displays the sensitivity/precision of each method in retrieving the indicated transcriptome elements (*x*-axis), as defined by CLS TMs.

(b-h) Comparison of *StringTie* and CLS transcript models (TMs)

Data shown in (b), (c) and (f) are for mouse; equivalent human data may be found in Figure 4. **(b)** Comparison of the numbers of unique transcript models present in each collection. Shared transcripts are defined by having identical intron chains. **(c)** Spliced length distributions of indicated non-redundant transcript catalogues. "FL" indicates the subset of transcripts from each catalogue that has 5′ support from CAGE, and 3′ support from PacBio-identified polyA sites. The median spliced length of each population is indicated by a vertical dotted line. **(d)** and **(e)**: *StringTie* models 5′ and 3′ completeness in human and mouse, respectively, compared to CLS merged models. "HiSeq PolyA" / "PacBio PolyA": comparison with polyA sites called using captured HiSeq / PacBio reads, respectively (see Methods). **(f)** Comparing completeness of transcript annotations: 5′ and 3′ completeness as estimated by CAGE overlap and upstream polyadenylation signal (PAS), with respect to 5′ (left) and 3′ ends (right), respectively. Neighbouring transcript ends were merged within each individual set (maximum distance: +/- 5nts on the same strand). "*GENCODE lncRNA*": subset of probed GENCODE lncRNAs detected by CLS or *StringTie*. "*GENCODE protein-coding (confident)*": 5′/3′ boundaries of high-confidence GENCODE protein-coding transcripts (see Methods). Control sites represent a random sample of internal exons′ middle coordinate. Represented is the proportion of transcript ends with CAGE or PAS support in each set (mouse). CAGE(+) 5′ ends are those TSSs having a CAGE cluster within a +/-50 bases window around them. Similarly, PAS(+) 3′ ends correspond to 3′ ends falling 10 to 50 bases downstream of a PAS motif. Note that full length (FL) CLS models have, by definition, a CAGE signal at 5′ end, and thus have 100% 5′ completeness. Corresponding counts of CAGE- or PAS-supported features are indicated on each bar. **(g-h)** CAGE TSS (top panel) and PAS (bottom panel) density aggregate plots, in human (g) and mouse (h). The mean density of CAGE TSSs and PAS (AATAAA and ATTAAA motifs) over each genomic position around various sets of transcript ends is represented. CAGE TSSs and PAS were required to overlap tested genomic regions on the same strand. Sample sizes (number of transcripts and number of merged ends after clustering) are indicated within each graph. Grey fringes represent the standard error of the mean. Transcript ends were merged as in (f), except for 3′ ends, for which a maximum clustering distance of 50 nucleotides was applied.

(i) Genome repeat coverage in CLS, *StringTie* and GENCODE exons

Shown is the fraction of exonic nucleotides covering genome repeats of various classes in each set of transcripts (left: human; right: mouse).



Supplementary Figure 13

Supplementary Figure 13: Full-length lncRNA transcripts: properties and genomic environment

(a-b) Comparison of lncRNA and mRNA transcript structure in mouse

(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) Exon length distributions

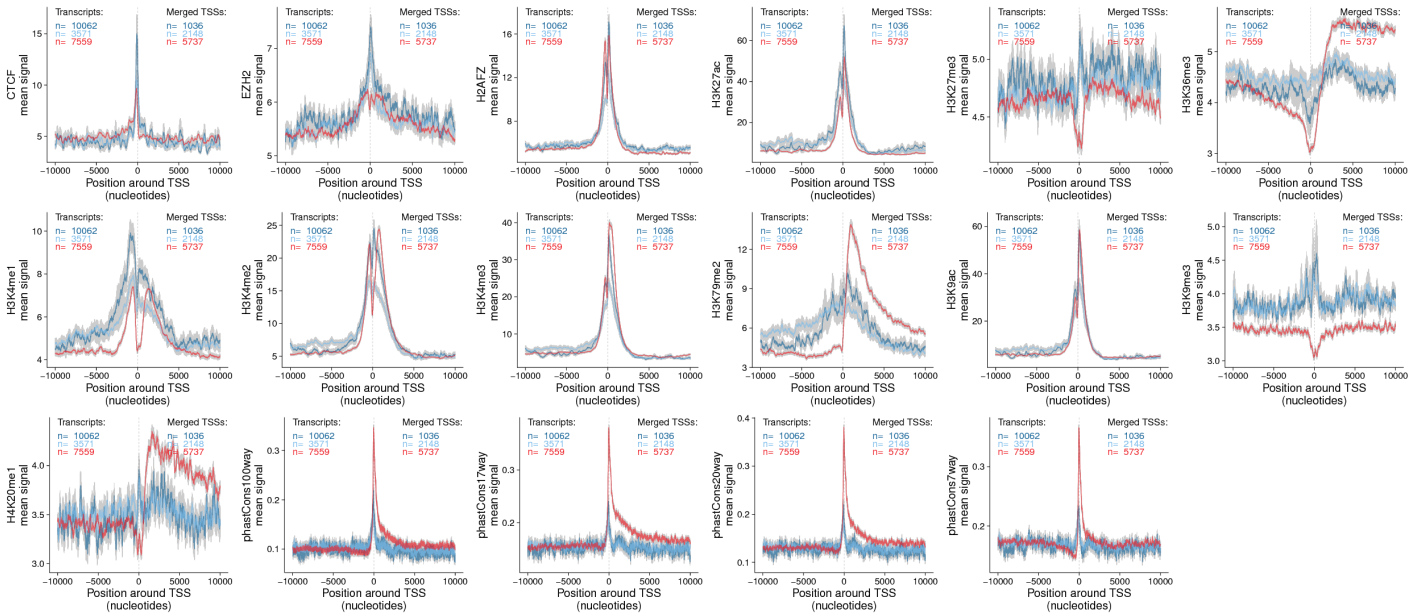
The distribution of exon lengths of: CLS full-length transcript models from targeted lncRNA loci (blue); transcript models from the targeted and detected GENCODE lncRNA loci (grey); CLS full-length transcript models from protein-coding loci (red). Left: human; right: mouse.

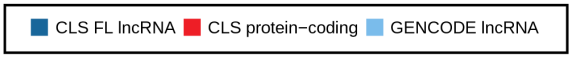
(d-e) Example of an expression QTL at lncRNA RP11-65J2

(d) The RP11-65J21.3 (ENSG00000262454) locus, showing phenotype-associated SNP rs246185. Existing GENCODE v20 annotation is shown in green, novel full-length transcript models in red. **(e)** Expression of ENSG00000262454 in muscle of GTEx individuals, broken down by genotype of rs246185. eQTL analysis was obtained from the GTEx Portal (http://www.gtexportal.org/home/).

(f-g) Creating an expression-matched set of protein-coding genes

Panels (f) and (g) show the distribution of whole-cell RNA levels for indicated transcript sets in HeLa and K562 cells, respectively. Note the log scale of the x-axis. Data are shown for CLS full-length lncRNA transcript models (dark blue), as well as the original GENCODE annotations to which they map (light blue). Also shown are data for all protein-coding genes (light red). From the latter, a subset was sampled with a similar expression distribution as the CLS lncRNAs (dark red).

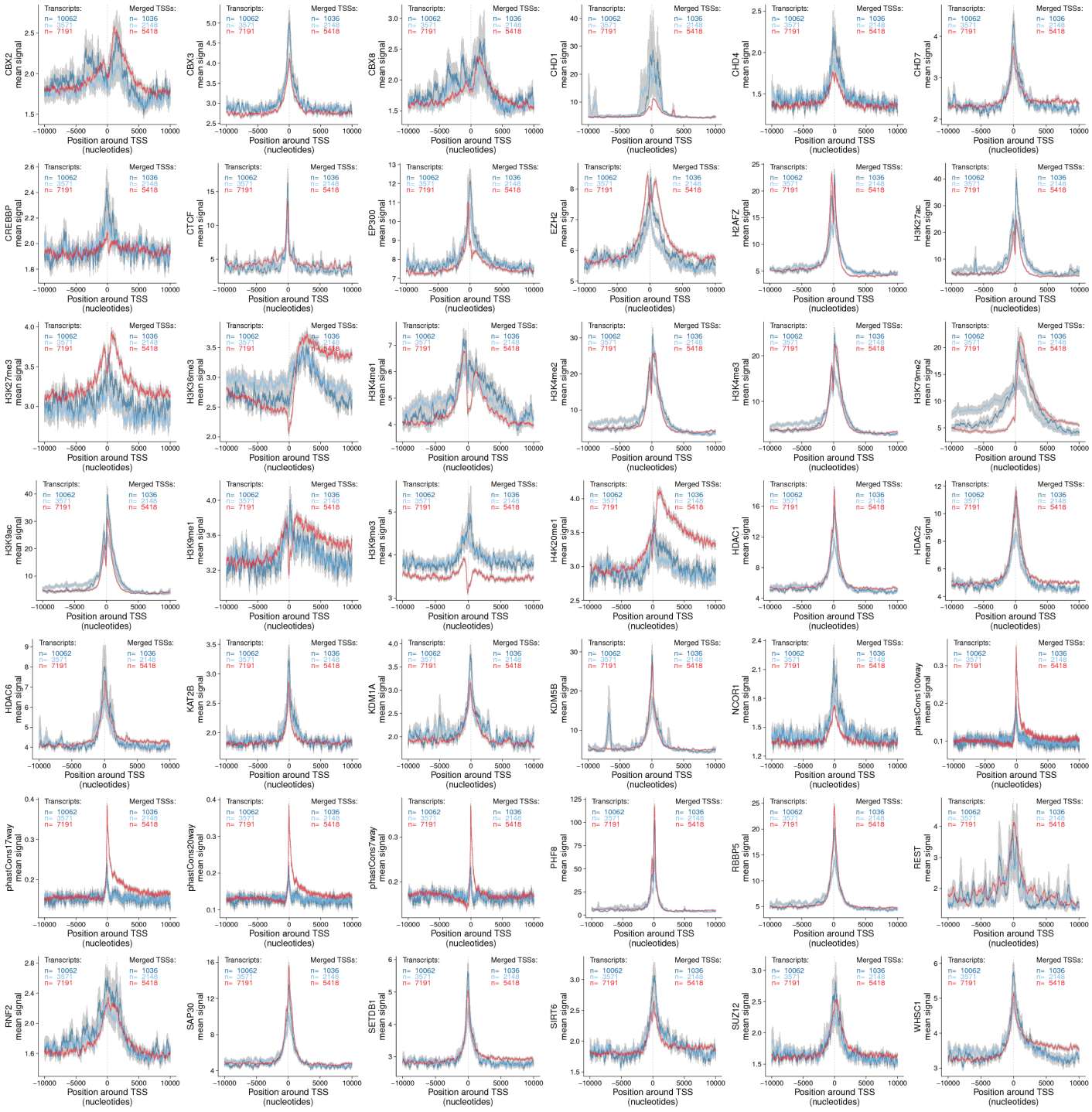




Supplementary Figure 14

Supplementary Figure 14: Characteristics of "standalone" promoters in HeLa

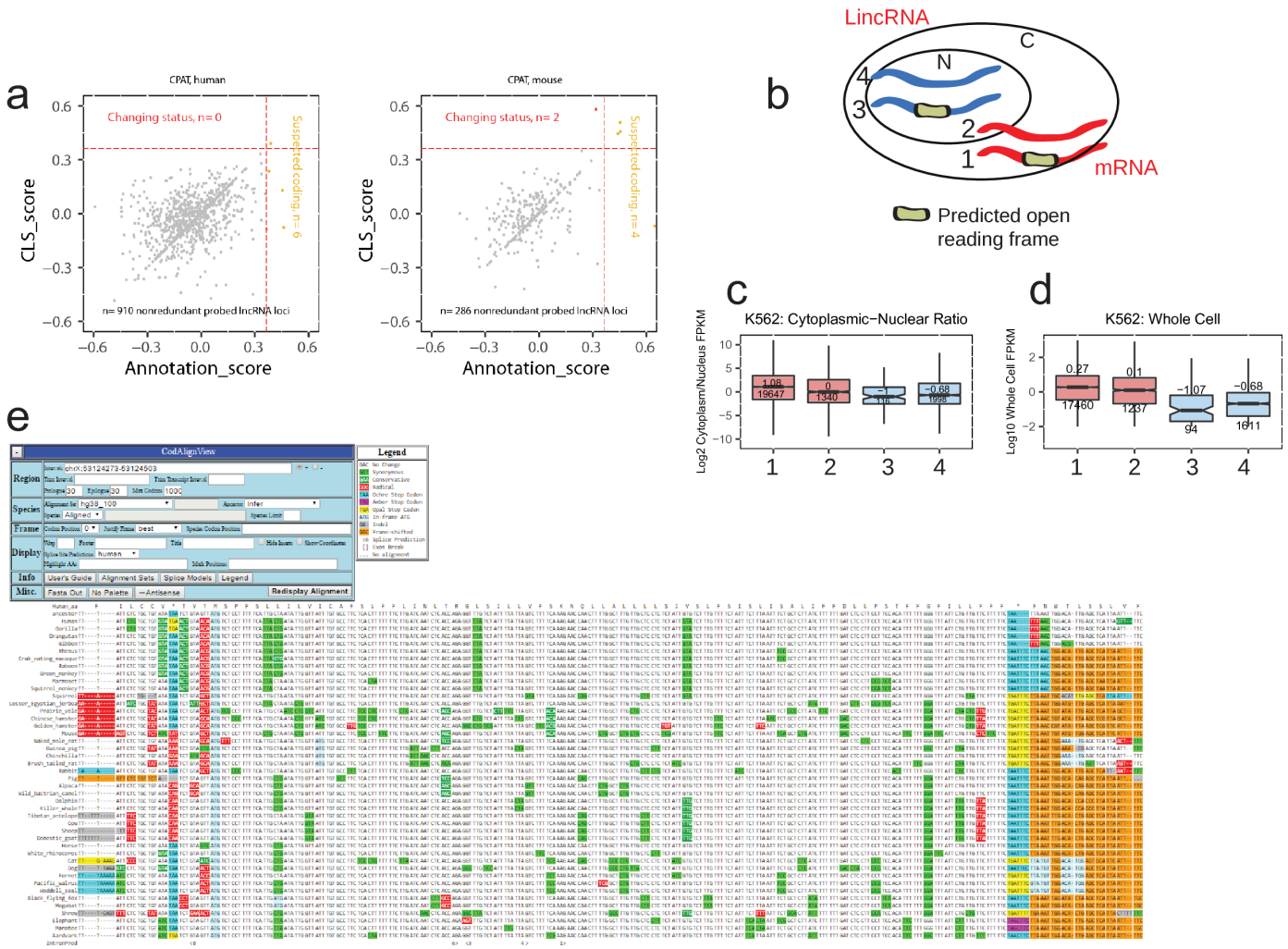
Montage of all signal density plots produced in HeLa across sets of "standalone" (*i.e.*, non-bidirectional) TSSs. The aggregate density of various features is shown across the TSS of indicated gene classes. Note that overlapping TSSs were merged within classes. The *y*-axis denotes the mean signal per 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.



Supplementary Figure 15

Supplementary Figure 15: Characteristics of "standalone" promoters in K562

Montage of all signal density plots produced in K562 across sets of "standalone" (*i.e.*, non-bidirectional) TSSs. The aggregate density of various features is shown across the TSS of indicated gene classes. Note that overlapping TSSs were merged within classes. 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 K562 as the CLS lncRNAs.



Supplementary Figure 16

Supplementary Figure 16: Analysis of protein-coding potential and subcellular localization

(a) Changes in protein-coding status due to long read extension

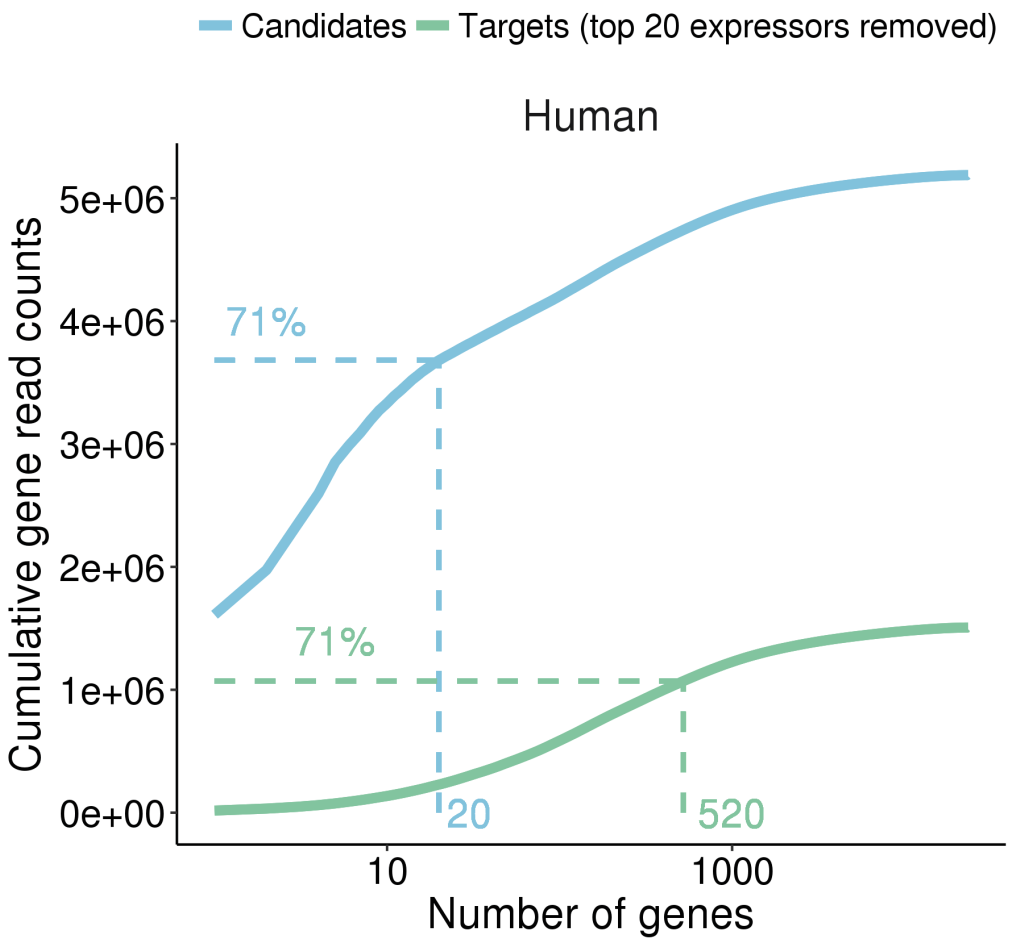
Changes in CPAT-predicted protein-coding potential in lncRNAs due to extension by CLS. Each point represents a probed and detected lncRNA gene. For each gene, the highest-scoring associated transcript model is used. The x-axis denotes the CPAT score of original GENCODE annotation, and the y-axis the score of associated full-length read models from CLS. Red lines indicate the prediction threshold dividing coding and non-coding. In yellow are shown gene loci that may be protein-coding, prior to CLS. In red are shown gene loci whose status changes following CLS.

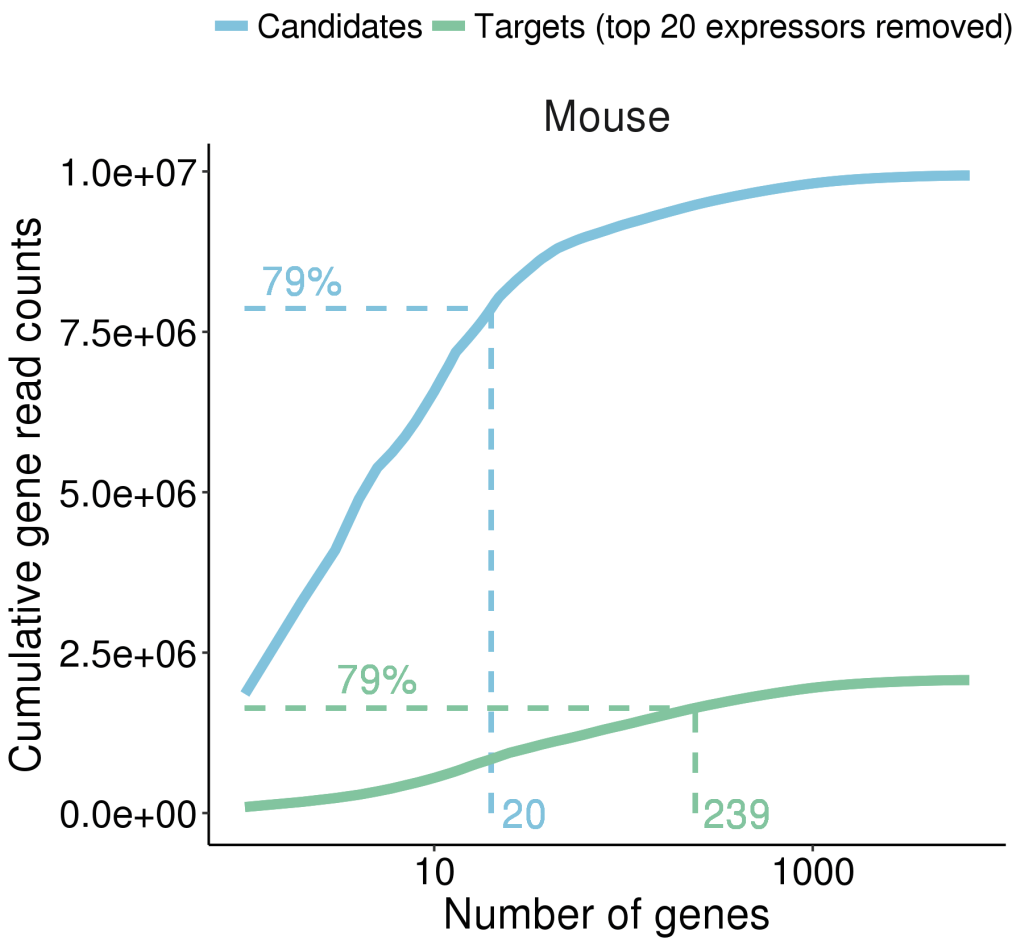
(b-d) Expression and localisation properties of full-length transcript models in K562 cells, broken down by annotated and predicted coding potential

(b) Schematic of subcellular localisation of annotated lncRNAs (blue) and mRNAs (red). Indicated are identified ORFs in these transcripts in beige colour. **(c)** Subcellular localisation of transcripts in K562 cells. Localisation (y-axis) is estimated from RNAseq data by the log2 ratio of cytoplasmic RPKM / nucleus RPKM. Inside each box are displayed the median value (above) and the number of transcript models considered (below). Samples are numbered as in (b). **(d)** Similar to (c), but showing whole-cell expression values. Note that here, ORFs are defined to be present if predicted by either PhyloCSF or CPAT.

(e) Detailed view of KANTR short ORF

This corresponds to region chrX:53124273-53124488 in the hg38_100 alignment set, using CodAlignView (https://data.broadinstitute.org/compbio1/cav.php).





Supplementary Figure 17

Supplementary Figure 17: Removing high-expressed genes that dominate sequencing

Graphs show the cumulative number of sequencing reads originating from ranked lists of GENCODE lncRNA genes before (blue, "Candidates") and after (green, "Targets") removing the 20 most highly expressed genes, emphasizing the high fraction of all reads originating from the top 20 genes (note the logarithmic scale on the *x*-axis). The remaining gene models ("Targets") were used for capture probe design. Blue dashed lines represent the percentage of reads accounted for by the top 20 genes in the "Candidates" set. Green dashed lines depict the number of genes accounting for that percentage of reads (*i.e.*, 71% in human, 79% in mouse) in the "Targets" set. These plots were produced using matched, public Illumina short-read RNAseq data corresponding to the organs studied here. Upper panel: human; lower panel: mouse.

Supplementary Tables

Species	Sample	Total # uniquely mapped double-bounded reads	# polyA reads	% polyA reads	# on-genome polyA site clusters (+/-5nts), min 2 reads
Human	**Brain**	170,012	106,505	63%	9,607
	Heart	153,214	115,973	76%	8,502
	HeLa	150,196	109,023	73%	8,211
	K562	128,994	98,758	77%	6,097
	Liver	118,868	94,739	80%	5,786
	Testes	278,929	206,457	74%	16,850
	Total	**1,000,213**	**731,455**	**73%**	**35,092**
Mouse	**Brain**	150,371	85,679	57%	7,903
	E15	185,837	69,564	37%	5,271
	E7	131,314	93,459	71%	6,419
	Heart	117,908	79,320	67%	5,608
	Liver	123,941	96,774	78%	4,867
	Testes	232,386	176,318	76%	12,852
	Total	**941,757**	**601,114**	**64%**	**27,152**

Supplementary Table 1

Supplementary Table 1: Statistics on polyA site identification

Statistics on polyA site identification in double-bounded, genome-mapped reads (*i.e.* excluding ERCC spike-ins).

a	b	c	d	e	f	g	h
Species	**Biotype**	**# merged transcript models**	**# merged FL transcript models**	**% merged transcript models that are FL (d/c)**	**# novel, FL transcript models**	**% FL transcript models that are novel (f/d)**	**# loci affected by novel, FL transcript models**
Human	enhancer	634	1	0%	1	100%	1
	lncRNA	47,002 (42,463)	13,071 (11,429)	28%	8,494	65%	1,220 (812)
	microRNA	172	2	1%	2	100%	1
	misc_RNA	19	5	26%	3	60%	1
	Mt_rRNA	28	18	64%	0	0%	0
	Mt_tRNA	4	3	75%	1	33%	1
	multiBiotype	8,616	3,742	43%	1,916	51%	1,027
	neg_region	45	3	7%	3	100%	1
	nonExonic	18,751	548	3%	287	52%	0
	prot_coding	102,156	47,672	47%	11,076	23%	4,294
	pseudogene	2,344	655	28%	429	66%	103
	snoRNA	71	0	0%	0	N/A	0
	snRNA	137	16	12%	5	31%	2
	uce	14	0	0%	0	N/A	0
	Total	**179,993**	**65,736**	**37%**	**22,217**	**34%**	**N/A**
Mouse	enhancer	364	5	1%	5	100%	1
	lncRNA	15,580 (13,130)	5,329 (4,350)	34%	3,168	59%	448 (249)
	microRNA	266	27	10%	9	33%	6
	misc_RNA	15	0	0%	0	N/A	0
	Mt_rRNA	42	0	0%	0	N/A	0
	Mt_tRNA	7	2	29%	0	0%	0
	multiBiotype	3,075	1,258	41%	619	49%	337
	neg_region	37	0	0%	0	N/A	0
	nonExonic	20,469	623	3%	419	67%	0
	piper	433	9	2%	8	89%	3
	prot_coding	88,177	37,244	42%	8,973	24%	3,608
	pseudogene	791	167	21%	40	24%	19
	snoRNA	131	9	7%	1	11%	1
	snRNA	73	0	0%	0	N/A	0
	uce	96	0	0%	0	N/A	0
	Total	**129,556**	**44,673**	**35%**	**13,242**	**30%**	**N/A**

Supplementary Table 2

Supplementary Table 2: Breakdown of captured transcripts by gene biotype and novelty

Numbers refer to transcript models merged across all tissue samples. Counts corresponding to lncRNA probed regions are reported between parentheses where appropriate. The number of annotated loci originating these transcript models is indicated in the rightmost column.

Species	Merging method	End support level	Total # TMs	# HiSeq-supported TMs	% HiSeq-supported TMs
Human	anchored	any	179,993	155,617	86%
		CAGE+polyA	65,736	60,046	91%
		CAGE	78,968	71,363	90%
		polyA	151,174	134,537	89%
	standard	any	117,258	94,163	80%
		CAGE+polyA	28,186	23,070	82%
		CAGE	35,122	28,494	81%
		polyA	95,545	80,135	84%
Mouse	anchored	any	129,556	113,186	87%
		CAGE+polyA	44,673	40,693	91%
		CAGE	53,078	47,924	90%
		polyA	110,775	99,362	90%
	standard	any	87,939	72,413	82%
		CAGE+polyA	18,934	15,261	81%
		CAGE	23,572	19,027	81%
		polyA	73,535	62,757	85%

Supplementary Table 3

Supplementary Table 3: HiSeq support of merged CLS transcript models

Numbers refer to transcript models (TMs) merged across all tissue samples using the "anchored" and "standard" (*i.e.*, non-anchored) methods. HiSeq-supported TMs refer to those TMs whose entire set of introns are supported by at least one split short read in the captured HiSeq libraries. These transcript models are referred to as "HiSeq-supported TMs" elsewhere in the paper. "CAGE+polyA" end support level corresponds to full-length TMs. "Any" end support level refers to all merged TMs, including the ones without CAGE/polyA end support.

Feature	Source	Number of targeted transcripts	Comments
lncRNAs (intergenic)	GENCODE v20	9,560	
microRNA	mirBase v20	785	Tiled 1kb
snoRNA	GENCODE v20	401	Tiled 1kb
snRNA	GENCODE v20	838	Tiled 1kb
VISTA enhancers	http://enhancer.lbl.gov/	1,908	
Ultraconserved elements	UCNEbase	316	Any UCE less than 500 bp long were removed.
Protein-coding	GENCODE v20	100	Expression matched to lncRNAs
E. coli (random genomic)		100	Identical in human and mouse libraries
ERCC sequences (selected)	https://www.thermofisher.com/order/catalog/product/4456740	42	Identical in human and mouse libraries

Supplementary Table 4

Supplementary Table 4: Target regions for capture library design (human)

Feature	Source	Number of targeted transcripts	Comments
lncRNAs (intergenic)	GENCODE vM3	2,817	
Orthologues of human lncRNAs	PipeR	2,469	
microRNA	mirBase v20	494	Tiled 1kb
snoRNA	GENCODE vM3	850	Tiled 1kb
snRNA	GENCODE vM3	721	Tiled 1kb
VISTA enhancers	http://enhancer.lbl.gov/	406	
Ultraconserved elements	UCNEbase	312	
Protein-coding (expression matched)	GENCODE vM3	100	
E. coli (random genomic)		100	Identical in human and mouse libraries
ERCC sequences (selected)	https://www.thermofisher.com/order/catalog/product/4456740	42	Identical in human and mouse libraries

Supplementary Table 5

Supplementary Table 5: Target regions for capture library design (mouse)

Species	Sample	ERCC mix
Human	**Heart**	Mix1
	Testes	Mix2
	Liver	Mix2
	Brain	Mix1
	HeLa	Mix2
	K562	Mix1
Mouse	**Heart**	Mix1
	Testes	Mix2
	Liver	Mix1
	Brain	Mix2
	E7	Mix1
	E15	Mix2

Supplementary Table 6

Supplementary Table 6: ERCC spike-in mixes used per library

Species	Sample type	Illumina index ID	Index Sequence
Human	**Heart**	AD020	GTGGCC
	Testes	AD021	GTTTCG
	Liver	AD022	CGTACG
	Brain	AD023	GAGTGG
	HeLa	AD025	ACTGAT
	K562	AD027	ATTCCT
Mouse	**Heart**	AD013	AGTCAA
	Testes	AD014	AGTTCC
	Liver	AD015	ATGTCA
	Brain	AD016	CCGTCC
	E7	AD018	GTCCGC
	E15	AD019	GTGAAA

Supplementary Table 7

Supplementary Table 7: Index / barcode sequences

See Supplementary Data 3 for full adapter sequences.

Label	Size Range	# SMRT-cells	Loading Concentration (pM)	Loading Method	Read Bases of Insert	Mean Read Length of Insert	Mean Read Quality of Insert	Mean Number of Passes
MM_1	240 - 646	1	500	diff	3,479,177	329	98.9%	44
MM_2	438 - 3400	1	500	diff	13,436,069	594	99.0%	30
MM_3	896 - 5931	21	250	Mag	784,093,824	1304	99.0%	15
MM_4	672 - 6841	21	400	Mag	1,205,620,473	1561	99.1%	13
MM_5	500 - 5000	21	25	diff	51,313,641	986	99.0%	18
HS_1	253 - 698	1	500	diff	10,808,941	332	98.9%	45
HS_2	388 - 3262	1	500	diff	12,585,406	606	98.9%	28
HS_3	503 - 12138	21	25	Mag	64,413,386	1087	98.8%	17
HS_4	551 - 11636	21	25/35	Mag	1,049,441,562	1486	99.3%	13
HS_5	558 - 5000	21	40	Mag	989,112,311	1147	99.2%	16

Supplementary Table 8

Supplementary Table 8: Summary of PacBio sequencing

MM: mouse; HS: human.

Properties common to all samples/fractions:

- PacBio Kit: #100-259-100
- Polymerase used: P6/C4 (except HS_4: P5/C4 and HS_5: P4/C3)
- Movie length: 4h
- Post-run analysis: RS_ReadsOfInsert.1
- Files generated: FASTQ

	a	b	c	d	e
Dataset	**# Mapped reads**	**# Uniquely mapped reads (UMD-ROIs)**	**# double-bounded UMD-ROIs**	**# genome-mapped, double-bounded UMD-ROIs**	**% double-bounded UMD-ROIs (c/b)**
Hs Brain	274,732	265,170	200,767	170,012	76%
Hs Heart	232,699	224,910	176,357	153,214	78%
Hs HeLa	233,303	220,340	169,847	150,196	77%
Hs K562	198,890	185,684	140,307	128,994	76%
Hs Liver	185,797	180,379	136,492	118,868	76%
Hs Testes	423,727	405,447	306,445	278,929	76%
Hs - total	**1,549,148**	**1,481,930**	**1,130,215**	**1,000,213**	**76%**
Mm Brain	231,189	219,521	168,322	150,371	77%
Mm E15	280,718	266,837	201,578	185,837	76%
Mm E7	208,421	194,473	146,325	131,314	75%
Mm Heart	207,954	193,872	133,684	117,908	69%
Mm Liver	181,337	174,843	137,294	123,941	79%
Mm Testes	342,414	329,350	252,582	232,386	77%
Mm - total	**1,452,033**	**1,378,896**	**1,039,785**	**941,757**	**75%**

Supplementary Table 9

Supplementary Table 9: Summary statistics on UMD-ROIs and double-bounded reads

Hs: Human; Mm: Mouse.

UMD-ROIs: Uniquely Mapped and Demultiplexed ROIs

Undetermined (*i.e.*, non-demultiplexed) reads are not reported, as they do not bear a recognizable index sequence, by definition. Genome-mapped reads refer to reads not mapped to ERCC spike-in sequences.

	a	b	c	d	e	f	g	h	i
Species	**Total # uniquely mapped reads**	**# reads stranded by at least one method**	**# reads stranded by both methods**	**# reads with same strand inferred by both methods**	**# reads stranded by polyA method only**	**# reads stranded by SJ method only**	**% reads stranded by at least one method (b/a)**	**% reads stranded by both methods (c/a)**	**% reads with same strand inferred by both methods (d/c)**
Human	2,053,424	1,446,986	566,109	564,258	168,398	712,479	70.5%	27.6%	99.7%
Mouse	1,870,681	1,255,423	491,493	490,110	111,877	652,053	67.1%	26.3%	99.7%

Supplementary Table 10

Supplementary Table 10: Comparison/integration of polyA and SJ strand inference approaches

"Undetermined" (*i.e.*, non-demultiplexed) ROIs are included in the total.

Species	TSS type	# TSSs	# CAGE-supported TSSs	% CAGE-supported TSSs
Human	novel	200,425	16,305	8.1%
	known	44,736	30,352	67.9%
Mouse	novel	155,083	11,255	7.3%
	known	32,230	23,195	72.0%

Supplementary Table 11

Supplementary Table 11: CAGE support of novel vs known PacBio TSSs

A TSS is considered supported if a FANTOM "true" TSS is found within 50 bases around it on the same genomic strand (see Methods).

Cell line	ChIP-Seq antibody target	ENCODE portal file accession
Helas3	CTCF	ENCFF000BAN
Helas3	EZH2	ENCFF000BAV
Helas3	H2AFZ	ENCFF000BAZ
Helas3	H3K27ac	ENCFF000BBR
Helas3	H3K27me3	ENCFF000BBX
Helas3	H3K36me3	ENCFF000BCD
Helas3	H3K4me1	ENCFF000BBF
Helas3	H3K4me2	ENCFF000BCJ
Helas3	H3K4me3	ENCFF000BCP
Helas3	H3K79me2	ENCFF000BCV
Helas3	H3K9ac	ENCFF000BDB
Helas3	H3K9me3	ENCFF000BBL
Helas3	H4K20me1	ENCFF000BDI
K562	CBX2	ENCFF000BVA
K562	CBX3	ENCFF000BVE
K562	CBX8	ENCFF000BVI
K562	CHD1	ENCFF000BVO
K562	CHD4	ENCFF000BVT
K562	CHD7	ENCFF000BVW
K562	CREBBP	ENCFF000BUW
K562	CTCF	ENCFF000BWF
K562	EP300	ENCFF000CAL
K562	EZH2	ENCFF000BWL
K562	H2AFZ	ENCFF000BWT
K562	H3K27ac	ENCFF000BWY
K562	H3K27me3	ENCFF000BXD
K562	H3K36me3	ENCFF000BXJ
K562	H3K4me1	ENCFF000BXQ
K562	H3K4me2	ENCFF000BXV
K562	H3K4me3	ENCFF000BYB
K562	H3K79me2	ENCFF000BYH
K562	H3K9ac	ENCFF000BYN
K562	H3K9me1	ENCFF000BYR
K562	H3K9me3	ENCFF000BYX
K562	H4K20me1	ENCFF000BYZ
K562	HDAC1	ENCFF000BZF
K562	HDAC2	ENCFF000BZL
K562	HDAC6	ENCFF000BZR
K562	KAT2B	ENCFF000CAO
K562	KDM1A	ENCFF000BZV
K562	KDM5B	ENCFF000CBA
K562	NCOR1	ENCFF000BZZ
K562	PHF8	ENCFF000CAT
K562	RBBP5	ENCFF000CBL
K562	REST	ENCFF000CBP
K562	RNF2	ENCFF000CBQ
K562	SAP30	ENCFF000CBV
K562	SETDB1	ENCFF000CBY
K562	SIRT6	ENCFF000CCC
K562	SUZ12	ENCFF000CCH
K562	WHSC1	ENCFF000CAD

Supplementary Table 12

Supplementary Table 12: Datasets used in the TSS vs ChIP-Seq analysis

All files are of "signal" type, in bigWig format, and were obtained from the official ENCODE portal (https://www.encodeproject.org).

All corresponding experiments were performed in Bradley Bernstein's lab at the Broad Institute.

Dataset name	Description	Population size	
		# Transcripts	# Merged TSSs
CLS_FL lncRNA	Merged, full-length captured human transcripts	**10062**	**1036**
GENCODE lncRNA	GENCODE v.20 transcript models of simplified biotype "lncRNA", overlapping exons of CLS_FL lncRNAs on the same genomic strand (obtained using bedtools intersect -split -s -u -a GENCODE_lncRNAs -b CLS_FL_lncRNA)	**3571**	**2148**
GENCODE protein-coding	Subset of GENCODE v.20 transcript models, of transcript_type "protein_coding", matched to CLS_FL lncRNAs for transcript expression in K562 and HeLaS3 cell lines, with CAGE-supported TSSs (i.e., +/- 20 bases from a FANTOM "true" TSS on the same strand)	7,559 (HeLaS3) 7,191 (K562)	5,737 (HeLaS3) 5,418 (K562)

Supplementary Table 13

Supplementary Table 13: Transcript collections used in the TSS vs ChIP-Seq and TSS conservation analyses

Supplementary Methods

Abbreviations

- **FL**: full length
- **HCGM**: High-Confidence Genome Mapping
- **ROI**: read of insert, *i.e.* PacBio read
- **SJ**: splice junction
- **SS**: splice site
- **TM**: transcript model
- **TSS**: Transcription Start Site
- **UMD-ROI**: Uniquely Mapped and Demultiplexed ROI

Post-processing of ROI alignments

Selection of uniquely mapped ROIs

Demultiplexed ROIs mapped uniquely on the genome were selected from the BAM files using the bamflag utility (https://github.com/pervouchine/bamflag) with the "*-m2 -u*" options. This procedure resulted in a set of 1,481,930 (human) and 1,378,896 (mouse) reads, referred to as UMD-ROIs (Uniquely Mapped and Demultiplexed ROIs) hereafter.

Identification of "double-bounded" ROIs

We defined a set of double-bounded reads, namely, UMD-ROIs bounded by a Universal Adapter at one end, and an Indexed Adapter at the other (See schema in Supplementary Figure 2c). We reasoned that such reads should contain the entire cDNA sequence inserted between the two library adapters, and therefore be enriched in fully sequenced inserts.

Globally, about three quarters of uniquely mapped reads were found to be double-bounded both in human (1,130,215 reads, *i.e.* 76%, of which 1,000,213 on-genome) and mouse (1,039,785 reads, *i.e.* 75%, of which 941,757 on-genome). More detailed statistics on double-bounded UMD-ROIs are provided in Table 9.

Identification of poly-adenylated ROIs, on-genome polyA sites and signals

PolyA site calling

We identified poly-adenylated UMD-ROIs and on-genome polyA sites using the *samToPolyA* utility (https://github.com/julienlag/samToPolyA), developed in-house, with the following options: *minClipped=20*, *minAcontent=0.9*, *minUpMisPrimeAlength=10*. That is, we searched for read alignments where a genome match was immediately followed by a final stretch of more than 20 unaligned As or Ts (ignoring adapter sequences, and allowing up to 10% of non-A/non-T nucleotides over the total length of the tail), resulting in a set of potential poly-adenylated reads and on-genome polyA sites. Hits immediately preceded by an upstream A-rich genomic sequence ($>$ 10bp, with $\leq$ 1 non-A bp) were discarded, in order to avoid erroneously calling polyA sites from internally RT-primed cDNAs.

Using this conservative procedure, 731,455 (73%) and 601,114 (64%) reads were found to be poly-adenylated in human and mouse, respectively. Resulting on-genome polyA sites were subsequently merged into clusters using the *bedtools merge* utility [1], using a maximum clustering distance of 5 bases ("*-d 5*"), and forcing strandedness ("*-s*"). Only on-genome polyA site clusters supported by a minimum of 2 reads were kept for further analysis. In total, 35,092 (human) / 27,152 (mouse) non-redundant polyA sites were identified with this procedure. Table 1 summarizes the results of the polyA calling pipeline.

Proximity of polyA signals

We scanned the immediate 5' proximity of our polyA sites for the presence of poly-adenylation signals mentioned in Lopez *et al.*, 2006 [2] (Supplementary Figure 10a). Specifically, we extracted the [-50, -10] sequence window upstream of each non-redundant polyA site, and checked if at least one of those motifs was present in it. We performed the same operation with a collection of negative sites. This latter set was obtained by extracting the middle coordinate of each of our non-terminal PacBio captured exons, distal (+/- 100 bases) to any identified 3' end in our data, and subsequently merging them ("*bedtools merge -d 5 -s*").

Using this method, we established that globally, 86% of observed polyA sites were preceded by a polyA signal in both human and mouse, compared to 12/15% for negative sites, respectively (Supplementary Figure 10c). The same analysis was performed separately on "known" (*i.e.*, sites falling within +/- 50 bases of a GENCODE-annotated 3' end on the same strand) and "novel" (*i.e.*, sites falling more than 50 bases away from of a GENCODE-annotated 3' end on the same strand) polyA sites. We found that although novel sites were slightly more depleted in polyA signals when compared to known ones, they were overall far above the 12/15% random expectation (Supplementary Figure 10d).

ROI genomic strand inference

As PacBio SMRT cDNA sequencing is not directional, we inferred the genome strand of all (including non-demultiplexed) 2,053,424 (human) / 1,870,681 (mouse) uniquely mapped ROIs using the following two methods, in parallel.

"PolyA" approach

We used the *samToPolyA* utility (https://github.com/julienlag/samToPolyA, see PolyA site calling) to assign a genomic strand to poly-adenylated ROIs. Reads where a polyA tail was detected at their 3' end were assigned a '+' genomic strand, whereas reads with a polyT tail at their 5' end were deduced to originate from the '-' strand.

"Splice Junction" (SJ) approach

We extracted part of the SJ sequences (*i.e.* the first and last two nucleotides of each intron) of all ROI unique spliced mappings. We identified, when possible, canonical SJ motifs (GT and AG at the donor and acceptor site, respectively) in each intron of this dataset, and assigned it a genomic strand accordingly: '+' (plus) for GT/AG introns, and '-' (minus) for CT/AC (*i.e.*, the reverse-complement of GT/AG) introns. Each spliced ROI was then assigned a genomic strand based on the inferred strand of the majority of its constituting introns.

Integration of the polyA and SJ approaches

When an ROI could be assigned a genomic strand with both approaches, we found that the agreement between the two methods was 99.7%. Overall, 1,446,986 (70.5%, human) / 1,255,423 (67.1%, mouse) ROIs could be stranded (*i.e.*, assigned a genomic strand) based on at least one method (See Table 10). In rare cases of conflict, priority was given to the strand information obtained *via* the polyA method over the SJ one.

ROI-to-locus/biotype assignment

We assigned each mapped and stranded ROI an originating annotated locus by comparing PacBio mappings to the reference Gencode annotations, Gencode v.20 (human) and v.M3 (mouse), using the *bedtools intersect* program [1] with the following options: *-split* (ignore introns, *i.e.* only exonic overlaps were considered) *-s* (force strandedness) *-wao* (output overlapping entries from both files), only on exon records in both datasets.

Based on this data, we could then assign a unique annotation biotype to each ROI, based on overlapping GENCODE annotations, where available. In most cases, we used the original GENCODE gene_type attribute for this purpose. To simplify, however, ROIs overlapping loci of the following GENCODE gene types were tagged "*lncRNA*", though: "*antisense*", "*lincRNA*", "*processed_transcript*", "*sense_intronic*" and "*sense_overlapping*". When falling outside of GENCODE exonic regions, biotypes were attributed according to the type of capture probe the ROIs overlapped (*e.g.* enhancer, UCE, PipeR, *etc.*). As a last resort, ROIs falling outside of any GENCODE-annotated exon or probed element were tagged "*nonExonic*". When an ROI overlapped exons of multiple biotypes, it was flagged as "*multiBiotype*".

Construction of a HCGM set (High-Confidence ROI Genome Mappings)

We built a collection of High-Confidence ROI Genome Mappings from 1,000,213 (human) and 941,757 (mouse) genome-mapped, double-bounded UMD-ROIs. HCGMs were defined as follows:

- If spliced, read mappings can be composed only of canonical splice junctions (GT or GC as donor site, AG as acceptor site) over their entire mapped length,
- If unspliced, reads need to bear a detectable polyA tail, using the procedure explained in PolyA site calling.

Using these criteria, we identified a set of 771,585 (*i.e.*, 77% of genome-mapped, double-bounded UMD-ROIs) and 604,199 (*i.e.*, 64%) HCGMs in human and mouse, respectively (see Supplementary Figure 2k).

Sequencing error rate estimation

We evaluated the sequencing error rate of the CLS PacBio and HiSeq sequencing output with *qualimap BAMQC* (version 2.2.1)[3]. This software relies on SAM's NM and MD optional attributes for error rate calculations, therefore we re-mapped our reads as detailed in the Online Methods, adding the "*–outSAMattributes NM MD*" option to *STAR*'s [4] parameters.

Qualimap bamqc was run with default options on these BAM files, and the following information was extracted from each library's *genome_results.txt* reports: number of mapped bases, number of mismatches, number of insertions, and number of deletions. We then computed the mismatch, insertion and deletion rates per mapped base in each library (Supplementary Figure 2m). The global error rate was calculated as the sum of mismatches, insertions and deletions, divided by the total number of mapped bases. We observed that PacBio libraries had a $\sim$ 2.1 times higher global error rate than HiSeq ones (1.37x10-3 *vs* 6.5x10-4 errors per mapped base on average, across all human and mouse samples). Both HiSeq and PacBio global error rates were mainly accounted for by sequence mismatches. As expected, non-demultiplexed PacBio reads were enriched in sequencing errors, which might explain why their sample barcode could not be identified in the first place.

Strikingly, PacBio reads were characterized by a much higher rate of insertions (7.5x10-5 vs 4.7x10-6 per mapped base on average, *i.e.* 16 times higher) and deletions (2x10-4 vs 1x10-5 per mapped base on average, *i.e.* 20 times higher) than their HiSeq counterparts. The relatively high rate of PacBio deletion errors casts some doubt on introns detected with this technology, and highlights the need for their systematic confirmation by HiSeq, which was performed in this study (see Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment below).

It should be noted that this analysis considers any sequence difference between RNA-Seq reads and the reference genome as a sequencing error, and therefore does not account for genuine, non-artefactual genome sequence vari-

ation. Thus, the error rates reported here may be slightly over-estimated for both HiSeq and PacBio reads.

Read merging and creation of a full-length lncRNA catalog

Read redundancy was reduced by merging transcript structures with compatible intron chains using the *compmerge* program (https://github.com/sdjebali/Compmerge). We used an original strategy, named "*anchored merging*", which consists in preventing reads with high-confidence boundaries - in our case, supported by a FANTOM true TSS at their 5' end (see Identification of high-confidence Transcription Start Sites using CAGE data below) and/or a captured, PacBio-encoded polyA site at their 3' end - from being merged into another longer read, regardless of their intron chain structure (see Figure 4b). The goal of this extra anchoring step is to preserve all transcript structures with high-confidence TSSs/3' ends, including those falling within exonic regions, which would be lost otherwise.

We anchored polyA- and CAGE-supported HCGMs before merging them using the *anchorTranscriptsEnds* software utility (https://github.com/julienlag/anchorTranscriptsEnds). First, we adjusted all high-confidence 5'/3' ends into clusters. That is, we merged close and overlapping sites using the *bedtools merge* utility, with a maximum clustering distance of 5 bases ("*-d 5*"), and forcing strandedness ("*-s*"). Each individual 5'/3' end belonging to a cluster was assigned its start/end coordinate, respectively - meaning that terminal exons were sometimes extended by a few nucleotides when necessary. In doing so, we ensured that within a cluster, all sites aligned at the exact same position. We subsequently added an "anchor" to all high-confidence, adjusted sites. This step consisted in attaching an artificial, biologically implausible chain of exons (*i.e.*, four 1 nucleotide-long exons, separated by 3 nucleotide-long introns) to each transcript model, upstream or downstream of its high-confidence 5' or 3' end, respectively. These false exons served as anchors to supported start and termination sites during the merging step, and were discarded immediately afterwards.

For comparison, we also performed a standard, "*non-anchored*" merging of HCGMs in parallel. The results of both strategies, across and within our interrogated tissues, are summarized in Supplementary Figure 11a. Following this merging step, we assigned a parent gene as well as a biotype to all merged transcript models (TMs), using the procedure described in ROI-to-locus/biotype assignment.

The end support - *i.e.*, by CAGE true TSS at the 5' end, and poly-adenylation at the 3' end - of each anchor-merged TMs was then deduced from the properties of its constituting ROIs, obtained from the procedures detailed in PolyA site calling and Identification of high-confidence Transcription Start Sites using CAGE data. Accordingly, the full-length set of TMs (referred to as "CLS_FL") consists only of models bounded by such high-confidence 5' and 3' ends. In addition, all their splice junctions are canonical, as they constitute a subset of HCGMs. The results of the read merging and selection of full-length transcript structures are detailed in Table 2, columns a-e.

The end support of transcript models merged using the standard (*i.e.* non-anchored) procedure was deduced not from their constituting ROIs', but rather, from the on-genome comparison of their end coordinates to CAGE TSSs and captured polyA sites (obtained with the methods described in PolyA site calling). 5'/3' ends were considered supported if they laid less than 20/5 bases away from a CAGE TSS / polyA site, respectively, and on the same genomic strand. The results of this comparison are summarized in Supplementary Figures 12c-d (second bar from the left).

Identification of high-confidence Transcription Start Sites using CAGE data

We used CAGE (Cap Analysis of Gene Expression) data produced by the FANTOM consortium [5] to single out high-confidence Transcription Start Sites (TSSs) in our mapped data. To do so, we compared the 5' ends of our HCGMs to the CAGE TSSs identified as "true" TSSs by FANTOM's TSS classifier (http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/TSSpredictionREADME.pdf) across FANTOM-interrogated tissues. The CAGE TSS files were downloaded from http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/ and lifted to hg38 and mm10 using the *liftOver* command-line tool (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/).

Captured TSSs were considered high-confidence (*i.e.*, CAGE-supported) if a FANTOM "true" TSS was found within a window of +/- 20 bases around it, on the same genomic strand (using *bedtools closest* with options "*-s -D b -t first -a <HCGM TSSs> -b <FANTOM true TSSs>*").

In addition, we analyzed the CAGE coverage of "known" (*i.e.*, sites falling within +/- 50 bases of a GENCODE-annotated TSS on the same strand) and "novel" (*i.e.*, sites falling more than 50 bases away from of a GENCODE-annotated TSS on the same strand) PacBio TSSs separately. To do so, for each non-redundant TSS (obtained using *bedtools merge -n -s -d 5*) of the two populations, we computed the distance to the closest FANTOM "true" TSS (using *bedtools closest* with options "*-s -D b -t first -a <HCGM TSSs> -b <FANTOM true TSSs>*").

We observed that novel PacBio TSSs far outnumber known ones in both species (200,425 vs 44,736 in human, 155,083 vs 32,230 in mouse, respectively, see Supplementary Figure 10e). While the CAGE coverage of known sites was higher, thousands of novel TSSs found a CAGE cluster in their close vicinity (+/- 50 bases on the same genomic strand, see Table 11).

Splice Junction analysis

Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment

PacBio Splice Junctions (SJs) were gathered from HCGMs (see Construction of a HCGM set (High-Confidence ROI Genome Mappings)), and as such, they were all canonical (GT|GC / AG). They were assigned a biotype based on that of their originating reads (see ROI-to-locus/biotype assignment). The *IPSA* suite [6] (*Integrative Pipeline for Splicing Analyses*, https://github.com/pervouchine/ipsa-full) was employed to extract SJs and their read counts from *STAR* [4] alignments of Illumina HiSeq data. *IPSA* was run with the default parameters. GENCODE versions 20 and M3 were used as a reference for human and mouse, respectively. All operations were performed on a non-redundant set of distinct SJs, which were uniquely identified by their chromosome, start/end coordinates, and genomic strand. A PacBio SJ was defined as HiSeq-supported if the exact same intron was also observed in the post-capture HiSeq data. HiSeq SJ support was also computed at the level of entire merged CLS transcript models (TMs). Overall, 86.5 % (human) / 87% (mouse) of TMs displayed HiSeq support of their complete intron chain (Table 3). This rate amounted to 91% when considering full-length TMs only.

We proceeded similarly when comparing PacBio SJs to GENCODE-annotated introns: they were flagged as "*known*" when an exact equivalent was found in the comprehensive GENCODE set, and "*novel*" otherwise. The "*known/novel*" status of each SJ was also propagated to its constituting donor and acceptor splice sites (SSs).

A comparison of captured SJs to the human *miTranscriptome* catalog [7] was also performed. We downloaded the GTF data (version 2) from http://mitranscriptome.org/download/mitranscriptome.gtf.tar.gz, converted it to BED,

and mapped its original GRCh37 (hg19) coordinates to GRCh38 (hg38) using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/). At this stage, 377,382 of 384,066 transcripts (98.3%) were successfully lifted over. A biotype was then assigned to these transcripts, following the procedure described in ROI-to-locus/biotype assignment, and we kept only those models of biotype *lncRNA* that overlapped a CLS-probed genomic region (N=32,502). 35,582 unique, canonical (GT|GC / AG) SJs were subsequently extracted and compared to human CLS (Supplementary Figure 6c, top panel). The same analysis was performed on the union of GENCODE 20 and miTranscriptome SJs (N= 42,698) within probed lncRNA regions (Supplementary Figure 6c, bottom panel).

Analysis of splicing motifs

We analyzed PacBio donor and acceptor splice sites (SS) separately. We employed the *geneid* software [8] (version 1.4, with options *-a -d -G -P <parameter file>*) to score individual sites using Position Weight Matrices computed on annotated human genes (parameter file available at https://public_docs.crg.es/rguigo/CLS/data/human.param.Feb_22_2006_GC). The score calculated by *geneid* for a given site *S* corresponds to the log-likelihood ratio of *S* in an actual SS *vs.* *S* in a false SS. We built control ("*Random*") sets of splice sites separately for donor and acceptor sites. To do so, we selected all putative splice sites (GT and GC for donors, and AG for acceptors) within genomic regions overlapped by introns or exons of HCGMs. We then filtered out any site observed as spliced in GENCODE or our PacBio SJs, and scored the remainder with *geneid*, as explained above.

Human-mouse evolutionary conservation of splice sites

The conservation of HiSeq-supported PacBio splice sites between human and mouse was analyzed by mapping "strong" SSs (namely, SSs with positive *geneid* scores, see Analysis of splicing motifs) from one species to the other using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) with "*reciprocal best*" alignment chains (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/vsMm10/reciprocalBest/hg38.mm10.rbest.chain.gz and http://hgdownload.cse.ucsc.edu/goldenPath/mm10/vsHg38/reciprocalBest/mm10.hg38.rbest.chain.gz for human and mouse, respectively). Supplementary Figure 7d-e summarizes the results of the SS *liftOver* step.

A subset of the "*Random*" collection of splice sites (described in Analysis of splicing motifs, and matched for *geneid* scores) was used as a control. This sample was produced by using the *matchDistribution* tool (https://github.com/julienlag/matchDistribution, commit version "*a72706a*", 500 sequential passes, 500 bins, with default options) with the *geneid* score distribution of "*Random*" sites as the "subject" set, and the *geneid* scores of the union of protein-coding and lncRNA sites as the "target" distribution. The result was a sample of "*Random*", unspliced sites matching the splicing strength of both lncRNA and protein-coding SSs in our data (see Supplementary Figures 7f-g).

After mapping high-strength SSs from those three collections from one species' genome to the other, we counted the number of orthologous sites in the destination genome that were also scored positively by *geneid*, as explained in Analysis of splicing motifs. We observed that although much weaker than that of protein-coding sites, the conservation of lncRNA SS strength was overall significantly above background (Chi-square test of conserved/non-conserved sites compared to "*Random*" sites) (see Supplementary Figure 7c).

Intron retention

Intron retention (IR) rates were calculated using *bedtools intersect* [1] on the CLS and GENCODE transcript sets. Note that CLS transcript models whose entire set of introns was not HiSeq-supported (see Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment) were ignored in this analysis. An intron was considered retained if its boundaries were fully contained within at least one individual exon of the same transcript set (*bedtools* option: "*-f1*") and on the same strand (*bedtools* option: "*-s*"). The IR rates

reported in Supplementary Figure 6f were calculated as the proportion of transcripts with a least one intron retained.

Identification of novel transcript structures

We used the *comptr* program (https://github.com/sdjebali/Comptr) to compare the intron chains of our merged TMs (the "*assessed*" set, obtained as described in Read merging and creation of a full-length lncRNA catalog) to the comprehensive set of GENCODE 20 and M3 transcripts (the "*reference*" set). We considered novel only the transcripts categorized by *comptr* as "*Extension*" (*i.e.*, there is a reference transcript with all its introns equal to the assessed transcript but the assessed transcript has additional introns), "*Intergenic_or_antisense*" (*i.e.*, the assessed transcript is stranded and spliced but does not overlap any reference transcript on the same genomic strand) and "*Overlap*" (*i.e.*, there is a reference transcript overlapped by the assessed transcript on the same strand). The results of the full-length TM structure comparison are summarized in Figure 4e-f, and detailed in Table 2, which also reports the number of annotated loci giving rise to novel FL structures.

Simulated read depth versus discovery rate

In order to evaluate the completeness of our post-capture annotation - *i.e.*, how close it is to saturation - we calculated the number of novel SJs and novel transcript structures discovered at increasing ROI sequencing depths in each tissue sample. We randomly sampled ROIs from unfiltered BAM files (that is, including unmapped ROIs) at increasing depths, in increments of 20,000 reads until the total of available reads was reached in each tissue sample. A combination of samtools [9] and standard GNU Linux utilities (*head*, *shuf*) was employed for that purpose.

We then counted the number of novel individual SJs (procedure described in Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment), or novel intron chains (see Identification of novel transcript structures) the sampled ROIs gave rise to. Each randomization at a given read depth was repeated 50/100 times for individual SJ and full intron chain simulations, respectively. When assessing the novelty of individual introns, we stratified the SJs generated at each read depth by level of sequencing support (all PacBio junctions, PacBio junctions with HiSeq support, and PacBio junctions without HiSeq support). Results of the simulations are presented in Figure 3d, as well as Supplementary Figure 8a-b. Results of the simulations using HiSeq short-read sequencing of captured cDNA are reported in Supplementary Figure 8c.

Analysis of protein-coding potential

The set of full-length transcript models were used as input for the programs CPAT [10] and PhyloCSF [11]. CPAT uses intrinsic sequence properties to predict coding potential. PhyloCSF, in contrast, uses evolutionary signatures of selection on coding sequences.

CPAT was run according to creator's protocol [10]. Hexamer tables and logit models were created using the *Human_ORF.fa* and *Human_NONCODE.fa* files and used for both human and mouse analyses. We used the cutoff value of 0.364 to distinguish coding from noncoding transcripts.

PhyloCSF was run on spliced alignments of transcripts using a custom pipeline. GTF format annotations were used to extract multiple alignment file (MAF, obtained from the UCSC Genome Browser for hg38) format blocks, which were stitched together to recreate the multiple alignment of the processed transcript. This was converted to a FASTA file and input to PhyloCSF. The parameter file used was "*29 mammals*". The settings used were "*–dna –aa –frame=3*

–removeRefGaps –orf=ATGStop –minCodons=20". A score threshold of 100 was used to define protein-coding transcripts.

Analysis of cytoplasmic/nuclear localization

PolyA RNA sequencing data from whole cell, nucleus and cytoplasm of ten human cell lines was obtained from the ENCODE portal (https://www.encodeproject.org/) [12, 13]. An annotation in GTF format was constructed by combining the annotation of merged transcript models with GENCODE v20. The *GRAPE* pipeline [14] (https://github.com/guigolab/grape-nf) was used to quantify these models, using *STAR* [4] (v.2.4.0j) with RSEM [15] (v.1.2.21) for quantification. The following non-default parameter was specified for *STAR*: "*–outFilterMismatchNmax 4*". *RSEM* was used in transcriptome mode. Next, for every full length CLS model, the log2 ratio was calculated of cytoplasmic to nuclear RPKM values. Any transcript model with a zero value in either compartment was discarded.

Evaluation of Illumina-based transcript reconstruction methods in matched samples

Global assessment of reconstruction software accuracy

We comprehensively assessed the accuracy of the short-read transcript reconstruction algorithms *StringTie* [16] and *Cufflinks* [17] using CLS transcript models (TMs) in matched samples as a gold standard. Capture HiSeq reads were mapped to the corresponding reference genome (hg38 and mm10) using *STAR*, as described in the Online Methods, adding the "*–outSAMattributes XS*" parameter, in order to comply with *StringTie* and *Cufflinks* requirements when used with unstranded reads. Reads mapping to ERCC spike-in sequences were discarded.

StringTie (v1.3.3) was run with default parameters except "*-p6*" (*i.e.*, 6 CPU threads), and *Cufflinks* (v2.2.1) with the "*–multi-read-correct*" and "*-p6*" options. Running on human data, *Cufflinks* hung for more than 10 days on a single region (chr12:65981298-65981423 in hg38), and thus had to be restarted with the offending region masked.

StringTie and *Cufflinks*'s respective outputs were then compared to the full set of 94,163 (human) / 72,413 (mouse) HiSeq-supported, standard-merged CLS TMs (see Read merging and creation of a full-length lncRNA catalog and Extraction of Splice Junctions and Splice Sites, HiSeq support and novelty assessment) as the reference annotation file. We obtained the corresponding sensitivity and precision measures using *gffcompare* (v0.9.9c) (https://github.com/gpertea/gffcompare), run with options "*-N*" and "*-M*" (*i.e.*, ignore single-exon transcripts and transfrags in both reference and test sets). While fair at the "base" and "intron" levels, sensitivity and precision were particularly poor at the "intron chain" and "transcript" levels for both programs (Supplementary Figure 12a). *StringTie* substantially and consistently outperformed *Cufflinks* at all accuracy levels, and we therefore decided to further analyze only *StringTie* models for the sake of simplicity.

StringTie produced a total of 94,082 (human) and 171,439 (mouse) distinct transcripts, merged across all assayed tissues in each organism. Of those, 65,060 (human, *i.e.* 69%) and 52,412 (mouse, *i.e.* 31%) could be assigned a genomic strand by the program (all unstranded models were single-exon transfrags, and were ignored in the rest of the analysis).

Following the procedure used for CLS models (see ROI-to-locus/biotype assignment), we found that 13,930 (human) and 2,920 (mouse) stranded *StringTie* models originated from probed lncRNA genomic regions. We then extracted intron chains from these models and performed a 3-way comparison with CLS and GENCODE transcript models falling within targeted lncRNA regions (Figure 4h and Supplementary Figure 12b). Spliced length statistics of these *StringTie*

TMs are presented in Figure 4i (human) and Supplementary Figure 12c (mouse), side-by-side with the CLS TM set.

End support of CLS and *StringTie*-reconstructed transcripts

We analyzed the end support of stranded *StringTie* models and compared it to CLS TMs. Since polyA tails are not preserved in Illumina-based reconstructed transcripts, we independently called polyA sites by applying the method described in PolyA site calling to uniquely mapped capture HiSeq reads.

Requiring a minimum of 2 reads supporting a given site, we could detect only 2,572 (human) / 2,278 (mouse) distinct polyA sites, that is, 14 (human) / 12 (mouse) times less than when using PacBio ROIs in matched samples with the same parameters (Table 1). We attribute this lack of sensitivity to the well-documented relative depletion of HiSeq reads towards the ends of transcripts [18].

Failure to properly resolve polyA sites using HiSeq data led us to evaluate *StringTie* models' 3' end completeness with CLS-called polyA sites instead. The full-length status assessment of *StringTie*-reconstructed transcripts was, as a result, performed exactly as for standard (*i.e.*, non-anchored) -merged CLS TMs (see Read merging and creation of a full-length lncRNA catalog), for a fair comparison. Only 4,633 (*i.e.* 7%, human) / 3,646 (*i.e.* 7%, mouse) *StringTie* transcripts were considered full-length using our criteria, a much lower rate than the one observed in the standard-merged CLS TM set (28,186, *i.e.* 24% in human, and 18,934, *i.e.* 22% in mouse, see Supplementary Figure 12d-e for a side-by-side comparison). The fraction of full-length *StringTie* TMs decreased immensely within probed lncRNA regions (116/13,930, *i.e.* 0.8% in human, and 32/2,920, *i.e.* 1.1% in mouse).

The 5' and 3' completeness of *StringTie* and CLS TMs were further analyzed and compared with the following datasets: GENCODE lncRNAs (5' and 3' ends from annotated lncRNAs originating *StringTie* or CLS TMs), GENCODE protein-coding transcripts (a confident set of protein-coding transcripts, not tagged *mRNA_end_NF* nor *mRNA_start_NF* in the original GENCODE GTF files), and a control set of sites (middle coordinate of internal exons, as described in Proximity of polyA signals). All sets of sites were individually clustered to reduce redundancy ("*bedtools merge -d 5 -s*").

We then assessed the proximity of each TSS to FANTOM5 CAGE true TSSs (as described in Identification of high-confidence Transcription Start Sites using CAGE data), and of each 3' end to canonical polyA signal motifs (PAS, as described in Proximity of polyA signals) using a combination of *bedtools slop* and *bedtools intersect* [1]. We considered a TSS CAGE-supported ("CAGE(+)") if a CAGE cluster could be found +/-50 bases around them, on the same strand. Similarly, "PAS(+)" 3' ends are those CLS polyA sites falling 10 to 50 bases downstream of a PAS motif (Figure 4i and Supplementary Figure 12f).

In addition, we present aggregate plots of PAS and CAGE TSSs around various sets of transcript ends in Supplementary Figure 12g-h.

Genome repeat coverage

Repeat elements in both mm10 and hg38 were downloaded from the UCSC Table Browser (`https://genome.ucsc.edu/cgi-bin/hgTables`, *RepeatMasker* track) in tabular format. Repeat features were split into different classes according to their *repClass* attribute, converted to BED format and projected on the genome using *bedtools merge* [1]. Next, repeat elements were compared to projected exons from the CLS, GENCODE and *StringTie* TM sets of all biotypes, using *bedtools coverage*. Only stranded models were considered in the *StringTie* TM set. The fraction of exonic nucleotides covering genome repeats of various classes in each set of TMs is reported in Supplementary Figure 12i.

Estimating capture sensitivity using spike-ins

Inspection of sensitivity curves for spike-ins in individual tissues (Supplementary Figure 1e) shows a detection threshold around 5.6x10-2 attomol (-1.25 in log10 units) for captured molecules. In 4 µg of a 1:100 dilution of spike-in RNA that was added to 4 µg of each RNA sample, this threshold is equivalent to 1344 molecules. We assume here that the total RNA content of a single cell is 5 pg [19], making this threshold equate to 7x10-3 molecules per cell. Non-captured sequences' detection threshold lies approximately 30-fold higher (1.5 log10 units), or 0.21 molecules per cell.

TSS overlap analysis

Coordinates of indicated features were downloaded and mapped to hg38 using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) where appropriate. CpG islands were downloaded from the UCSC Table Browser (hg38) (https://genome.ucsc.edu/cgi-bin/hgTables).

Promoter and enhancer maps were downloaded as all files corresponding to 15-state *ChromHMM* predictions brain, heart and liver from Epigenome Roadmap [20]. These coordinates were merged separately by promoter (states: 1, 2, 3, 4, 14) / enhancer predictions (states: 7, 8, 9, 10, 11, 15). GWAS SNPs (gwas_catalog_v1.0-associations_e84_r2016-05-08.tsv) were obtained from the GWAS Catalog [21]. Conserved elements, obtained using *PhastCons* [22] 46-way primate alignments, were downloaded from the UCSC Genome Browser (hg38).

Comparison of human TSSs with DNase-Seq (DHS), ChIP-Seq and conservation tracks

Input datasets

We compared various sets of human TSSs to ENCODE ChIP-Seq (in cell lines K562 and HeLa, see "signal" BigWig file list in Table 12), DNase-Seq (HeLa-S3 DNase Hypersensitive Sites hotspots, downloaded from https://www.encodeproject.org/files/ENCFF968ECA/) and conservation (*phastCons* [22] scores downloaded from http://hgdownload-test.cse.ucsc.edu/goldenPath/hg38/) tracks. At the time of the study, the relevant ENCODE ChIP-Seq signal files were only available on human assembly hg19, therefore we mapped our TSSs to this genome version using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) before proceeding with the ChIP-Seq comparison.

The three collections of transcripts used as input were "*CLS_FL lncRNA*" (described in Read merging and creation of a full-length lncRNA catalog), "*GENCODE lncRNA*" (*i.e.*, GENCODE-annotated lncRNAs detected by CLS), and "*GENCODE protein-coding*" (defined in Transcript expression matching of the GENCODE protein-coding set below). The TSS sets analyzed consisted in a "*standalone*" version, where transcripts originating from bi-directional promoters were filtered out. These were generated by removing all transcripts whose TSS fell within 1,000 bases upstream of any GENCODE or captured FL TSS on the opposite genomic strand. Given the fuzzy nature of ChIP-Seq and DNase-Seq peaks, we merged TSSs within each set using a rather large maximum clustering distance of 200 bases ("*bedtools merge -s -d 200*" [1]).

The basic characteristics of each TSS dataset used in the ChIP-Seq and conservation analyses are reported in Table 13.

Transcript expression matching of the GENCODE protein-coding set

We selected a subset of protein-coding transcripts with expression similar to that of CLS_FL lncRNAs in K562 and HeLa cells. First, we merged the

CLS_FL transcript models with GENCODE v.20, then quantified the resulting transcripts in K562 and HeLaS3 ENCODE polyA+ whole-cell RNA-Seq experiments *ENCSR000CPH* and *ENCSR000CPR* (downloaded from the ENCODE portal, https://www.encodeproject.org), respectively.

The transcript quantifications were computed using *GRAPE* [14] (https://github.com/guigolab/grape-nf, commit version "*bcaa6688b9*", bundling *STAR* [4] version 2.4.0j and *RSEM* [15] version 1.2.21) running under the *NextFlow* framework [23] (version 0.17.3, https://www.nextflow.io/).

We then extracted the posterior mean estimate FPKM ("*pme_FPKM*") values of CLS_FL lncRNA and GENCODE protein-coding transcripts from *RSEM* output. The subset of expression-matched GENCODE protein-coding transcripts was obtained using the *matchDistribution* software tool (https://github.com/julienlag/matchDistribution, commit version "*a72706a*", 10 sequential passes, 500 bins, with option "*–transform=log10*") with GENCODE protein-coding *pme_FPKM*s as the "*subject*" set, and CLS_FL lncRNA as the "*target*" distribution, separately on K562 and HeLaS3.

The results of the expression matching are presented in Supplementary Figure 13f-g and Table 13.

Aggregate plots of signal density surrounding TSSs

We employed *bwtool* [24] (https://github.com/CRG-Barcelona/bwtool) to produce aggregate plots of ChIP-Seq read density and conservation scores on the aforementioned TSS collections, using the following command: "*bwtool agg -long-form -header -expanded -firstbase 10000:10000 <TSS set> <signal BigWig file> output.txt*".

The mean signal and standard error of the mean were extracted from *bwtool*'s output and plotted as a function of the nucleotide position around TSSs (Supplementary Figures 14 and 15).

Comparison of TSSs and DNase Hypersensitive Sites (DHS) in HeLa cells

HeLa-S3 DHS peak ("hotspot") BED files were first converted to the BigWig format using "*bedtools genomecov -bga*" [1] followed by *bedGraphToBigWig* (http://hgdownload.soe.ucsc.edu/admin/exe/).

The resulting BigWig files were subsequently compared only to the "*raw*" merged TSSs of transcripts detected in HeLa by CLS, using "*bwtool agg -long-form -header -expanded -firstbase 10000:10000*" to obtain the mean DHS hotspot density at each base surrounding TSSs. The "*GENCODE protein-coding*" set was "expression-matched" in HeLa-S3, as described in Transcript expression matching of the GENCODE protein-coding set. The "*GENCODE lncRNA*" set consisted in GENCODE v.20 lncRNA transcripts detected by CLS in HeLa-S3.

Testing predicted peptides

Using a published proteogenomics workflow [25], we searched the Human Proteome Project (C-HPP) [26] database of testis peptides for those matching predicted ORFs, but found no hits at a threshold of 0.01 PEP.

Identifying lncRNA orthologues

We defined orthology using *MultiZ* sequence alignments [27]. Taking the entire genomic span of GENCODE lncRNA gene annotations, we created human-to-mouse and mouse-to-human orthology mappings using *liftOver* (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/) with chain files from the UCSC Genome Browser (hg38 -> mm10, mm10 -> hg38). *liftOver* was run with "*-minMatch=0.8*" (minimum fraction of nucleotides mapped), and all other options set to default.

Orthology was then defined using strand-specific intersection, requiring a minimum of 5% of the genomic span of both elements to overlap. Orthologue lists were now defined in two ways: "*Reciprocal*" requires reciprocal mapping in both directions (131 pairs where both have an ID from GENCODE 20 / M3); "*Union*" is the union of both directions mappings, without requiring reciprocal hits (293). Note that these numbers refer to the entire lncRNA annotation. The subset of these reciprocal pairs was then obtained that map to the probed lncRNA annotation in either species. We proceeded with the larger *Union* set to boost statistical significance, being 101/84 orthologous and probed genes, for human and mouse respectively. This set we defined as "conserved" having an orthologue in the other species. Amongst these were cases such as *SNHG17* and *MALAT1*.

We asked whether conserved lncRNAs were more likely to be detected (having >0 mapping reads, using either all reads or only full-length reads). We also compared the number of reads between conserved and non-conserved lncRNAs. Both analyses were performed separately in each species, and presented in Supplementary Figure 11d-g.

RT-PCR experimental validation of CLS transcript models

500 ng of total RNA from HeLa, brain and testis samples (the same we used for the capture assays) were used for retrotranscription. Retrotranscription was performed with ReverseAid retrotranscriptase (Thermo Scientific), using both oligo-dT and random hexamers as primers, following the manufacturer's protocol.

For CARMN and KANTR, testis and brain cDNAs, respectively, were amplified for 40 cycles at 56°C annealing. For CASC19, HeLa cDNA was amplified for 40 cycles at 56°C annealing, gel purified and amplified for 40 more cycles to enrich for specific bands. The SAMMSON transcript was amplified from testis, for 40 cycles at 54°C annealing with Expand polymerase.

PCRs were performed with KOD DNA Polymerase (Novagen) using primers to be found in Supplementary Data 3. The amplicons were sequenced using Sanger sequencing and are available in Supplementary Data 4.

Bibliography

1. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47,** 11.12.1–11.12.34. ISSN: 1934-340X (2014).
2. Lopez, F, Granjeaud, S, Ara, T, Ghattas, B & Gautheret, D. The disparate nature of "intergenic" polyadenylation sites. *RNA* **12,** 1794–1801 (2006).
3. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32,** 292–294. ISSN: 1367-4803 (2016).
4. Dobin, A *et al.* STAR: ultrafast universal RNA-seq aligner. eng. *Bioinformatics* **29,** 15–21 (2013).
5. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–70. ISSN: 1476-4687 (2014).
6. Pervouchine, D. D., Knowles, D. G. & Guigó, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29,** 273–274. ISSN: 13674803 (2013).
7. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47,** 199–208. ISSN: 1061-4036 (2015).
8. Blanco, E., Parra, G. & Guigó, R. in *Curr. Protoc. Bioinforma.* Unit 4.3 (John Wiley and Sons, Inc., Hoboken, NJ, USA, 2007). doi:10.1002/0471250953.bi0403s18.
9. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079. ISSN: 1367-4803 (2009).
10. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41,** e74–e74. ISSN: 1362-4962 (2013).
11. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** i275–i282. ISSN: 1367-4803 (2011).
12. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8. ISSN: 1476-4687 (2012).
13. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 1476-4687 (2012).
14. Knowles, D. G., Roder, M., Merkel, A. & Guigo, R. Grape RNA-Seq analysis pipeline environment. *Bioinformatics* **29,** 614–621. ISSN: 1367-4803 (2013).
15. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323. ISSN: 1471-2105 (2011).
16. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295. ISSN: 1087-0156 (2015).
17. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. eng. *Nat Protoc* **7,** 562–578. ISSN: 1754-2189 (2012).
18. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32,** 915–925. ISSN: 1087-0156 (2014).

19. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21,** 1160–1167. ISSN: 1088-9051 (2011).

20. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330. ISSN: 0028-0836 (2015).

21. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901. ISSN: 0305-1048 (2017).

22. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050. ISSN: 1088-9051 (2005).

23. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35,** 316–319. ISSN: 1546-1696 (2017).

24. Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30,** 1618–9. ISSN: 1367-4811 (2014).

25. Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat. Commun.* **7,** 11778. ISSN: 2041-1723 (2016).

26. Zhang, Y. *et al.* Tissue-Based Proteogenomics Reveals that Human Testis Endows Plentiful Missing Proteins. *J. Proteome Res.* **14,** 3583–94. ISSN: 1535-3907 (2015).

27. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–15. ISSN: 1088-9051 (2004).