RNA-Seq and Single-Cell RNA-Seq Tertiary Analysis

Yue Zhang[1,2] and Ramesh Nair[2]

[1]Genetics Bioinformatics Service Center (GBSC) and [2]Stanford Center for Genomics and Personalized Medicine (SCGPM)

RNA-Seq Tertiary Analysis

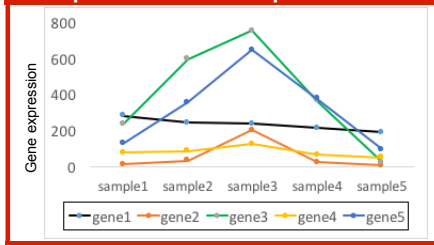
Weighted Gene Co-expression Network Analysis (WGCNA) R Package

1.WGCNA Publication

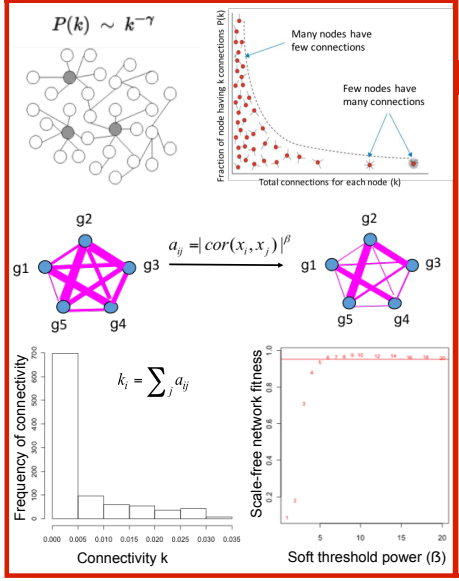
Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

Citations: **1354**

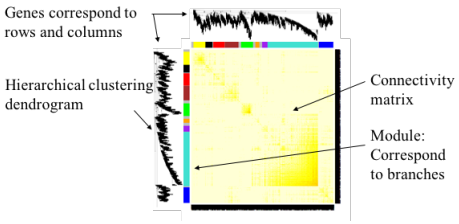
2. Input Data – Gene Expression Matrix



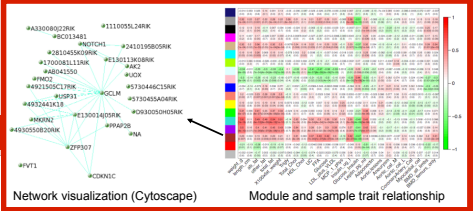
3. Scale-free Network Fitness



4. Module Detection



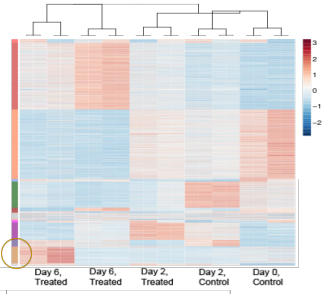
5. Relate Module With Sample Trait



6. BaaS Use Case

BaaS project:
Drug-induced pathway disruption for cardiac differentiation.

Input data and clean data:
FPKM matrix in log scale was generated from study samples.
~10K genes remained after filtering out low expression genes.



GO:0072358	cardiovascular system development
GO:0007507	heart development

Single-Cell RNA-Seq Tertiary Analysis

Single-Cell Differential Expression (SCDE) R Package

1. SCDE Publications

Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
Fan, J. *et al*. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13**, 241–244 (2016).

2. Input Data – Raw Count Matrix

	cell1	cell2	cell3	cell4	cell5
gene1	0	2	0	0	0
gene2	12	34	52	24	0
gene3	0	456	756	365	0
gene4	0	68	76	1	1
gene5	0	18	0	14	0

- Cells labeled green are similar and from same cell type.
- Genes labeled red are dropout events.

3. Statistical Methods

1. Find a group of nearest neighbor cells for each cell using Pearson correlation.
2. Estimate expected expression for each gene in each cell.
3. Build probabilistic model with the observed expression value of each gene in each cell modeled as a mixture of a dropout and an amplified component.

$$p(x \mid r_c, \Omega_c) = p_d(x) p_{Poisson}(x) + (1 - p_d(x)) p_{NB}(x \mid r_c)$$

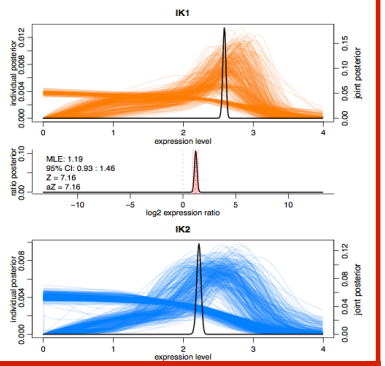
4. Generate a corresponding weight matrix measuring dropout event probability which is used for weighted PCA

	cell1	cell2	cell3	cell4	cell5
gene1	0	0.002	0	0	0
gene2	0.003	0.034	0.052	0.024	0
gene3	0	0.156	0.589	0.265	0
gene4	0	0.109	0.276	0.009	0
gene5	0	0.028	0	0.021	0

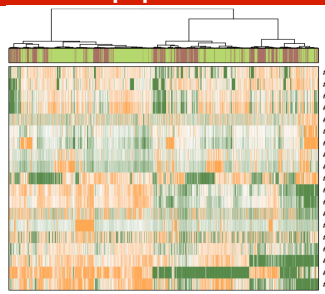
Prior distribution: Learned from expression data
Likelihood function: From probabilistic model
Posterior distribution: Prior * Likelihood
Joint posterior distribution of a gene being expressed at an average level *x* in a subpopulation of cells S is determined as an expected value from Bootstrap sampling

$$p_S(x) = E\left[\prod_{c \in B} p(x \mid r_c, \Omega_c)\right]$$

4. DE Gene Detection



5. Subpopulation Detection



1. Weighted PCA on noisy datasets with missing values make the result more sensitive to the true underlying signal variations.
2. Cell hierarchical clustering is performed on the eigengene for each geneset.

Bioinformatics-as-a-Service (BaaS)

- Available to any Stanford faculty or affiliate
- Hourly rate affordable by small labs
  - FY 2016: Subsidized by Dean Ann Arvin and Dean Harry Greenberg
  - FY 2017: Subsidized by Prof. Mike Snyder
- Contacts for BaaS
  Ramesh Nair, Bioinformatics Service Supervisor (rvnair@stanford.edu)
  Yue Zhang, Bioinformatics Data Scientist (yuezhang@stanford.edu)