Table of Contents **Page**

Supplemental Figure 1. A detailed description of our internal priming filtering model can be found in the Methods section of the main text. Briefly, we used a modified PolyA-Seq run where the reverse transcription primer lacked the 3' VN that is typically used to anchor the priming reaction at the beginning of a polyA tail. This resulted in random placement of PolyA-Seq reads and enabled direct observation of internal polyA tracks versus genuine polyA tails. Reads that ended with a stretch of As that were also found in the genome are likely internal priming events. *They may also include some genuine polyadenylation events and we sought to minimize these by removing known polyA sites. In any case, false-negatives can only result in an underestimate of our false-discover rate (See Methods and Supp. Fig 2). Genuine polyA sites were detected from reads that ended in stretches of As that were not found in the genome. *pA* and *I* are the genuine polyA site and internal-priming base frequencies respectively. x_i (A_i,C_i,G_i,T_i) denotes the base identity at position *i* within the downstream site being assessed, where A_i, C_i, G_i, T_i, are the frequencies of As, Cs, Gs, and Ts, at position *i* in the matrix. Thus any 10-base sequence can be used to compute a polyA score. The internal priming model takes as input 10 bases and is simply a log-odds ratio of the likelihood that it was sampled from the internal priming distribution of base frequencies vs. frequencies of genuine polyA sites. Unless otherwise noted, all nucleic acid sequences are shown in the 5'->3' orientation.

Supplemental Figure 2. Internal priming model performance and estimation of false-discovery. (*A*) ROC analysis of our internal priming site filtering model on independently generated UHR data. Sensitivity (proportion of real polyA sites) as a function of the false-positive rate (proportion of internal priming sites) is plotted for a range of polyA score thresholds. (*B*) Sensitivity and (*C*) specificity, as a function of the polyA score. A score of 3.0 was used as the threshold for calling polyA sites in this study (sensitivity = 85.6%, FP rate = 2.5%). (*D*) Precision (TP/(TP+FP)) as a function of the polyA score. This roughly translates to 1-FDR (false-discover rate) when the proportion of positives to negatives in the test set is similar to the proportion found in the entire genome (which is true; see Methods).

Supplemental Figure 3. (*A*) Discriminating internal priming sites from polyA tails using RT-PCR. A polyT oligo priming in random locations on a polyA tail produces a smear following RT-PCR, whereas an internal priming event (mostly resulting from a short stretch of As) yields a tight band. (*B*) Molecular protocol.

Supplemental Figure 4. (*A*) Validation of 4 known polyA sites, 6 sites shown in Figure 3A,C,F (main text), and 10 randomly selected internal priming sites by RT-PCR (using approach shown in Supp. Fig. 4). All reactions were run in parallel from mRNA through gel exposure. (*B*) Primer sequences and additional information for reactions shown in *(A)*. T(20) = oligonucleotide consisting of 20 Ts.

Supplemental Figure 5. (*A*) More detailed UCSC Genome Browser shots for polyA sites shown in Figure 3A,C,F (main text). Wiggle tracks depict all aligned reads (including unfiltered).

Supplemental Figure 6. (*A*) Base composition surrounding polyA sites before and after filtering. Genomic base frequencies are shown for ends of aligned reads (left of vertical line) and downstream sequences. Data for ends of UCSC alignments of RefSeq RNAs is included for reference; one-base shift is presumably due to the incomplete removal of polyA tails. (*B*) Impact of filtering method on polyA sites. Brain PolyA-Seq sites are shown in human *ELAVL1* for all reads, simple filtering based on downstream adenines as previously published (Lee et al. 2007), and the filtering procedure described here (see Methods), in descending order. Liver sites are included to show that the downstream site is specific to brain. For simplicity, only sense data is presented. Y-axes are on a log scale. (*C*) Assessment of reported polyA Sites. PolyA-Seq results and reported polyA sites (from PolyA-DB (Lee et al. 2007)) are shown within the 3'UTR of *ELAVL1*. PolyA-Seq filtering suggests that the penultimate reported site is a false-positive.

Supplemental Figure 7. Number of polyA-site clusters as a function of minimum distance between peaks. Distances greater than 30 bp do not significantly reduce the number of clusters, indicating that this is a good minimal approximation of wiggle in transcriptional cleavage surrounding a unique site.

Supplemental Figure 8. PolyA-Seq detects polyadenylation of ubiquitously expressed primary microRNA transcripts. The human *let-7a1/let-7f-1/let-7d* cluster is shown. Arrows indicate the direction of transcription, on the forward genomic strand in this case. Polyadenylation of the primary transcript is detected in all tissues assayed in all species (data not shown, but see Figure 3 for rhesus and mouse). Two adjacent polyA sites are observed in primates, but only one in other species, due to the advent of a second canonical polyA signal (AAUAAA) in this lineage (data not shown). In addition, antisense polyadenylation is observed downstream of *let-7f-1* in human brain, supported by spliced Genbank RNAs and ESTs, as well as rhesus ileum and testis (data not shown). Fwd, forward genomic strand; rev, reverse strand. PolyA-Seq Y-axis units are reads per million.

Supplemental Figure 9. PolyA-Seq detects polyadenylation of tissue-specific primary microRNA transcripts. A) Human *mir-122* is detected only in liver and B) *mir-124-1* principally in brain, in accordance with their known tissue specificities. Results are shown on both linear and log scales to highlight minor polyA events. Y-axis scales are in reads per millions. Fwd, forward genomic strand; rev, reverse strand. Arrows indicate the direction of transcription. C) Polyadenylation within the body of *mir-124-1*, immediately upstream of the position where the mature microRNA is excised, consistent with reported transient polyadenylation of degradation intermediate (Slomovic et al. 2010). D) The position of the mature *mir-124-1* transcript (pink) within the precursor, obtained from miRBase, shows that PolyA-Seq exactly identifies polyadenylation of the upstream product remaining after excision.

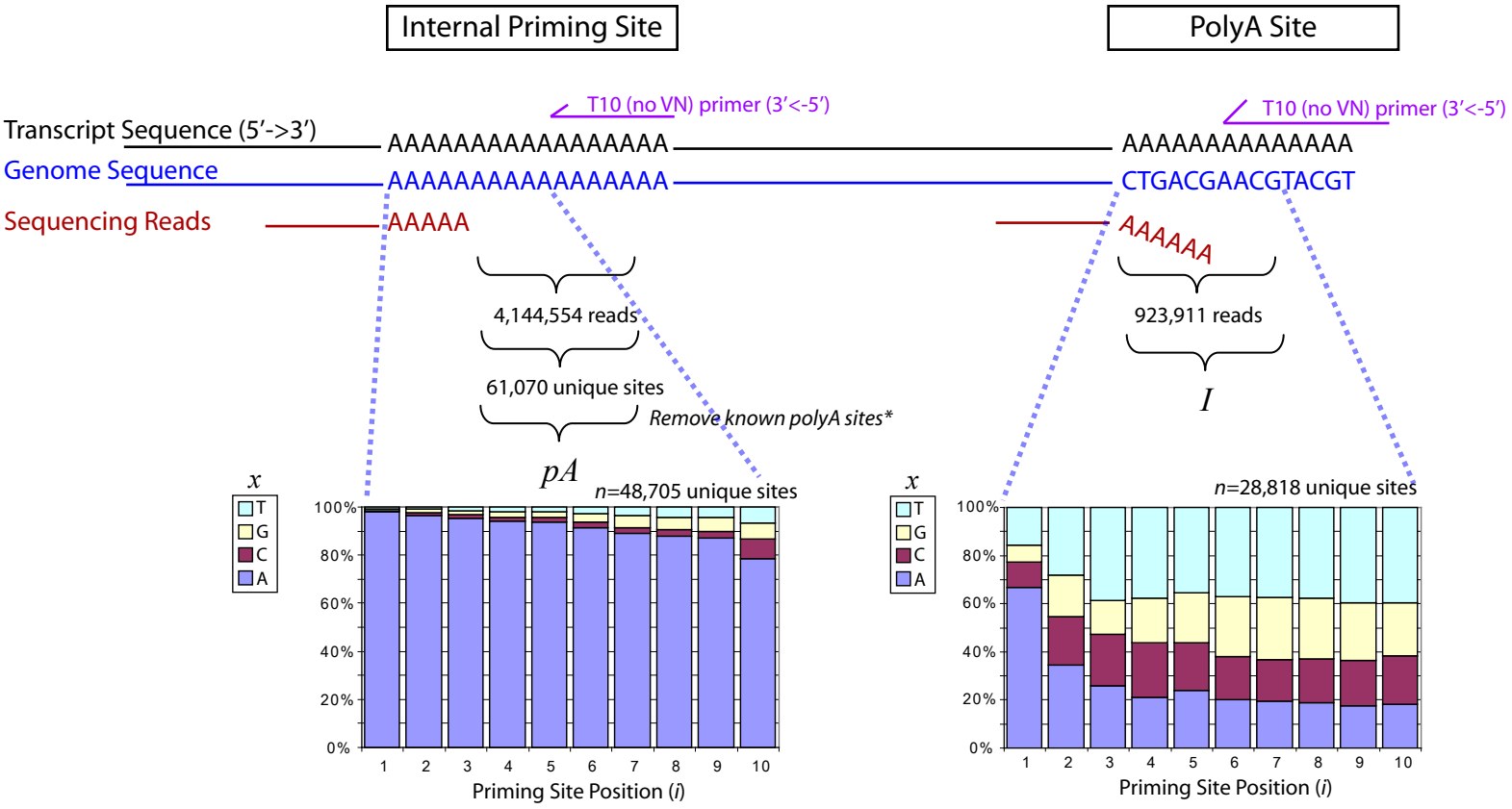
Supplemental Figure 10. PolyA-Seq detects polyadenylation of non-coding RNAs in a strand-specific manner. A) Human accelerated region transcript *HAR1A* is detected in all tissues, and *HAR1B* in brain and muscle. PolyA signals (AATAAA) are also shown, with the sign indicating the genomic strand (positive, forward strand; negative, reverse strand).

The conservation track reveals conservation of only the antisense segment of the transcripts. B) The X-inactive specific transcript (*Xist*) is detected at high levels in brain and muscle, and at a low level in testis. C) The HOX antisense intergenic RNA (*HOTAIR*) is detected in testis. Intriguingly, PolyA-Seq suggests the presence of a transcript antisense to *HOTAIR* in testis. For each locus, only tissues with data are shown. Y-axes for PolyA-Seq sites are in reads per million, and vary among tissues.

Supplemental Figure 11. PolyA-Seq detects polyadenylation of antisense transcripts. A) Human *HOXA11* is detected in kidney and testis, while its antisense transcript, *HOXA11AS*, is detected in kidney. B) *DLX1* and its antisense transcript, *DLX1AS*, are detected in human brain. *DLX1AS* was only annotated via its mouse (mus) ortholog. In both cases, the coding gene is on the reverse genomic strand and the antisense gene on the genomic strand, as indicated by the arrows. PolyA-Seq data (in reads per million) is shown separately for the forward (fwd) and reverse (rev) genomic strands; tissues without data for the given genes are not shown.

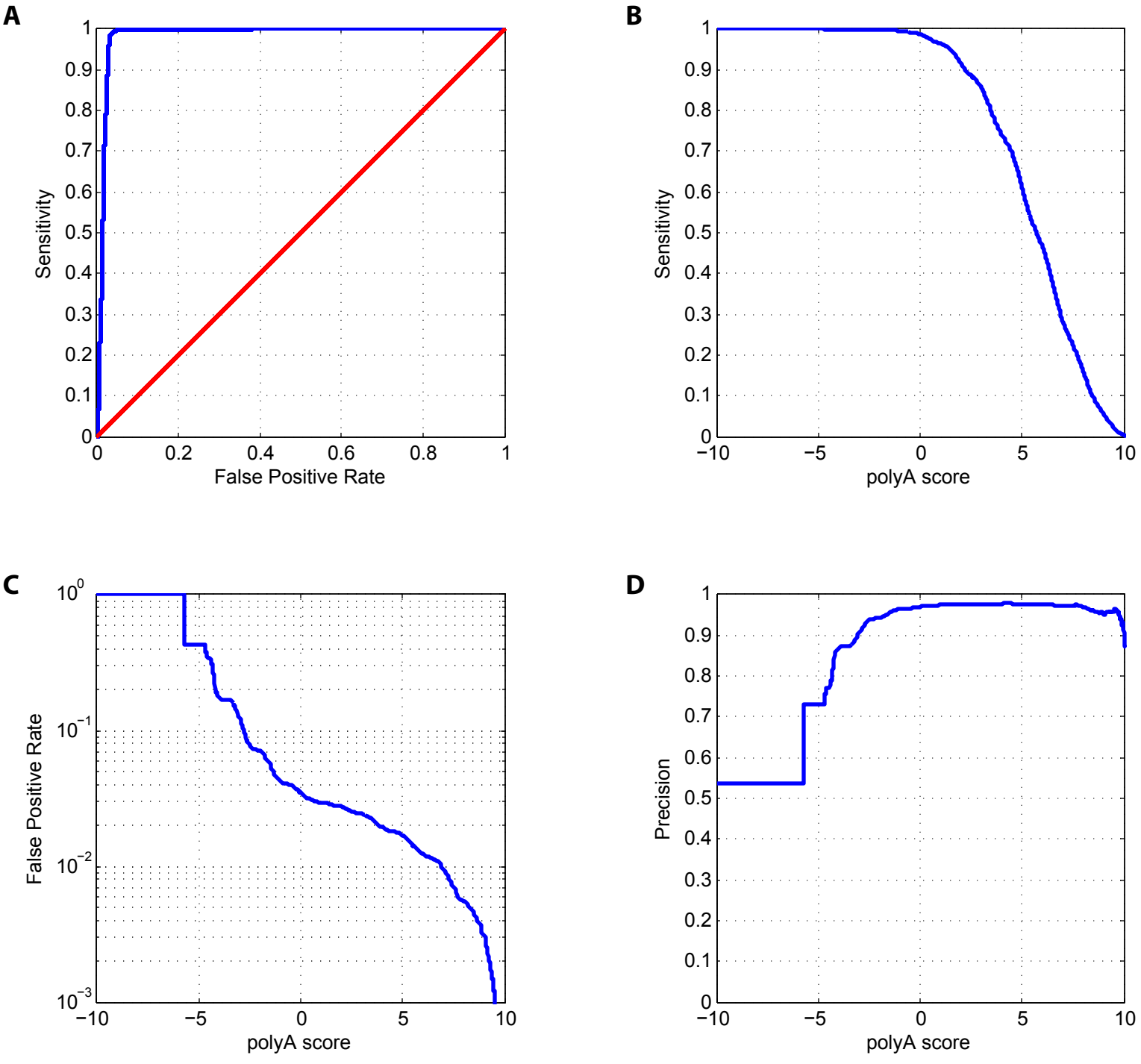
Supplemental Figure 12. PolyA-Seq identifies polyA sites in 3'UTRs of tail-to-tail genes. The human genes encoding ring finger protein 214 (*RNF214*, forward genomic strand) and β-site AβPP cleaving enzyme 1 (*BACE1*, reverse strand) are convergent. PolyA-Seq results support the annotated 3'UTRs but reveal the positions of the dominant polyA sites in each gene. PolyA-Seq sites on the forward genomic strand ("fwd") are shown above the transcripts, and those on the reverse strand ("rev") below the transcripts. Y-axis units are in reads per million; note that Y-axis scales vary.

Supplemental Figure 13. Splicing-dependent alternative polyadenylation in human *SLC11A2*. Two polyA sites are observed in all tissues. Refer to caption of Supplemental Fig. 9 for details.



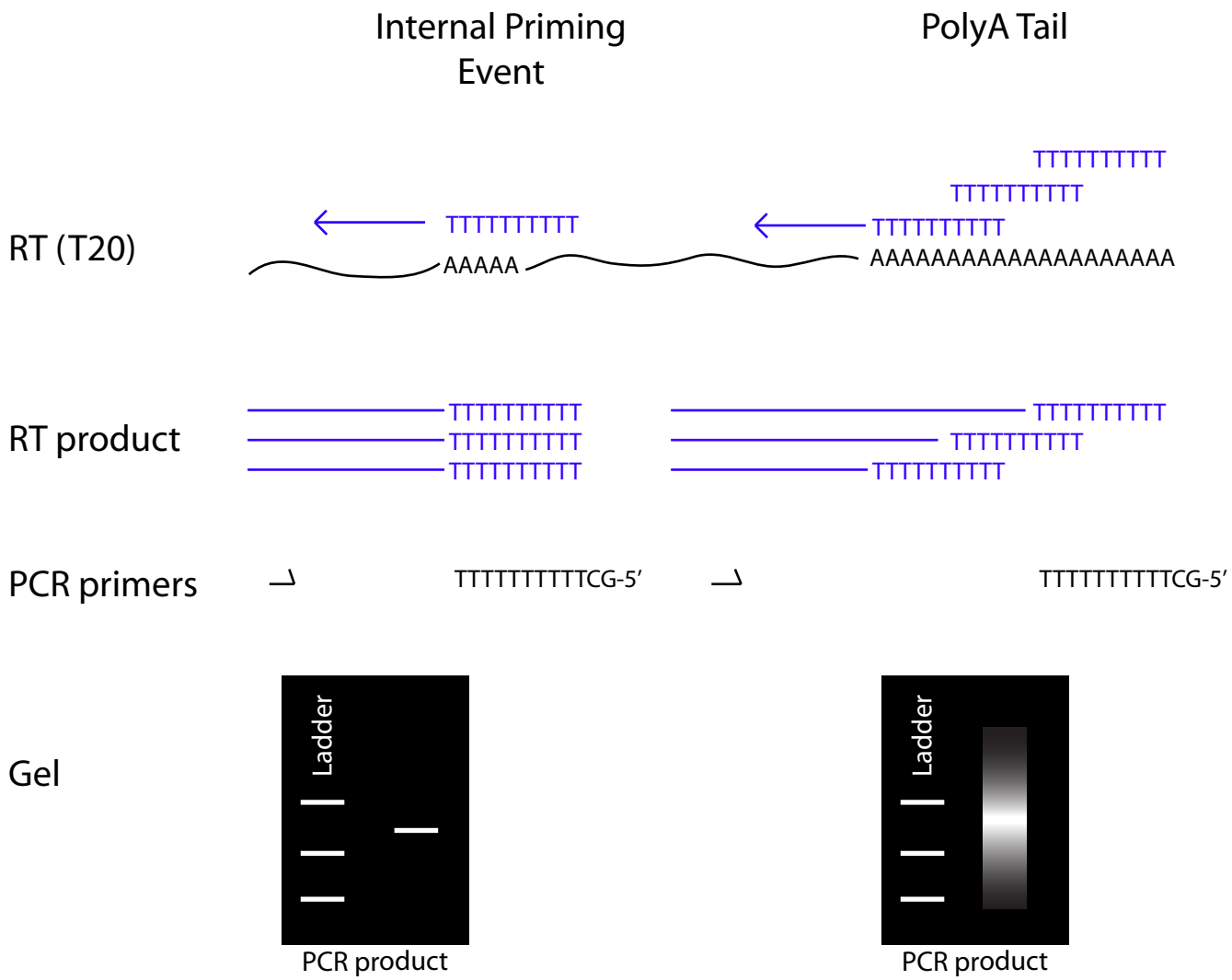
$$\text{polyA Score} = \log_{10}\left(\frac{\sum_{i=1}^{10} pA(x_i)}{\sum_{i=1}^{10} I(x_i)}\right)$$

Supplementary Figure 1



Supplementary Figure 2

A



B

Protocol

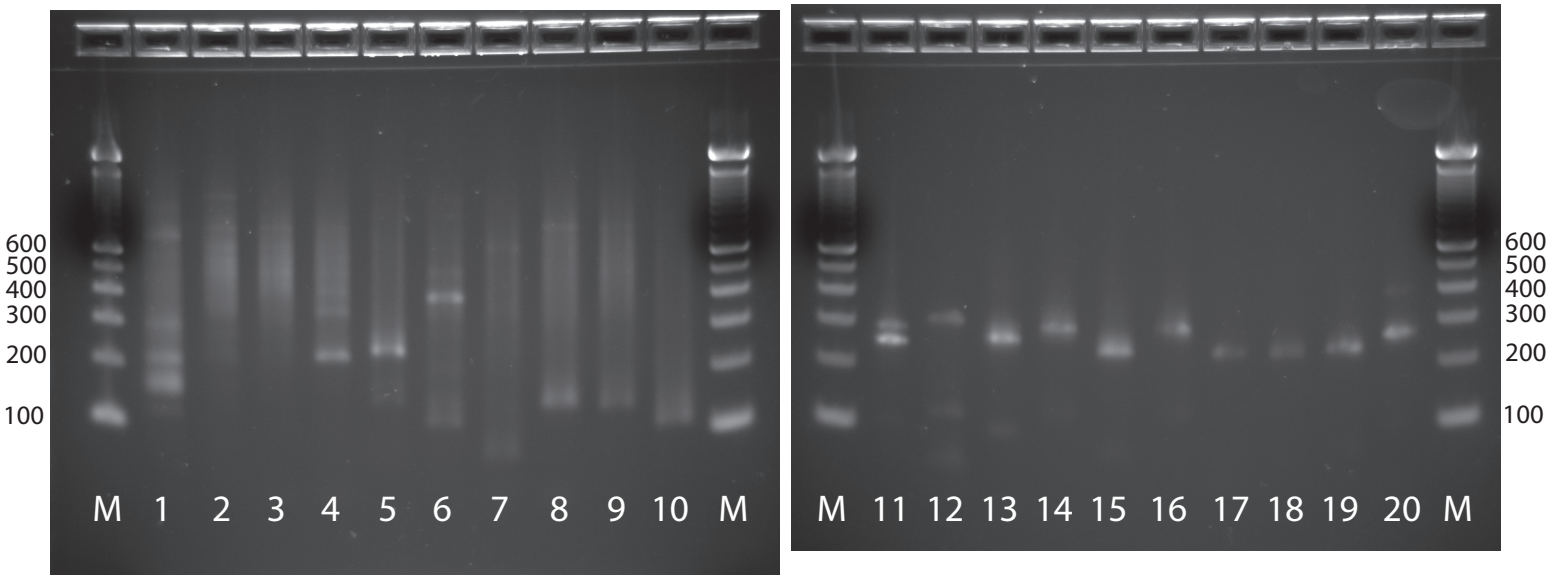
Reverse Transcription
200 ng mRNA input; superscript III protocol in 20 ul reaction with 1 ul of 1 uM T20, 1 hour at 42, 5 min @ 85, no clean-up.

PCR (Invitrogen HF Taq - run according to manufacturer's recommendations)
1 uL RT product
buffer, dNTP, MgSO4, Taq
40 nM (final concentration) forward primer
40 nM (final concentration) GCT20
water to 25 uL final volume

35 cycles of:
94 for 25 sec
50 for 25 sec
72 for 30 sec

Supplemental Figure 3

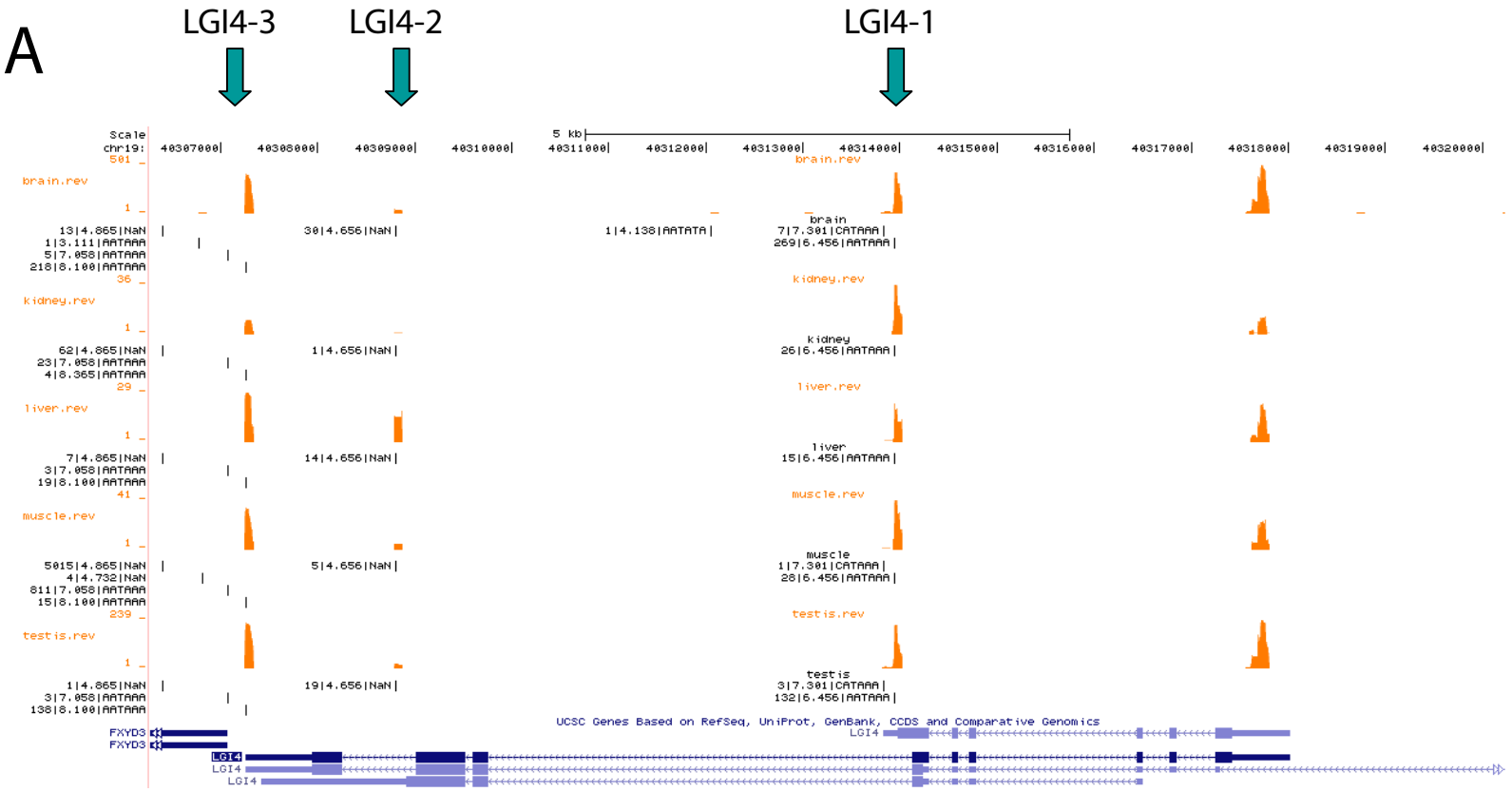
A

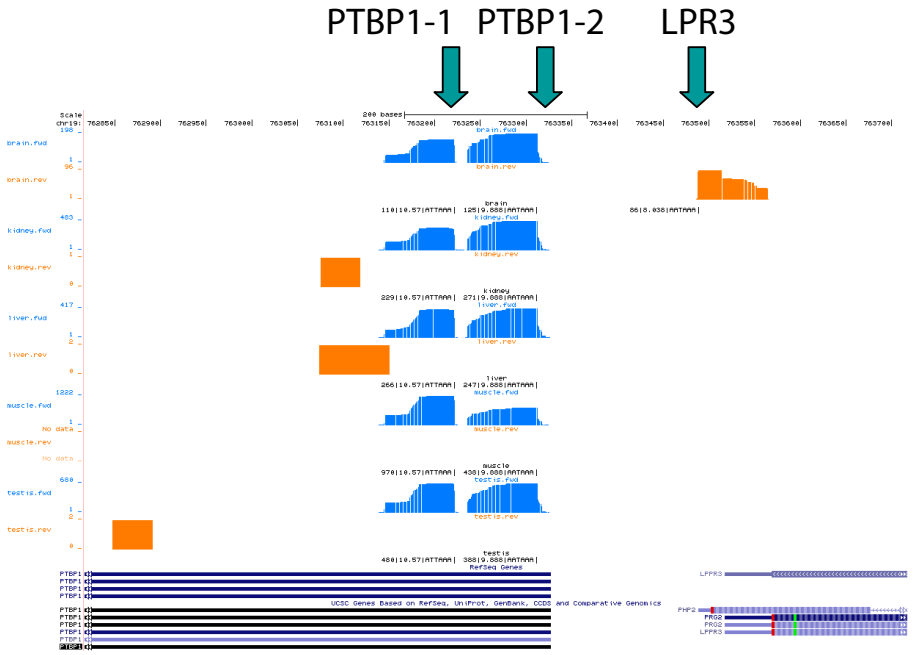


B

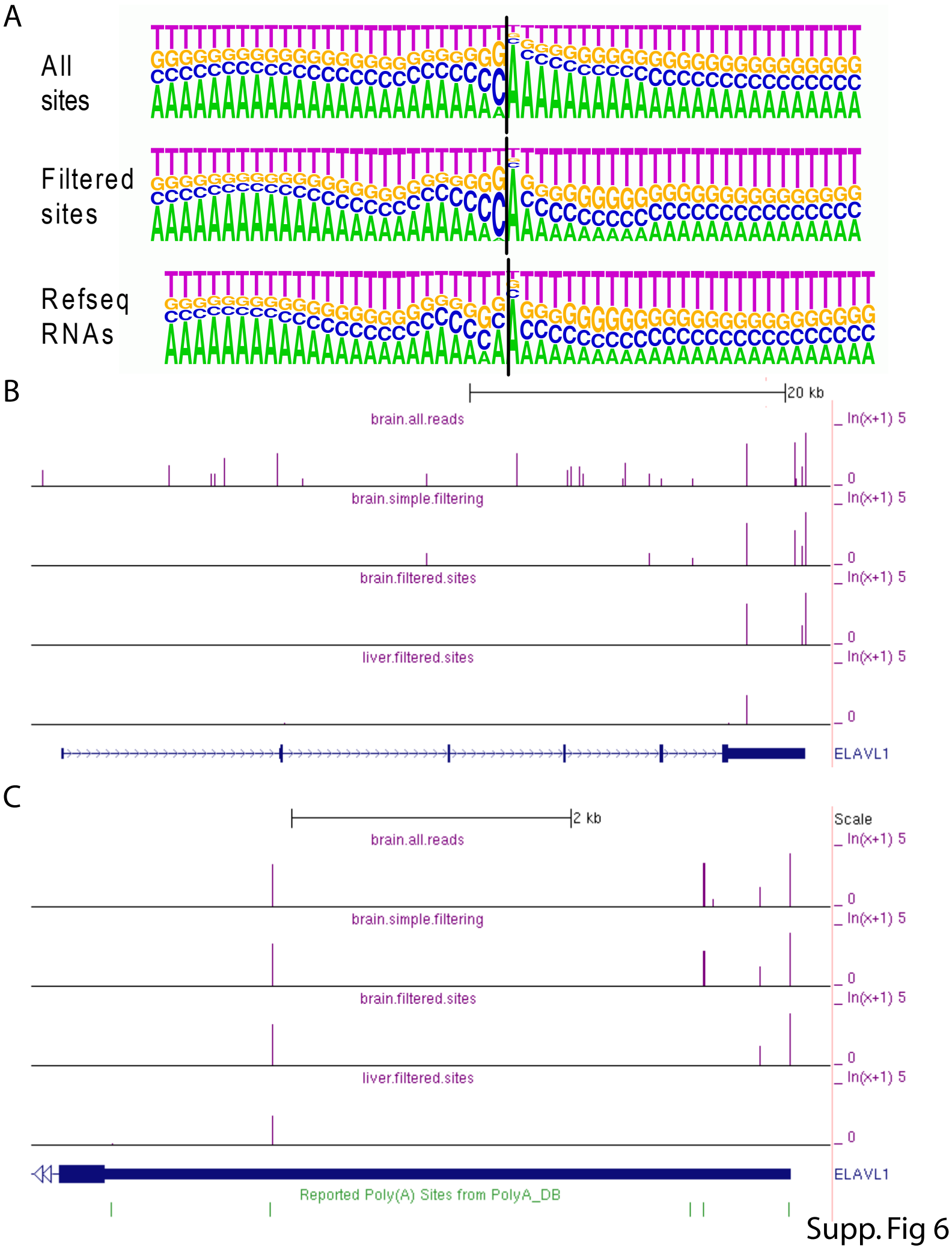
Gel ID	Assay Type	Gene	Site Genome Coord (hg18)	Strand	Expected Amplicon Length	RT Primer	PCR-Fwd	PCR-Rev
1	PolyA Site Validation (Fig 3/S6)	LGI4-1	chr19:40313945	-	>137	T(20)	CAAGAGGCTCGTCTCACTCC	GCT(20)
2	PolyA Site Validation (Fig 3/S6)	LGI4-2	chr19:40308795	-	>147	T(20)	ACACGCTACATTGGGGACTC	GCT(20)
3	PolyA Site Validation (Fig 3/S6)	LGI4-3	chr19:40307257	-	>148	T(20)	GTGCCTTTTGCGCCTCTT	GCT(20)
4	PolyA Site Validation (Fig 3/S6)	LPPR3/PHP2	chr19:763489	-	>147	T(20)	GCTACTTCCGCAAGATGCAG	GCT(20)
5	PolyA Site Validation (Fig 3/S6)	PTBP1-1	chr19:763222	+	>138	T(20)	ACAGCAATTCCAGGCTCAGT	GCT(20)
6	PolyA Site Validation (Fig 3/S6)	PTBP1-2	chr19:763312	+	>111	T(20)	TCCCTTGTCTAGCCCTGTGT	GCT(20)
7	Positive Control	BCAM	chr19:50016482	+	>60	T(20)	CCATCATCTGTGGACACTGG	GCT(20)
8	Positive Control	GRIK5	chr19:47194414	-	>123	T(20)	GGGGAGAAACCTCGGAATTT	GCT(20)
9	Positive Control	HCN2	chr19:568065	+	>115	T(20)	AACTTTGCATGTTCTTGTTTTGTT	GCT(20)
10	Positive Control	PRNP	chr20:4630172	+	>87	T(20)	CGCCGTGATGAATGTACTGA	GCT(20)
11	Internal Priming Site Validation	FRMD4A	chr10:14,282,048-14,282,346	-	239	T(20)	CTGAGTCCATCCCCATCAGT	GCT(20)
12	Internal Priming Site Validation	CELF2	chr10:11,119,674-11,119,942	+	309	T(20)	CCTAGCTTGGGCTGTTGAGT	GCT(20)
13	Internal Priming Site Validation	TMTC2	chr12:81,687,127-81,687,375	+	246	T(20)	AGGCGGGTGAAACACCTTAG	GCT(20)
14	Internal Priming Site Validation	PAFAH1B1	chr17:2,500,286-2,500,552	+	267	T(20)	CGCCTGTAATCCCAACACTT	GCT(20)
15	Internal Priming Site Validation	FBXL20	chr17:34,808,469-34,808,742	-	232	T(20)	GGAGTTCAAGACCAGCCTGA	GCT(20)
16	Internal Priming Site Validation	CIDECP	chr3:10,039,640-10,039,892	-	273	T(20)	CGCCTCTAATCCCAGCACT	GCT(20)
17	Internal Priming Site Validation	RBM6	chr3:49,975,305-49,975,533	+	225	T(20)	TCGAGACCATCCTGGCTATC	GCT(20)
18	Internal Priming Site Validation	WHSC1	chr4:1,858,121-1,858,363	+	215	T(20)	AGACCATCCTAGCCAACGTG	GCT(20)
19	Internal Priming Site Validation	SRPK1	chr6:35,919,779-35,920,038	-	225	T(20)	GATCGAGACCATCCTGGCTA	GCT(20)
20	Internal Priming Site Validation	MOSPD3	chr7:100,049,784-100,050,015	+	247	T(20)	GAGGTGGGCTGATAACCTGA	GCT(20)

Supplemental Figure 4

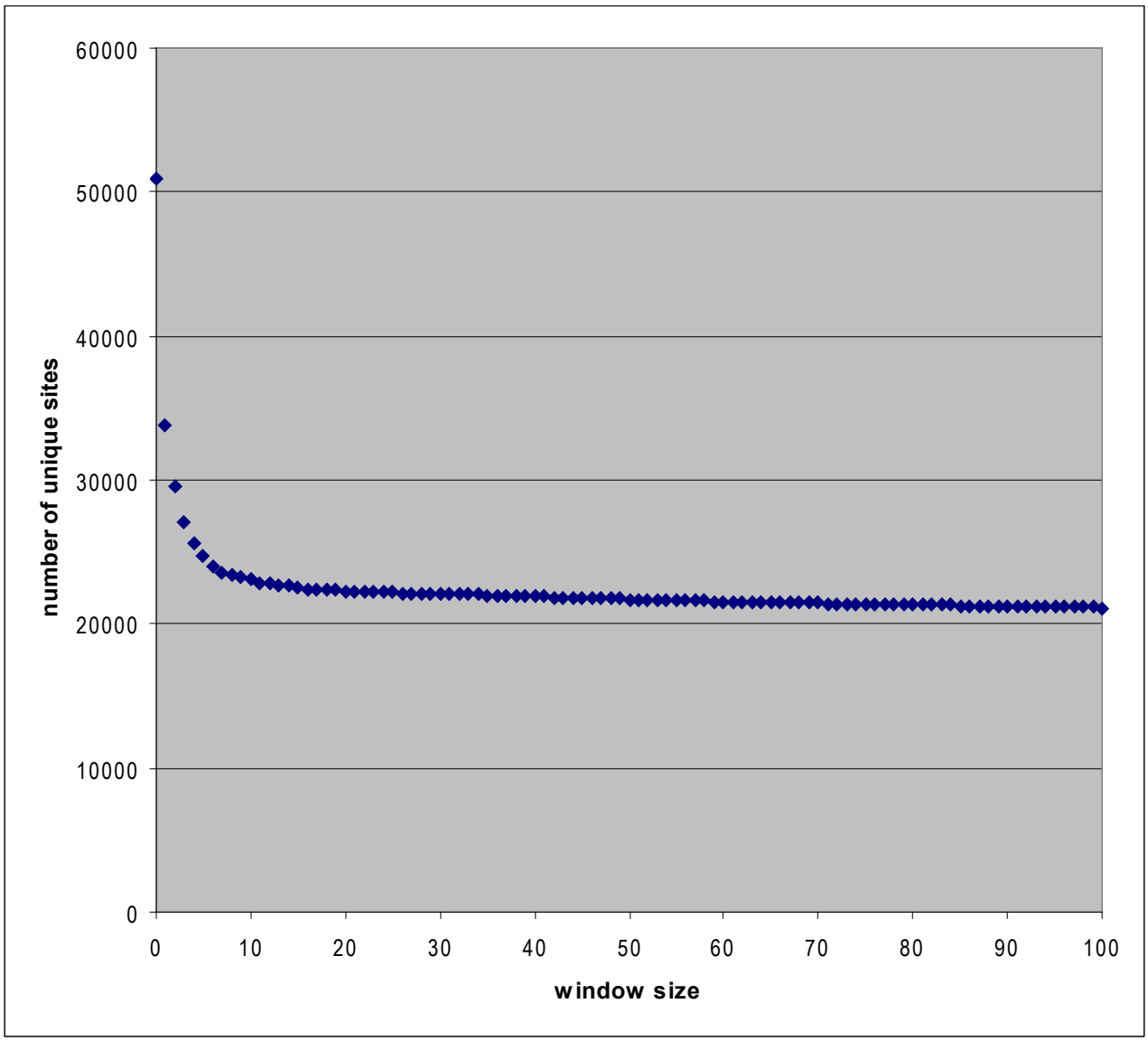




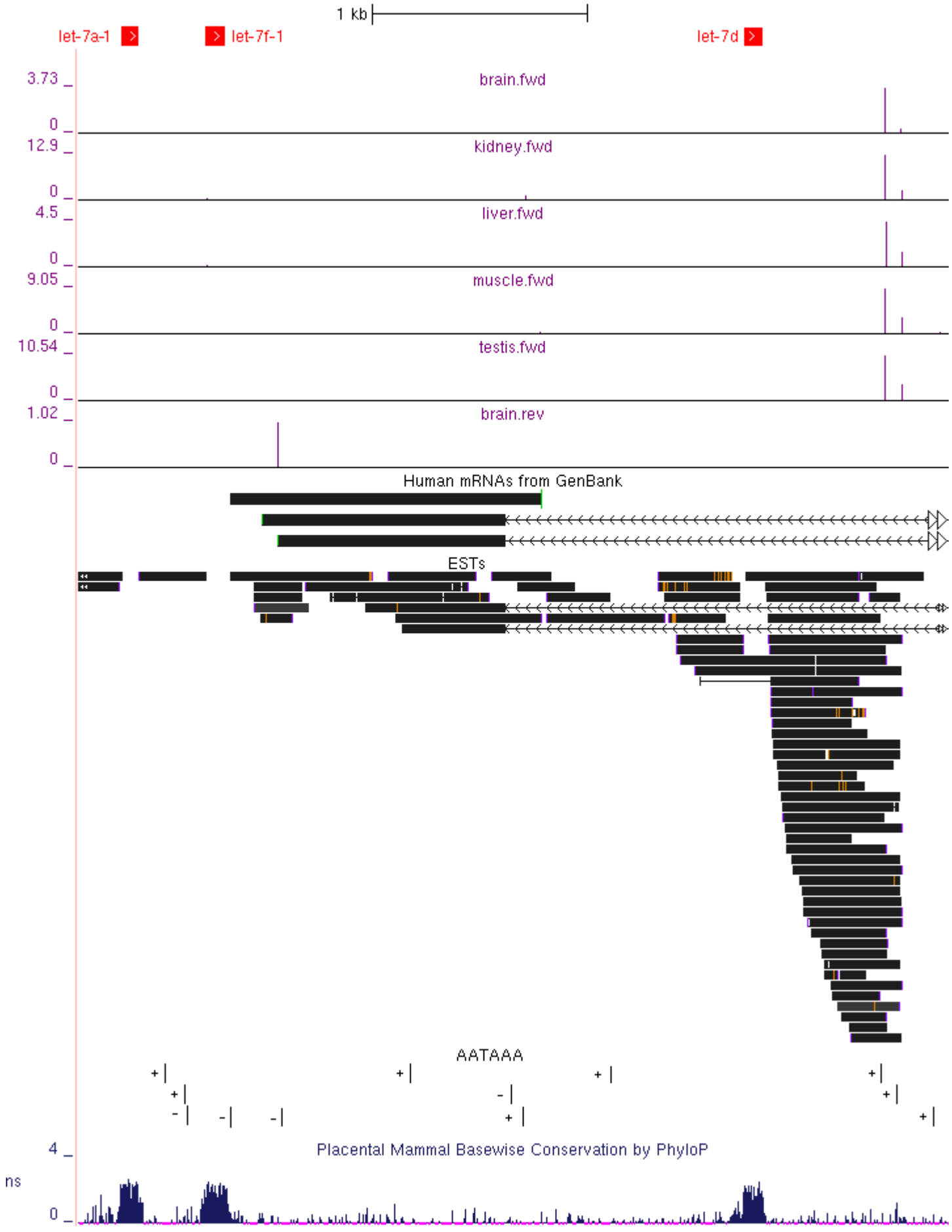
Supplemental Figure 5


A
All sites
Filtered sites
Refseq RNAs
B
20 kb
brain.all.reads
brain.simple.filtering
brain.filtered.sites
liver.filtered.sites
ln(x+1) 5
0
ELAVL1
C
2 kb
Scale
brain.all.reads
brain.simple.filtering
brain.filtered.sites
liver.filtered.sites
ln(x+1) 5
0
ELAVL1
Reported Poly(A) Sites from PolyA_DB

Supp. Fig 6

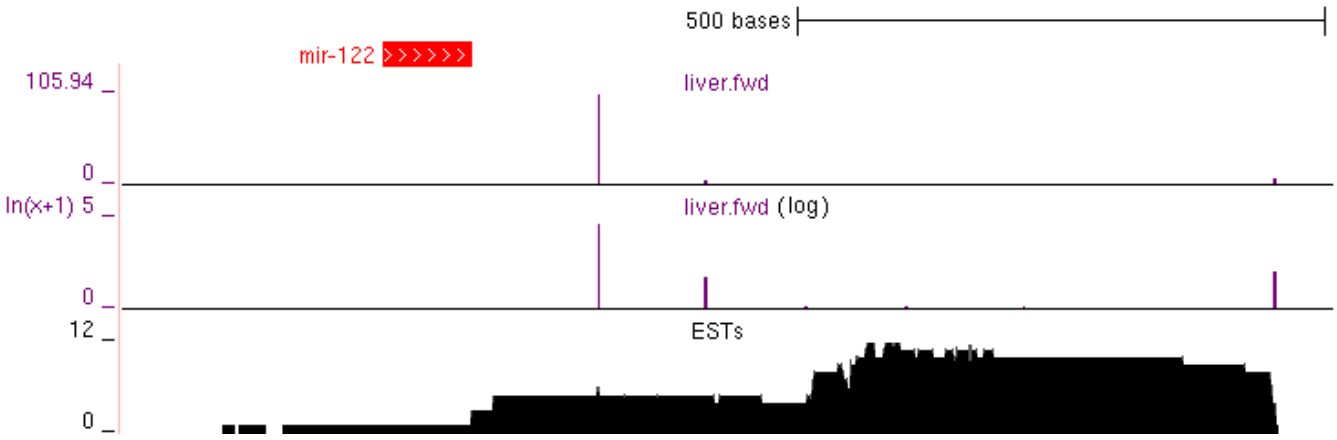


Supplementary Figure 7

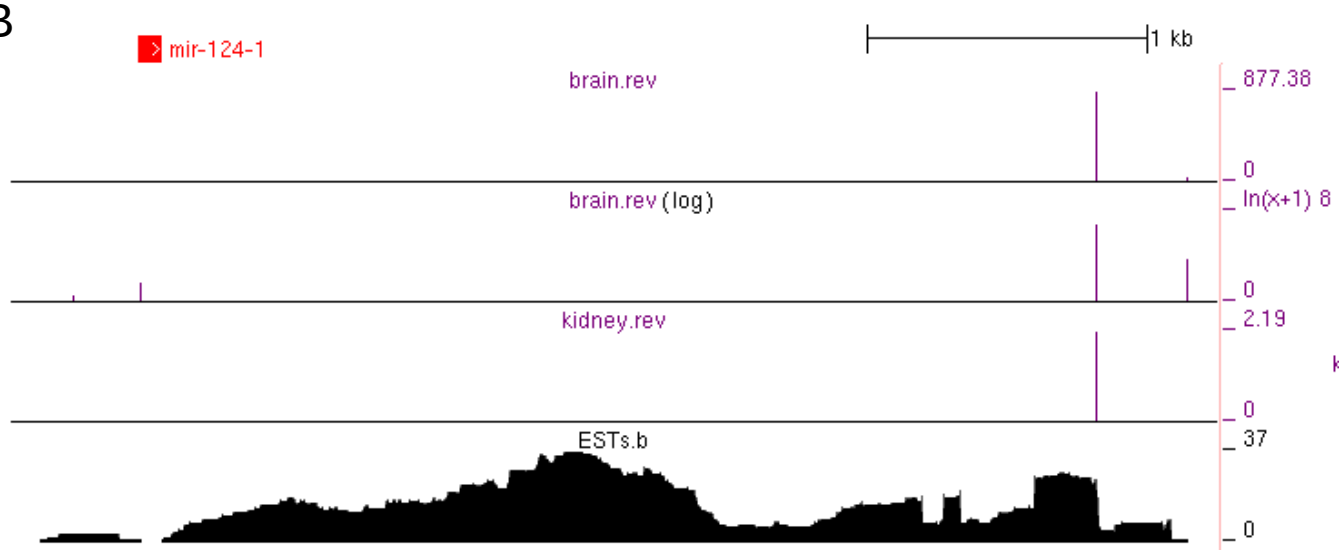

1 kb
let-7a-1
let-7f-1
let-7d
3.73
brain.fwd
0
12.9
kidney.fwd
0
4.5
liver.fwd
0
9.05
muscle.fwd
0
10.54
testis.fwd
0
1.02
brain.rev
0
Human mRNAs from GenBank
ESTs
AATAAA
4
Placental Mammal Basewise Conservation by PhyloP
ns
0

Supp. Fig. 8

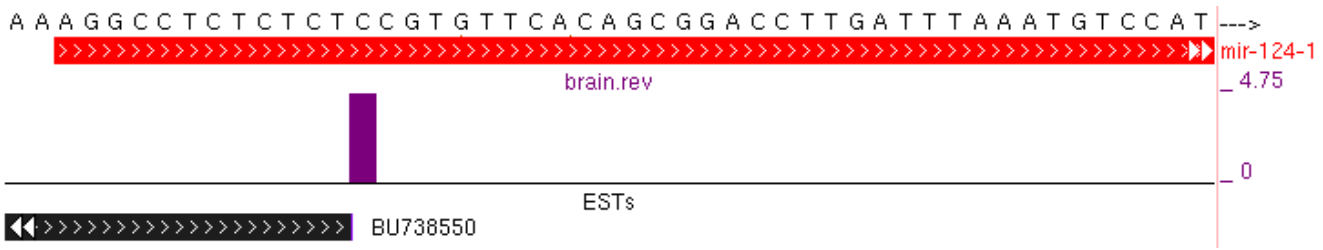
A



B

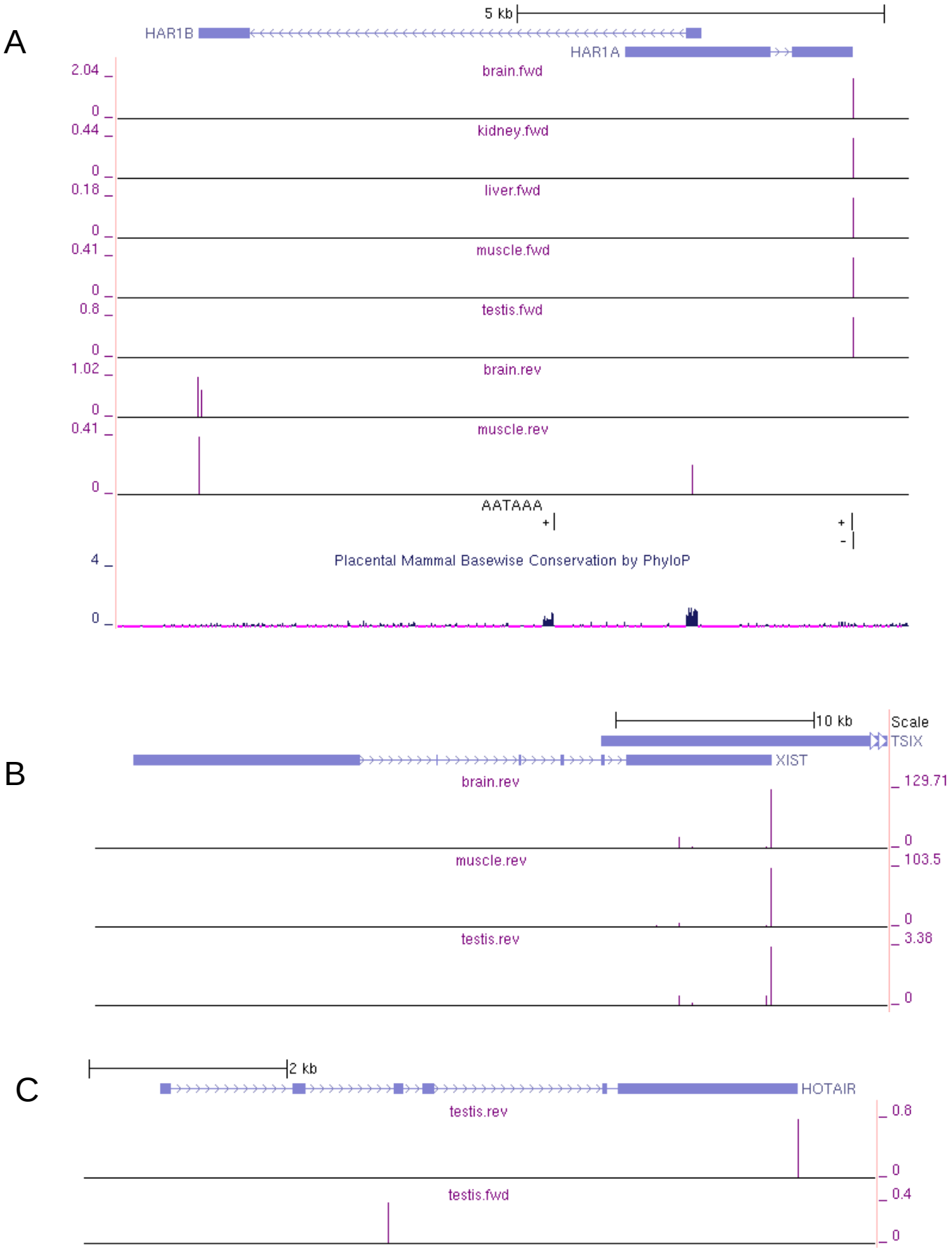


C

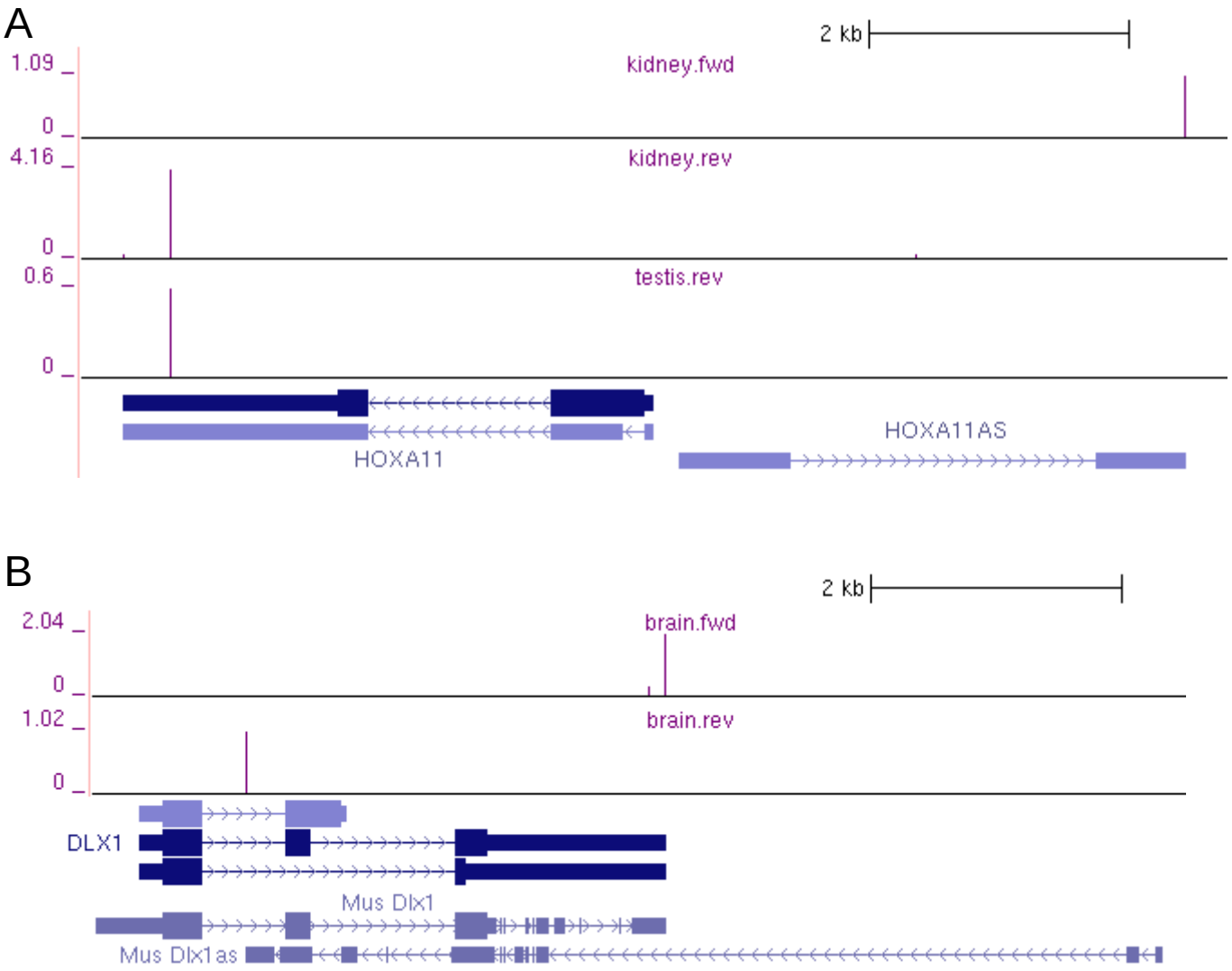


D

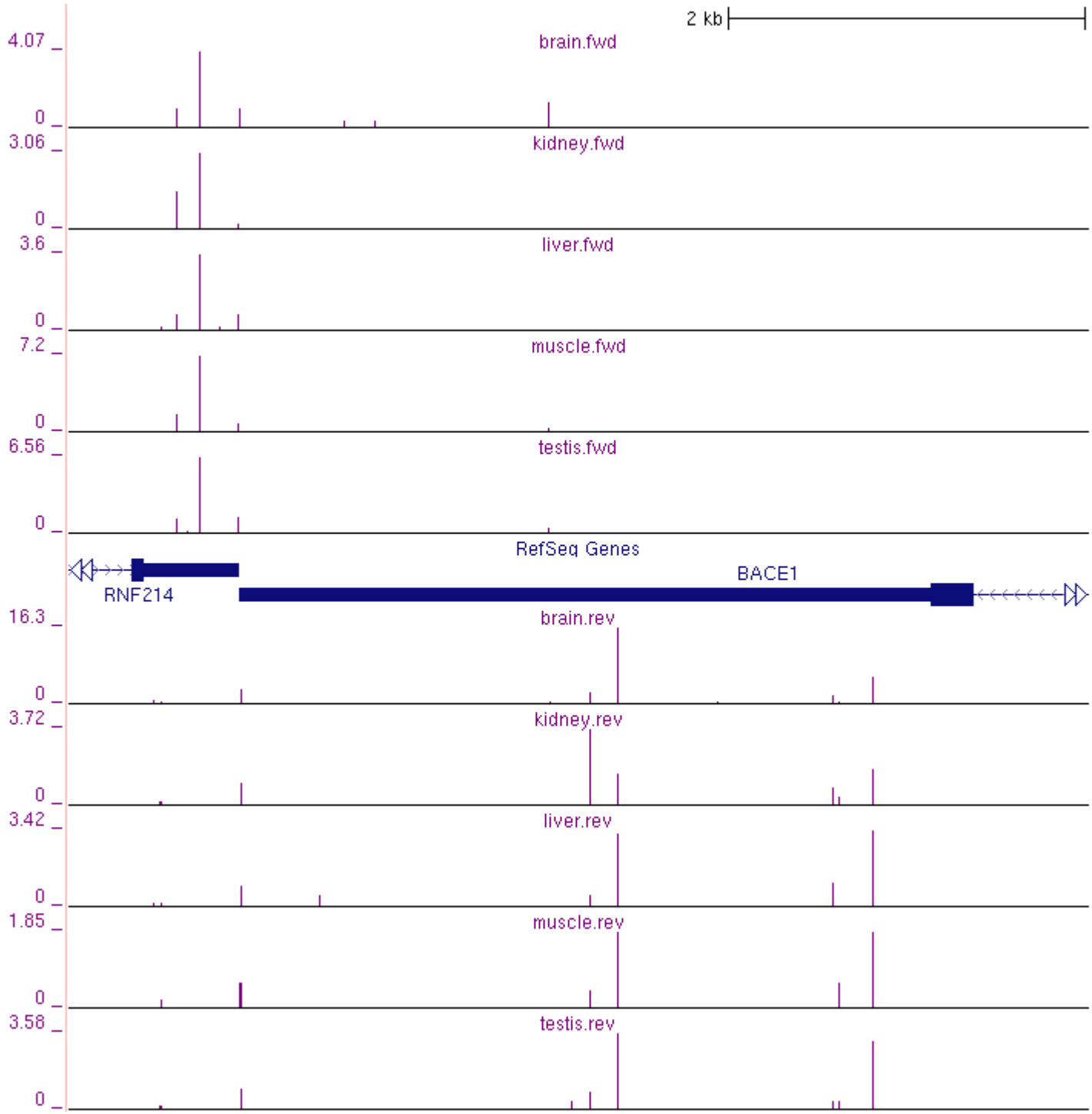
Supp. Fig. 9


A
5 kb
HAR1B
HAR1A
2.04
brain.fwd
0
0.44
kidney.fwd
0
0.18
liver.fwd
0
0.41
muscle.fwd
0
0.8
testis.fwd
0
1.02
brain.rev
0
0.41
muscle.rev
0
AATAAA
+
+
-
4
Placental Mammal Basewise Conservation by PhyloP
0
B
10 kb
Scale
TSIX
XIST
brain.rev
129.71
0
muscle.rev
103.5
0
testis.rev
3.38
0
C
2 kb
HOTAIR
testis.rev
0.8
0
testis.fwd
0.4
0

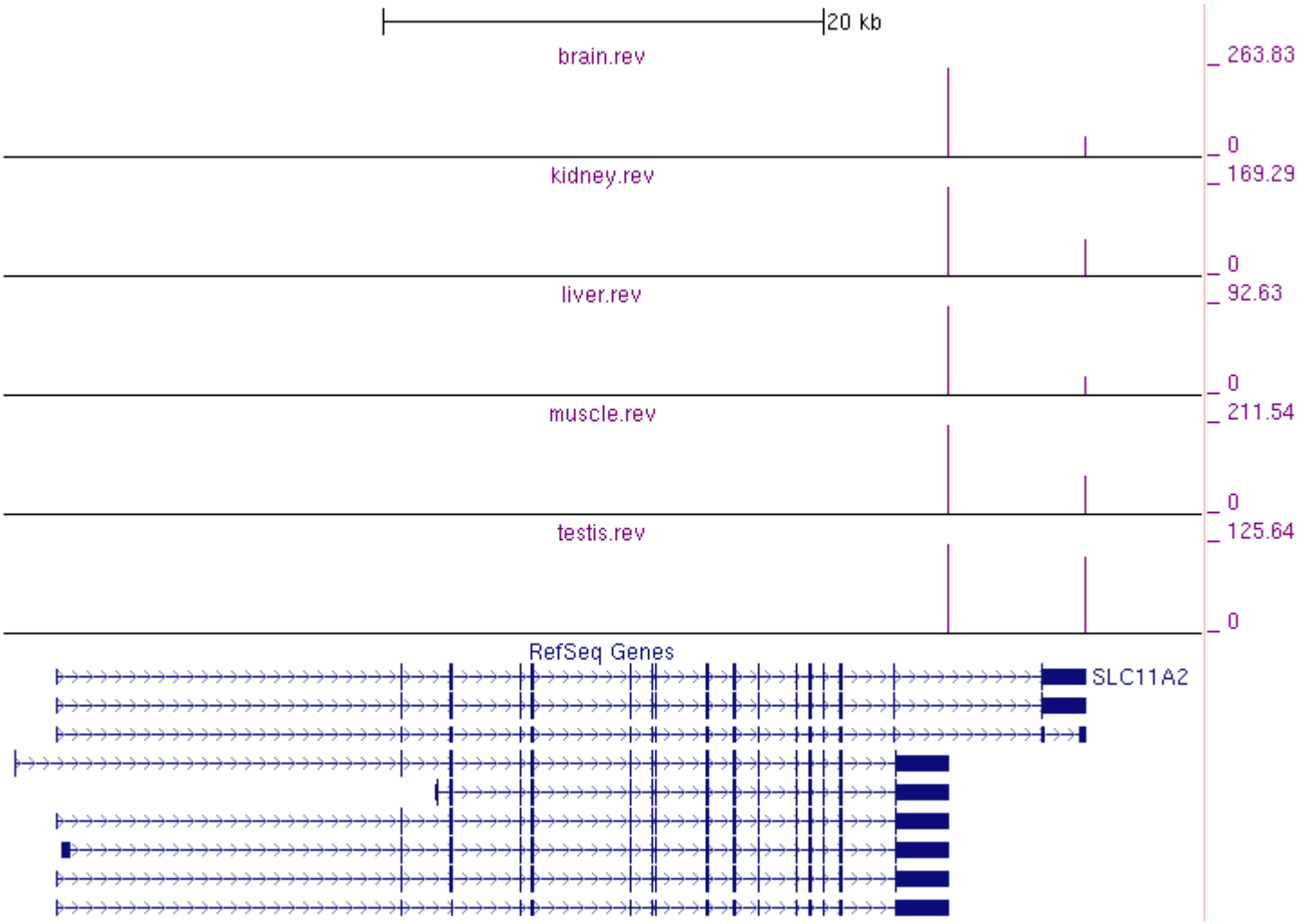
Supp. Fig. 10


A
2 kb
1.09
kidney.fwd
0
4.16
kidney.rev
0
0.6
testis.rev
0
HOXA11
HOXA11AS
B
2 kb
2.04
brain.fwd
0
1.02
brain.rev
0
DLX1
Mus Dlx1
Mus Dlx1as

Supp. Fig. 11


2 kb
4.07
brain.fwd
0
3.06
kidney.fwd
0
3.6
liver.fwd
0
7.2
muscle.fwd
0
6.56
testis.fwd
0
RefSeq Genes
RNF214
BACE1
16.3
brain.rev
0
3.72
kidney.rev
0
3.42
liver.rev
0
1.85
muscle.rev
0
3.58
testis.rev
0

Supp. Fig. 12


20 kb
brain.rev
263.83
0
kidney.rev
169.29
0
liver.rev
92.63
0
muscle.rev
211.54
0
testis.rev
125.64
0
RefSeq Genes
SLC11A2

Supp. Fig. 13

Supplemental References

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**(3): R25.

Lee, J.Y., Yeh, I., Park, J.Y., and Tian, B. 2007. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic acids research* **35**(Database issue): D165-168.

Shepard, P.J., Choi, E.A., Lu, J., Flanagan, L.A., Hertel, K.J., and Shi, Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA (New York, NY* **17**(4): 761-772.

Slomovic, S., Fremder, E., Staals, R.H., Pruijn, G.J., and Schuster, G. 2010. Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells. *Proceedings of the National Academy of Sciences of the United States of America* **107**(16): 7407-7412.

Supplemental Tables

Supplemental Table 1. Comparison of polyA site filtering methods. Two sets of mouse polyA-biased sequencing reads were aligned to the mouse genome (see Methods), one reported here (PolyA-Seq) and another reported previously (PAS-Seq, Shepard et al.; see text for further details). Following Shepard et al. (Shepard et al. 2011), leading Ts were clipped from PAS-Seq reads prior to alignment, reads containing more than 12 Ts were excluded, and reads were separately aligned with Bowtie (Langmead et al. 2009). Alignments were then filtered using one of two methods. First, as previously reported (Shepard et al. 2011), reads were excluded if there were six consecutive adenines or at least seven adenines total in the 10 nt immediately downstream of the read alignment (downstream adenines). Second, alignments were filtered based on the log-odds ratio of base frequency matrices for validated and false-positive polyA sites, described in this work (base frequencies). In all cases, the downstream adenines method excluded a subset of reads filtered by the base frequency method (union). Blat was used to align the 3P-Seq since it is better able to cope with overhangs (i.e. soft-clip) and we consequently observed better alignment sensitivity. We note that our ability to measure the internal priming rate of PAS-Seq and 3P-Seq may not be accurate since our filtering model was calibrated using a different molecular approach. For 3P-Seq in particular we are likely overestimating the false-negative rate since manual inspection of about a dozen sites revealed that most map to the ends of previously annotated transcripts. The major conclusion is that our base-frequencies filter captures all sites flagged by simple filtering and that all methods likely have some degree of internal priming.

Species	Sequencing Method	Sample	Aligner	Reads * 10^6		Fraction excluded by filtering		
				Total	Aligned uniquely	Downstream adenines*	Base frequencies	Union
Mouse	PolyA-Seq	Kidney	Soap2	10.7	8.1	30%	40%	40%
Mouse	PAS-Seq	ES cells	Soap2	6.2	2.5	24%	44%	44%
Mouse	PAS-Seq	ES cells	Bowtie	6.2	1.7	23%	40%	40%
C. elegans	3P-Seq	Multiple	Blat	136.9	94.2	3%	17%	17%

*Downstream adenines failed to identify ~100,000 PAS-Seq reads aligned within an rRNA, due to non-contiguous adenine triplets. In contrast, the most abundant PAS-Seq site obtained with the base frequency method was the polyA site of 60S ribosomal protein L8.

Supplemental Table 2. Information related to sample/RNA sources used in this study.

Sample	Species	Vendor	Catalog #	Lot #	Age of Organism
MAQC-UHR1	Human	Stratagene(Agilent)	740000-41	6054210	Na
MAQC-UHR2	Human	Stratagene(Agilent)	740000-41	6054210	Na
MAQC-Brain1	Human	Ambion (ABI)	AM6051	105P055201A	12 individuals (age range 23-86)
MAQC-Brain2	Human	Ambion (ABI)	AM6051	105P055201A	12 individuals (age range 23-86)
Brain	Human	Zyagen	HR-201	na	Na
Kidney	Human	Zyagen	HR-901	na	Na
Liver	Human	Zyagen	HR-314	na	na
Testis	Human	Zyagen	HR-401	na	na
Muscle	Human	Zyagen	HR-102	na	na
Liver	Rhesus	Merck	na	na	na
Brain	Rhesus	Merck	na	na	na
Kidney	Rhesus	Merck	na	na	na
Testis	Rhesus	Merck	na	na	na
Ileum	Rhesus	Merck	na	na	na
Brain	Dog	BioChain	R1734035-50	A902122	4 yr (male, single donor)
Kidney	Dog	BioChain	R1734142-50	A804152	4 yr (male, single donor)
Testis	Dog	BioChain	R1734260-50	A909069	4 yr (male, single donor)
Brain	Mouse	BioChain	R1334035-50	A807034	12 weeks old, female, single donor
Kidney	Mouse	BioChain	R1334142-50	A509325	3-4 weeks old, female, 20 donors
Liver	Mouse	BioChain	R1334149-50	A601587	8 weeks, female, single donor
Testis	Mouse	BioChain	R1334260-50	A502092	6-8 weeks old, male, 50 donors
Muscle	Mouse	BioChain	R1334171-50	B307160	6-8 weeks old, female, 30 donors
Brain	Rat	Zyagen	RR-201	na	na
Testis	Rat	Zyagen	RR-401	na	na