Hurdle Models for Single Cell Gene Expression

Andrew McDavid

Department of Statistics, University of Washington
and
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center
anmcd@uw.edu

@EquivMeasures

July 1, 2016

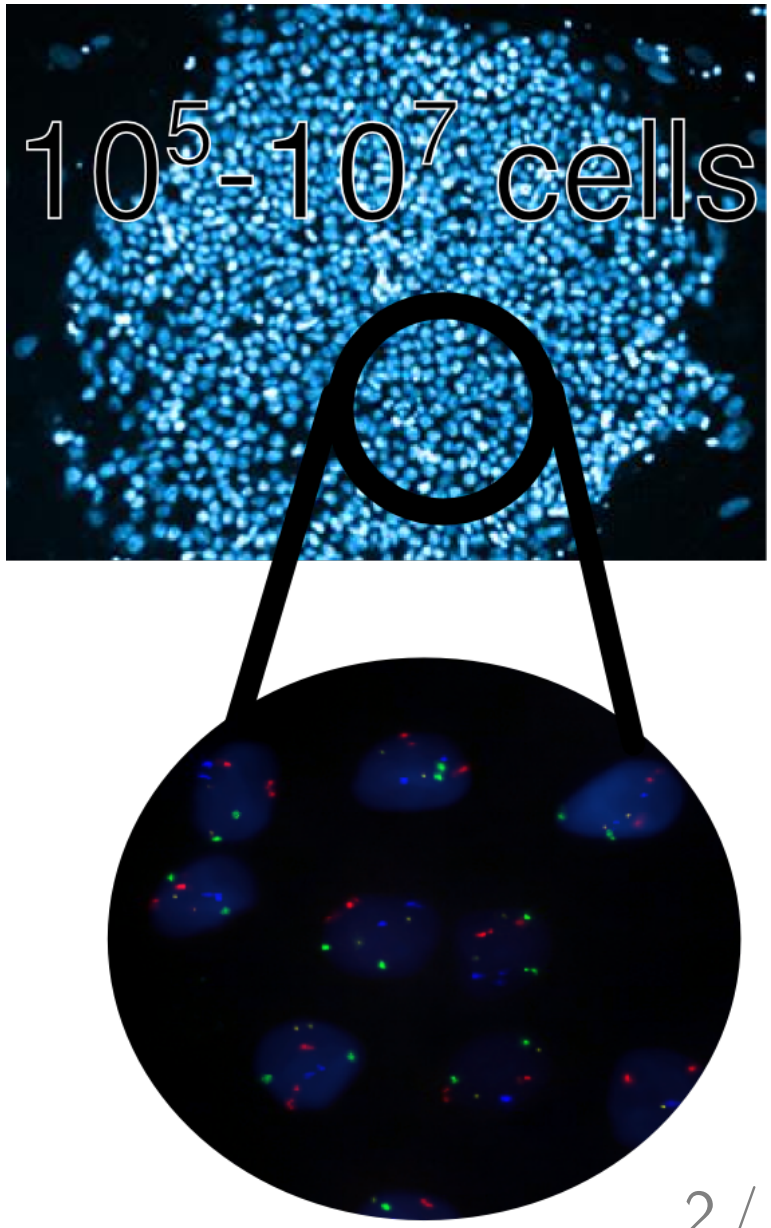
FRED HUTCH
CURES START HERE™

Why single cells?

$\mathbf{Y}_i$ vector of expression values.
Bulk gene expression: $\sum_i \mathbf{Y}_i$.
But what about:

- The cell-to-cell variance of each gene (Var Y_j)?
- Clusters of cells or latent structure (E[$\mathbf{Y}|Z$])?
- Cellular coexpression (Cov $\mathbf{Y}$)? or probabilistic independences?

Biological averaging has convolved over variables of interest.



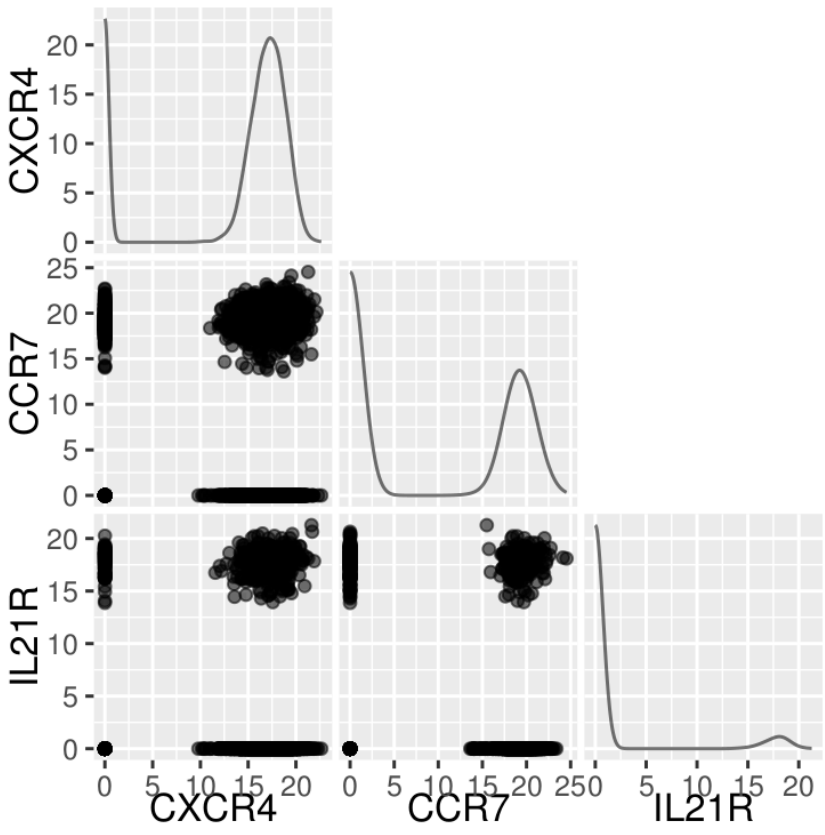
Bimodality and single cell gene expression

A defining characteristic is **bimodality** in expression (Flatz 2011, Powell 2012, McDavid 2013, Marinov 2014).

Some (gene dependent) fraction of the time, little or no expression is detected.

Given detection, expression is symmetric and bounded away from zero.

Fluidigm qPCR



40 - Cycle threshold (Ct)

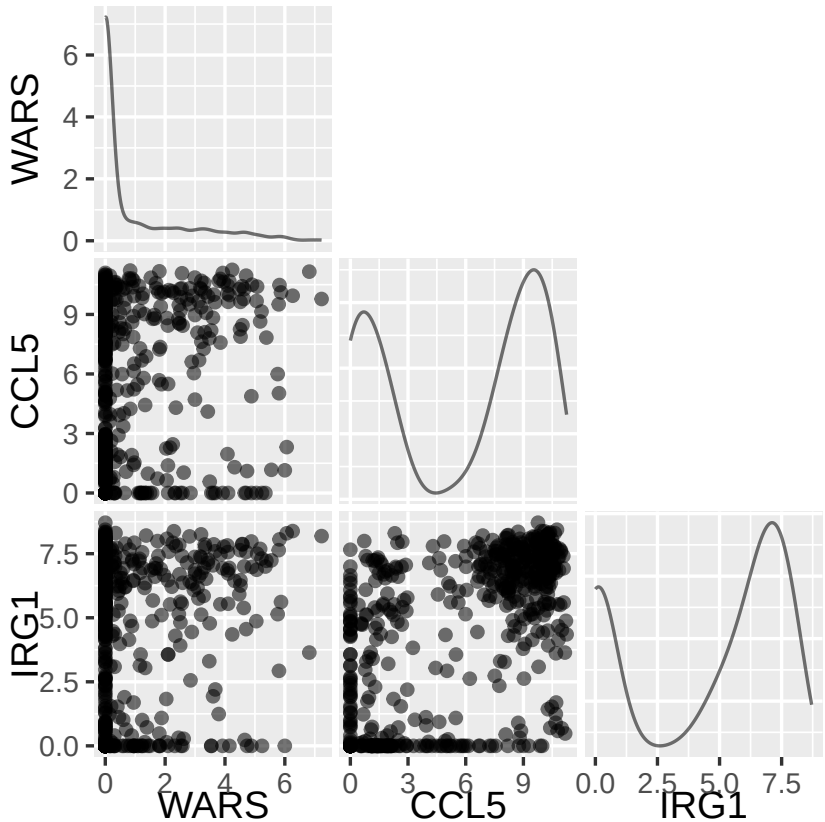
Bimodality and single cell gene expression

A defining characteristic is **bimodality** in expression (Flatz 2011, Powell 2012, McDavid 2013, Marinov 2014).

Some (gene dependent) fraction of the time, little or no expression is detected.

Given detection, expression is symmetric and bounded away from zero.

RNAseq



log_2(transcripts per million +1).

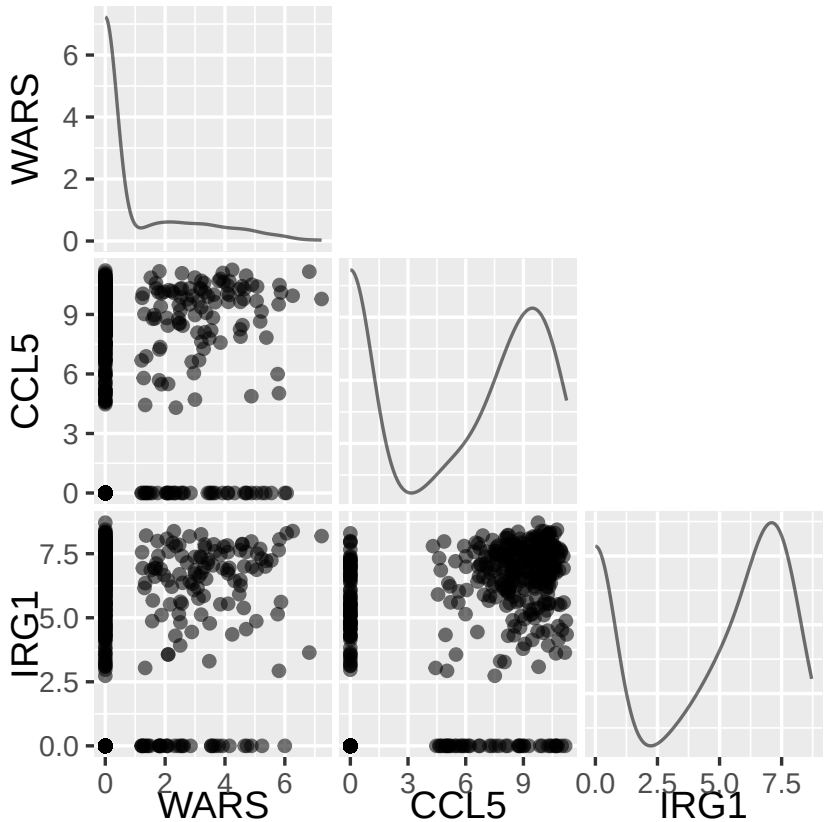
Bimodality and single cell gene expression

A defining characteristic is **bimodality** in expression (Flatz 2011, Powell 2012, McDavid 2013, Marinov 2014).

Some (gene dependent) fraction of the time, little or no expression is detected.

Given detection, expression is symmetric and bounded away from zero.

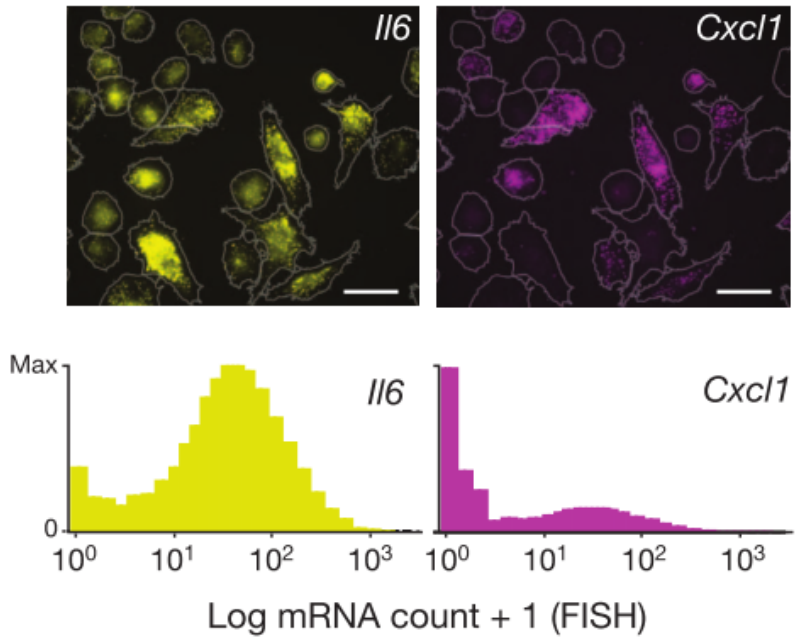
RNAseq, thresholded



log_2(transcripts per million +1)

Cause of bimodality

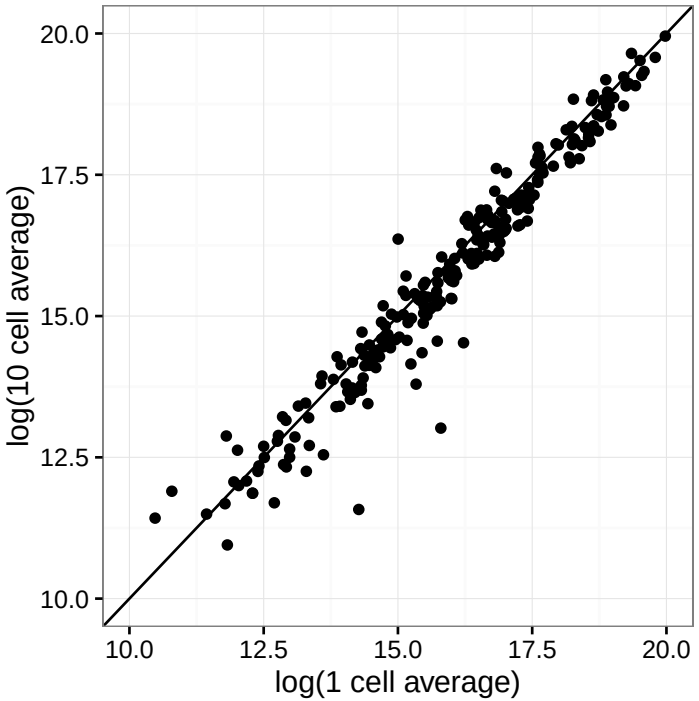
- Fluorescent *in-situ* hybridization experiments: mRNA often **zero-inflated, log-normal** distributed.
- Transcription occurs in bursts while DNA is uncoiled and accessible, followed by stochastic decay.
- Consistent with zero-inflation of single-cell qPCR and sequencing.



Shalek, *et al*, 2013

Are zeros limits of detection or censoring?

- N 10-cell equivalents $\Rightarrow$ $10N$ the expression of a single cell equivalent
- Single molecule capture efficiency varies from 90% to 20%

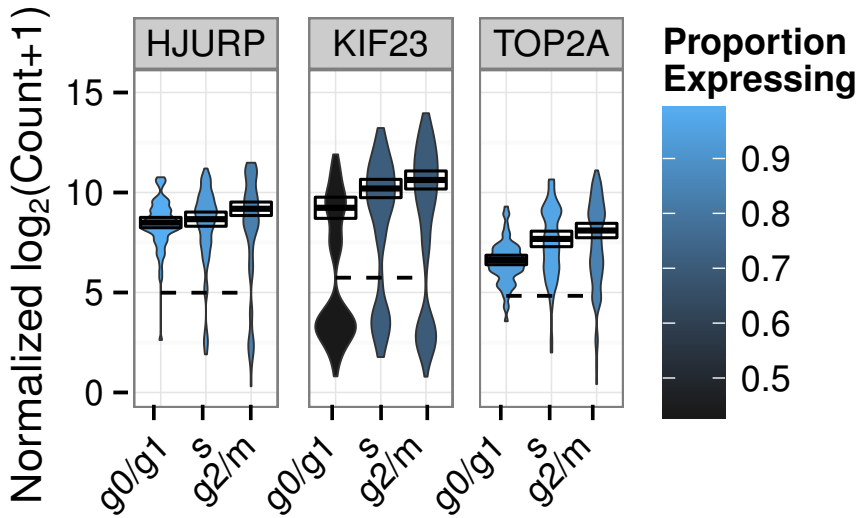


$$E\left(y_{(1)}\right) = E\left(\frac{1}{10N}\sum_{i}^{N} 2^{y_{(10),i}}\right)$$

▸ Similar relationships for the frequency of expression

Hurdle models

- Both rate of zeros and mean of log-normal vary according to biological treatments, generally in tandem.
- Phenomenological model: accommodate, rather than explain.

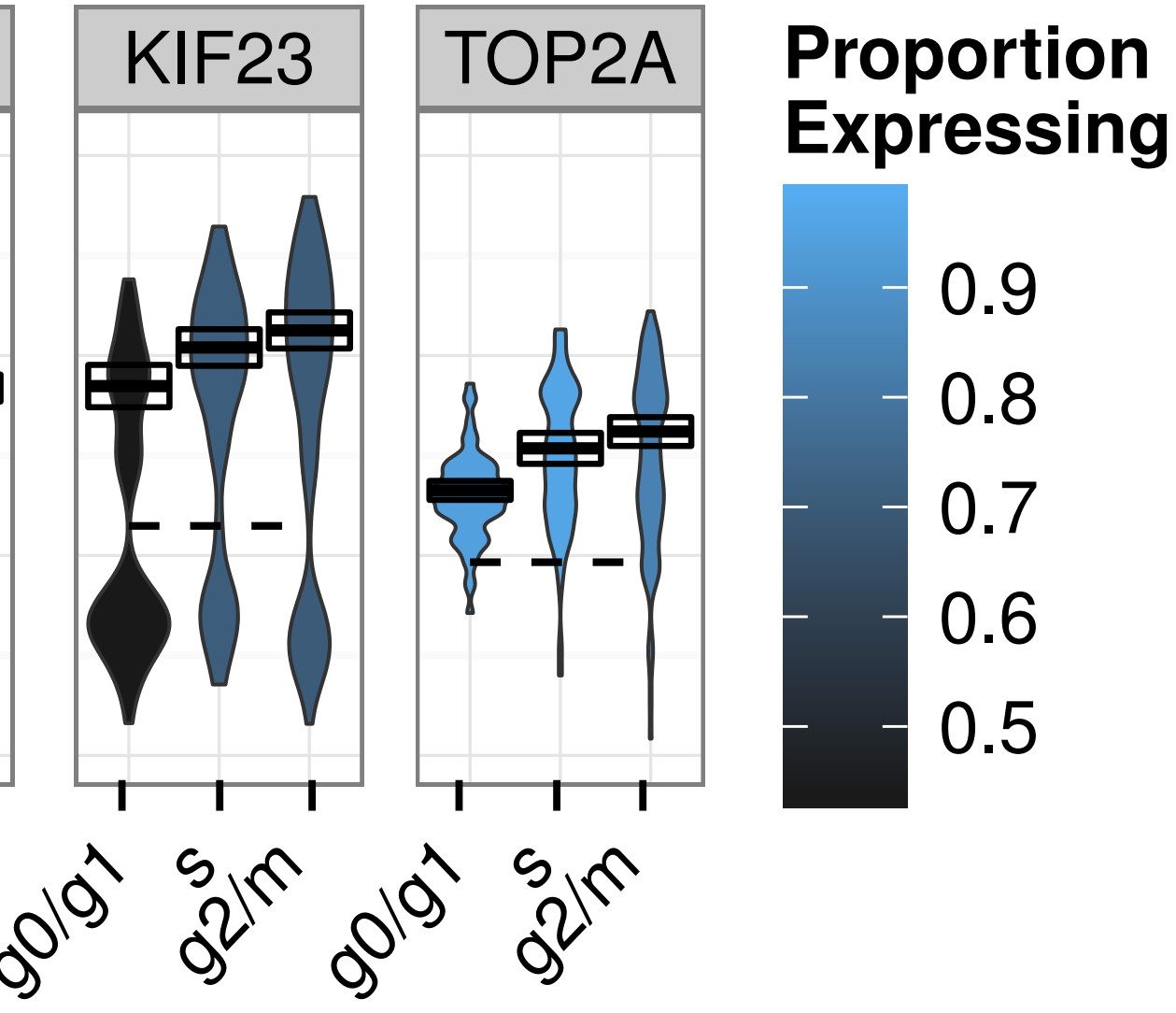


Hurdle model

Y_i is log-expression in cell i. Then

$$Y_i = U_i V_i \quad \text{and} \quad U_i \perp V_i,$$
$$U_i \sim \text{Normal}(\mu_i, \tau^2),$$
$$V_i \sim \text{Bernoulli}(p_i).$$

Hurdle models



Hurdle linear model

Let

$$\mu_i = \mathbf{X}_i^T \beta,$$
$$\text{logit}\, p_i = \mathbf{X}_i^T \beta'$$

be linear functions of covariates. Then we can do ANOVA and linear regression using the Hurdle model.
The log-likelihood of a sample of n cells, given $\mu_i(\beta)$ and $p_i(\beta')$ is

$$\mathcal{L}(\mu_i, p_i; \mathbf{y}) = \sum_{i=1}^{n} \underbrace{\left[1_{[y_i \neq 0]} \text{logit}\, p_i + \log(1 - p_i)\right]}_{\text{Bernoulli}} +$$
$$\sum_{i:y_i \neq 0} \underbrace{-1/2 \log\left(\tau^2 2\pi\right) - 1/2 \left[\frac{y_i - \mu_i}{\tau}\right]^2}_{\text{Normal}}$$

Cell cycle experiment (Dennis, *et al* [2014])

- 333 genes, 930 cells, sorted by cell cycle (G0/G1, S, G2/M)
- 119 known, **ranked** genes associated with cell cycle from a bulk expression data base (cyclebase.org)
- Compare number of ranked and unranked genes discovered at a given P-values using:

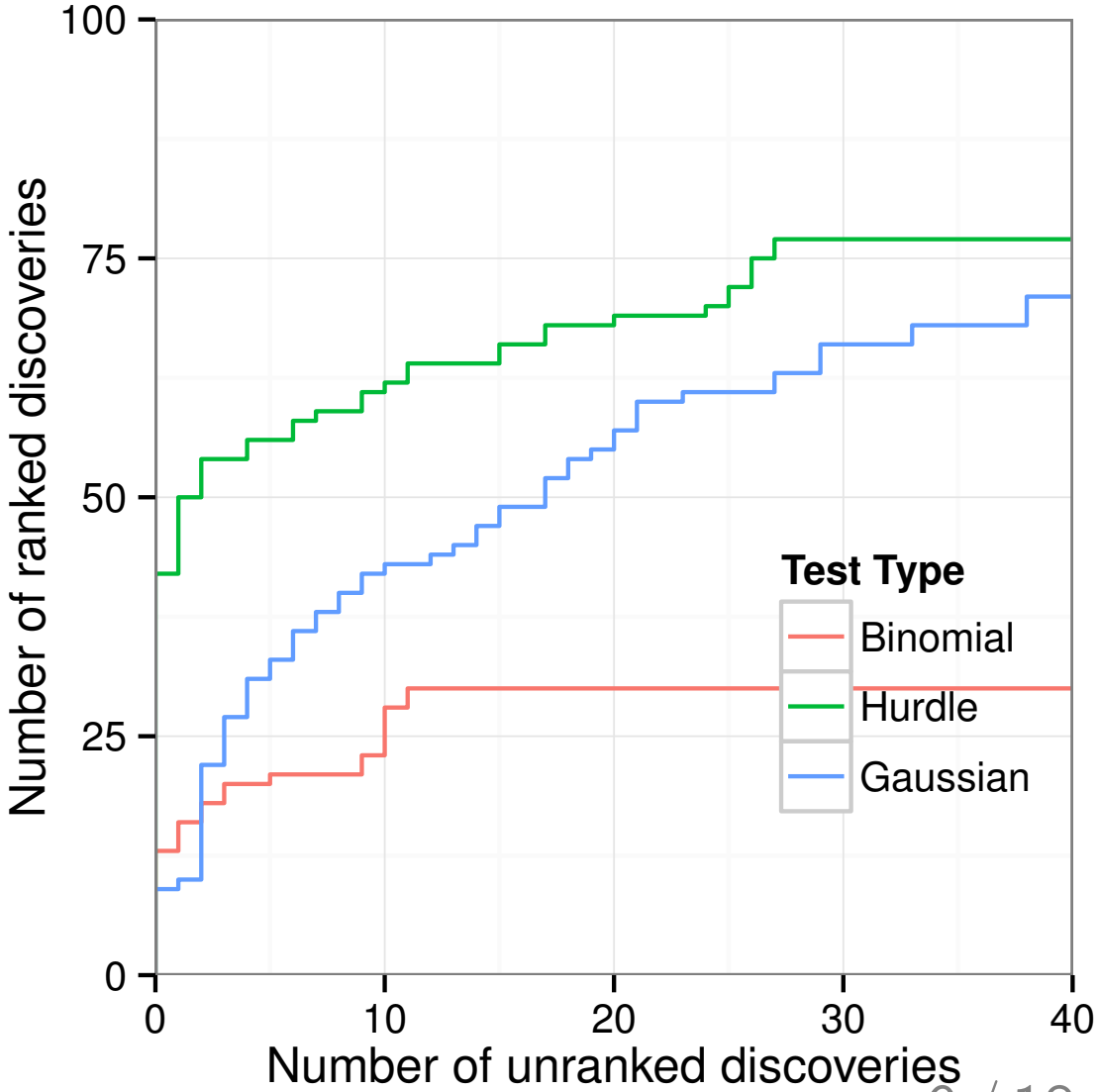
  Binomial: logistic regression on $1_y \equiv 1_{[y \neq 0]}$
  Gaussian: linear regression on y
  Hurdle: joint regressions on 1_y and y

Performance

Binomial logistic regression on $1_y \equiv 1_{[y \neq 0]}$

Gaussian linear regression on y

Hurdle joint regressions on 1_y and y



Hurdle model extensions and applications

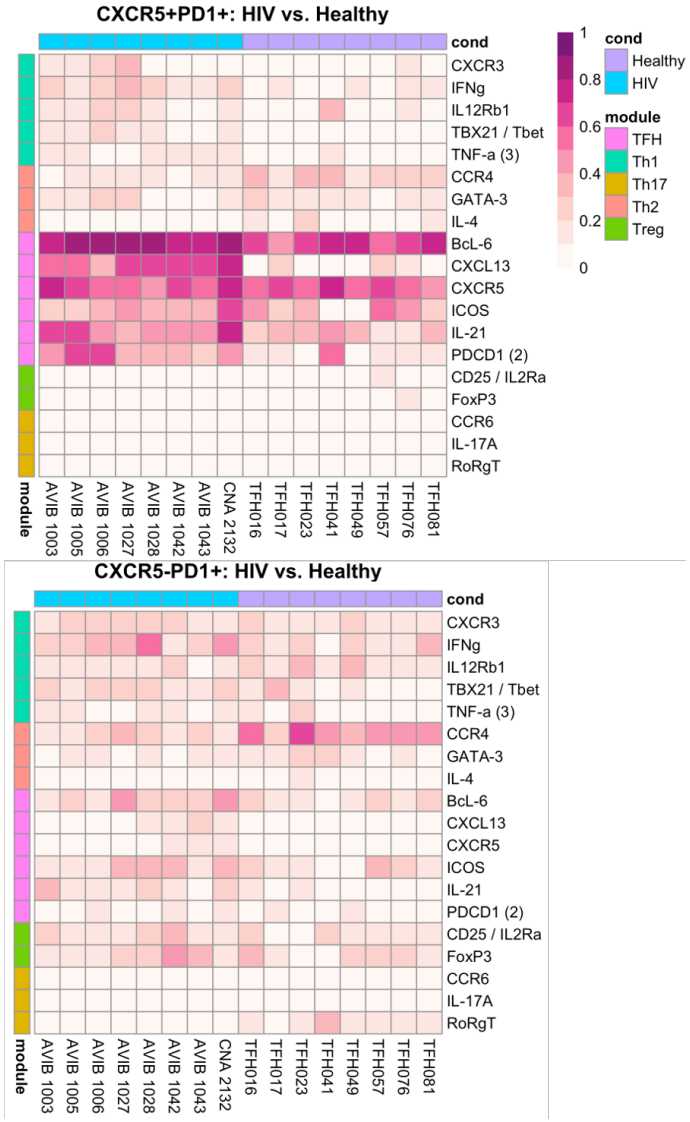
1. **Empirical Bayesian regularization** to borrow strength across genes.

$$U_{ij} \sim \text{Normal}(\mu_{ij}, \tau_j^2),$$
$$\tau_j^2 \sim \text{Inverse-Gamma}(a, b).$$

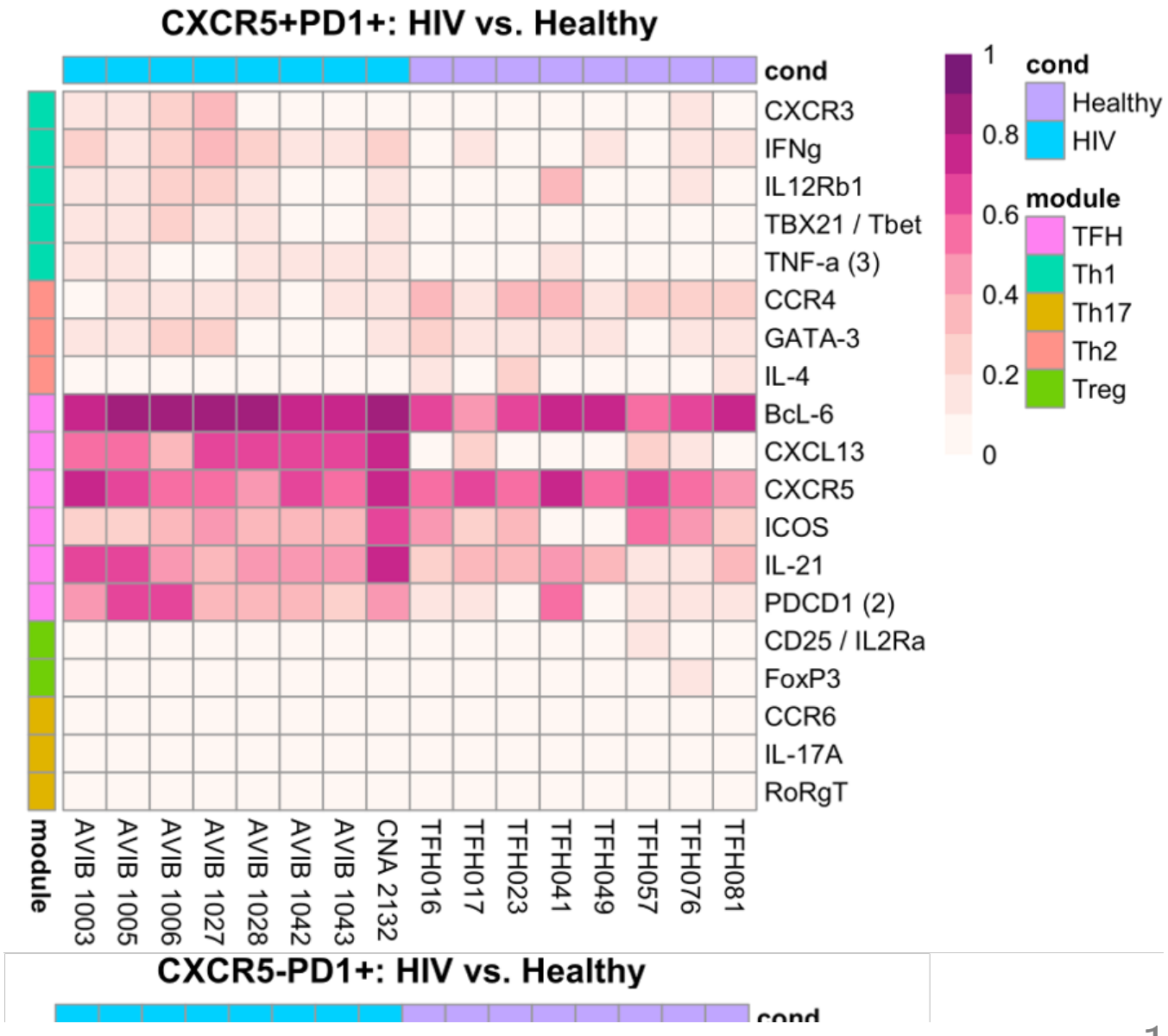
2. **Stablity under linear separation** with Cauchy prior on logistic coefficients.
3. **Mixed models**, in which the between-individual and within-individual variability is parametrized.
4. Parametric graphical modeling on zero-inflated data to estimate gene-gene interactions
5. Competitive **gene set enrichment** analysis.

Tfh and HIV (Swiss Institute for Vaccine Research)

- Scientific question: how does HIV alter the expression profile of Tfh-maturation and signaling genes?
- 16 donors, recent HIV naive to anti-retroviral therapy, and healthy controls. Lymph biopsies.
- Two cell populations: $CXCR5^{-}PD1^{+}$, $CXCR5^{+}PD1^{+}$ (Tfh)

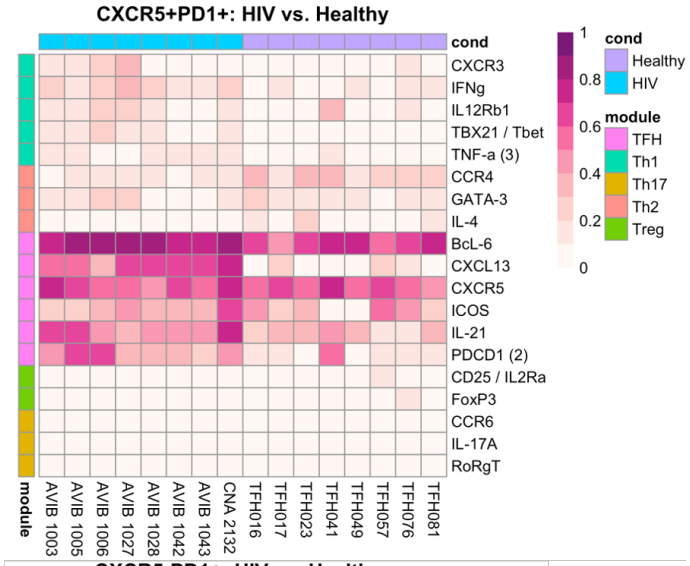


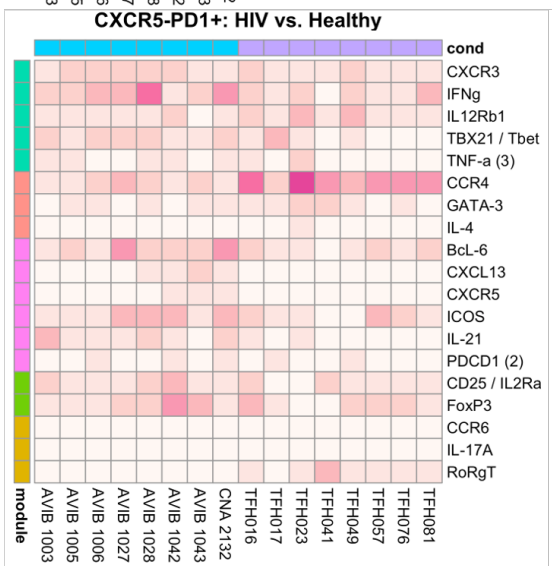
Tfh and HIV (Swiss Institute for Vaccine Research)



Tfh and HIV (Swiss Institute for Vaccine Research)

- Scientific question: how does HIV alter the expression profile of Tfh-maturation and signaling genes?
- 16 donors, recent HIV naive to anti-retroviral therapy, and healthy controls. Lymph biopsies.
- Two cell populations: $CXCR5^{-}PD1^{+}$, $CXCR5^{+}PD1^{+}$ (Tfh)
- Statistical question: do Tfh genes differ **on average** compared to non-Tfh genes in HIV+ vs healthy controls?





Competitive Gene Set Enrichment

1. Vector of expression estimates $\hat{\beta}_g$ and $\hat{\beta}'_g$ for genes $g = 1, \ldots, N_G$.
2. Geneset $\mathbf{C}$

$$[C_g] = \begin{cases} 1 & \text{Gene g is in set} \\ 0 & \text{Else} \end{cases}$$

   and its complement $\mathbf{D} = 1 - \mathbf{C}$.
3. Expression in the set vs expression outside the set:

$$\delta = \frac{\mathbf{C}^T \hat{\boldsymbol{\beta}}}{\|\mathbf{C}\|_1} - \frac{\mathbf{D}^T \hat{\boldsymbol{\beta}}}{\|\mathbf{D}\|_1}$$

   by comparing δ to Normal$(0, \text{Var}(\delta))$.
4. Need an estimate of Var(δ).

Var(δ) and non-independence

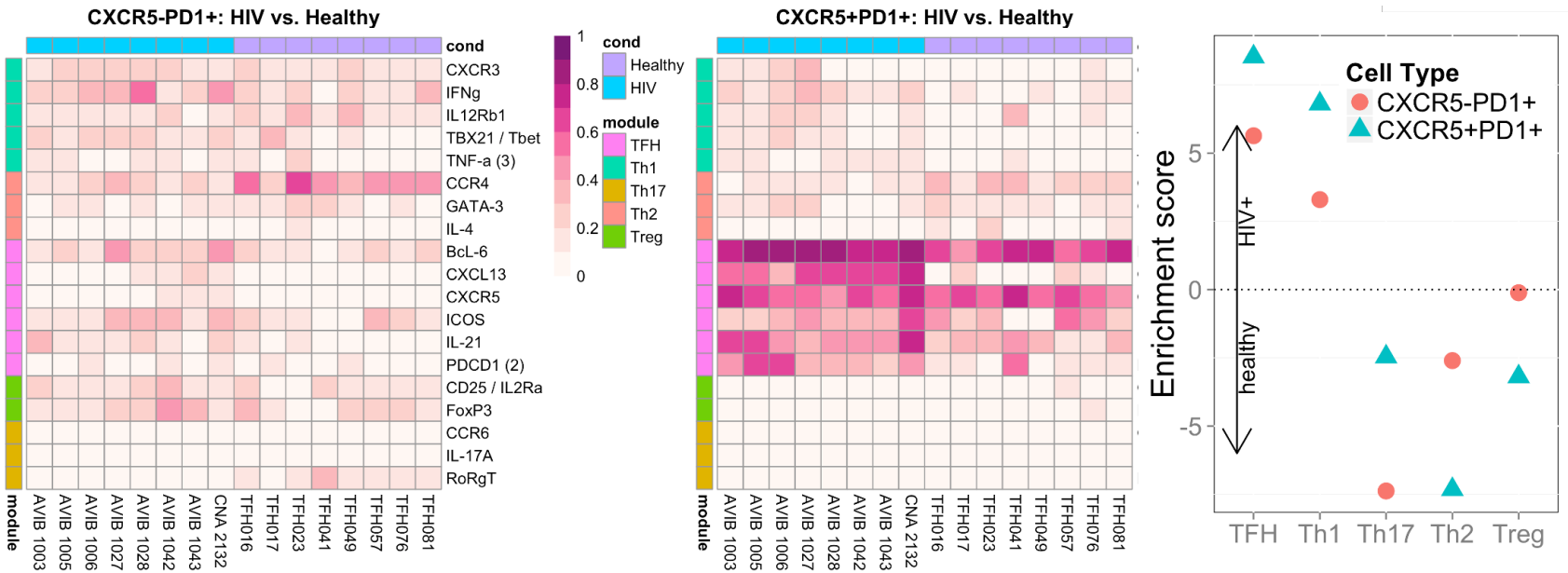
Expression between genes dependent, so $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) \neq 0$ in general.
Estimate covariance matrix $\mathbf{\Lambda} = [\lambda_{ij}] = \text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$, then

$$\text{Var}\left(\frac{\mathbf{C}^T\hat{\boldsymbol{\beta}}}{\|\mathbf{C}\|_1}\right) = \frac{\mathbf{C}^T\mathbf{\Lambda}\mathbf{C}}{\|\mathbf{C}\|_1^2}.$$

Var(δ) and non-independence

Expression between genes dependent, so $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) \neq 0$ in general.
Estimate covariance matrix $\mathbf{\Lambda} = [\lambda_{ij}] = \text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$, then

$$\text{Var}\left(\frac{\mathbf{C}^T\hat{\beta}}{\|\mathbf{C}\|_1}\right) = \frac{\mathbf{C}^T\mathbf{\Lambda}\mathbf{C}}{\|\mathbf{C}\|_1^2}.$$



Proximate and future work

- Power/sample size calculations
- Gene expression matrices $\mathbf{Y}_{ik}$ in donor k for condition i over genes $j = 1, \ldots, J$.

  Test condition effect $\beta_i \neq 0$ **over the donor super-population**.

  Super-population variability $\text{Var}(\beta_{ij})$ might be similar between genes, like shrinkage models for dispersions: `limma`, `deseq2`, `edgeR`, etc.
- More useful decompositions of parameters for Hurdle models
- Clustering on zero-inflated data

More Reading

- Finak G., McDavid A., Yajima M, *et al* (2015). *MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome Biology.
- Dennis, L., McDavid, A., Danaher, P., *et al* (2014). *Modeling bi-modality improves characterization of cell cycle on gene expression in single cells.* PLoS Computational Biology.
- McDavid, A., Finak, G., Chattopadyay, P. K., *et al* (2013). *Data Exploration, Quality Control and Testing in Single-Cell qPCR-Based Gene Expression Experiments.* Bioinformatics.
- `http://github.com/RGLab/MAST` – use branch `summarizedExpt`

Goal

- Learn how to filter, explore and test for differential expression.
- Join me in eating this delicious dog food.
- Package to be submitted to Bioconductor for fall release, on github in the meantime.

MAITAnalysis Vignette

```
devtools::install_github('RGLab/MAST@summarizedExpt')
library(MAST)
vignette('MAITAnalysis')
file.edit(system.file('doc/MAITAnalysis.R',
          package='MAST'))
```

Infelicities

- Want a unique key for rows (ENSEMBLE ids vs entrez ids vs UCSC transcript ids)
- Also want a human-readable default key for plots

Infelicities

- Want a unique key for rows (ENSEMBLE ids vs entrez ids vs UCSC transcript ids)
- Also want a human-readable default key for plots
- Hard to tell what your contrasts are with `model.matrix`. Easy to get the wrong answer. Important for power calculations.

Infelicities

- Want a unique key for rows (ENSEMBLE ids vs entrez ids vs UCSC transcript ids)
- Also want a human-readable default key for plots
- Hard to tell what your contrasts are with `model.matrix`. Easy to get the wrong answer. Important for power calculations.
- `heatmap`, `heatmap2`

Infelicities

- Want a unique key for rows (ENSEMBLE ids vs entrez ids vs UCSC transcript ids)
- Also want a human-readable default key for plots
- Hard to tell what your contrasts are with `model.matrix`. Easy to get the wrong answer. Important for power calculations.
- `heatmap`, `heatmap2`, `pheatmap`

Infelicities

- Want a unique key for rows (ENSEMBLE ids vs entrez ids vs UCSC transcript ids)
- Also want a human-readable default key for plots
- Hard to tell what your contrasts are with `model.matrix`. Easy to get the wrong answer. Important for power calculations.
- `heatmap`, `heatmap2`, `pheatmap`, `aheatmap`

Infelicities

- Want a unique key for rows (ENSEMBLE ids vs entrez ids vs UCSC transcript ids)
- Also want a human-readable default key for plots
- Hard to tell what your contrasts are with `model.matrix`. Easy to get the wrong answer. Important for power calculations.
- `heatmap`, `heatmap2`, `pheatmap`, `aheatmap`, `complexheatmap`

Infelicities

- Want a unique key for rows (ENSEMBLE ids vs entrez ids vs UCSC transcript ids)
- Also want a human-readable default key for plots
- Hard to tell what your contrasts are with `model.matrix`. Easy to get the wrong answer. Important for power calculations.
- `heatmap`, `heatmap2`, `pheatmap`, `aheatmap`, `complexheatmap`, `heatmapTheAwakening`

Acknowledgments

External Collaborators

Mario Roederer
Lucas Dennis
Martin Prlic
Giuseppe Pantaleo
Dan Lu and Bill Robinson

Fred Hutch and UW Statistics

Raphael Gottardo
Greg Finak
Masanao Yajima

R01 EB008400 from the National Institute of Biomedical Imaging and Bioengineering