Supplementary materials

Improved definition of the mouse transcriptome via targeted RNA sequencing

Giovanni Bussotti, Tommaso Leonardi , Michael B. Clark, Tim R. Mercer, Joanna Crawford, Lorenzo Malquori, Cedric Notredame, Marcel E. Dinger, John S. Mattick and Anton J. Enright °

° Correspondence:
Anton J. Enright
Tel: + 44 (0) 1223 492 668
Fax: + 44 (0) 1223 492 620
Email: aje@ebi.ac.uk

Supplemental Table 1

Table summarizing the overlap between the non-GENCODE HQ genes and other mouse datasets (Pruitt et al. 2012; Karolchik et al. 2004; Necsulea et al. 2014; Hezroni et al. 2015; Lin et al. 2014). Genes are deemed to be overlapping if they share at least one exonic base in the same orientation. For datasets without strand information (Necsulea et al. 2014) the overlap between genes is deemed if they share at least one exonic base, regardless the exon orientation. The last column to the right shows the results considering a less permissive overlap criterion. Genes are deemed to overlap if they share at least a transcript with a reciprocal coverage of at least 50%.

Dataset	Overlapping genes	Overlapping genes (min 0.5 fraction)
RefSeq	257	156
UCSC	481	291
Necsulea A et al. (stranded)	545	345
Necsulea A et al. (unstranded)	625	404
Lin S et al.	3,443	1051
Hezroni H et al.	274	8
Non redundant Total	**3,624**	**1,244**

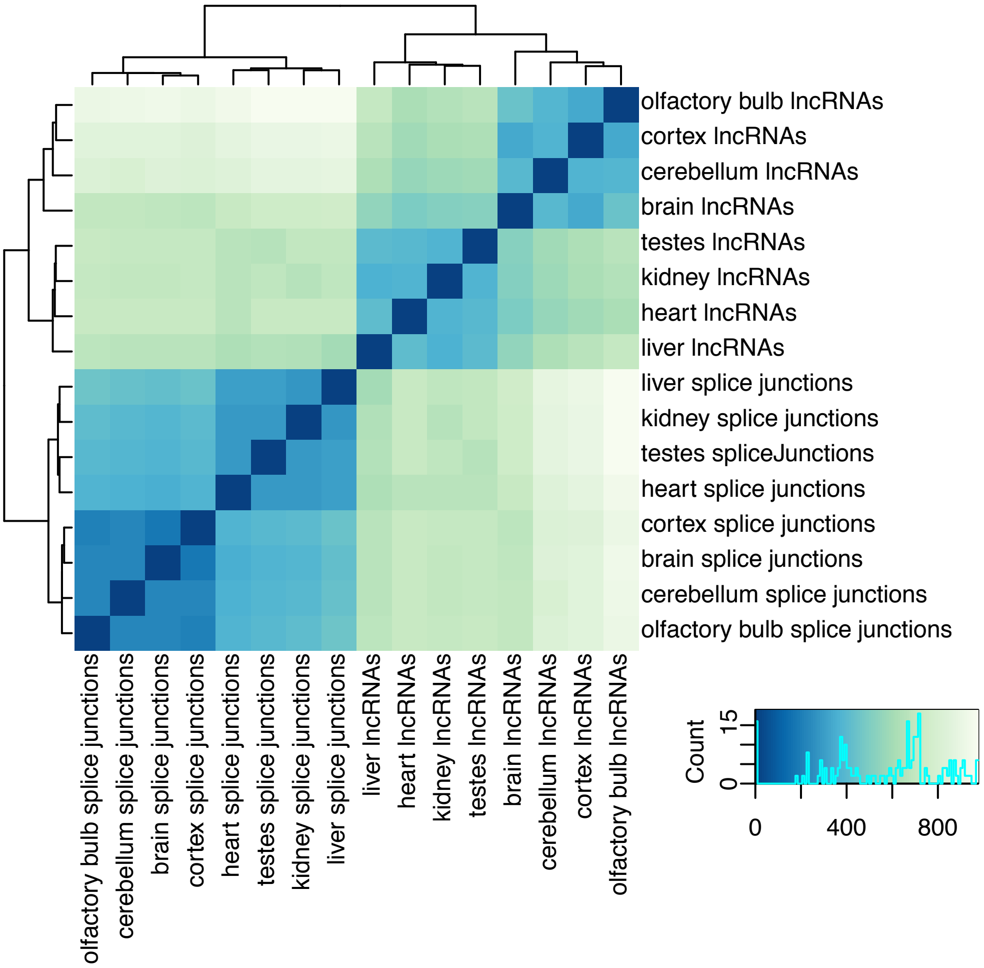
Supplemental Table 2

Table recapitulating relevant genomic and expression features of the transcripts in the HQ set. We used the set of mouse enhancer available from https://www.ebi.ac.uk/research/flicek/publications/FOG15 (Villar et al. 2015). We defined promoters the regions 1 kb upstream GENCODE TSSs. The field "genes co-expressed with neighbor GENCODE genes" reports the gene having a Pearson expression correlation of at least 0.8 with any of the neighbor GENCODE genes (considering max 3 neighbors on each side).

Not overlapping GENCODE exons	**Genes**	**Transcripts**
intergenic transcripts	4,963	5,731
intergenic transcripts, non enhancer	4,609	5,194
transcripts overlapping enhancers	354	537
antisense transcripts	818	1,075
intronic transcripts	2,279	2,311
promoter overlapping transcripts	1,063	1,181
genes with at least one spliced isoform	1,511	2,785
genes with at least one junction supported by at least 5 reads	1,235	2,030
genes co-expressed with a neighbor GENCODE gene	420	463
overlapping GENCODE exons		
transcripts with at least one spliced isoform	8,351	16,366
transcripts with at least one junction supported by at least 5 reads	6,796	11,306
genes co-expressed with a neighbor GENCODE gene	359	651
transcripts connecting two or more annotated GENCODE genes	400	1,174
protein coding transcripts previously annotated as lncRNAs	104	104
transcripts extending upstream of annotated GENCODE gene starts	816	1,914
transcripts extending downstream of annotated GENCODE gene ends	1,080	2,544
transcripts overlapping lncRNA CaptureSeq probes but not splice junction probes	1,208	1,682
transcripts overlapping splice junction CaptureSeq probes but not lncRNA probes	3,775	5,702
transcripts overlapping both splice junction probes and lncRNA probes	4,670	9,324

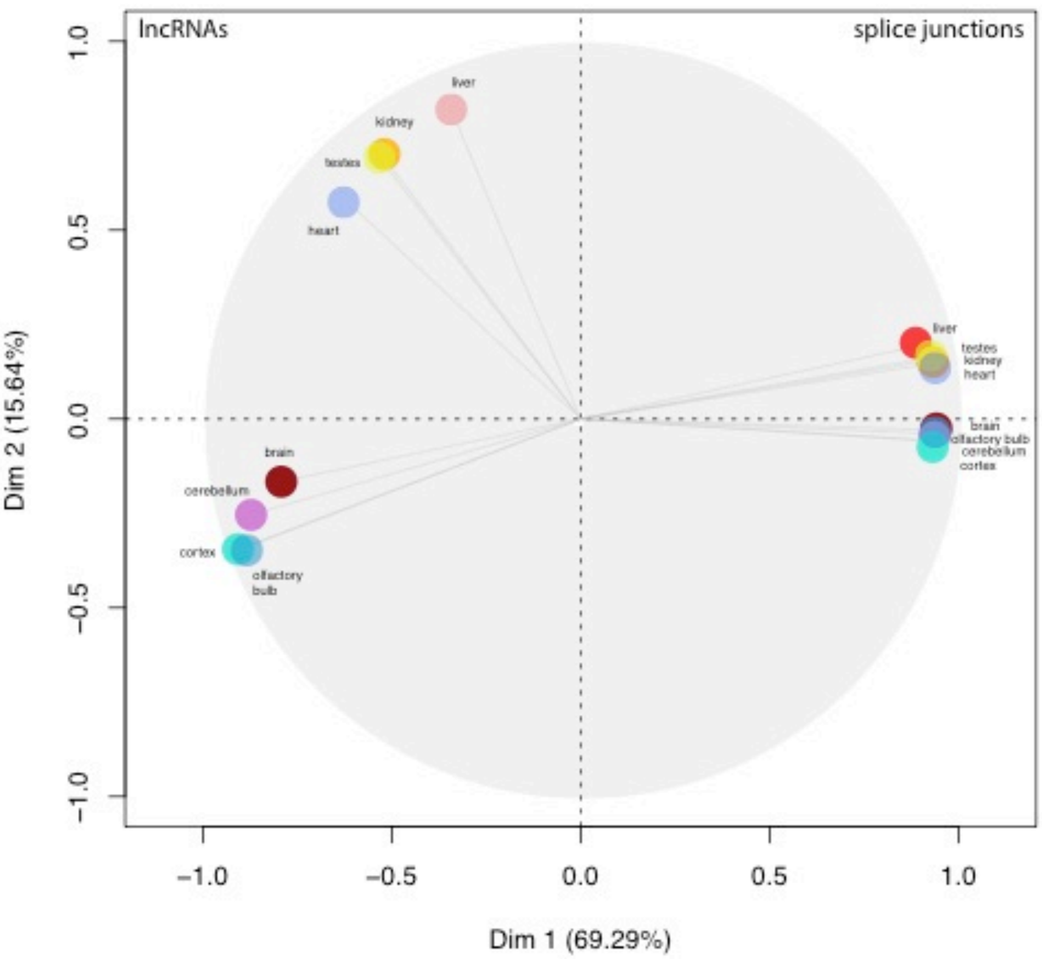
Supplemental Figure 1

Heatmap representing the Euclidean distances between the samples as calculated from the regularised log counts.



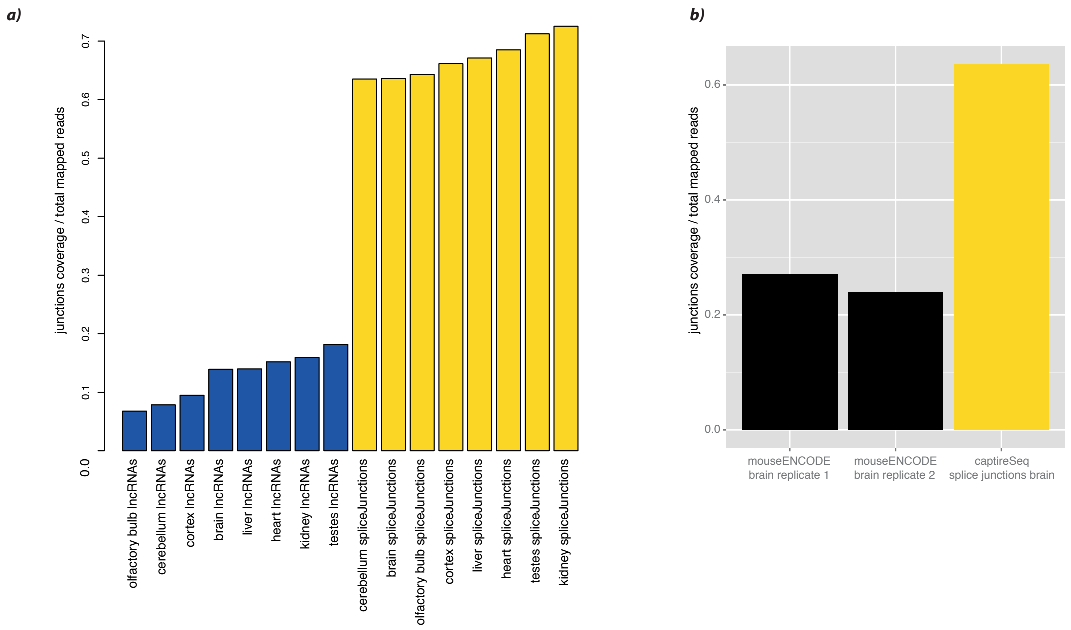
Supplemental Figure 2

PCA loadings of each sample. The first dimension explains 69.29% of the variation and separates the lncRNA and splice junction CaptureSeq designs. The second dimension explains 15.64% of the variation and separates brain related samples from the others.



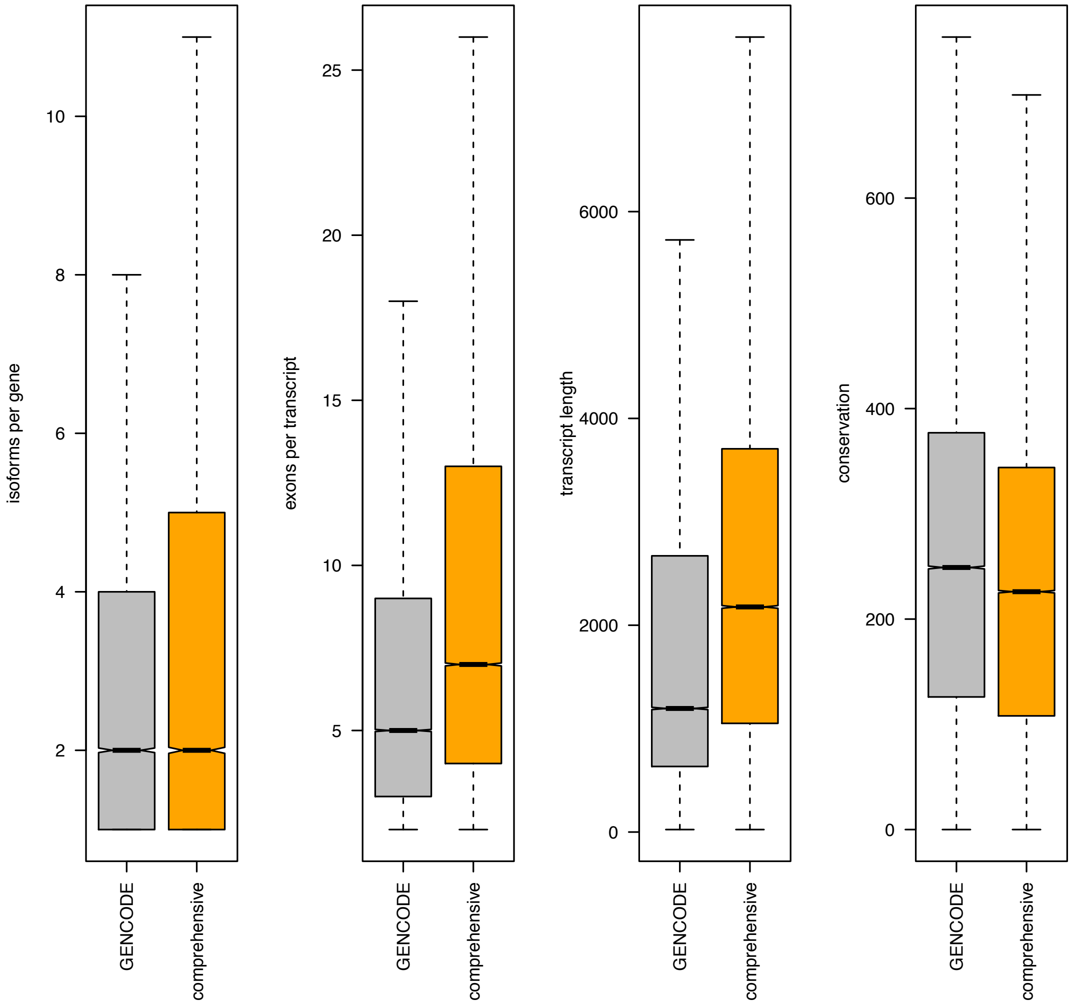
Supplemental Figure 3

Barplots showing the fraction of reads crossing splice junctions with respect to total mapped reads for each of our samples ***a)*** and two mouseENCODE brain RNAseq samples (see methods) ***b)***.



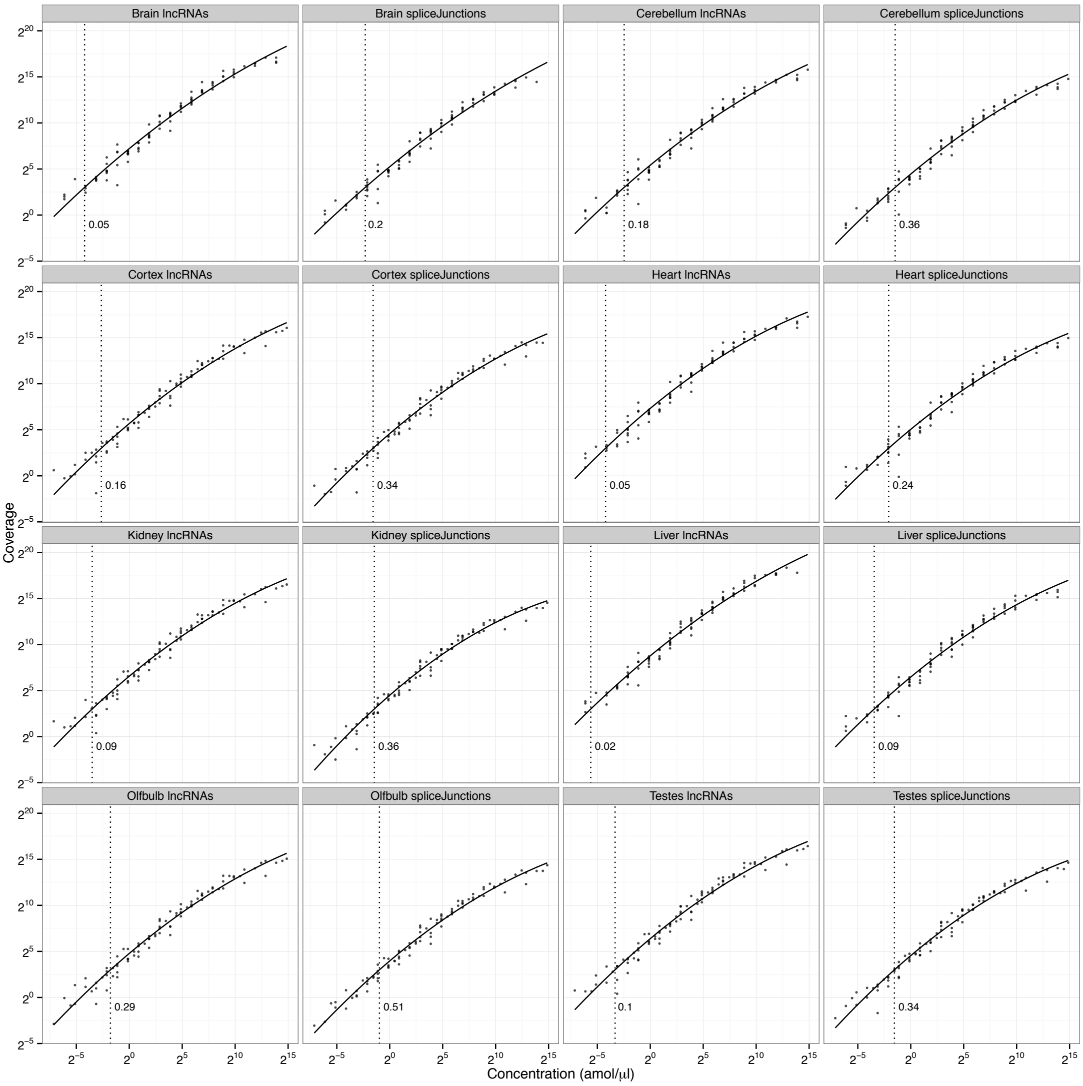
Supplemental Figure 4

Boxplot panels comparing GENCODE (M4) (grey) versus the merged comprehensive assembly (orange). Monoexonic transcripts are not shown. From left to right the first and the second panels report the number of isoforms and the number of exons per gene locus. The third panel indicates the mature transcript lengths. The last panel indicates the phastCons evolutionary conservation, which is based on a hidden Markov model that estimates the conservation probability of each nucleotide based on its context.



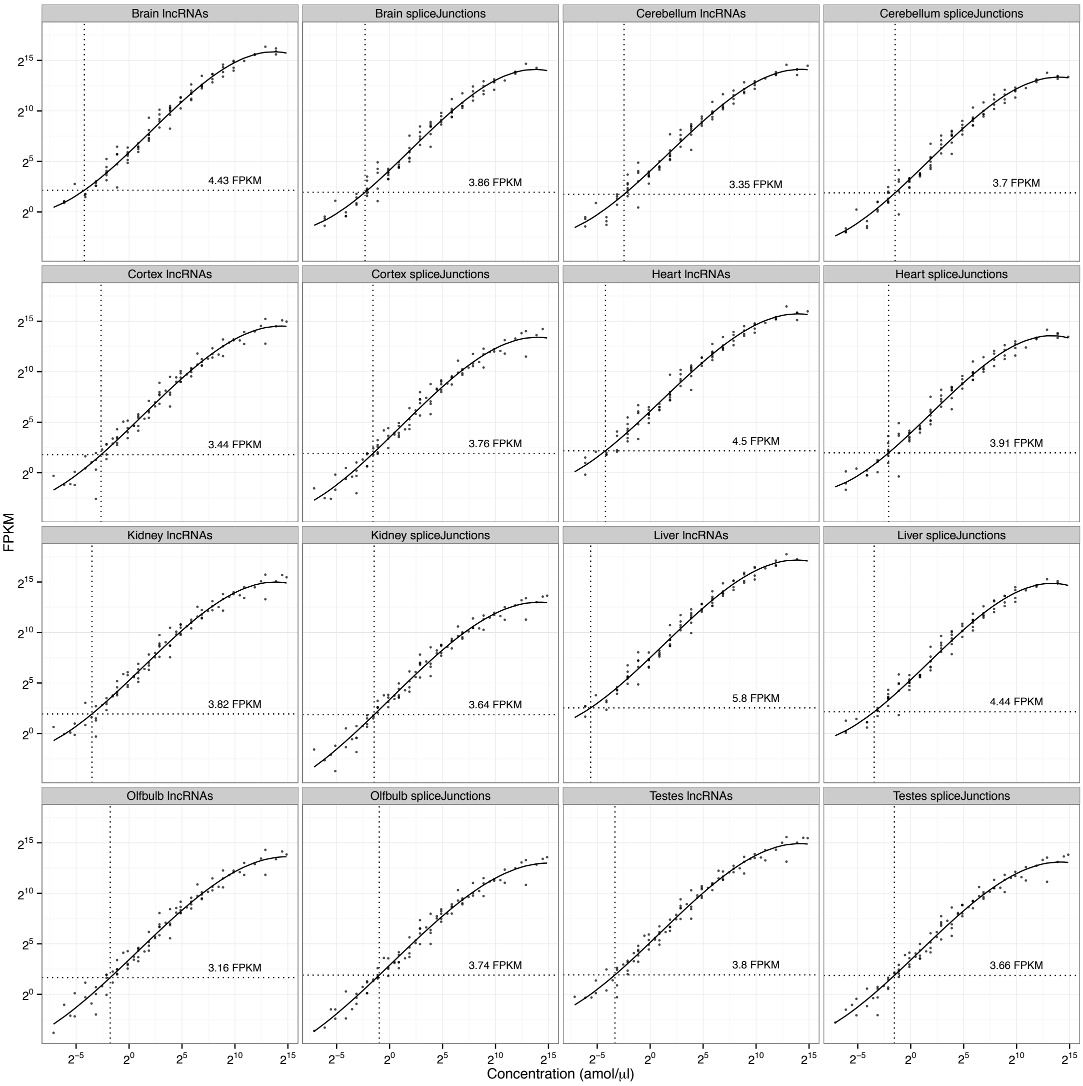
Supplemental Figure 5

Scatterplots showing for each sample the ERCC spike-in known concentrations (x-axes) versus the mean sequencing coverage (y-axes). The black lines are polynomial fits of second degree between the coverage (log2) and the known concentrations (log2) of the ERCCs. The vertical dotted lines indicate the concentrations (also reported beside the lines in attomol (amole)/µl) at which the fitted coverage is 8.



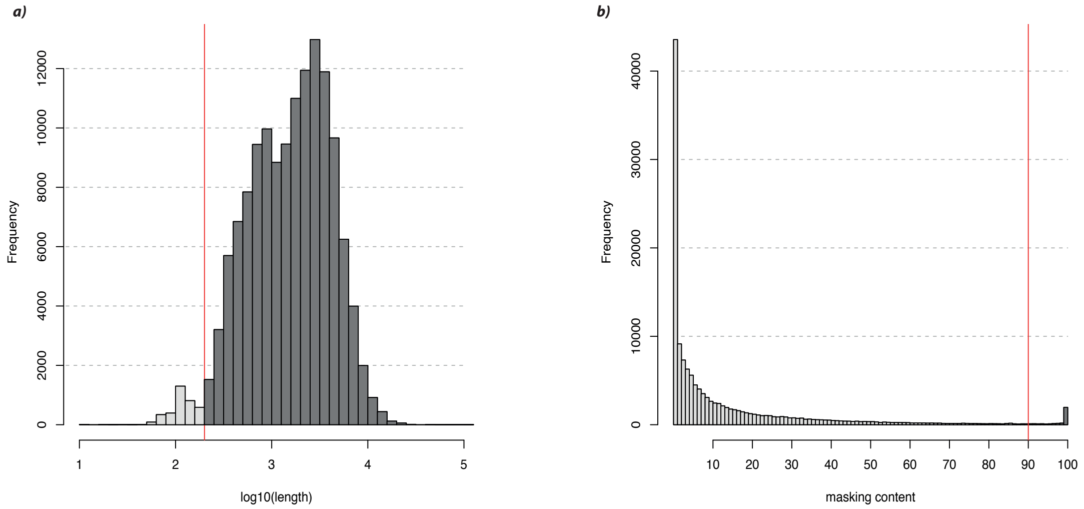
Supplemental Figure 6

Scatterplots showing for each sample the ERCC spike-in known concentrations (x-axes) versus the FPKM values calculated by *Cufflinks* (y-axes). The black lines are polynomial fits of third degree between the FPKM (log2) and the known concentrations (log2) of the ERCCs. The vertical dotted lines indicate the ERCC concentrations at which coverage is 8, while the horizontal dotted lines indicate the corresponding FPKM values calculated from the fit.



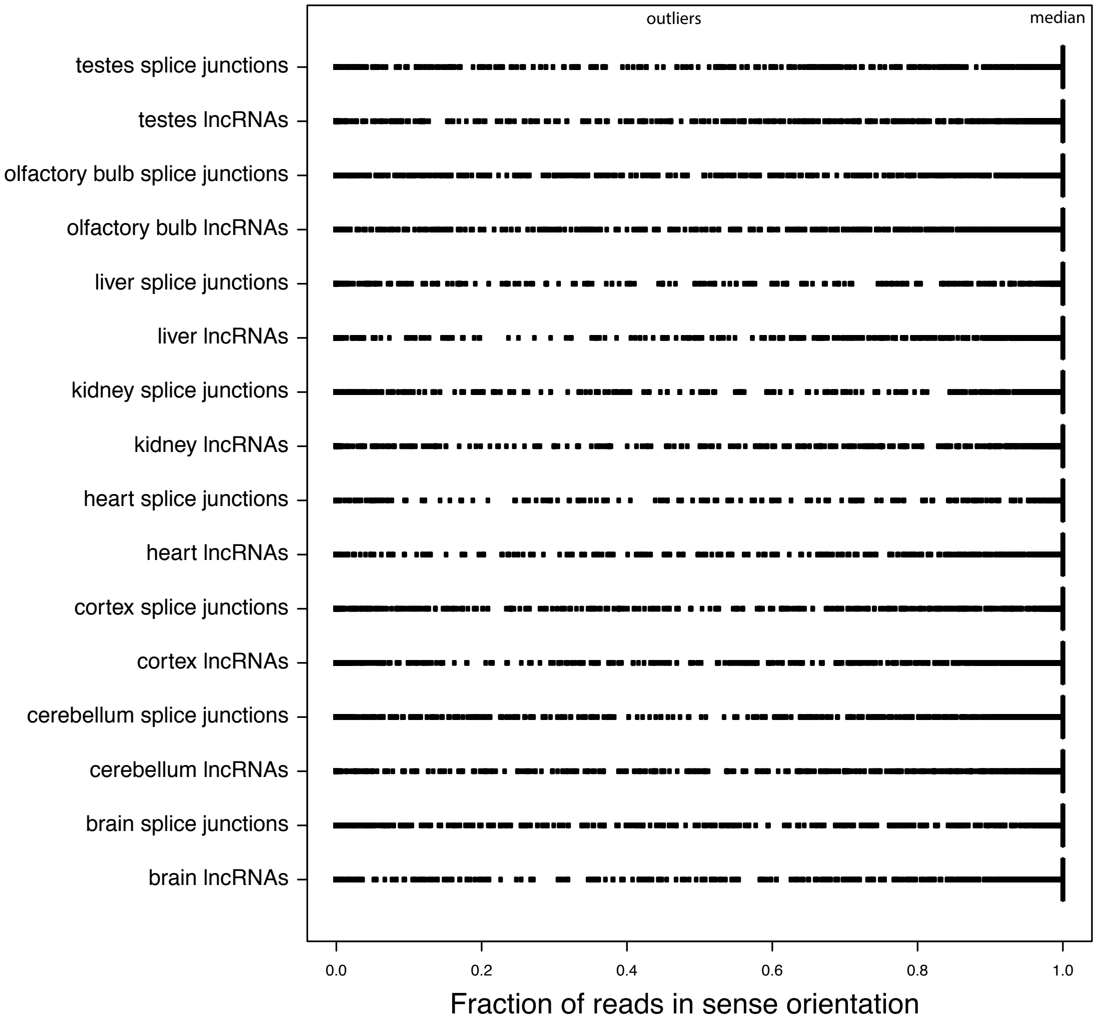
Supplemental Figure 7

a) Histogram showing the log10 distribution of transcript lengths in the comprehensive assembly. The red line indicates the minimum accepted length (200bp) we use to select the HQ transcripts. ***b)*** Histogram showing the distribution of masking content of the transcripts in the comprehensive assembly. The red line indicates the maximum masking threshold (90%) we use to select the HQ transcripts.



Supplemental Figure 8

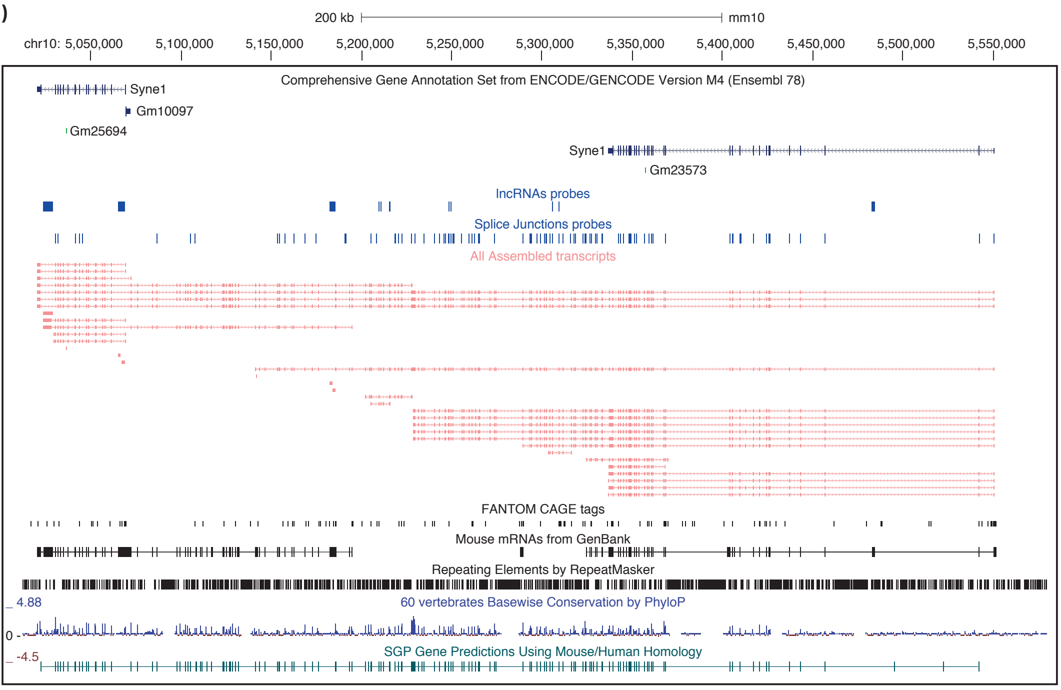
Strand specificity of the monoexonic transcripts not overlapping any other exon in any orientation. For each monoexonic transcript the ratio between the reads mapping in the sense orientation versus the reads mapping in the antisense orientation is shown. The figure shows the boxplots of the distributions of such scores across the 16 samples. A score of 1 indicates that the reads fully support the reported transcript orientation. The vertical black lines indicate the medians of each boxplot, while the dots indicate the outliers. This shows that while there are many outliers, the vast majority of reads are specific to the sense strand and not indicative of DNA contamination.



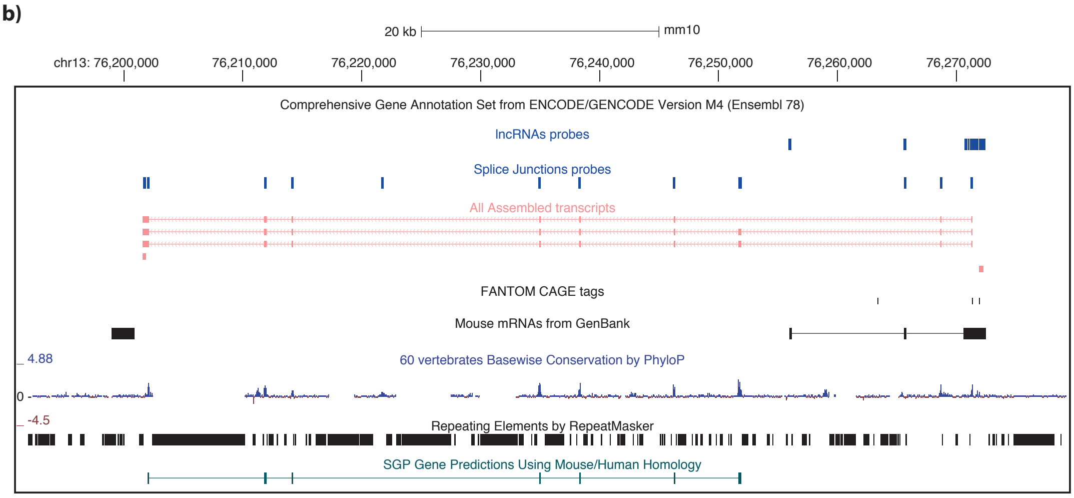
Supplemental Figure 9

Genomic tracks edited from UCSC genome browser (Kent et al. 2002) snapshots. Example of new transcript isoforms derived from GENCODE (M4) (***a***) or intergenic loci (***b***). From top to the bottom the GENCODE (M4) annotations, the CaptureSeq lncRNA and splice junction probes, all the assembled transcripts, FANTOM CAGE tags, GenBank mRNAs, *repeatMasker* masked elements, *phyloP* 60 vertebrates conservation, and human homology based gene predictions tracks from SGP2 (Guigo et al. 2003).

a)

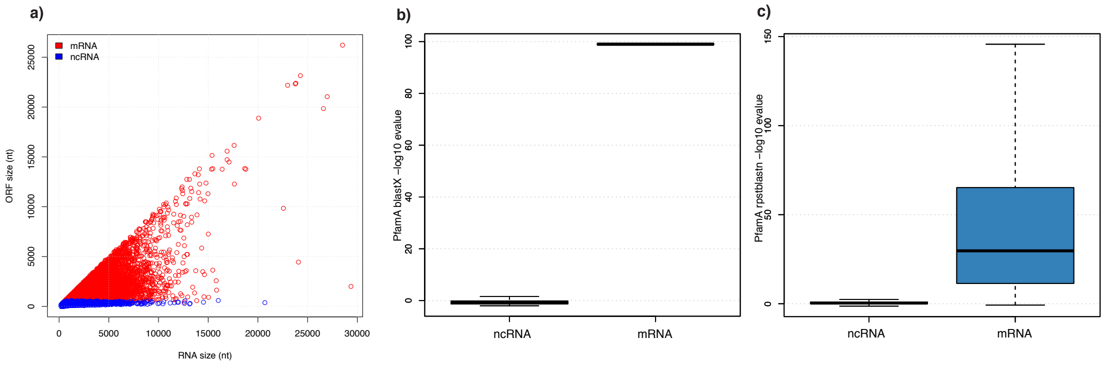


b)



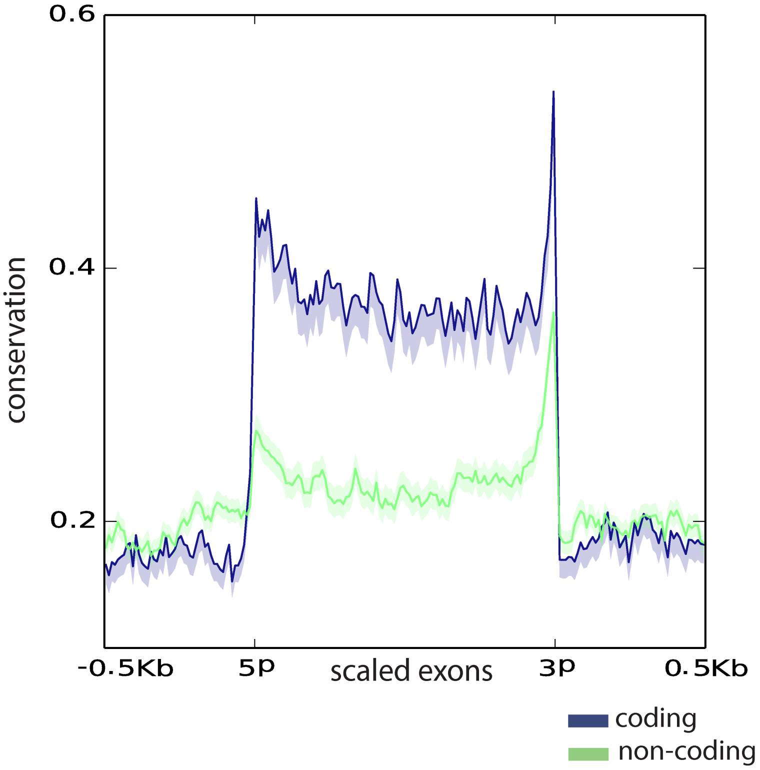
Supplemental Figure 10

Coding potential analyses of the HQ transcripts. ***a)*** Scatterplot comparing the transcript size (x-axis) and the *CPAT* predicted ORF size (y-axis). Each dot represents a transcript. Each transcript was labelled as mRNA (red) or ncRNA (blue) following *CPAT* coding probability predictions. Transcripts with a coding probability below 0.44 were considered as ncRNAs (blue dots) while all the transcripts above this threshold were considered mRNAs. ***b)*** –log10 E-Values boxplots as estimated by NCBI BlastX when scanning predicted mRNAs and ncRNAs against PFAM A (Finn et al. 2008). ***c)*** –log10 E-Values boxplots as estimated by NCBI *rpstBlastn* when scanning predicted mRNAs and ncRNAs against PFAM A (Marchler-Bauer et al. 2002).



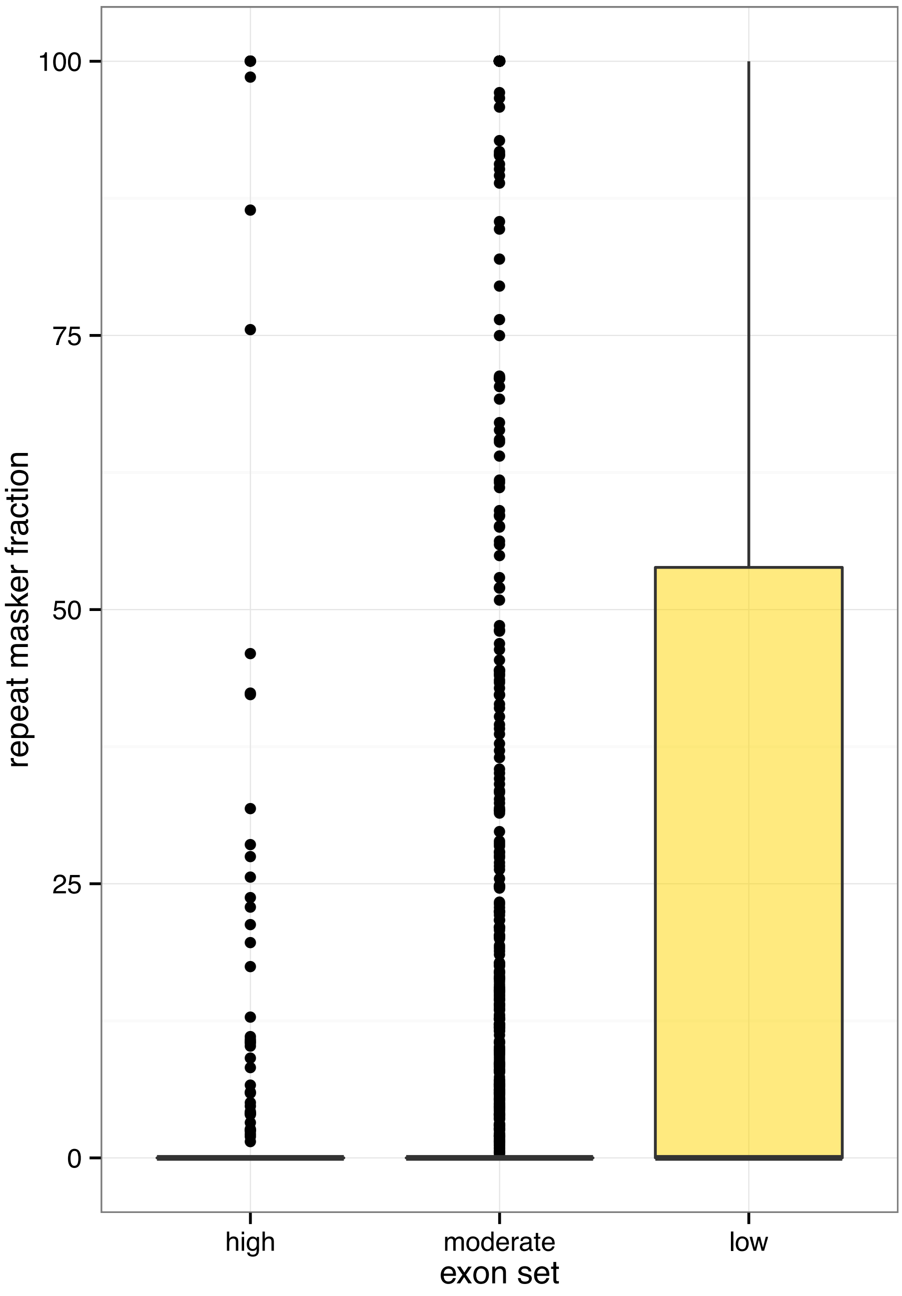
Supplemental Figure 11

PhyloP evolutionary conservation profile plot. Same layout as Figure 2. The ribbons indicate the standard error. The green curve represents the exons having less than 70% ORF coverage as estimated by *CPAT*, while the purple line indicates a coverage >20%.



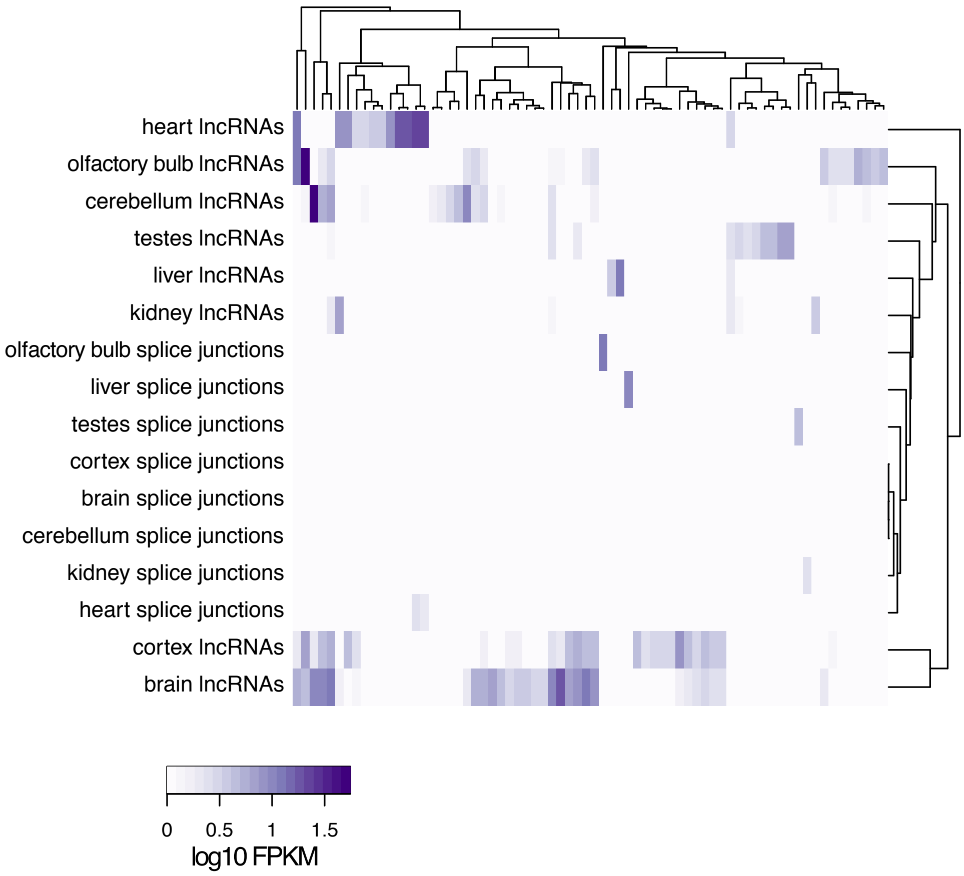
Supplemental Figure 12

Boxplots showing the content of repeats and low complexity regions in the three groups of exons shown in **Figure 2c**. The vertical axis indicates the fraction of soft masked nucleotides for each exon in the three groups.



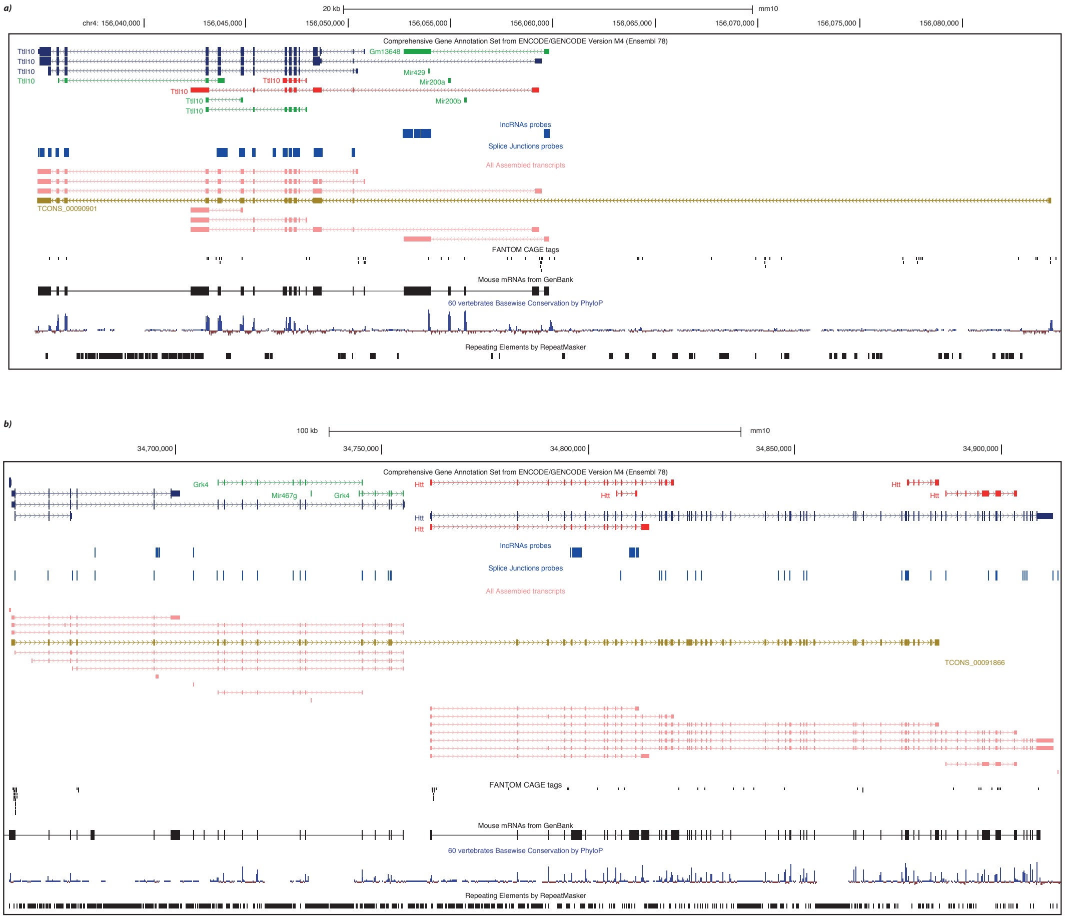
Supplemental Figure 13

Expression of novel intergenic HQ genes. Rows are the 16 samples we considered in the paper. Columns are the 70 intergenic non-coding spliced transcripts. The colour scale represents the log10 FPKMs. Darker shades of purple indicate higher expression.



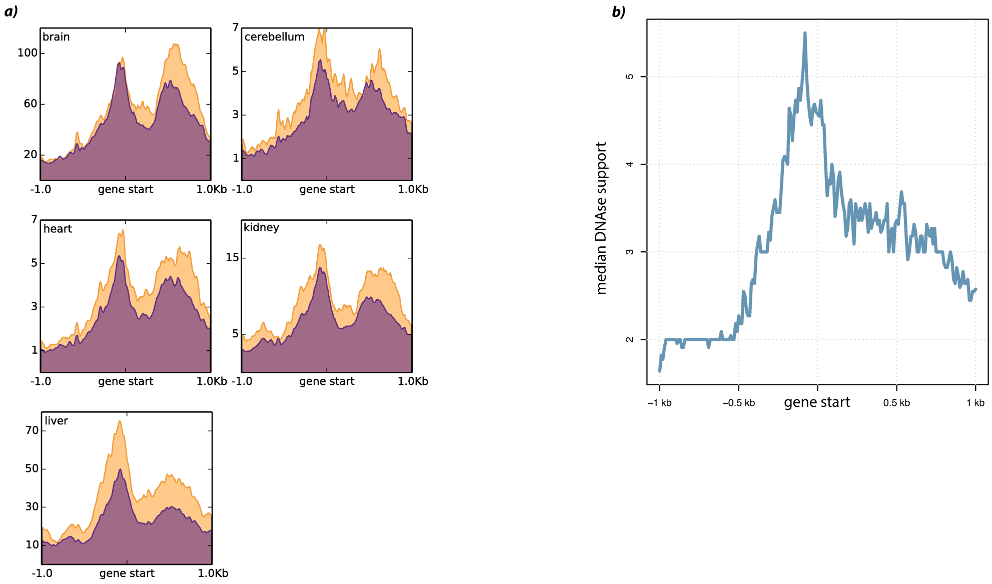
Supplemental Figure 14

Genomic tracks edited from UCSC genome browser snapshots. ***a)*** Alternative promoter of the CaptureSeq targeted gene *Ttll10*. The HQ transcript adding a 5' exon is highlighted in gold. ***b)*** HQ transcript joining together the G protein-coupled receptor kinase 4 (*Grk4* - *ENSMUSG00000052783*) and the huntingtin gene (*Htt* - *ENSMUSG00000029104*). Highlighted in gold the HQ transcript that connects *Grk4* to *Htt*. From top to the bottom the GENCODE (M4) annotations, the CaptureSeq lncRNA and splice junction probes, all the assembled transcripts, GenBank mRNAs, *phyloP* 60 vertebrates conservation and *repeatMasker* tracks (Smit et al.).



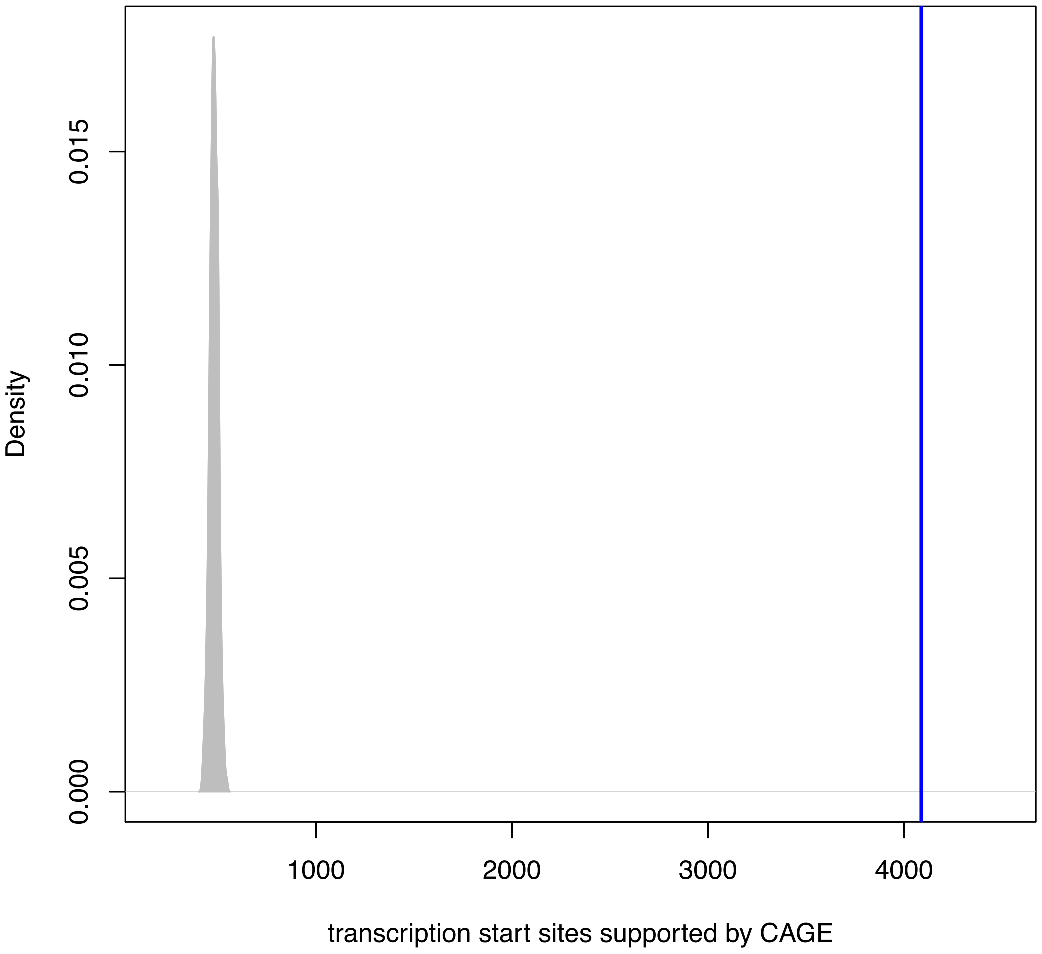
Supplemental Figure 15

DNAseHS support of the HQ gene start sites. Same layout as figure in Figure 4. ***a)*** The purple curve represents the mean DNAseHS support across all the HQ genes. The orange curve represents the mean DNAseHS support across just the HQ genes expressed in that specific tissue. ***b)*** The curve shows the median DNAseHS support of the HQ genes in the tissue where they are best expressed.



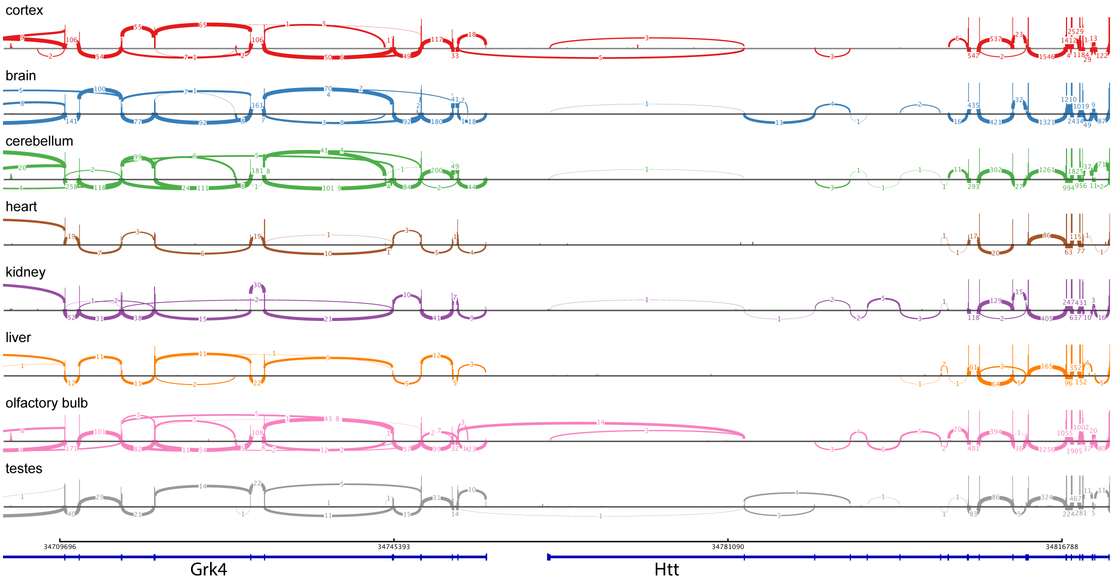
Supplemental Figure 16

CAGE support of the new transcription start sites. On the x-axis the number of transcription start sites that overlap at least a CAGE peak. On the vertical axis it is shown the density function. The grey distribution on the left represents the one thousand random projections of the transcription start sites to the genome. The blue line on the right shows the observed number of transcription sites supported by CAGE peaks.



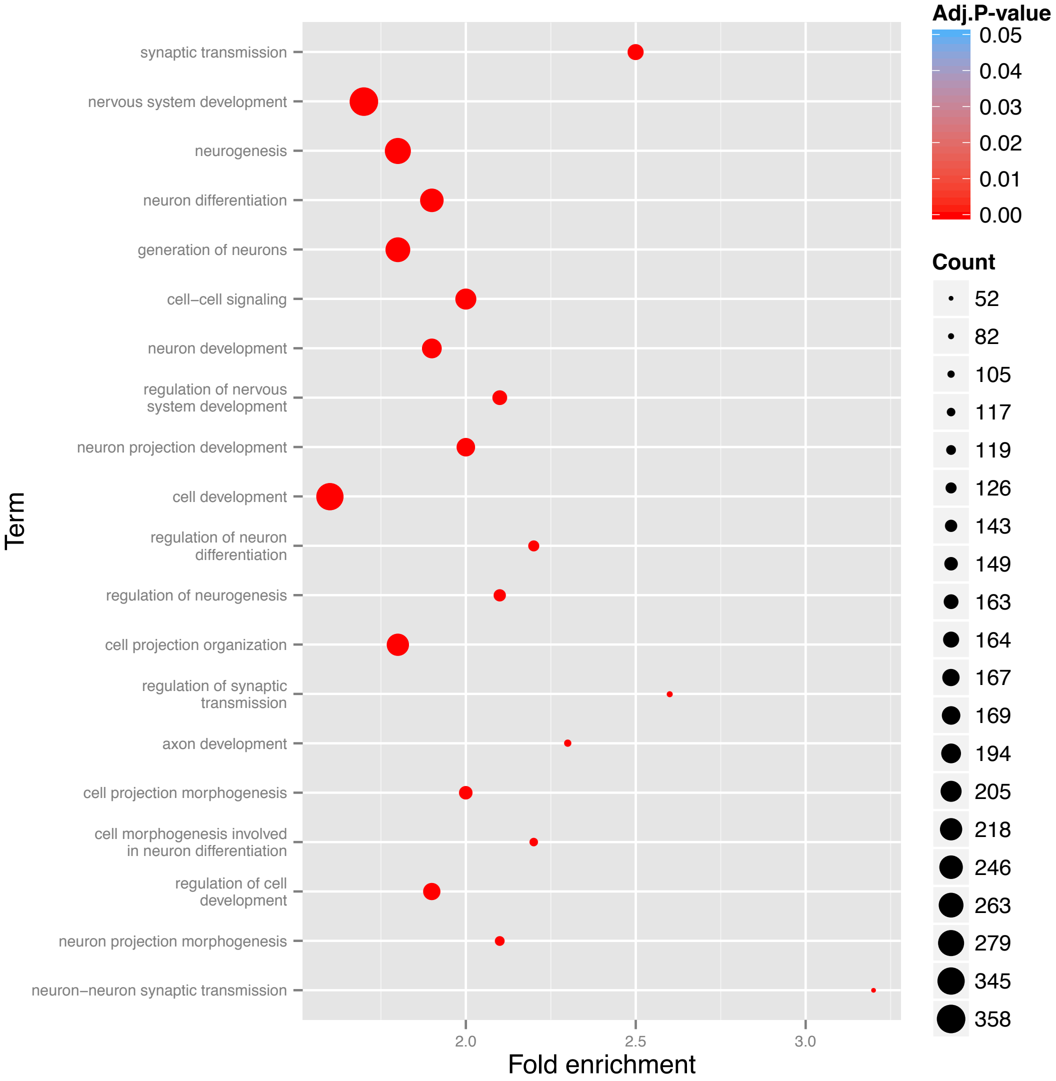
Supplemental Figure 17

Sashimi plot edited from IGV (Thorvaldsdóttir et al. 2013; Robinson et al. 2011). The top tracks represent the junction coverage in the CaptureSeq splice junction experiments. The bottom track shows the *Grk4* and *Htt* genes.



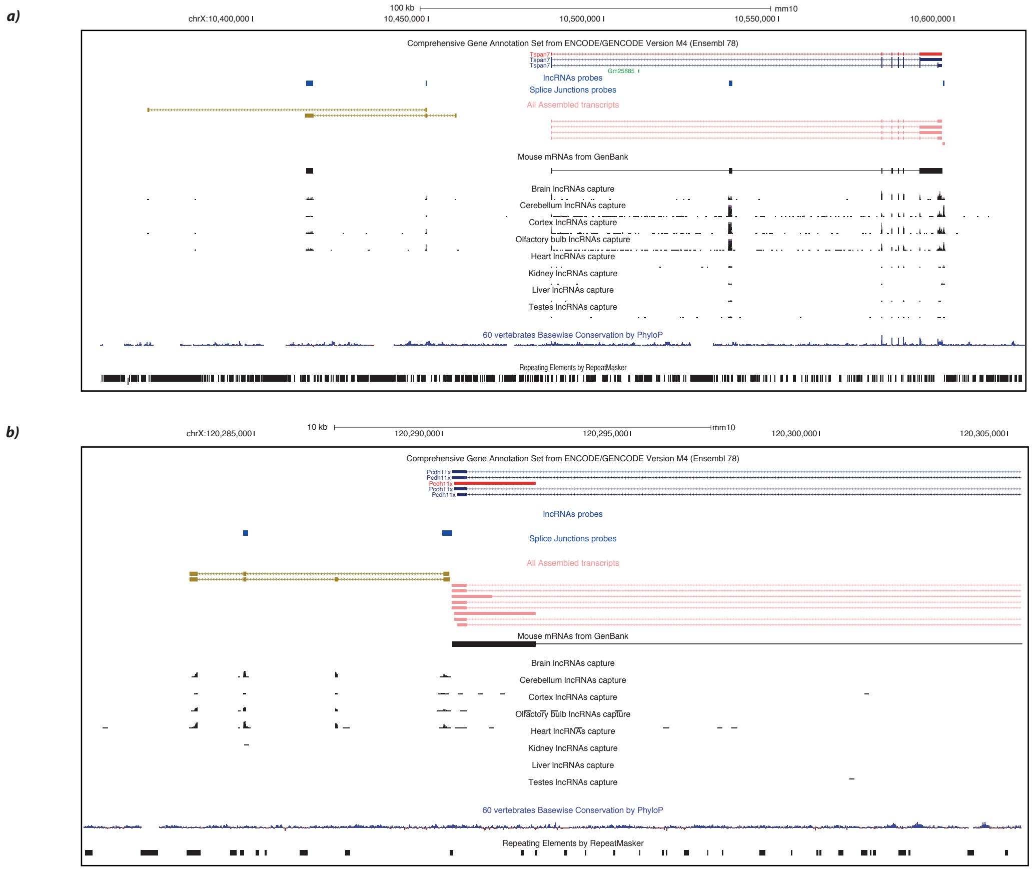
Supplemental Figure 18

GO biological process term enrichment of the genes overlapping or close to HG genes enriched or depleted in brain tissues. The x-axis represents the fold enrichment of the observed versus expected number of genes in each ontology. The vertical axis shows the 20 best terms sorted by increasing P-value. The size of each dot indicates the actual number of genes assigned to each ontology. The colour reflects the Benjamini-Hochberg multiple testing adjusted P-value.



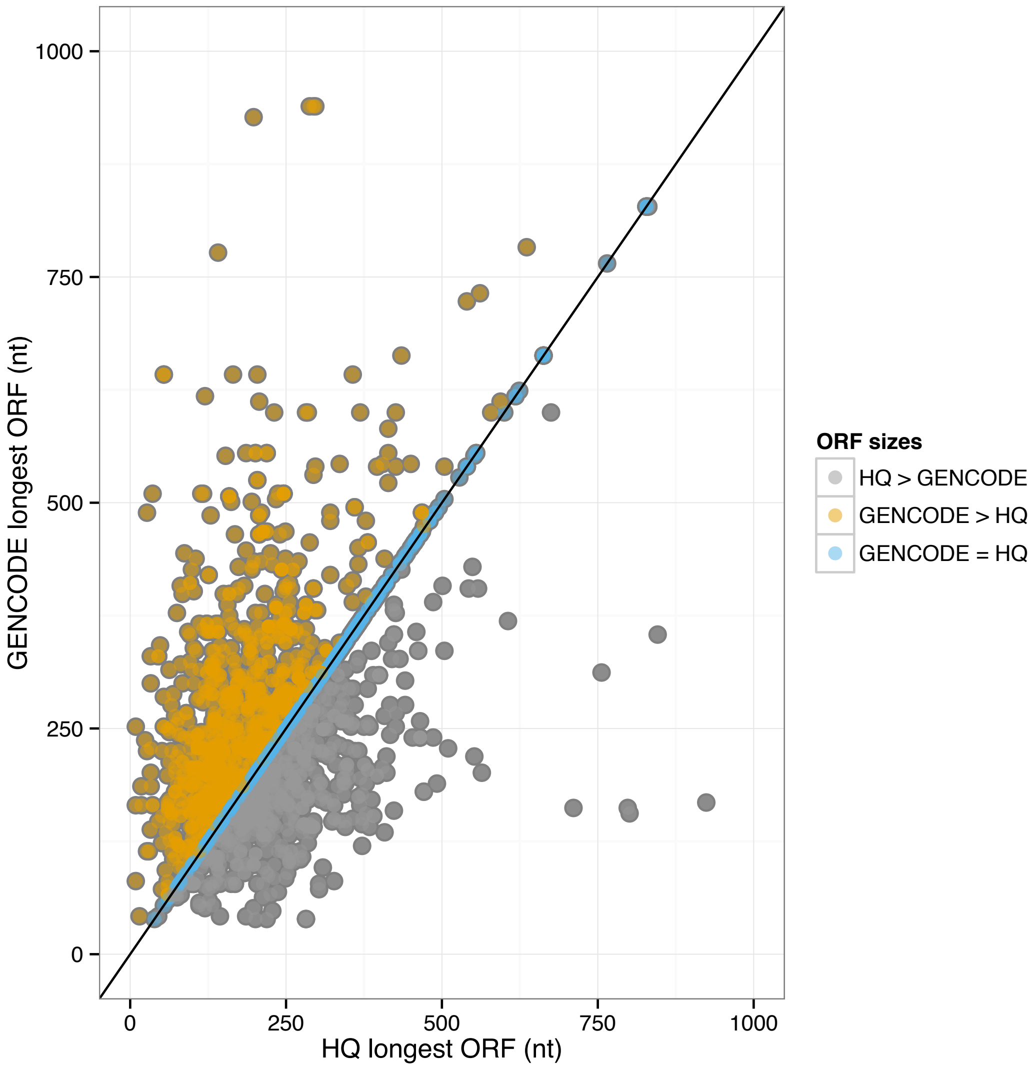
Supplemental Figure 19

Genomic tracks edited from edited from UCSC genome browser snapshots of brain specific HQ transcripts (highlighted in gold). TCONS_00132850 and TCONS_00132849 (***a***) and TCONS_00134637 TCONS_00134636 (***b***). From top to the bottom the GENCODE (M4) annotations, the CaptureSeq lncRNA and splice junction probes, all the assembled transcripts, GenBank mRNAs, the read densities in the CaptureSeq lncRNA design of the considered tissues, *phyloP* 60 vertebrates conservation and *repeatMasker* tracks.



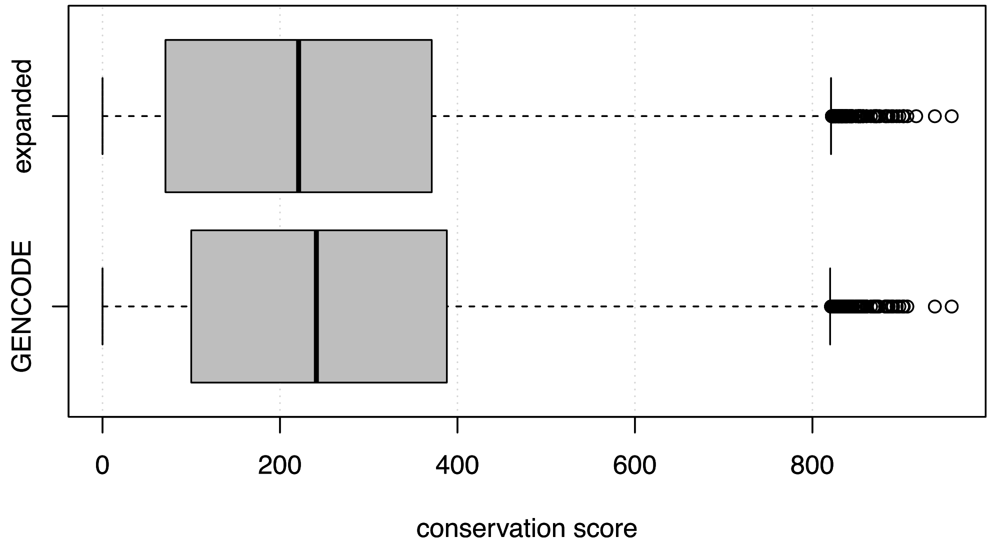
Supplemental Figure 20

Scatterplot showing the longest *CPAT* predicted ORF size in corresponding HQ and GENCODE lncRNA. The black line indicates the bisector. Grey dots indicate that the ORF is longer in HQ. Orange dots indicate that the ORF is longer in GENCODE. Blue dots indicate that the ORF does not change. Outlier ORFs above 1 kb are discarded.



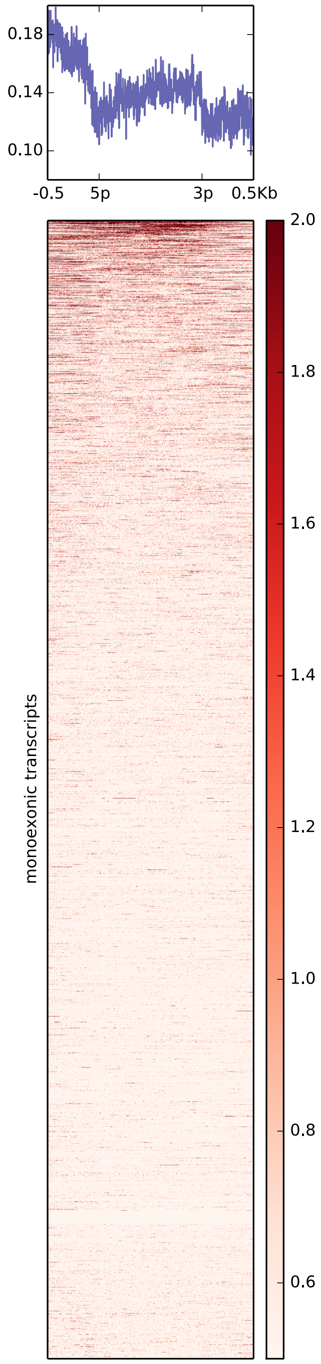
Supplemental Figure 21

Boxplot showing the phastCons conservation score (see methods) for GENCODE and expanded assemblies.



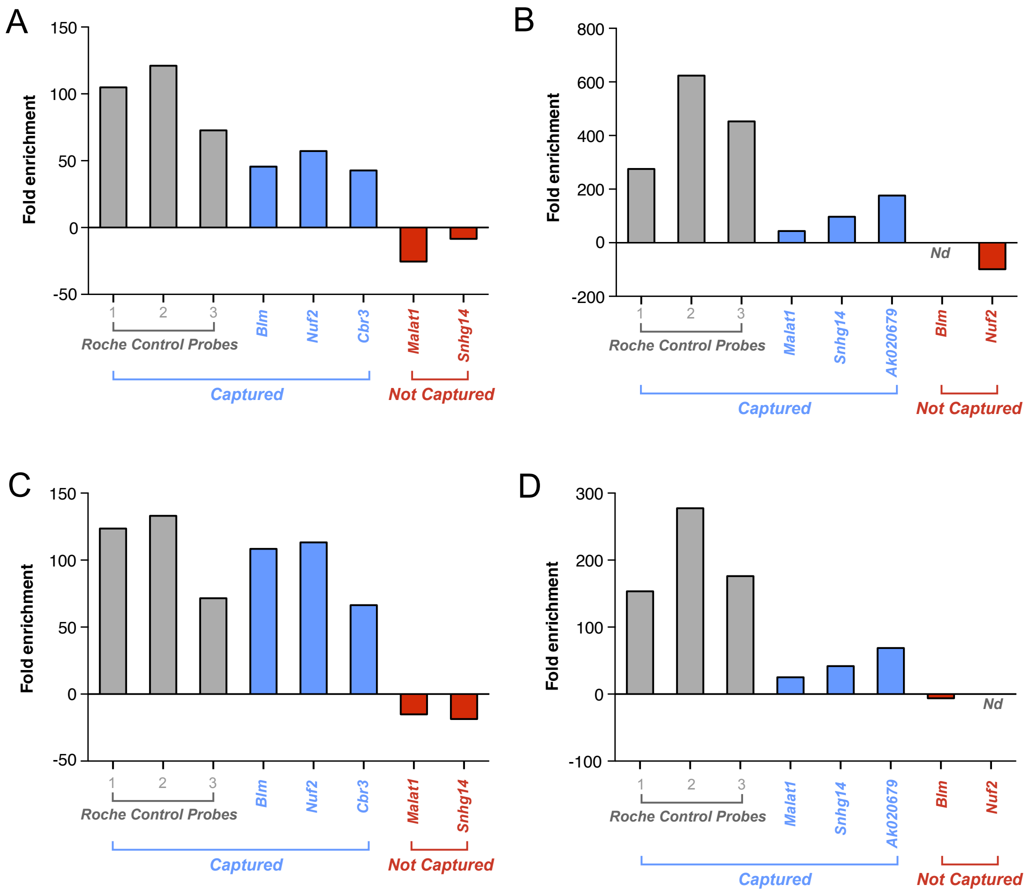
Supplemental Figure 22

PhyloP evolutionary conservation of the monoexonic transcripts not overlapping any annotated exon and not included in known introns with the same orientation. Same layout as Figure 2c.



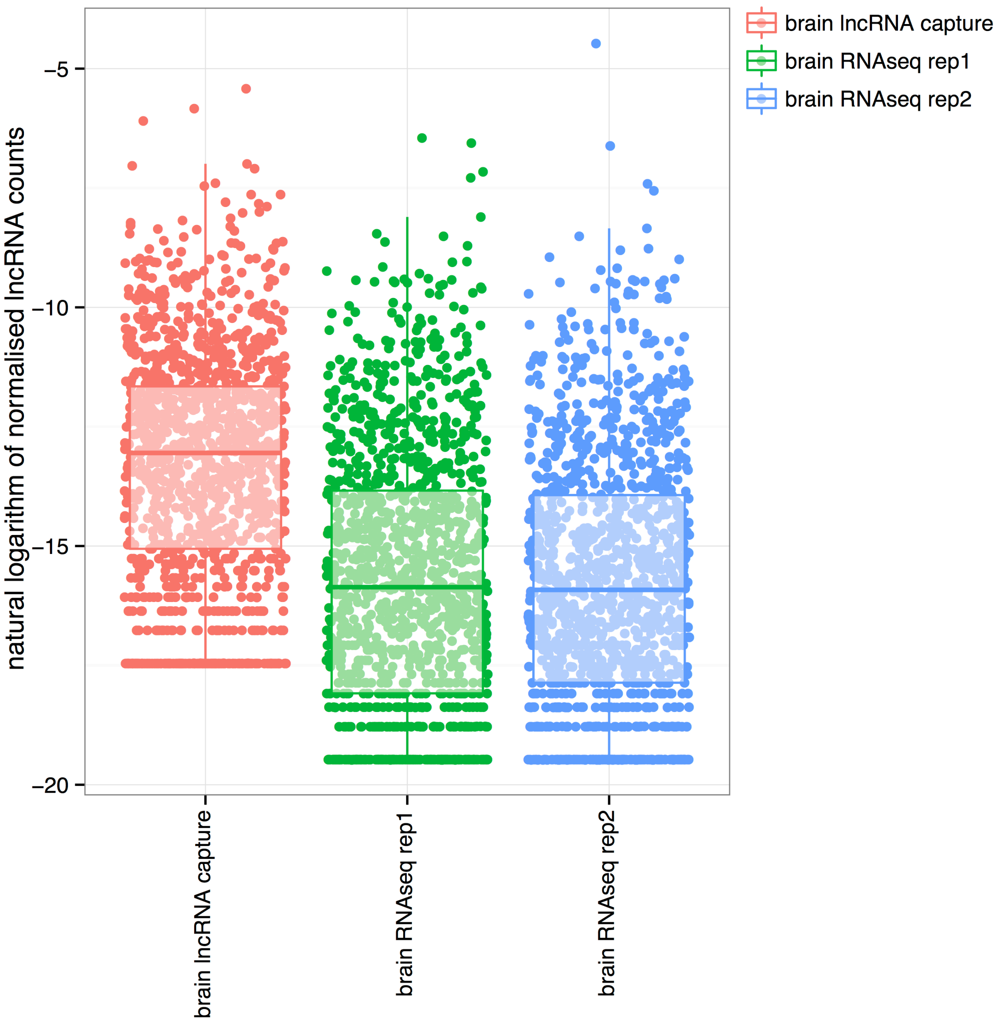
Supplemental Figure 23

Quantitative RT-PCR shows fold-enrichment following CaptureSeq for targeted Roche controls (grey; 1-3), genes targeted for capture (blue) and depletion of non-targeted genes (red). Nd: no expression detected post-capture. A) Mouse splice junction capture, non-brain organs. B) Mouse LncRNA capture, non-brain organs. C) Mouse splice junction capture, brain regions. D) Mouse LncRNA capture, brain regions.



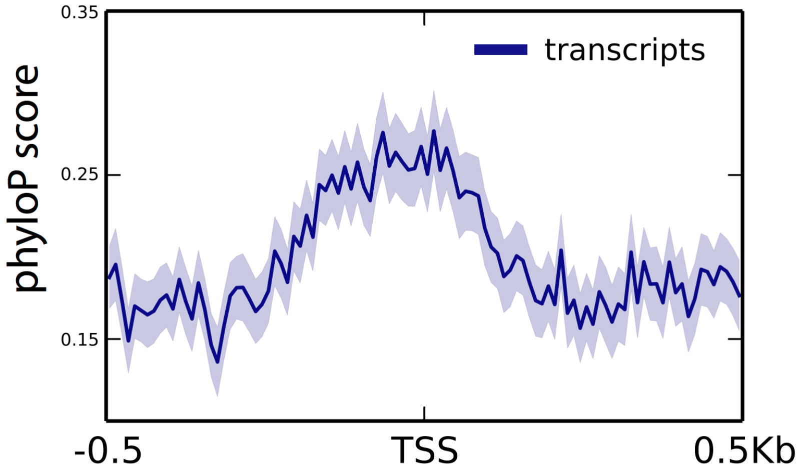
Supplementary figure 24

Comparison of the normalized read count of GENCODE M4 lncRNAs in the brain CaptureSeq and two brain mouseENCODE RNAseq runs. The mouseENCODE datasets were mapped to the GRCm38 assembly as described in the methods. The normalized read count of the GENCODE M4 lncRNAs were estimated as described in the methods. Each dot in the plot indicates a lncRNA gene. The vertical axis reports the normalized read counts. The colour scheme demonstrates in red, green and blue respectively the CaptureSeq, RNAseq replicate1 and RNAseq replicate 2 experiments.



Supplementary figure 25

Promoter conservation profile. The set of HQ transcript corresponding to known GENCODE genes with upstream start (see methods) was considered in this analysis. The x-axis reports the positions +/- 500 bp around the TSSs. The vertical axis reports the mean *phyloP* 60way vertebrate conservation scores. The ribbon represents the standard error.



SUPPLEMENTAL MATERIAL

Supplemetary Data 1: The comprehensive set of putative and predicted lncRNA that was targeted. The coordinates refer to MGSCv37 genome assembly and are reported in bed12 format.

Supplemetary Data 2: Annotation table of targeted lncRNAs referring to Supplemetary Data 1.

Supplemetary Data 3: Comprehensive unfiltered mouse transcriptome assembly as returned cuffmerge. The coordinates refer to GRCm38 genome assembly and are reported in gtf format.

Supplemetary Data 4: HQ gene set generated after applying all the filters described in the text. The coordinates refer to GRCm38 genome assembly and are reported in gtf format.

Supplementary Data 5: Table recapitulating relevant genomic and expression properties of the HQ transcripts. The “gene_id” and the “transcript_id” fields are the gene and the transcripts identifiers as returned by cufflinks. The “length” field correspond the transcript nucleotide length. “CG%” is the CG content of the transcripts. “masking%” is the fraction of the transcripts that is covered either by repeats or low complexity regions. “exons” is the number of exons in the transcript. “codingProbabilityCPAT” represents the probability as estimated by CPAT that the transcript is coding. The fields labelled with “TRFPKM” represent the FPKM expression of the transcripts across the 16 samples. “ORFsize” indicates the ORF size as estimated by CPAT. “PFAMblastX” indicates the lowest E-Value as reported by BlastX when scanning the transcript against PFAM A. “PFAMrpstblastn” indicates the lowest E-Value as reported by rpstBlastN when scanning the transcript against PFAM A. “phastCons” indicates scores reflecting the evolutionary conservation (see methods).

Supplemetary Data 6: HQ coding transcript. The coordinates refer to GRCm38 genome assembly and are reported in gtf format.

Supplemetary Data 7: HQ non-coding transcript. The coordinates refer to GRCm38 genome assembly and are reported in gtf format.

Supplemetary Data 8: HQ genes bridging two or more GENCODE genes. The coordinates refer to GRCm38 genome assembly and are reported in gtf format.

Supplemetary Data 9: HQ coding transcripts corresponding to previous non-coding transcripts. The coordinates refer to GRCm38 genome assembly and are reported in gtf format.

Supplemetary Data 10: Nimblgen lncRNA capture probe coordinates. The coordinates refer to GRCm38 genome assembly and are reported in bed format.

Supplemetary Data 11: Nimblgen splice junctions capture probe coordinates. The coordinates refer to GRCm38 genome assembly and are reported in bed format.

Supplemetary Data 12: PCR primer sequences used to confirm Capture enrichment.