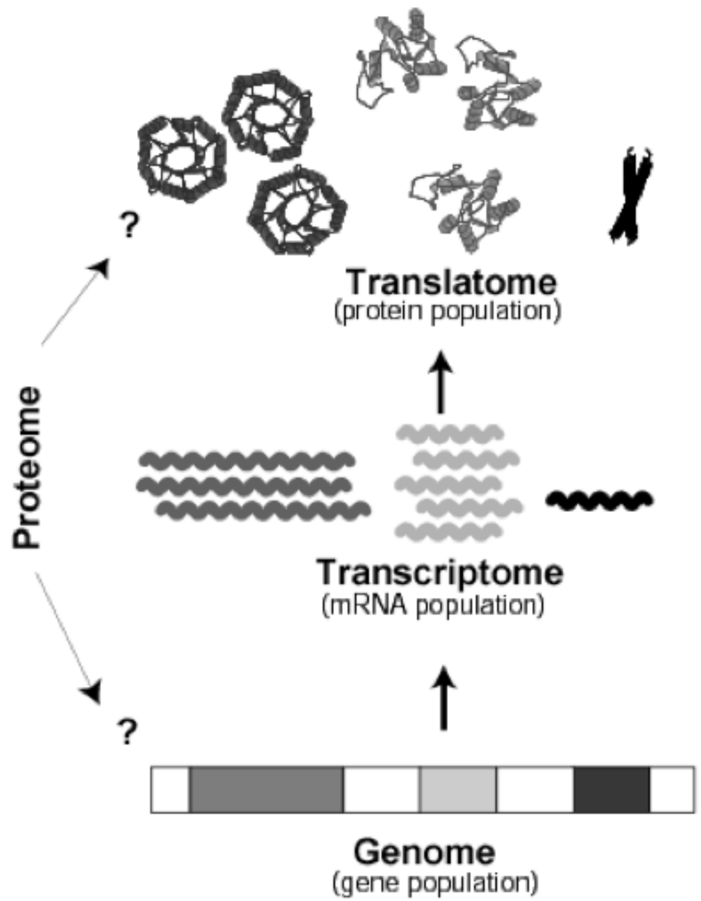
Comparing RNA & Protein Abundance

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?



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

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

vs.

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

Expectations from simple kinetic models:

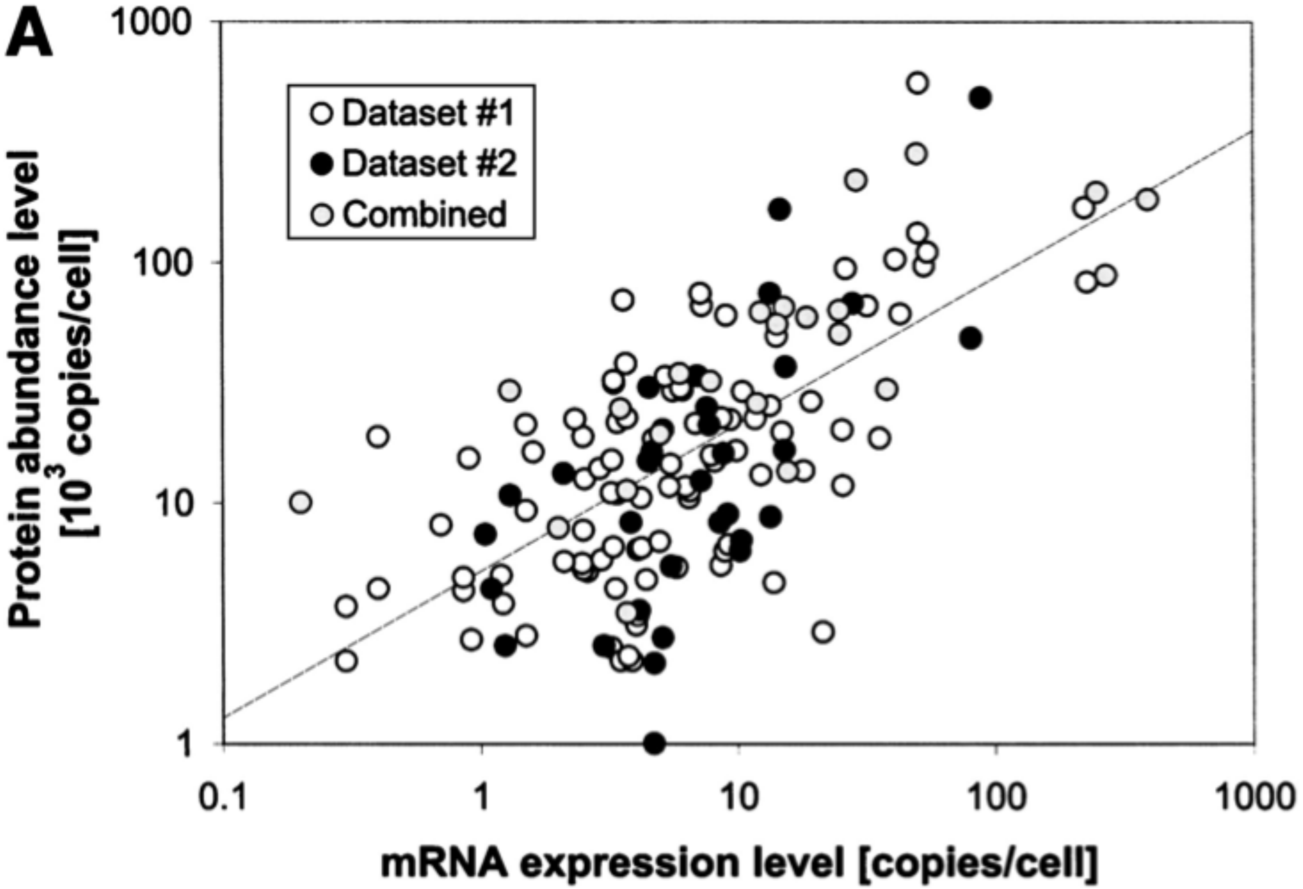
$$\frac{dP_i}{dt} = k_{s,i}[\mathrm{mRNA}_i] - k_{d,i}P_i$$

At steady state: $P_i = \dfrac{k_{s;i}[\mathrm{mRNA}_i]}{k_{d,i}}$

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

Outliers from trend interesting

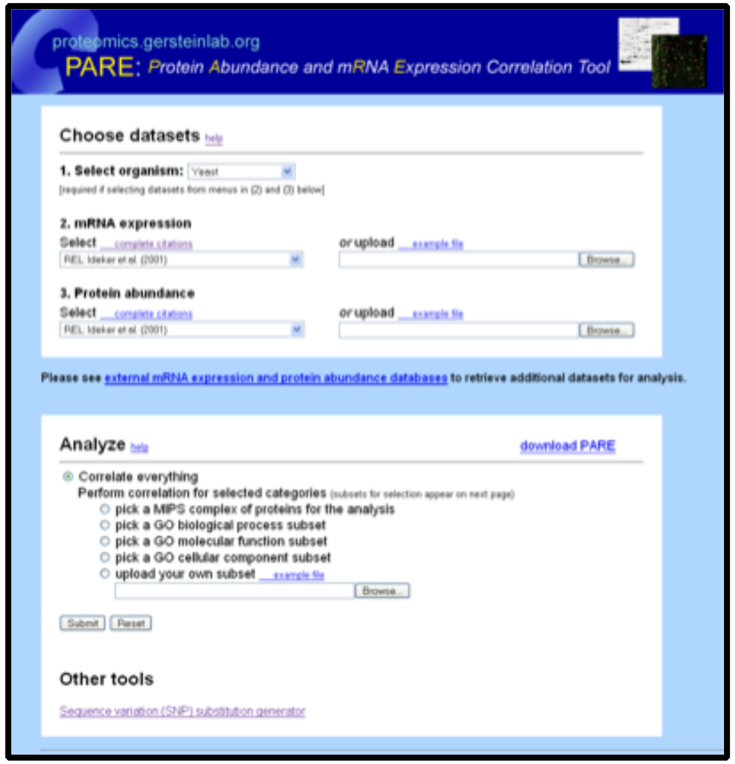
Early result on mRNA vs Protein, using 2D gels



r =0.67

[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: REL: Ideker et al. (2001) or upload example file [Browse...]

3. Protein abundance

Select complete citations: REL: Ideker et al. (2001) or upload example file [Browse...]

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)

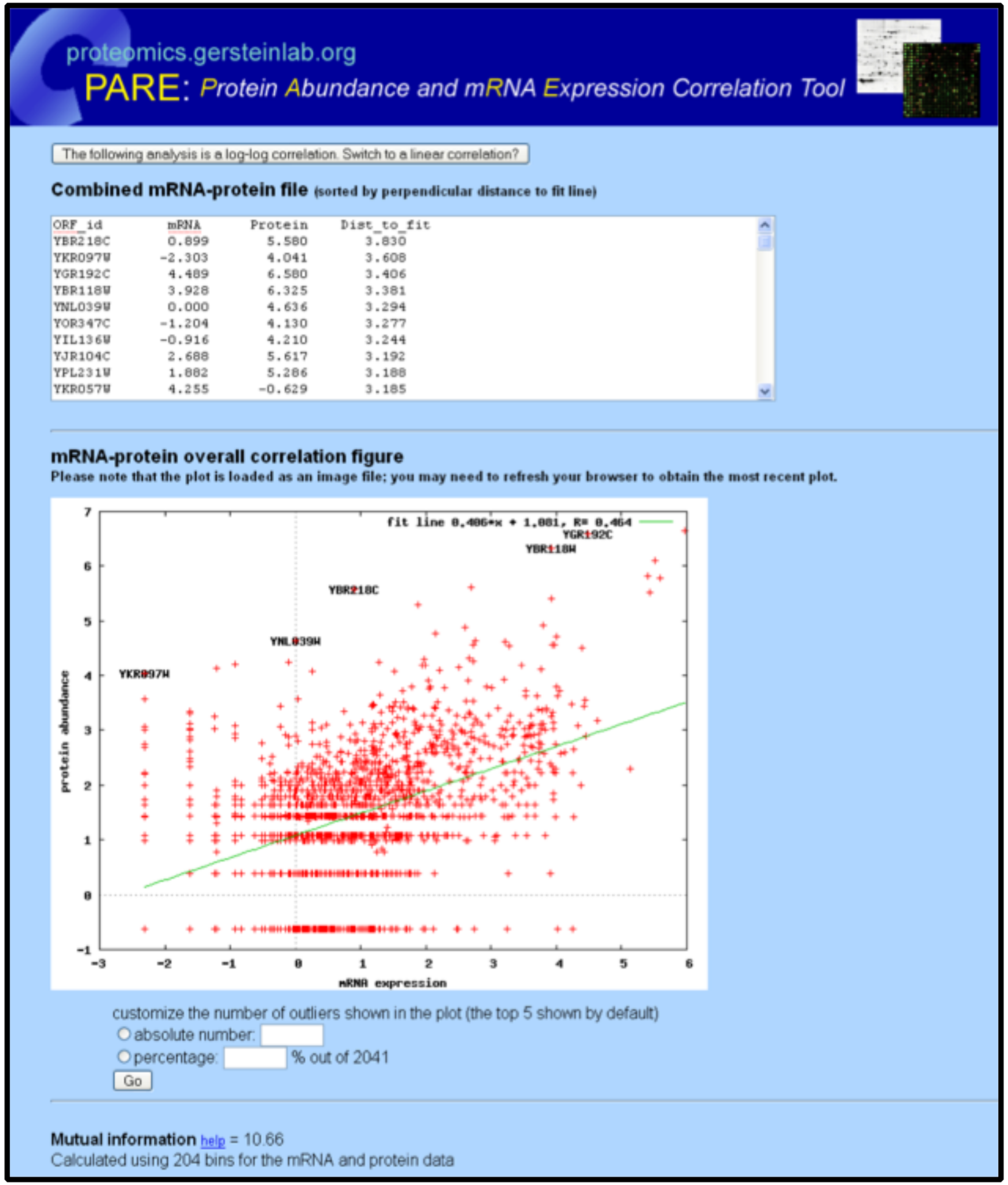
○ pick a MIPS complex of proteins for the analysis
○ pick a GO biological process subset
○ pick a GO molecular function subset
○ pick a GO cellular component subset
○ upload your own subset example file [Browse...]

[Submit] [Reset]

Other tools

Sequence variation (SNP) substitution generator

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



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

[Go]

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

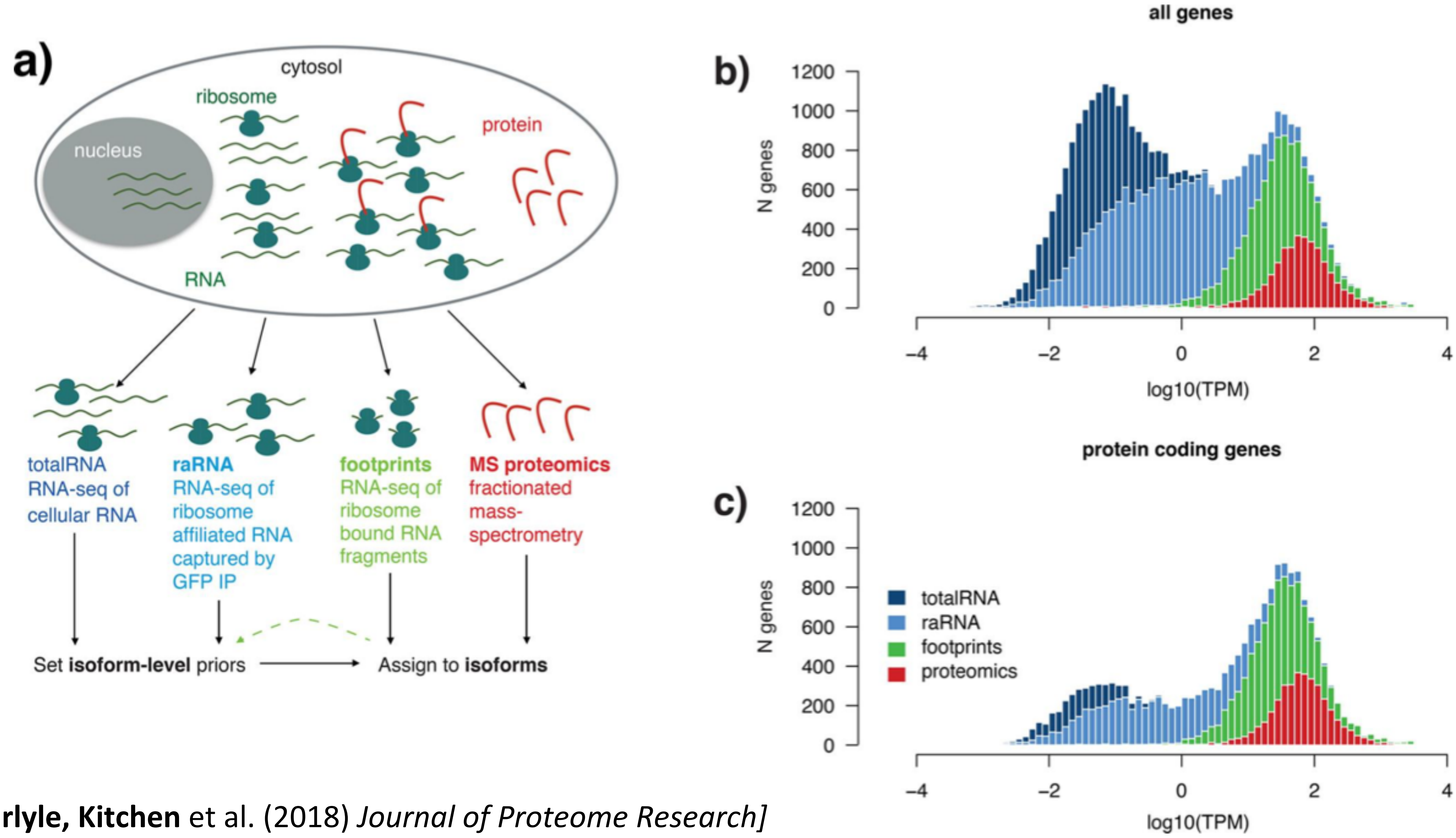
Calculation of mutual information

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

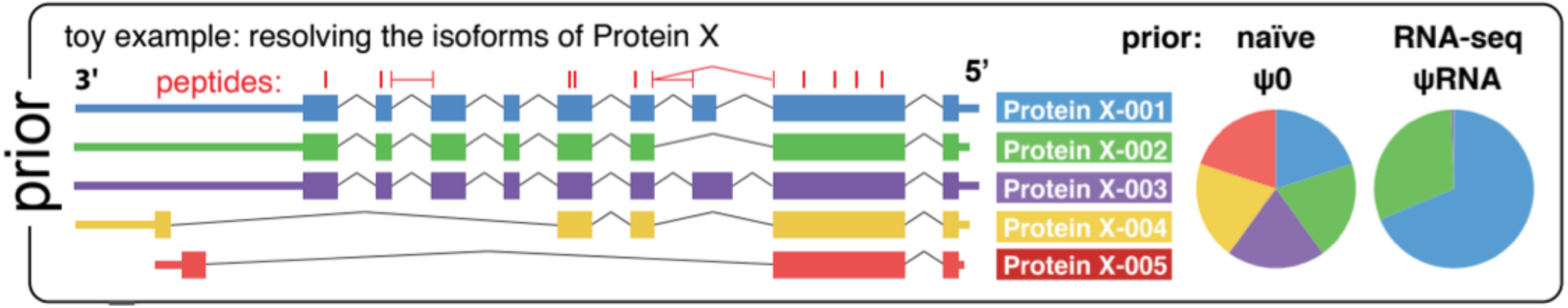
Integration of RNA-seq and Proteomic Data for Isoform Interpretation



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

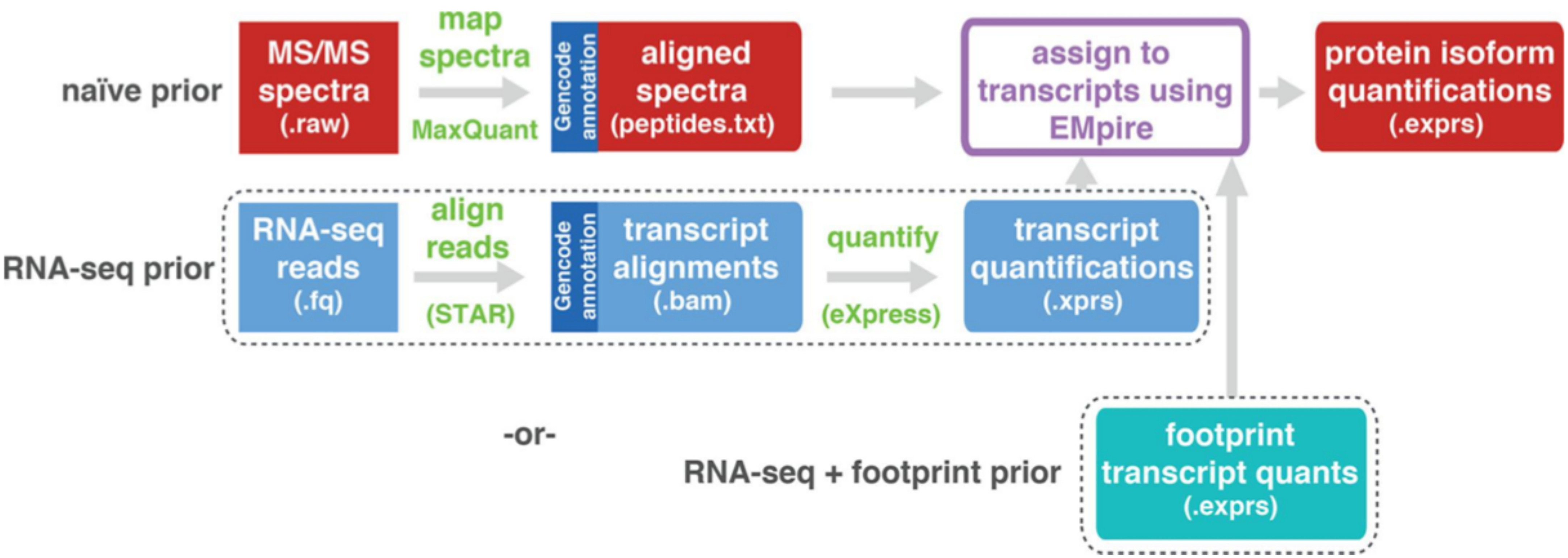
Challenge for Isoform-Level Interpretation of Proteomics Data
Multimapping

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



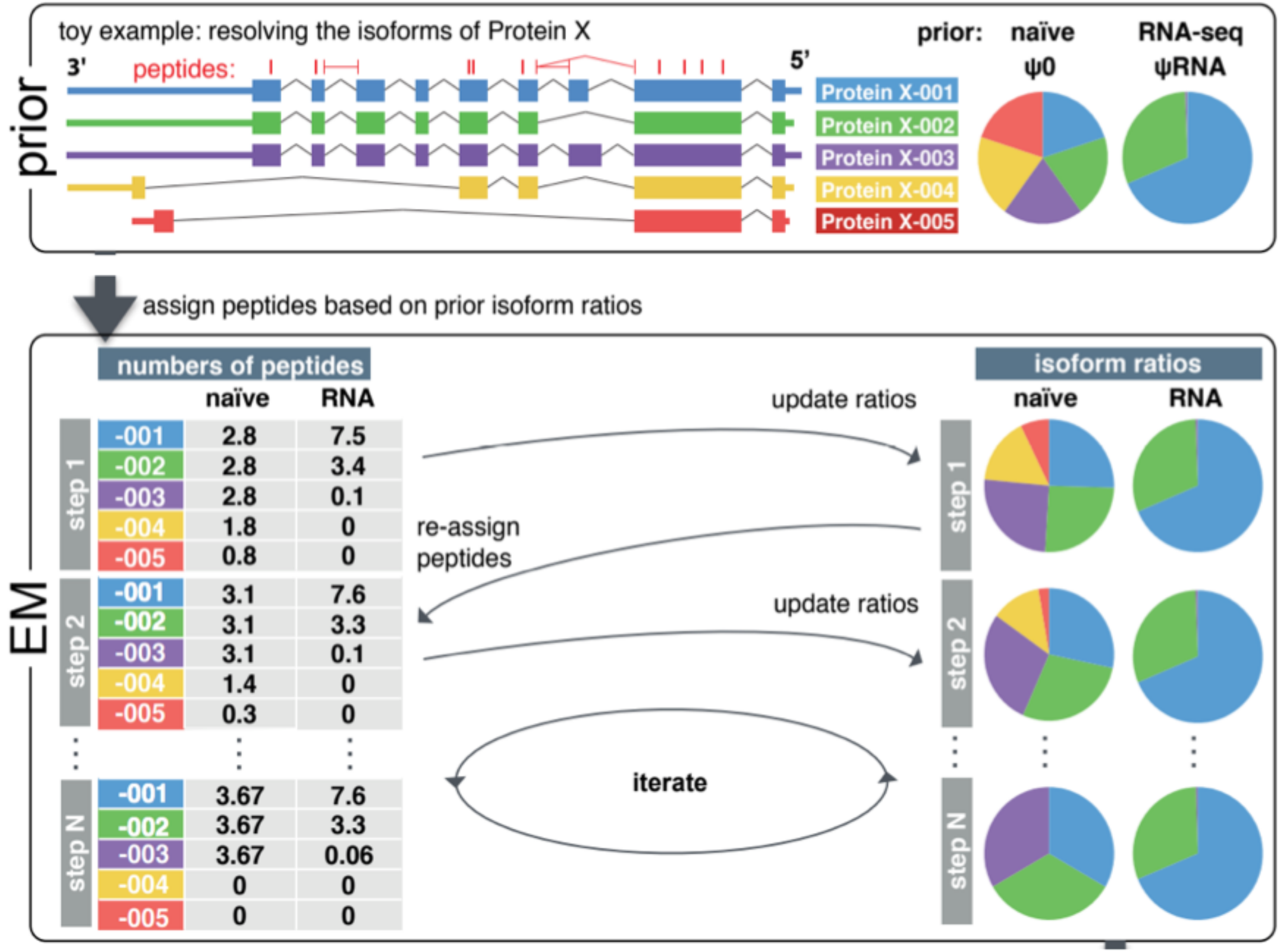
[**Carlyle, Kitchen** et al. (2018) *Journal of Proteome Research]*

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


naïve prior
MS/MS spectra (.raw)
map spectra
MaxQuant
Gencode annotation
aligned spectra (peptides.txt)
assign to transcripts using EMpire
protein isoform quantifications (.exprs)
RNA-seq prior
RNA-seq reads (.fq)
align reads
(STAR)
Gencode annotation
transcript alignments (.bam)
quantify
(eXpress)
transcript quantifications (.xprs)
-or-
RNA-seq + footprint prior
footprint transcript quants (.exprs)

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

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

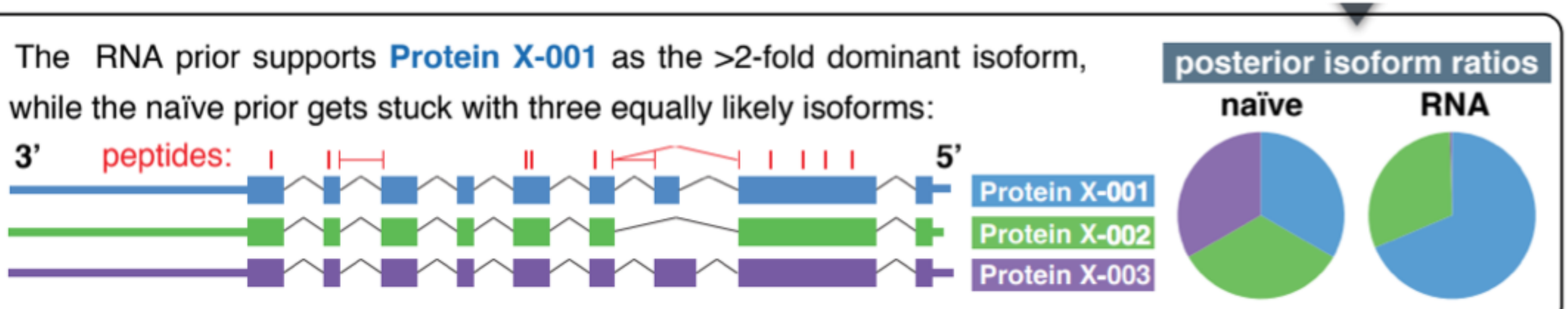


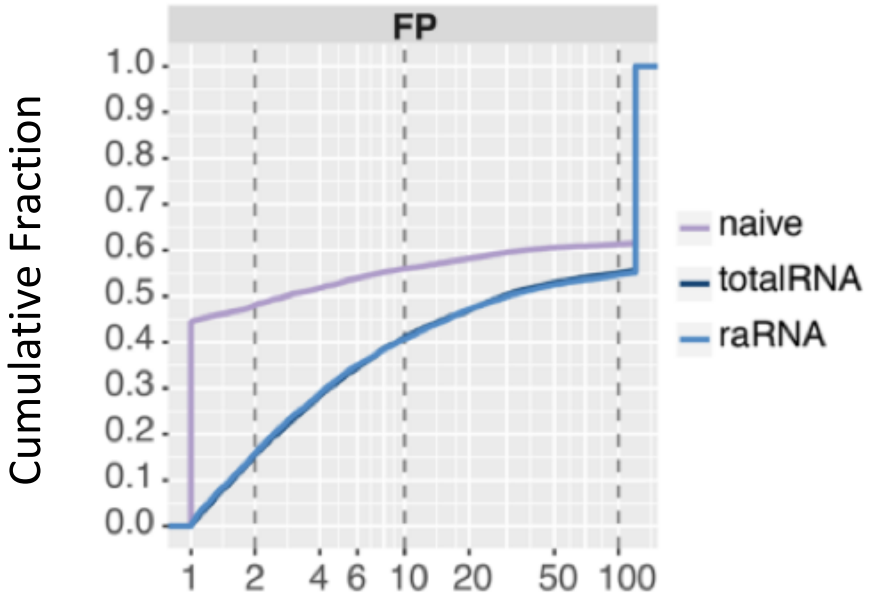
		naïve	RNA
step 1	-001	2.8	7.5
	-002	2.8	3.4
	-003	2.8	0.1
	-004	1.8	0
	-005	0.8	0
step 2	-001	3.1	7.6
	-002	3.1	3.3
	-003	3.1	0.1
	-004	1.4	0
	-005	0.3	0
⋮		⋮	⋮
step N	-001	3.67	7.6
	-002	3.67	3.3
	-003	3.67	0.06
	-004	0	0
	-005	0	0

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

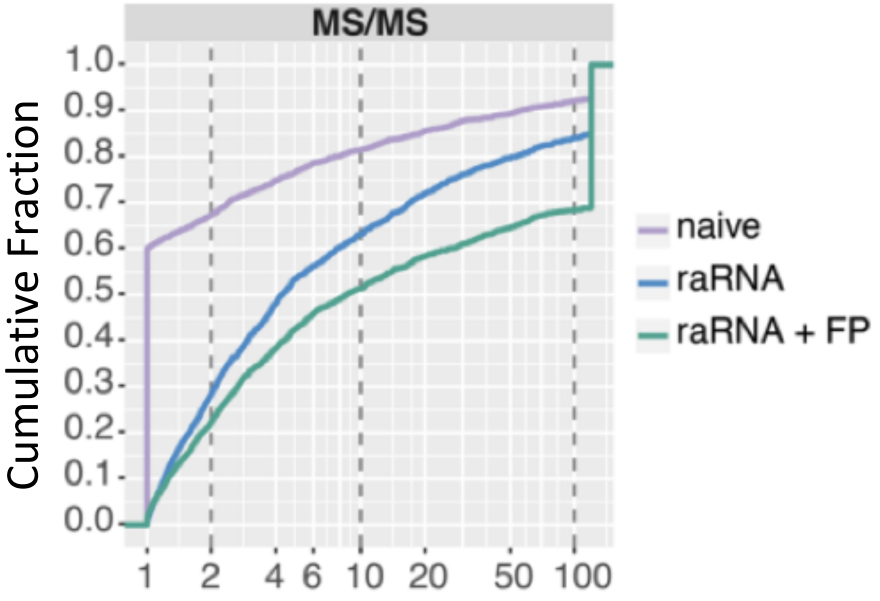
result

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





≤ principal isoform dominance
=principal isoform/second isoform

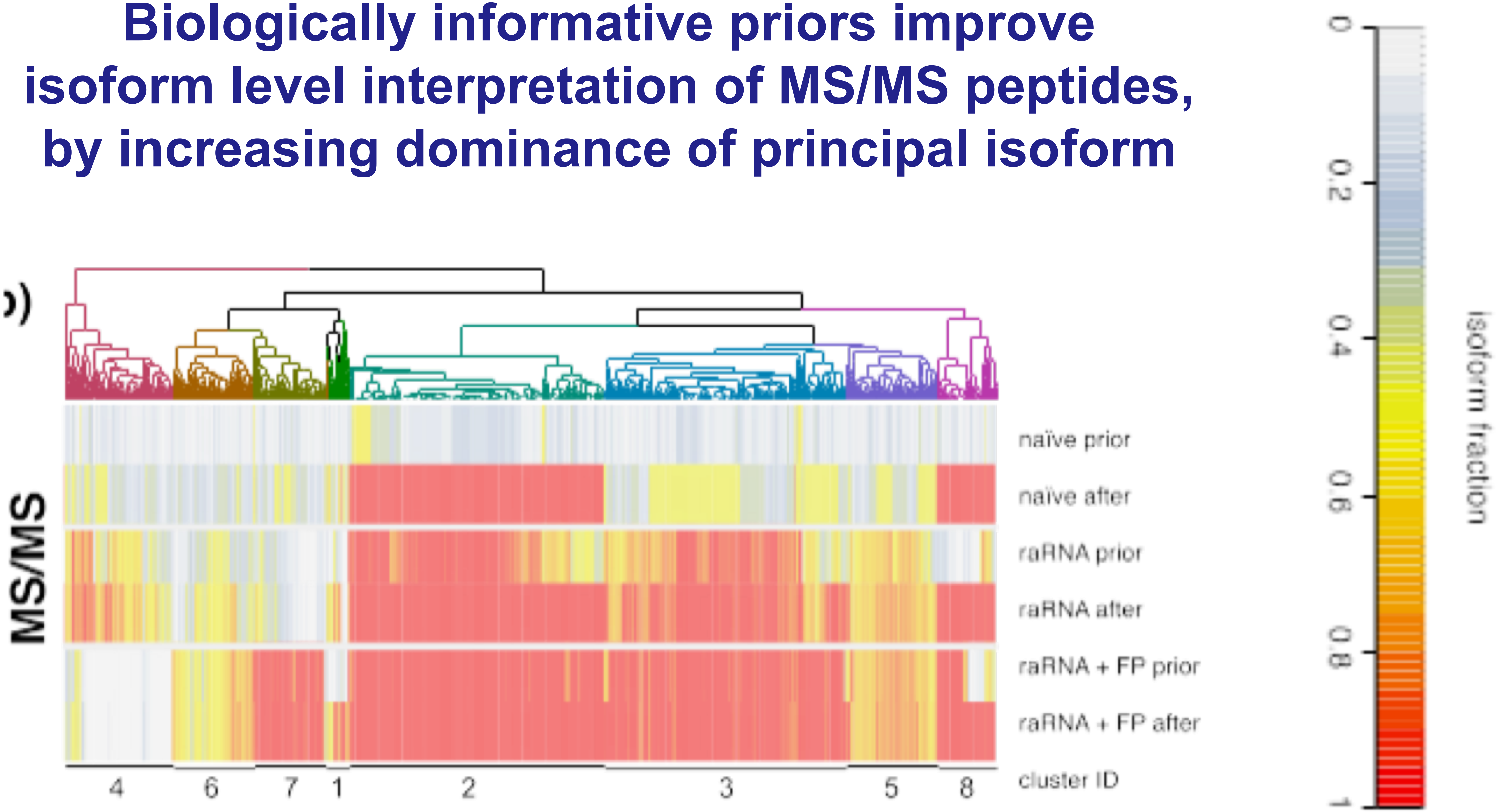


≤ principal isoform dominance
=principal isoform/second isoform

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

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

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

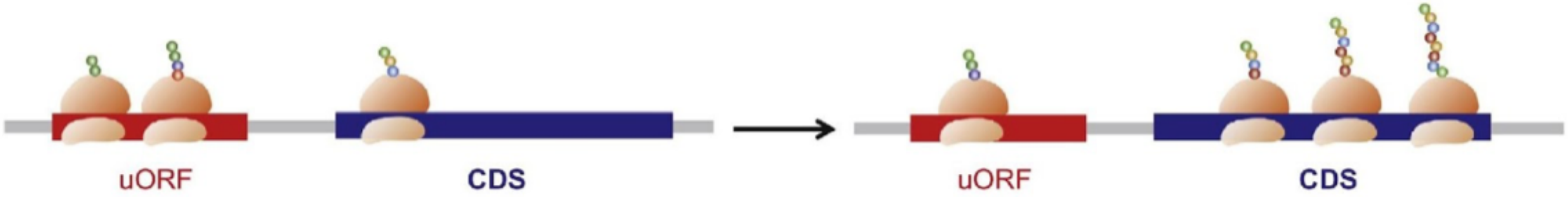


[**Carlyle, Kitchen** et al. (2018) *Journal of Proteome Research*]

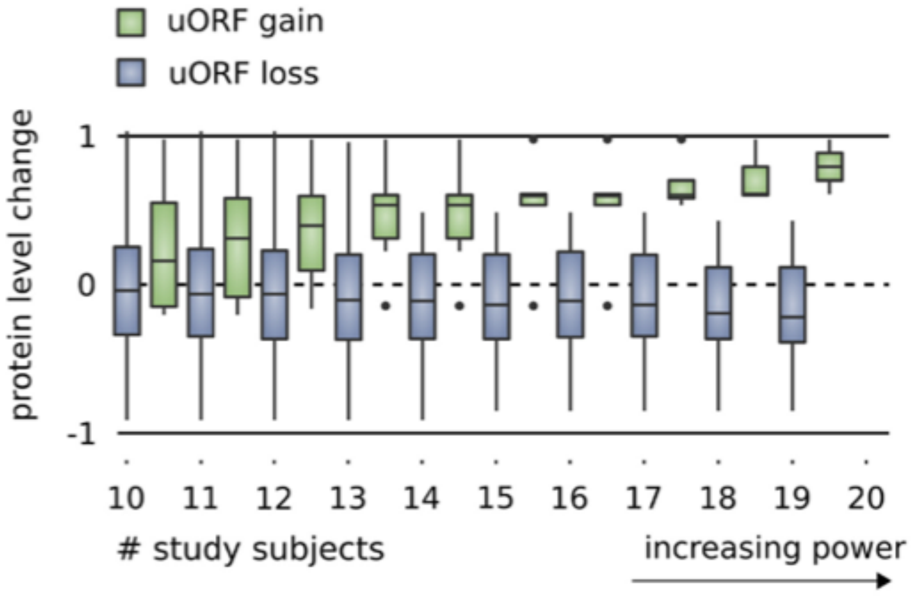
Outline: Comparing Protein & RNA Abundance

- **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

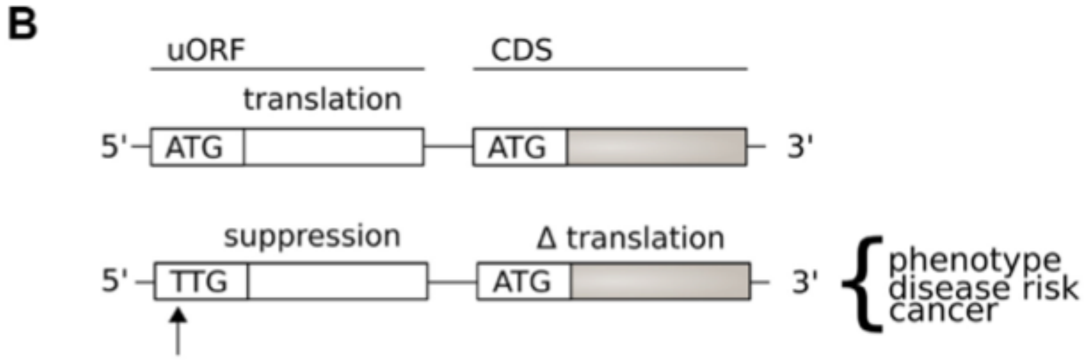
Upstream open reading frames (uORFs) may shift the expected balance between mRNA & protein



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



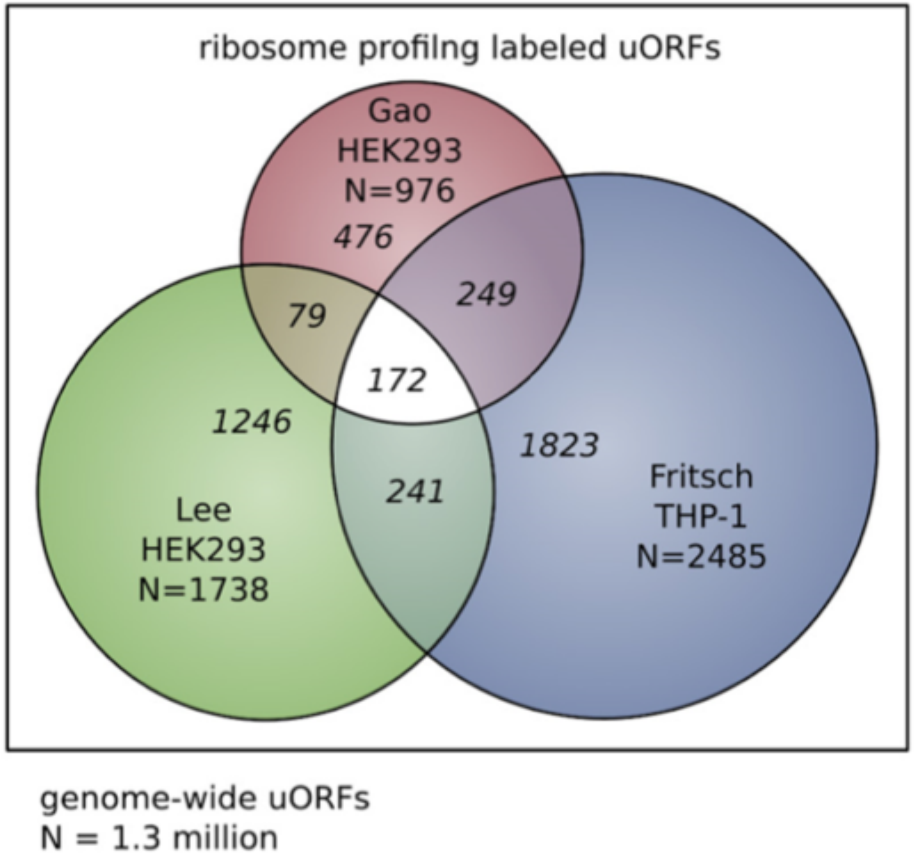
In Battle et al. 2014 data uORF gain & loss assoc. protein level change.



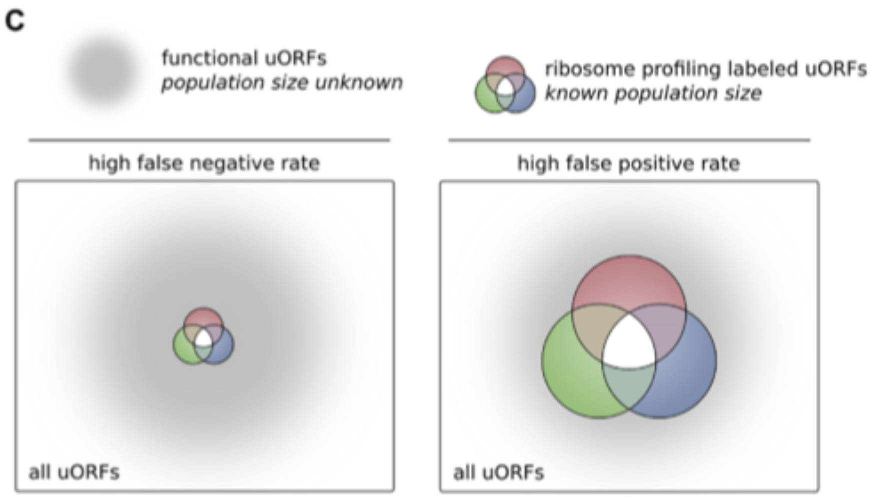
uORF regulation can be affected by mutation

[McGillivray et al., *NAR* (‘18)]

The population of functional uORFs may be significant



From a “Universe” of 1.3 M pot. uORFs

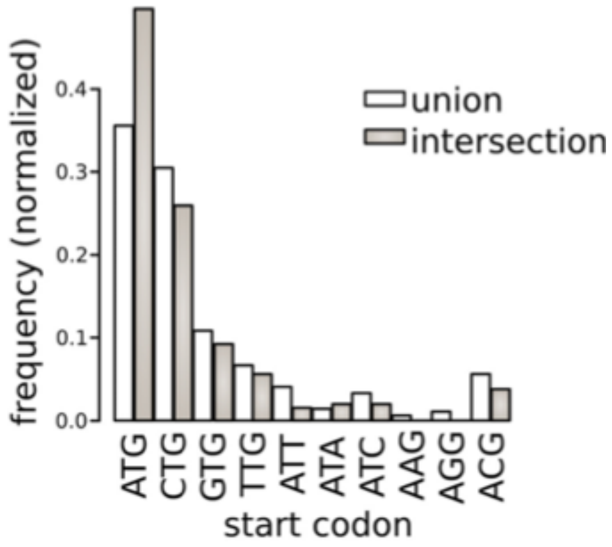


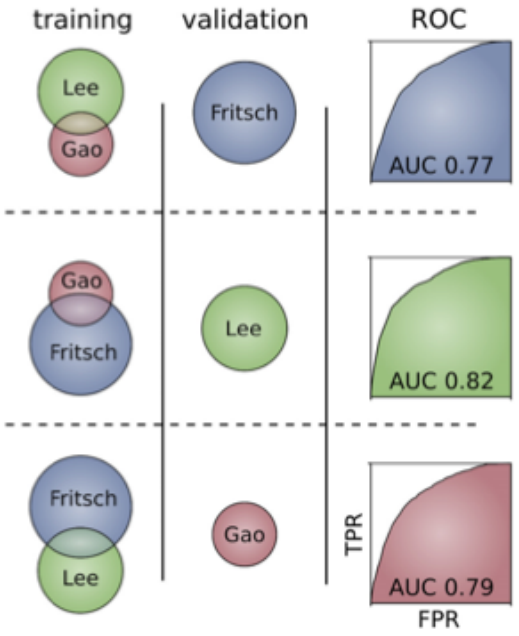
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

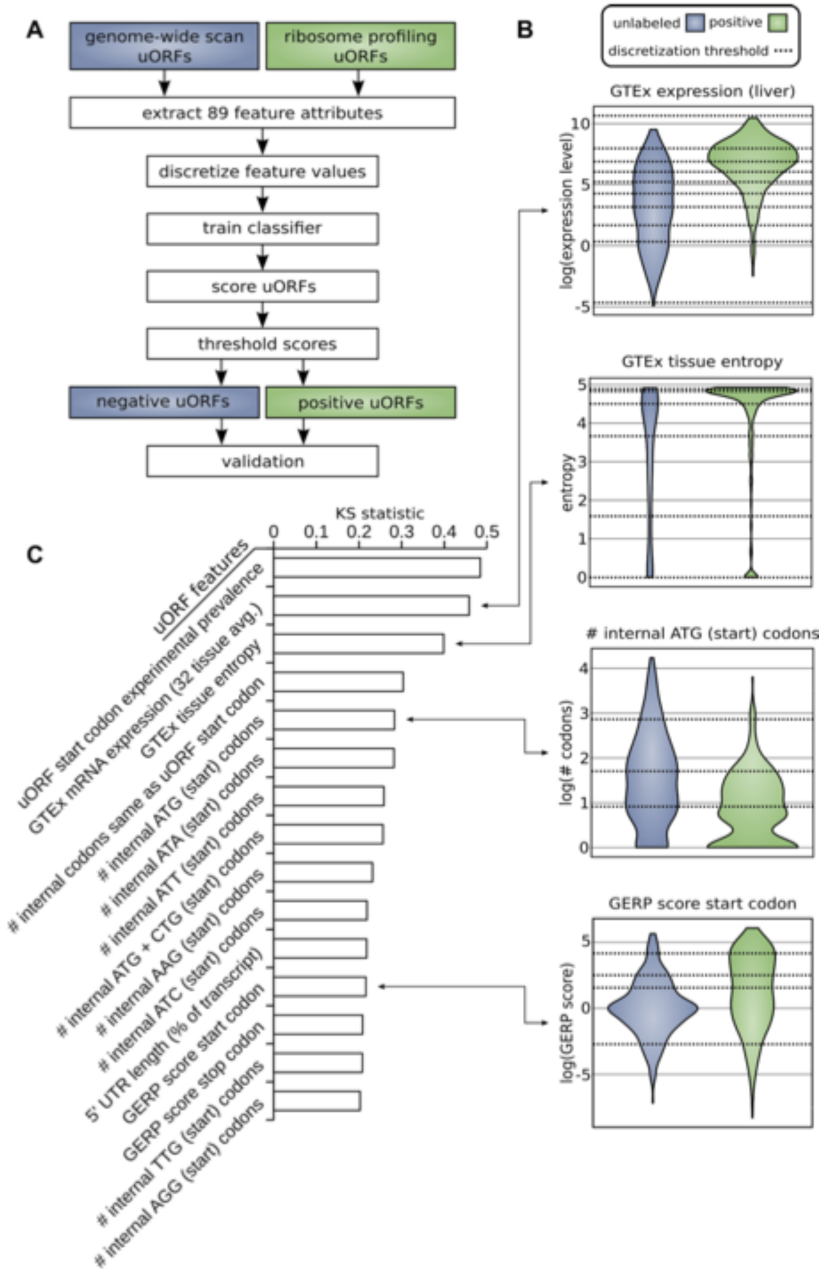
[McGillivray et al., *NAR* (‘18)]

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).







Expr. Level

Tissue Dist.

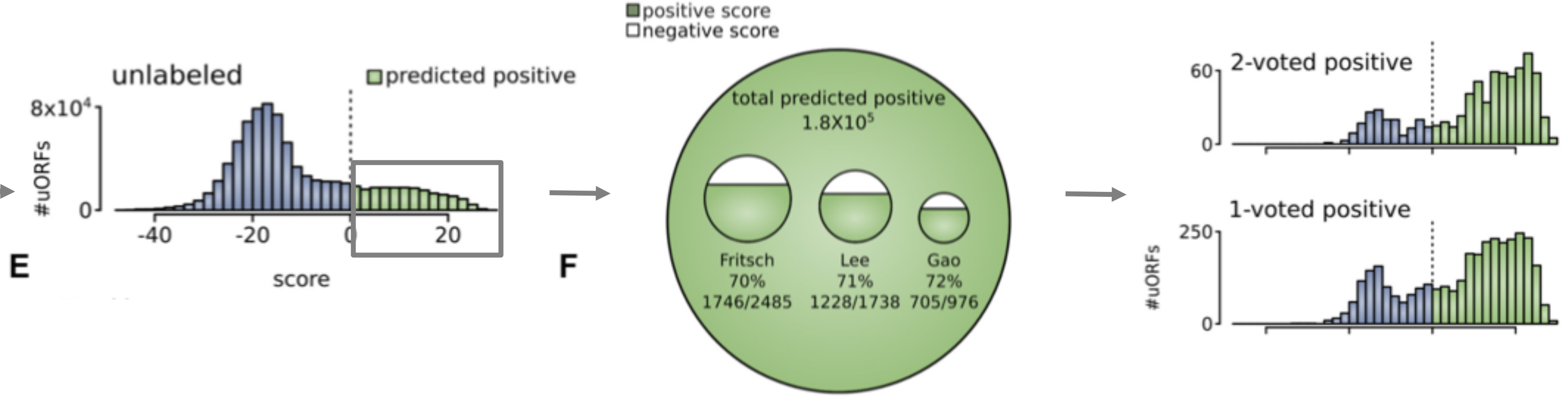
Int. ATG Start

Cons-ervation

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

A comprehensive catalog of functional uORFs

Universe of **1.3M** uORFs scored via Simple Bayes algo.



- 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

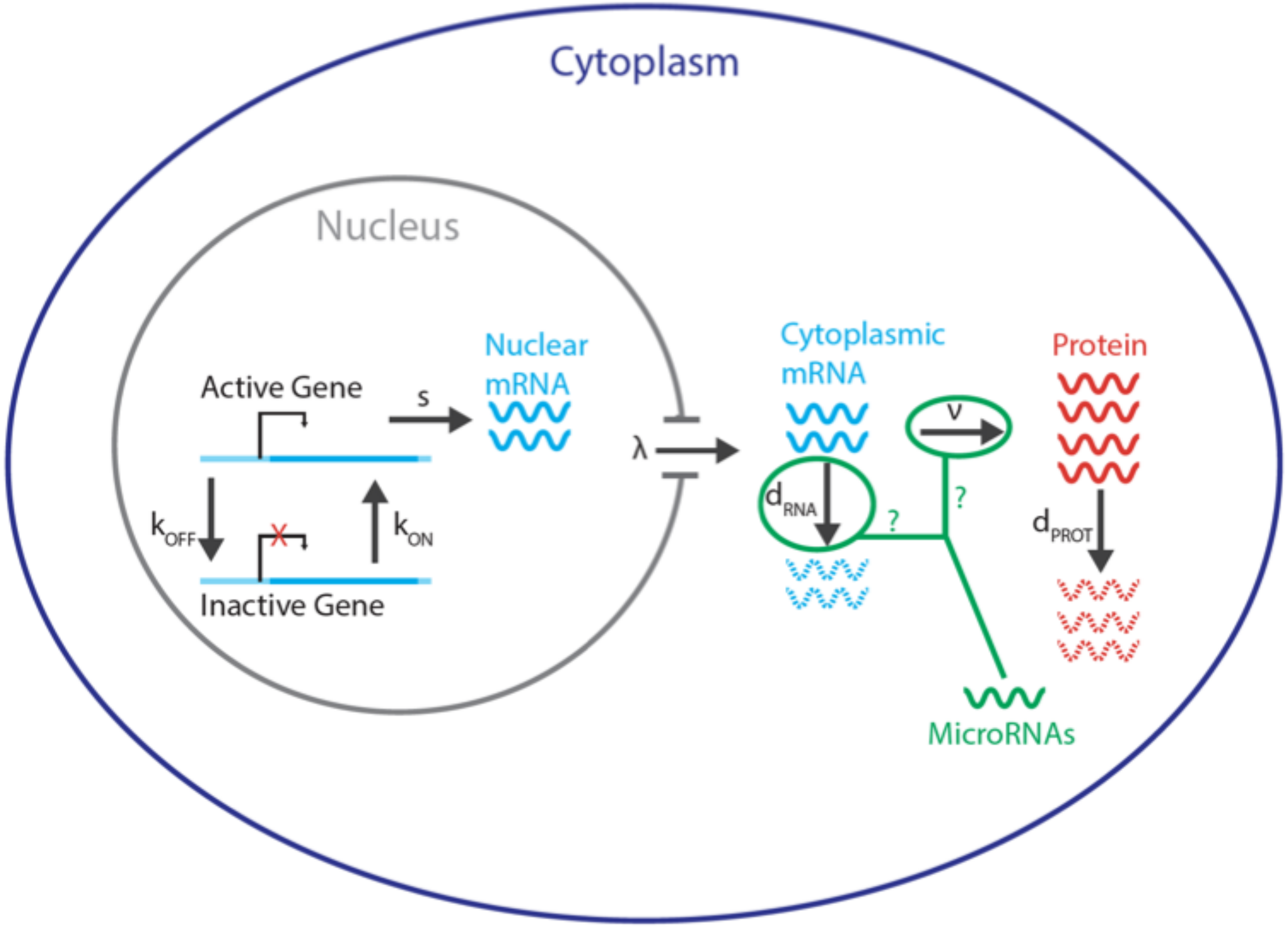
Outline: Comparing Protein & RNA Abundance

- **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

Schematic workflow

MicroRNA intervention



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**

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

Brainspan data

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