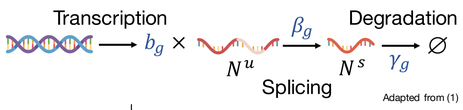
Biophysically Informed Modeling for Mapping Effects of Genetic and Environmental Perturbation on Cell State

Tara Chari[1], Gennady Gorin[2], Lior Pachter[1,3]

[1]Division of Biology and Biological Engineering, California Institute of Technology
[2]Division of Chemistry and Chemical Engineering, California Institute of Technology
[3]Department of Computing and Mathematical Sciences, California Institute of Technology

CME-Based Models of Transcription



Chemical model of splicing and degradation with bursty transcription with **gene (g) specific parameters**

Uses unspliced (u) and spliced (s) molecule counts

$$M^u \sim Poiss(N^u C^u L_g)$$
$$M^s \sim Poiss(N^s \lambda^s)$$

Model for sampling of molecules during cDNA library creation (10x Genomics), with **gene-length dependent** capture[1]

$$X^u \sim Bin(M^u p^u)$$
$$X^s \sim Bin(M^s p^s)$$

Model for sampling of molecules during the PCR amplification/sequencing process with **global sampling parameters**[1]

Sampled molecule count random variable

Solution for probability of molecule counts at steady state:

From $\frac{dP(x^u,x^s,t)}{dt}$ obtain $P(x^u,x^s,t;b_g,\beta_g,\gamma_g,C^u,\lambda^s,L^g)$

Obtain PGF of $P(x^u,x^s,t)$: $H(y^u,y^s,t)$ where $\lim_{t\to\infty} H(y^u,y^s,t) = H(y^u,y^s)$

$$P(x^u,x^s) \approx iFFT\left(H\left(e^{\frac{-2\pi u}{U}}, e^{\frac{-2\pi s}{S}}\right)\right)$$

where $u = 0,..,U-1, s = 0,...,S-1$ and U,S = observed max of X^u, X^s respectively[1-3]

Perturbation Dataset

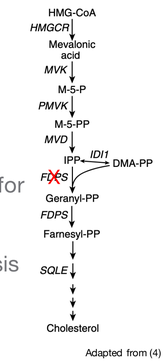
Combinatorial Perturb-Seq CRISPR screen[4]:

- 37,238 cells
- 86 perturbation guides (single or double targets)
- 6 control (non-targeting) guides

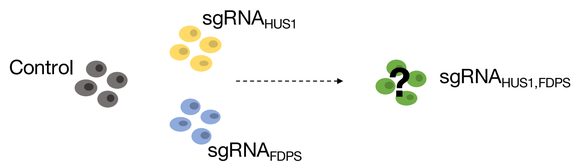
HUS1/FDPS Combinatorial Knockouts:

HUS1 – Component of 9-1-1 cell-cycle checkpoint for cell cycle arrest in response to DNA damage

FDPS – Intermediate gene in cholesterol biosynthesis pathway



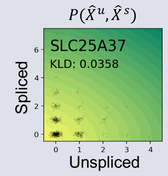
Effect of Perturbation in HUS1/FDPS Knockouts:



Methods

1. Combine all control samples to **fit global parameters** using a grid search over parameter values
   a) Set C^u, λ^s to pair of grid values
   b) **Fit gene parameters** for combined sample conditional on C^u, λ^s by optimizing KL divergence from the observed molecule count distributions

$$KL(P(\hat{X}^u,\hat{X}^s), P(X^u,X^s;b_g,\beta_g,\gamma_g,C^u,\lambda^s,L_g)) \longrightarrow$$

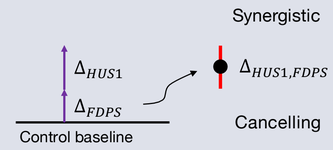


   c) Determine **best global parameters** with min KL divergence

2. **Fit gene parameters**, under the global parameter set, for Control, sgRNA$_{HUS1}$, and sgRNA$_{FDPS}$ cells. Optimize KL divergence from the observed molecule count distributions

3. **Predict change in parameters** with log-additive effect model:
*All parameter values are in $\log_{10}$

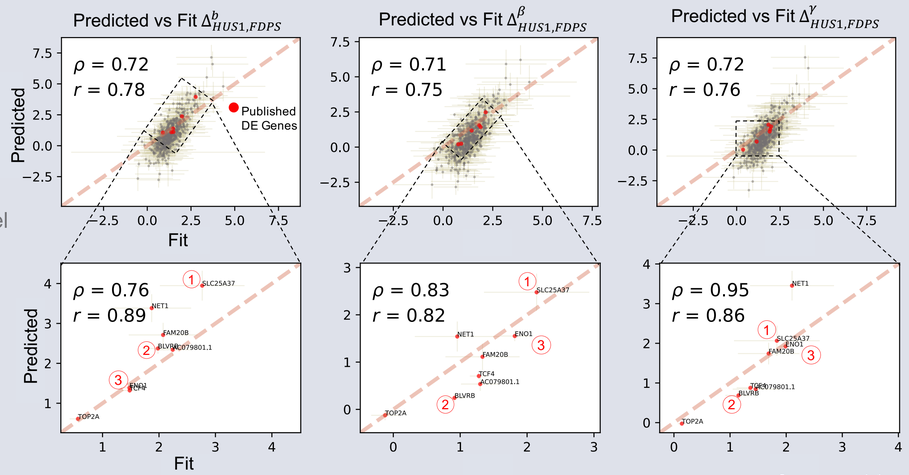
$$(1)\ \Delta^b_{HUS1,FDPS} = (b_{HUS1} - b_{CTRL}) + (b_{FDPS} - b_{CTRL})$$
$$(2)\ \Delta^\beta_{HUS1,FDPS} = (\beta_{HUS1} - \beta_{CTRL}) + (\beta_{FDPS} - \beta_{CTRL})$$
$$(3)\ \Delta^\gamma_{HUS1,FDPS} = (\gamma_{HUS1} - \gamma_{CTRL}) + (\gamma_{FDPS} - \gamma_{CTRL})$$



Results

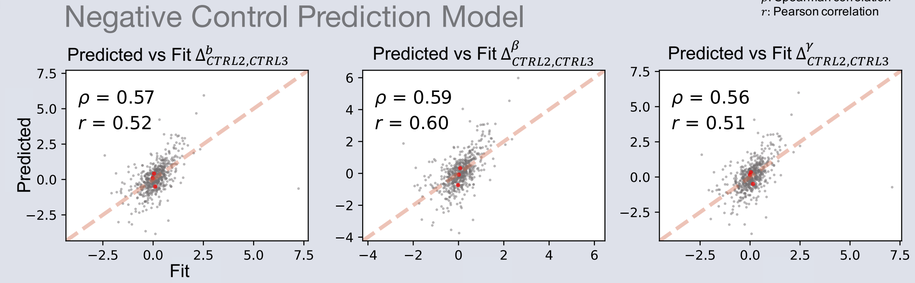
- Predicted $\Delta_{HUS1,FDPS}$ ~75% correlated with $\Delta_{HUS1,FDPS}$ calculated from the CME model fit
- Differentially expressed (DE) genes from original study display parameter shifts
- Negative control prediction model, where $\Delta_{CTRL2,CTRL3}$ calculated for two control samples, has ~50% correlation to calculated model fit
- Highlights other candidate genes with distinct (predicted) $\Delta_{HUS1,FDPS}$

Log-Additive Prediction Model



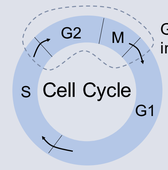
ρ: Spearman correlation
r: Pearson correlation

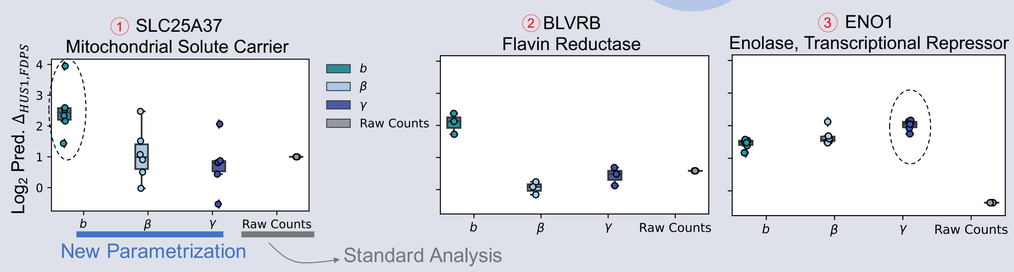
Negative Control Prediction Model



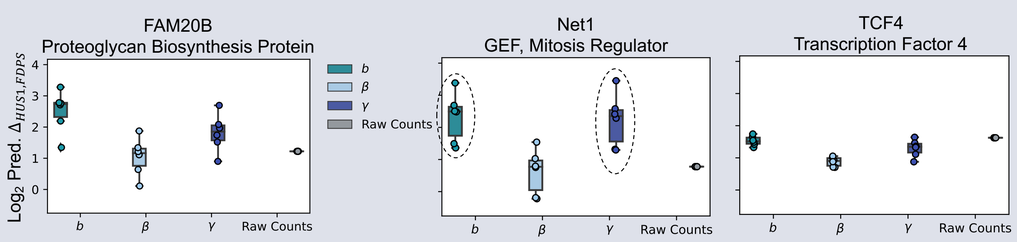
Perturbed Gene Candidates

DE Candidates from Study





Other Candidates from $\Delta_{HUS1,FDPS}$ Prediction Model



Discussion

We **expand gene analysis from one feature (spliced counts) to three** physically meaningful parameters. This is a step towards **interpretable perturbation modeling/prediction**, where predictions have physical meaning beyond phenomenological expression counts[5].

Predicting combinatorial perturbations limits the number of experiments necessary, as instead we can **use *in silico* calculations to predict mechanistic changes**.

In future, we plan to **expand prediction** to other CRISPR datasets, drug/environmental perturbations, cross-species prediction, and interpolation of perturbation effects across time. We can also expand beyond the naïve, log-additive model demonstrated here.

References

1. Gorin, Gennady, and Lior Pachter. 2021. https://doi.org/10.1101/2021.07.30.454514
2. Bokes, Pavol, John R. King, Andrew T. A. Wood, and Matthew Loose. 2012. https://doi.org/10.1007/s00285-011-0433-5
3. Singh, Abhyudai, and Pavol Bokes. 2012. https://doi.org/10.1016/j.bpj.2012.07.015
4. Replogle, Joseph M., Thomas M. Norman, Albert Xu, Jeffrey A. Hussmann, Jin Chen, J. Zachery Cogan, Elliott J. Meer, et al. 2020. https://doi.org/10.1038/s41587-020-0470-y
5. Gorin, Gennady, John J. Vastola, Meichen Fang, and Lior Pachter. 2021. https://doi.org/10.1101/2021.09.06.459173

Data Availability

All code and analysis can be found at https://github.com/tarachari3/perturbCME.git

Contact:
tchari@caltech.edu