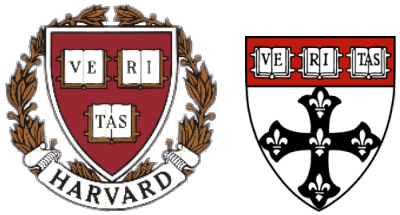
High-Sensitivity Pattern Discovery in High-Dimensional Heterogeneous Datasets

Gholamali (Ali) Rahnavard
Postdoctoral Associate/ The Huttenhower Lab

SAMSI Bioinformatics Workshop: Discovering Patterns in Human Microbiome Data (HMD)

17 March 2015



Harvard T.H. Chan School of Public Health
Department of Biostatistics

Outline

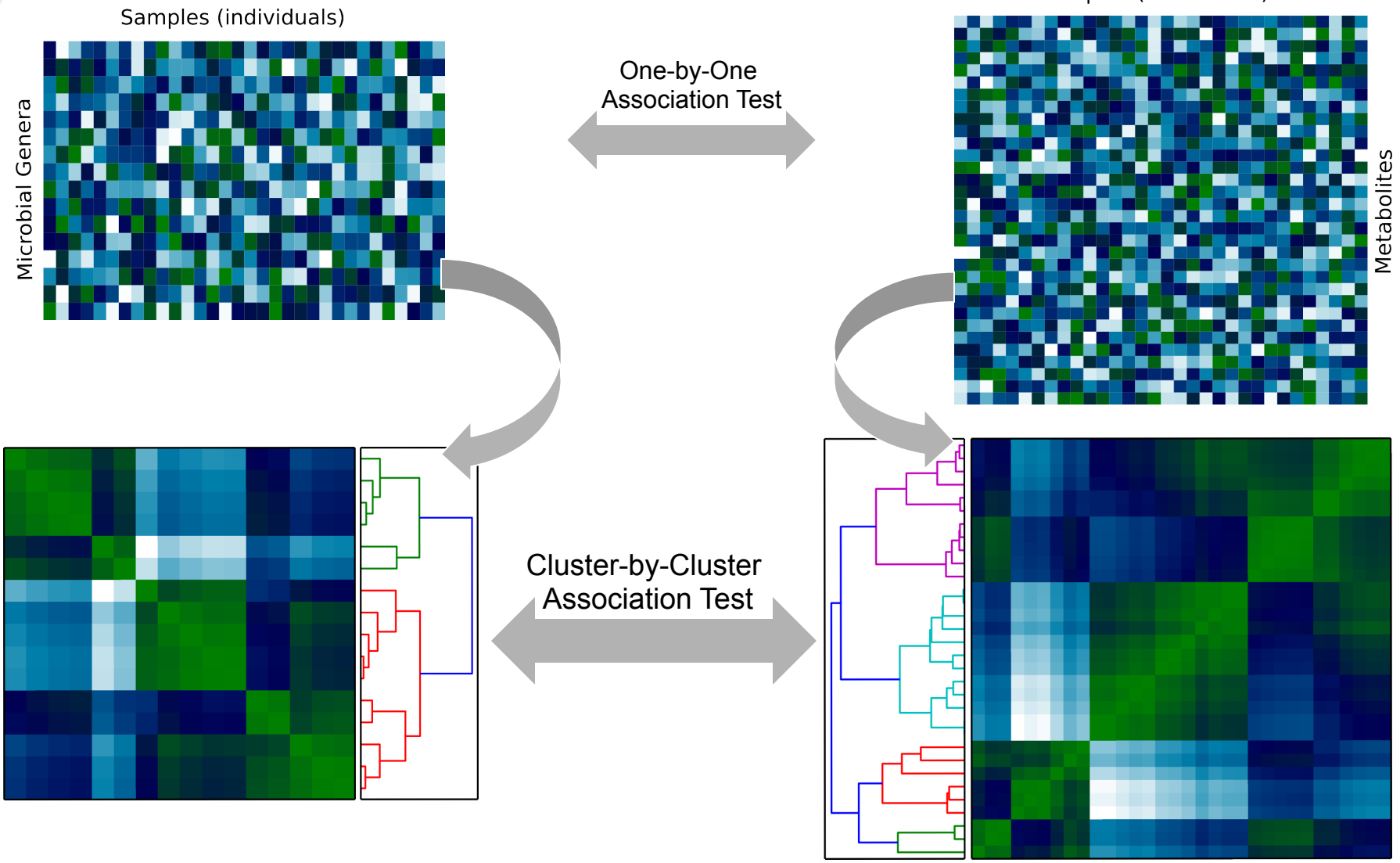
- Background
- Proposed Method
- Validation
- Evaluation
- Application
- Software

Fundamental Problem

How to discover quickly and confidently latent associations within high-dimensional, heterogeneous datasets in the presence of noise and collinearity?

Image is from: http://freshgypsy.com/post/21824134701/rainbowpointelismart

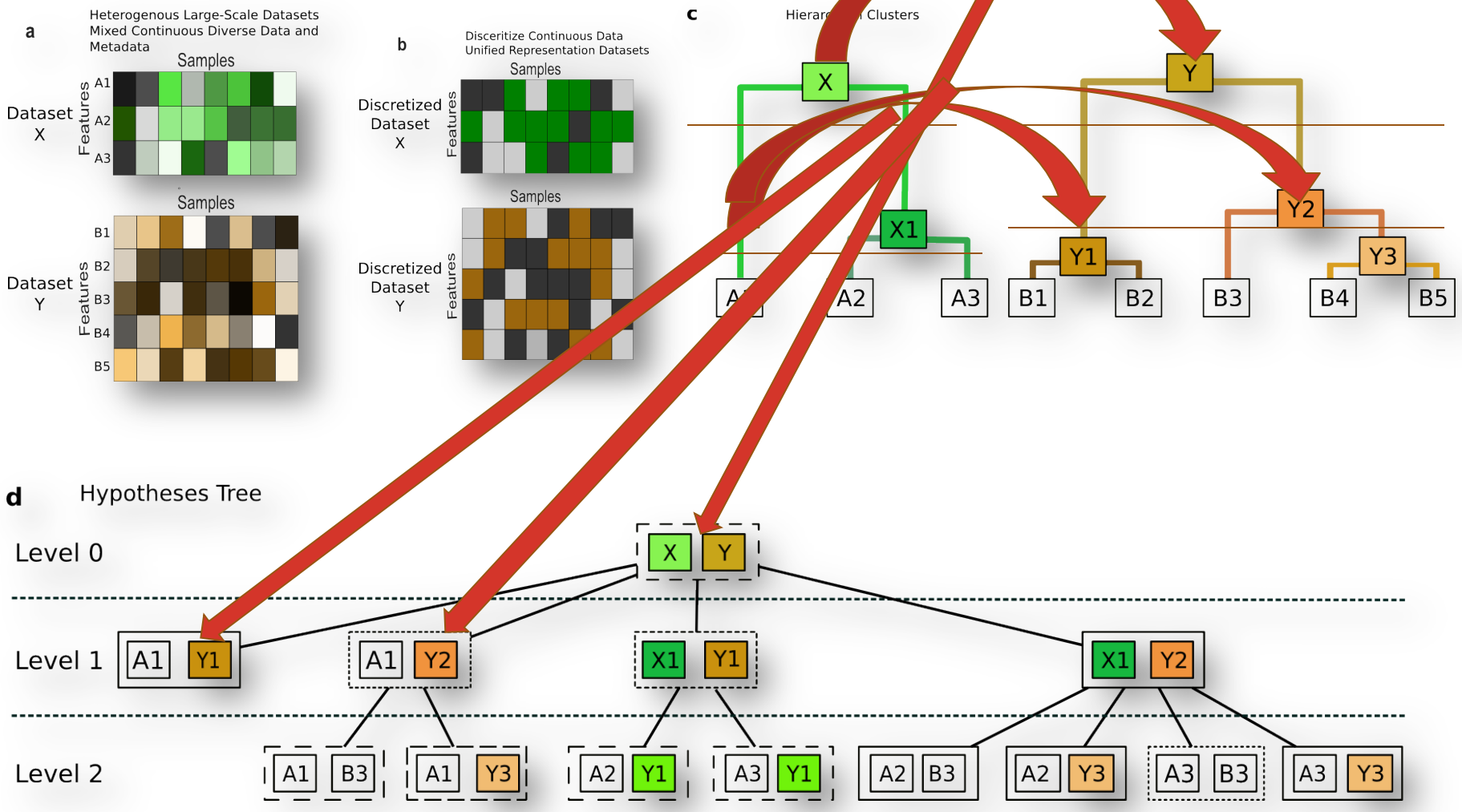
Example: Multi-omics Integration



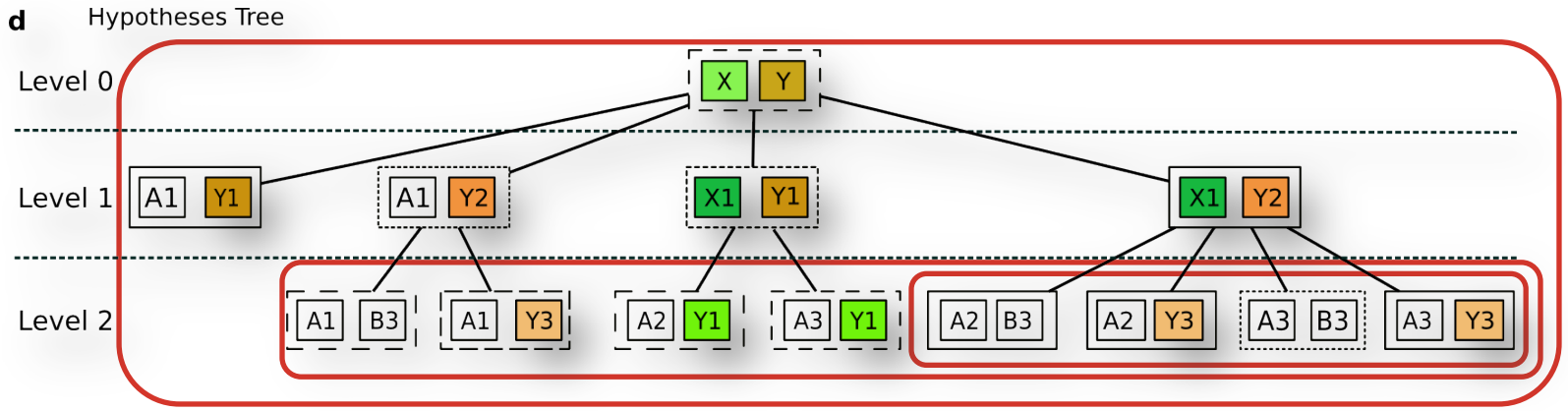
Proposed method

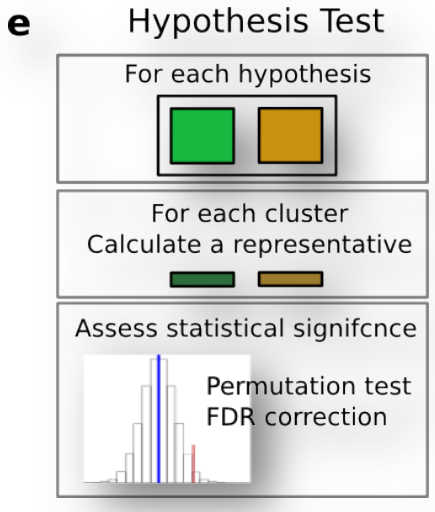
HIERARCHICAL ALL-AGAINST-ALL ASSOCIATION TESTING

Discretizing, Clustering, and Coupling

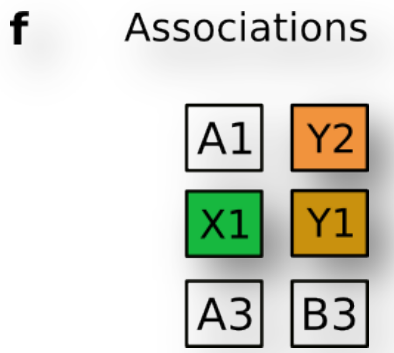


Hypothesis Testing, FDR Correction

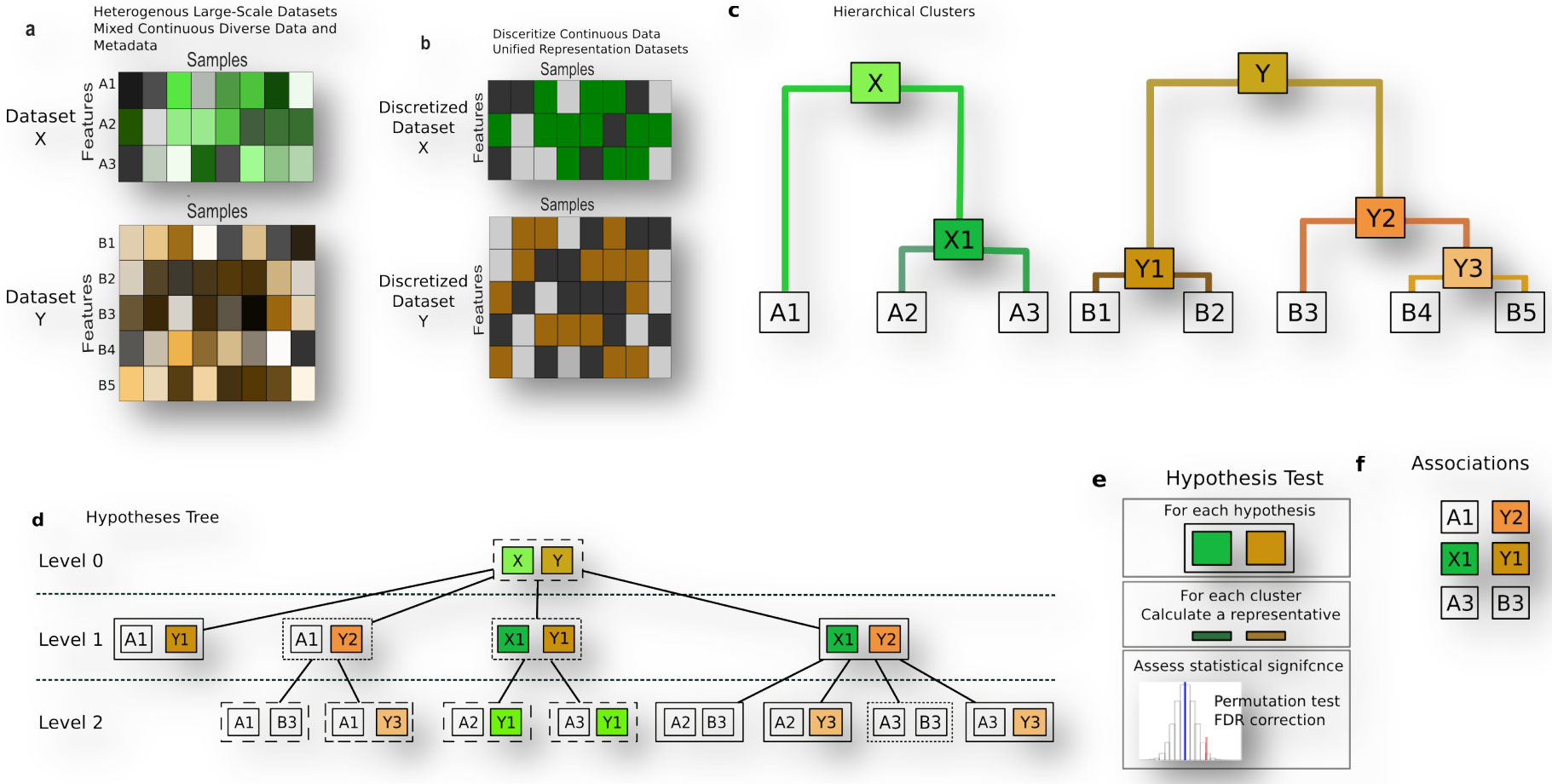




- Benjamini-Hochberg for FDR correction
  - Family of hypotheses
  - Level of hypotheses
  - All tested hypotheses



Method Overview

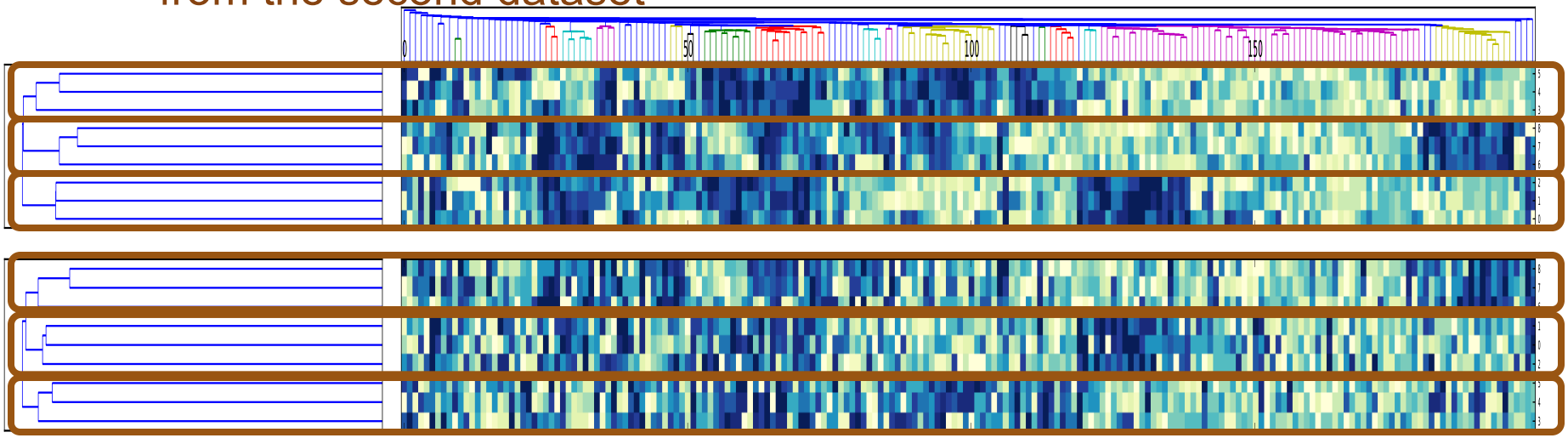


Validation

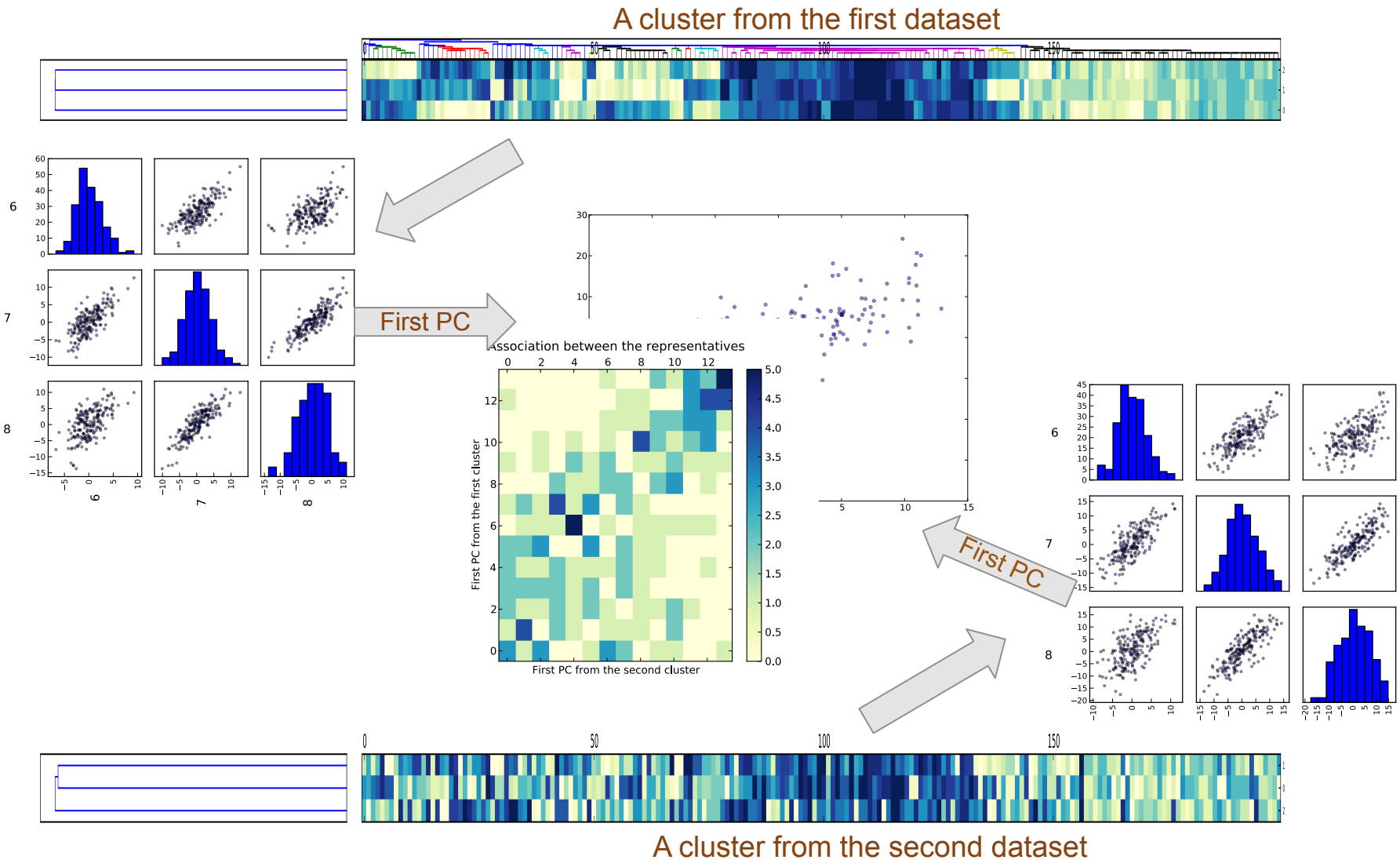
SYNTHETIC DATA

Simulated datasets

- Two synthetic datasets:
  - 12 features and 200 samples
  - 3 block of features with each datasets
  - Features within blocks are linearly correlated
  - The second dataset is spiked linearly from the first dataset
  - Each block from the first dataset is linearly correlated with one block from the second dataset



Associations



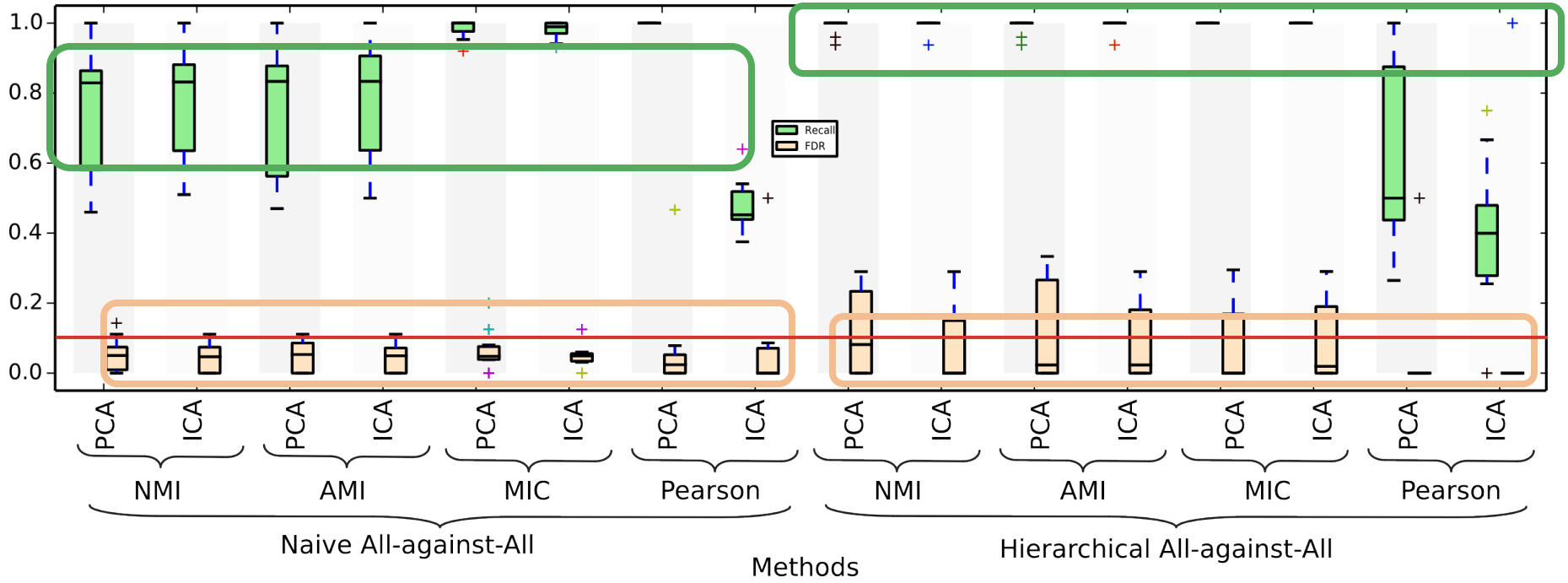
Evaluation

PERFORMANCE

Simulated Datasets

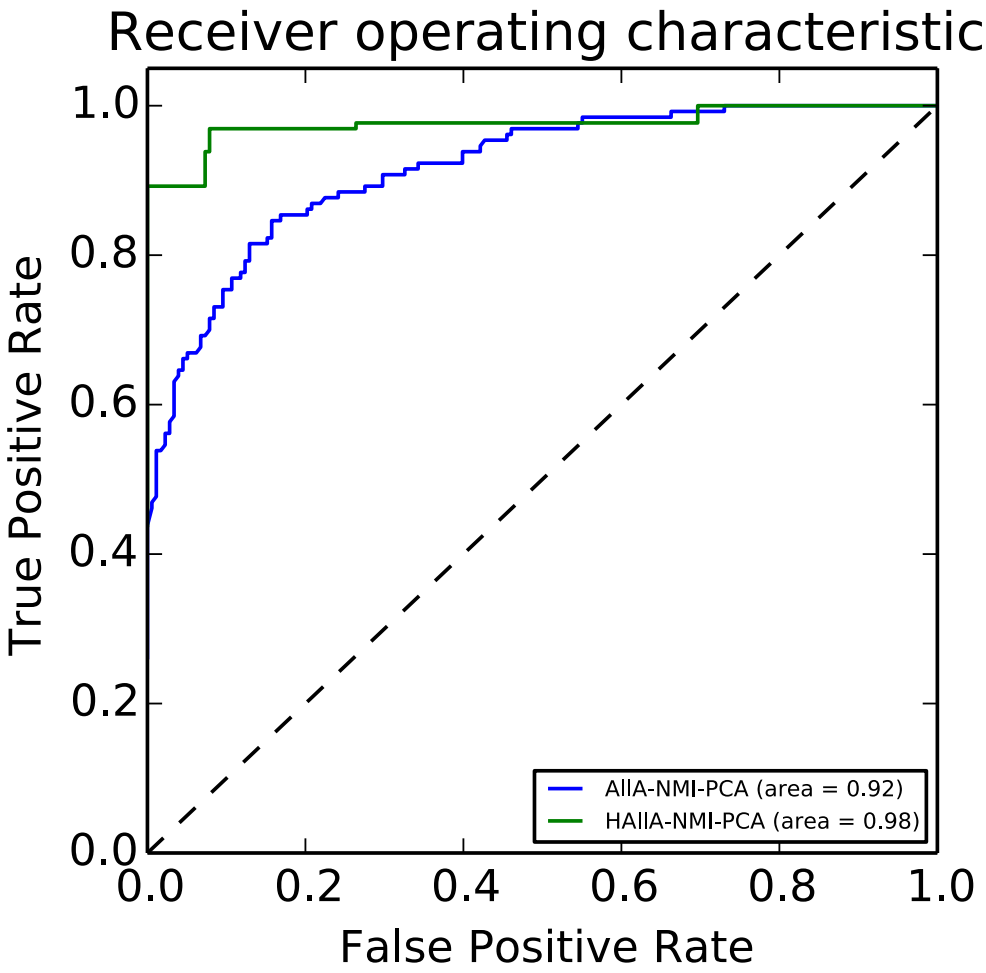
- 10 different simulated datasets are used:
  - Datasets scaled from 4 features and 100 samples to 24 features and 500 samples.
  - These datasets have simulated clusters in their features with linear association within and between them.

Naïve AllA vs. HAllA



- Higher statistical power in HAllA
- Benjamini-Hochberg has controlled the FDR in the naïve All-against-All approach

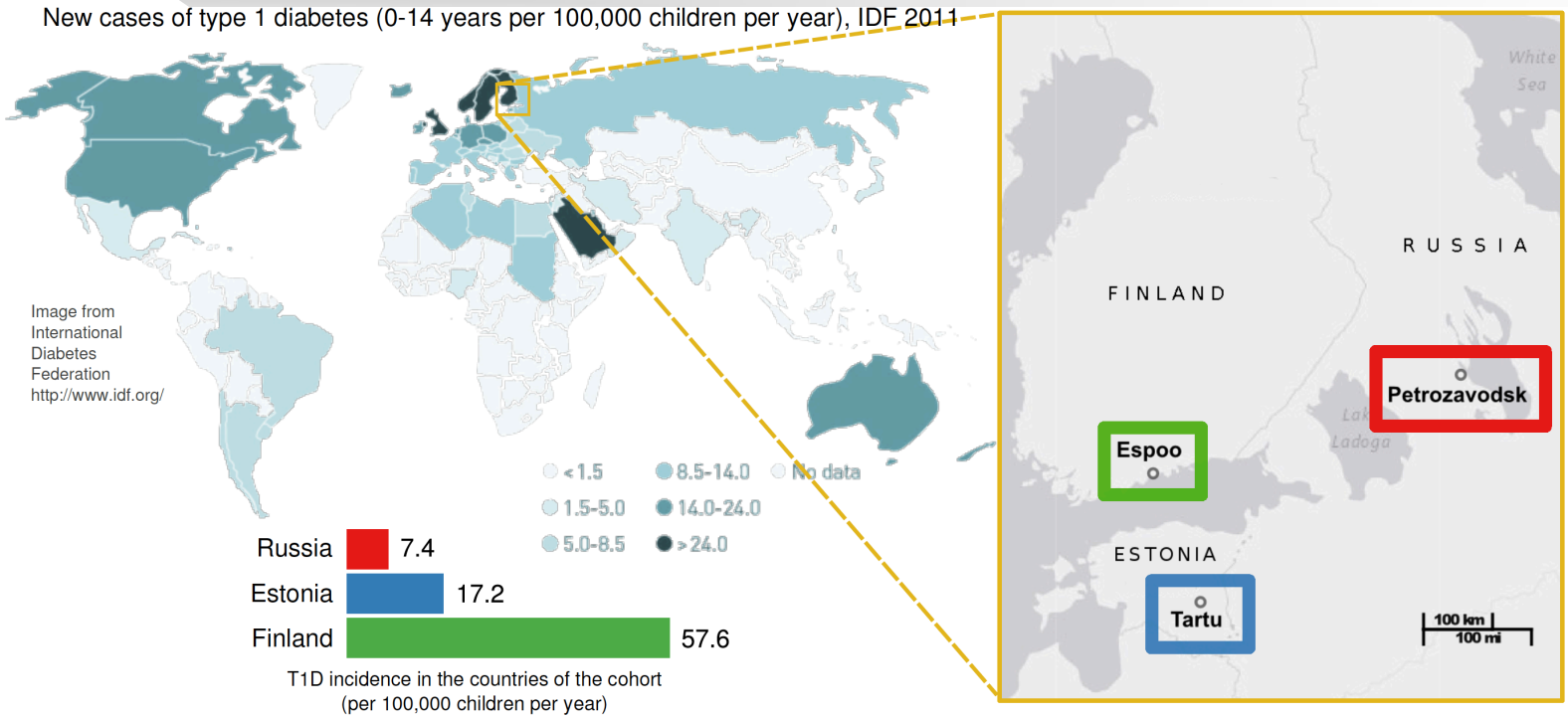
HAllA vs. AllA: ROC Curve

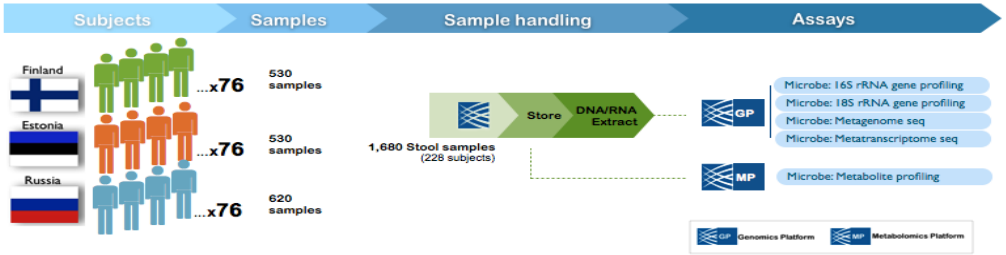


Application

GENERA VS. METABOLITES

The Multi-Country DIABIMMUNE Study





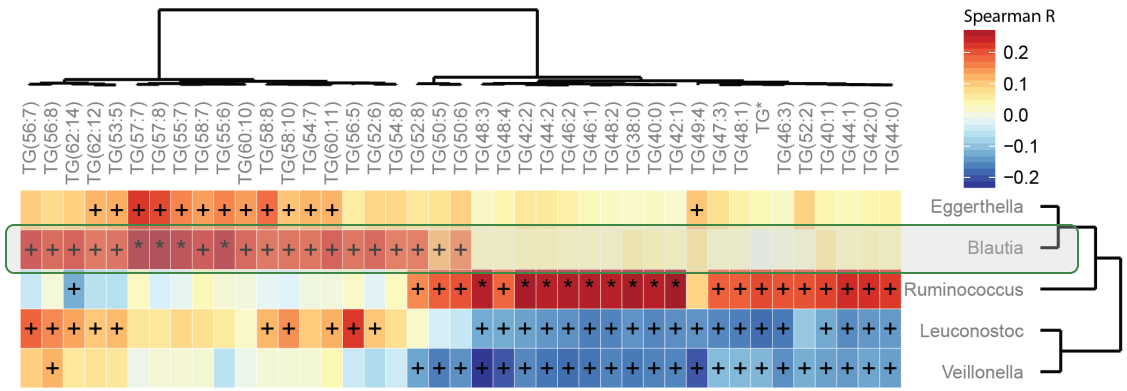
Two datasets for 104 samples :

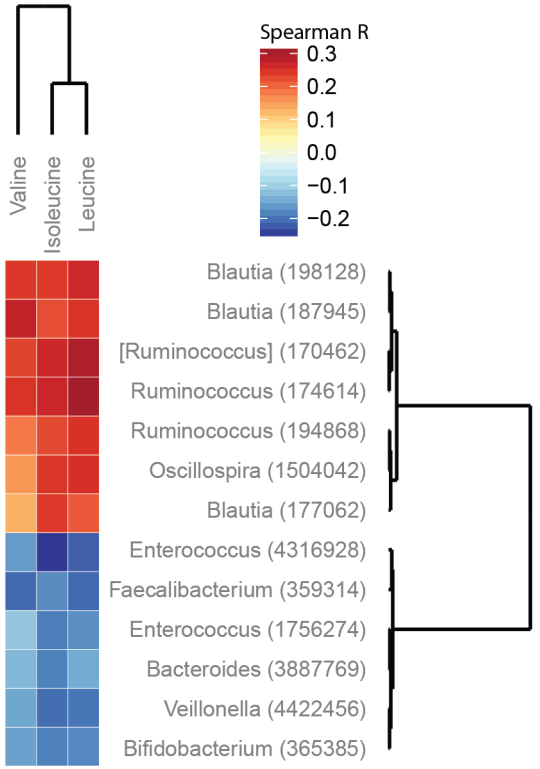
- 16S rRNA gene sequencing to profile 20 genera
- 284 labeled metabolites

Aleksandar Kostic

Tommi Vatanen

Serum lipids and metabolites are associated with T1D-related microbial taxa



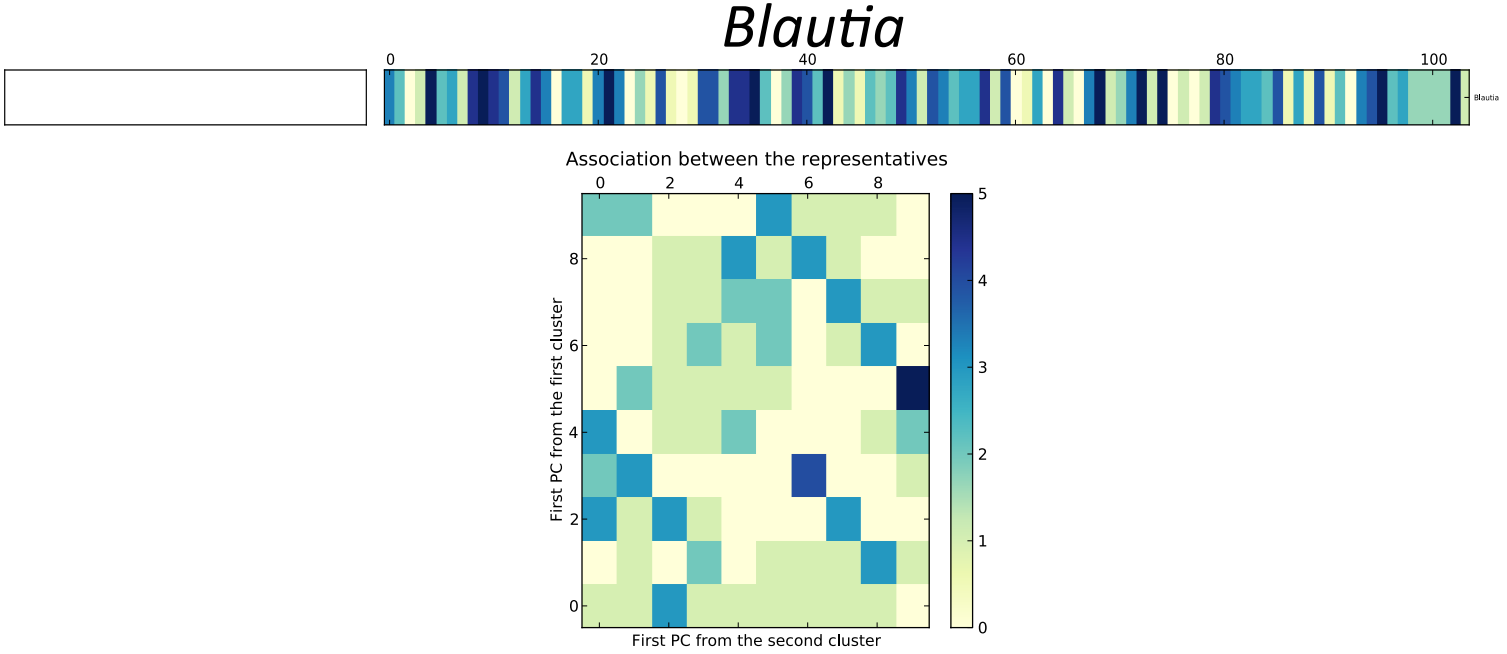


- Positive correlation
  - *Blautia* and long-chain triglycerides
  - *Ruminococcus* and short-chain triglycerides
- Negative correlation
  - *Veillonella* and short-chain triglycerides

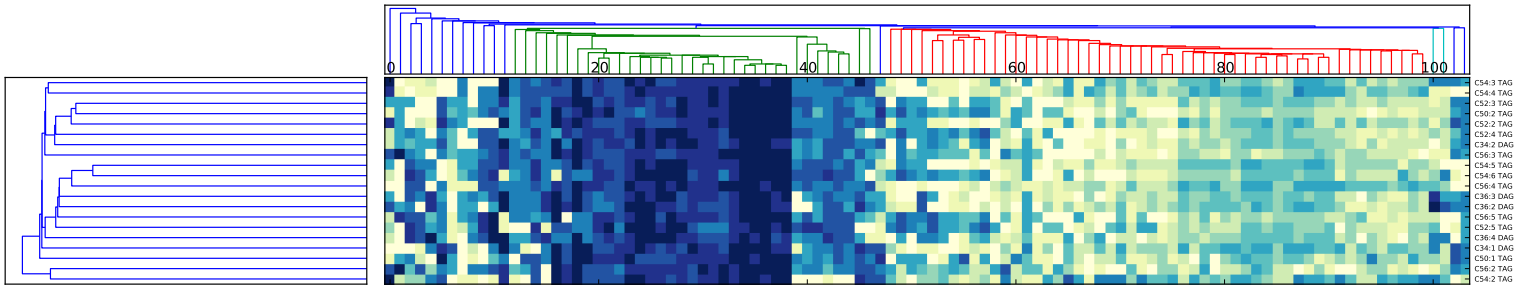
- Positive correlation
  - *Blautia* and branched-chain amino acids
  - *Ruminococcus* and branched-chain amino acids
- Negative correlation
  - *Veillonella* and branched-chain amino acids

Tommi Vatanen

Blautia and long-chain triglycerides



long-chain triglycerides



Software

FEATURES

Extensibility

Testing Approaches

- Naïve AllA
- Hierarchical AllA

Similarity Metrics

- NMI
- MIC
- Pearson
- AMI

Decomposition Methods

- PCA
- ICA
- KPCA
- PLS
- CCA

Packaging

- Python package
  - Software
  - Unit tests
  - Evaluation Modules
- Online References
  - Hierarchical All-against-All(HAllA)
    - http://huttenhower.sph.harvard.edu/halla
  - The Huttenhower Lab
    - http://huttenhower.sph.harvard.edu

Installation and Running

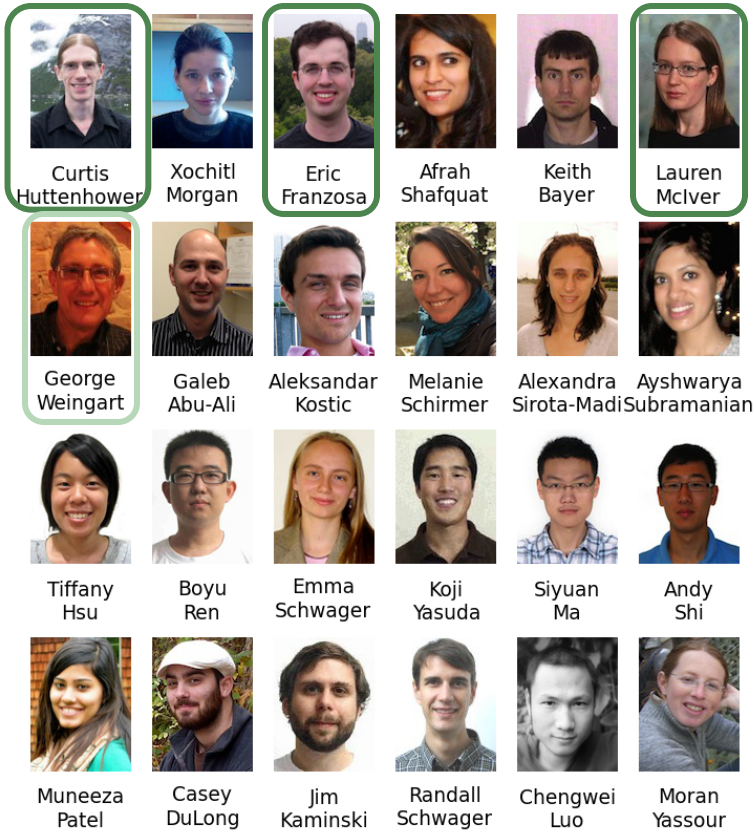
- Download
- python setup.py install
- Command line
  - halla –X dataset1.txt –Y datsest2.txt –o output --plotting-results

Summary Advantages

- Works with heterogeneous data (continuous+ categorical)
- Able to detect individually weak, but collectively strong signals
- Interpretable approach to data dimension reduction
- Assigns q-values to associations with high sensitivity
- Extensible to the use of different clustering and association methods

Acknowledgments

The Huttenhower Lab



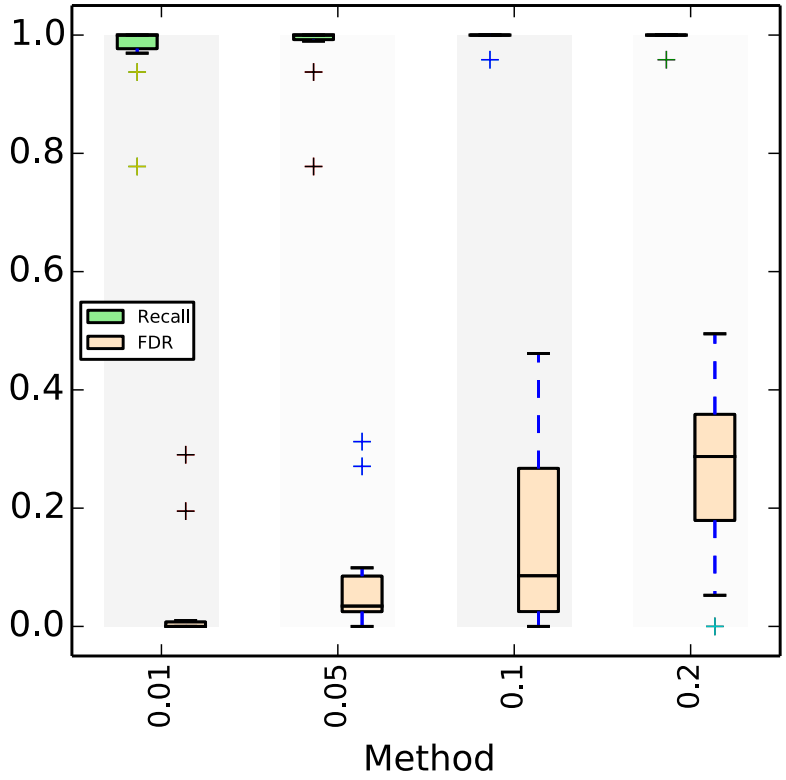
Alumni

- Levi Waldron
- Yo Sup Moon

Harvard T.H. Chan School of Public Health
Department of Biostatistics

FDR Controlling

Benjamini-Hochberg-Family FDR correction



Benjamini-Hochberg-Level FDR correction