

Comparing RNA & Protein Abundance

M Gerstein
&
P Emani

Outline: Comparing Protein & RNA Abundance

- **Past Context:**
to work in the Center
 - Quantifying the moderate **statistical correlation between protein & RNA**
 - PARE server
- **EMpire** (Current result)
 - Leveraging the correlation to **better assign peptides to isoforms**
 - EM algorithm better assigns **dominant isoforms**, with greater interpretability
- **uORFs** (Current result)
 - Affect translation & relationship between protein & RNA
 - Feature integration to find **small subset of uORFs that most alter translation**
- **Future Direction:**
Protein v RNA using matched samples in the Brainspan dataset + single-cell data

Why relate amounts of protein & mRNA?

Gene expression -
major place for **regulation**
(easy to measure)

VS.

Concentration of protein -
major determinant of **activity**

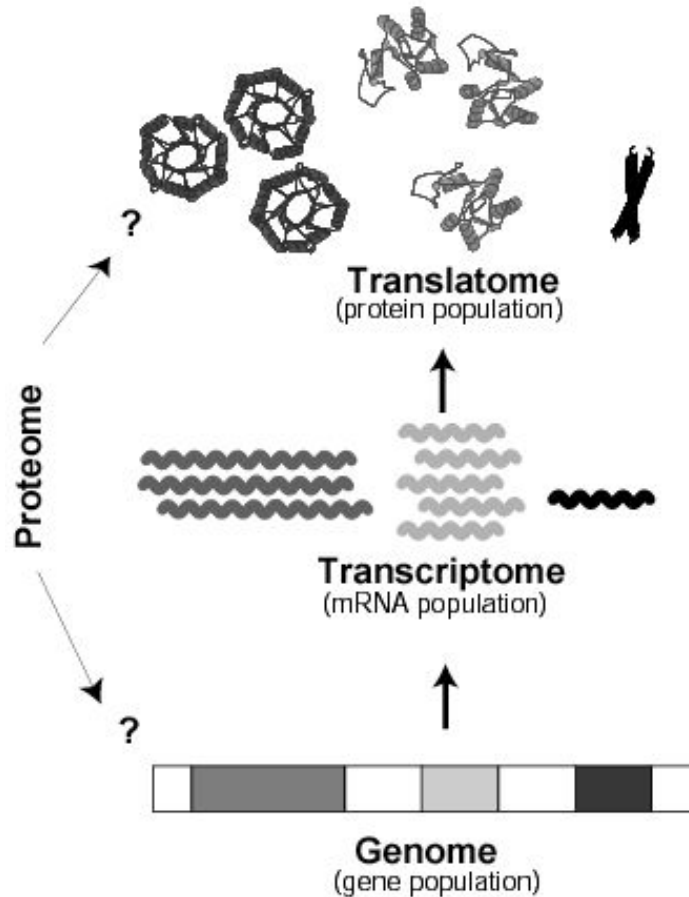
Expectations from simple kinetic models:

$$\frac{dP_i}{dt} = k_{s,i} [\text{mRNA}_i] - k_{d,i} P_i$$

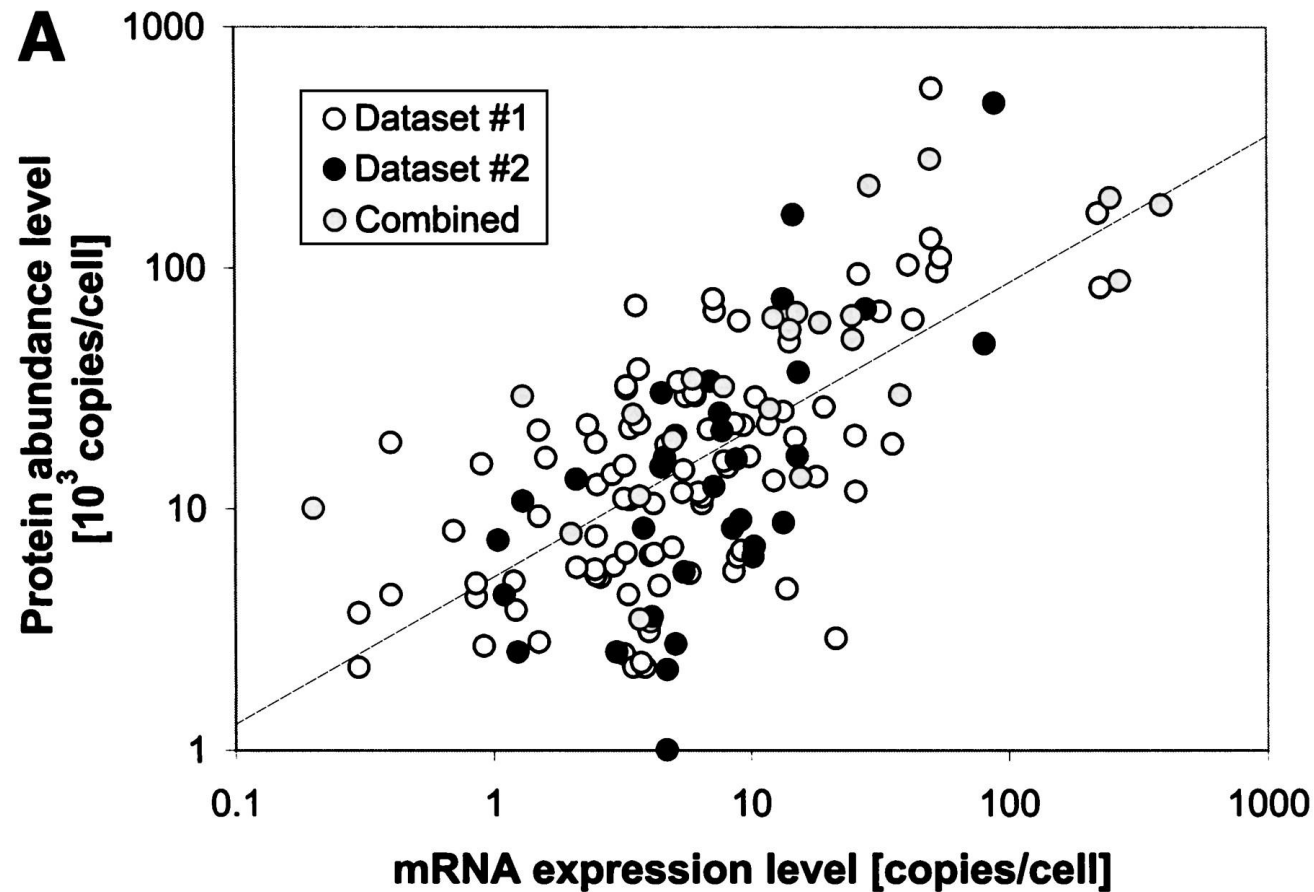
At steady state: $P_i = \frac{k_{s,i} [\text{mRNA}_i]}{k_{d,i}}$

where $k_{s,i}$ and $k_{d,i}$ are the protein synthesis and degradation rate constants

Outliers from trend interesting



Early result on mRNA vs Protein, using 2D gels



[Greenbaum et al. *Bioinformatics* 2002, 18, 587]

PARE

proteomics.gersteinlab.org
PARE: Protein Abundance and mRNA Expression Correlation Tool

Choose datasets [help](#)

1. **Select organism:** Yeast
[required if selecting datasets from menus in (2) and (3) below]

2. **mRNA expression**
Select [complete citations](#) or upload [example file](#)
REL: Ideker et al. (2001)

3. **Protein abundance**
Select [complete citations](#) or upload [example file](#)
REL: Ideker et al. (2001)

Please see [external mRNA expression and protein abundance databases](#) to retrieve additional datasets for analysis.

Analyze [help](#) [download PARE](#)

Correlate everything
Perform correlation for selected categories (subsets for selection appear on next page)

- pick a MIPS complex of proteins for the analysis
- pick a GO biological process subset
- pick a GO molecular function subset
- pick a GO cellular component subset
- upload your own subset [example file](#)

Other tools
[Sequence variation \(SNP\) substitution generator](#)

proteomics.gersteinlab.org
PARE: Protein Abundance and mRNA Expression Correlation Tool

The following analysis is a log-log correlation. Switch to a linear correlation?

Combined mRNA-protein file (sorted by perpendicular distance to fit line)

ORF_id	mRNA	Protein	Dist_to_fit
YBR218C	0.899	5.580	3.830
YKR097W	-2.303	4.041	3.608
YGR192C	4.489	6.580	3.406
YBR118W	3.928	6.325	3.381
YNL039W	0.000	4.636	3.294
YOR347C	-1.204	4.130	3.277
YIL136W	-0.916	4.210	3.244
YJR104C	2.688	5.617	3.192
YPL231W	1.882	5.286	3.188
YKR057W	4.255	-0.629	3.185

mRNA-protein overall correlation figure
Please note that the plot is loaded as an image file; you may need to refresh your browser to obtain the most recent plot.

customize the number of outliers shown in the plot (the top 5 shown by default)

absolute number:

percentage: % out of 2041

Mutual information [help](#) = 10.66
Calculated using 204 bins for the mRNA and protein data

Open-source code
Downloadable

Analyze all or GO
subset

Log-log plot of
correlation
-linear fit
-outliers labeled

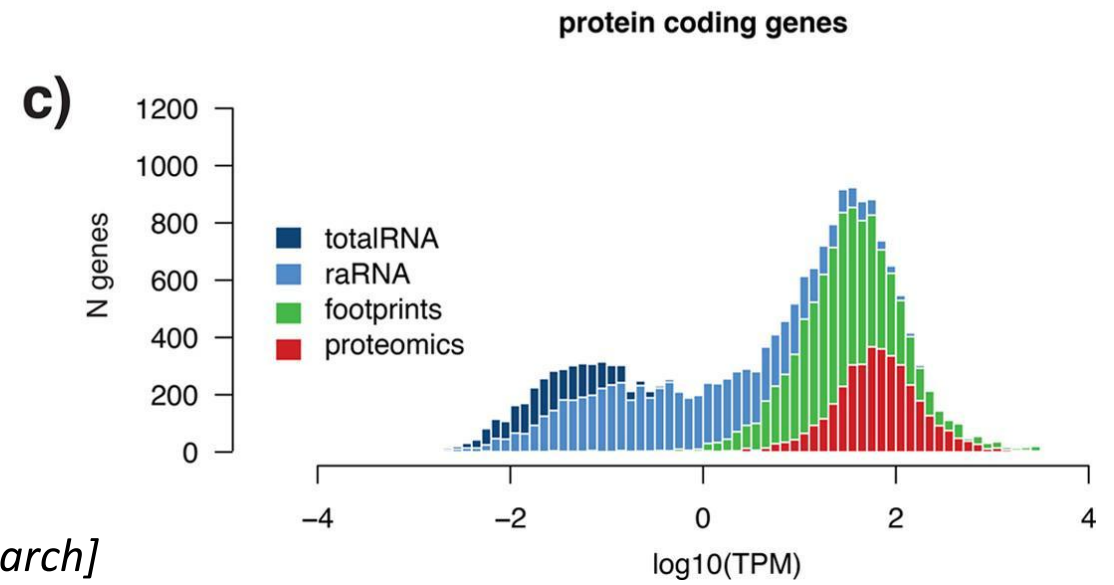
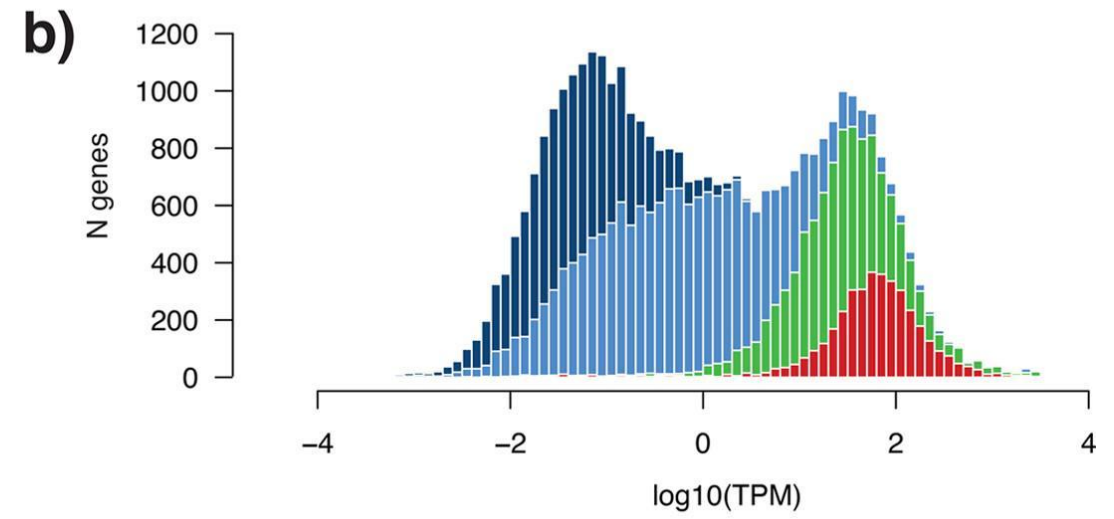
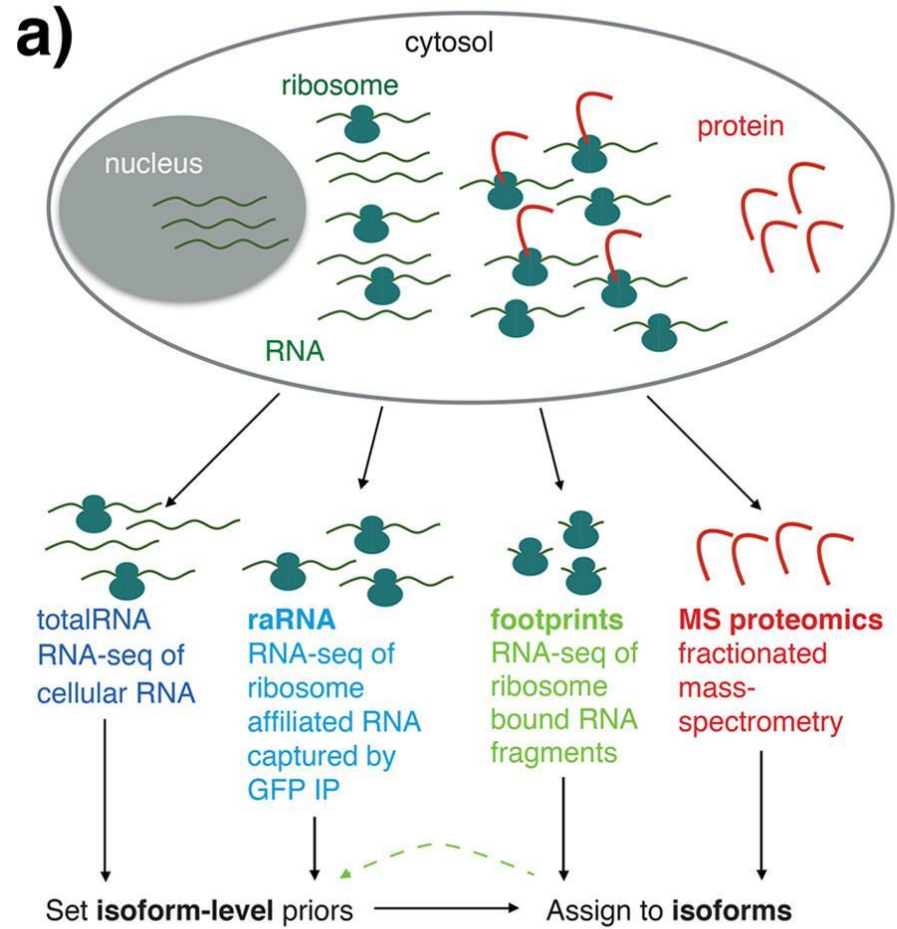
Calculation of
mutual information

[Yu et al., BMC Bioinfo. '07]

Outline: Comparing Protein & RNA Abundance

- **Past Context:**
to work in the Center
 - Quantifying the moderate **statistical correlation between protein & RNA**
 - PARE server
- **EMpire** (Current result)
 - Leveraging the correlation to **better assign peptides to isoforms**
 - EM algorithm better assigns **dominant isoforms**, with greater interpretability
- **uORFs** (Current result)
 - Affect translation & relationship between protein & RNA
 - Feature integration to find **small subset of uORFs that most alter translation**
- **Future Direction:**
Protein v RNA using matched samples in the Brainspan dataset + single-cell data

Integration of RNA-seq and Proteomic Data for Isoform Interpretation

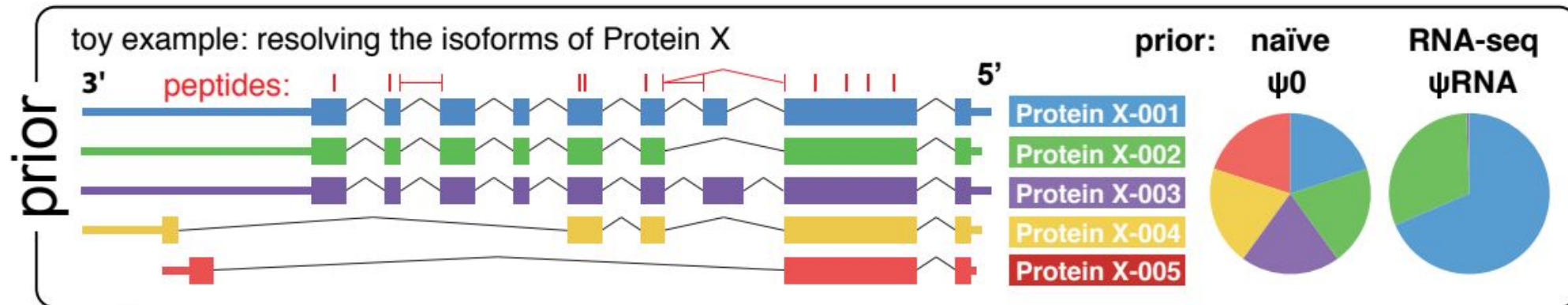


[Carlyle, Kitchen et al. (2018) *Journal of Proteome Research*]

Challenge for Isoform-Level Interpretation of Proteomics Data

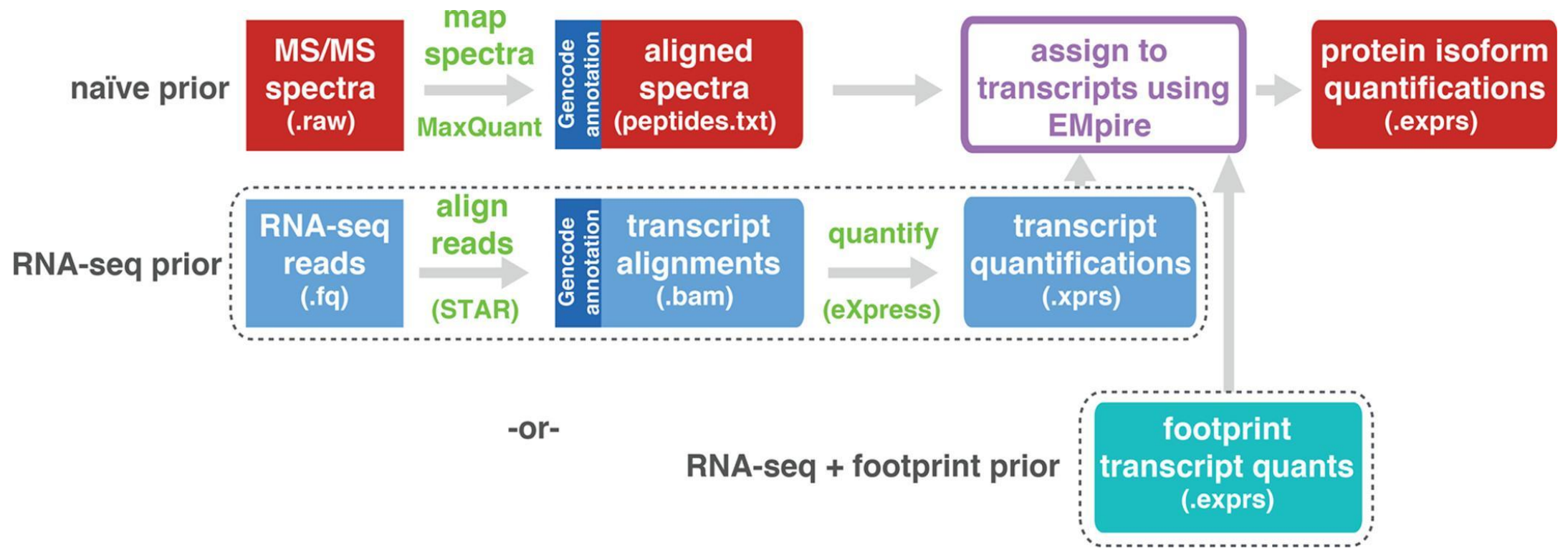
Multimapping

- Different assays reflecting expression at various levels
- More reads at earlier stage assay (RNA-Seq > FP > MS)
- Leverage other assays for better estimation



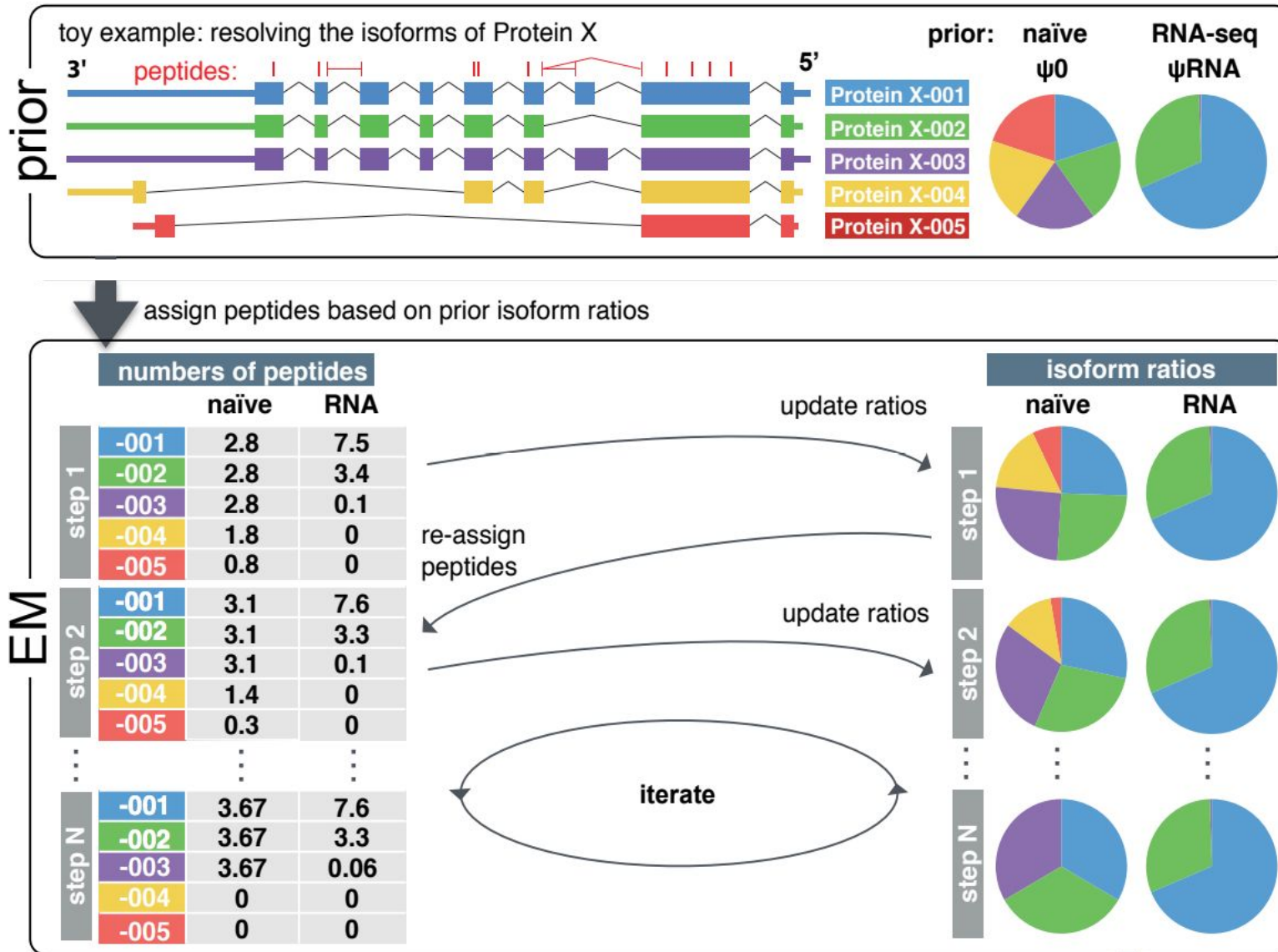
[Carlyle, Kitchen et al. (2018) *Journal of Proteome Research*]

EMpire (Expectation Maximisation Propagation of Isoform abundance from RNA Expression)



[Carlyle, Kitchen et al. (2018) *Journal of Proteome Research*]

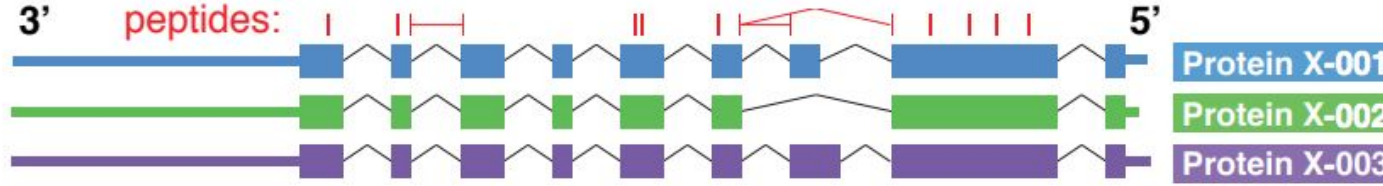
EMpire (Expectation Maximisation Propagation of Isoform abundance from RNA Expression)



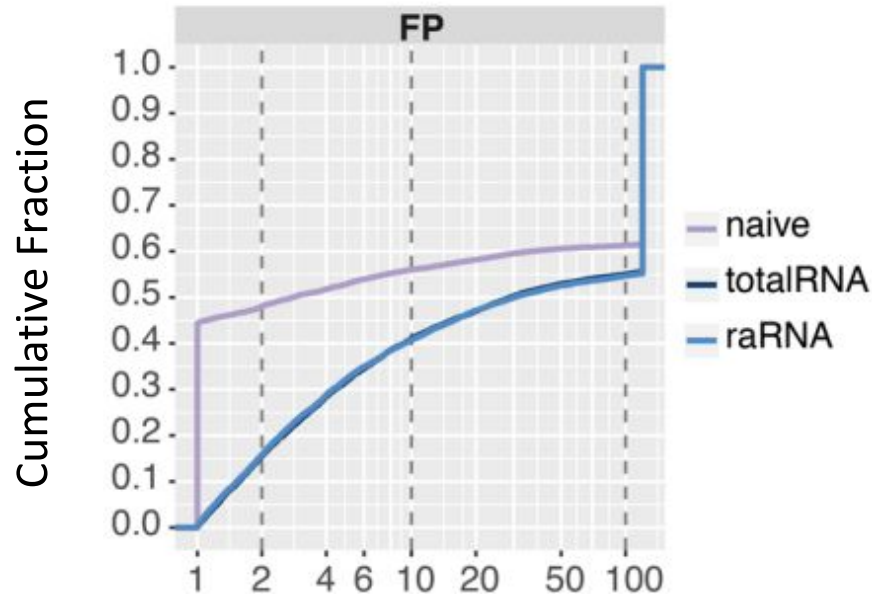
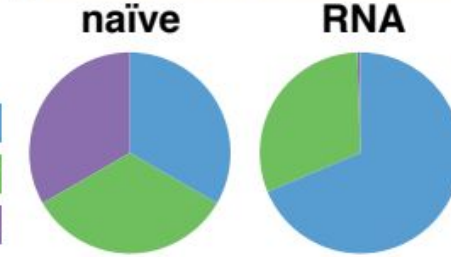
[Carlyle, Kitchen et al. (2018) *Journal of Proteome Research*]

result

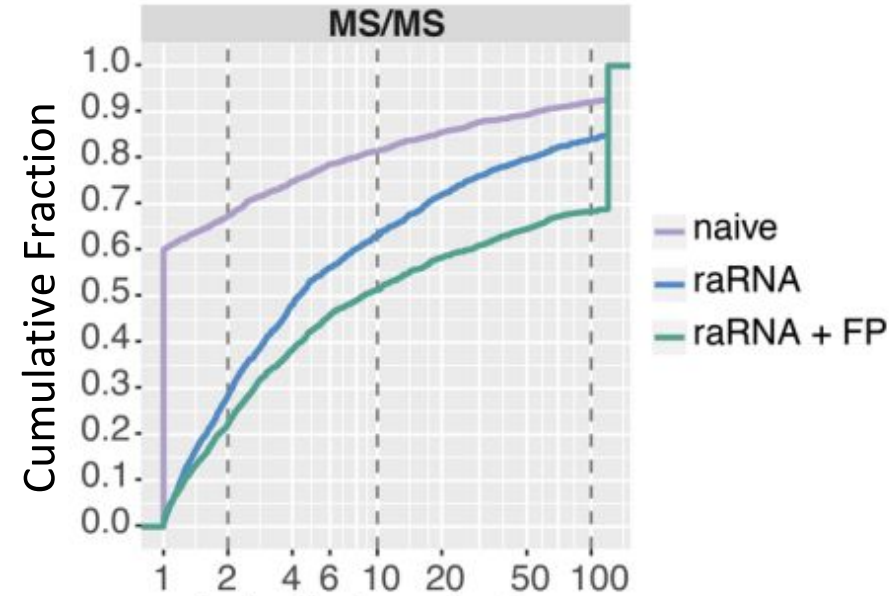
The RNA prior supports **Protein X-001** as the >2-fold dominant isoform, while the naïve prior gets stuck with three equally likely isoforms:



posterior isoform ratios



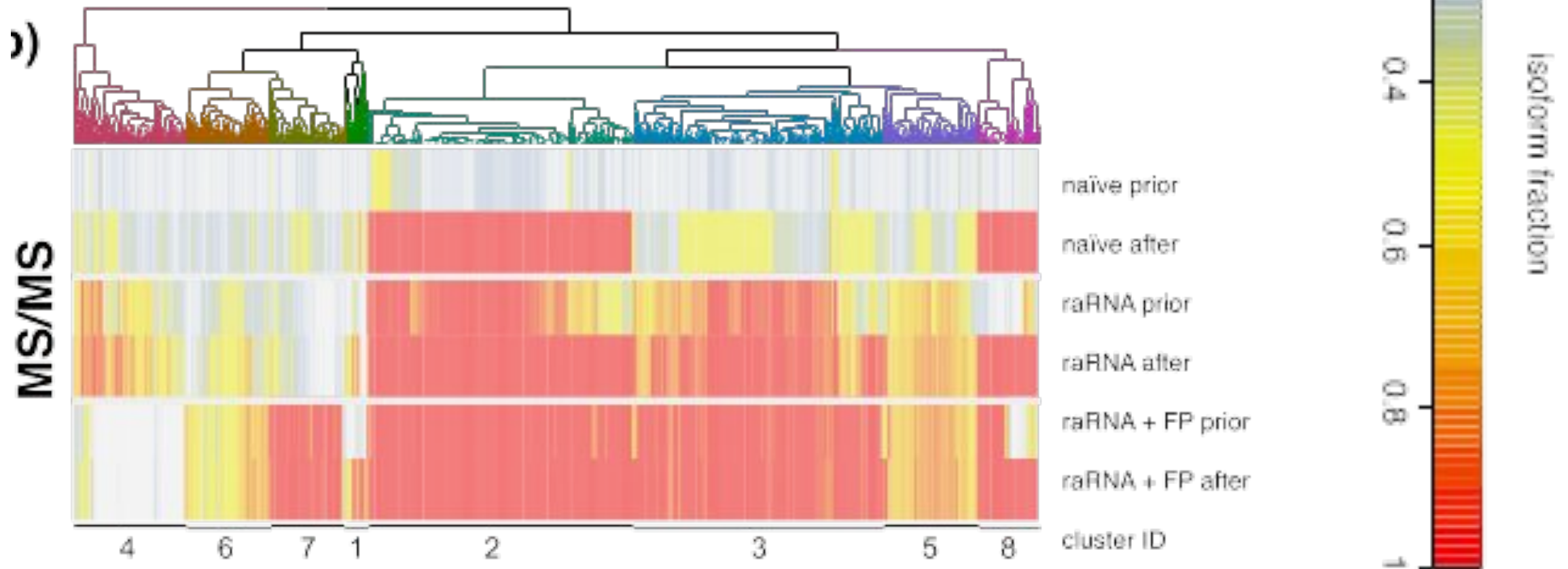
\leq principal isoform dominance
=principal isoform/second isoform



\leq principal isoform dominance
=principal isoform/second isoform

Larger principal isoform dominance = **Less** ambiguity in major isoform identification

Biologically informative priors improve isoform level interpretation of MS/MS peptides, by increasing dominance of principal isoform

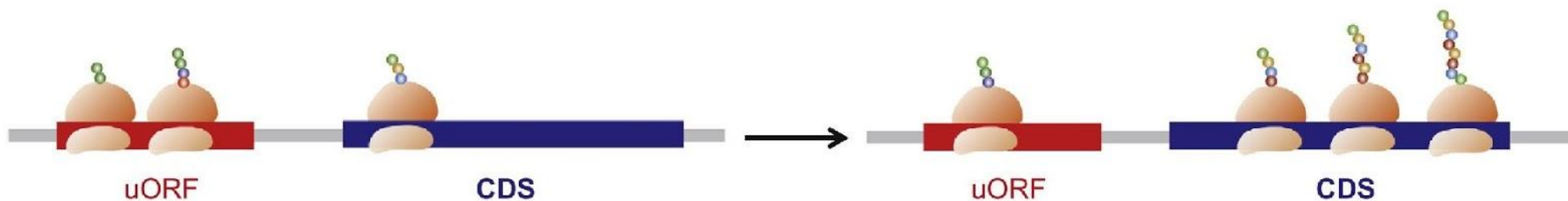


[Carlyle, Kitchen et al. (2018) *Journal of Proteome Research*]

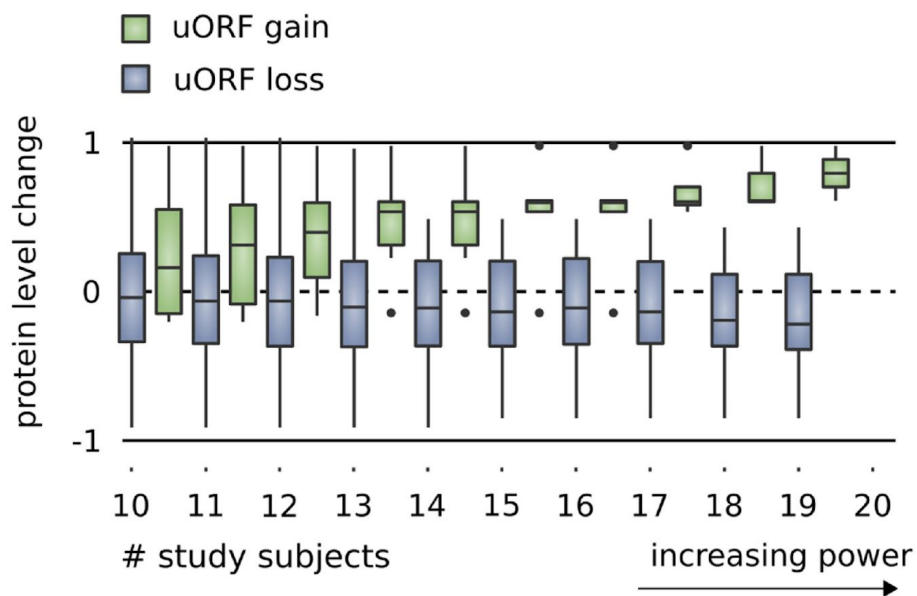
Outline: Comparing Protein & RNA Abundance

- **Past Context:**
to work in the Center
 - Quantifying the moderate **statistical correlation between protein & RNA**
 - PARE server
- **EMpire** (Current result)
 - Leveraging the correlation to **better assign peptides to isoforms**
 - EM algorithm better assigns **dominant isoforms**, with greater interpretability
- **uORFs** (Current result)
 - Affect translation & relationship between protein & RNA
 - Feature integration to find **small subset of uORFs that most alter translation**
- **Future Direction:**
Protein v RNA using matched samples in the Brainspan dataset + single-cell data

Upstream open reading frames (uORFs) may shift the expected balance between mRNA & protein

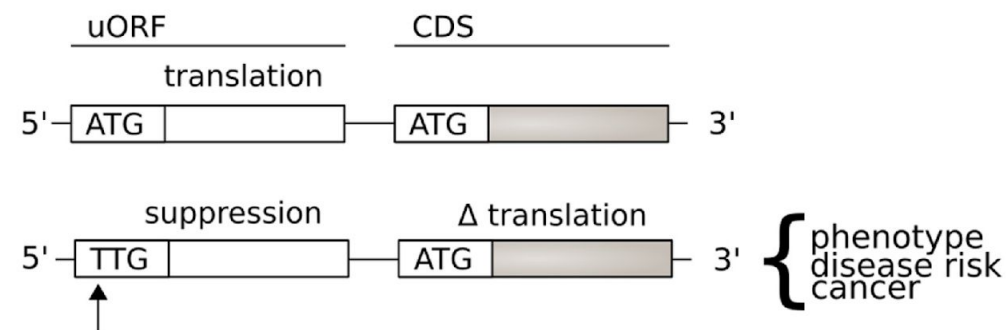


[Zhang et al., Trends in Biochemical Sciences ('19)]



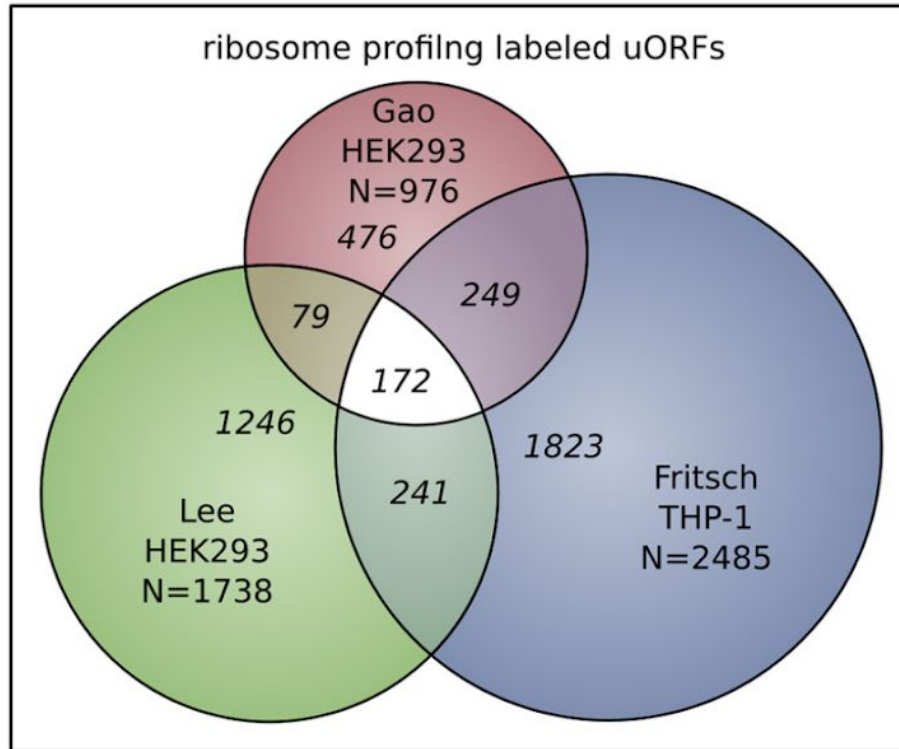
In Battle et al. 2014 data uORF gain & loss assoc. protein level change.

B



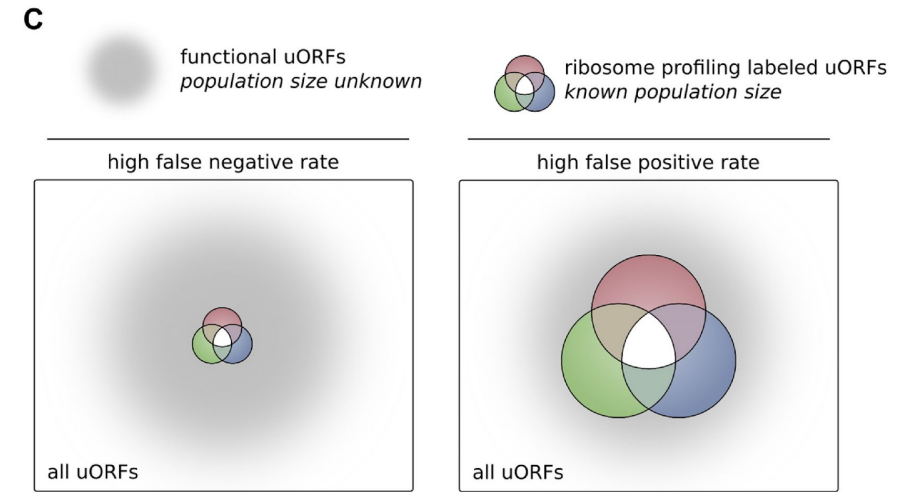
uORF regulation can be affected by mutation

[McGillivray et al., NAR ('18)]



genome-wide uORFs
N = 1.3 million

The population of functional uORFs may be significant



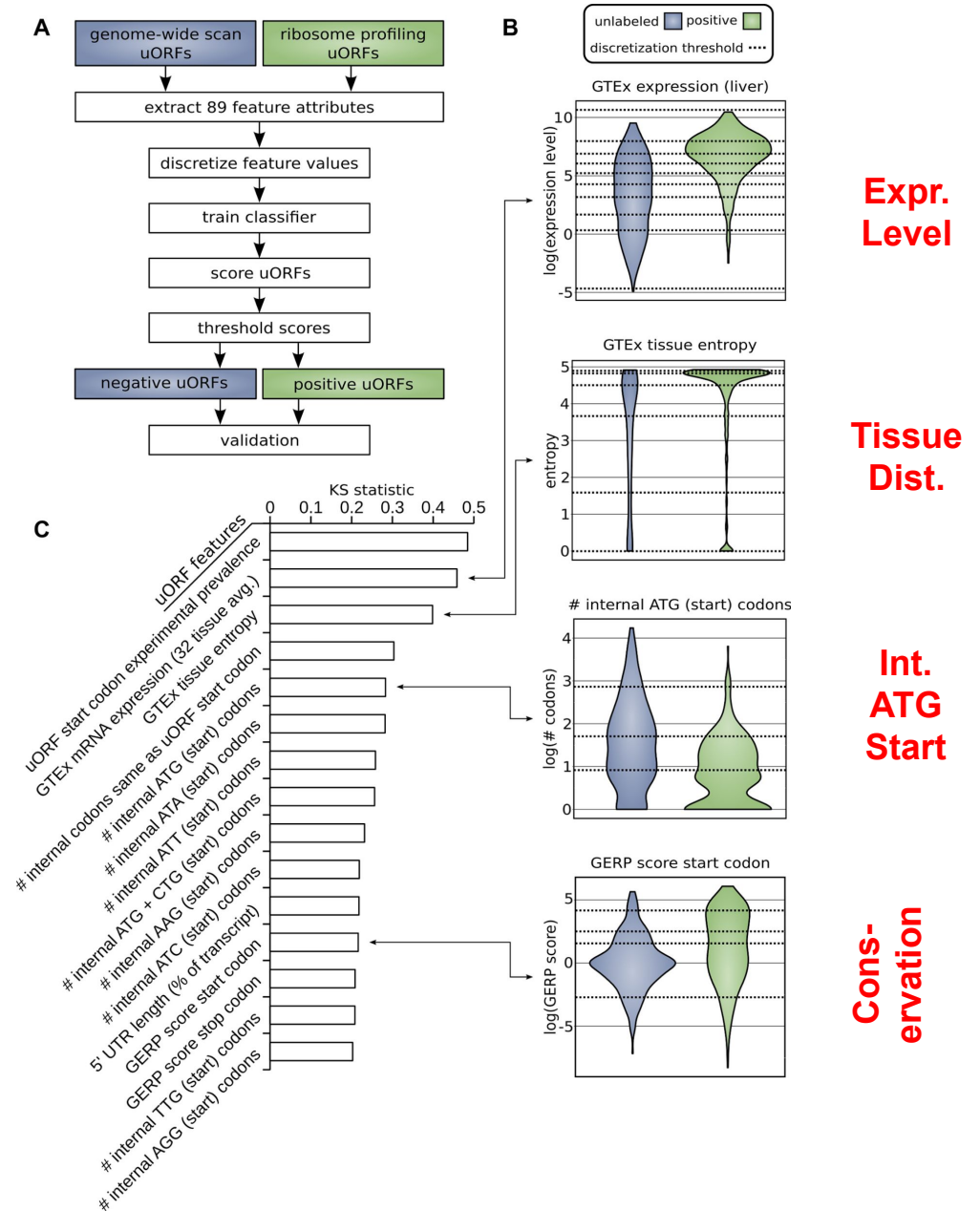
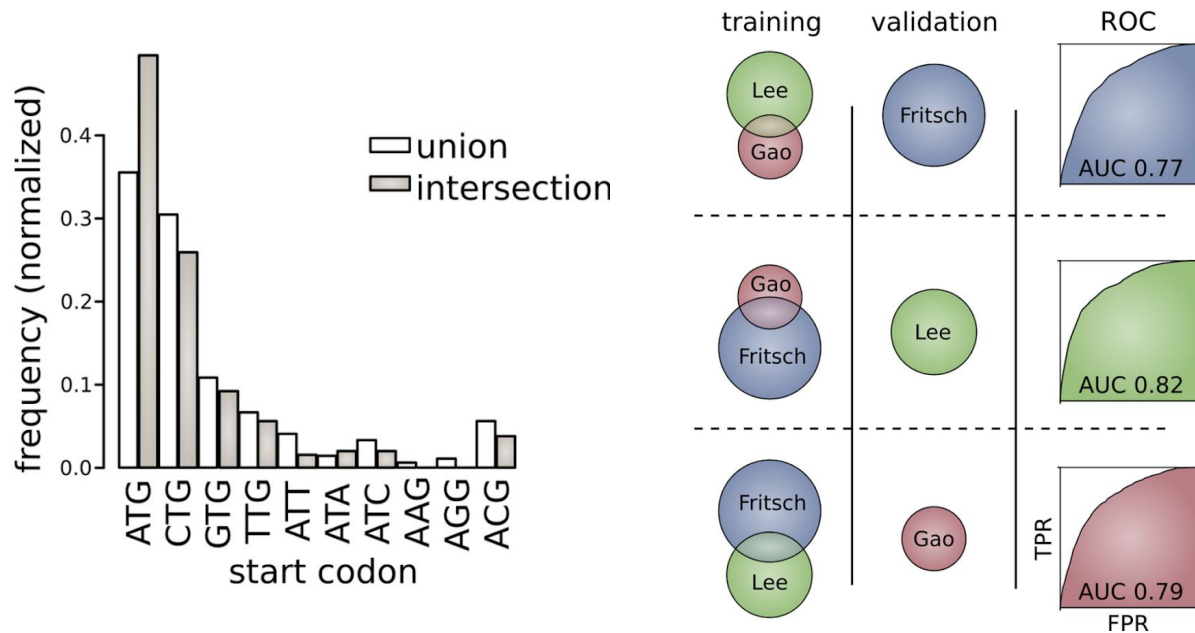
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

**From a “Universe” of 1.3 M
pot. uORFs**

[McGillivray et al., *NAR* ('18)]

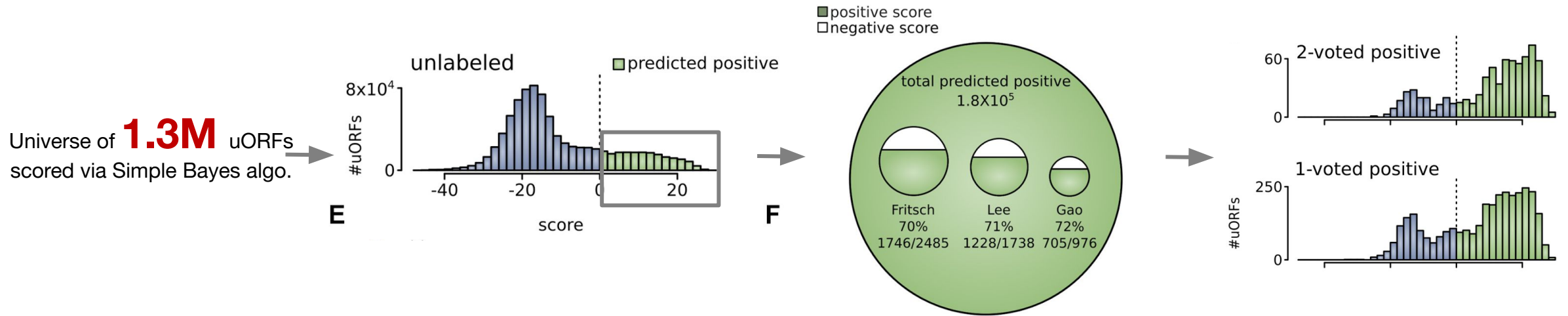
Prediction & validation of functional uORFs using 89 features

- All near-cognate start codons predicted.
- Cross-validation on independent ribosome profiling datasets and validation using in vivo protein levels and ribosome occupancy in humans (Battle et al. 2014).



[McGillivray et al., NAR ('18)]

A comprehensive catalog of functional uORFs



- Predicted functional uORFs may be intersected with disease associated variants.

[McGillivray et al., *NAR* ('18)]

- **180K**: Large predicted positive set likely to affect translation
- Calibration on gold standards, suggests getting **~70%** of known

Outline: Comparing Protein & RNA Abundance

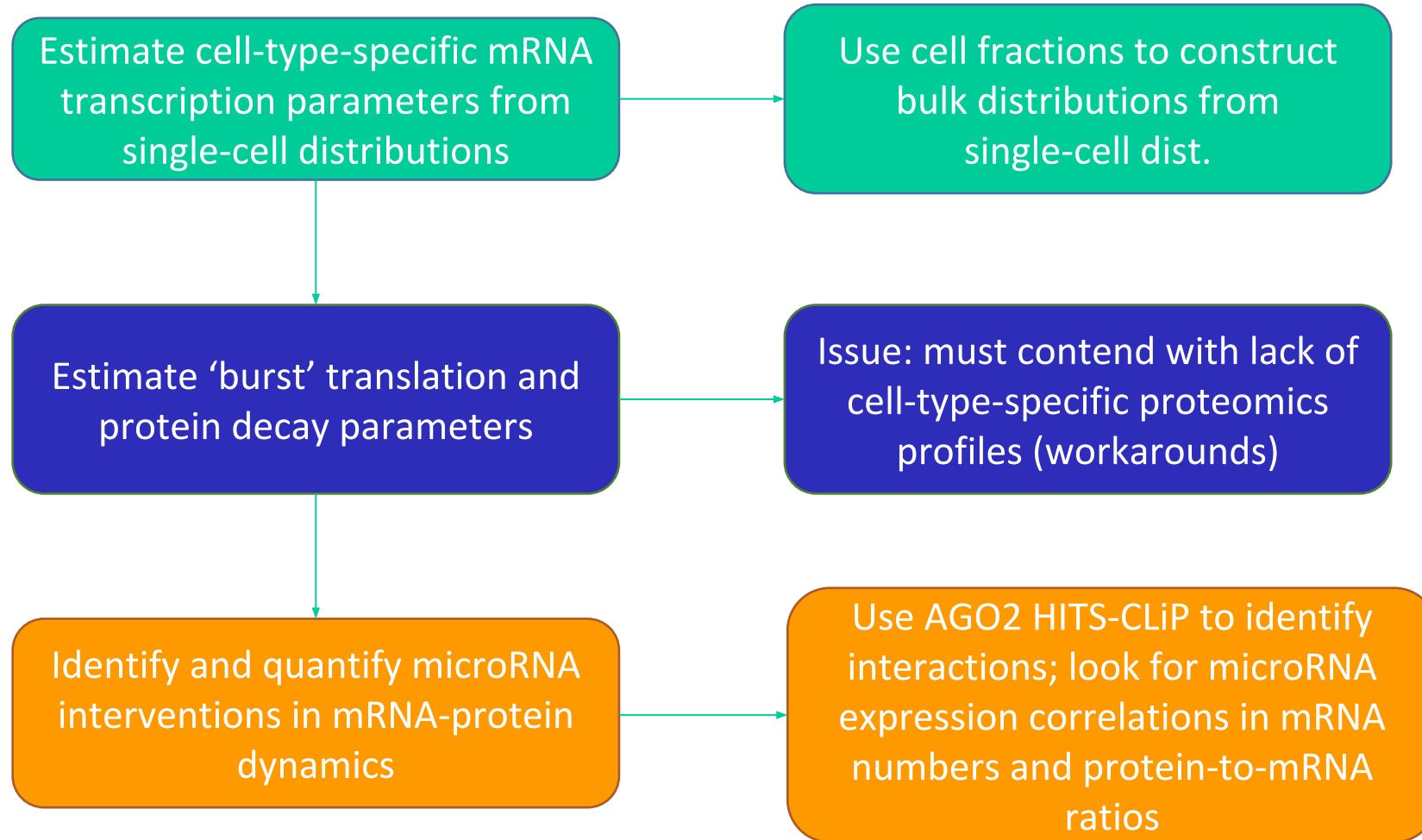
- **Past Context:**
to work in the Center
 - Quantifying the moderate **statistical correlation between protein & RNA**
 - PARE server
- **EMpire** (Current result)
 - Leveraging the correlation to **better assign peptides to isoforms**
 - EM algorithm better assigns **dominant isoforms**, with greater interpretability
- **uORFs** (Current result)
 - Affect translation & relationship between protein & RNA
 - Feature integration to find **small subset of uORFs that most alter translation**
- **Future Direction:**
Protein v RNA using matched samples in the Brainspan dataset + single-cell data

Leveraging New Datasets

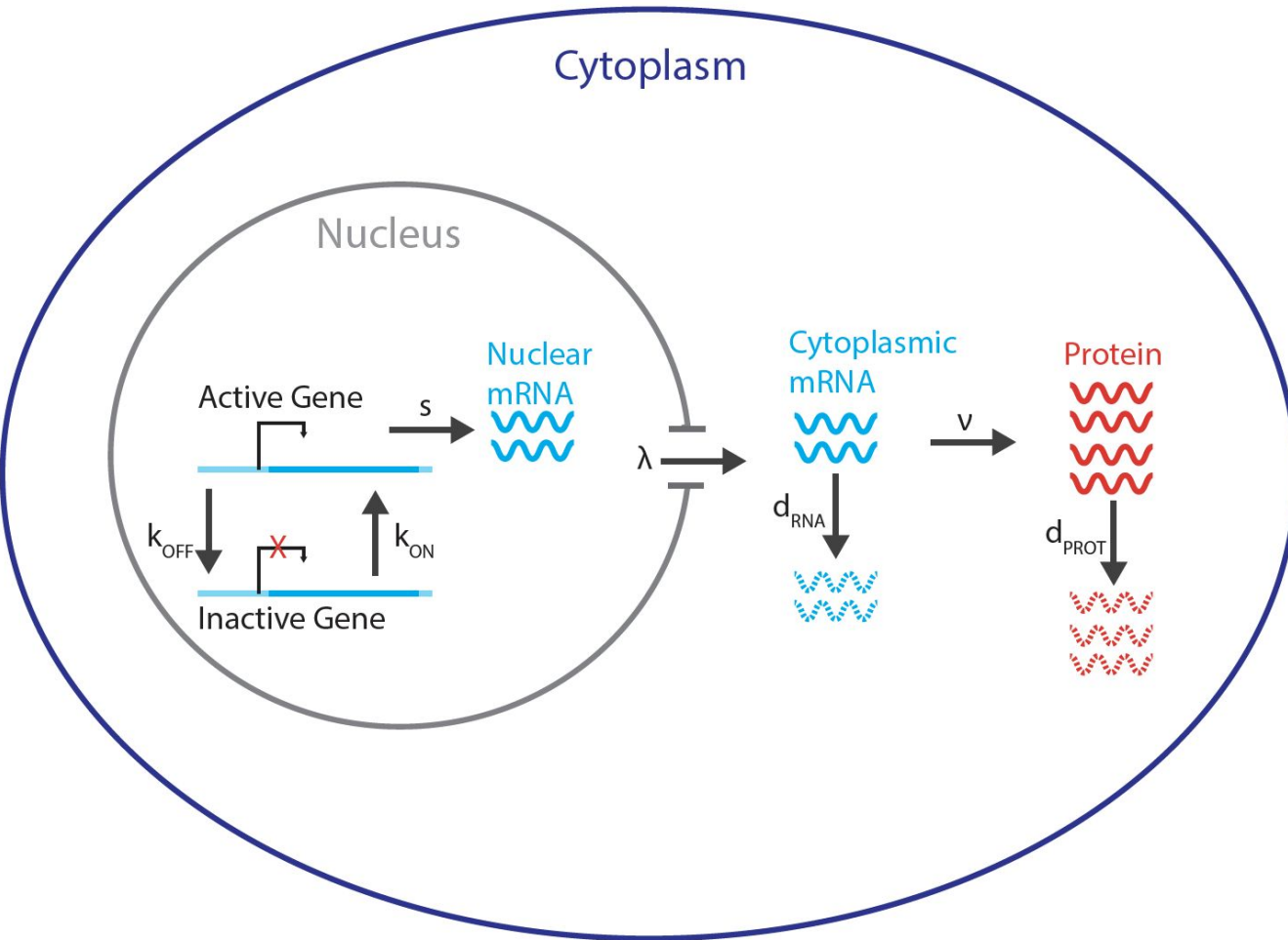
1. Availability of single-cell RNA-seq datasets warrants revisiting mRNA-protein dynamics with an eye towards cell-type-specific characterizations:
 - Can the rich information on mRNA distributions from single-cell data provide clues to the origin of mRNA-protein discrepancies in the bulk data?
 - What can we learn about cell-type specifics of the mRNA-to-protein relationship?
2. The Brainspan dataset considered in the NIDA grant research represents a unique opportunity for individual-specific study:
 - Individual-matched bulk RNA-seq, smallRNA-seq and proteomics data are available for several individuals in several brain regions
 - AGO2 HITS-CLIP data is also available to link microRNA to potential mRNA targets

Combining all these modalities into a coherent framework suggests a return to first principles

Schematic workflow



mRNA-Protein dynamics

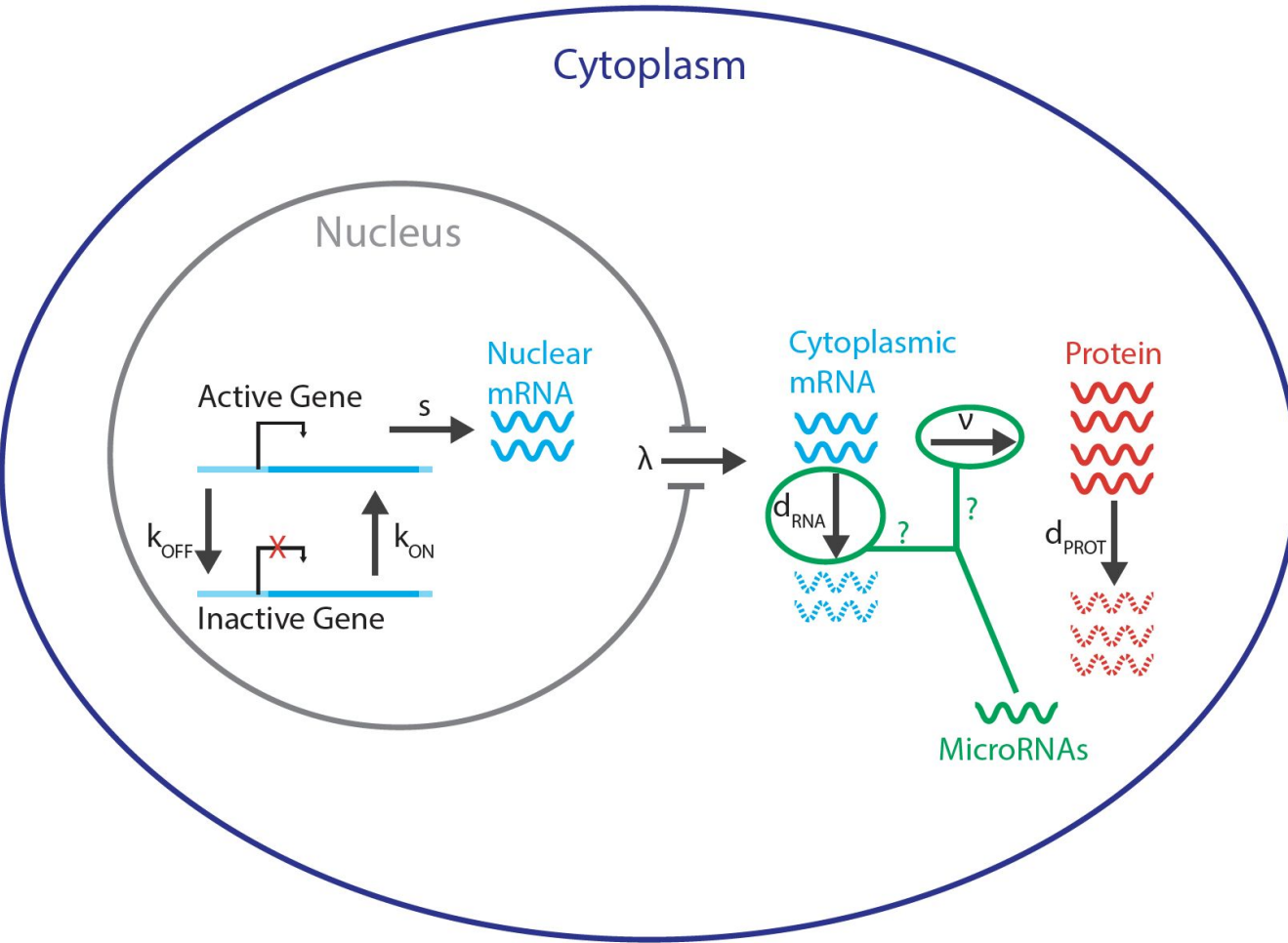


1. Use normalized single-cell mRNA distributions:
 - a. Infer cell-type-specific k_{ON} , k_{OFF} and s (Kim and Marioni 2013) using Poisson-Beta model of 'burst' transcription
 - b. Replace decay rate by nuclear export rate λ
2. Use two-state model of protein translation (Shahrezaei and Swain 2008):
 - a. Obtain cell-type-specific cyto. RNA and Protein numbers in terms of λ , v , d_{RNA} , d_{PROT}
 - b. Use bulk mRNA and Protein data as constraints

Kim and Marioni, *Genome Biology* **2013**, 14:R7.

Shahrezaei and Swain, *PNAS* **2008**, 105(45), Pgs. 17256–17261.

MicroRNA intervention



1. Use AGO2 HITS-CLIP binding data on microRNA-mRNA interactions in the brain (Sousa et al 2017):
 - a. Compare mRNAs with multiple targeting microRNAs to those with one or none; expression levels and inferred parameters
2. Use smallRNA-seq expression data:
 - a. Directly correlate with protein-to-mRNA ratios and mRNA expression levels
 - b. Correlate with inferred parameters to determine whether microRNA impacts mRNA degradation, protein translation, or both

Sousa et al., *Science* **2017**, 358, Pgs. 1027–1032.

Outline: Comparing Protein & RNA Abundance

- **Past Context:**
to work in the Center
 - Quantifying the moderate **statistical correlation between protein & RNA**
 - PARE server
- **EMpire** (Current result)
 - Leveraging the correlation to **better assign peptides to isoforms**
 - EM algorithm better assigns **dominant isoforms**, with greater interpretability
- **uORFs** (Current result)
 - Affect translation & relationship between protein & RNA
 - Feature integration to find **small subset of uORFs that most alter translation**
- **Future Direction:**
Protein v RNA using matched samples in the Brainspan dataset + single-cell data

Acknowledgments!

genecensus.org/expression/translatome

D **Greenbaum**, R Jansen, M Gerstein

Proteomics.gersteinlab.org (**PARE**)

E **Yu**, A Burba, M Gerstein

github.com/rkitchen/EMpire

B **Carlyle**, R **Kitchen**, J Zhang, R Wilson, T Lam, J Rozowsky,
K Williams, N Sestan, M Gerstein, A Nairn

[github.gersteinlab.org/uORFs](https://github.com/gersteinlab.org/uORFs)

P **McGillivray**, R Ault, M Pawashe, R Kitchen,
S Balasubramanian, M Gerstein

Brainspan data

P **Emani**, T Galeev, N Sestan, A Nairn