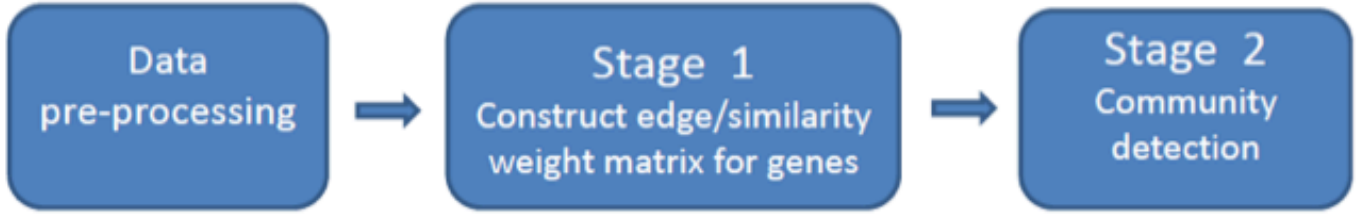
# Inferring Gene-Gene Interactions and Functional Modules Beyond Standard Models

Haiyan Huang
Department of Statistics, UC Berkeley

Feb 7, 2018

# Background

- **Gene regulatory networks**
  - Directed, genes as nodes, a directed edge from gene A to gene B if A encodes a transcription factor that regulates the expression of B.
  - Need time-course data at small time intervals.
  - Study undirected networks as a simplification (when inferring gene networks using expression data).
- **Task of our interest**: To identify gene functional groups, many of which encode biological pathways, using gene expression data.
  - A network of relations needs to be learned using gene expression levels as covariates with functionally related genes having denser connections (stage 1).
  - Detect tightly knit sub-structures (stage 2).



# Background


- **High dimensionality** (p >> n) often results in a high rate of false positives, severely limiting a genomic scale network exploration.
- Measures based on **pairwise marginal relationships** are often inappropriate for detecting higher level interactions.
  - A gene may interact with a group of genes despite having weak marginal correlations with the individual genes.
- **Partial correlation** takes the influence of other genes into account, e.g., Gaussian Graphical Models (GGM).
  - Inclusion of non-pathway genes in the conditional set can lead to spurious dependencies.
  - “Seed genes” (Kim et. al. (2012)), i.e. known pathway genes are useful to define an appropriate conditional set, but they are not always available.
  - Under GGM, the elements in the precision matrix are proportional to partial correlations, i.e., correlation between two genes conditioned on the rest genes.

# Problem Formulation

**Our formulation was motivated by considering GGM in a linear regression setting:**

The partial correlation between gene *i* and gene *j* conditioned on a set of genes *Z* is simply the correlation cor(*ε1,ε2*) of the residuals *ε1* and *ε2* resulting from linearly regressing gene *i* and gene *j* against the genes in *Z*, respectively.

For instance, assume that genes A, B and C are interacting as a group, and that $G_A = \beta_A G_C + \varepsilon 1$ and $G_B = \beta_B G_C + \varepsilon 2$, where $G_A$, $G_B$, $G_C$ denote the expression profile for genes A, B and C, respectively.

If $\text{cor}(\varepsilon 1, \varepsilon 2) \approx 1$, then $G_A + \Gamma_B G_B + \Gamma_C G_C + \Gamma \approx 0$

# Problem Formulation

**Our formulation was motivated by considering GGM in a linear regression setting:**

The partial correlation between gene *i* and gene *j* conditioned on a set of genes *Z* is simply the correlation cor(*ε1,ε2*) of the residuals *ε1* and *ε2* resulting from linearly regressing gene *i* and gene *j* against the genes in *Z*, respectively.

**Therefore, under a GGM, it is reasonable to assume that functionally related genes form strong linear relationships, with one linear equation for one gene module (a set of functionally related genes), suggesting a new formulation for finding gene modules:**

**Finding strong linear models that involve a very small set of genes among the thousands of candidate genes.**

# An attempted solution

- To **find strong linear models** that involve a very small set of genes among the thousands of candidate genes, we introduce a procedure based on **a novel application of sparse canonical correlation analysis (SCCA)** to the gene expression data matrix.

# Sparse Canonical Correlation Analysis (SCCA)

- SCCA finds vectors $\mathbf{a}$ and $\mathbf{b}$ that solve

$$\max_{\mathbf{a},\mathbf{b}} \mathbf{a}^T\mathbf{Y}^T\mathbf{X}\mathbf{b} \quad \text{subject to } \mathbf{a}^T\mathbf{a} \leq 1, \mathbf{b}^T\mathbf{b} \leq 1, p_1(\mathbf{a}) \leq c_1, p_2(\mathbf{b}) \leq c_2,$$

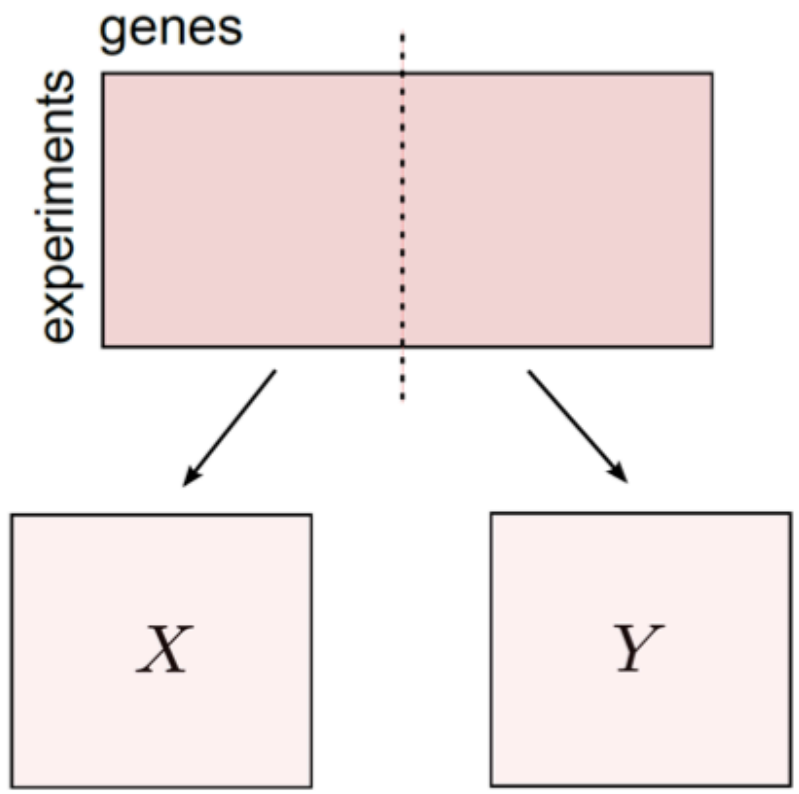
  where $\mathbf{X} \in \mathbb{R}^{\mathbf{n}\times\mathbf{p}}$, $\mathbf{Y} \in \mathbb{R}^{\mathbf{n}\times\mathbf{q}}$ have centered and scaled columns, $p_1$ and $p_2$ are convex functions. (Witten et al. (2009))

- For high-dimensional biological data, sparsity is necessary and we impose this using $L_1$ penalty, $p_1(\cdot) = p_2(\cdot) = \|\cdot\|_1$.

- Modified NIPALS algorithm (Li et al. (2011)) with tuning parameters $\lambda = (\lambda_1, \lambda_2)$.

# SCCA with random partition and subsampling

- Average **a**, **b** over a large number of subsampling and random partitions.
  - Subsample a fixed fraction of the genes
  - Random split the matrix (by genes) into x and y and run SCCA



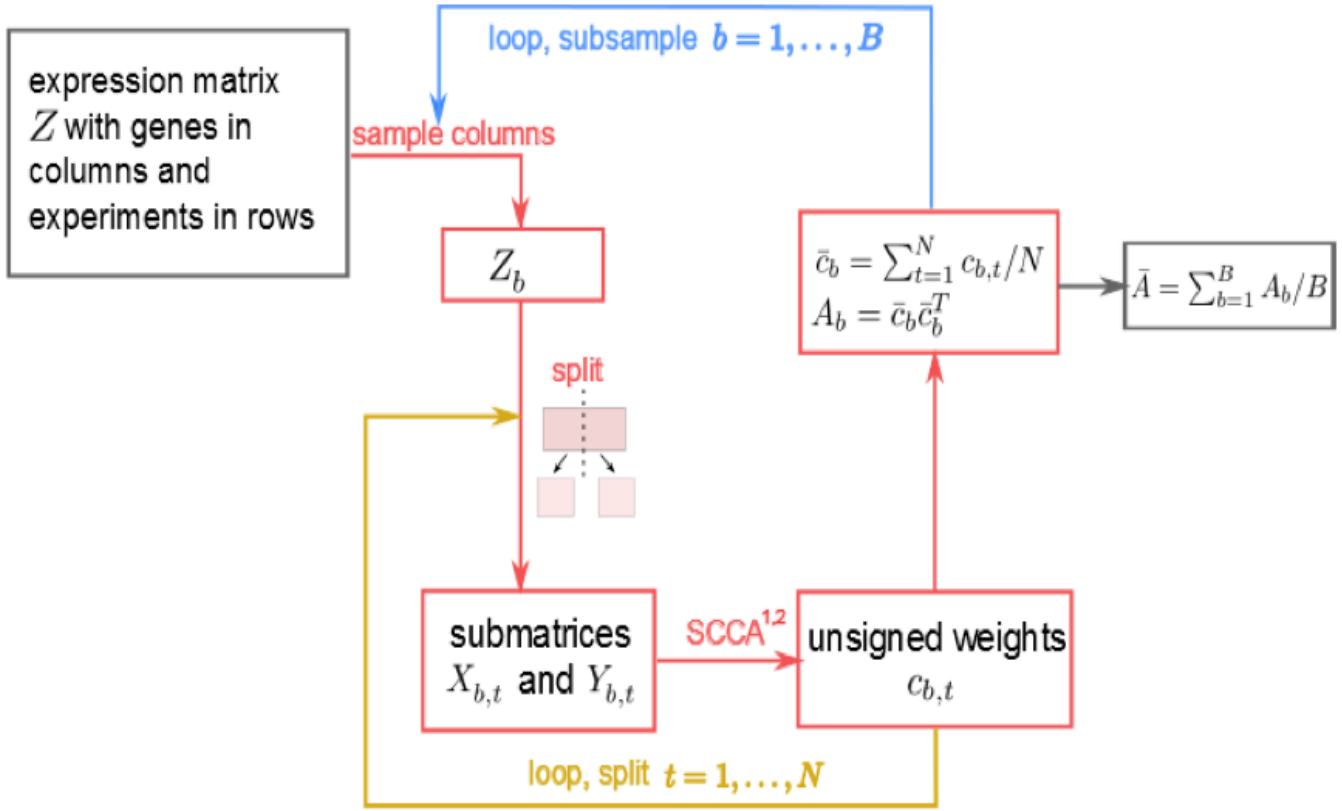

| Genes | \|a\| | \|b\| | c |
|---|---|---|---|
| g1 | 0.4 | / | 0.4 |
| g2 | / | 0 | 0 |
| g3 | / | 0 | 0 |
| g4 | 0 | / | 0 |
| g5 | / | 1.3 | 1.3 |
| g6 | 0 | / | 0 |
| g7 | 0 | / | 0 |
| g8 | / | 1.1 | 1.1 |
| g9 | / | 0 | 0 |
| g10 | 1.2 | / | 1.2 |
| g11 | / | 0.5 | 0.5 |
| g12 | 0.6 | / | 0.6 |

## Proposition

*Assuming only one pathway $K$ and only genes in $K$ are correlated. For every partition $t$, let $\mathbf{c}_t$ be the list of the absolute values $|\mathbf{a}_t|$ and $|\mathbf{b}_t|$ ordered according to the gene list. Let $\bar{\mathbf{c}} = \sum_{t=1}^{N} \mathbf{c}_t / N$, where $N$ is the number of partitions, then $\exists\ B > 0$*

$$\lim_{N\to\infty} \lim_{n\to\infty} (\min_{i\in K} \bar{c}_i - \max_{j\notin K} \bar{c}_j) = B.$$

# Summary of Procedure




1. SCCA implemented by a modified NIPALS algorithm involving iterative penalized regression with L1 penalty.

2. If some genes are known to operate in the same pathway, the penalty parameter corresponding to those genes could be lowered.

# Subsampling helps the identification of weaker functional group

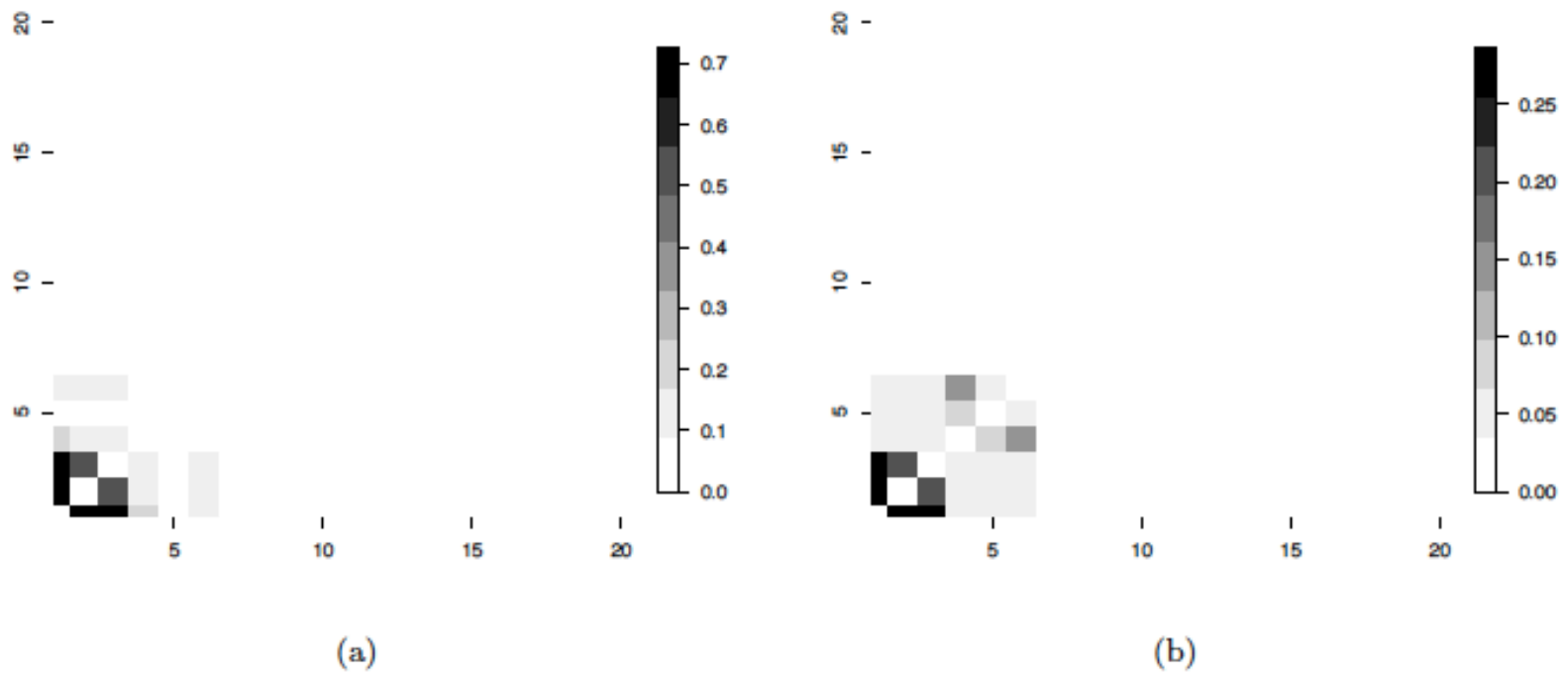



Figure 2: Asymptotic values of $\bar{A}$ with (a) no subsampling and (b) 50% subsampling.

- We simulated the expressions of 20 genes.
- 6 of them form two disjoint functional gene groups.
- One group has 3 genes that are perfectly linearly related (see (a)).
- The other group has 3 genes that have relatively weaker linear relationships (see (b).
- The left 14 genes are independent from each other.

# Subsampling helps the identification of overlapping functional groups

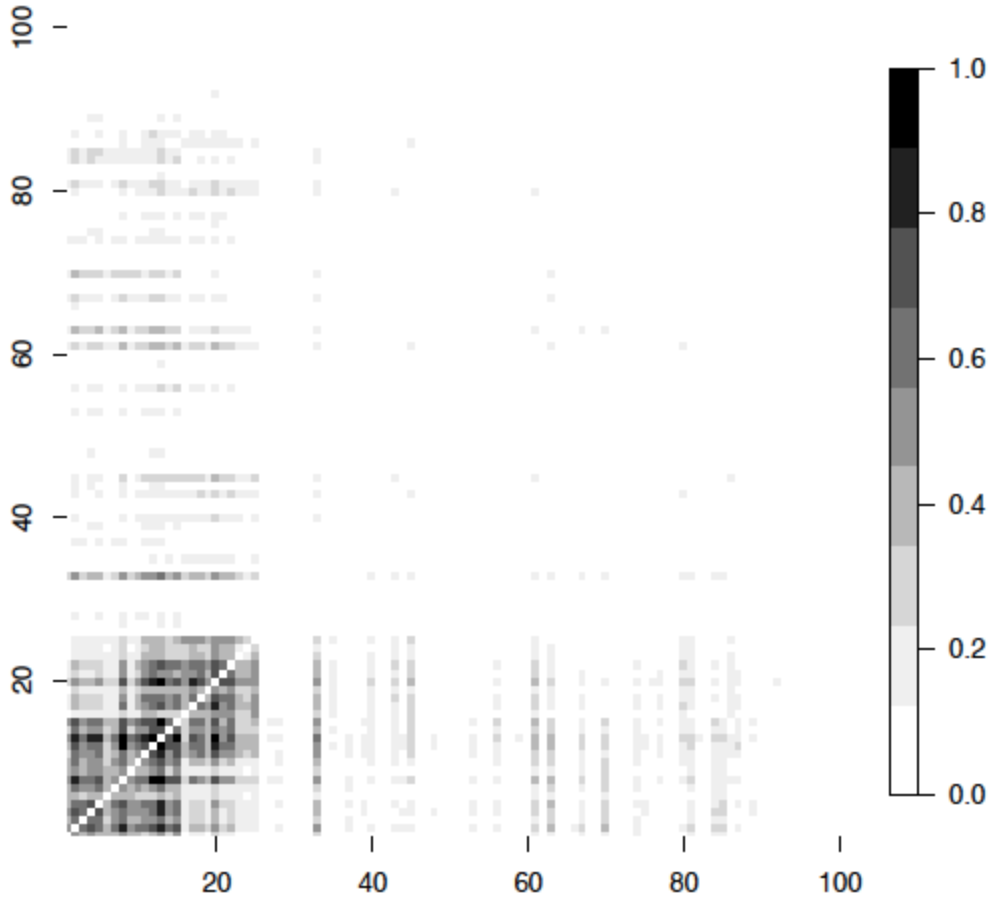
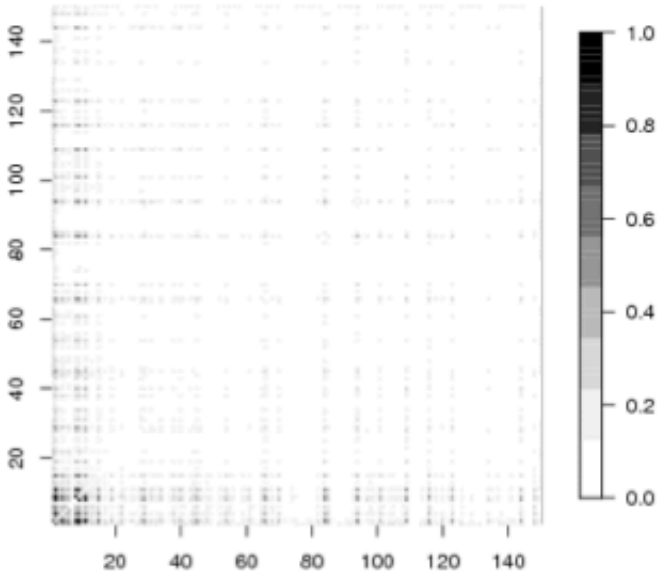



Figure 1: Heatmap of $\bar{A}$ with 50% subsampling.

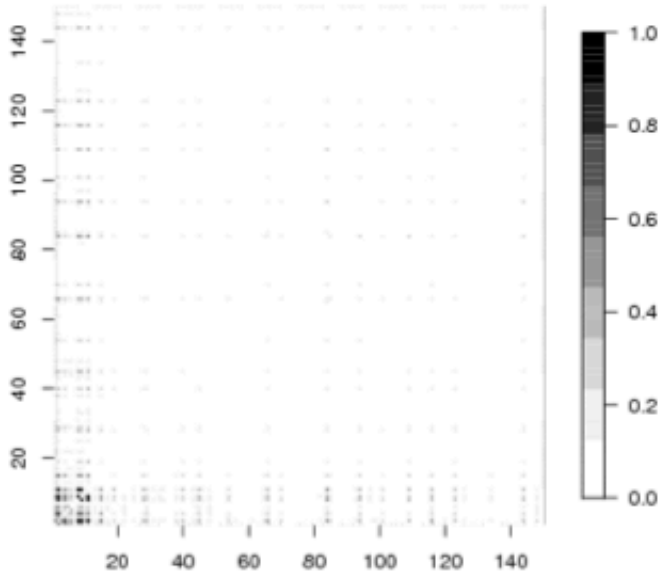
We simulated the expressions of 150 genes.

There are two overlapping functional gene groups, with each functional group having 15 genes and 5 genes in common.

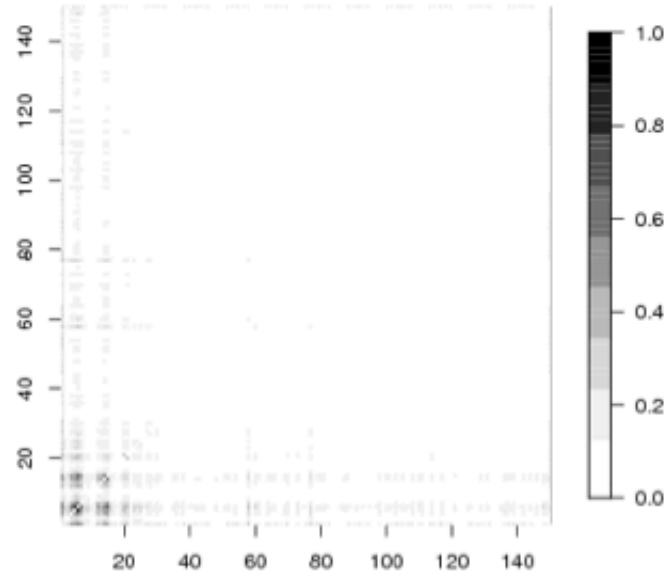
# Choose the amount of regularization




- $p = 150,\ n = 30$
- functional group: 1-15
- subsample 70%, $\lambda_a = 3, \lambda_b = 3$




- $p = 150,\ n = 30$
- functional group: 1-15
- subsample 70%, $\lambda_a = 9, \lambda_b = 9$

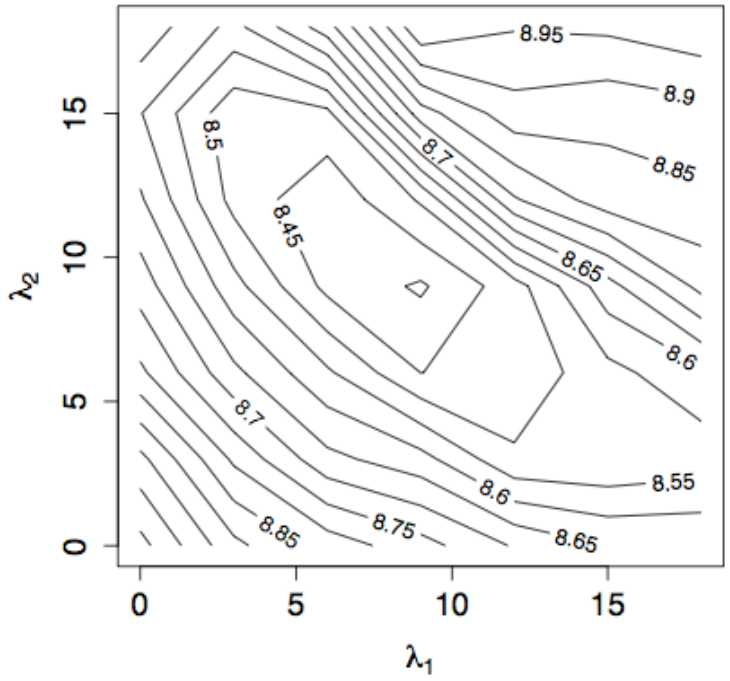


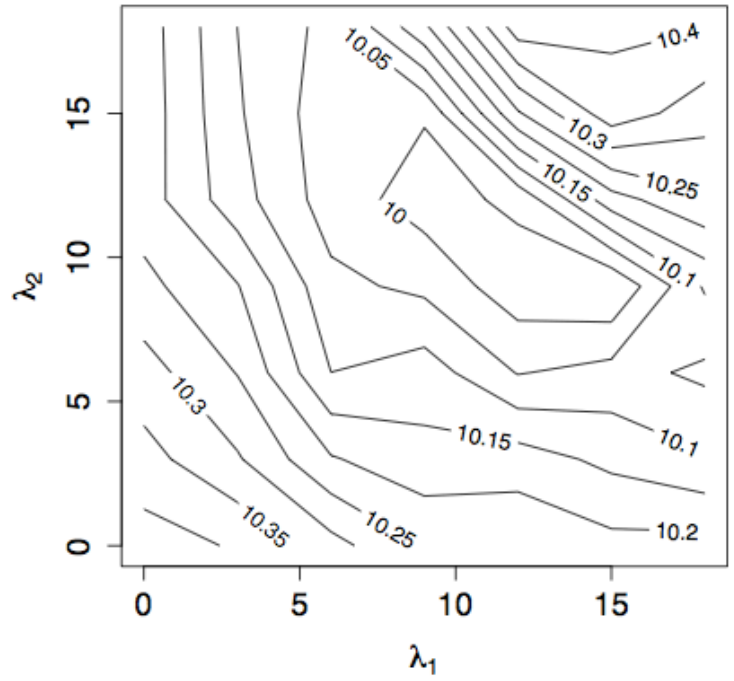

- $p = 150,\ n = 30$
- functional group: 1-15
- subsample 70%, $\lambda_a = 18, \lambda_b = 18$

The entropy of an edge weight matrix $\bar{A}$ is defined as

$$H(A) = -\sum_{i<j, A_{ij}>0} (A_{ij}/S_A) \log(A_{ij}/S_A),$$

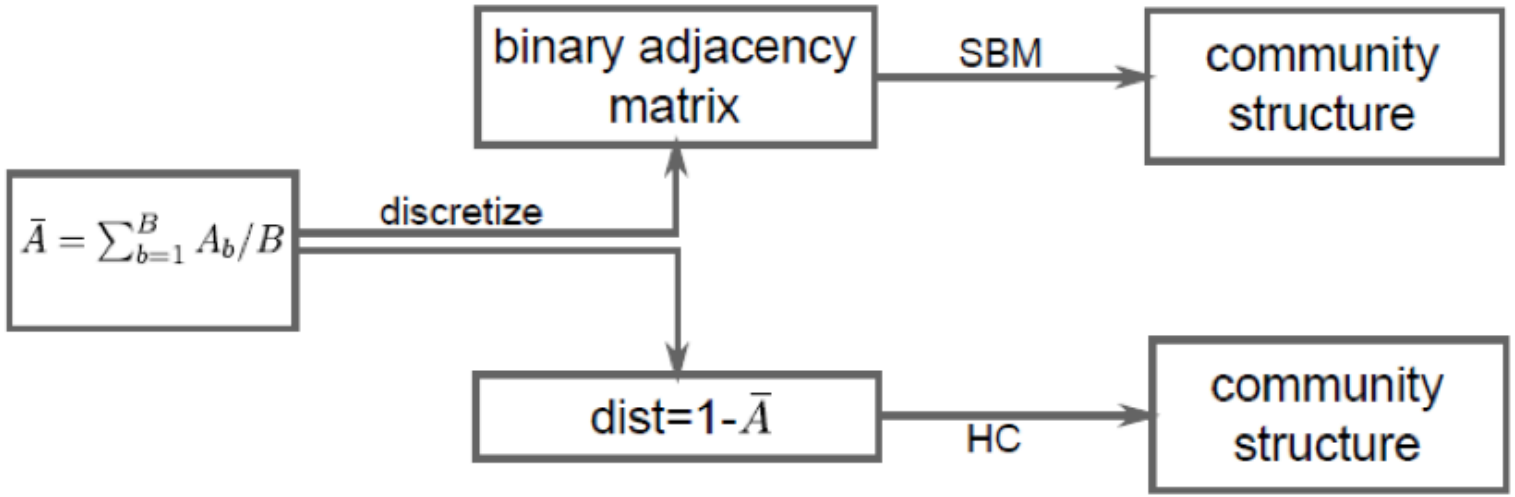
where $S_A = \sum_{i<j} A_{ij}$.
$(\lambda_1, \lambda_2)$ leading to small entropy values are desirable.



# Identify community structures in $\bar{A}$

- Gene functional groups as communities in a network.
- Community detection methods (many are available; we implemented the below two)
  - Greedy algorithms — hierarchical clustering (HC)
  - Probabilistic models — stochastic block models (SBM) fitted using the pseudo-likelihood algorithm Chen et al. (2012)



# Performance comparison for simulated data

- Classification rates: **precision** $= \frac{TP}{TP+FP}$, **recall** $= \frac{TP}{TP+FN}$.
- Cross-mix the following to allow comparisons
  - Stage 1: building edge weight matrices $\bar{A}$
    - *scca*
    - *pearson*
    - *module*: Transformed Pearson's correlation matrix used in Langfelder et al. (2007).
  - Stage 2: detecting block structures from $\bar{A}$
    - *sbm*
    - *hc*: Cutting the dendrogram when clusters of size less than 25 start to appear.
    - *dynamic*, *hybrid*: HC with dendrogram cutting methods in `dynamicTreeCut` (Langfelder et al. (2008)).
- Six methods for comparison: *scca.hc*, *pearson.hc*, *scca.sbm*, *pearson.sbm*, *module.dynamic*,*module.hybrid*.

Table: Classification performance of different methods using datasets with $p$ = 500, two functional groups, subsampling level 70%, and various levels (0%, 33% and 67%) of experiment dependency.

| | Pathway 1 | | | | | |
|---|---|---|---|---|---|---|
| | 0% | | 33% | | 67% | |
| | Precision | Recall | Precision | Recall | Precision | Recall |
| *scca.hc* | 0.861 | 0.533 | 0.831 | 0.441 | 0.811 | 0.433 |
| *module.dynamic* | 0.718 | 0.3 | 0.742 | 0.333 | 0.764 | 0.38 |
| *module.hybrid* | 0.439 | 0.407 | 0.544 | 0.447 | 0.453 | 0.385 |
| *pearson.hc* | 0.238 | 0.233 | 0.497 | 0.427 | 0.471 | 0.393 |
| | Pathway 2 | | | | | |
| | 0% | | 33% | | 67% | |
| | Precision | Recall | Precision | Recall | Precision | Recall |
| *scca.hc* | 0.808 | 0.487 | 0.890 | 0.489 | 0.833 | 0.420 |
| *module.dynamic* | 0.758 | 0.4 | 0.808 | 0.347 | 0.8 | 0.4 |
| *module.hybrid* | 0.565 | 0.473 | 0.529 | 0.387 | 0.455 | 0.46 |
| *pearson.hc* | 0.438 | 0.387 | 0.323 | 0.307 | 0.460 | 0.273 |

# Arabidopsis data

- Shoot tissue dataset from *A.thaliana* subject to oxidation stress, composed of 22810 genes, 13 experiments and two replicates for each experiment.
- Select genes that have
  - a reasonable large variance across the experiment. Also remove genes with a suspiciously high experiment variance.
  - small replicate discrepancy
  - reasonably large expression levels

  $\Rightarrow p = 2718.$
- *scca.hc* vs. *pearson.hc*, *module.dynamic*, *module.hybrid*

Of the 13 groups found,

*GO enrichment of groups*

| Group ID | Enriched GO term | Number of genes with enriched terms | P-values |
|---|---|---|---|
| 1 | Chloroplast organellar gene | 10 out of 15 | $1.10 \times 10^{-4}$ |
| 2 | Phenylpropanoid-flavonoid biosynthesis | 3 out of 4 | $6.65 \times 10^{-7}$ |
| 3 | Glucosinolate biosynthsis | 7 out of 7 | $1.95 \times 10^{-14}$ |
| 4 | Chloroplast organellar gene | 3 out of 3 | $7.83 \times 10^{-3}$ |
| 5 | Ribosome | 10 out of 15 | $7.20 \times 10^{-13}$ |
| 8 | Ribosome | 5 out of 6 | $8.31 \times 10^{-8}$ |
| 10 | Photosystem I or II | 8 out of 10 | $2.87 \times 10^{-14}$ |
| 12 | Endomembrane system | 3 out of 4 | $2.35 \times 10^{-3}$ |

The Arabidopsis Information Resource
(http://www.arabidopsis.org/tools/bulk/index.jsp)

# Results using *module.dynamic* and *module.hybrid*

GO enrichment of groups — first cut

| Group ID | Enriched GO term | Number of genes with enriched terms | P-values |
|---|---|---|---|
| 9 | Cell wall | 16 out of 81 | $4.46 \times 10^{-6}$ |
| 10 | Defense response | 29 out of 78 | $1.58 \times 10^{-2}$ |
| 11 | Phenylpropanoid-flavonoid biosynthesis | 11 out of 76 | $5.42 \times 10^{-12}$ |

**13 groups in total with sizes from 60 to 293**

GO enrichment of groups — second cut

| Group ID | Enriched GO term | Number of genes with enriched terms |
|---|---|---|
| 62 | NA | 0 out of 6 |
| 63 | Chloroplast | 4 out of 6 |
| 64 | Located in plasma membrane | 2 out of 5 |
| 65 | Located in plasma membrane | 3 out of 5 |
| 66 | Pyridoxine biosynthetic process | 2 out of 5 |

**66 groups in total with sizes from 5 to 81**

# Conclusion on the SCCA method

- High precision
- More conceptually appealing.
  - Under a regression setting, it captures group interactions by providing an aggregated measure of gene partial correlations when the correct conditional set is unknown
- Flexible enough to incorporate prior knowledge when available

***Gene-Fishing:***

A semi-supervised, non-parametric clustering procedure with bagging

# Background

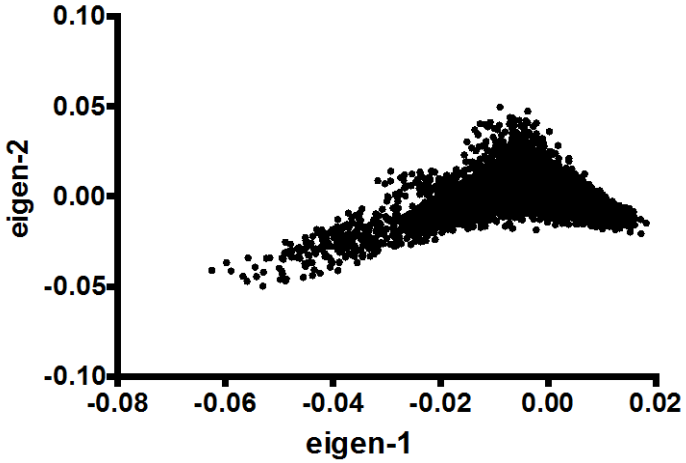
**Biomedical questions of interest**

1. Identifying novel cholesterol-metabolism regulators that may contribute (or be related) to cardiovascular disease risk.
2. Characterizing the diversity of cholesterol metabolism across different tissues.
3. Understanding mechanisms underlying adverse effects of statin treatment.

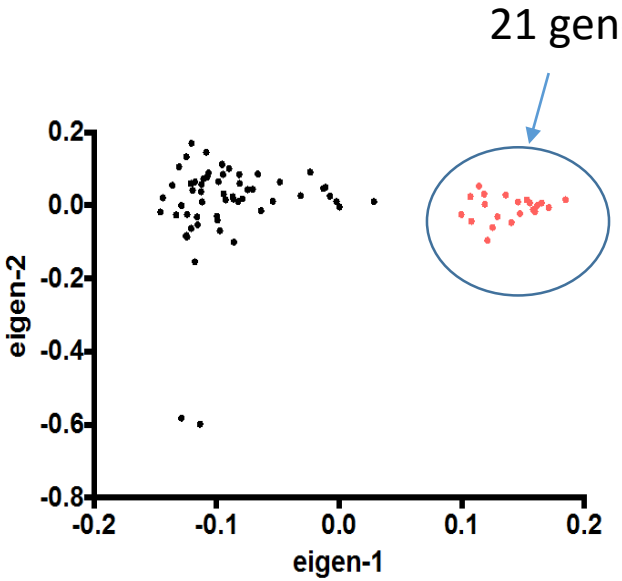
**Datasets**:

1. CAP LCLs Dataset (n=426): RNA-seq and expression array of subject-derived lymphoblastoid cell lines (LCLs) were established from CAP subjects; demographic information and plasma LDLC are also available.
2. GEUVADIS Dataset (n=465; publicly available): the transcriptome of 465 LCL samples (from the 1000 Genome Project) were deeply sequenced on Illumina HiSeq2000.
3. GTex Dataset (publicly available): 7051 samples for 43 different tissues were collected for RNA-Seq analysis.

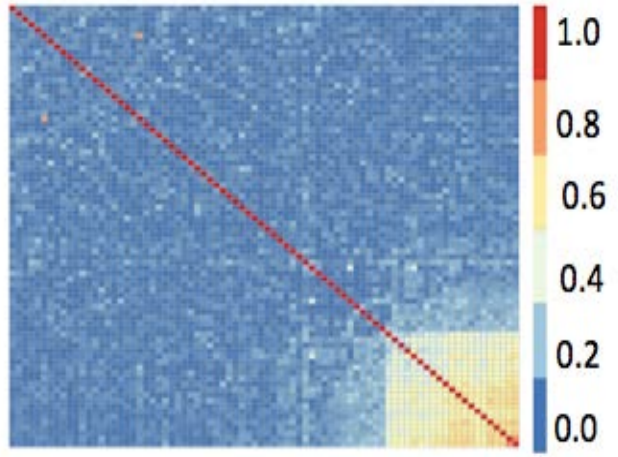
# Our Initial Analysis:




Spectral analysis of all 14028 genes (Cap LCL expression dataset)

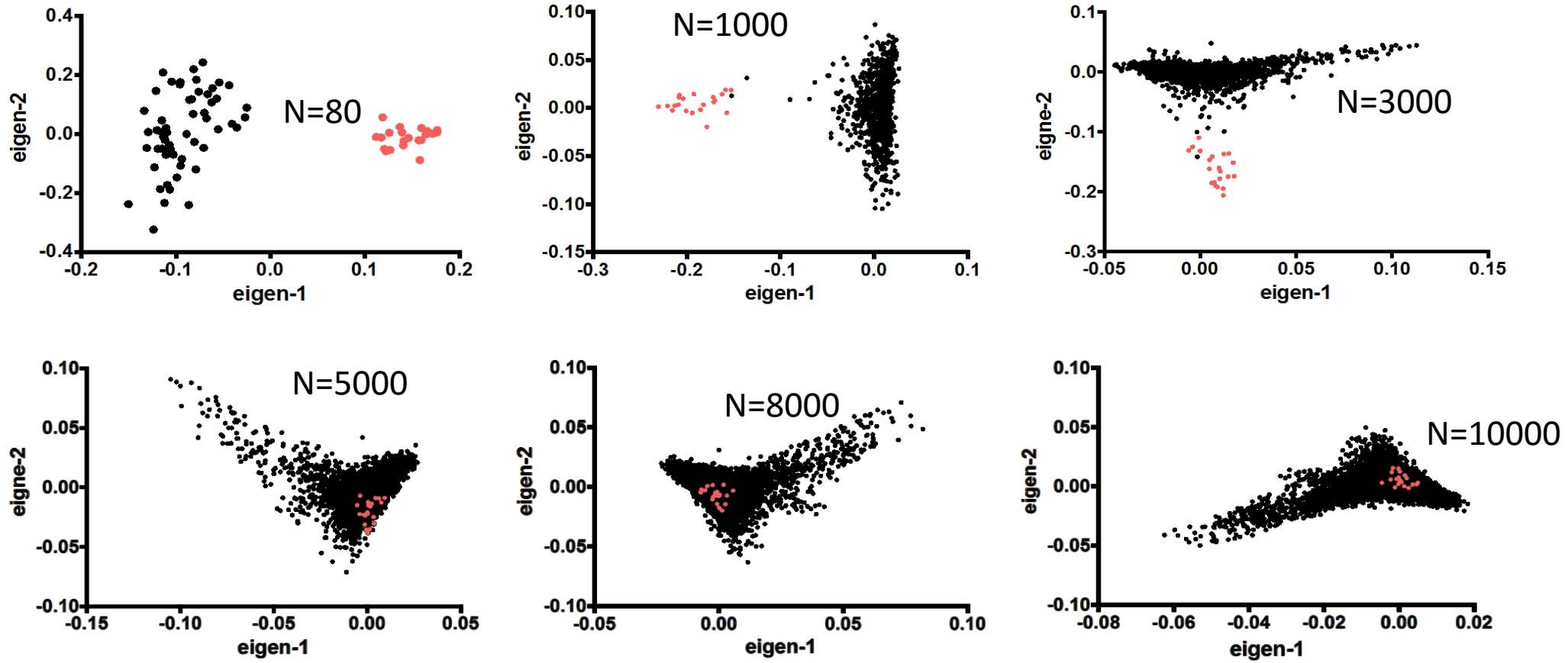



Spectral analysis of 80 (GO-annotated) cholesterol metabolism associated genes




Visualization of gene-gene correlation matrix

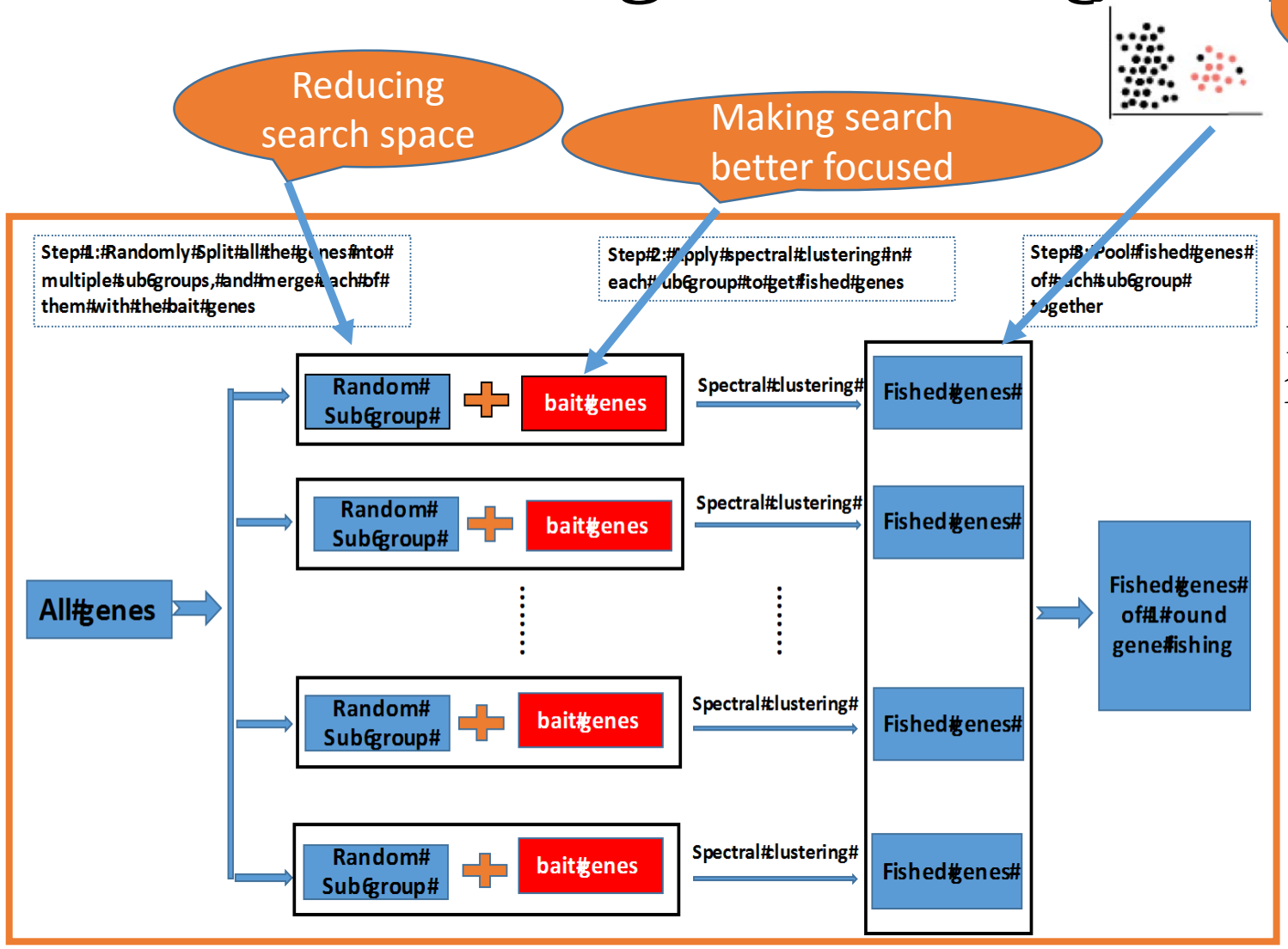
# Interesting patterns become hidden as noise increases




The 21 cholesterol synthetic pathway genes

Random genes

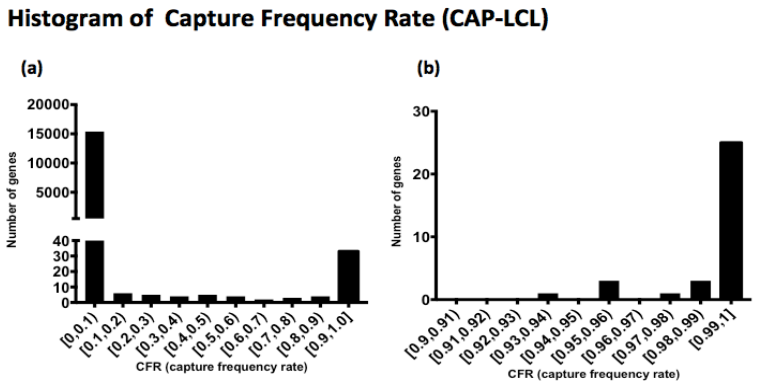
# Workflow of gene fishing



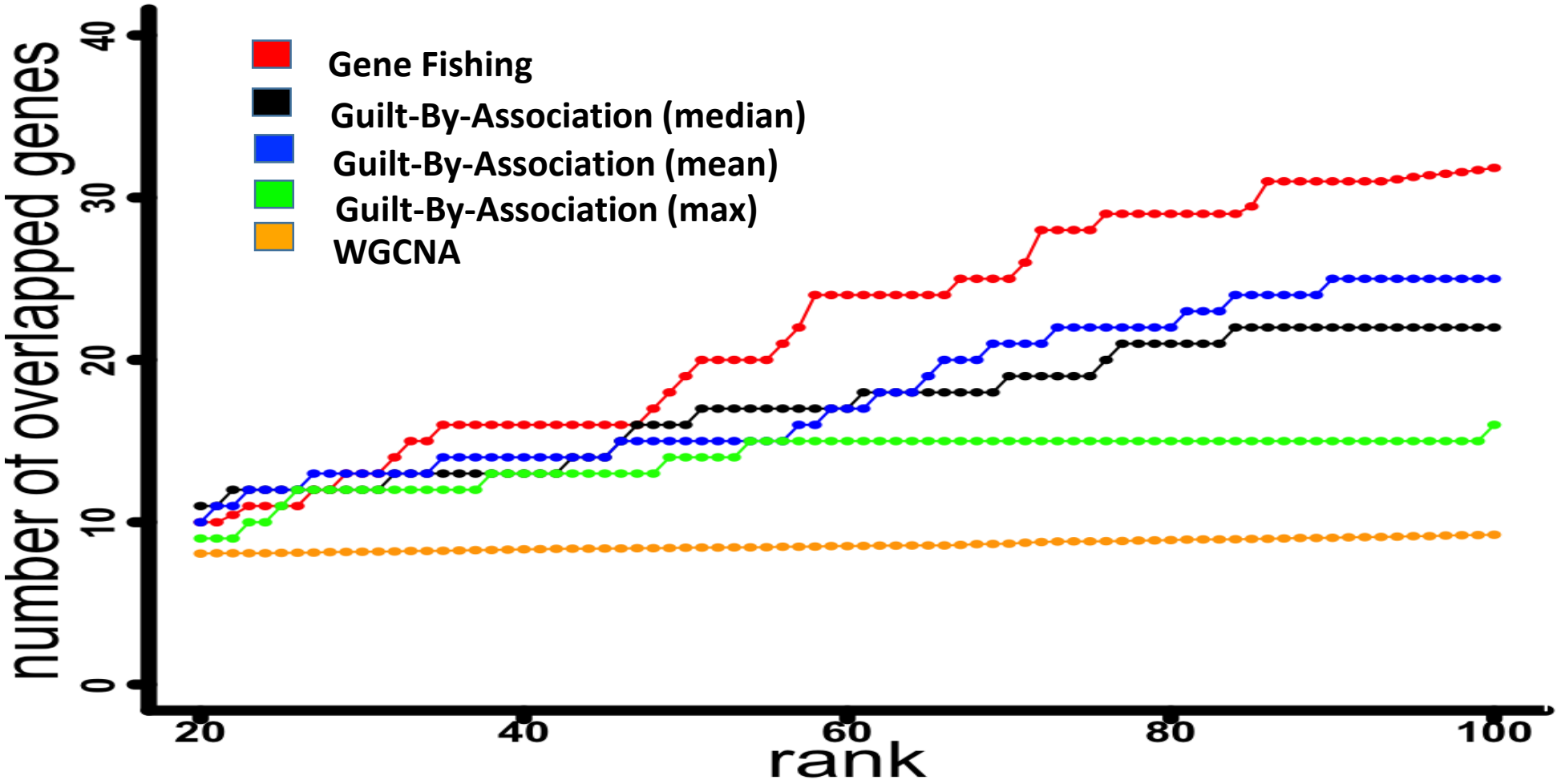

Repeat for 1000 times

A ranked gene list

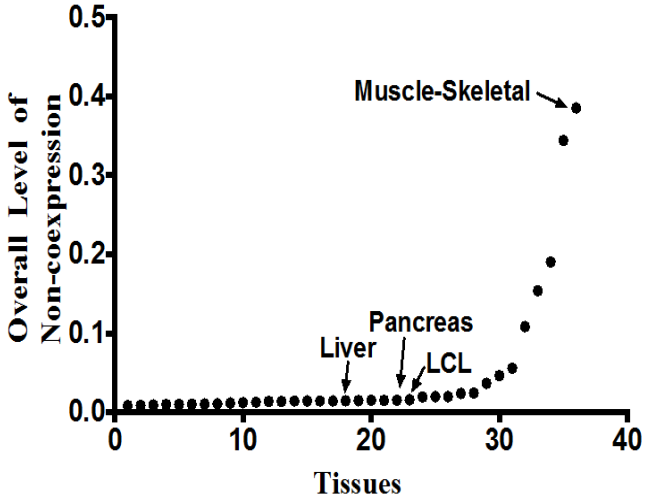
| Gene name | Capture Frequency Rate |
|---|---|
| $G_1$ | $CFR_1$ |
| $G_2$ | $CFR_2$ |
| … | … |
| $G_n$ | $CFR_n$ |

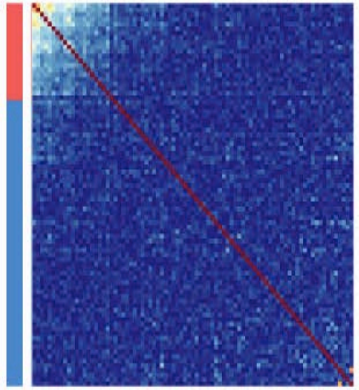


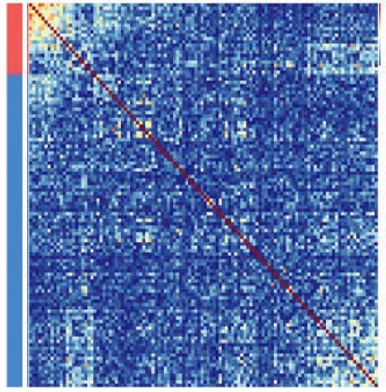
# Comparison of Reproducibility between the CAP LCL dataset and the GEUVADIS dataset

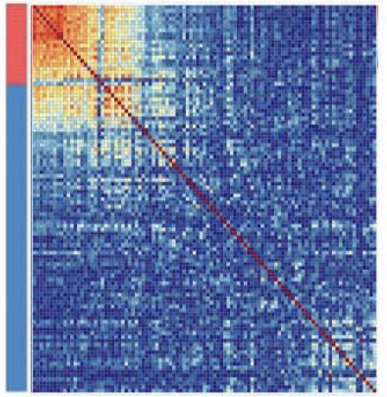


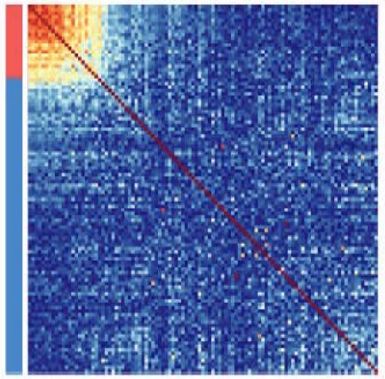
# Application to GTex dataset

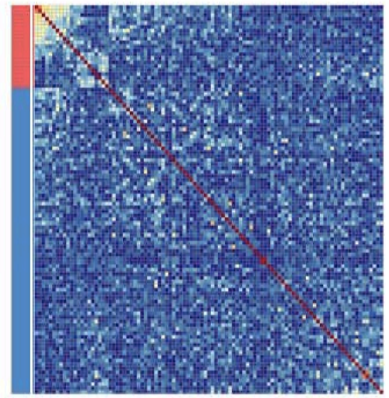




GEUVADIS-LCL

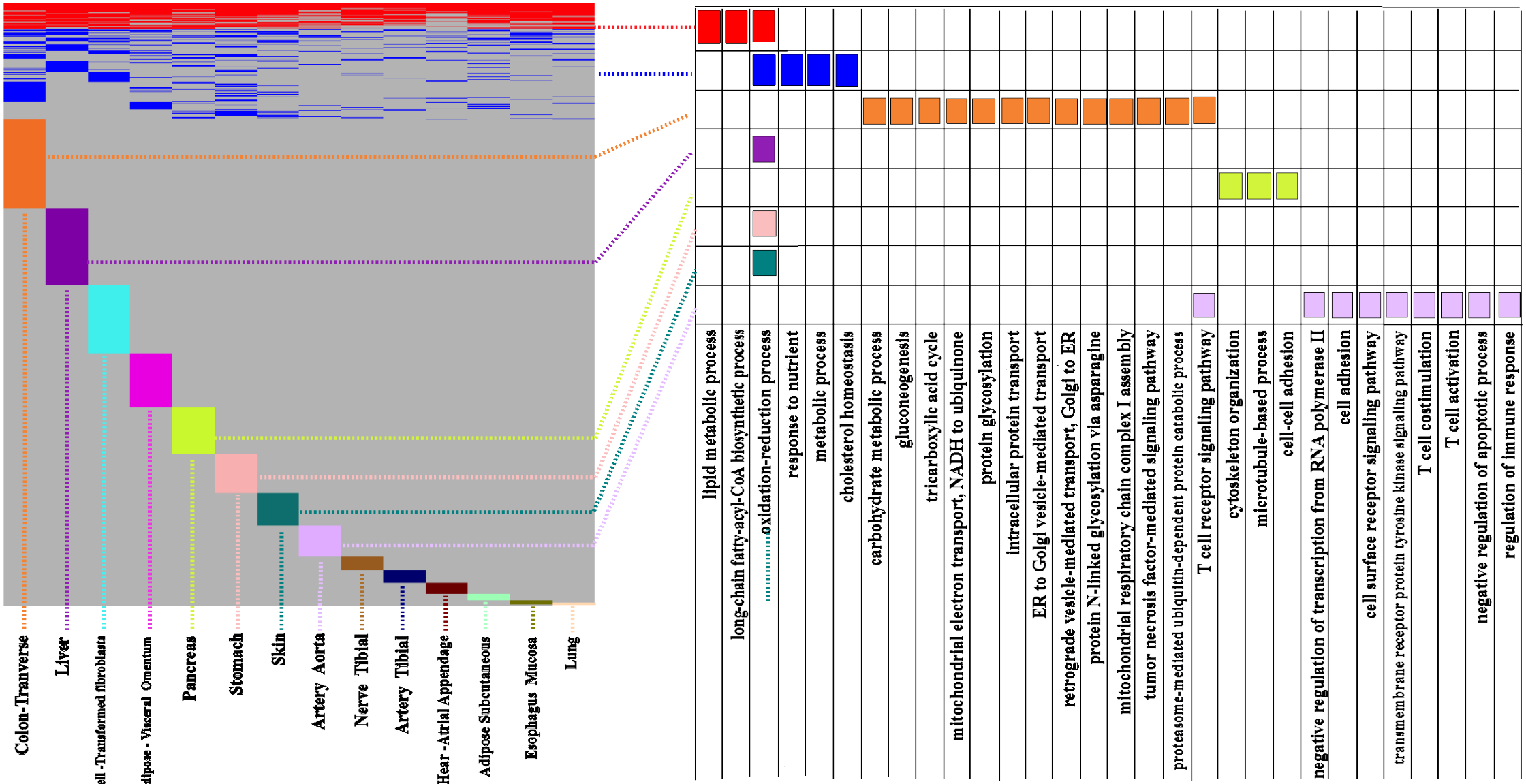

Testis


Adrenal Gland


Colon-Transverse


Colon-Sigmoid

# Application to GTex dataset



# Acknowledgements


- SCCA project
  Wang YXR, Jiang K, Feldman LJ, Bickel PJ, Huang H.
  Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis.
  *Annals of Applied Statistics*. 2015. 9(1): 300-323.
- Gene-Fishing project
  - UC Berkeley: Dr. Ke Liu (Postdoc); Yun Zhou (Student in Biostatistics); Tal Ashuach (Student in Computational Biology); Peter Bickel (Professor, Statistics)
  - Children's Hospital at Oakland Research Institute: Dr. Elizabeth Theusch (Postdoc) Dr. Marisa Wong Medina (Scientist)