bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

XPRESSyourself: Enhancing, Standardizing, and Automating Ribosome Profiling Computational Analyses Yields Improved Insight into Data

Jordan A. Berg,[1*] **Jonathan R. Belyeu,**[2] **Jeffrey T. Morgan,**[1] **Yeyun Ouyang,**[1] **Alex J. Bott,**[1] **Aaron R. Quinlan,**[2,4,5] **Jason Gertz,**[3] **Jared Rutter**[1,6*]

[1]Department of Biochemistry, University of Utah, Salt Lake City, UT, USA, 84112.

[2]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA, 84112.

[3]Department of Oncological Sciences, University of Utah, Salt Lake City, UT, USA, 84112.

[4]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA, 84112.

[5]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA, 84112.

[6]Howard Hughes Medical Institute, University of Utah, Salt Lake City, UT, USA, 84112.

*Address correspondence to: jordan.berg@biochem.utah.edu, rutter@biochem.utah.edu

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Ribosome profiling, an application of nucleic acid sequencing for monitoring ribosome activity, has revolutionized our understanding of protein translation dynamics. This technique has been available for a decade, yet the current state and standardization of publicly available computational tools for these data is bleak. We introduce XPRESSyourself, an analytical toolkit that eliminates barriers and bottlenecks associated with this specialized data type by filling gaps in the computational toolset for both experts and non-experts of ribosome profiling. XPRESSyourself automates and standardizes analysis procedures, decreasing time-to-discovery and increasing reproducibility. This toolkit acts as a reference implementation of current best practices in ribosome profiling analysis. We demonstrate this toolkit's performance on publicly available ribosome profiling data by rapidly identifying hypothetical mechanisms related to neurodegenerative phenotypes and neuroprotective mechanisms of the small-molecule ISRIB during acute cellular stress. XPRESSyourself brings robust, rapid analysis of ribosome-profiling data to a broad and ever-expanding audience and will lead to more reproducible and accessible measurements of translation regulation. XPRESSyourself software is perpetually open-source under the GPL-3.0 license and is hosted at https://github.com/XPRESSyourself, where users can access additional documentation and report software issues.

Introduction

High-throughput sequencing data has revolutionized biomedical and biological research. One such application of this consequential technology is ribosome profiling, which, coupled with bulk RNA-Seq, measures translation efficiency, translation pausing, novel protein translation products, and more [1–3]. Though the experimental procedures for ribosome profiling have matured, an abundance of biases and peculiarities associated with each analytical method or tool are still present and may often be obscured to a new user of this methodology [4–8]. Additionally, standardized methods for handling this unique data type remain elusive. This has been problematic and evidenced by various studies using vague or opaque methods for data analysis (for examples, see [9–13]), or methods rely on outdated tools [5]. Very few labs have the tools necessary to separate the biological signals in ribosome profiling data from the inherent biases of the experimental measurements, and these tools are not readily accessible by the community. This is a critical time in the rapidly expanding influence of ribosome profiling. For too long, the bioinformatic know-how of this incredibly powerful technique has been limited to a small handful of labs. As more and more ribosome profiling studies are performed, more and more labs will lack the ability to analyze their data with ease and fidelity. Few, if

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

any, extant pipelines or toolkits offer a thorough set of integrated tools for assessing standard quality control metrics or performing proper reference curation to reduce systematic biases across any organism, particularly with ribosome profiling data [14–18].

For example, one issue in ribosome profiling is the pile-up of ribosomes at the $5'$- and $3'$- ends of coding regions within a transcript, a systematic biological signal arising from the slower kinetics of ribosome initiation and termination compared to translation elongation and is generally regarded to not accurately reflect measurements of translation efficiency. These signals are further exacerbated by pre-treatment with cycloheximide during ribosome footprint harvesting [4, 19, 20]. These pile-ups can dramatically skew ribosome footprint quantification and measurements of translational efficiency. Current practices in the field recommend excluding pile-up-prone regions when quantifying ribosome profiling alignments as they lead to noisier estimations of translation efficiency [3, 21]; however, no publicly available computational tools currently exist to facilitate these automated adjustments to reference transcripts. Curating references properly and robustly requires advanced implementations. In addition, downstream data visualization methods presently available are often not optimized to analyze and compare translation regulatory regions of a gene.

To address deficiencies in the public ribosome profiling computational toolkit, we developed XPRESSyourself, a computational toolkit and adaptable, end-to-end pipeline that bridges these and other gaps in ribosome profiling data analysis. XPRESSyourself implements the complete suite of tools necessary for comprehensive ribosome profiling and bulk RNA-Seq data processing and analysis in a robust and easy-to-use fashion, often packaging tasks that would typically require hundreds to thousands of lines of code into a single command. For instance, XPRESSyourself creates the mRNA annotation files necessary to remove confounding systematic factors during quantification and analysis of ribosome profiling data, allowing for accurate measurements of translation efficiency. It provides the built-in capacity to quantify and visualize differential upstream open-reading frame (uORF) usage by generating IGV-like, intron-less plots for easier visualization [22]. The ability to visualize (and in another XPRESSyourself module, quantify) the usage of micro-uORFs is important in exploring regulatory events or mechanisms in a wide array of biological responses and diseases. XPRESSyourself also introduces a tool for efficient identification of the most problematic rRNA fragments for targeted depletion, which provides immense financial and experimental benefits to the user by amplifying ribosome footprint signal over rRNA noise. Tools like this will become vital as ribosome profiling moves into development in new organisms.

XPRESSyourself aims to address the lack of consensus in analytical approaches used to process ribosome profiling data by acting as a reference implementation of current best practices for ribosome profiling analysis. While

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a basic bioinformatic understanding is becoming more commonplace amongst the scientific community, the intricacies of processing RNA-Seq data remain challenging for many. Moreover, many users are often not aware of the most up-to-date tools or the appropriate settings for their application [23, 24]. Even for the experienced user, developing robust automated pipelines that accurately process and assess the quality of these datasets can be laborious. The variability that inevitably arises with each lab or core facility designing and using distinct pipelines is also a challenge to reproducibility in the field. XPRESSyourself curates the state-of-the-art methods for use and where a required functionality is unavailable, introduces a thoroughly tested module to fill that gap. While some tasks in these pipelines may be considered mundane, we eliminate the need of each user to rewrite even simple functionality and promote reproducibility between implementations. To aid users of any skill-level in using this toolkit, we provide thorough documentation, walkthrough videos, and interactive command builders to make usage as easy as possible, while allowing for broad use of this toolkit from personal computers to high-performance clusters.

Finally, the most broadly relevant aspect of our update and streamlining of ribosome-profiling analysis is the novel biological insights we are able to obtain from published datasets. We highlight this in the ISRIB ribosome-profiling study discussed in this manuscript, where we are able to observe significant translation regulation that was missed previously when the data were initially analyzed using now outdated techniques. This analysis generates novel hypotheses for genes potentially involved in neurodegeneration in humans, but more broadly emphasizes the benefit of analysis and re-analysis of data using the complete and up-to-date benchmarked methodology provided within XPRESSyourself.

Design and Implementation

Architecture and Organization

XPRESSyourself is currently partitioned into two software packages, XPRESSpipe and XPRESSplot. XPRESSpipe contains automated, end-to-end pipelines tailored for ribosome profiling, single-end RNA-Seq, and paired-end RNA-Seq datasets. Figure 1 outlines the tasks completed by these pipelines. Individual sub-modules can be run automatically through a pipeline or manually step-by-step. Modules optimize available computational resources where appropriate to deliver results as quickly as possible. XPRESSplot is available as a Python library and provides an array of analytical methods specifically for sequence data, but tractable to other data types. For a comparison of how XPRESSyourself compares to other available software packages available at the time of writing, we refer the reader

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to Figure S1 [14, 15, 18, 25–49].

To make analysis as easy and accessible as possible, an integrated command builder for reference curation and sample analysis can be run by executing `xpresspipe build`. This command builder will walk the user through potential considerations based on their library preparation method and build the appropriate command for execution on their personal computer or a supercomputing cluster. The builder will then output the requested command for use on a computational cluster, or the command can be executed immediately on a personal computer.

The software is designed such that updating and testing of a new module, or updating dependency usage, are facile tasks for a trained bioinformatician. More details on current and future capabilities can be found in each package's documentation [50, 51] or their respective `versions` page on each toolkit's repository page [52].

Automated Reference Curation

The first step of RNA-Seq alignment is curating an organism reference to which the alignment software will map sequence reads. XPRESSpipe uses STAR [53] for mapping reads as it has been shown consistently to be the best performing RNA-Seq read aligner for the majority of cases [54, 55]. The appropriate reference files are automatically curated by providing the appropriate GTF file saved as `transcripts.gtf` and the directory path to the genomic FASTA file(s). Additional modifications to the GTF file required for ribosome profiling or desired for RNA-Seq are discussed in the next section. We recommend organizing these files in their own directory per organism.

GTF Modification

For ribosome profiling, frequent read pile-ups are observed at the $5'$- and $3'$- ends of an open reading frame which are largely uninformative to a gene's translational efficiency [4]. While these pile-ups can be indicative of true translation dynamics [56], current best-practices have more recently settled on ignoring these regions during read quantification and calculations of translation efficiency [3, 21]. By providing the `--truncate` argument during reference curation, the $5'$- and $3'$- ends of each coding region will be recursively trimmed until the specified amounts are removed from coding space. A recursive strategy is required here as GTF file-formats split the CDS record into regions separated by introns. By default, 45 nt will be trimmed from the $5'$-ends and 15 nt from the $3'$-ends recursively until the full length is removed from coding space, as is the current convention within the ribosome profiling field [3]. The resulting output file will then be used to process ribosome footprint libraries and their corresponding bulk RNA-Seq libraries. If

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

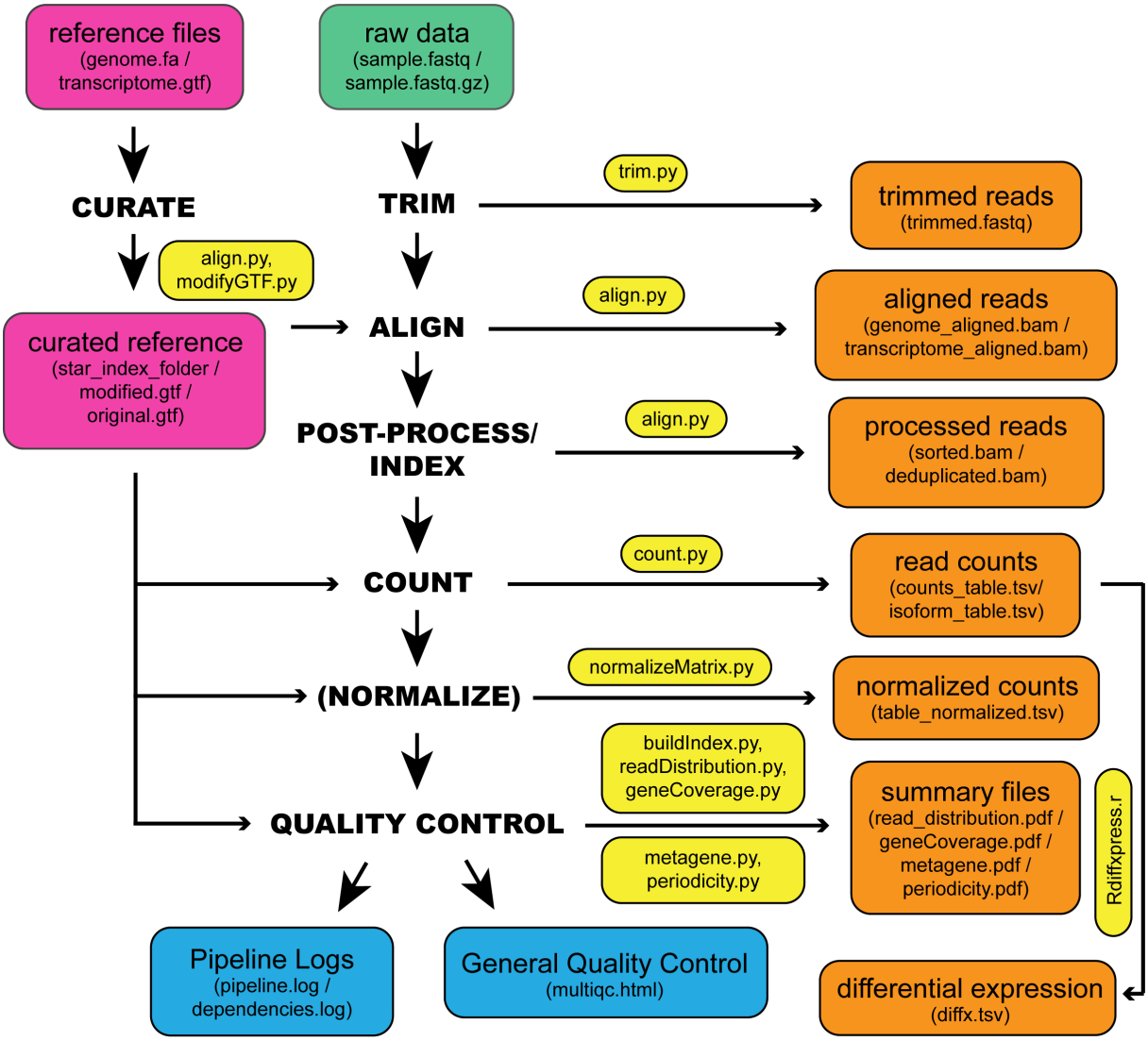

Figure 1: **Workflow schematic of the inputs, outputs, and organization of XPRESSpipe.** Representation of the general steps performed by XPRESSpipe with data and log outputs. Steps in parentheses are optional to the user. Input and output file types are in parentheses for each input or output block. The main script(s) used for a given step are in yellow blocks. The green block indicates input sequence file(s). Pink blocks indicate reference input files and curated reference. Orange blocks indicate output files. Blue blocks indicate general quality control and log file outputs. Differential expression analysis is run independently from the pipeline as the user will need to ensure count table and metadata table formatting are correct before use.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

generating a GTF file for use solely with general bulk RNA-Seq datasets, this file should not be truncated.

Optionally, the GTF can be parsed to retain only protein-coding genes records. This acts as a read masking step to exclude non-protein coding transcripts. In particular, overabundant ribosomal RNAs resulting during library preparation are excluded from downstream analyses using this modified reference file. Parameters can also be provided to retain only the Ensembl canonical transcript record. This can be useful for some tools that penalize reads that overlap multiple isoforms of the same gene. If using HTSeq with default XPRESSpipe parameters or Cufflinks to quantify reads, this is not necessary as they do not penalize a read mapping to multiple isoforms of the same gene or are capable of handling quantification of different isoforms of a gene [57, 58].

Read Processing

Pre-Processing. In order for sequence reads to be mapped to the genome, reads generally need to be cleaned of artifacts from library creation. These include adaptors, unique molecular identifier (UMI) sequences, and technical errors in the form of low-quality base calls. Parameters, like minimum acceptable quality or length, can be modified, or features such as unique molecular identifiers (UMIs) can be specified to identify and group PCR artifacts for later removal [59, 60].

Alignment. Reads are aligned to the reference genome with STAR, which, despite being more memory-intensive, is one of the fastest and most accurate sequence alignment options currently available [53–55]. XPRESSpipe is capable of performing a single-pass, splice-aware, GTF-guided alignment or a two-pass alignment of reads wherein novel splice junctions are determined and built into the genome index, followed by alignment of reads using the updated index. Both coordinate- and transcriptome-aligned BAM files are output by STAR. We abstain from rRNA negative alignment at this step as downstream analysis of these mapped reads could be of interest to some users. When rRNA alignment is preferred, a protein-coding-only GTF file should be provided during quantification. A STAR-compatible VCF file can also be passed to this step to allow for genomic variant consideration during alignment.

Post-Processing. XPRESSpipe further processes alignment files by sorting, indexing, and optionally parsing unique alignments based on UMIs for downstream analyses. PCR duplicates are also detected and marked or removed for downstream analyses; however, these files are only used for relevant downstream steps or if the user specifies to use these de-duplicated files in all downstream steps. Use of de-duplicated alignment files may be advisable in situations where the library complexity profiles (discussed below) exhibit high duplication frequencies. However, generally the

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

abundance of PCR-duplicates is low in properly-prepared sequencing libraries; thus, doing so may be overly stringent and unnecessary [59]. Optionally, BED coverage files can also be output.

Quantification. XPRESSpipe quantifies read alignments for each input file using HTSeq with the `intersection-nonempty` method by default [57, 61]. We use this quantification method as it conforms to the current TCGA processing standards and is favorable in the majority of applications [62]. If masking of non-coding RNAs or quantification to truncated CDS records is desired, a `protein_coding` modified GTF file should be provided to the `--gtf` argument. HTSeq importantly allows selection of feature type across which to quantify, thus allowing for quantification across the CDSs of a transcript instead of entire exons. If a user is interested in quantifying ribosome occupancy of transcript uORFs in ribosome footprint samples, they can provide `five_prime_utr` or `three_prime_utr` for the `--feature_type` parameter if such annotations exist in the organism of interest's GTF file. If the user is interested in isoform abundance estimation of reads, Cufflinks is alternatively available for quantification [58, 61].

Normalization. Methods for count normalization are available within XPRESSpipe by way of the XPRESSplot package. For normalizations correcting for transcript length, the appropriate GTF must be provided. Sample normalization methods available include reads-per-million (RPM), Reads-per-kilobase-million (RPKM) or Fragments-per-kilobase-million (FPKM), and transcripts per million (TPM) normalization [63]. For samples sequenced on different chips, prepared by different individuals, or on different days, the `--batch` argument should be provided along with the appropriate metadata matrix [64].

Quality Control

Read Length Distribution. The lengths of all reads are analyzed after trimming. By assessing the read distribution of each sample, the user can ensure the expected read size was sequenced. This is particularly helpful in ribosome profiling experiments for verifying the requisite 17-33 nt ribosome footprints were selectively captured during library preparation [3, 65]. Metrics here, as in all other quality control sub-modules, are compiled into summary figures for easy pan-sample assessment by the user.

Library Complexity. Measuring library complexity is an effective method for analyzing the robustness of a sequencing experiment in capturing various, unique RNA species. As the majority of RNA-Seq preparation methods involve a PCR step, sometimes particular fragments will be favored and over-amplified in contrast to others. By plotting the number of PCR replicates versus expression level for each gene, one can monitor any effects of limited transcript

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

capture diversity and high estimated PCR duplication rate on the robustness of their libraries. This analysis is performed using dupRadar [66] using the duplicate-tagged alignment files output during post-processing. Metrics are then compiled and plotted by XPRESSpipe.

Metagene Estimation Profile. To identify any general biases for the preferential capture of the $5'$- or $3'$- ends of transcripts, metagene profiles are generated for each sample. This is performed by determining the meta-genomic coordinate for each aligned read in exon space. Coverage is calculated for each transcript, normalized, and combined to eliminate greediness of super-expressors in profile coverage. Required inputs are an indexed BAM file and an un-modified GTF reference file. Outputs include metagene metrics, individual plots, and summary plots. Parameters can be tuned to only analyze representation along CDS regions.

Gene Coverage Profile. Extending the metagene estimation analysis, the user can focus on the coverage profile across a single gene. Although traditional tools like IGV [22] offer the ability to perform such tasks, XPRESSpipe offers the ability to collapse the introns to observe coverage over exon space only. This is helpful in situations where massive introns spread out exons and make it difficult to visualize exon coverage for the entire transcript in a concise manner. CDS feature annotations are displayed to aid ribosome profiling data users in identifying CDS coverage and uORF translation events. When running a XPRESSpipe pipeline, a housekeeping gene will be automatically processed and output for the user's reference. Figure S2 provides a comparison with the output of IGV [22] and XPRESSpipe's `geneCoverage` module over a similar region for two genes to demonstrate the compatibility between the methods. We note that while the published superTranscripts tool offers similar functionality, it lacks integration and automation and must be manually paired with IGV for multi-sample comparisons and visualization [31]. Other tools, such as Rfeet and riboStreamR [25, 47], suffer from similar integration and automation shortcomings. XPRESSpipe's `geneCoverage` module offers easy and automated functionality for this task.

Codon Phasing/Periodicity Estimation Profile. In ribosome profiling, a useful measure of a successful experiment is obtained by investigating the codon phasing of ribosome footprints [3]. To do so, the P-site positions relative to the start codon of each mapped ribosome footprint are calculated using riboWaltz [67]. The same inputs are required as in the `metagene` sub-module.

Identify Problematic rRNA Fragments from Ribosome Footprinting for Depletion. rRNA depletion is intrinsically complicated during the preparation of ribosome-footprint profiling libraries: poly(A) selection is irrelevant, and kit-based rRNA depletion is grossly insufficient. Especially in the case of ribosome profiling experiments, where RNA is digested by an RNase to create ribosome footprints, many commercial depletion kits will not target the most abundant

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

rRNA fragment species produces during the footprinting step of ribosome profiling. The sequencing of these RNAs becomes highly repetitive, wasteful, and typically biologically uninteresting in the context of gene expression and translation efficiency. The depletion of these sequences is therefore desired to increase the depth of coverage of ribosome footprints. Depending on the species and condition being profiled, custom rRNA-depletion probes for a small subset of rRNA fragments (generally 2-5) can easily account for more than 90% of sequenced reads [1, 3]. `rrnaProbe` analyzes the over-represented sequences within a collection of footprint sequence files that have already undergone adaptor and quality trimming, compiles conserved sequences across the overall experiment, and outputs a rank-ordered list of these sequences for probe design.

Analysis

XPRESSpipe provides a DESeq2 command-line wrapper for performing differential expression analysis of count data. We refer users to the original publication for more information about uses and methodology [68].

More analytical features are available in XPRESSplot, which requires as input a gene count table as output by XPRESSpipe and a meta-sample table (explained in the documentation [51]). Analyses with limited to no options in Python libraries include principle components plotting with confidence intervals and automated volcano plot creation for RNA-Seq or other data. Other instances of analyses can be found in the documentation [51].

Results

Benchmarking Against Published Ribosome Profiling Data and New Insights

The integrated stress response (ISR) is a signaling mechanism used by cells and organisms in response to a variety of cellular stresses [69]. Although acute ISR activation is essential for cells to properly respond to stresses, long periods of sustained ISR activity can be damaging. These prolonged episodes contribute to a variety of diseases, including many resulting in neurological decline [70]. A recently discovered small-molecule inhibitor of the ISR, ISRIB, has been demonstrated to be a potentially safe and effective neuroprotective therapeutic for traumatic brain injury and other neurological diseases. Interestingly, ISRIB can suppress the damaging chronic low activation of the ISR, while it does not interfere with a cytoprotective acute, high-grade ISR, adding to its wide pharmacological interest [9, 71–76].

A recent study (data available under Gene Expression Omnibus accession number GSE65778) utilized ribosome profiling to better define the mechanisms of ISRIB action on the ISR, modeled by 1-hour tunicamycin (Tm) treatment

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in HEK293T cells [9]. A key finding of this study is that a specific subset of stress-related transcription factor mRNAs exhibits increased translational efficiency (TE) compared to untreated cells during the tunicamycin-induced ISR. However, when cells were co-treated with tunicamycin and ISRIB, the TE of these stress-related mRNAs showed no significant increase compared to untreated cells, which indicates that ISRIB can counteract the translational regulation associated with the ISR.

To showcase the utility of XPRESSpipe in analyzing ribosome profiling and sequencing datasets, we re-processed and analyzed this dataset using the more current *in silico* techniques included in the XPRESSpipe package to further query the translational mechanisms of the ISR and ISRIB. All XPRESSpipe-processed biological replicate samples exhibited a strong correlation between read counts per gene when thresholded similarly to count data available with the original publication (Spearman ρ values 0.991-0.997) (Figure 2A shows representative plots; Figure S3A shows all replicate comparisons; Figure S4B shows the corresponding plots using the count data provided with the original publication for reference).

Compared to the count data made available with the original manuscript, when XPRESSpipe-processed samples were thresholded as in the original published count data, samples showed generally comparable read counts per gene between the two analytical regimes (Spearman ρ values 0.937-0.951) (Figure 2B shows representative plots; Figure S3B shows all comparisons). This is in spite of the fact that the methods section of the original publication employed software that was current at the time but is now outdated, such as TopHat2 [77], which has a documented higher false-positive alignment rate, generally lower recall, and lower precision at correctly aligning multi-mapping reads compared to STAR [53–55]. Many of the genes over-represented in the original count data as compared to data processed by XPRESSpipe appear to be due to the over-estimation of pseudogenes or other gene paralogs. Figure S4A highlights a sampling of some extreme cases where particular genes with paralogs are consistently over-represented between samples in the original processed data. This suggests a programmatic difference in how these transcripts are being treated. As these genes share high sequence similarity with each other, reads mapping to these regions are difficult to attribute to a specific genomic locus and are often excluded from further analyses due to their multi-mapping nature. The benchmarking study [54] that examined these and other aligners described how TopHat2 had a disproportionally high rate of incorrectly aligned bases or bases that were aligned uniquely when they should have been aligned ambiguously, at least partially explaining the observed overcounting effect with TopHat2. Had TopHat2 marked problematic reads as ambiguous, they would have been excluded from later quantification.

Additionally, when TopHat2 and STAR were tested using multi-mapper simulated test data of varying complexity,

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TopHat2 consistently suffered in precision and recall. These calls are increasingly more difficult to make with smaller reads as well, and this is evident from Figure 2B, where ribosome footprint samples consistently showed more over-counted genes than the corresponding RNA-Seq samples. When dealing with a ribosome footprint library of about 50-100 million reads, and with TopHat2's simulated likelihood of not marking an ambiguous read as such being about 0.5% higher than STAR, this would lead to around 250,000 to 500,000 spuriously aligned reads, which is in line with our observations (statistics were derived from [54]; analyses are available in the manuscript scripts repository [78]).

Another potential contributor to this divergence is that the alignment and quantification within XPRESSpipe use a current human transcriptome reference, which no doubt contains updates and modifications to annotated canonical transcripts and so forth when compared to the version used in the original study. However, in practice, these effects are modest for this dataset (Figure S5). Additionally, the usage of the now outdated DESeq1 [79] appears to contribute significantly to the outcome in differential expression analysis (Figure S6). While differences in processing between the outdated and current methods may not always create systematic differences in output, key biological insights may be missed. The analysis that follows is exploratory and only meant to suggest putative targets identifiable by re-analyzing pre-existing, publicly available data.

We first looked at the canonical targets of translation regulation during ISR, as identified in the original study within the XPRESSpipe-processed data. These targets include ATF4, ATF5, PPP1R15A, and DDIT3 (Figure 3A-C, highlighted in purple) [9]. Of note, the fold-change in ribosome occupancy of ATF4 (6.83) from XPRESSpipe-processed samples closely mirrored the estimate from the original publication (6.44). Other targets highlighted in the original study [9], such as ATF5, PPP1R15A, and DDIT3 also demonstrated comparable increases in their ribosome occupancy fold-changes to the original publication count data (XPRESSpipe: 5.90, 2.47, and 3.94; respectively. Original: 7.50, 2.70, and 3.89; respectively) (Figure 3A). Similar to the originally processed data, all of these notable changes in ribosome occupancy return to untreated levels during Tm + ISRIB co-treatment (Figure 3B). Additional ISR targets containing micro-ORFs described in the study (highlighted in green in Figure 3A-C) were also similar in translational and transcriptional regulation across conditions between the two analytical regimes.

Both the original study and our XPRESSpipe-based re-analysis show that ISRIB can counteract the significant increase in TE for a set of genes during ISR. To further build upon the original analysis and explore TE regulation during ISR, we asked if ISRIB has a similar muting effect on genes with significant decreases in TE induced by the ISR. In the original study, genes with significant decreases in TE were reported in a source-data table and not

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

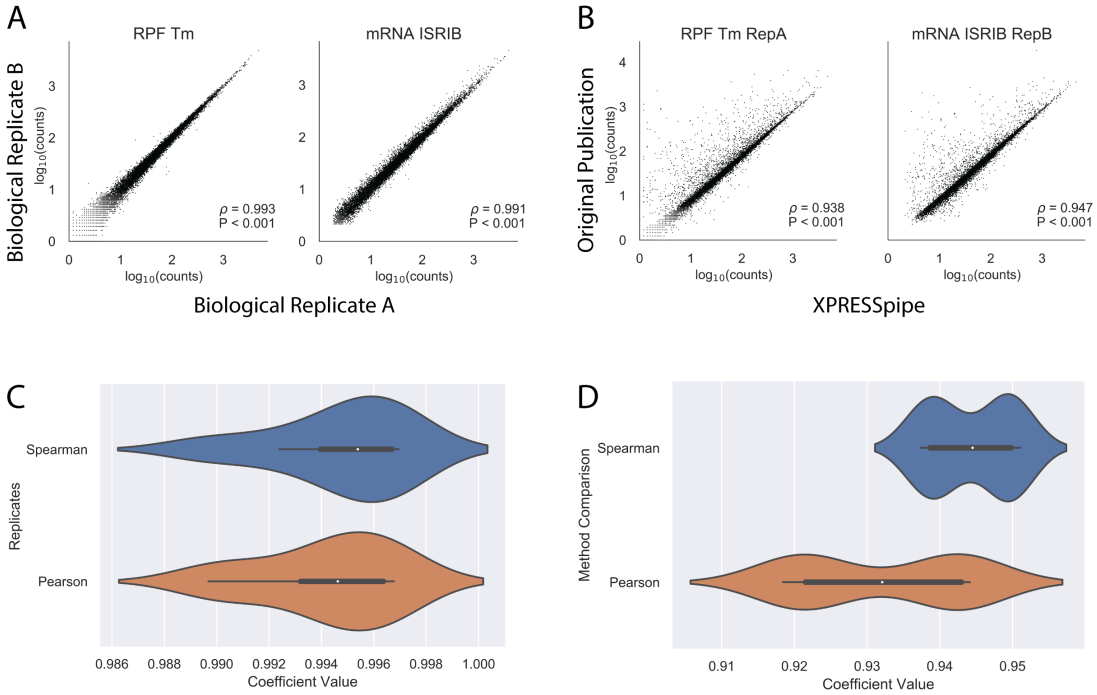

Figure 2: **Representative comparisons between processed data produced by XPRESSpipe and original study.** Genes were eliminated from analysis if any RNA-Seq sample for that gene had fewer than 10 counts. A) Representative comparisons of biological replicate read counts processed by XPRESSpipe. B) Representative comparisons of read counts per gene between count data from the original study and the same raw data processed and quantified by XPRESSpipe. C) Boxplot summaries of Spearman ρ and Pearson r values for biological replicate comparisons. D) Boxplot summaries of Spearman ρ and Pearson r values for between method processing. RPF, ribosome-protected fragments. Tm, tunicamycin. All ρ values reported in A and B are Spearman correlation coefficients using RPM-normalized count data. Pearson correlation coefficients were calculated using $\log_{10}$(rpm(counts) + 1) transformed data. XPRESSpipe-processed read alignments were quantified to *Homo sapiens* build CRCh38v98 using a protein-coding only, truncated GTF.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

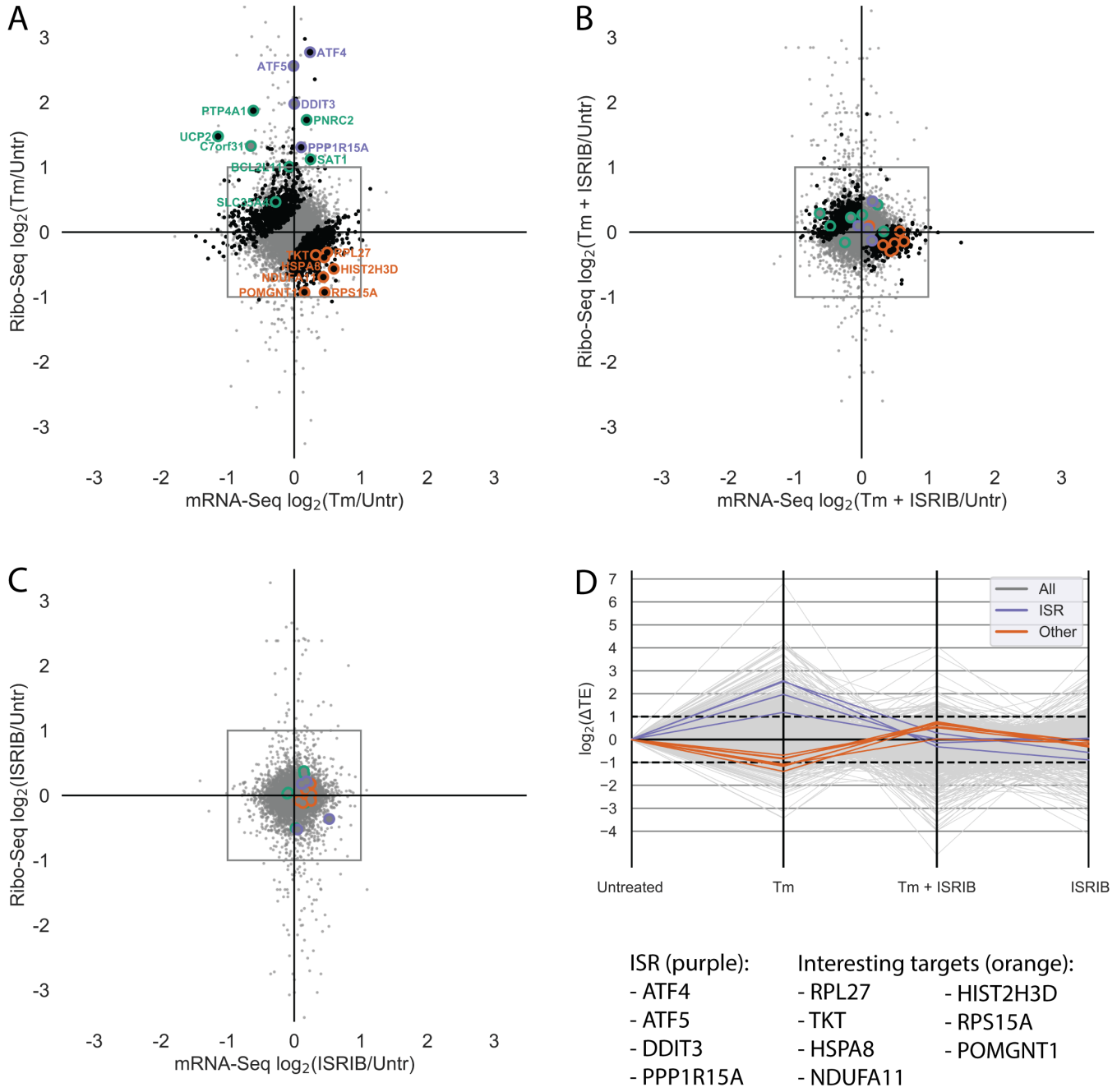

Figure 3: **Analysis of previously published ISR TE data using XPRESSpipe.** A-C) $\log_2$(Fold Change) for each drug condition compared to untreated for the ribo-seq and RNA-Seq data. Purple, ISR canonical targets highlighted in the original study. Green, genes with uORFs affected by ISR as highlighted in the original study. Orange, genes fitting a strict TE thresholding paradigm to identify genes that display a 2-fold or greater increase in TE in Tm + ISRIB treatment compared to Tm treatment. Black, genes with statistically significant changes in TE. Grey, all genes. Changes in ribo-seq and mRNA-Seq were calculated using DESeq2. TE was calculated using DESeq2. Points falling outside of the plotted range are not included. D) Changes in $\log_2$(TE) for each drug condition compared to untreated control. Grey, all genes. Purple, ISR targets identified in the original study. Orange, genes fitting a strict TE thresholding paradigm to identify genes that display a 2-fold or greater increase in TE in Tm + ISRIB treatment compared to Tm treatment. XPRESSpipe-processed read alignments were quantified to *Homo sapiens* build CRCh38v98 using a protein-coding only, truncated GTF.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

focused on in the study. However, re-analysis of these data with the updated XPRESSpipe methodology identifies genes with apparent translational down-regulation that may play a role in the neurodegenerative effects of ISR and the neuroprotective properties of ISRIB [73–76]. Importantly, several of these genes were not identified as having significantly down-regulated TEs in the original analysis, which suggests a rationale for not focusing on translational downregulation. In all, we identified seven genes with the regulatory paradigm of interest: significant decreases in TE during tunicamycin-induced ISR that are restored in the ISR + ISRIB condition (Table 1, descriptions sourced from [80–82] (Figure 3D)). RNA-Seq and ribosome-footprint coverage across these genes show that the significant changes in their TE are due to neither spurious, high-abundance fragments differentially present across libraries nor variance from an especially small number of mapped reads (Figure S7). This is an important consideration as the commonly suggested use of the CircLigase enzyme in published ribosome profiling library preparation protocols, which circularizes template cDNA before sequencing, can bias certain molecules' incorporation into sequencing libraries based on read-end base content alone [83].

Five (POMGNT1, RPL27, TKT, HSPA8, NDUFA11) out of the seven identified genes have annotated neurological functions or mutations that cause severe neurological disorders. Mutations in one other gene (RPS15A) generally result in metabolic disorders. While none of these genes were identified as being of interest in the original study using the original methods, by re-processing the original manuscript count data with DESeq2 [68] and the same expression pattern thresholding, four of these genes are now present in the analysis (RPL27, TKT, HSPA8, RPS15A) (see Figure S6 for a systematic comparison). NDUFA11 and TKT are protein-coding genes whose functions are integrally tied to successful central carbon metabolism and mitochondrial electron transport chain function, respectively. NDUFA11 encodes a subunit of mitochondrial respiratory complex I [84], and TKT encodes a thiamine-dependent enzyme that channels excess sugars phosphates into glycolysis as part of the pentose phosphate pathway [85]. Mutations in NDUFA11 cause severe neurodegenerative phenotypes such as brain atrophy and encephalopathy [84], and mutations in TKT cause diseases associated with neurological phenotypes [86]. These regulatory and phenotypic observations raise the possibility that their role may be functionally relevant to the neurodegenerative effects of ISR and the neuroprotective properties of ISRIB.

While at this stage speculative, it is interesting that the processing of these data with updated methods provides a very conservative list of differentially expressed genes, and that the majority of those genes are associated with severe neurological phenotypes. It is therefore easy to speculate that TE regulation of these targets' abundance might be important in the neurodegeneration observed in prolonged ISR conditions. ISRIB's neuroprotective effects

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: **Translationally down-regulated genes during acute Tm treatment with restored regulation during Tm + ISRIB treatment.** Gene names succeeded by an asterisk indicate these genes were identified in the original data when re-analyzed with DESeq2 [68]. Gene names succeeded by an ampersand indicate genes with strong neurological phenotype annotations. None of these genes were present in the original analysis tables.

Gene Name	Relevant Description
POMGNT1[&]	Participates in O-mannosyl glycosylation. Mutations have been associated with muscle-eye-brain diseases and congenital muscular dystrophies. Expressed especially in astrocytes, as well as in immature and mature neurons. Expressed across brain.
RPL27*[&]	Subunit of ribosome catalyzing protein synthesis. Expressed in cerebral cortex in embryonic tissue and/or stem cells. Mutations associated with Diamond-Blackfan Anemia 16, a metabolic disease, which may present with microcephaly.
TKT*[&]	Encodes thiamine-dependent enzyme that channels excess sugars phosphates to glycolysis. Mutations associated with developmental delays and Wernicke-Korsakoff Syndrome, a metabolic and neuronal disease and associated with encephalopathy and dementia-like characteristics.
HSPA8*[&]	Encodes heat shock protein 70 member. Facilitates protein folding and localization. Diseases associated with mutations include Auditory System Disease and Brain Ischemia, both neurological disorders. Expressed in cerebral cortex in embryonic tissue and/or stem cells.
NDUFA11[&]	Encodes subunit of mitochondrial complex I, a vital component of the electron transport chain. Mutations are associated with severe mitochondrial complex I deficiency. Related pathways include the GABAergic synapse. Associated diseases include brain atrophy, encephalopathy, and leber hereditary optic neuropathy. Overexpressed in frontal cortex.
HIST2H3D	Responsible for nucleosome structure. No neurological phenotypes currently annotated.
RPS15A*	Subunit of ribosome catalyzing protein synthesis. Diseases associated include Diamond-Blackfan Anemia, an inborn error of metabolism disease.

may stem from a restoration of one or more of these entities' protein expression. Though speculative without further validation, these ISRIB-responsive neuronal targets act as interesting cases for further validation and study in a model more representative of neurotoxic injury and disease than the HEK-293T model used in the original study. In all, this comparison demonstrates the utility of XPRESSpipe for rapid, user-friendly analysis and re-analysis of ribosome-profiling experiments in the pursuit of biological insights and hypothesis generation.

Cost Analysis and Performance

XPRESSpipe functions can be computationally intensive. Super-computing resources are recommended, especially when handling large datasets or when aligning to larger, more complex genomes. Many universities provide super-computing resources to their affiliates; however, in cases where these resources are not available, servers such as Amazon Web Services (AWS) [87] can be used to process sequencing data using XPRESSpipe. Table 2 summarizes the runtime statistics for the ISRIB dataset used in this study. The ISRIB ribosome profiling dataset contained a total of 32 raw sequence files that were aligned to *Homo sapiens*; thus it acts as a high-end estimate of the time required to process data with XPRESSpipe. For a comparable dataset, cost to use an AWS computational node with similar

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

specifications for the specified pipeline elapsed time in Table 2 would be approximately 25.76 USD using an Amazon EC2 On-Demand m5.8xlarge (however, significantly reduced rates are available if using Spot instances or by using the free tier) and storage cost would amount to around 17.41 USD/month for all input and output data on Amazon S3 storage (storage costs could be reduced as much of the intermediate data may not be relevent for certain users; however, raw input data should always be archived by the user).

Table 2: **XPRESSpipe sub-module statistics for dataset GSE65778.** `geneCoverage` module performed on high-coverage gene. While some memory footprints are large in this test case, steps will scale based on available user resources. Input raw FASTQ files were uncompressed. The `metagene` and `geneCoverage` sub-modules used a conservative BAM file multiprocessing limit to avoid out-of-memory errors. XPRESSpipe v0.3.1 was used to generate these statistics.

Process	Command	Wallclock Time	Max RAM
Curate STAR reference	`curateReference`	00h 38m 34s	34.03 GB
Truncate GTF	`modifyGTF -t`	00h 03m 40s	03.27 GB
Read Pre-processing	`trim`	00h 08m 50s	00.48 GB
Alignment / Post-processing	`align`	07h 57m 44s	38.03 GB
Read Quantification	`count -c htseq`	03h 13m 04s	00.16 GB
Isoform Abundance	`count -c cufflinks`	00h 56m 44s	02.36 GB
Differential Expression (n=9)	`diffxpress`	00h 07m 50s	00.65 GB
Read Distributions	`readDistribution`	00h 05m 00s	00.28 GB
Metagene Analysis	`metagene`	01h 45m 35s	35.52 GB
Gene Coverage (n=1)	`geneCoverage`	01h 24m 00s	19.32 GB
Periodicity	`periodicity`	01h 08m 22s	54.16 GB
Library Complexity	`complexity`	01h 02m 57s	01.52 GB
rRNA probe	`rrnaProbe`	00h 00m 55s	00.15 GB
Pipeline	`riboseq`	16h 46m 19s	54.16 GB

Attribute	Value
Total Raw Input	257 GB
Total Output	500 GB
Allocated Virtual CPUs	32
Minimum Allocated Memory	62.50 GB

Availability and Future Directions

We have described the software suite, XPRESSyourself, an automated reference implementation of best-practices in ribosome profiling data analysis built upon a synthesis of new tools, old tools, and pipelines. XPRESSyouself is perpetually open source and protected under the GPL-3.0 license. Updates to the software are version controlled and maintained on GitHub [52]. Jupyter notebooks and video walkthroughs are included within the README files at [52]. Documentation is hosted on readthedocs [88] at [50] and [51]. Source code for associated analyses and figures for this manuscript can be accessed at [78]. The data used in this manuscript are available under the Gene

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Expression Omnibus persistent identifier GSE65778 [89] for ribosome profiling data and under the dbGaP Study Accession persistent identifier phs000178 [90] for the TCGA data.

Although RNA-Seq technologies are quite advanced, standardized computational protocols are far less established for ribosome profiling. As we discussed in this manuscript, this becomes problematic when individuals or groups are not using best practices in analysis or may not be aware of particular biases or measures of quality control required to produce reliable, high-quality sequencing analyses. XPRESSpipe handles these issues through on-going curation of benchmarked software tools and by simplifying the required user input. It also outputs all necessary quality control metrics so that the user can quickly assess the reliability of their data and identify any systematic problems or technical biases that may compromise their analysis. Video walkthroughs, example scripts, and interactive command builders are available within this software suite to make these analyses accessible to experienced and inexperienced users alike. XPRESSyourself will enable individuals and labs to process and analyze their own data, which will result in quicker turnaround times of experiments and immediate financial savings.

One particular benefit of XPRESSyourself is that it consolidates, streamlines, and introduces many tools specific to ribosome profiling processing and analysis. This includes curating GTF files with $5'$- and $3'$- truncated CDS annotations, rRNA probe design for subtractive hybridization of abundant rRNA contaminants, automated quality-control analysis and summarization to report ribosome footprint periodicity, metagene coverage, and intron-less gene coverage profiles. These tools will help to democratize aspects of ribosome profiling analysis for which software have not been previously publicly available or difficult to access.

We demonstrated the utility of the XPRESSyourself toolkit by re-analyzing a publicly available ribosome profiling dataset. From this analysis, we identified putative translational regulatory targets of the integrated stress response (ISR) that may contribute to its neurodegenerative effects and their rescue by the small-molecule ISR inhibitor, ISRIB. This highlights the importance of re-analyzing published datasets with more current methods, as improved analysis methodologies and updated organism genome references may result in improved interpretations and hypotheses.

With the adoption of this flexible pipeline, the field of high-throughput sequencing, particularly ribosome profiling, can continue to standardize the processing protocol for associated sequence data and eliminate the variability that comes from the availability of a variety of software packages for various steps during sequence read processing. Additionally, XPRESSpipe consolidates various tools used by the ribosome profiling and RNA-Seq communities into a single, end-to-end pipeline. With these tools, genome reference formatting and curation are automated and accessible to the public. Adoption of this tool will allow scientists to process and access their data independently

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3: **Software Description**

Project Name	XPRESSyourself
Project Home Page	https://github.com/XPRESSyourself
Archived Versions DOIs	10.5281/zenodo.3338669, 10.5281/zenodo.3337897
Operating Systems	macOS, Linux, centOS
Programming Languages	Python, R, Julia
Other Requirements	Anaconda
License	GNU General Public License v3.0

and quickly, guide them in understanding key considerations in processing their data, and standardize protocols for ribosome profiling and other RNA-Seq applications, thus increasing reproducibility in sequencing analyses.

List of Abbreviations

AWS - Amazon Web Services, BAM - Binary Sequence Alignment Map, BED - Browser Extensible Data, cDNA - complementary DNA, CDS - coding sequence of gene, ChIP-seq - chromatin immunoprecipitation sequencing, CPU - central processing unit, dbGaP - Database of Genotypes and Phenotypes, DNA - deoxyribonucleic acid, FDR - false discovery rate, FPKM - fragments per kilobase of transcript per million, GEO - Gene Expression Omnibus, GTF - General Transfer Format, IGV - Integrative Genomics Viewer, ISR - integrated stress response, ISRIB - ISR inhibitor, mRNA - messenger RNA, nt - nucleotide, PCA - principal component analysis, PCR - polymerase chain reaction, RAM - random access memory, RNA - ribonucleic acid, RNA-Seq - RNA sequencing RPKM - reads per kilobase of transcript per million, RPM - reads per million, rRNA - ribosomal RNA, TCGA - The Cancer Genome Atlas, TE - translation efficiency, TPM - transcripts per million, UMI - unique molecular identifier, UTR - untranslated region, VCF - Variant Call Format

Ethics Approval and Consent to Participate

Protected TCGA data were obtained through dbGaP project number 21674 and utilized according to the associated policies and guidelines.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Consent for Publication

Protected TCGA data were obtained through dbGaP project number 21674 and utilized according to the associated policies and guidelines.

Competing Interests

The authors declare that they have no competing interests.

Funding

J.A.B. received support from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Inter-disciplinary Training Grant T32 Program in Computational Approaches to Diabetes and Metabolism Research, 1T32DK11096601 to Wendy W. Chapman and Simon J. Fisher. J.T.M. received support as an HHMI Fellow of the Jane Coffin Childs Memorial Fund for Medical Research. A.J.B received support from the National Cancer Institute (NCI) Predoctoral to Postdoctoral Fellow Transition Award, K00CA212445. This work was supported by NIDDK fellowship 1T32DK11096601 (to J.A.B.) and NIH grant R35GM13185 (to J.R.). The computational resources used were partially funded by the NIH Shared Instrumentation Grant 1S10OD021644-01A1.

Contributions

J.A.B. conceptualized and administered the project; performed all investigation, analysis, visualization, and data curation; provisioned resources; and wrote the original draft of this manuscript. J.A.B., J.R.B., J.T.M., and J.G. and developed the methodology. J.A.B. and J.R.B. designed and wrote the software. J.A.B., J.T.M., A.J.B., and Y.O. performed software validation. J.A.B. and J.R. acquired funding. J.R., A.R.Q., and J.G. supervised the study. All authors reviewed and edited this manuscript.

Acknowledgments

The authors wish to thank Michael T. Howard for helpful discussions concerning ribosome profiling and sequencing analysis. The authors also wish to thank Mark E. Wadsworth, Ryan Miller, and Michael J. Cormier for helpful

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 4: **Author ORCIDs**

Author	ORCID
J.A.B.	0000-0002-5096-0558
J.R.B.	0000-0001-5470-8299
J.T.M.	0000-0002-3017-8665
Y.O.	0000-0001-9523-1044
A.J.B.	0000-0003-2273-8922
A.R.Q.	0000-0003-1756-0859
J.G.	0000-0001-7568-6789
J.R.	0000-0002-2710-9765

discussions on pipeline design. They also wish to thank T. Cameron Waller for helpful discussions related to pipeline design and biological analysis. The support and resources from the Center for High-Performance Computing at the University of Utah are gratefully acknowledged. The results published here are in whole or part based upon data generated by the TCGA Research Network [62].

References

[1] N. Ingolia, S. Ghaemmaghami, J. Newman, J. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218 (2009). Available from: https://doi.org/10.1126/science.1168978.

[2] G. Brar, J. Weissman. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* **16**, 651 (2015). Available from: https://doi.org/10.1038/nrm4069.

[3] N. McGlincy, N. Ingolia. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112 (2017). Available from: https://doi.org/10.1016/j.ymeth.2017.05.028.

[4] M. Gerashchenko, V. Gladyshev. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res* **42** (2014). Available from: https://doi.org/10.1093/nar/gku671.

[5] A. Bartholomäus, C. D. Campo, I. Z. Mapping the non-standardized biases of ribosome profiling. *Biol Chem* **397** (2016). Available from: https://doi.org/https://doi.org/10.1515/hsz-2015-0197.

[6] J. Hussmann, S. Patchett, A. Johnson, S. Sawyer, W. Press. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet* **11** (2015). Available from: https://doi.org/https://doi.org/10.1371/journal.pgen.1005732.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[7] A. Diament, T. Tuller. Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biol Direct* **11** (2016). Available from: https://doi.org/https://doi.org/10.1186/s13062-016-0127-4.

[8] M. Gerashchenko, V. Gladyshev. Ribonuclease selection for ribosome profiling. *Nucleic Acids Res* **45** (2017). Available from: https://doi.org/https://doi.org/10.1093/nar/gkw822.

[9] C. Sidrauski1, A. McGeachy, N. Ingolia, P. Walter. The small molecule ISRIB reverses the effects of eIF2ã phosphorylation on translation and stress granule assembly. *eLife* (2015). Available from: https://doi.org/10.7554/eLife.05033.

[10] F. Mohammad, C. Woolstenhulme, R. Green, A. Buskirk. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep* **14** (2016). Available from: https://doi.org/https://doi.org/10.1016/j.celrep.2015.12.073.

[11] G. Li, E. Oh, J. Weissman. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484** (2012). Available from: https://doi.org/https://doi.org/10.1038/nature10965.

[12] A. Lecanda, *et al.*. Dual randomization of oligonucleotides to reduce the bias in ribosome-profiling libraries. *Methods* **107** (2016). Available from: https://doi.org/https://doi.org/10.1016/j.ymeth.2016.07.011.

[13] X. Gao, *et al.*. Quantitative profiling of initiating ribosomes in vivo. *Nat Methods* **12** (2015). Available from: https://doi.org/https://doi.org/10.1038/nmeth.3208.

[14] E. Afgan, *et al.*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537 (2018). Available from: https://doi.org/10.1093/nar/gky379.

[15] A. Michel, *et al.*. RiboGalaxy: A browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA Biol* **13**, 316 (2016). Available from: https://doi.org/10.1080/15476286.2016.1141862.

[16] Nextflow. `https://www.nextflow.io/example4.html`.

[17] DNAnexus. `https://github.com/dnanexus/tophat_cufflinks_rnaseq`.

[18] O. Carja, T. Xing, E. Wallace, J. Plotkin, P. Shah. riboviz: analysis and visualization of ribosome profiling datasets. *BMC Bioinformatics* **18** (2017). Available from: https://doi.org/10.1186/s12859-017-1873-8.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[19] C. Artieri, H. Fraser. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res* **24**, 2011 (2014). Available from: https://doi.org/10.1101/gr.175893.114.

[20] J. Hussmann, S. Patchett, A. Johnson, S. Sawyer, W. Press. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet* **11** (2015). Available from: https://doi.org/10.1371/journal.pgen.1005732.

[21] D. Weinberg, *et al.*. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep* **14**, 1787 (2016). Available from: https://doi.org/10.1016/j.celrep.2016.01.043.

[22] J. Robinson, *et al.*. Integrative Genomics Viewer. *Nat Biotechnol* **29**, 24 (2011). Available from: https://doi.org/10.1038/nbt.1754.

[23] Z. Costello, H. Martin. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ Syst Biol Appl* **4** (2018). Available from: https://doi.org/10.1038/s41540-018-0054-3.

[24] V. Funari, S. Canosa. The Importance of Bioinformatics in NGS: Breaking the Bottleneck in Data Interpretation. *Science* **344**, 653 (2014). Available from: https://doi.org/10.1126/science.344.6184.653-c.

[25] R. Kumari, A. Michel, P. Baranov. PausePred and Rfeet: webtools for inferring ribosome pauses and visualizing footprint density from ribosome profiling data. *RNA* **24** (2018). Available from: https://doi.org/10.1261/rna.065235.117.

[26] C. Oertlin, *et al.*. Generally applicable transcriptome-wide analysis of translation using anota2seq. *Nucleic Acids Res* **47** (2019). Available from: https://doi.org/10.1093/nar/gkz223.

[27] A. Popa, *et al.*. RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Res* **5** (2016). Available from: https://doi.org/10.12688/f1000research.8964.1.

[28] W. Li, W. Wang, P. Uren, L. Penalva, A. Smith. Riborex: fast and flexible identification of differential translation from Ribo-seq data. *Bioinformatics* **33** (2017). Available from: https://doi.org/10.1093/bioinformatics/btx047.

[29] S. Verbruggen, G. Menschaert. mQC: A post-mapping data exploration tool for ribosome profiling. *Comput Methods Programs Biomed* (2018). Available from: https://doi.org/10.1016/j.cmpb.2018.10.018.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[30] Å. Birkeland, K. ChyŻyńska, E. Valen. Shoelaces: an interactive tool for ribosome profiling processing and visualization. *BMC Genomics* **19** (2018). Available from: https://doi.org/10.1186/s12864-018-4912-6.

[31] N. Davidson, A. Hawkins, A. Oshlack. SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biol* **18** (2017). Available from: https://doi.org/10.1186/s13059-017-1284-1.

[32] T. Backman, T. Girke. systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics* **17** (2016). Available from: https://doi.org/10.1186/s12859-016-1241-0.

[33] H. Tjeldnes, K. Labun. ORFik: Open Reading Frames in Genomics. `https://github.com/JokingHero/ORFik` (2017). Available from: https://doi.org/10.18129/B9.bioc.ORFik.

[34] T. Martin, I. Erte, P. Tsai, J. Bell. coMET: an R plotting package to visualize regional plots of epigenome-wide association scan results. *QG14* (2014). Available from: http://quantgen.soc.srcf.net/qg14/.

[35] T. Martin, I. Yet, P. Tsai, J. Bell. coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC Bioinformatics* **16** (2015). Available from: https://doi.org/10.1186/s12859-015-0568-2.

[36] T. Hardcastle. riboSeqR. Available from: https://doi.org/10.18129/B9.bioc.riboSeqR.

[37] F. Ramírez, F. Dündar, S. Diehl, B. Grüning, T. Manke. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42** (2014). Available from: https://doi.org/10.1093/nar/gku365.

[38] Picard. `https://broadinstitute.github.io/picard/`.

[39] S. Zhang, *et al.*. Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning. *Cell Syst* **5** (2017). Available from: https://doi.org/10.1016/j.cels.2017.08.004.

[40] P. O'Connor, D. Andreev, P. Baranov. Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat Commun* **7** (2016). Available from: https://doi.org/10.1038/ncomms12915.

[41] Z. Xiao, Q. Zou, Y. Liu, X. Yang. Genome-wide assessment of differential translations with ribosome profiling data. *Nat Commun* **7** (2016). Available from: https://doi.org/10.1038/ncomms11194.

[42] Y. Zhong, *et al.*. RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* **33** (2017). Available from: https://doi.org/10.1093/bioinformatics/btw585.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[43] L. Calviello, *et al.*. Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* **13** (2016). Available from: https://doi.org/10.1038/nmeth.3688.

[44] H. Wang, J. McManus, C. Kingsford. Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. *Bioinformatics* **32** (2016). Available from: https://doi.org/10.1093/bioinformatics/btw085.

[45] P. Spealman, H. Wang, G. May, C. Kingsford, C. McManus. Exploring Ribosome Positioning on Translating Transcripts with Ribosome Profiling. *Methods Mol Biol* **1358** (2016). Available from: https://doi.org/10.1007/978-1-4939-3067-8_5.

[46] J. Dunn, J. Weissman. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics* **17** (2016). Available from: https://doi.org/https://doi.org/10.1186/s12864-016-3278-x.

[47] P. Perkins, S. Mazzoni-Putman, A. Stepanova, J. Alonso, S. Heber. RiboStreamR: a web application for quality control, analysis, and visualization of Ribo-seq data. *BMC Genomics* **20** (2019). Available from: https://doi.org/https://doi.org/10.1186/s12864-019-5700-7.

[48] H. Fang, *et al.*. Scikit-ribo Enables Accurate Estimation and Robust Modeling of Translation Dynamics at Codon Resolution. *Cell Syst* **6** (2018). Available from: https://doi.org/https://doi.org/10.1016/j.cels.2017.12.007.

[49] S. Chun, C. Rodriquez, P. Todd, R. Mills. SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data. *BMC Bioinformatics* **17** (2016). Available from: https://doi.org/https://doi.org/10.1186/s12859-016-1355-4.

[50] XPRESSpipe documentation. `https://xpresspipe.readthedocs.io/en/latest/`.

[51] XPRESSplot documentation. `https://xpressplot.readthedocs.io/en/latest/`.

[52] XPRESSyourself. `https://github.com/XPRESSyourself/`.

[53] A. Dobin, *et al.*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15 (2013). Available from: https://doi.org/10.1093/bioinformatics/bts635.

[54] G. Baruzzo, *et al.*. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods* **14**, 135 (2017). Available from: https://doi.org/10.1038/nmeth.4106.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[55] I. Raplee, A. Evsikov, C. M. de Evsikova. Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research. *J Pers Med* **9** (2019). Available from: https://doi.org/10.3390/jpm9020018.

[56] T. Tuller, H. Zur. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res* **42** (2015). Available from: https://doi.org/https://doi.org/10.1093/nar/gku1313.

[57] S. Anders, P. Pyl, W. Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166 (2015). Available from: https://doi.org/10.1093/bioinformatics/btu638.

[58] C. Trapnell, *et al.*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7** (2012). Available from: https://doi.org/10.1038/nprot.2012.016.

[59] Y. Fu, P. Wu, T. Beane, P. Zamore, Z. Weng. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19** (2018). Available from: https://doi.org/10.1186/s12864-018-4933-1.

[60] T. Smith, A. Heger, I. Sudbery. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27** (2017). Available from: https://doi.org/10.1101/gr.209601.116.

[61] C. Robert, M. Watson. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol* **16** (2015). Available from: https://doi.org/10.1186/s13059-015-0734-x.

[62] The Cancer Genome Atlas. `https://portal.gdc.cancer.gov`.

[63] C. Evans, J. Hardin, D. Stoebel. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform* **19**, 776–792 (2018). Available from: https://doi.org/10.1093/bib/bbx008.

[64] J. Leek, W. Johnson, H. Parker, A. Jaffe, J. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28** (2012). Available from: https://doi.org/10.1093/bioinformatics/bts034.

[65] C. Wu, B. Zinshteyn, K. Wehner, R. Green. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. *Mol Cell* **73** (2019). Available from: https://doi.org/10.1016/j.molcel.2018.12.009.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[66] S. Sayols, D. Scherzinger, H. Klein. dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC Bioinformatics* **17** (2016). Available from: https://doi.org/10.1186/s12859-016-1276-2.

[67] F. Lauria, *et al.*. riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. *PLoS Comput Biol* **14** (2018). Available from: https://doi.org/10.1371/journal.pcbi.1006169.

[68] M. Love, W. Huber, S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15** (2014). Available from: https://doi.org/10.1186/s13059-014-0550-8.

[69] H. Harding, *et al.*. An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. *Mol Cell* **11** (2003). Available from: https://doi.org/10.1016/S1097-2765(03)00105-9.

[70] D. Santos-Ribeiro, L. Godinas, C. Pilette, F. Perros. The integrated stress response system in cardiovascular disease. *Drug Discov Today* **23** (2018). Available from: https://doi.org/10.1016/j.drudis.2018.02.008.

[71] H. Rabouw, *et al.*. Small molecule ISRIB suppresses the integrated stress response within a defined window of activation. *Proc Natl Acad Sci U S A* **116** (2019). Available from: https://doi.org/10.1073/pnas.1815767116.

[72] J. Tsai, *et al.*. Structure of the nucleotide exchange factor eIF2B reveals mechanism of memory-enhancing molecule. *Science* **359** (2018). Available from: https://doi.org/10.1126/science.aaq0939.

[73] A. Choua, *et al.*. Inhibition of the integrated stress response reverses cognitive deficits after traumatic brain injury. *Proc Natl Acad Sci U S A* **114** (2017). Available from: https://doi.org/10.1073/pnas.1707661114.

[74] M. Halliday, *et al.*. Partial restoration of protein synthesis rates by the small molecule ISRIB prevents neurodegeneration without pancreatic toxicity. *Cell Death Dis* **6** (2015). Available from: https://doi.org/10.1038/cddis.2015.49.

[75] C. Sidrauski, *et al.*. Pharmacological brake-release of mRNA translation enhances cognitive memory. *Elife* **2** (2013). Available from: https://doi.org/10.7554/eLife.00498.

[76] Y. Sekine, *et al.*. Stress responses. Mutations in a translation initiation factor identify the target of a memory-enhancing compound. *Science* **348** (2015). Available from: https://doi.org/10.1126/science.aaa6986.

[77] D. Kim, *et al.*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14** (2013). Available from: https://doi.org/10.1186/gb-2013-14-4-r36.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[78] Manuscript code. https://github.com/XPRESSyourself/xpressyourself_manuscript/tree/master/supplemental_files. Available from: https://doi.org/DOI: 10.5281/zenodo.3337599.

[79] S. Anders, W. Huber. Differential expression analysis for sequence count data. *Genome Biol* **11** (2010). Available from: https://doi.org/10.1186/gb-2010-11-10-r106.

[80] GeneCards. https://www.genecards.org/. Accessed 20 October 2019.

[81] National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/. Accessed 20 October 2019.

[82] UniProt. https://www.uniprot.org/uniprot/. Accessed 20 October 2019.

[83] R. Tunney, *et al.*. Accurate design of translational output by a neural network model of ribosome distribution. *Nat Struct Mol Biol* **25**, 577 (2018). Available from: https://doi.org/10.1038/s41594-018-0080-2.

[84] I. Berger, *et al.*. Mitochondrial complex I deficiency caused by a deleterious NDUFA11 mutation. *Ann Neurol* **63** (2008). Available from: https://doi.org/https://doi.org/10.1002/ana.21332.

[85] L. Mitschke, *et al.*. The crystal structure of human transketolase and new insights into its mode of action. *J Biol Chem* **285** (2010). Available from: https://doi.org/https://doi.org/10.1074/jbc.M110.149955.

[86] L. Boyle, *et al.*. The crystal structure of human transketolase and new insights into its mode of action. *Am J Hum Genet* **98** (2016). Available from: https://doi.org/https://doi.org/10.1016/j.ajhg.2016.03.030.

[87] Amazon Web Services. https://aws.amazon.com.

[88] Read the Docs. https://readthedocs.org/.

[89] Ribosome Profiling GEO Accession. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65778.

[90] TCGA dbGaP Accession. (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v10.p8).

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

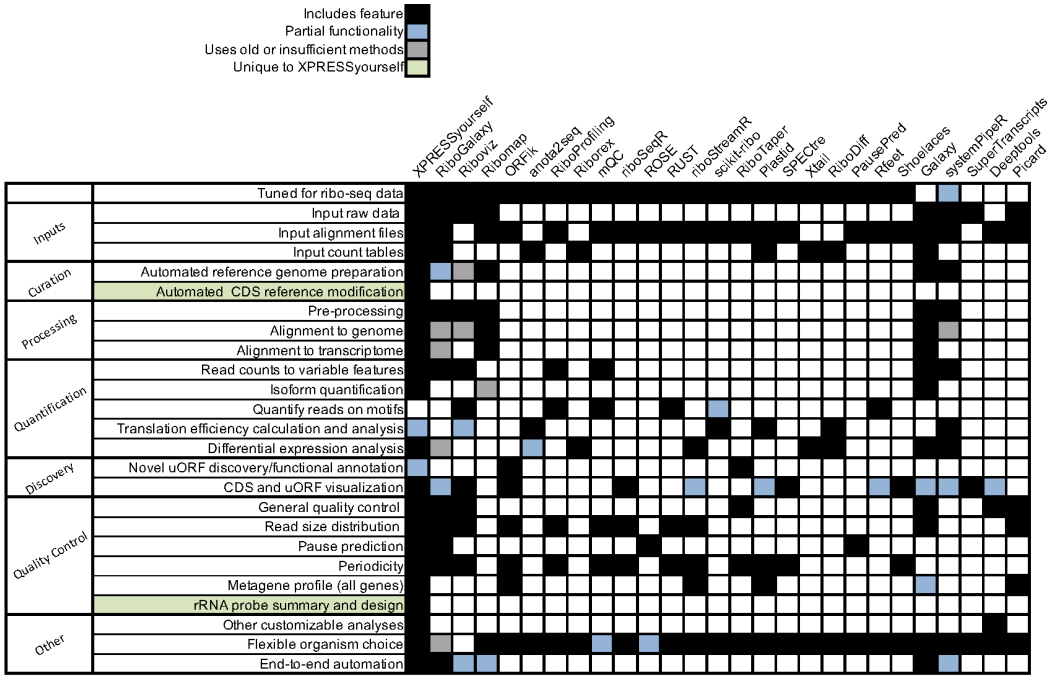

Figure S1: **Comparison between XPRESSyourself and other available software packages for ribosome profiling data analysis.** Black boxes indicate full functionality, blue boxes indicate partial functionality, grey boxes indicate incomplete or outdated functionality, and blank boxes indicate no functionality for the specified task. Rankings were compiled using the tools' documentation, manuscript, and codebase. If a function was not clearly described in any of these venues, a blank box was given.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

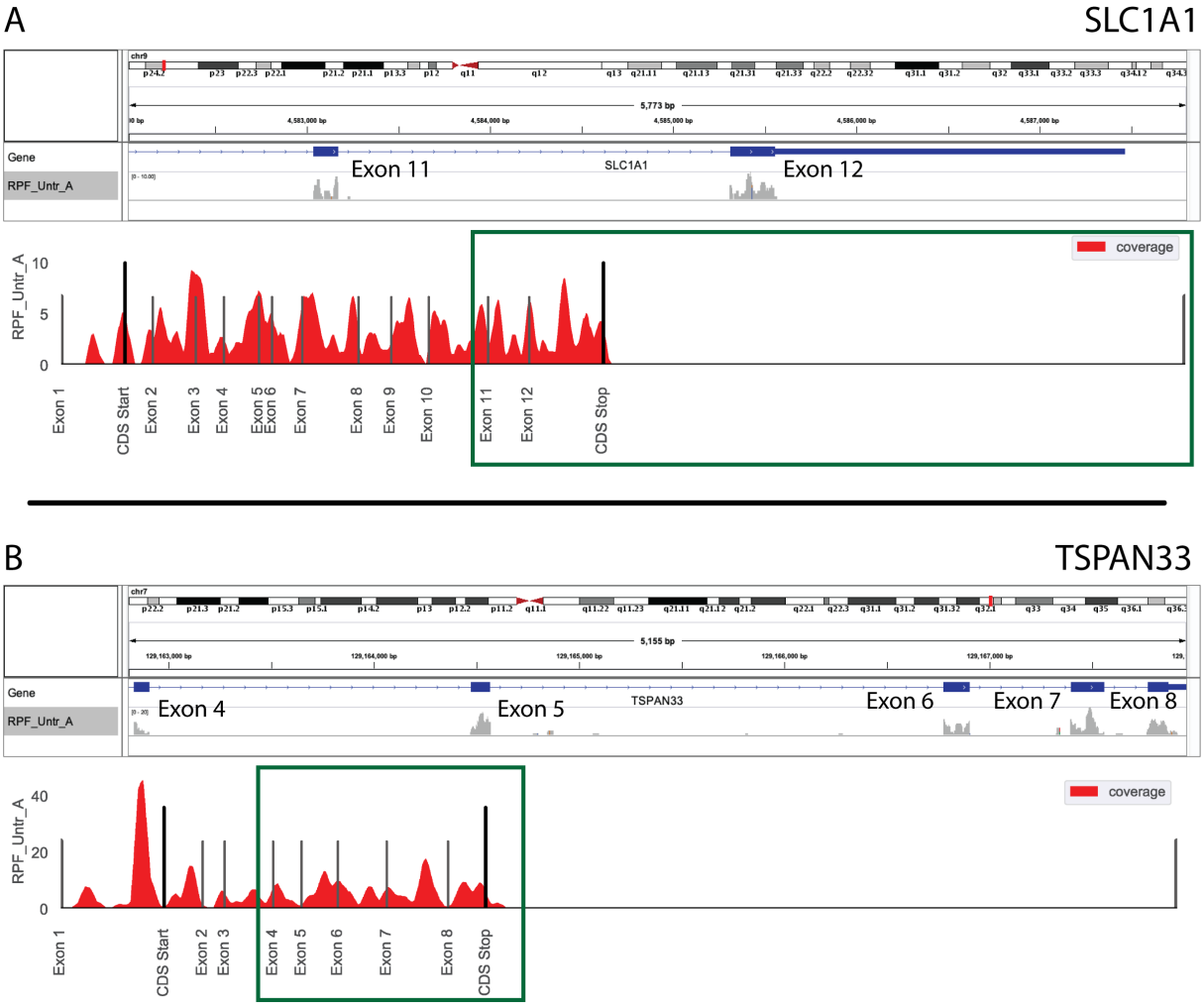

Figure S2: **Comparison between IGV browser and geneCoverage output.** A) Gene coverage from IGV (above) and XPRESSpipe (below) for SLC1A1. B) Gene coverage from IGV (above) and XPRESSpipe (below) for TSPAN33. Introns collapsed by XPRESSpipe. Green box, region displayed in corresponding IGV window.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

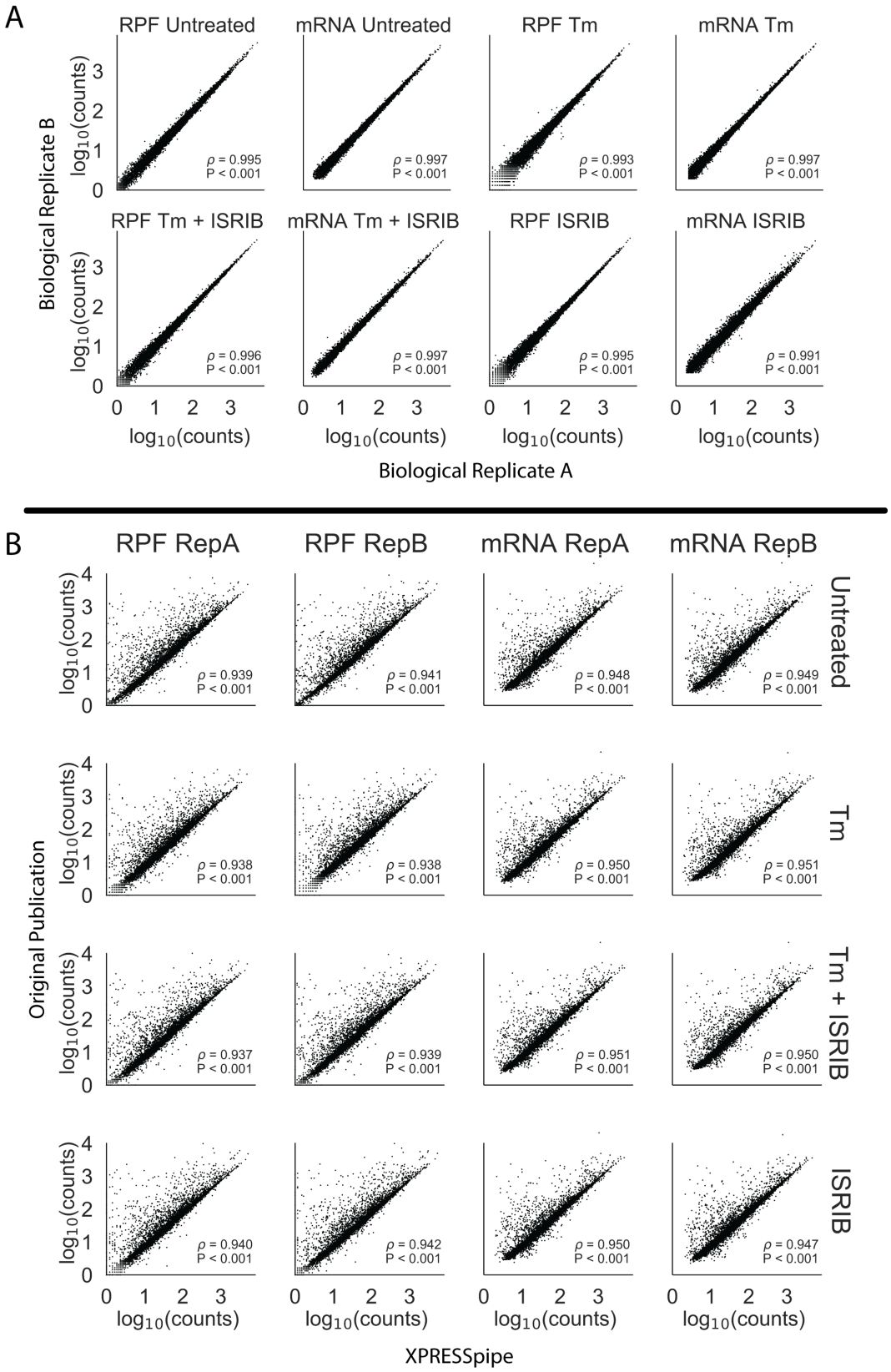

Figure S3: **Comparison between processed data produced by XPRESSpipe and original study.** Genes were eliminated from analysis if any RNA-Seq sample for that gene had fewer than 10 counts. A) Comparison of biological replicate read counts processed by XPRESSpipe. B) Comparison of read counts per gene between count data from the original study and the same raw data processed and quantified by XPRESSpipe. RPF, ribosome-protected fragments. Tm, tunicamycin. All ρ values reported are Spearman correlation coefficients. XPRESSpipe-processed read alignments were quantified to *Homo sapiens* build CRCh38v98 using a protein-coding only, truncated GTF.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

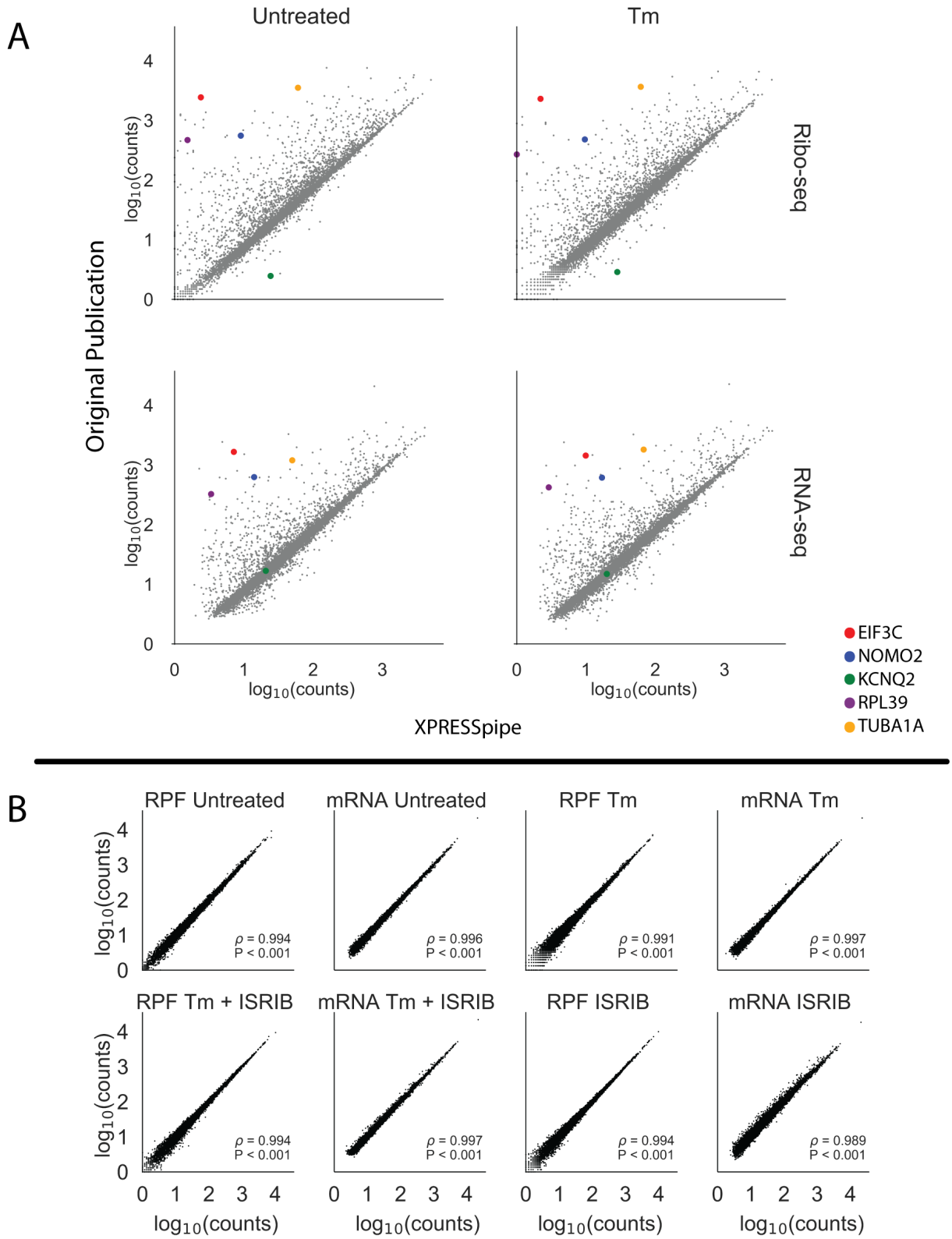

Figure S4: **Original ISRIB count data plotted against XPRESSpipe-processed data reveals systematic differences between the analytical regimes.** A) Selected highlighted genes show consistent differences between processing methods. B) Spearman correlation plots using the data table provided as supplementary data with the original ISRIB manuscript comparing biological replicates. RPF, ribosome-protected footprint. Tm, tunicamycin. All ρ values reported are Spearman correlation coefficients.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

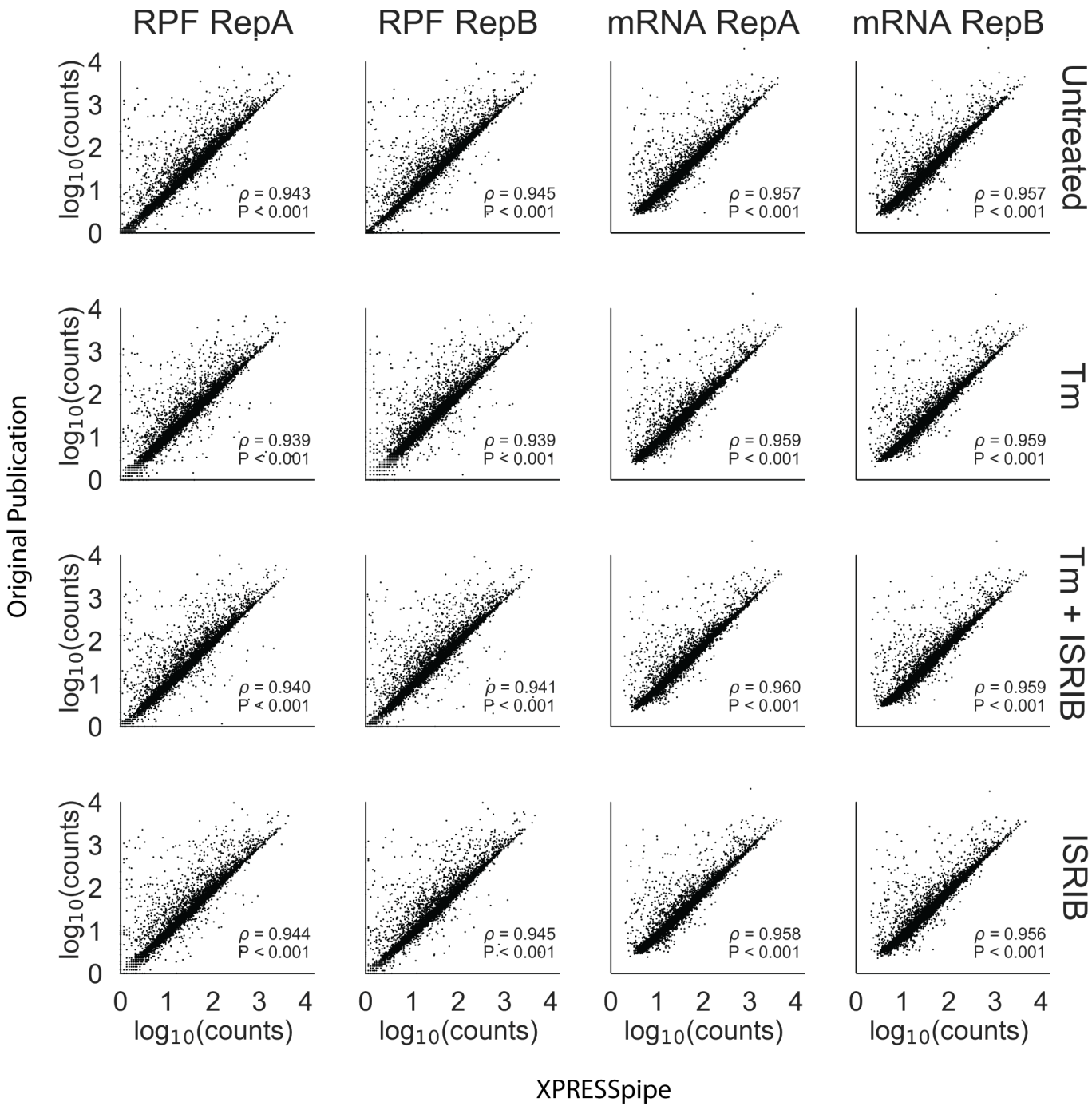

Figure S5: **Original ISRIB count data plotted against XPRESSpipe-processed data quantified using same reference version reveals mild improvement in comparability between the analytical regimes.** Original samples were processed using Ensembl human build GRCh38v72, as in the original manuscript, and compared with the original count data provided with the manuscript. XPRESSpipe-prepared counts were thresholded similarly as the original data (each gene needed to have at least 10 counts across all mRNA samples). RepA, biological replicate A. RepB, biological replicate B. RPF, ribosome-protected footprint. Tm, tunicamycin. All ρ values reported are Spearman correlation coefficients.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

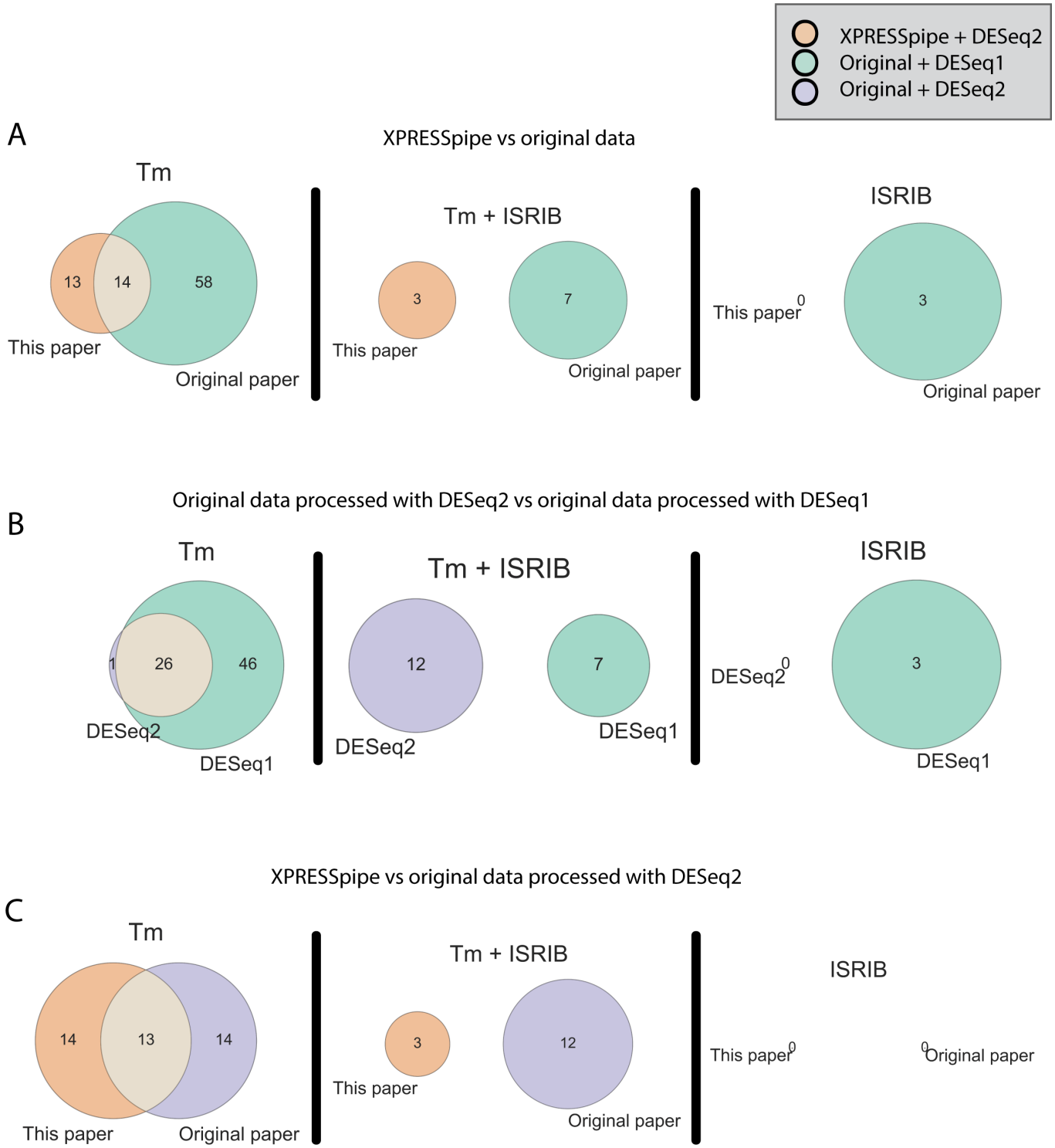

Figure S6: **Cross-method analysis comparisons.** A) XPRESSpipe-processed data (orange) versus data as originally presented within original manuscript using original methods (green). B) Comparison of analyses using provided count table in original publication using DESeq2 (purple) versus original analysis provided in manuscript using DESeq1 (green). C) XPRESSpipe-processed (orange) versus originally-processed data (purple), both using DESeq2 for differential expression analysis. Tan regions indicate overlap between gene lists. Thresholds used were the same as those used in the original study: $|\log_2(\text{Fold Change})| > 1$, FDR < 0.1.

bioRxiv preprint doi: https://doi.org/10.1101/704320; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

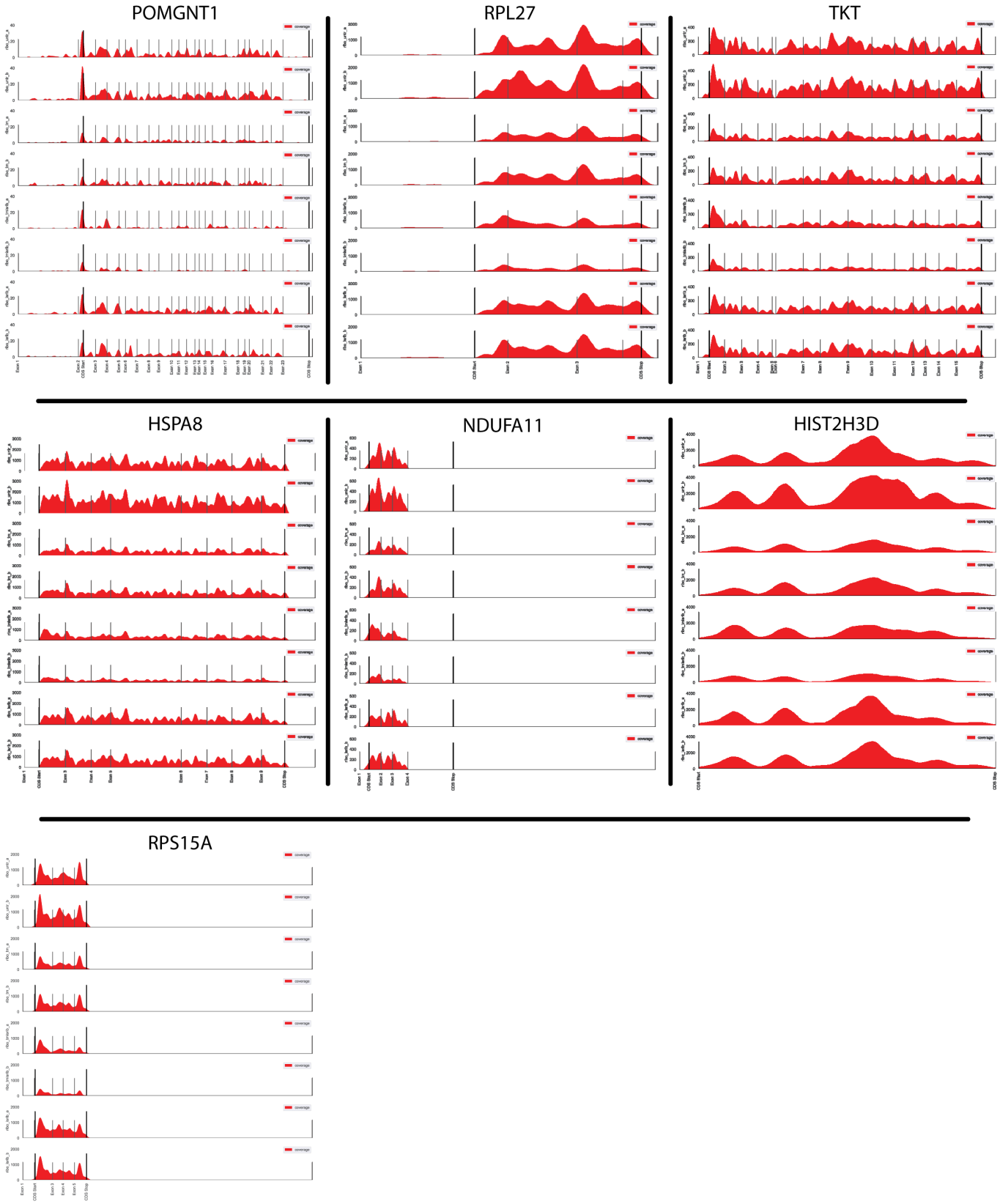

Figure S7: **Gene coverage plots for neurologically annotated genes passing strict thresholding.** Coverage plots were generated using XPRESSpipe's geneCoverage module, which collapses introns within the representation.