Supporting Information: Isoform level interpretation of high-throughput proteomic data enabled by deep integration with RNA-seq

[+]Carlyle, Becky. C[1]., [+]Kitchen, Robert .R[1,2]., Zhang, Jing[2]., Wilson, Rashaun S[3]., Lam, Tukiet T[2,3,4]., Rozowsky, Joel S[2]., Williams, Kenneth R[2,3]., Sestan, Nenad[5]., *Gerstein, Mark B[2]., and *Nairn, Angus C[1].

[1] Department of Psychiatry, Yale School of Medicine, Connecticut Mental Health Center, 34 Park St, New Haven, CT 06519

[2] Department of Molecular Biophysics & Biochemistry, Yale School of Medicine, PO Box 208114, New Haven, CT, 06520

[3] Yale/NIDA Neuroproteomics Center, Yale School of Medicine, 300 George Street, New Haven, CT 06510

[4] W.M. Keck Biotechnology Resource Laboratory, Yale School of Medicine, 300 George Street, New Haven, CT 06510

[5] Department of Neuroscience and Kavli Institute for Neuroscience, Departments of Genetics and Psychiatry, Section of Comparative Medicine, and Yale Child Study Center, Program in Cellular Neuroscience, Neurodegeneration and Repair, Yale School of Medicine, New Haven, CT 06510

[+] These authors contributed equally to this work

* Correspondence should be addressed to A.C.N. (angus.nairn@yale.edu, 203 974 7725) and M.B.G. (mark.gerstein@yale.edu, 203 432 6105)

Contents of Supporting Information File

Figure S1

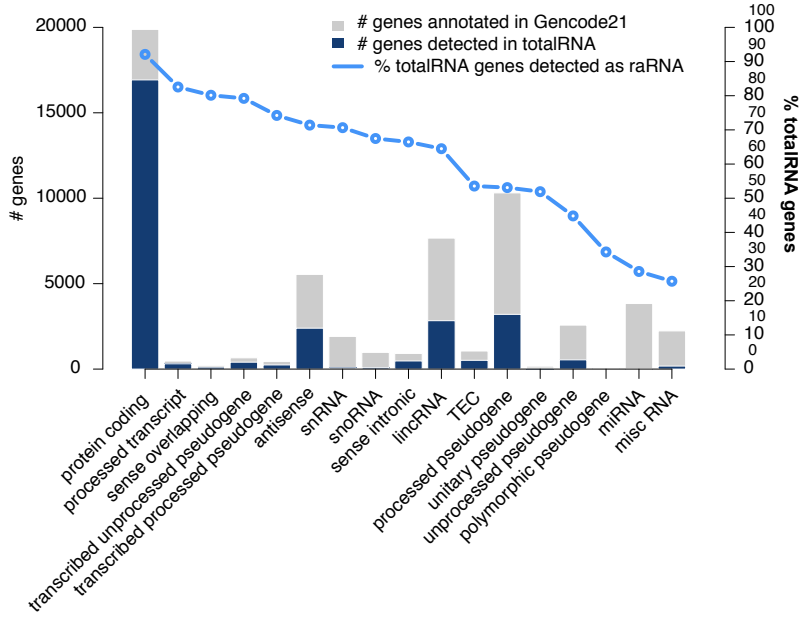

Figure S1 I Non-coding RNA biotypes are depleted in raRNAs

The number of genes detected by totalRNA (dark blue bars, left hand axis, #genes) is compared to the number of genes in the annotation (light grey bars) for each biotype. The overlaid line shows the percent of genes observed in totalRNA samples also detected as raRNA (right hand axis, % totalRNA genes). 95% of mRNAs ('protein coding') detected in totalRNA were observed as raRNA. This fraction decreased for other non-coding RNA biotypes such as lincRNAs, where 65% of those observed in total RNA were detected as raRNA, and processed pseudogenes, where 53% of those observed in totalRNA were present in raRNA data

Figure S2

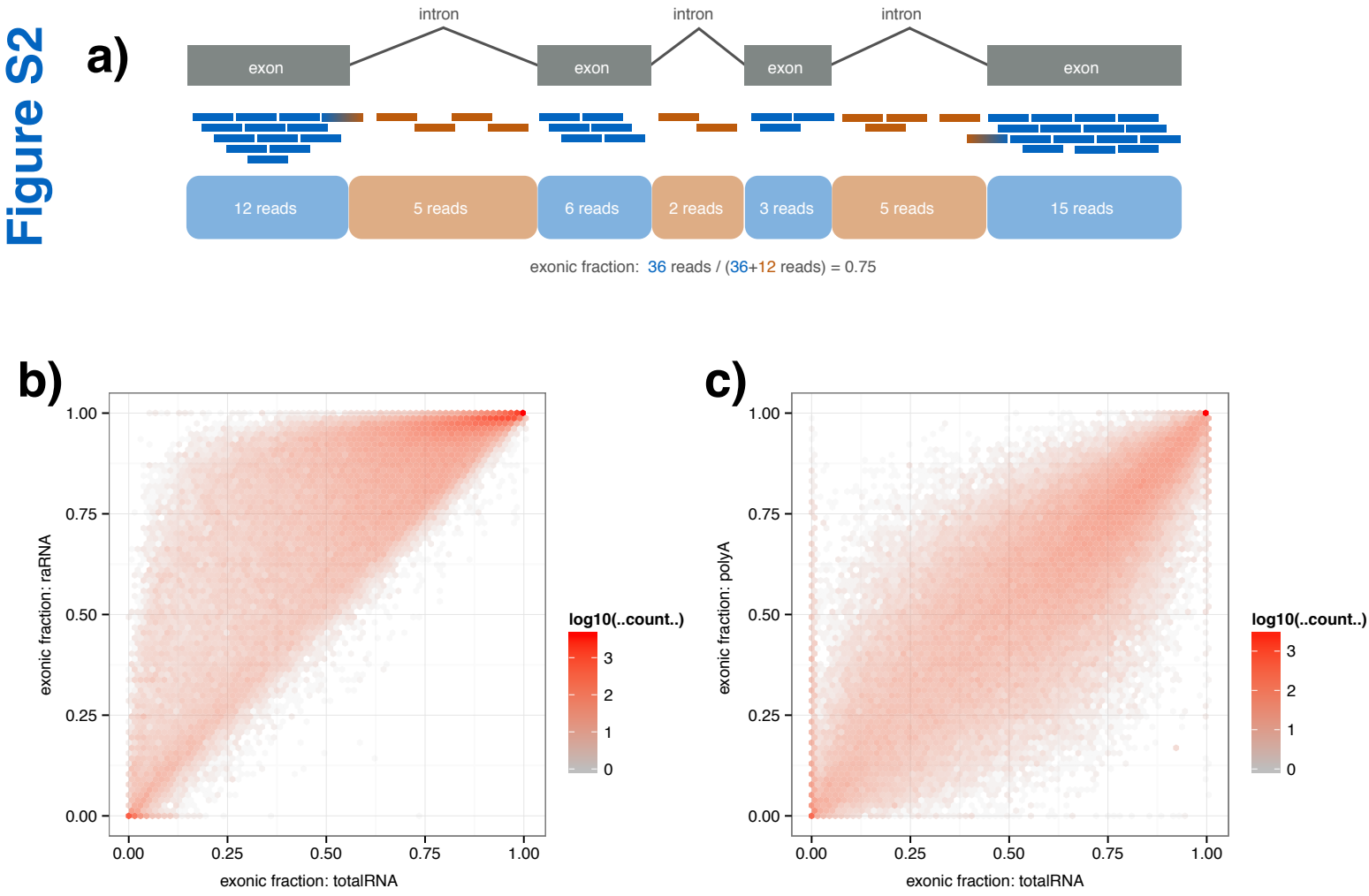

Figure S2 | RNA-seq of raRNA suffers much less intronic 'contamination' than totalRNA

a) Exonic signal is calculated as a ratio: number of exonic reads per gene / total exonic + intronic reads per gene. In this example 36 exonic reads out of a total 48 reads gives an exonic fraction of 0.75. A value of 1 indicates all reads derived from a selected gene are exonic.

b) Density plot comparing exonic fraction from totalRNA (x-axis) with exonic fraction from raRNA (y-axis). Data was skewed towards a higher exonic fraction from raRNA, reflecting capture of mature, cytosolic mRNAs by the ribosome.

c) Density plot comparing exonic fraction from between whole-cell totalRNA (x-axis) and whole-cell poly-A+ RNA-seq (y-axis), both data from the ENCODE K562 cell-line (www.encodeproject.org). Poly-A+ capture did not show an equivalent reduction in intronic signal compared to the raRNA capture in a), likely due to the presence of nuclear polyadenylated pre-mRNA fragments.

Figure S3

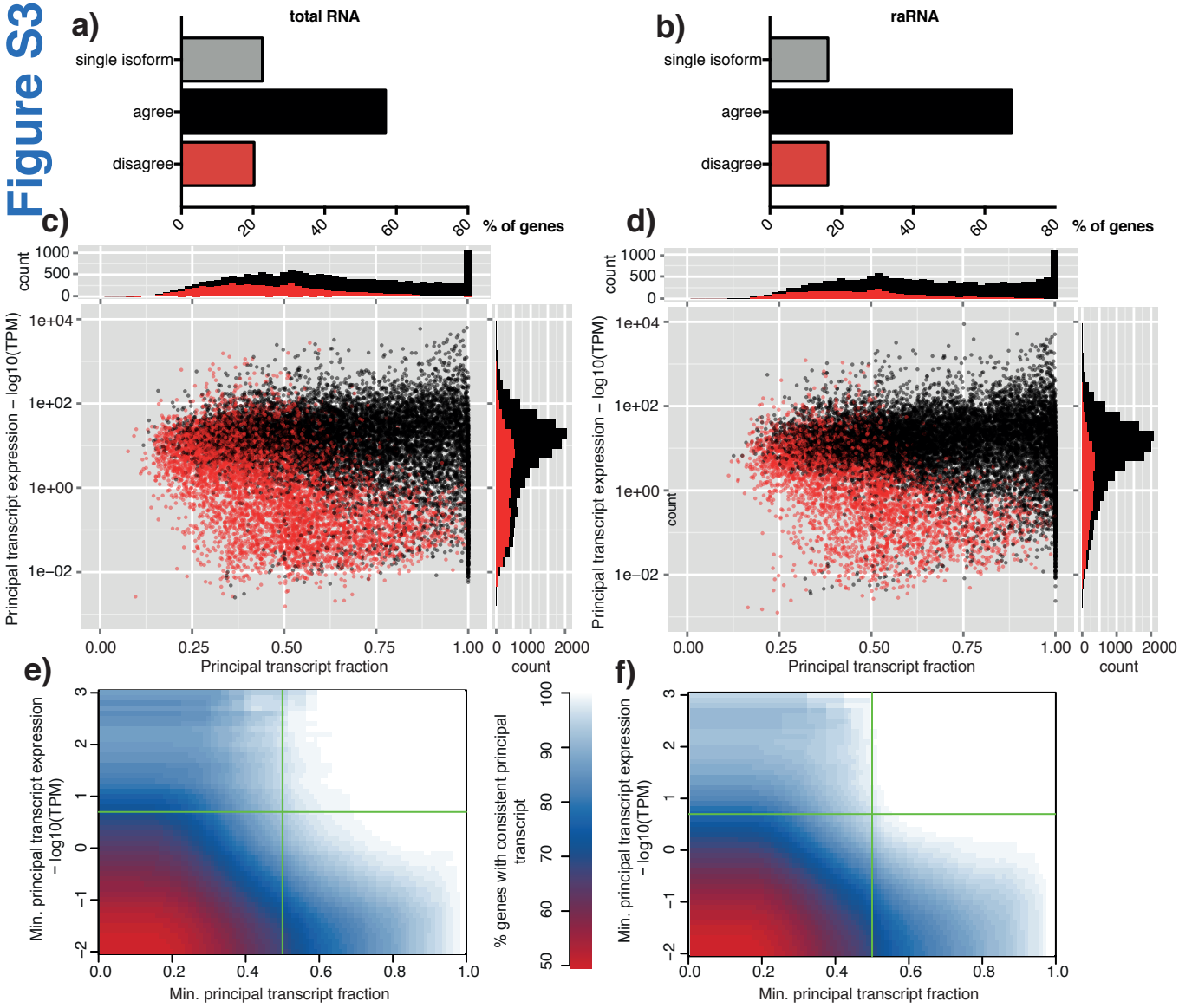

Figure S3 | Principal isoform identification is consistent across biological replicates

The majority of multi-isoform genes from both a) totalRNA-seq (31465 genes) and b) raRNA-seq (25,211 genes) agreed on the same principal isoform (i.e. the highest expressed isoform for each gene) in all three biological replicate samples. Grey = single isoform genes; black: all three replicates agreed; red: at least one replicate showed a different principal isoform. A lack of reliability in detecting the principal isoform may indicate either biological or technical variability between samples, thus these genes are not useful in informing downstream analyses.

For both c) total RNA and d) raRNA, agreement on the same principal isoform across the three replicate samples increased with both increasing dominance of this isoform (as a fraction of the gene expression explained by this isoform; x-axis) and absolute expression of the gene (y-axis; log10 transcripts per million). The lower the TPM and the lower the dominance, the more likely it was that there was a disagreement on the principal isoform between samples.

e) and f) Heat maps show the effect of varying minimum thresholds of gene expression and principal isoform dominance on agreement between replicates. Greater consistency was evident in f) raRNA compared to totalRNA (e), represented by the increased area of white in the upper right quadrant defined by green lines. However, in both cases, more than 90% of genes with a principal isoform (>50% of reads) at more than 5 transcripts per million were consistently defined as having a principal transcript.

Figure S4

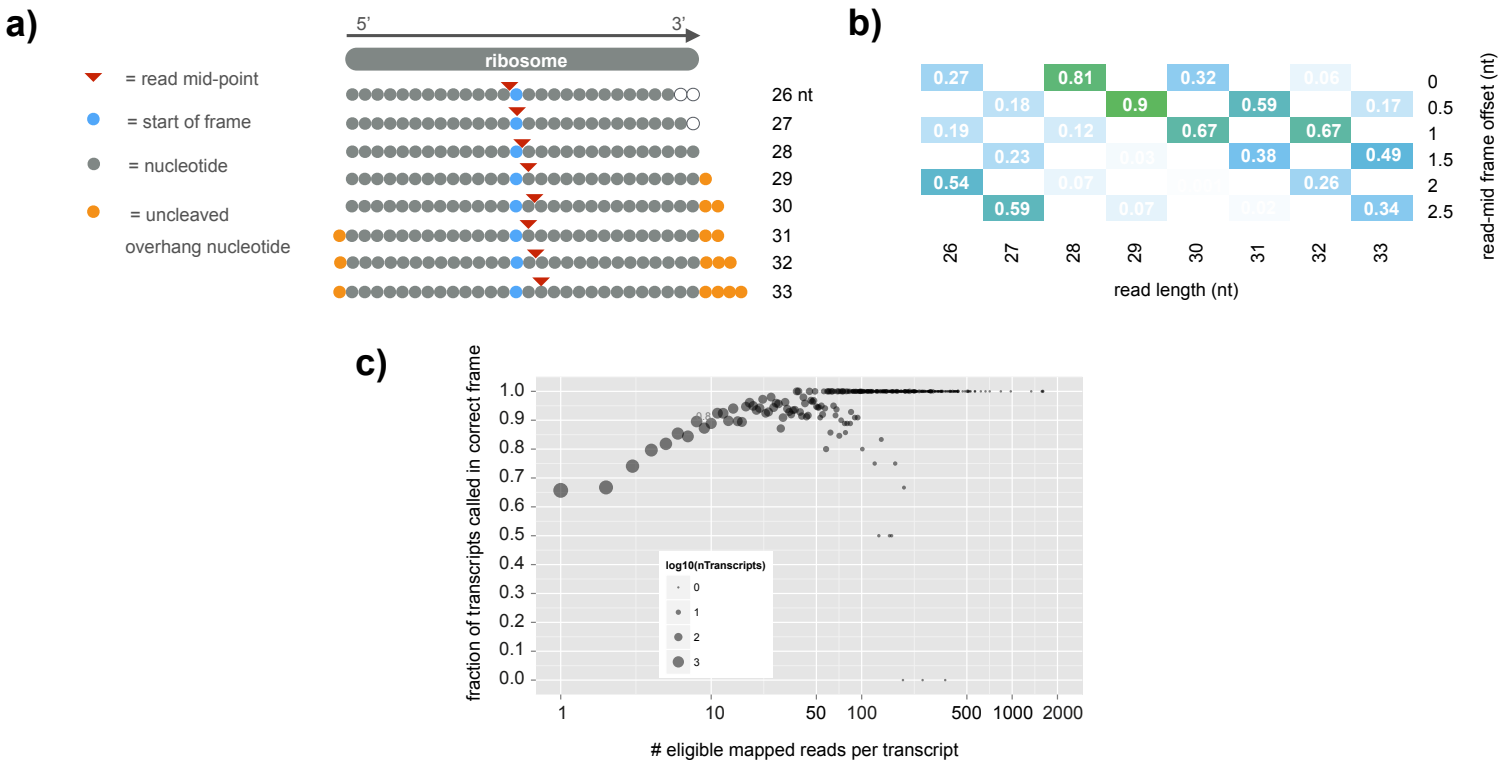

Figure S4 I Ribosome footprints predict the open reading frame

a) In a perfect footprint preparation, all reads would be 28 nt in length (the exact number of nucleotides physically protected by a cycloheximide halted ribosome), and the read midpoint (red arrowhead) would be situated at a predictable number of nucleotides from the start of the nearest codon (zero offset, blue circles). In actuality, due to the variations inherent in the technique such as incomplete RNase cleavage, a range of fragment lengths was obtained. In our hands, incomplete RNase digestion tended to leave uncleaved nucleotides at the 3' end of the footprint (orange circles).

b) Rather than discard these "imperfect" reads, it was possible to compute a position-weight matrix (PWM) for single isoform genes showing the fraction of reads of a given length at each of the three possible offsets to the annotated codon position. Over 59% of all 28-32 nt reads were in frame with each other following ribosome IP, rising to 90% for the most populous 29 nt fragments.

c) The fraction of transcripts called in the correct frame (compared to the annotation) increased with increasing number of observed footprints per transcript. With ≥10 reads it was possible to accurately predict the correct frame for >90% of transcripts.

Figure S5

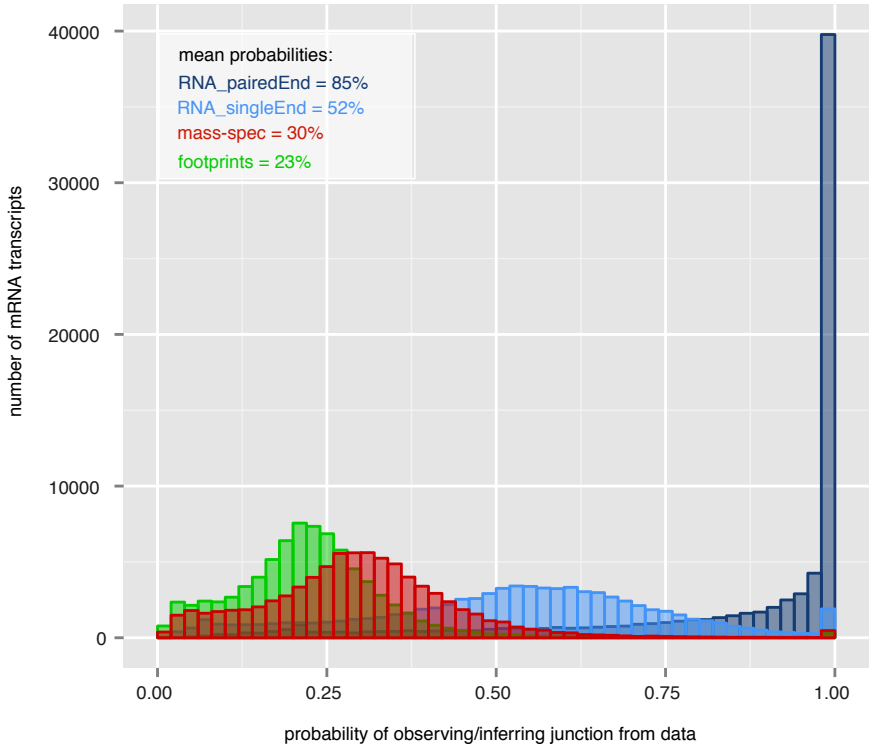

Figure S5 | Paired end 200 nt RNA-seq fragments have a large probability of crossing an exon junction
Probability distribution, over all ~80,000 mRNA transcripts annotated in Gencode, of a randomly selected RNA-seq read, 28 nt ribosome footprint, or 13 amino acid mass spectrometry peptide overlapping a junction between two or more coding exons. Paired-end RNA-seq produces reads from each end of a ~200nt insert sequence and, as such, it is possible to infer the presence of an exon-exon junction anywhere within the insert, even if the reads themselves do not contain the junction. Thus, the likelihood of any given 200nt insert sequence spanning an exon junction within the CDS of an mRNA is extremely high for the vast majority of transcripts (~85%; dark blue bars). Reading 75nt from only a single end of the insert, as for older RNA-seq experiments, leads to a marked reduction in the likelihood of observing an exon junction (~52%; light blue bars) as the insert size can no longer be imputed without the read's pair. Assuming a peptide length of 13 amino-acids, mass-spectrometry produces observations of peptides with a much lower likelihood of spanning a CDS exon junction (~30%; red bars). The 28nt ribosome footprints are the least likely to produce exon-spanning reads (~23%; green bars).

Figure S6

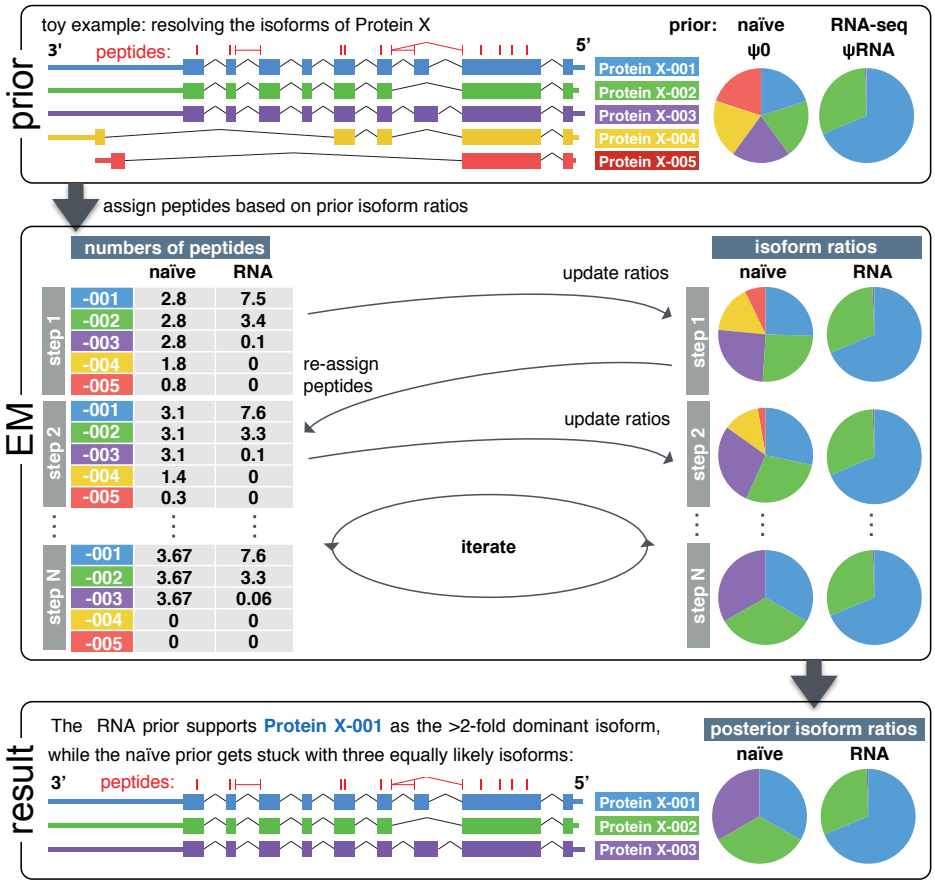

Figure S6 | Analytical approach to integrated analysis of the transcriptome, translatome, and proteome
Schematic diagram depicting the isoform assignment process for a toy example of Protein X, a five isoform gene. The EM algorithm (middle panel) iterates through a process of fractionally assigning each read or peptide to each of the possible isoforms, weighted by the likelihood and length of each isoform. At each step these isoform likelihoods are updated based on this new assignment of the peptides. The uninformative 'naive' prior, in which each transcript is equally likely to generate these observed footprints, converges on three equally likely transcripts which the peptides cannot discriminate between (lower panel). The use of a biologically informed prior, obtained directly from the relative transcript abundances from RNA-seq overcomes ambiguity (lower panel) as the peptides are fully consistent with the transcript abundances for this gene. The biological prior supports Protein X-001 as the greater than two-fold dominant isoform, with Protein X-002 as a minor secondary isoform.

Figure S7

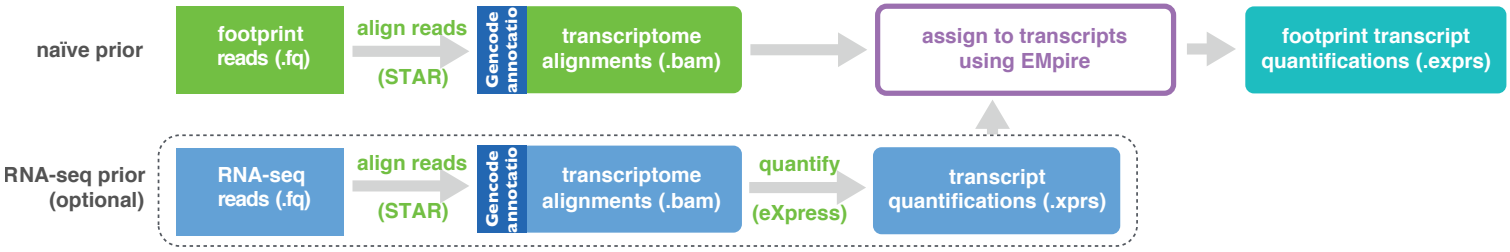

Figure S7 | Analytical workflow for isoform assignment of ribosome footprints
Isoform prediction and assignment for HEK cell ribosome footprints was performed without (top) or with (bottom) an RNA-seq informed biological prior. Here, RNA-seq transcript quantifications are produced by the eXpress tool37 and all footprint alignments are in transcriptome coordinates.

Figure S8

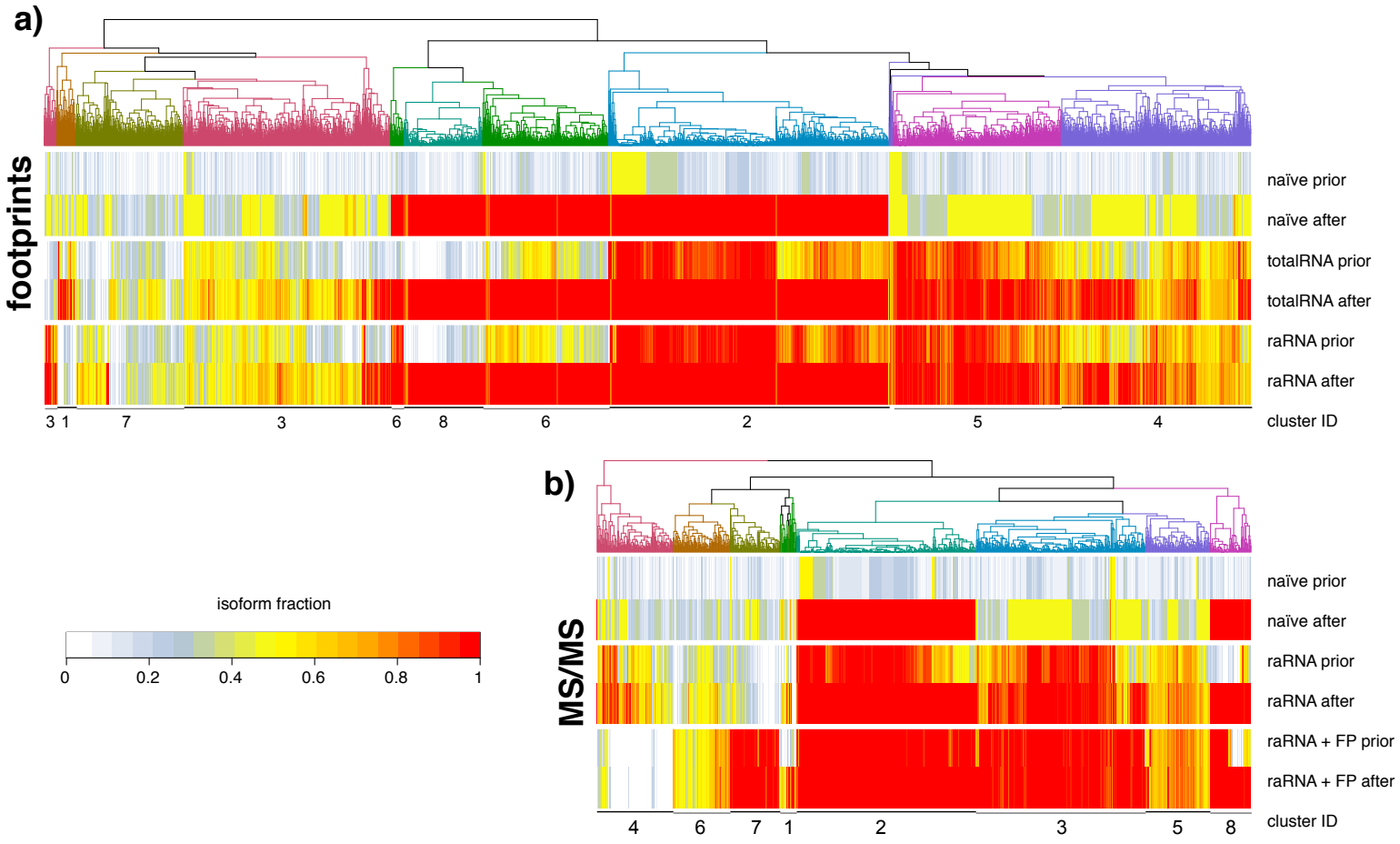

Figure S8 | Use of a biologically informative prior improves isoform level interpretation of ribosome footprints and MS/MS peptides

Heatmaps show the effect of using different priors on the dominance of the principal isoform following the EM. In each doublet row, the top row represents assignment of reads/peptides before EM, the second row represents updated ratios after iterations of EM with a naive or biological prior. Clusters of principal isoforms (x-axes) are indicated by dendrogram colours (top) and numeric IDs (bottom), the latter matching the cluster IDs in Figure 4.

a) The heatmap plots the principal isoform fraction for the 6,650 multi-isoform genes with at least 3 footprint reads. Using a biological prior improved the ability to resolve the principal isoform in 3,795 genes (57.1%) and converged on the same isoform as the naive prior in 2,747 genes (41.3%).

b) The heatmap plots the principal isoform fraction for the 1,212 multi-isoform genes with at least 2 peptides. Using a biological prior improved the ability to resolve the principal isoform in 663 genes (54.7%) and converged on the same isoform as the naive prior in 408 genes (33.7%).

Figure S9

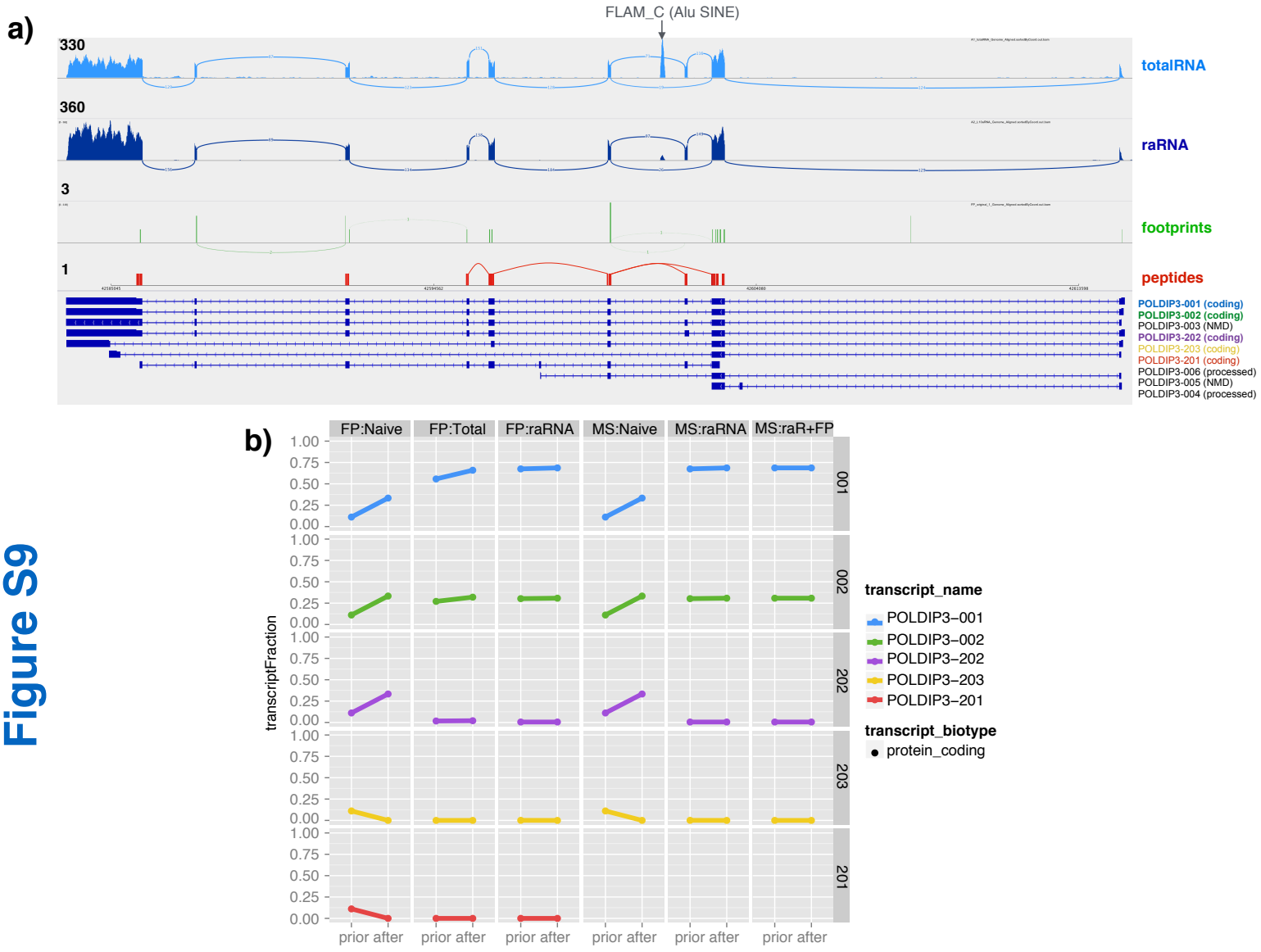

Figure S9 | Detailed summary of alignments and EM performance for POLDIP3

a) Browser track of totalRNA-seq, raRNA-seq, and ribosome footprint alignments to the POLDIP3 gene in genome coordinates. The totalRNA and raRNA reads clearly support two transcripts, POLDIP3-001 and POLDIP3-002, while the ribosome footprints are unable to discriminate between these isoforms. A significant number of total RNA reads mapped to an intronic Alu repeat region (labeled), but this is not a novel POLDIP exon. POLDIP transcript IDs reflect the Oct 2014 build of Ensembl.

b) Expanded EM results highlighted the necessity of the RNA prior (either totalRNA or raRNA) to be able to resolve the difference between these isoforms for both ribosome footprinting and MS/MS proteomics