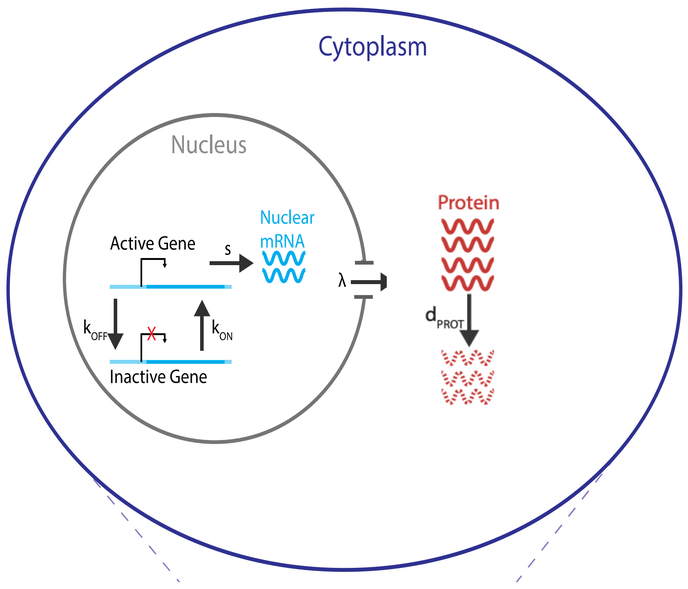
A joint model of cell-type-specific RNA and protein expressions in the human brain

Donglu Bai, Prashant S. Emani , Mark Gerstein*, Departments of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

Motivation and Workflow



- No proteomic amplification methods as the same resolution as single-cell RNA-seq
- To address this, we have leveraged scRNA-seq to aid in identification of cell-type-specific proteomic signatures by incorporating the stochastic nature of gene expression
- Our goal is to better understand the biological variation between cell-type specific mRNA and protein distributions in human brain tissue samples

Inference Target	Genomics Assay(s)
Gene activation & transcriptional rates	Single-cell RNA-seq (ROSMAP)
Deconvolution of cell-type proportions	Single-cell RNA-seq + Bulk-tissue RNA-seq (ROSMAP)
Cell-type protein translational rates	Proteomics (ROSMAP)
Train/test linear ML for inferring cell-type protein signatures	Proteomics (ROSMAP) + Cell-type proportions +transcriptional parameters

Methods and Approaches

Noisy Beta-Poisson Model

Based our model on the BPSC model of Vu et al*, but with a noise term $H(Y|x)$ added

$$Q(Y) = \sum_x H(Y|x).P(x)$$

$$P(x) = p_0 I(x=0) + (1-p_0)\sum_{m=1}^{M} p_m.BP\left(\frac{x}{\lambda_2^{(m)}}\middle|\alpha^{(m)},\beta^{(m)},\lambda_1^{(m)}\right)$$

$$H(Y|x) = Binomial(Y|x,p_T) = \binom{x}{Y}.(1-p_T)^{x-Y}.{p_T}^{Y}$$

Y = Experimentally Observed RNA counts
x = Underlying "true" RNA counts
p_T = Probability of counts dropping out

Transcriptional bursting rate: $s \equiv \lambda_1$
Probability of gene being "on": $p \equiv u$
Gene activation rate: $k_{on} \equiv \alpha$
Gene inactivation rate: $k_{off} \equiv \beta$

For our current implementation(m), we only consider a single mode

Deconvolution of cell-type proportions in mixtures

We deconvolve the bulk RNA-seq expression using the same snRNA-seq data we used for the parameter inference

Cell-type proportions ⟸ Signature matrix ⟸ CIBERSORTx / Bisque

Inference of cell-type specific protein translation rates

For solving the master equation for RNA and protein numbers, the equilibrium protein numbers are*:

$$\langle n\rangle_{gc} = \nu'_{Transl,gc}.s'_{gc}.\frac{1}{1+\frac{k'_{Off,gc}}{k'_{On,gc}}} = A_{gc}.\nu'_{Transl,gc}, \text{ where } A_{gc} = s'_{gc}.\frac{1}{1+\frac{k'_{Off,gc}}{k'_{On,gc}}}$$

For a complex tissue with multiple cell types:

$$\langle n\rangle_g = \sum_{c=1}^{N_{cell\ types}} w_c.\langle n\rangle_{gc} = \sum_{c=1}^{N_{cell\ types}} w_c.A_{gc}.\nu'_{Transl,gc}$$

Assumption: The translation rate is the same across individuals, but different across cell types and genes

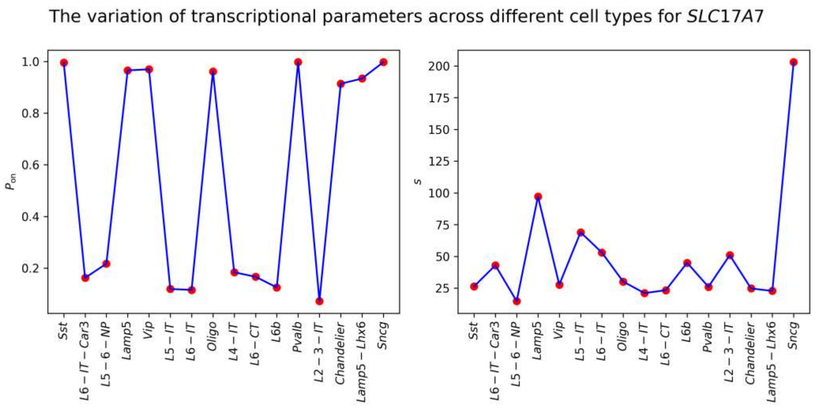
Train and test linear ML for inferring cell-type protein signatures

L2 regularization: Train → Regularization parameter tuning → Validation → Best regularization parameter (smallest RMSE) → Test

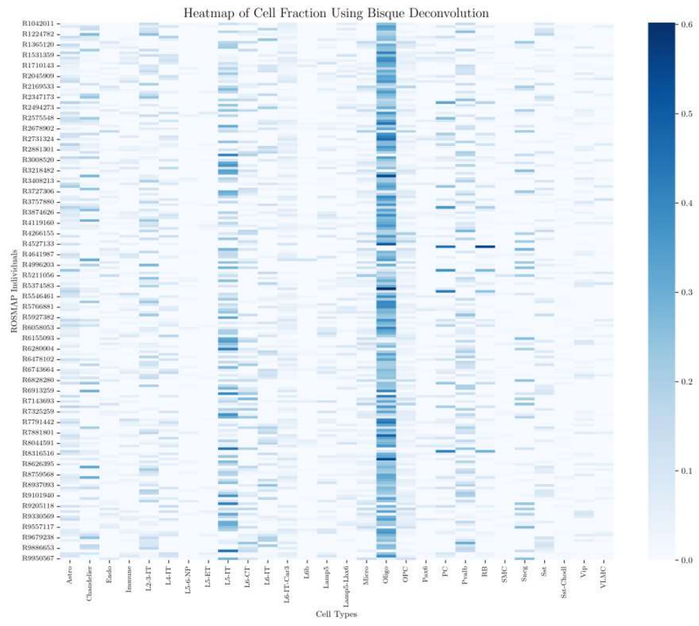
Results

Transcriptional parameters for marker gene *SLC17A7*

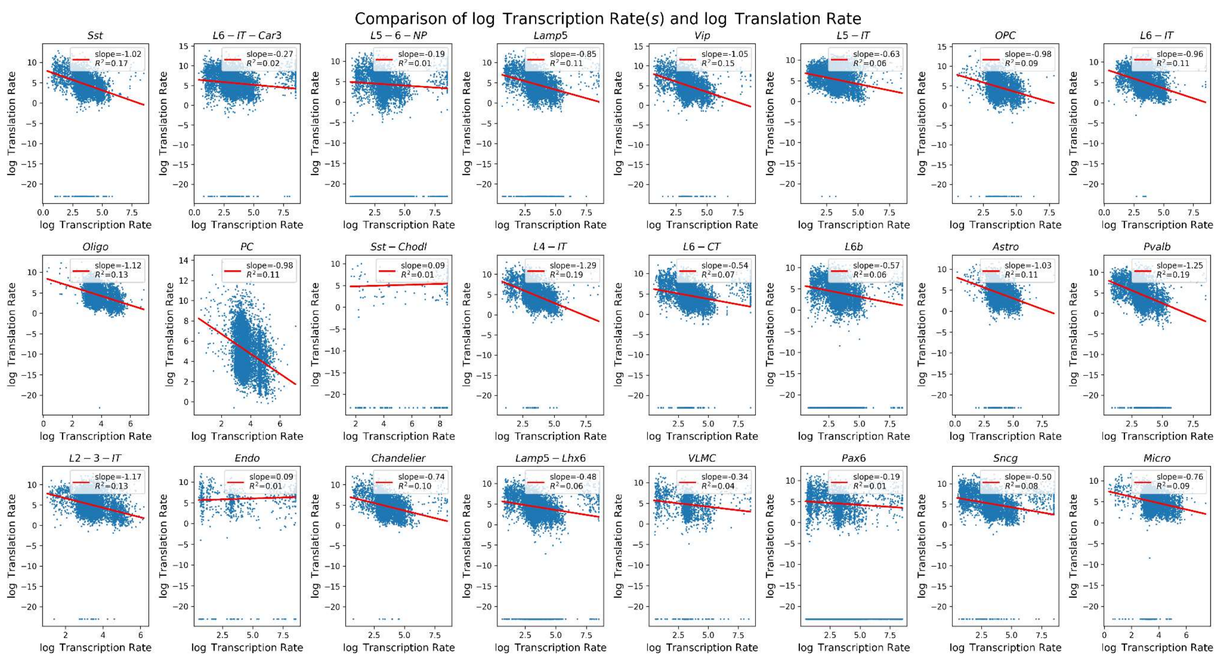
The probability that the gene is in the "on" state: $P_{on}=\frac{K_{on}}{K_{on}+K_{off}}$



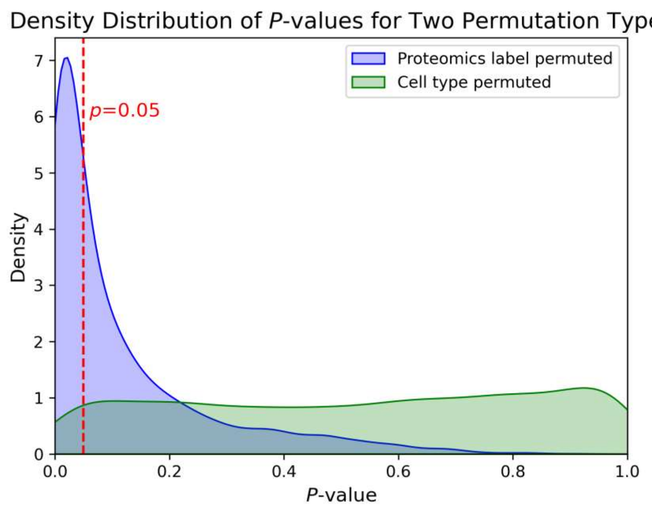
Deconvolution of cell-type proportions



Cell-type specific protein translation rates



Assessing test error (RMSE) in linear ML



Conclusions and Future Directions

- *SLC17A7* is an excitatory neuron marker, yet it shows high P_{on} and s for inhibitory neurons. This shows that our inferred transcriptional rates may lack biological accuracy. We may need to adjust parameter fitting for low-expression profiles or use single-cell ATAC-seq for better constraint
- Bisque deconvolution confirms expected cell-type proportions
- Low slopes and R square indicate a minor negative correlation between transcriptional and translational rates, as expected
- The permuted labels in proteomics have a significant effect given the high density of low p-values. In contrast, the permuted cell type labels doesn't show as strong an effect given its broader p-value distribution

Bennett, David A., et al. "Religious orders study and rush memory and aging project." *Journal of Alzheimer's disease* 64.s1 (2018): S161-S189.
Vu, Trung Nghia, et al. "Beta-Poisson model for single-cell RNA-seq data analyses." *Bioinformatics* 32.14 (2016): 2128-2135.
Peccoud, Jean, and Bernard Ycart. "Markovian modeling of gene-product synthesis." *Theoretical population biology* 48.2 (1995): 222-234.
Shahrezaei, Vahid, and Peter S. Swain. "The stochastic nature of biochemical networks." *Current opinion in biotechnology* 19.4 (2008): 369-374.