## EMpire: Deep isoform level integration of the transcriptome, translatome, and proteome

\*Carlyle B,C1., \*Kitchen R.R1,2., Zhang J2., Wilson R. S3., Lam T.T2,4., Rozowsky, J.S2., Williams K. R3., Sestan N5., Gerstein M.B2+., and Nairn A.C1+.

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

+ Correspondence should be addressed to M.B. Gerstein (mark.gerstein@yale.edu) and A.N. (angus.nairn@yale.edu)

**Acknowledgements:** This work was supported by NIDA grant R21DA040454 and NIMH grants U01MH103339, MH106934, and MH11092, by the Yale/NIDA Neuroproteomics Centre (DA018343-12), by NIH SIG grants 1S10OD019967-0 & 1S10ODOD018034-01, and by the State of Connecticut, Department of Mental Health & Addiction Services. B.C. was supported by a 2014 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation.

Mark /Joel/Jing – could you please add the appropriate grants to this list for your lab? Thanks

#

# Abstract

Protein abundance as the outcome of gene expression is subject to regulation at multiple levels through a complex balance of production and turnover. Although extremely sensitive, high throughput mRNA measurements are not an infallible proxy for protein abundance. Complete understanding of gene expression can be gained only through a combination of multi-omic assays. A particular strength of mRNA-seq is in its ability to quantify isoforms and as a result it has been established that most cell types, and even tissues, predominantly express a single principal isoform. However, despite large improvements in the sensitivity of ribosome footprinting and LC-MS/MS proteomics technology over recent years, the similarity of coding sequences limits our ability to allocate footprints or peptides to isoforms. The approach we describe here, *EMpire*, seeks to incorporate isoform-level mRNA-seq quantifications as priors in an Expectation Maximization algorithm to guide and improve allocation of footprints or peptides to isoforms. We show that EMpire identifies a principal isoform in over 80% of genes in homogenous HEK293 cell culture, and over 70% of proteins in complex human brain tissue, while still allowing disagreement with RNA-seq if the evidence is sufficiently strong. The use of EMpire to define isoforms in experiments with matched RNA-seq and proteomic data increases the functional relevance of such datasets, and will further broaden our understanding of multi level control of gene expression.

# Background

Over the last two decades, genome-wide analysis of nucleic acids has rapidly advanced to the point where we can routinely survey the entire genome, epigenome, and RNA transcriptome of any cellular system. Ambitious projects such as 1000 Genomes1,2, The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov), and the Genotype-Tissue Expression project (GTEx)3 have performed deeply integrative analyses of the genome and transcriptome to better understand the impact of DNA variants on human health and disease. While these efforts focus on breadth over a large number of individuals, other high profile projects such as the Encyclopedia of DNA Elements (ENCODE)4, the Roadmap Epigenome Project5, and International Human Epigenome Consortium (IHEC)6 have attempted deeper characterization of the multi-omic landscape of specific cell types, tissues, and species.

Due to ever simpler and cheaper sample preparation, RNA analysis remains the de-facto approach for a genome-wide survey of gene expression and as such features heavily in all of these projects. The incorporation of proteomics into these large studies has lagged behind nucleic acid based techniques. Since the discovery that electrospray ionization could be used to analyze proteins by mass spectrometry in 1989, liquid chromatography mass spectrometry (LC-MS/MS) technology has improved at a remarkable rate, and the analysis of complete cellular proteomes is now within reach7,8. For these large datasets to maximize the potential for interesting functional discoveries, integration with proteomic data is critical, given that it is arguably protein abundance and modification that most closely reflect the biosynthetic state of the cell, not RNA. Recently, studies of the translatome have started to bridge this gap between transcriptome and proteome. By measuring the dynamic profiles of ribosomes as they are translating mRNA to protein through ribosomal footprinting9, as well as steady state mRNA and protein levels, additional and often crucial insight can be obtained related to expression regulation within a cellular system10–13. The potential benefits of a principled integration of a study, from the experimental level through to the analytical, of these highly complementary data modalities are very enticing14,15.

To date the majority of high-throughput studies probing translational dynamics have been in ‘lower’ organisms such as bacteria and yeast16–20, with more recent studies moving via cell-culture to complex eukaryotes12,13,21–25. A major challenge in moving to mammalian systems is the splicing complexity of the transcriptome; current estimates are that over 90% of human multi-exon protein coding genes can transcribe alternatively spliced mRNAs26. Despite the challenging nature of assigning relatively short reads with non-uniform coverage to isoforms, tools for quantifying RNA-seq at the isoform level have become extremely advanced. In a 2015 systematic review, 11 of the commonly used tools were found to have comparable consistency and accuracy, particularly for higher abundance transcripts27. As a result, the most recent improvements to these workflows have focussed on increasing processing speed, decreasing memory usage, and correcting for biases inherent in RNA-seq experiments, such as 3’ bias.

Isoforms should also be an important consideration for analysis of ribosome footprint data, as one can easily envision a situation in which a gene that is producing multiple mRNA transcripts at different abundances could lead to incorrect footprint density calculations, and as a result inappropriate estimates of translational efficiencies. Tools for analysis of ribosome footprint data have proliferated in the last two years, but the majority do not explicitly address allocation of footprints to isoforms28–31, instead focusing on improving sensitivity in analysis of translation efficiency. In such tools, it is common to abstract to the gene-level by choosing a single ‘representative’ transcript; typically one with either the longest coding sequence or the highest density of footprint reads. Recently, Floor and Doudna used Cufflinks in an attempt to obtain transcript-level assignments of footprints12, while *Ribomap* uses mRNA transcript abundance to fractionally assign footprints to isoforms in a single step32. The Cufflinks approach is inherently limited by the coverage of the footprints and the *Ribomap* method does not allow for footprints to disagree with mRNA-seq data in the case of significant post-transcriptional regulation.

Despite vast improvements in the technical implementation of LC-MS/MS in the last decade7,33, distinguishing between isoforms remains a major challenge for data analysis. In a basic mass spectrometry analysis workflow, acquired spectra are searched against a protein sequence database with a set number of entries. Isoforms are then distinguished by unique peptide identification from the resulting peptide spectral matches. Sequence similarity across isoforms can ultimately limit the number of unique peptides and corresponding enzymatic cleavage sites within the protein sequence. Additionally, low-abundance peptides can be present below the detection limit of the mass spectrometer and are often masked by peptides of higher abundance34,35. Finally, the amino acid sequence and charge states of the peptide influence chromatographic elution and ionization efficiency, which can also affect peptide detection36–38. Collectively, these issues can result in identification of primarily redundant peptides, leading to isoform ambiguity.

Current mass spectrometry analysis software use a protein grouping feature, which organizes spectral counts for all redundant peptides into groups representing the entire isoform family39–43. Protein grouping is beneficial because is retains spectral information, but does not provide a solution to isoform ambiguity. Alternatively, one can perform a targeted spectral search against a database containing only sequences corresponding to the isoforms of interest. In this way, the search algorithm is manipulated with fewer database entries in an attempt to increase the probability of identifying unique peptides. This method is not always successful, and can complicate the analysis when performing quantitative comparisons. Top-down proteomics is another useful strategy for isoform identification, but analysis of intact proteins is often challenging44,45. Recently, targeted mass spectrometry approaches such as parallel reaction monitoring (PRM) have been employed to distinguish between isoforms46–52. This method relies on targeting specific masses for peptide identification, and its sensitivity allows identification of low-abundance peptides. While PRM is a promising advancement toward solving the problem of isoform ambiguity, it is still limited by the availability of unique peptides within the protein sequence.

As the ability to profile experimental systems on a multi-omic level rapidly increases, tools must be developed for deep analytical integration of methodologies53–57. While it is undoubtably powerful to perform multi-omic experiments on the same tissues and compare data post hoc, exploitation of the vastly greater isoform resolving power of RNA-seq to study the translatome and proteome should be considered. The tool presented here identifies the major isoform(s) for each gene from transcriptome sequencing data, using this data to guide footprints and peptides iteratively towards these same isoforms. This iterative process is critical as it still allows footprints and peptides to diverge from the RNA-seq prediction if there is sufficient evidence against that particular mRNA. Specifically, we combine extraction of total-RNA, purified transcripts engaged by the ribosome, ribosome footprints, and proteomics with a novel analytic approach to integrating these data at the isoform level. Our software tool, **EMpire** (**EM** **P**ropagation of **I**soform abundance from **R**NA **E**xpression) employs an expectation-maximisation algorithm to resolve transcripts producing ribosome footprints and protein isoforms producing peptides. We show that use of a ‘biologically informative’ mRNA prior is extremely effective at resolving isoform ambiguity.  Although the tool was designed for ribosome footprint and peptide data, it could be easily modified to take data from any paired sample type, such as HTS CLiP and ChIP-seq data.

Results

### Integrated experiments for profiling the transcriptome, translatome, and proteome

In order to fully explore the possibilities for isoform-level integration of RNA, ribosome footprint, and proteomic data, we designed a series of assays to be run in parallel on the same cellular sample (Figure 1a). For this proof of principle we used a modified stable human cell-line (HEK293-L10a), collecting RNA-sequencing (RNA-seq) data at two levels, total cellular RNA (‘**totalRNA**’) as well as only those transcripts engaged by the ribosome (ribosome associated RNA; ‘**raRNA**’). We also obtained ribosome footprinting (‘**FP**’) through an immunoprecipitation (IP)-based approach. Finally we obtained mass-spectrometry (LC-MS/MS**;** ‘**MS**’) proteomic data in ‘discovery’ mode, which relies on fractionating samples to be able to identify more peptides. Each of these assays was carried out over three biological replicate samples.

These assays differ not only in their molecular target but also their sensitivity. The depth of coverage, in terms of genes detected at all levels of expression, is unsurprisingly by far the greatest in totalRNA (Figure 1b). The bimodal distribution of totalRNA gene expression broadly reflects a distinction between the cohort of mostly low-abundance non-coding RNAs and the higher-expressed protein-coding transcriptome. In our dataset, there are 19,881 protein-coding genes, of which 11,286 are expressed above 5 transcripts per million (TPM (Figure 1c). raRNA gene expression captures greater than 90% of these protein coding genes (Figure 1c), while depleting for lower abundance non-coding RNA biotypes, including linc RNAs and processed pseudogenes (Figure S1). We detect at least 5 ribosome footprints from ~67% of these protein-coding genes, with fractionated ‘discovery’ proteomics identifying 2 or more peptides from 39.5% (Figure 1c).

According to current attempts to annotate the human genome, the average coding gene has the potential to express ~4 distinct mRNA transcripts, with complex genes potentially generating more than 1058. However, several recent observations suggest that the majority of human cell-types and tissues tend to predominantly express a single ‘principal’ RNA transcript59,60. Identification of principal isoforms can yield important biological insights because they dictate the sequence, structure, regulation, and function of the majority of the protein produced by the gene. Each modality of \*omic data suffers from different biases and confounds with regards to isoform identification. By integrating multiple modes of such data we can attempt to overcome some of these effects to provide confident principal isoform identification.

Intronic reads are one of the largest confounds to transcript quantification by RNA-seq. (Figure S2a). A likely source of these reads is from pre-spliced transcripts in the nucleus. We find that RNA-seq reads derived from raRNA indeed contain far fewer intronic reads than totalRNA (Figure S2b). We hypothesised that poly-A purification may bring a similar benefit over total RNA, but inspection of data from an ENCODE K562 cell-line (www.encodeproject.org) shows no such reduction in intronic reads (Figure S2c). Notably, the ‘cleaner’ exonic signal from raRNA data leads to more consistent transcript quantification across the three biological replicates. If agreement is defined as identification of the same principal isoform in all three replicate samples, raRNA provides an improvement in the proportion of genes that agree on the major isoform compared to totalRNA (Figure S3a,b) This agreement is clearly dependent on both the expression of the gene and on the magnitude of the dominance of the principal isoform (Figure S3c,d). For genes expressed above 5 TPM and with a principal isoform that accounts for more than 50% of the mRNA produced by the gene, principal isoform agreement increases to 97% for raRNA and 93% for totalRNA (Green lines, Figure S3e,f). While conservative in terms of excluding many non-coding genes, the 5 TPM threshold in fact includes 96% of protein coding genes for which we observe footprint reads and/or peptides (Figure 1c).

In the case of ribosome footprints, despite their short fragment size, they contain other useful information that can be leveraged when assigning them to isoforms. The “perfect” cycloheximide frozen ribosome footprint, (consisting of only those nucleotides directly physically protected from the RNase enzyme by the ribosome) is a 28 nucleotide fragment, with the read midpoint situated between nucleotide 1 and 2 of the nearest codon (zero offset, Figure S4a). This consistent fragmentation pattern allows identification of codons from ribosome footprints, and as a result prediction of the open reading frame61 (ORF). Due to variations in the ribosome profiling technique such as incomplete RNase digestion, a range of fragment lengths are obtained. Rather than discard these “imperfect reads,” reads from genes with a single ORF can be used to create a position-weight-matrix (PWM) of read-midpoints to codon-offsets (Figure S4b). The PWM shows the consistency of ORF prediction within each biological replicate, and can be used as a quality control metric to highlight samples with inconsistent RNase digestion. In our hands, consistently digested samples produce a majority of 29 nucleotide length fragments, which correctly predict the transcript frame 90% of the time (Figure S4b). Incomplete RNase digestion usually results in a 3’ base overhang (Figure S4a). As a result, 3 footprints per transcript is sufficient to call the correct frame 75% of the time, rising to greater than 90% accuracy with at least 10 footprint reads (Figure S4c).

### A Bayesian approach to isoform-level integration of RNA-seq, ribosome footprints, and MS peptides

Computational tools for RNA-seq transcript quantification, such as the eXpress algorithm used to assess isoform consistency in our data62, typically employ an expectation maximization (EM) approach to determine the optimal abundances of each transcript so as to best explain the set of observed sequence reads. It is possible to employ a similar approach to quantifying isoforms based either on ribosome footprint reads or on LC-MS/MS peptides12. Unfortunately, compared to RNA-seq, ribosome footprint and peptide data are much more limited in their capacity to identify specific isoforms due to their smaller size, lower yield, and confinement to the coding sequence (CDS) of the gene. Given that isoform discrimination relies largely on those reads or peptides that span one or more exon-exon boundaries, identifying the correct isoform from the footprints or peptides alone is problematic. For protein analysis, a random 13 amino-acid peptide has an average probability of just 30% to cross an exon-exon boundary (Figure S5). Despite their increased number, ribosome footprints fare even worse due to their smaller size leading to an average probability of just 23% to cross an exon:exon junction. RNA-seq reads, however, are much longer (especially with paired-end data) and, as such, have a much higher probability (on average 85%) of spanning at least one junction. It is therefore strongly advantageous to inform an EM model by using totalRNA or raRNA transcript quantifications to set biologically informative priors.

In order to exploit this increased ability of RNA-seq to quantify the set of expressed transcripts in a sample, we therefore designed an EM algorithm to take RNA-seq as input to form biologically informative priors, that would be used to assign short length FPs and peptides to isoforms (Figure 2). Unlike a naïve prior (which initially assumes a equal likelihood of any isoform) the RNA-seq derived priors are able to overcome a large amount of the ambiguity in the set of transcripts/isoforms likely responsible for the observed footprints and/or peptides. The choice of biological prior used in the EM algorithm is extremely flexible and depends largely on the available data for a particular experiment. For example, a study with RNA-seq and MS proteomic data may use the RNA-seq as the prior, as would a study involving RNA-seq and ribosome footprinting (Figure S6a). For more elaborate experiments such carried out in the current study, where RNA, ribosome footprint, and MS proteomic data are available, the choice of prior for the proteomic EM could also be the isoform likelihoods output by the EM on the footprint reads (Figure S6b). We have implemented this EM algorithm in a novel software tool, **EMpire** (**EM** **P**ropagation of **I**soform abundance from **R**NA **E**xpression; URL). The tool can be run on any combination of RNA-seq, ribosome footprint, and mass-spec proteomics data (Methods and Figure S6) or a naïve prior, although use of RNA-seq to set isoform priors greatly improves the ability of the algorithm to assign footprints and peptides to specific isoforms (see below). When ribosome footprint data is input to the EM algorithm, an in frame footprint is weighted higher than an out of frame footprint, according to the values in the sample-specific PWM. By default, EMpire restricts footprints/peptide alignments to annotated coding sequences (CDS), however the tool can also include untranslated region (UTR) reads/peptides for characterization of upstream open reading frames (uORFs) or novel translated peptides.

As an example, assignment of ribosome footprints to the transcripts of POLDIP3 is vastly improved using the RNA-seq prior compared to the naïve prior. Use of the RNA-seq prior overwhelmingly suggests the presence of a two-fold abundant dominant transcript, POLDIP3-001, with a minor transcript of POLDIP3-002, an outcome which is fully consistent with both footprint read locations (Figure 2,S7a), peptide data (Figure S7b), and PCR data for this gene (see below and Figure 4b).

The POLDIP3 gene is representative of the majority of multi-isoform genes, most of which have an ambiguous principal isoform following EM using a naive prior (Figure S8a). In 58% of genes neither footprints nor peptides alone can distinguish a single principal isoform, instead settling on two or more equally likely isoforms. In genes for which the naive EM does converge on a single principal isoform, this isoform is typically extremely dominant and at least 5-fold more likely than the ‘next-best’ isoform (Figure S8b,c). Use of a biological prior (RNA-seq for footprints; RNA-seq or RNA-seq+footprints for MS proteomics) typically resolves isoform ambiguity for over 80% of genes, compared to only 42% able to be resolved using a naïve prior (Figure S8d-g). Furthermore, we can define clusters of genes that behave similarly in terms of the ability to resolve a principal isoform with different priors. Using an unsupervised hierarchical clustering and dynamic tree-cut we define 8 clusters of genes in each of the footprint EM and the proteomic EM that behave similarly based on the prior (Figure S9 and Figure 3a-b). These clusters can be further generalized into genes for which the biological prior is necessary or beneficial for principal isoform identification (footprint EM: 57.1% of genes, proteomic EM: 54.7%), whether it has no effect compared to the naive prior (footprint EM: 41.3% of genes, proteomic EM: 33.7%), or is inconsistent, picking a different principal isoform with different biological priors (footprint EM: 1.6% of genes, proteomic EM: 11.6%).

The use of EMpire provides two means by which isoform selection can improve; the use of relevant prior, as described above, and the expectation maximization process using the footprint or proteomic data (Figure 3, gradient of arrows). There are clusters in both the footprint (Figure 3a: 3, 4, 6 & 8) and proteomic (1 & 8) data for which the principal isoform achieves dominance as a result of the EM on the footprint and proteomic data, and not solely as a result of using relevant priors. There is therefore the potential for ribosome footprinting and/or proteomics to assist RNA-seq with isoform quantification, or at least principal isoform identification.

We selected a variety of genes with isoforms containing a single skipped exon, for which the naive EM was unable to identify a principal isoform but the biological priors appeared to resolve this ambiguity (Figure 4a). We designed PCR primers to amplify the region containing the prospective skipped exon, resulting in products of defined sizes dependent on the presence or absence of the exon. The example highlighted in Figure 2, POLDIP3, shows evidence of a principal transcript at 563 bp, the product size for transcript 001, with a minor product at 476 bp (transcript 002). For the remaining four cases (ALDH2, PDHB, COPE, MOGS), there is only evidence for the proposed dominant transcript, which is consistent with the RNA-seq (Figure 4b).

Finally, in order to assess the performance of the EMpire software in a more complex proteomics dataset, we applied the RNA-seq prior to a sample set from our recent study of human brain regions. We analyzed 5197 proteins detected by single-shot LC/MS-MS from the dorsolateral prefrontal cortex (dlPFC) of 5 adult humans, using publicly available mRNA-seq data from the same powdered samples as a relevant prior (www.brainspan.org/). In these complex brain samples, using the RNA-seq prior both increases the number of genes where a consistent principal isoform is called across all 5 samples (Figure 5a, blue bars), and allows for selection of a two-fold dominant principal isoform in an extra 40.6% of proteins (Figure 5b). In 29.8% of proteins the peptide data alone were sufficient for selecting a principal isoform. After EM with an mRNA prior it was not possible to call a two-fold dominant principal isoform in 29.6% of proteins. Unsurprisingly, as we saw with HEK293 cells, the more dominant a principal isoform is, the more consistently it is called in all five samples (Figure 5b, blue bars).

# Discussion

The eukaryotic genome is capable of producing an enormous repertoire of mRNA products. Isoforms are mRNA variants arising from the same gene that may differ in their transcription start site, exon usage, and untranslated regions26. Each of these variant features may be subject to different regulatory mechanisms, and result in variable function of the final protein product. It is therefore critical to our understanding of gene expression to consider abundance of isoforms, and not simply genes, in high throughput data. The presence of UTRs in mRNA, and the relatively long reads used in paired end mRNA-seq allows for increasingly reliable definition of isoforms in these data. While ribosome footprints may be found in UTRs, peptides are not produced from these regions, and the shorter sequences output from both techniques are much less likely to cross isoform defining exon boundaries. The EMpire tool leverages the improved isoform-level information acquired from mRNA-seq27 experiments to allocate ribosome footprints and peptides to isoforms using a process of Expectation Maximisation.

In our proof of concept study using HEK293 cells, almost 60% of genes have no clear principal isoform following naïve EM. By using an mRNA-seq prior, we can confidently assign a two-fold dominant isoform to ~80% of genes detected by ribosome footprints and peptides. The addition of footprint data to the mRNA-seq prior further increases the fraction of proteins with a principal isoform at all levels of dominance (2, 10 and 100-fold). For ~55% of these genes, the biological prior is necessary for definition of the principal isoform. For the vast majority of genes detected by ribosome footprints, the different modalities consistently select the same isoform. Approximately one tenth of genes detected by proteomics harbor a disagreement between the RNA-seq and peptide data. This is likely a function of the coverage of the proteomic data, and the observability of individual peptides. Where only a small number of peptides are detected from a single protein-coding gene, a peptide that disagrees with the RNA prior has significant power to change the outcome of the EM. While the ability of EMpire to allow the modalities to disagree is a critically important feature that may highlight interesting forms of downstream isoform regulation, in this case it is a likely that a substantial amount of the disagreement from proteomic data arises from low protein coverage, and will decrease as the sensitivity of these tools increases.

As cell cultures are relatively homogenous model systems, we explored the performance of the EMpire algorithm on proteomic data from a more complex system. Human dlPFC is an extremely complex tissue, with a layered cytoarchitecture and the presence of a large number of individual cell types. We hypothesized that it may be more difficult for the algorithm to define isoforms in this data, as different cell types may contribute different isoforms of the same proteins. However, the use of mRNA-seq as a prior increased the number of genes with a 2-fold dominant isoform from 30% (naïve prior) to 70%. It is likely that a substantial portion of the 30% of genes remaining uncertain can be accounted for by the contributions of isoforms from varying cell types. While mRNA-seq in individual cells is now a rapidly expanding reality, techniques to assess the diversity of cell types on the proteomic level are still in their infancy. We expect that as this technology improves, EMpire will be a useful tool for investigating the isoform diversity of cell types, as well as tissues. It is also interesting that the mRNA-seq data we used for this study was relatively old, comprising 50 nucleotide single end reads. Despite the presumed limitations of these shorter single end reads in defining isoforms, use of this mRNA-seq still provided a clear benefit for isoform selection from the proteomic data. As we saw in the HEK cell experiment, the principal isoform selection was consistent, with the same principal isoform selected in all five biological replicates in 65% of genes. As expected, there is a clear relationship where the more dominant the principal isoform is, the more likely it is consistently selected.

In addition to this improvement in isoform identification, EMpire maps all features back to the original genomic co-ordinates. Currently, it is surprisingly difficult to combine mRNA-seq data with protein data post-hoc. This is partly a result of using non-comparable reference databases in the mRNA-seq versus proteomic data. This feature of EMpire therefore provides stability for comparisons through multiple versions of reference databases, and makes comparison between different modalities easier. With small adaptations, it will be possible to use this pipeline for other high throughput data formats such as HTS-CLIP, methylation and ChIP-seq data. EMpire may also prove useful in model systems with less complete reference annotation than human and mouse, using RNA-seq based technologies to define isoforms identified by proteomics. In these systems using ribosome footprinting may be particularly helpful, as the weightings derived from their frame prediction capabilities can be used to define open reading frames. Integration of these data modalities from the beginning of the analysis pipeline will increase the functional salience of these data, minimize artifacts arising from poor comparability of reference databases, and enable us to more fully understand the relationship between mRNA, translation and protein.

# Methods

### HEK293 cell culture – Generation of a stable cell line expressing eGFP-L10a

HEK293 cells were transiently transfected with pCMV-EGFP-L10a using Effectene transfection reagent (Qiagen) according to the manufacturer’s protocols.  Stably expressing colonies were selected by growth in media containing G418.  The pCMV-EGFP-L10a contains the mouse L10a coding sequence, which diverges from the human coding sequence at 71 out of 653 bases, despite ultimately producing the same 100% conserved protein product.  This enabled us to assess the ratio of exogenous GFP-L10a to endogenous L10a from our RNA-seq data, which was approximately 1:5 (data not shown) in the cell line (HEK293-L10a) used to produce all data.

### Obtaining ribosome-associated RNA (raRNA) by eGFP-L10a immunoprecipitation

raRNAs were obtained according to a modified version of the original bacTRAP immunoprecipitation (IP) protocol63.  HEK293-L10a cells were grown to 80% confluency and lysed by rotor homogenisation in bacTRAP lysis buffer (20 mM HEPES, 5 mM MgCl2, 150 mM KCl, 0.5 mM dithiothreitol, 100 μg/ml cycloheximide, protease inhibitors, and recombinant RNase inhibitors) plus 2% n-dodecyl-beta-maltoside (n-dodec, Thermo Fisher Scientific).  Addition of 2% n-dodec ensures capture of ribosome footprints from both cytosolic and endoplasmic reticulum (ER) associated ribosomes, which in the latter case are otherwise depleted in these preparations64 (data not shown).

Lysates were cleared by centrifugation at 13,400 x g for 10 min at 4oC, then subjected to IP.  For EGFP-L10a IP, 450 μL of BSA blocked MyOne Streptavidin T1 Dynabeads (Thermo Fisher Scientific) were coated with 180 μL biotinylated Protein L, and pre-conjugated to a combination of 75 μg each of the mouse monoclonal antibodies 19F7 and 19C8 (Sloan Kettering Memorial Hospital).  Dynabead-antibody complexes were added to the cell lysate and immunoprecipitated overnight. The next day beads were washed 4 times with a high salt wash buffer (10 mM HEPES [pH 7.4], 350 mM KCl, 5 mM MgCl2, 1% NP-40, 0.5 mM dithiothreitol, 100 μg/ml cycloheximide).  Bound mRNAs were eluted by resuspending the beads into 700 μL Qiazol and following the manufacturer’s instructions for RNA purification using the miRNeasy kit (Qiagen). Full length total RNA was also prepared by lysing a pellet of HEK293-L10a cells in 700 μL of Qiazol, and using the miRNeasy kit.

### Ribosome profiling sample preparation

Ribosome footprints were prepared as described9, with some modifications.  Briefly, 80% confluent HEK293- L10a cells were lysed as above in bacTRAP lysis buffer plus 2% n-dodec.  Ribosomes were collected by immunoprecipitation (IP) with 20 μg of biotin conjugated eGFP monoclonal antibody (Sloan Kettering, as above) complexed with 160 μL Streptavidin Dynabeads.  Bead-associated ribosomes were resuspended in 300 μL bacTRAP lysis buffer without RNase inhibitors, and treated with RNaseI as for cell lysates.  Digestion was stopped by addition of 10 μL of Superasin (Ambion).  Ribosomes were collected by reattachment to the magnet and resuspended in 700 μL Qiazol, and processed as per Qiagen miRNeasy kit instructions.  Ribosome footprints were eluted from RNeasy columns in 30 μL RNase free water, then extracted overnight at -80oC following addition of a further 38.5 μL RNase free water, 1.5 μL GlycoBlue (ThermoFisher), 10 μM sodium acetate, and finally 150 μL isopropanol.  Footprints were collected by centrifuging at maximum speed on a desktop centrifuge for 30 min at 4oC.  Pelleted RNA was air-dried, then run on a 15% TBE Urea Gel (ThermoFisher).  A band was cut containing nucleotides of 26-32 nt size.  Overnight RNA extraction from the gel pieces, followed by T4 Polynucleotide Kinase (New England Biolabs Inc) treatment of fragments was performed as described9.

### RNA-seq library preparation and rRNA depletion

Full length RNA from total cells and raRNA underwent rRNA removal by RiboZero kit (EpiCentre, Illumina), to remove the ~90% of cellular RNA they represent.   rRNA depleted RNA was prepared for sequencing according to TruSeq library preparation protocols (Illumina), using random primers to synthesize cDNA.  Libraries were run on an Illumina HiSeq 2500 at the Yale Center for Genome Analysis, and paired end 75 nucleotide reads obtained.

Following T4 PNK treatment, ribosome footprints were prepared for sequencing using the NEBNext Small RNA Library Prep kit and the manufacturer’s instructions. This resulted in the use of a single gel extraction step, unlike previous protocols9.  After testing various rRNA depletion protocols, we made the decision not to remove rRNA , simplifying the workflow and decreasing the opportunity for investigator introduced variability or end bias.

### Mass-spectrometry (MS) proteomics

Frozen pellets of HEK293-L10a cells were lysed by sonication in RIPA buffer plus protease inhibitors.  Protein was precipitated from the lysate to remove detergents by chloroform/methanol precipitation.  Protein pellets were resuspended in 8 M urea, 0.4 M ammonium bicarbonate, reduced for 30 min at 37oC with 4 mM dithiothreitol, alkylated by incubating for 30 min with 8 mM iodoacetamide, before dilution to 2 M urea and addition of trypsin at a ratio of 1 μg:20 μg  total protein.  Samples were digested overnight at 37oC, then acidified and desalted on a C18 Macro Spin Column (Nest Group).  Peptides were eluted in 80% acetonitrile/0.1% trifluoroacetic acid (TFA), then dried by Speedvac.  The dried pellet was resuspended in 10 mM potassium phosphate in 25% acetonitrile solution (pH 3.0) and separated in the first dimension by Strong Cation Exchange on a 2.1 x 200 mm PolySULFOETHYL ATM column (PolyLC Inc.) via an HP 1090 HPLC.  Separation was carried out over a linear 118 min gradient with increasing Buffer B (10 mM potassium phosphate, 25% acetonitrile pH 3.0, 1 M potassium chloride) at a flow rate of 200 μL/min.  Twenty fractions were collected, pooled into 10 tubes, and each tube desalted using a Microspin C18 column (The Nest Group) prior to LC MS/MS.  Analysis was performed using an LTQ Orbitrap Elite equipped with a Waters NanoACQUITY ultra-performance liquid chromatography (UPLC) system using a Waters Symmetry C18 180 µm by 20 mm trap and a 1.7 µm (75 µm-inner-diameter by 250 mm) NanoACQUITY UPLC column (at 35°C) for peptide separation.

### RNA-seq read alignment and transcript quantification

Due to our IP of ribosomes, and the decision not to deplete footprint samples of rRNA, we carried out an explicit alignment of the reads to known human rRNA before alignment to the genome.

**TotalRNA- and raRNA-seq reads** were mapped to the annotated 5S and 45S (chrUn\_gl000220) rRNAs using STAR to remove any remaining rRNA contamination in our samples. Based on this alignment we observed that residual rRNA could explain on average ~20% of the sequence reads across all totalRNA and raRNA samples. We mapped the remaining ~80% of the RNA-seq reads to the human genome (hg38) and annotated transcriptome (gencode v21) again using the STAR aligner, following roughly the ENCODE alignment parameters (github.com/ENCODE-DCC/long-rna-seq-pipeline/blob/master/DAC/STAR\_RSEM.sh).  Of the non-default options in STAR, the following are the most important to ensure compatibility with EMpire:

‘--outSAMtype BAM SortedByCoordinate’ for visualisation in IGV

‘--quantMode TranscriptomeSAM’ for alignments in transcriptome coordinates for eXpress

‘--outFilterMismatchNoverLmax  0.05’ to ensure # mismatches to <5% of the # of mapped bases

Transcript aligned reads were quantified using eXpress.

**Ribosome footprint reads** were clipped of their 3’ adapter and aligned, like the totalRNA- and raRNA-seq samples above, to the annotated 5S and 45S (chrUn\_gl000220) ribosomal RNAs. Removal of rRNA reads was very important to reduce the effect of spurious alignments to the genome. Non-rRNA reads were aligned to the human genome (hg38) and annotated transcriptome (gencode v21) again using the STAR aligner.

### Mass-spectrometry spectra alignment

The entire human transcriptome (as defined in gencode21) was in-silico translated, in three frames, to amino acid sequences using the transseq function within the EMBOSS65 software library.  Also included in this ‘target’ database were CRAPome66 sequences of likely contaminants (such as Bovine Albumen).

Spectra obtained from the Orbitrap Elite, in proprietary ThermoFisher ‘.raw’ files, were processed using MaxQuant67 (v1.5.2.1). The standard peptides.txt output file from MaxQuant is used as input to EMpire. We also include support for X!Tandem68 following conversion to .mzXML by MSConvert69. Spectra were searched against these transcriptome-derived protein sequences common contaminant sequences, and a library of reverse ‘decoy’ sequences68.

The distribution of expectation values for spectra with legitimate database hits was compared to the equivalent distribution for spectra assigned to the reverse ‘decoy’ database and a [maximum] expectation threshold was selected for each sample that limited the false discovery rate (FDR) to 1%.  I.e. 1% of spectra below this expectation value mapped to the reverse database while 99% mapped to the real database.

Peptides that mapped to more than one distinct genomic locus or to contaminant sequences were discarded from further analysis, however this is not to be confused with peptides assigned to multiple potential isoforms of the same gene which were retained. It is worth noting that many of the peptides identified using the SWISS-Prot reference could be assigned to multiple distinct genes and thus constituted the majority of the data loss attributed with the use of this reference (data not shown). The transcriptome-derived reference did not suffer as much from multi-mapping, due mainly to less ambiguous gene-isoform relationships.

The EMpire software described below performs a second round of FDR correction based on decoy mappings and discards multi-locus mapping sequences.

### EMpire: software implementation and testing

The EMpire software is freely available as both source code LINK WILL BE ADDED.

required inputs to the software are all in common data formats, examples are available in the GitHub repo:

**footprinting:**

gencode/ENSEMBL genome annotation (.gtf)

footprint read -> transcript alignments (.bam)

[optional] eXpress RNA-seq transcript quants (.xprs)

**mass-spec:**

gencode/ENSEMBL genome annotation (.gtf)

translated amino-acid sequences for each transcript (.fasta)

spectra -> transcript X!Tandem output (.xml)

[optional] spectra MS1 intensities for quantification (.mzXML)

[optional] eXpress RNA-seq transcript quants (.xprs) [OR] EMpire footprint EM output (.exprs)

The memory requirements of the EM algorithm and the frame analysis, which are described below, generally increase with increasing number of footprint reads or spectra. For all proteomics data and for most footprint experiments (with fewer than ~10million reads), EMpire will run on a modest modern machine with >8GB of memory. For larger footprinting datasets the required memory scales with the number of input reads, but even the deepest experiments should run with 20-30GB memory.

### EMpire: expectation maximization algorithm

Footprint reads or MS peptides are defined in terms of their binary compatibility to the set of isoforms of a given gene. If a read or peptide has a valid alignment to a transcript it is given a value of 1, or else it is assigned 0. Using this compatibility matrix, ***I***, of all reads/peptides against all transcripts as well as the relative transcript abundances from RNA-seq we can write down the likelihood, , of observing all ***1..N*** footprint reads, ***R***, given the distribution of the abundances,, of ***1..K*** isoforms:



In the simplest case of the naïve prior, we have no information about which isoform(s) may be responsible for generating the footprint reads and/or peptides as so we define the initial distribution of isoform abundances as uniform:

*k = K-1*

In the case of a non-uniform prior (i.e. from the RNA-seq expression data),  is set as the ratios of isoform expressions to the total (cumulative) expression of all transcripts in their parent gene.

Due to the fact that at the same expression isoforms with longer open reading frames will produce more footprint reads (and below saturation longer proteins will produce more identifiable peptides) we also define the probability that the ***j***th isoform will contribute a footprint or peptide based on the length of its coding sequence, ***l***, and its abundance:



For the peptide EM, the effective isoform length is calculated from an in-silico digestion of the protein to the constituent peptides. Currently Trypsin and LysC digestions are supported but cleavage rules can be added for more enzymes quite straightforwardly. The user has the option to specify any number of missed-cleavage events to match the parameters used for the spectral search.

For each gene we can update the isoform abundances (by MLE) from the original RNA-seq to new values that best explain the observed footprints. Sampling from these new isoform abundances allows us to assign footprint reads to specific isoforms and so on.

For the ribosome footprints we further modify the compatibility matrix, ***I***, to reflect the likelihood that a read of this length would be observed with its offset from the coding frame of the transcript. The calculation of this read-position weight matrix is described in the next section. Essentially if 90% of the reads of this read’s length have an observed frame offset of 0.5nt and this read has the same offset to the current transcript then the compatibility is set to 0.9. This allows for the down-weighting of reads that have a spurious frame-offset to the current transcript; in this example if only 2% of all reads of the same length have a frame offset of 1.5nt, which matches that of the current alignment to the current transcript, then the compatibility is set to just 0.02 and this read will have very little positive support for this isoform.

### EMpire: ribosome footprint frame analysis

Here, we calculated the frame using the offset of the mid-point of the footprint read to the start of the middle-nucleotide of the closest codon triplet. Using this metric and the resulting position-weight-matrix (PWM) of the footprint size vs. codon offset we can do two things; first we can infer that the result of incomplete RNase digestion, which will likely differ between footprint preps, tends to leave additional nucleotides at the 3’ end of the footprint (Figure S4a). Second, we can use the PWM of read-mids to codon-offsets from single isoform genes to allow the reads to decide for themselves the optimal translation frame for each coding sequence and then ask, as a function of the number of reads mapped to a transcript, what fraction of transcripts are called in the correct frame (see Methods).

### Other bioinformatic & statistical analysis

All RNA-seq statistical analyses, pre-processing, and normalization was been performed within the R/Bioconductor scripting environment70. Gene clustering for Figure 3/Figure S9 was performed using the dynamicTreeCut package, with default parameters except for the minimum module sizes which were adjusted for aesthetics (250 genes for footprints, and 200 genes for peptides). Cluster profiles were computed from the median major isoform fractions of the genes within each cluster.

**PCR confirmation of principal isoforms from RNA-seq**

Total-RNA was extracted from a HEK-L10a cell pellet as described previously. cDNA was synthesized using SuperScriptIII Reverse Transcriptase (ThermoFisher Scientific) according to Manufacturer’s instructions. PCR primers for selected genes where isoforms were defined by a skipped exon were designed according to the scheme below.

Primers for individual genes were:

POLDIP3: AAGTGCAGGATGCCAGAGAG Fw

 CAATGGGCTGAGAACAGGCT Rv

ALDH2: CCGAGGTCTTCTGCAACCAG Fw

 TTGCATCAGGAGCGGGAAAT Rv

PDHB: CTGGCTTGGTGCGGAGAC Fw

 CCAGCAAAGCCCATCTCTGA Rv

COPE: AGAGAGACGTGGAGAGGGAC Fw

 CCACTATCCTTGTCTAGCGCC Rv

MOGS: CAGGTGTCGCTAACCGGAC Fw

 CGGGTCTTCATGCCGAAGTA Rv

Standard PCR was performed using Taq DNA polymerase (Invitrogen, Thermo Fisher Scientific). cDNA was diluted 1:10 , using 1 μL per 20 μL reaction. 30 PCR cycles were performed, before running the samples on a 1.5% agarose gel.

# References

1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

2. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

3. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science (80-. ).* **348,** (2015).

4. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

5. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

6. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30,** 224–226 (2012).

7. Mann, M., Kulak, N. A., Nagaraj, N. & Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* **49,** 583–90 (2013).

8. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.* **24,** 11–17 (2015).

9. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7,** 1534–50 (2012).

10. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15,** 205–13 (2014).

11. Piccirillo, C. A., Bjur, E., Topisirovic, I., Sonenberg, N. & Larsson, O. Translational control of immune responses: from transcripts to translatomes. *Nat. Immunol.* **15,** 503–11 (2014).

12. Floor, S. N. *et al.* Tunable protein synthesis by transcript isoforms in human cells. *Elife* **5,** 1276–1280 (2016).

13. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).

14. Kitchen, R. R., Rozowsky, J. S., Gerstein, M. B. & Nairn, A. C. Decoding neuroproteomics: integrating the genome, translatome and functional anatomy. *Nat. Neurosci.* **17,** 1491–1499 (2014).

15. Ebrahim, A. *et al.* Multi-omic data integration enables discovery of hidden biological regularities. *Nat. Commun.* **7,** 13091 (2016).

16. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–23 (2009).

17. Brar, G. A. *et al.* High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. *Science (80-. ).* **335,** (2012).

18. Mohammad, F., Woolstenhulme, C. J., Green, R. & Buskirk, A. R. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep.* **14,** 686–94 (2016).

19. Woolstenhulme, C. J. *et al.* High-Precision Analysis of Translational Pausing by Ribosome Profiling in Bacteria Lacking EFP. *Cell Rep.* **11,** 13–21 (2015).

20. Gene expression: Ribosome profiling in bacteria. *Nat. Methods* **9,** 126–126 (2012).

21. Tanenbaum, M. E., Stern-Ginossar, N., Weissman, J. S. & Vale, R. D. Regulation of mRNA translation during mitosis. *Elife* **4,** e07957 (2015).

22. Gonzalez, C. *et al.* Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J. Neurosci.* **34,** 10924–36 (2014).

23. Lee, M. T. *et al.* Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature* **503,** 360–364 (2013).

24. Janich, P., Arpat, A. B., Castelo-Szekely, V., Lopes, M. & Gatfield, D. Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames. *Genome Res.* **25,** 1848–59 (2015).

25. Hornstein, N. *et al.* Ligation-free ribosome profiling of cell type-specific translation in the brain. *Genome Biol.* **17,** 149 (2016).

26. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).

27. Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16,** 150 (2015).

28. Menschaert, G. & Fenyö, D. Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrom. Rev.* n/a-n/a (2015). doi:10.1002/mas.21483

29. Koch, A. *et al.* A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* **14,** 2688–98 (2014).

30. Xiao, Z., Zou, Q., Liu, Y., Yang, X. & Cowling, V. H. Genome-wide assessment of differential translations with ribosome profiling data. *Nat. Commun.* **7,** 11194 (2016).

31. Zhong, Y. *et al.* RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* **33,** 139–141 (2017).

32. Wang, H. *et al.* Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. *Bioinformatics* **32,** 1880–1882 (2016).

33. Mann, M. Origins of mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **17,** 678–678 (2016).

34. Sterner, J. L., Johnston, M. V., Nicol, G. R. & Ridge, D. P. Signal suppression in electrospray ionization Fourier transform mass spectrometry of multi-component samples. *J. Mass Spectrom.* **35,** 385–391 (2000).

35. Qian, W.-J. *et al.* Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. *Mol. Cell. Proteomics* **5,** 1727–44 (2006).

36. Meek, J. L. & Rossetti, Z. L. Factors affecting retention and resolution of peptides in high-performance liquid chromatography. *J. Chromatogr. A* **211,** 15–28 (1981).

37. Meek, J. L. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 1632–6 (1980).

38. Guo, D., Mant, C. T., Taneja, A. K., Parker, J. M. R. & Rodges, R. S. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides. *J. Chromatogr. A* **359,** 499–518 (1986).

39. Searle, B. C. Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **10,** 1265–1269 (2010).

40. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–58 (2003).

41. Zhang, J. *et al.* PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol. Cell. Proteomics* **11,** M111.010587-M111.010587 (2012).

42. Old, W. M. *et al.* Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Mol. Cell. Proteomics* **4,** 1487–1502 (2005).

43. Wenger, C. D. & Coon, J. J. A Proteomics Search Algorithm Specifically Designed for High-Resolution Tandem Mass Spectra. *J. Proteome Res.* **12,** 1377–1386 (2013).

44. Neil L. Kelleher *et al.* Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry. (1999). doi:10.1021/JA973655H

45. Siuti, N. & Kelleher, N. L. Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **4,** 817–821 (2007).

46. Simon, R. *et al.* Total ApoE and ApoE4 Isoform Assays in an Alzheimer’s Disease Case-control Study by Targeted Mass Spectrometry (n = 669): A Pilot Assay for Methionine-containing Proteotypic Peptides. *Mol. Cell. Proteomics* **11,** 1389–1403 (2012).

47. Picard, N., Ratanasavanh, D., Prémaud, A., Le Meur, Y. & Marquet, P. IDENTIFICATION OF THE UDP-GLUCURONOSYLTRANSFERASE ISOFORMS INVOLVED IN MYCOPHENOLIC ACID PHASE II METABOLISM. *Drug Metab. Dispos.* **33,** 139–146 (2004).

48. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **4,** 222 (2008).

49. Krastins, B. *et al.* Rapid development of sensitive, high-throughput, quantitative and highly selective mass spectrometric targeted immunoassays for clinically important proteins in human plasma and serum. *Clin. Biochem.* **46,** 399–410 (2013).

50. Peggion, C. *et al.* Absolute quantification of myosin heavy chain isoforms by selected reaction monitoring can underscore skeletal muscle changes in a mouse model of amyotrophic lateral sclerosis. *Anal. Bioanal. Chem.* **409,** 2143–2153 (2017).

51. Ghosh, S., González-Mariscal, I., Egan, J. M. & Moaddel, R. Targeted proteomics of cannabinoid receptor CB1 and the CB1b isoform. *J. Pharm. Biomed. Anal.* (2016). doi:10.1016/j.jpba.2016.11.003

52. Barthélemy, N. R. *et al.* Differential Mass Spectrometry Profiles of Tau Protein in the Cerebrospinal Fluid of Patients with Alzheimer’s Disease, Progressive Supranuclear Palsy, and Dementia with Lewy Bodies. *J. Alzheimer’s Dis.* **51,** 1033–1043 (2016).

53. Zickmann, F. & Renard, B. Y. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics* **31,** i106–i115 (2015).

54. Woo, S. *et al.* Proteogenomic Database Construction Driven from Large Scale RNA-seq Data. *J. Proteome Res.* **13,** 21–28 (2014).

55. Wang, X. & Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29,** 3235–3237 (2013).

56. Kumar, D. *et al.* Integrating transcriptome and proteome profiling: Strategies and applications. *Proteomics* **16,** 2533–2544 (2016).

57. Kumar, D., Yadav, A. K., Jia, X., Mulvenna, J. & Dash, D. Integrated Transcriptomic-Proteomic Analysis Using a Proteogenomic Workflow Refines Rat Genome Annotation. *Mol. Cell. Proteomics* **15,** 329–339 (2016).

58. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–15 (2010).

59. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8 (2012).

60. Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* **14,** R70 (2013).

61. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13,** 165–170 (2015).

62. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10,** 71–3 (2013).

63. Heiman, M., Kulicke, R., Fenster, R. J., Greengard, P. & Heintz, N. Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). *Nat. Protoc.* **9,** 1282–91 (2014).

64. Reid, D. W. & Nicchitta, C. V. Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. *J. Biol. Chem.* **287,** 5518–27 (2012).

65. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–7 (2000).

66. Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nat. Methods* **10,** 730–736 (2013).

67. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).

68. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467 (2004).

69. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24,** 2534–2536 (2008).

70. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).

# Figure Legends

**Figure 1 | Experimental approach to integrated analysis of the transcriptome, translatome, and proteome**

**a)** Schematic diagram of the experimental approach to multi-modal profiling of the translatome in HEK293-L10a cells. Total-RNA and protein were obtained from lysing whole cells, while ribosome-associated (raRNA) and ribosome footprint RNA were obtained following immunoprecipitation of intact ribosomes from detergent extracted post nuclear supernatant (capturing cytosolic and ER associated ribosomes).

**b)** RNA-seq of totalRNA (dark blue) captures 60,155 genes (#genes, y-axis) that vary widely in abundance (log10(TPM), x-axis) and biotype. Shown on these histograms are genes also observed when profiling raRNA (light blue, 27,977 genes), ribosome footprints (green, lower threshold of 5 ribosome footprints/gene, 11,230 genes), and protein (red, genes with at least 2 peptides, 3833). Genes with at least five ribosome footprints are generally expressed above ~1 transcript per million (TPM). Fractionated LC-MS/MS, where peptides are first separated into pools offline, then analyzed serially, was used to identify proteins.

**c)** Of the 60,155 total genes captured by total RNA, 19,881 are annotated as protein coding.

Ribosome footprints are identified from ~56% of these protein coding genes, and proteomics samples ~19%.

**Figure 2 | Analytical approach to integrated analysis of the transcriptome, translatome, and proteome**

**a)** Schematic diagram depicting the expectation maximisation (EM) process for the POLDIP3 gene using the 11 ribosome footprint reads that map to the coding sequence (CDS) of at least one of its 5 possible mRNA transcripts. The uninformative ‘naive’ prior (upper panel), in which each transcript is equally likely to generate these observed footprints, converges on three equally likely transcripts which the footprint reads cannot discriminate between. 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 reads/peptides. The use of a biologically informed prior, obtained directly from the relative transcript abundances from totalRNA-seq or raRNA-seq, overcomes ambiguity (lower panel) as the footprint reads are fully consistent with the transcript abundances for this gene. The biological prior supports POLDIP3-001 as the greater than two-fold dominant isoform, with POLDIP-002 as a minor secondary isoform.

**Figure 3 | A biologically realistic prior improves isoform level interpretation of ribosome footprints and MS/MS peptides**

**a)** Genes with at least three ribosome footprint reads cluster into 8 main groups (Fig S9a) based on the result of the EM. Each plot shows for each of the 8 groups the principal isoform fraction before (prior) and after the EM. The results based on the three available priors are illustrated by the three columns of plots (naive, left; totalRNA, centre; raRNA, right). Genes fall into three main groups, those that are aided by the use of a biologically informed RNA-seq prior (green, 3795 genes total), those for which the principal isoform is decided entirely by the footprints (blue, 2747 genes total), and those in which the principal isoform is unstable (red, 108 genes).

**b)** As **a)** following EM using peptides obtained from mass spectrometry (MS/MS). Peptides cluster into 8 main groups based on the result of the EM (Figure S9b) Here the priors are naive (left), raRNA (centre), and raRNA+footprints (right); where the latter is the isoform abundance output generated by the ribosome footprint EM using the raRNA prior - the right column in **a)**. Genes fall into three main groups, those that are aided by the use of the biologically informed RNA-seq priors (green, 663 genes total), those for which the principal isoform is decided entirely by the peptides (blue, 408 genes total), and those in which the principal isoform is unstable (red, 141 genes).

**Figure 4 | A RNA-seq prior improves isoform assignment in specific genes where naive EM ties**

**a)** Detailed illustration of the EM result for 5 selected genes showing differences in the relative isoform abundances of each. In all cases, the biological prior is necessary to resolve the principal isoform (red) and the second isoform (blue) where applicable. To the right, the isoform names are shown along with the expected product size for the PCR validation in **b)** (see also supplemental methods).

**b)** PCR analysis of the 5 genes selected in panel **a)** show that, at least at the mRNA level, all agree with the principal isoform inferred by RNA-seq. POLDIP3 (left) also shows evidence for the expression of the second isoform predicted by RNA-seq.

**Figure 5 | Consistent principal isoform identification in complex human brain samples using a RNA-seq prior**

**a)** Using an mRNA-seq prior (right) increases consistency of identification of a single dominant isoform compared to a naïve prior (left). Red bars indicate genes where the EM is unable to break a tie between equally likely isoforms, or where different principal isoforms are called in the 5 biological replicates **b)** Following a naïve prior (left), up to 70% of genes are ambiguous in terms of their principal isoform; peaks are evidence of the algorithm’s failure to break a tie between 2 (x=0.50), 3 (x=0.33) or 4 (x=0.25) equally likely isoforms per gene. Use of the mRNA prior (right) substantially reduces the number of genes with an ambiguous principal isoform, with ~70% of genes reporting a 2-fold or greater dominant isoform (x>0.66).

**Figure S1 | 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 are observed as raRNA. This fraction decreases for other non-coding RNA biotypes such as lincRNAs, where 65% of those observed in total RNA are detected as raRNA, and processed pseudogenes, where 53% of those observed in totalRNA are present in raRNA data

**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 signal of 0.75. A value of 1 indicates all reads derived from a selected gene are exonic.

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

**c)** Density plot comparing exonic signal 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 does not show anywhere near the same reduction in intronic signal compared to the raRNA capture in **a)**, likely due to the presence of nuclear polyadenylated pre-mRNA fragments.

**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) agree 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 agree; red: at least one replicate shows a different principal isoform. A lack of reliability in detecting the principal isoform may indicate either biological or technical variability between samples, as a result 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 increases 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 is that there will be a disagreement on the principal isoform between samples.

**e)** Heat maps show the effect of varying minimum thresholds of gene expression and principal isoform dominance on agreement between replicates. Greater consistency is 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 at least a 50% dominant principal isoform expressed at more than 5 transcripts per million are consistently defined as having a principal transcript.

**Figure S4 | Ribosome footprints can be used to 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 reality, due to the variations inherent in the technique such as incomplete RNase cleavage, a range of fragment lengths is obtained. In our hands, incomplete RNase digestion tends to leave uncleaved nucleotides at the 3’ end of the footprint (orange circles). **b)** Rather than discard these “imperfect” reads, it is 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 are in frame with each other following 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, see **methods**) increases with increasing number of observed footprints per transcript. With ≥10 reads it is possible to accurately predict the correct frame for >90% of transcripts.

**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, 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. As a result, 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 | The EMpire tool can be used for experiments with any combination of RNA-seq, ribosome footprinting, and MS/MS**

**a)** Isoform prediction and assignment for experiments using ribosome footprints can be performed without (top) or with (bottom) an RNA-seq informed biological prior. Here, RNA-seq transcript quantifications are produced by the eXpress tool62 and all footprint alignments are in transcriptome coordinates.

**b)** As **a)**, but for mass-spectrometry experiments with peptide alignments (produced by MaxQuant67 or X!Tandem68) in transcript coordinates. The top row is a simple peptide input with no prior and the middle is the RNA-seq informed biological prior for experiments with RNA-seq and MS/MS. The last row shows the tool in use as here where RNA-seq, footprint, and MS/MS proteomics are available from the same samples; in this situation the output from the EM in **a)** is input to the MS/MS EM and can also support the less common situation where no RNA-seq data are available for a given experiment.

**Figure S7 | 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. As shown in the schematic in **Figure 3**, 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 map to an intronic Alu repeat region (labeled), this is not a novel POLDIP exon.

**b)** Expanded EM results highlight 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.

**Figure S8 | Using an RNA-seq prior robustly decreases the ambiguity of footprint and peptide assignment to the principal isoform in multi-isoform genes**

**a)** Histogram showing the number of genes where the naive EM is unable to break a tie between multiple equally-likely (‘best’) isoforms. Single isoform genes are shown as zero on the x-axis. EM on footprint data (grey bars) with a naive prior is able to settle on a single isoform for over 2000 genes (shown at 1 on the x axis); the remaining 58.2% of multi-isoform genes are tied between 2 or more equally-likely isoforms. For MS/MS peptide data (blue bars), naive EM is ambiguous as to the likely principal isoform in 58.1% of multi isoform genes.

Re-plotted principal isoform ambiguity for naive EM on **b)** ribosome footprint data and **c)** MS/MS peptides (right) in terms of fold-dominance of the principal isoform over the second most abundant isoform. On these histograms, a value of 1 reflects a gene in which a single principal isoform cannot be determined. For clarity, principal isoforms more than 5-fold more abundant than the second isoform are capped to 5-fold dominance.

Using **d)** totalRNA- or **e)** raRNA-seq as a prior substantially improves our ability to resolve a dominant principal isoform from ribosome footprint EM.

Using **f)** RNA-seq and/or **g)** ribosome footprint data as a prior greatly improves ability to resolve a dominant principal isoform from MS/MS peptides.

**Figure S9 | Use of a biologically informative prior can dramatically improve 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 EMpire 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 3**.

**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 improves the ability to resolve the principal or isoform in 3,795 genes (57.1%) and converges on the same isoform as the naive prior in 2,747 genes (41.3%); in the remaining 108 genes (1.6%) the different RNA priors (totalRNA and raRNA) disagree on the major isoform.

**b)** The heatmap plots the principal isoform fraction for the 1,212 multi-isoform genes with at least 2 peptides. Using a biological prior improves our ability to resolve the principal isoform in 663 genes (54.7%) and converges on the same isoform as the naive prior in 408 genes (33.7%); in the remaining 141 genes (11.6%) the different RNA priors (raRNA and raRNA+footprints) disagree on the principal isoform.