miBAT: Multi-modal profiling of the translatome at isoform resolution

*Kitchen R.R., *Carlyle B,C., Zhang J., Overton J.D., McPeek J.D., Lam T.T., Rozowsky, J.S., Gerstein M.B., and Nairn A.C.

Abstract

TBD

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 every imaginable cellular system. Ambitious projects such as 1000 Genomes, The Cancer Genome Atlas (TCGA), and the Genotype-Tissue Expression project (GTEx) 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), the Roadmap Epigenome Project, and BrainSpan have attempted deeper characterisation of the multi-omic landscape of specific cell types, tissues, and species. Thanks 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. Biologically however, RNA is far from the end of the story.

Protein levels arguably most closely reflect the biosynthetic state of the cell but, compared to nucleic acids, are much more difficult to measure in a high-throughput assay thanks primarily to complex chemistry of amino acids and our current inability to amplify polypeptides. For this reason mass-spectrometry proteomics would greatly benefit from the type of deep analytical integration seen in studies of the transcriptome and the genome [REF any ASE/RDD paper] or epigenome [REF ENCODE-nets?]. The handful of studies claiming integrated analysis of the transcriptome generally refer to having obtained RNA and protein in parallel from the same samples, and comparing results post hoc. [24290761, etc, etc]. While it is undoubtedly powerful and important to perform multi-omic experiments on the same samples/tissues the potential insights from a fully integrated analysis are often lost.

Studies that include a direct comparison of mRNA expression and protein abundance typically report poor correlation, as post-transcriptional regulation often leads to differential synthesis and turnover of both RNA transcripts and proteins. Recently, studies of the translatome as a whole have started to bridge the gap between transcriptome and proteome. By measuring the dynamic profiles of ribosomes as they are translating mRNA to protein, as well as steady state mRNA

and protein levels, we get an additional and often crucial [REFs Ingolia, Piccirillo, etc] insight into expression regulation within a cellular system. The potential benefits of a principled integration of every level of the experiment (from sample extraction, to alignment, to quantification) of these highly complementary data modalities are very enticing.

To date the majority of high-throughput studies probing translational dynamics have been in 'lower' organisms such as bacteria and yeast, with a few recent studies moving via cell-culture to more complex eukaryotes. The main problem with studies of the mammalian translatome is the splicing complexity of the transcriptome; current estimates are that over 90% of human multi-exon protein coding genes transcribe alternatively spliced mRNAs [18978772]. However, several recent observations suggest that in fact the majority of human cell-types, tissues, or even organs tend to predominantly express a single isoform of each gene [22955620, 23815980]. Identification of these principal isoforms by RNA-sequencing can yield important functional insights, such as the absence or presence of protein-protein interaction (PPI) domains and the use of alternative splice and promoter sites that may be specific to an experimental condition. It is further possible that due to post-transcriptional regulation and the presence of non coding RNAs, the principle RNA transcript may not in fact be the dominant protein isoform, however most proteomic analysis is performed at the abstracted level of the gene and does not distinguish between individual isoforms. Here, multi-modal profiling of the various levels of the translatome is well positioned to resolve such isoform ambiguity with parallel measurements of mRNA, ribosome occupancy, and protein abundance.

In this study 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, **miBAT** (**m**ulti-modal **i**soform-level **B**ayesian **A**nalysis of the **T**ranslatome), 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 and that footprints obtained by immunoprecipitation of ribosomes provide significantly cleaner data compared to standard sucrose-cushion based purification.

Results

Multi-modal profiling of the transcriptome and translatome

Given the potential scope for post-transcriptional regulation in any given cellular system, we designed a series of assays to be run in parallel on the same sample (Figure 1a). For this proof of principle we used a human cell-line (HEK293), 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 through two methods of purification, the standard sucrose cushion and a novel immunoprecipitation based approach. Finally we obtained proteomic data in both 'discovery' and 'quantification' mode, where the former relies on heavily fractionated runs to be able to identify the maximum

number of peptides while the latter is a more standard proteomic prep focussing on providing quantitative measurements of a smaller number of peptides and proteins.

Each of these assays differs not only in their molecular target but also their sensitivity. The depth of coverage, in terms of genes detected at various levels of expression, is unsurprisingly by far the greatest in totalRNA and raRNA (Figure 1b), extending deep into the poorly expressed non-coding transcriptome. The sensitivity and specificity of raRNA to selecting protein coding genes is extremely high, with over 90% of the protein coding genes detected in totalRNA also reflected in the raRNA-seq data (Figure S1a-b). Ribosome footprints cover around two-thirds of the genes in the expression regimen typically reported as producing protein, in our dataset this constitutes the 11,268 protein coding genes expressed above 5 transcripts per million (TPM). Fractionated 'discovery' proteomics hits 39.5% of these genes, while standard quantitative proteomics represents only 10.2% of the most highly abundant genes (Figure 1b and Figure S1a).

IP is superior to sucrose cushion for collecting ribosome footprints

We observed a marked difference in data quality resulting from the two methods of purifying ribosomes for footprinting. Compared to the footprints obtained by sucrose cushion, the footprints obtained by immunoprecipitation (IP) contained a larger fraction of ribosomal RNA (rRNA) which was carefully removed (see Methods), however the remaining reads were of substantially higher quality and purity. Notably, 74% of the IP footprint reads were in the expected 28-32nt size range (compared to an average of 51% from the cushion) and 79% of those 28-32nt reads could be uniquely mapped to the genome (compared to an average of just 48% from the cushion; Figure S1c). Finally, of the uniquely mapped reads the vast majority (97%) of IP footprints mapped to the coding sequence (CDS) compared to 75% of the sucrose cushion footprints; even after explicitly removing nuisance small-RNA genes this fraction remained at 86% for the cushion footprints (Figure S1d). This last point especially, reflecting a relatively high level of nuisance contamination in footprint samples obtained by sucrose cushion, may have resulted in the recent burst of academic activity with several papers claiming [REFs], then refuting [REFs], widespread ribosome occupancy and potential translation of non-coding transcripts.

The IP sample also substantially outperformed the cushion samples in terms of frame fidelity, in which multiple footprint reads aligned to the same mRNA agree on the same frame of translation. Of the IP footprint sizes, 81% of the 28nt and 90% of the 29nt reads were in-frame with each other, compared to just 45% and 57%, respectively, for the cushion reads (Figure S2a-b). 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 S2c). Second, we can use the PWM of read-mids to codon-offsets 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). Resulting from this, we observe that again the IP footprints substantially outperform the sucrose cushion in that a transcript with just three footprints is called in the correct frame 75% of the time, rising to over 90% accuracy in transcripts with at least 10 footprint reads (Figure S2e). Equivalent values for the sucrose cushion footprints show that 3 footprints can call the correct frame in only 55% of transcripts and 10 footprints only manage 65% accuracy (Figure S2d).

Challenges and opportunities in quantifying transcripts and isoforms

One of the major challenges of these frame analyses, and analyses of ribosome footprints and mass-spec peptides in general, in complex eukaryotic systems is the issue of alternate isoforms of a given gene. For the analysis above, as is commonly the metric used to select the likely isoform in footprint analyses [REFs], the mRNA transcript selected for each gene was that with the largest number of footprints. One can easily envision a situation where a gene was producing multiple mRNA transcripts at different abundances, which would lead to incorrect frame calculations for genes with isoforms resulting from frame shifting. Fortunately, several recent publications have found that the majority of human genes tend to predominantly express a single isoform in a given cell-type, tissue, or even organ [REF ENCODE + gencode folks]. The major challenge in studies of the translatome, i.e. ribosome footprinting and mass-spec proteomics, is reliably determining, for each gene, what this major isoform actually is.

According to current attempts to annotate the human genome, the average protein coding gene has been observed to express 4 distinct mRNA transcripts, with more complex genes containing between 10 and 61 mRNAs. Discrimination of these different splice variants relies largely on observations (be they peptides, footprints, or RNA-seq reads) that span one or more exon-exon boundaries. Unfortunately for mass-spec proteomics, identifying the correct isoform from the peptides alone is problematic due to the small size of the peptides (on average 13 amino-acids), their low abundance, and their confinement to the CDS of the gene. These factors contribute to giving any random peptide an average probability of 30% to hit an exon-exon junction (Figure S3). Despite their increased number, ribosome footprints fare even worse due to their smaller size (on average 28-29nt) leading to an average probability of just 23% to hit an exon-exon junction (Figure S3). However RNA-seq reads, which are much longer (especially in modern paired-end data) have a much higher probability (on average 85%) of spanning at least one exon-exon junction (Figure S3). Our goal is to exploit the vastly greater power of RNA-seq to study the translatome at isoform resolution by identifying the major isoform(s) for each gene using the RNA-seq data to assist the assignment of footprints and peptides to these same isoforms, while still allowing them to diverge from the RNA-seq prediction if there is sufficient evidence against that particular mRNA.

One of the major confounders to transcript quantification by RNA-seq is intronic reads, i.e. reads that align either from an exon across into an intron or reads that align entirely inside an intron (Figure S4a). A likely source of these reads is from pre-spliced transcripts in the nucleus and we find that RNA-seq reads derived from raRNA indeed contain far fewer intronic reads than the totalRNA data (Figure S4b). We hypothesised that poly-A purification may bring a similar benefit

over total RNA, but inspection of data from an ENCODE K562 cell-line shows no such reduction in intronic reads (Figure S4c).

This 'cleaner' exonic signal from the raRNA does in fact lead to more consistent transcript quantification across our three biological replicate samples. If we define agreement as the same major isoform identified in all three replicate samples, raRNA provides a slight improvement in the fraction of genes that agree on the major isoform compared to totalRNA (Figure 2a). This agreement is clearly dependent on the expression of the gene and on the magnitude of the dominance of the major isoform (Figure 2b). By selecting realistic/sensible lower thresholds on expression and major isoform dominance we can substantially improve the fraction of genes that agree between all three replicate samples on the major isoform (Figure 2c). Selecting genes with a major isoform that is expressed above 5 TPM (an expression range that includes 96% of genes with footprint reads and/or peptides) and that accounts for more than 50% of the total number of RNA molecules produced by the gene, agreement increases to 97% for raRNA and 93% for totalRNA.

Bayesian approach to isoform-level integration of transcriptome and translatome

In order to exploit the higher resolution of RNA-seq, we designed an expectation maximisation (EM) algorithm that uses transcript specific RNA-seq expression values to be input as biologically informed prior expectations. These informed priors are used to update the likely abundance of each transcript, and assign footprint reads or peptides to their most likely transcript of origin. In Figure 3, we compare the iterations the algorithm progresses through when trying to assign ribosome footprints in POLDIP3 using a naive prior (assumes equal likelihood of all transcripts) compared to an RNA-seq prior (detailed browser tracks for these data - Figure S5a). Use of the naive prior results in a three way tie between three equally likely isoforms. Use of an RNA-seq prior overwhelmingly suggests the presence of a two fold abundant dominant transcript, 001, with a minor transcript of 002, an outcome which is fully consistent with both footprint read locations (Figure 3,S5a) and peptide data (Figure S5b).

We have implemented both the EM algorithm and the footprint frame prediction software into a novel software tool, miBAT (**m**ulti-modal **i**soform-level **B**ayesian **A**nalysis of the **T**ranslatome; https://github.com/gersteinlab/Thunder). The tool can be run on any combination of RNA-seq, ribosome footprint, and mass-spec proteomics data (Methods and Figure S6) although, as we discuss below, use of RNA-seq to set isoform priors greatly improves the ability of the algorithm to assign footprints and peptides to specific isoforms.

- by default, restricts footprints/peptides to the CDS, but can be run on whole transcripts for potential identification of uORFs or novel translated peptides

Biologically informed prior substantially improves footprint and MS/MS isoform assignments

The majority of multi-isoform genes still have an ambiguous major isoform following EM using a naive prior (Figure S7a). In 58% of genes neither footprints or peptides alone can distinguish a single major isoform, instead settling on two or more equally likely isoforms. In genes for which the naive EM does converge on a single major isoform, this isoform is typically extremely dominant and at least 5-fold more likely than the 'next-best' isoform (Figure S7b). Use of a biological prior (RNA-seq for footprints; RNA-seq or RNA-seq+footprints for mass-spec) resolves this ambiguity for a large fraction of these genes, reducing the ambiguity to less than 20% (Figure S7c-d).

We can define clusters of genes that behave similarly in terms of the ability to resolve a major isoform with different priors. Using an unsupervised hierarchical clustering and dynamic tree-cut (see Methods) we define 8 clusters of genes in each of the footprint EM and the proteomic EM that behave similarly based on the prior (Figure S8 and Figure 4a-b). These clusters can be further generalised into genes for which the biological prior is necessary or beneficial for major 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 major isoform with different biological priors (footprint EM: 1.6% of genes, proteomic EM: 11.6%). As a brief validation exercise, we selected a variety of genes with isoforms containing a single skipped exon, for which the naive EM was unable to identify a major isoform but the biological priors resolved this ambiguity. We designed PCR primers to amplify the region containing the prospective skipped exon, resulting in products of different sizes dependent on the presence or absence of the exon (See Methods for primer design). POLDIP3 shows evidence of a major transcript at 563 bp, the product size for transcript 001, with a minor product at 476 bp (transcript 002). For the remaining four cases, there is no evidence of a PCR product equating to the non-dominant transcript, which is consistent with the RNA-seq (Figure 4c-d).

There are a substantial number of genes for which the major isoform becomes clearer following the addition of footprint and/or proteomic data. The major isoform in genes in footprint EM clusters 3, 4, 6, and 8 in particular (Figure 4a and S8a) gets stronger with the addition of the footprint reads, as does the major isoform in clusters 1, 6, 7, and 8 (Figure 4b and S8b) following addition of the proteomic data. Therefore there is the potential for the footprints and/or peptides to assist RNA-seq with isoform quantification, or at least major isoform identification which may be of benefit to some studies, especially in organisms with a less complete annotation than human.

Post analysis of our complete data sets, there are two interesting outcomes to note. Firstly, when analysing the footprints, we see that the vast majority (>98%) of IP reads are assigned to mRNA transcripts by the EM compared to only ~85% of the sucrose cushion reads (Figure S9). Potentially nuisance biotypes including lincRNA, miRNA, snoRNA, and retained intron transcripts are clearly present in the cushion while completely absent from the IP sample,

suggesting these are artefact of sample preparation rather than reflective of real biology. Second, quantification of mRNA abundance, ribosome density, translational efficiency, and protein abundance at the gene- (Figure S10a) and isoform-level (Figure S10b) reveals little difference beyond the increased resolution of the set of expressed isoforms. In these rapidly dividing cells therefore, the major determinant of protein abundance is transcription, not translation, as has been reported by others in similar cell-systems [REF Biggin etc].

Finally, in order to assess the stability of the assignment of footprints to transcripts in a broader context, we applied both the naive EM and the RNA-seq prior to a recent large-scale ribosome footprinting analysis of 54 Yoruban individuals from the 1000 Genomes project [REF Battle] with complementary RNA-seq data [REF Pickrell]. We see that in this dataset, even though the footprint sequencing is substantially newer and, in most samples, contains between 2-5 times more reads than the single-end RNA-seq, the RNA-seq prior still significantly reduces the variation in major isoform fraction across the 54 individuals ($p<1E-16$; Figure S11).

Discussion

To our knowledge this approach represents the first experimentally and analytically integrated analysis of the translatome. We and others have previously reported on sequence-level integration of transcriptome and proteome [REFs] and related software tools; for example our 'RNAseq Translator' app on BaseSpace, a collaboration with ABSciex and Illumina, for mapping mass-spectra to a transcriptome derived directly from a Cufflinks [REF] RNA-seq transcriptome assembly. However our miBAT algorithm is the first to fully leverage the power of multi-level transcriptome and translatome profiling of the same cells by integrating abundance measurements in a statistically rigorous way through the transcriptome-translatome-proteome part of the central dogma.

For this integration we find that the profiling of mRNA by RNA-seq is critical for overcoming isoform ambiguity in ribosome footprints or mass-spec peptides; principally because RNA-seq reads are longer, but also because footprints and peptides are blind to differences in the 5' or 3' UTR of transcripts. Our analysis also highlights the benefit of clean data, both in terms of the reduced intronic contamination from raRNAs to the reduction of nuisance smallRNA contamination in the ribosome footprinting data.

Our enhanced, immunoprecipitation-based ribosome footprints also support a lack of evidence published by others [REF] for association of ribosomes with non-coding RNAs. In fact the transcript-level results from our analysis with miBAT suggests only a very small number of footprint reads cannot be explained by an annotated coding sequence, and it is highly likely that the remaining ~1% of reads result either from coding sequences that are not fully annotated or from upstream open reading frames. In non-transgenic systems it is still possible to obtain ribosomes via IP through the use of antibodies against endogenous ribosomal components, such as the Y10B antibody first discovered in Lupus patients [REF Joan-y-baby] that targets the 5.8S ribosomal-RNA.

Finally, it is worth noting that in these cultured, rapidly dividing cells, transcription is clearly the major determinant of protein abundance, as seen by others [REFs- biggin 25745146 etc]. It is well established however that translational control is very important in a large range of complex cellular systems. Aberrant control of protein translation may contribute to every stage of cancer cell transformation, through mechanisms including commandeering of initiation factors, decreased IRES-dependent translation of critical tumor suppressors, and cap-independent translation of angiogenic factors. [PMID:22767671,20332778]. In the immune system, translational control allows specific cell types to respond differentially to the same stimulus, through activation of mTOR signaling, phosphorylation of initiation factors, and synthesis of ribosome binding proteins [PMID:24840981]. In the nervous system, rapid non-transcription-dependent translation of proteins is responsible for synaptic plasticity, the process that underlies formation of long-term memories, and protein translation may be disrupted in neurodegenerative, neurodevelopmental, and psychiatric conditions as wide ranging as Alzheimer's disease, autism, and major depressive disorder [25032491]. In all these systems, it is clear that translation control mechanisms must be specific to certain cell types, but little progress has been made towards elucidating targets on a genome wide scale. Successful investigations are currently underway to investigate the translatomes of specific groups of cells in the nervous systems (BACTRAP, Ribotag) and immune systems (FACs sorting). For this purpose, all of the IP methods presented in this paper are fully compatible with transgenic systems which allow for specific ribosome IPs to enable deep investigation of targets of translation control which are specific to certain cell types.

Figure Legends

Figure 1 | experimental approach to integrated analysis of the transcriptome and translatome

a: Schematic diagram of the experimental approach to multi-modal profiling of the translatome in Human Embryonic Kidney (HEK293) cells. Total-RNA and protein are obtained from lysing whole cells, raRNA and ribosome footprint RNA are obtained by immunoprecipitation of intact ribosomes from the cytosol.

b: RNA-seq of both totalRNA (dark blue) and raRNA (light blue) capture a large range of molecules that vary widely in abundance and biotype. Genes with at least two ribosome footprints (green) are generally expressed above ~1 transcript per million (TPM). Standard, unfractionated mass spectrometry can detect only the most abundant proteins (with at least two peptides per protein) while highly fractionated MS/MS, where peptides are separated and run over serial injections, delves much deeper into the lower-abundance proteome.

Figure 2 | major isoform identification is consistent across biological replicates

a: The majority of multi-isoform genes from both totalRNA-seq and raRNA-seq agree on the same major isoform in all three biological replicate samples. Grey = single isoform genes. Black = all three replicates agree, red = at least one replicate shows a different major isoform.

b: Agreement on the same major 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 major isoform between samples.
c: Heat maps show the effect of varying minimum thresholds of gene expression and major isoform dominance on this agreement between replicates. Greater consistency is evident in raRNA compared to totalRNA, represented by the increased area of white in the upper right quadrant. However, in both cases, more than 90% of genes with at least a 50% dominant major isoform expressed at more than 5 transcripts per million will consistently call the same major transcript.

Figure 3 | analytical approach to integrated analysis of the transcriptome and translatome
Schematic diagram depicting the expectation maximisation process for the POLDIP3 gene using the 11 ribosome footprint reads that map to the coding sequence (CDS) of one or more of its mRNA transcripts. The uninformative 'naive' prior, 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 use of a biologically informed prior, obtained directly from the relative transcript abundances from totalRNA-seq or raRNA-seq, overcomes this ambiguity 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 a minor secondary isoform of POLDIP-002.

Figure 4 | a biologically realistic prior dramatically 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 based on the result of the EM. Each plot shows the major 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), those for which the major isoform is decided entirely by the footprints (blue), and those in which the major isoform is unstable (red).
b) As **a)** following EM using peptides obtained from mass-spectrometry. Here the priors are naive (left), raRNA (centre), and raRNA+footprints (right); where the latter is the isoform abundances output by the ribosome footprint EM using the raRNA prior - the right column in **a)**. Generally the use of the biologically informative prior is more useful as the peptides are less capable of resolving the major isoform than the more abundant ribosome footprints.
c) Detailed illustration of the EM result for five selected genes showing differences in the relative isoform abundances of each. In all cases, the biological prior is necessary to resolve the major isoform (red) and the second isoform (blue) where applicable. To the right, the isoform names are shown along with the relative abundance (as a percent of the gene expression) and expected product size for the PCR validation in **d)** (see also supplemental methods).
d) PCR validation of the five genes selected in panel **c)** show that, at least at the mRNA level, all agree with the major isoform inferred by RNA-seq. POLDIP3 (left) also shows evidence for the expression of the second isoform predicted by RNA-seq.

Figure S1 | summary of totalRNA, raRNA, and ribosome footprint data
a) Comparison of totalRNA and raRNA in terms of detecting low-abundance non-coding transcripts (ncRNA); Low abundance ncRNA are depleted in raRNA suggesting these transcripts are not engaged by the ribosome (top graph), however raRNA still detects the majority of mRNAs at every level of expression (bottom graph).
b) Further comparison of the RNA [gene] biotypes detected by totalRNA and raRNA-seq. Detection is liberally defined as any gene with a non-zero mean average expression over the three biological replicate samples. The number of genes detected by totalRNA (dark blue bars) 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; >95% of mRNAs ('protein_coding') detected in totalRNA are also observed as raRNA. This fraction decreases for non-coding other 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.
c) Comparison of biochemical methods for obtaining ribosome footprints reveals immunoprecipitation (IP) by eGFP-L10a produces much cleaner profiles of reads as assessed by read length and multi-mappability to the genome (left histogram). The standard method of purifying footprints (by sucrose cushion density gradient; middle and right histograms) produces a much wider range of 'off-target' (28nt < L > 32nt) read lengths (L) and a much lower fraction of reads that can be uniquely mapped to the genome (green bars) than IP.
d) Of the reads that can be uniquely mapped to the genome, the IP sample shows slightly reduced yield but far greater purity in terms of coding sequence mapped reads compared to the two sucrose cushion samples (left plot). The purity of the cushion samples improves a little after explicitly removing likely-contaminant genes based on biotype: snRNA, snoRNA, miRNA, tRNA, Mt_rRNA, Mt_tRNA, misc_RNA.

Figure S2 | ribosome footprints obtained by IP have increased fidelity to the coding frame
a) In a perfect footprint preparation, all reads would be 28 nucleotide in length (the exact number of nucleotides physically protected by a cycloheximide halted ribosome), and the read midpoint would sit a predictable nucleotides from the start of the nearest codon. In reality, due to the variations inherent in the technique such as incomplete RNase cleavage, a range of fragment lengths is obtained. Rather than discard these "imperfect" reads, it is possible to compute a position-weight matrix (PWM) showing the fraction of reads of a given length at each of the three possible read-mid-point to nearest codon offsets. Here is plotted the PWM for the ribosome footprint reads obtained by sucrose cushion, showing that 57% of 29nt reads and 61% of 30nt reads lie a predictable distance from the nearest codon.
b) As **a)**, however showing the PWM for ribosome footprint reads obtained by eGFP-L10a IP. A much greater fraction of 28-32nt reads are in frame with each other following IP compared to the sucrose cushion, again reflecting the superior quality of this method of preparation.
c) Graphical summary of the PWMs in **a)** and **b)**, where the implication is that incomplete RNase digestion tends to leave uncleaved nucleotides at the 3' end of the footprint. Given that

this digestion is likely specific to each preparation we use experiment-specific PWMs as a means to determine the quality of a ribosome footprint experiment.
d) Using the footprints obtained by sucrose cushion and the PWM in **a)** we show that the fraction of transcripts called in the correct frame (compared to the annotation, see **methods**) increases with increasing number of observed footprints.
e) As **d)**, except using the footprints obtained by eGFP-L10a IP and the PWM in **b)**. Far fewer reads are required to correctly determine the frame of a transcript compared to sucrose cushion; 75% of transcripts with ≥3 IP footprint reads are called in the correct frame, compared to just 55% by sucrose cushion. This percentage rises rapidly for the IP reads and with ≥10 reads it is possible to accurately predict the correct frame for >90% of transcripts.

Figure S3 | probability distribution, over all ~80,000 mRNA transcripts, of a randomly selected RNA-seq read, ribosome footprint, or mass-spec 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, a la 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 mate. With an average 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). Finally, the 28nt ribosome footprints are the least likely to produce exon-spanning reads (~23%; green bars).

Figure S4 | 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 [REF]. 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 S5 | More detailed summary of alignments and EM performance for POLDIP3
a) Browser track of totalRNA-seq, raRNA-seq, and ribosome footprints 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.

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 S6 | Schematic diagram showing the variety of ways of using the miBAT tool 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 tool [REF Robertson] and all footprint alignments are in transcriptome coordinates.
b) As **a)**, but for mass-spectrometry experiments with peptide alignments (produced by X!Tandem [REF]) in transcript coordinates. The top row is 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 directly 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 | use of a biologically informative prior reduces isoform ambiguity in footprint and MS proteomic data
Using an RNA-seq prior robustly decreases the ambiguity of footprint and peptide assignment to the major 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 major isoform in 58.1% of multi isoform genes.
b) Re-plotted major isoform ambiguity for naive EM on ribosome footprint data (left) and MS/MS peptides (right) in terms of fold-dominance of the major isoform over the second most abundant isoform. On these histograms, a value of 1 reflects a gene in which a single major isoform cannot be determined. For clarity, major isoforms more than 5-fold more abundant than the second isoform are capped to 5-fold dominance.
c) Using totalRNA- or raRNA-seq as a prior substantially improves our ability to resolve a dominant major isoform from ribosome footprint EM.
d) Using RNA-seq and/or ribosome footprint data as a prior greatly improves ability to resolve a dominant major isoform from MS/MS peptides.

Figure S8 | 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 major isoform following the miBAT 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 major isoforms (x-axes) are indicated by dendrogram colours (top) and numeric IDs (bottom), the latter matching the IDs in **Figure 4**.
a) The heatmap plots the major isoform fraction for the 6,650 multi-isoform genes with at least 3 footprint reads. Using a biological prior improves our ability to resolve the major 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 major isoform fraction for the 1,212 multi-isoform genes with at least 2 peptides. Using a biological prior improves our ability to resolve the major 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 major isoform.

Figure S9 | ribosome footprints obtained by IP are more reliably assigned to coding transcripts
Following the EM, >98% of reads from ribosome footprints obtained by IP are assigned to protein coding (mRNA) transcripts. Compared to ~85% of reads from footprints obtained by sucrose cushion, this again reflects the superior purity resulting from this method of footprint extraction. Transcript biotypes that are over-represented in the cushion reads include miRNA, lincRNA, and snoRNA; these likely represent non-specific small-RNA contaminants rather than useful ribosome-protected fragments.

Figure S10 | quantification of RNA expression, translational efficiency, and protein abundance
Comparison of expression at each level, RNA expression vs. ribosome density (**left**), RNA expression vs. protein abundance (**middle**), and translational efficiency vs. protein abundance (**right**) at both **a)** the gene-level, and **b)** the isoform-level shows that in this system the abundance of protein is largely driven by the RNA expression rather than any wide scale regulation of translation.

Figure S11 | data from 54 Yoruban individuals shows a biological prior from RNA-seq improves major isoform consistency even with very deep footprint sequencing
a) Major isoforms determined from a public ribosome footprint data derived from 54 Yoruban individuals in the 1000 Genomes project [REF Battle] show lower variability when the EM is run with RNA-seq data from the same individuals [REF Pickrell]. This dataset is particularly interesting as the ribosome footprint sequencing is of substantially higher quality than the older RNA-seq data. Despite this, the use of the RNA-seq data as priors for the EM performed by miBAT significantly reduced the variability in transcript fraction across the 54 biological replicate samples. The plots show the distribution of major transcript fractions for all genes (x-axis) across all 54 individuals in terms of median/IQR (**top**) and mean/stddev (**bottom**). The variability across the individuals is noticeably lower with the RNA prior (**right**) compared to the naive prior (**left**).

b) Boxplots of the major isoform fraction standard deviations (grey vertical error bars from the bottom row of plots in **a)**) show that the RNA-seq prior significantly reduces the variability across these 54 individuals ($p<1E-16$).