Data Integration and Iterative Testing

Andrew Nobel

Department of Statistics and Operations Research

Department of Biostatistics

UNC Chapel Hill

CATS Workshop, June 8, 2016

Outline

Overview of Data Integration

Iterative Hypothesis Testing

- Community Detection
- Mining of Differential Correlation

Data Integration

Classical: Horizontal Integration

- Common measurement platform
- Multiple experiments or sample groups

Modern: Vertical Integration

- Common samples
- Multiple measurement platforms/technologies
- Data with and without response variables

Example: The Cancer Genome Atlas (TCGA)

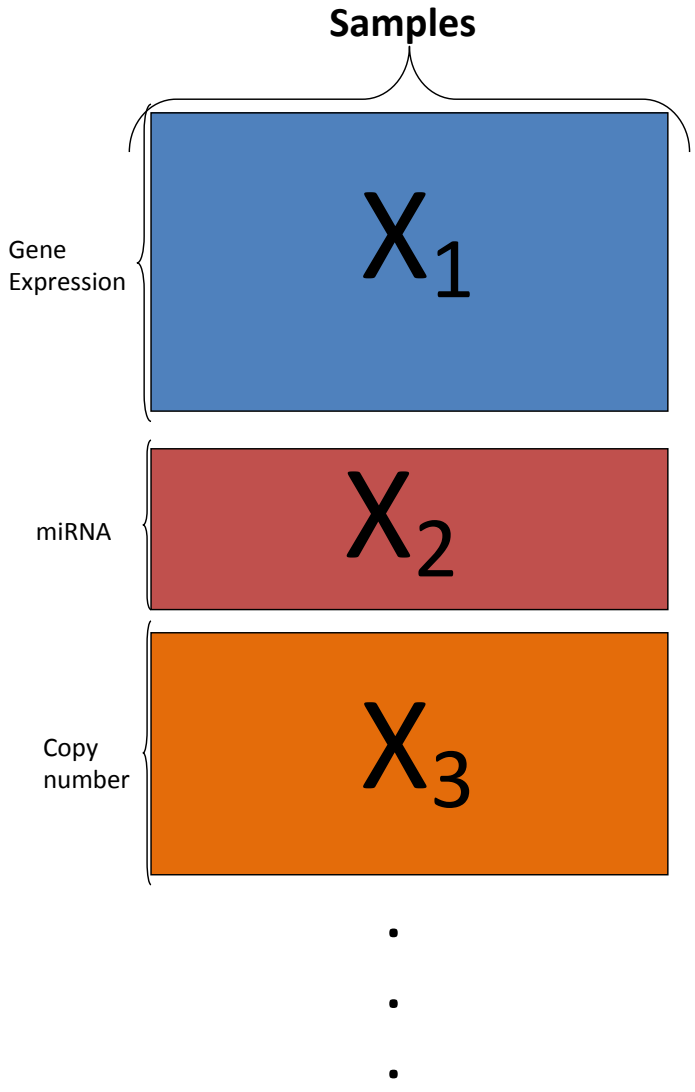
Representative Dataset: 350 breast cancer tumors

- Gene expression data (18K genes)
- miRNA data (650 miRNAs)
- Copy number data (200K probes)
- Methylation data (22K CG regions)
- Mutation data (12K genes)

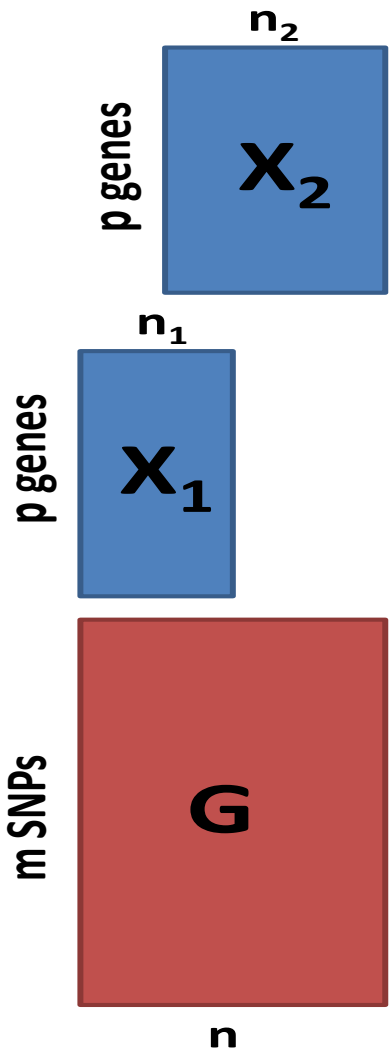
Responses

- Survival
- Response to therapy

TCGA: Data Matrix View



Multi-Tissue eQTL Analysis (GTEx)



Potential

- Borrowing strength across tissues
- Genetic basis of tissue variation

Complications

- Donors vary by tissue
- Configurations of association
  - Common across tissues
  - Tissue specific
- Large number of tissues

Potential of Vertical Integration

Statistics: Enhanced power, improved prediction

- Borrowing information across platforms
- Analysis of shared and individual variation across platforms

Genomics: New/enhanced insights into underlying biology

- Relationships between measured genomic features
- Role of genomic features in predicting outcome of interest

The Fine Print (in practice)

Genomics (upstream)

- Preprocessing: imputation, normalization, scaling
- Identification and removal of appropriate covariates

Statistics (midstream)

- Verifying distribution and sparsity assumptions
- Selecting free parameters
- Availability of robust/efficient software

Networks as Models or Summaries

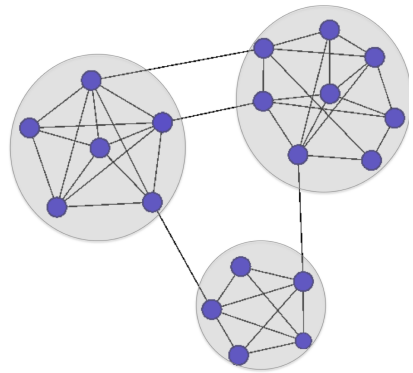
Positives

- Natural representation for pairwise (two-way) relationships
- Amenable to ready interpretation and visualization

Limitations

- Do not capture higher order interactions
- May not capture heterogeneity among samples

Motivating Example: Community Detection



Given network $G = (V, E)$ identify sets $C_1, \ldots, C_k \subseteq V$ such that

- Edge density within sets C_i is large
- Edge density between sets C_i is small

Iterative Testing for Community Detection

Given: $B_t \subseteq V$, level $\alpha \in (0, 1)$

For each $u \in [n]$ compute p-value

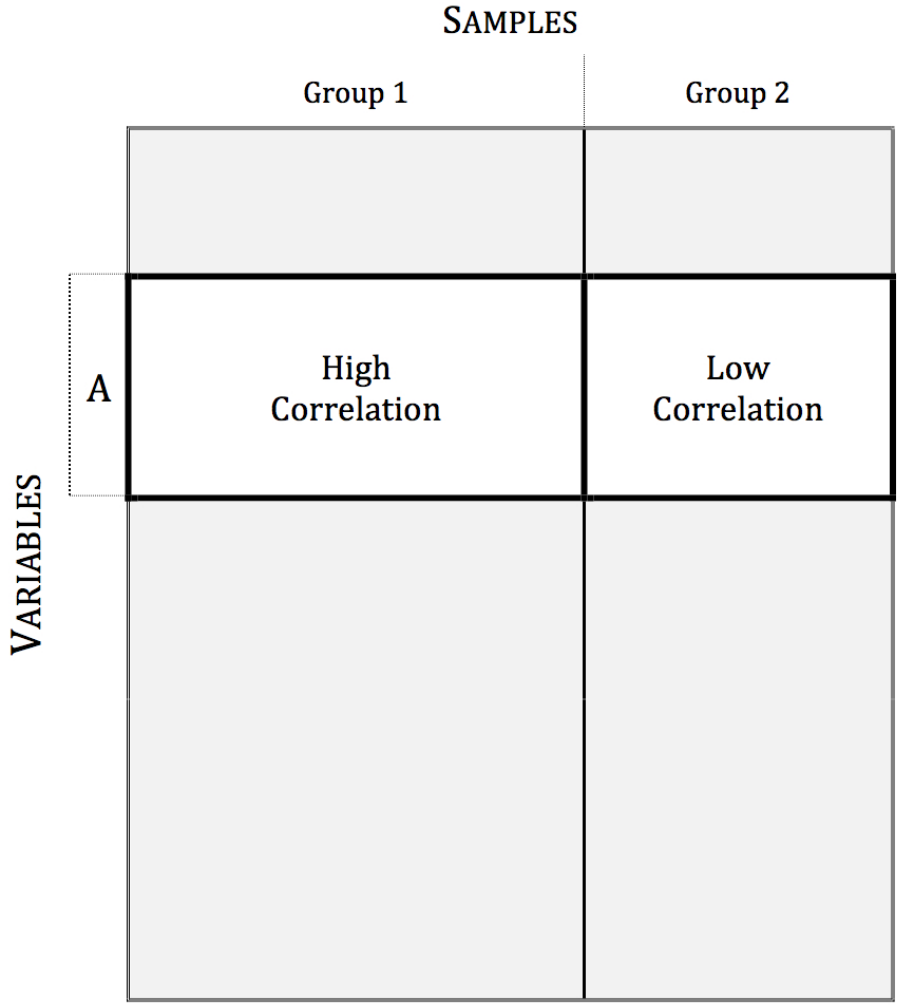
$$p(u : B_t) \approx \text{significance of connection between } u \text{ and } B_t$$

using configuration null model. Then

- Order nodes V s.t. $p(u_1 : B_t) \leq \cdots \leq p(u_n : B_t)$
- $k_0 :=$ largest k such that $p(u_k : B_t) \leq (k/n)\,\alpha$
- $B_{t+1} := \{u_1, \ldots, u_k\}$

Repeat until $B_{t+1} = B_t$

Differential Correlation Mining



- Two sample/treatment groups
- Common set of variables
- Identify sets of variables that are differentially correlated across groups
- Examples: genomic data, word usage frequencies, brain activity, etc.
- Distinct from differential expression and clustering

Example: Breast Cancer Subtypes in TCGA

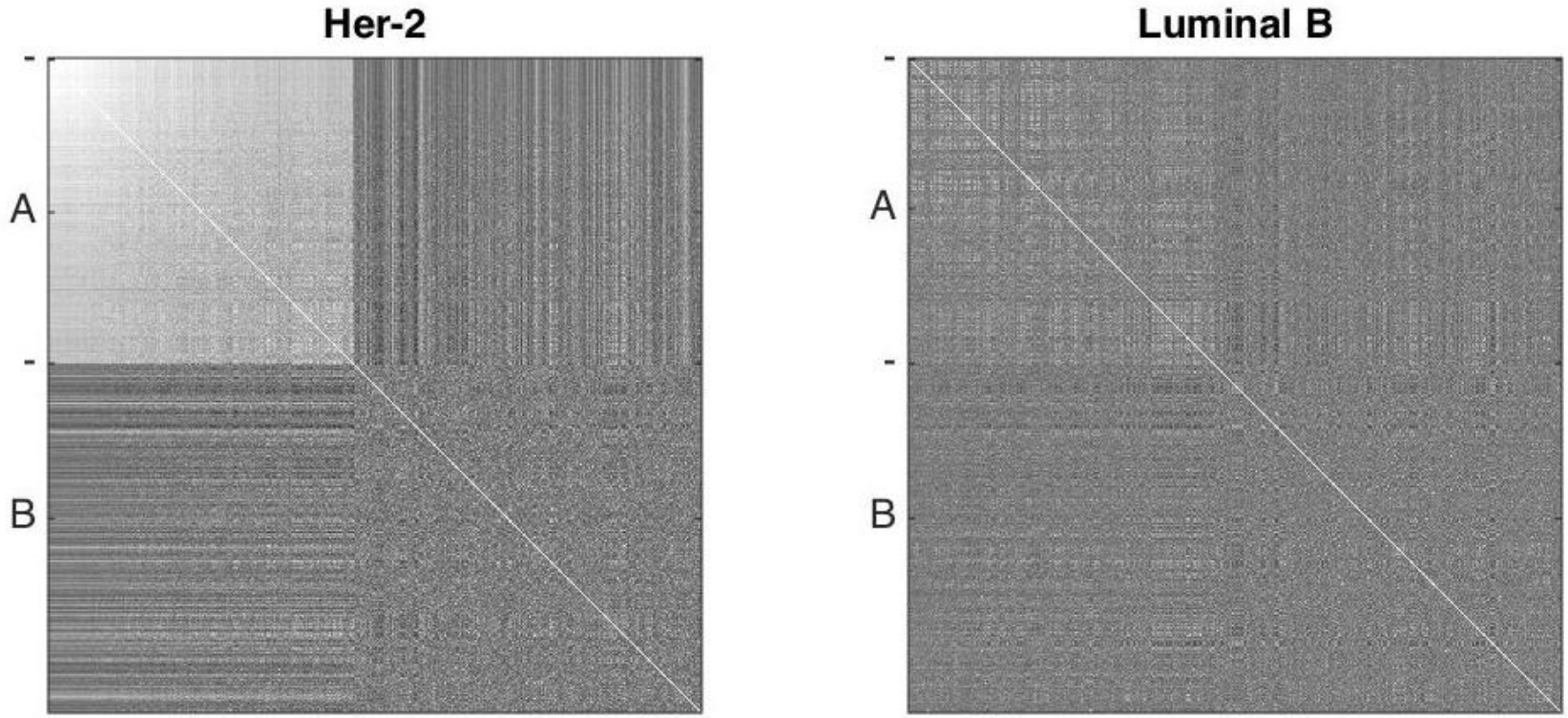


Figure: Sample correlation from Her-2 and Luminal B cancer subtype samples. Differentially correlated set of 165 variables (A) and 200 randomly chosen variables (B).

Example: Brain Connectome

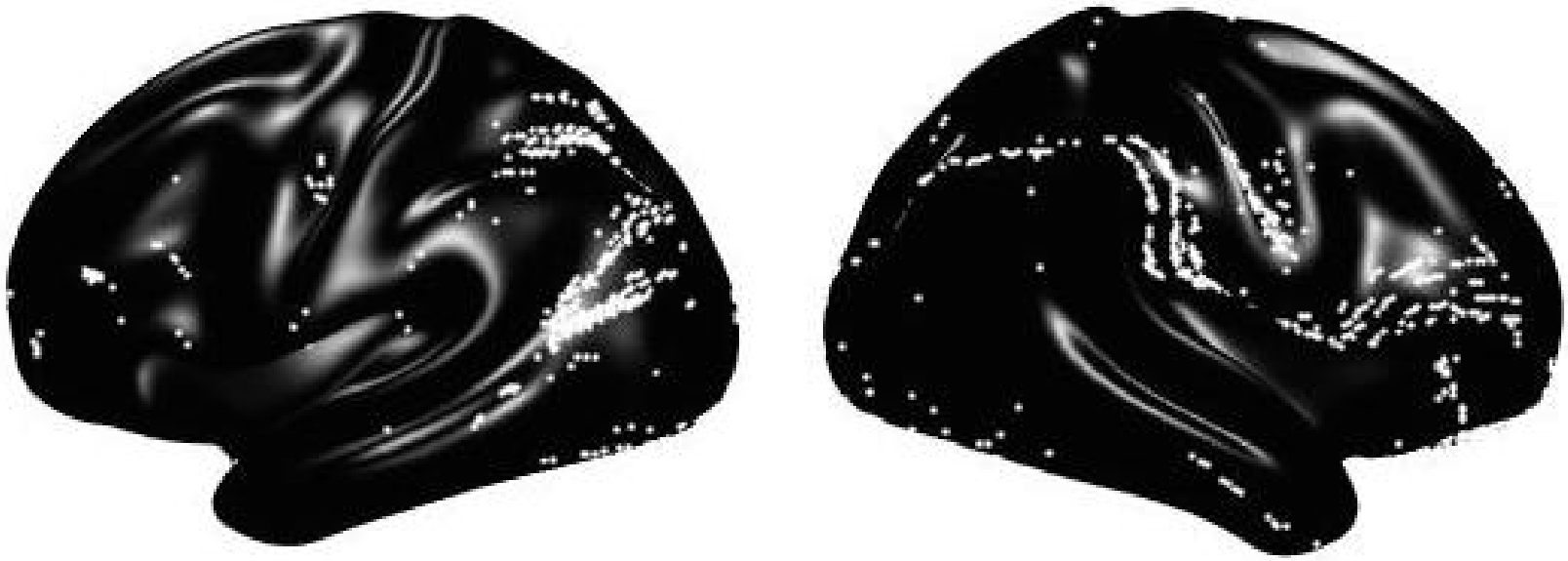


Figure: Brain locations of differentially correlated set for languages tasks versus motor tasks (Data: Human Connectome Project).

Overview

1. Data Integration
   - Horizontal
   - Vertical

2. Iterative Testing
   - Community Detection
   - Differential Correlation Mining

Thanks

Collaborators

- Kelly Bodwin, Eric Lock, James Wilson, Simi Wang
- Shankar Bhamidi, Steve Marron, Peter Mucha, and Kai Zhang

Support

- NIH MH090936 and MH101819
- NSF DMS-0907177 and DMS-1310002