Statistical tests and feature selection in unsupervised learning

Neo Christopher Chung

Lecture 6, 1000-719bMSB

Inference vs. Prediction

There are, generally speaking, two sides of data analysis (or data science).

Statisticians care more about **inference**

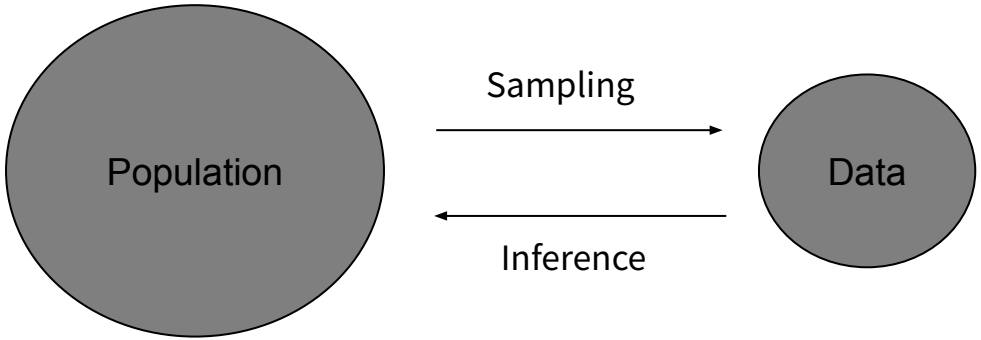
> to learn about the data generating process.

Machine learning focuses more on **prediction**

> to predict the outcomes of the new data.

Our modeling choices depends on our goals

Inference

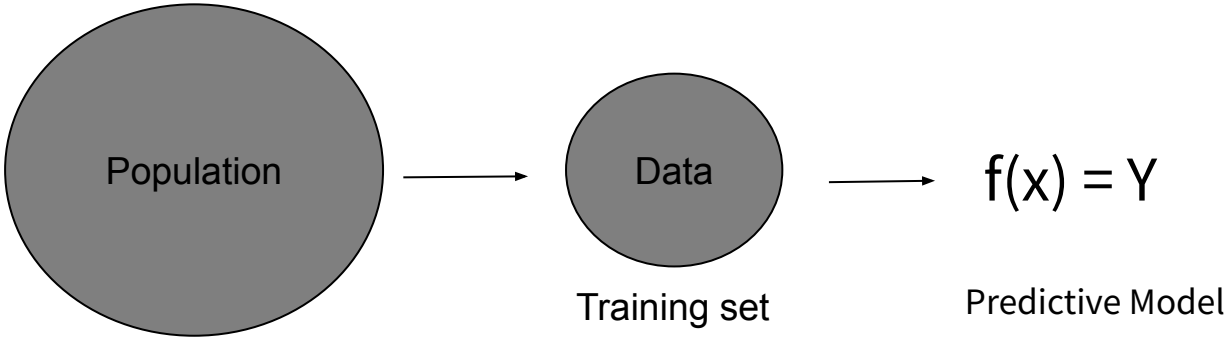


- “Data Modeling Culture” by Leo Breiman
- Extract information from the data about the underlying mechanism producing the data.
- Inferring the important characteristics of the population relies on accurate and robust models, estimation methods, and interpretations.
- Model interpretability is the most critical for accurate inference.
- E.g., Generalized linear models (linear or logistic regression), generalized additive model

Inference

1. **Modeling:** Reason about the data generation process and choose the stochastic model that approximates the data generation process best.
2. **Model validation:** Evaluate the validity of the stochastic model using residual analysis or goodness-of-fit tests.
3. **Inference:** Use the stochastic model to understand the data generation process.

Prediction



- “Algorithmic Modeling Culture” by Leo Breiman
- Predictive models are created solely based on their performance in predicting the testing set (not in the training set). Thus, it does not need to understand or elucidate the actual data generation process.
- Interpretability isn’t a focus, although it often becomes important.
- E.g., support vector machines, decision trees, neural networks, random forests

Prediction

1. **Modeling:** Consider several different models and different parameter settings.
2. **Model selection:** Identify the model with the greatest predictive performance using validation/test sets; select the model with the highest performance on the test set.
3. **Prediction:** Apply the selected model on new data with the expectation that the selected model also generalizes to the unseen data.

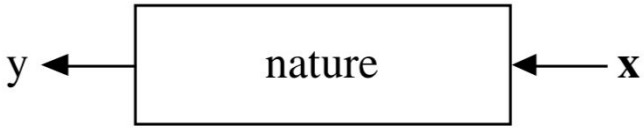
Statistical Modeling: The Two Cultures

1. Rashomon: the multiplicity of good models
    a. Small perturbation to the data may result in different good models.
    b. Different subset of features could be selected with similar performances.
2. Occam's Razor: the conflict between simplicity and accuracy
    a. The simple models are often more robust or generalizable
    b. The complicated models (e.g., neural nets) would provide higher accuracy (LMs)
3. Richard Bellman: dimensionality - curse or blessing?
    a. "The curse of dimensionality" advises us to reduce the features
    b. We can leverage weak predictors in many features to obtain a strong predictor

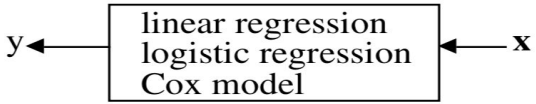
→ "The goal is not interpretability, but accurate information."

Leo Breiman (2001)

Data Generation Process



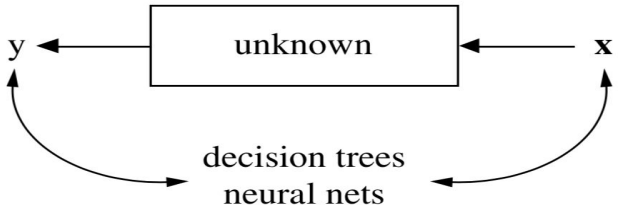
Data Modeling / Inference



Medical doctors care about interpretability, even if an algorithm may provide a higher accuracy.

Genetics researchers may focus only on inference as the goal is to understand the mechanisms.

Algorithmic Modeling / Prediction



Algorithms for image classification and localization do not need to have interpretability.

The “unknown” nature may be too complex to make reasonably simple models.

Leo Breiman (2001)

Prediction and Inference, in practice

Research at the intersection of those two cultures, since Breiman's 2001 paper.

Some methods are used in both ways -- e.g., a logistic regression

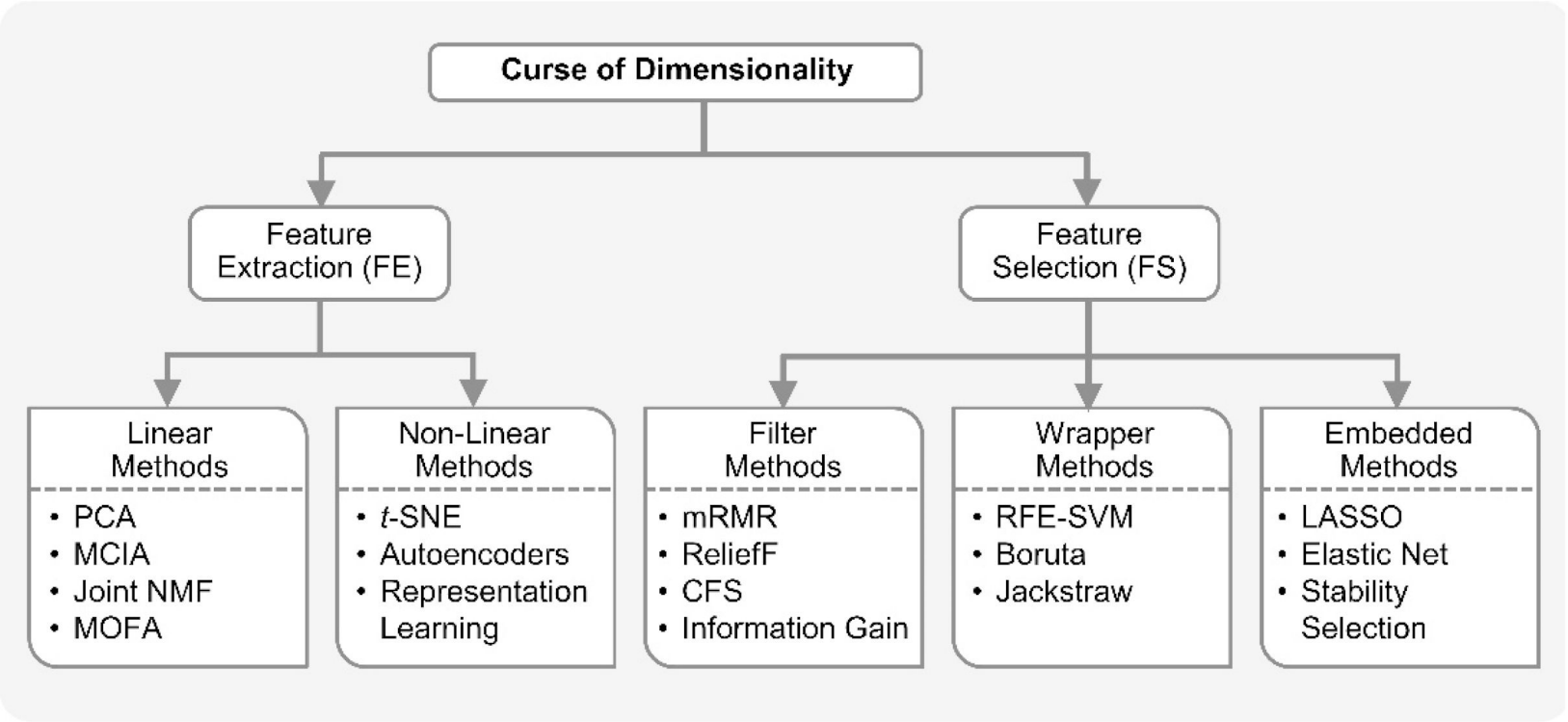
The ultimate goal is to bridge this gap from either side.

In machine learning, interpretability is one of the most important topics currently.

In statistical learning, sensitivity and specificity mirrors that in prediction.

Feature selection can be seen in both cultures, with an increasing importance

Feature selection in the context



Mirza et al. 2019

Feature selection

With a large dataset with many features, we want to carry out feature selection:

- Occam’s Razor
- Training or converging faster
- Reduces the complexity of a model.
- Easier to interpret and to generalize.
- Potentially improves the performance / accuracy of a model
- Prevent overfitting.

Filter methods

Filter methods are used to select a subset of relevant features independent of any model, oftening as a preprocessing step.

Selected simply on their statistics or metrics with respect to the outcome variable.

Many of the filter methods are univariate and often dont consider co-linearity

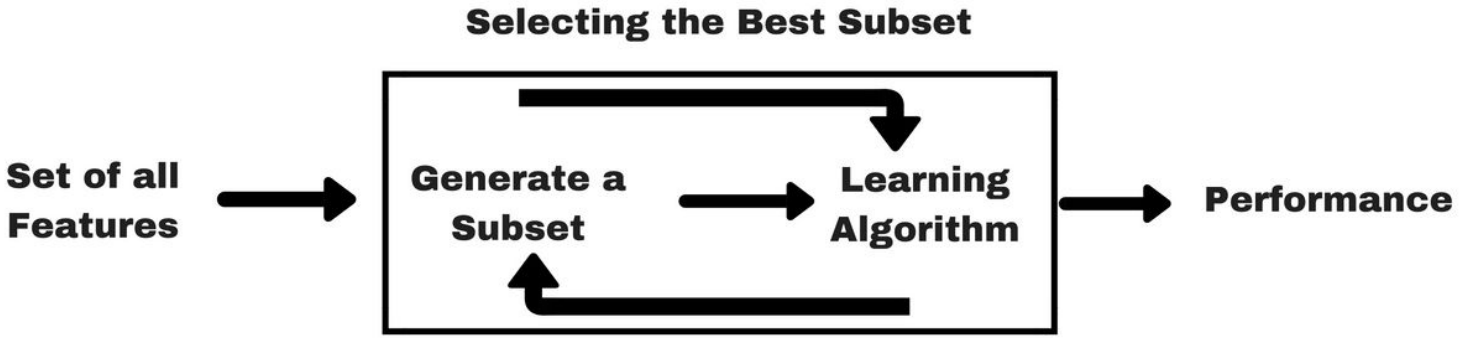
Mirza et al. 2019;
http://analyticsvidhya.com

Wrapper methods

Search for the best feature combination by training a particular predictive model repeatedly

Keep the best or worst performing subsets

Boruta with random forests; and Jackstraw with latent variables



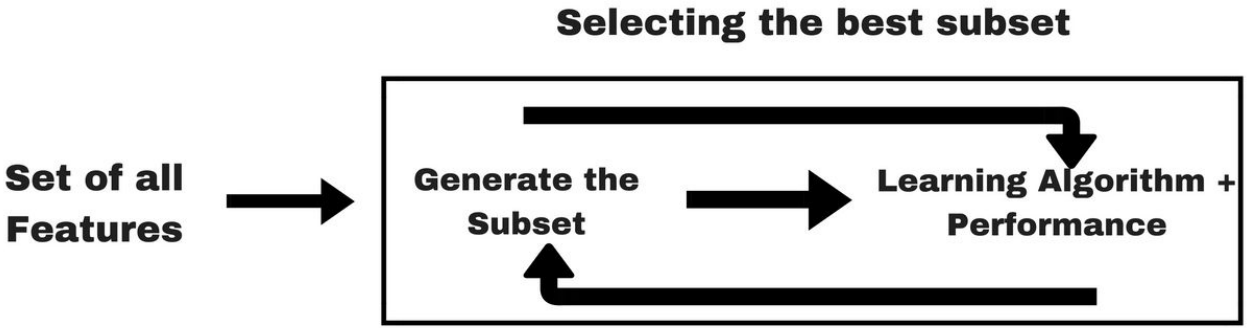
Mirza et al. 2019;
http://analyticsvidhya.com

Embedded methods

The algorithms with built-in feature selection methods such that they perform feature selection as a step toward predictive model building.

Embedded methods are in between filter and wrapper methods in terms of computational complexity.

E.g., Lasso, elastic net, penalized matrix decompositions



Mirza et al. 2019;
http://analyticsvidhya.com

Linear Model

Consider *m* variables of *n* observations

As a convention, we are stacking vectors ($\mathbf{y_1}\ \mathbf{y_2}\ \ldots\ \mathbf{y_m}$) as rows into a *m* x *n* matrix Y:

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} + \mathbf{E}$$

$$\begin{Bmatrix} y_{11} \cdots y_{1n} \\ y_{21} \cdots y_{2n} \\ \vdots \\ y_{m1} \cdots y_{mn} \end{Bmatrix} = \begin{Bmatrix} b_{11}\, b_{12} \cdots b_{1p} \\ b_{21}\, b_{22} \cdots b_{2p} \\ \vdots \\ b_{m1}\, b_{m2} \cdots \\ b_{mp} \end{Bmatrix} \begin{Bmatrix} x_{11} x_{12} \ldots x_{1n} \\ x_{21} x_{22} \ldots x_{2n} \\ \vdots \\ x_{p1} x_{p2} \ldots x_{pn} \end{Bmatrix} + \begin{Bmatrix} e_{11} e_{12} \ldots e_{1n} \\ e_{21} e_{22} \ldots e_{2n} \\ \vdots \\ e_{m1} e_{m2} \ldots \\ e_{mn} \end{Bmatrix}$$

Linear Model

Consider *m* variables of *n* observations

As a convention, we are stacking vectors ($\mathbf{y_1}$ $\mathbf{y_2}$ ... $\mathbf{y_m}$) as rows into a *m* x *n* matrix Y:

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} + \mathbf{E}$$

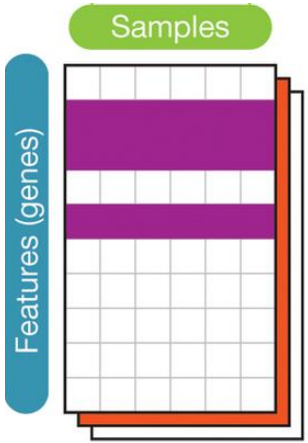
Known and measured

A linear model is an example of supervised learning

Estimation and significance testing on $\mathbf{b_i}$ help us identify $\mathbf{y_i}$ that are associated with **X**

Supervised vs. unsupervised learning

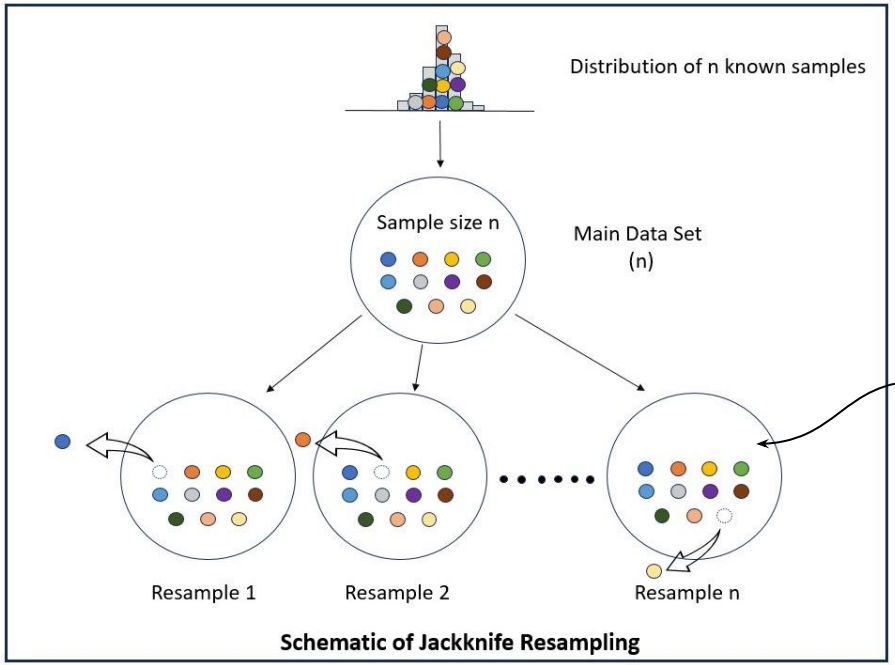
Microarray, RNA-seq, and others



When independent (e.g., biological) variables are **known,** we conduct supervised learning.

With **unknown independent variables,** use unsupervised learning

Jackknife



jackknife replicate

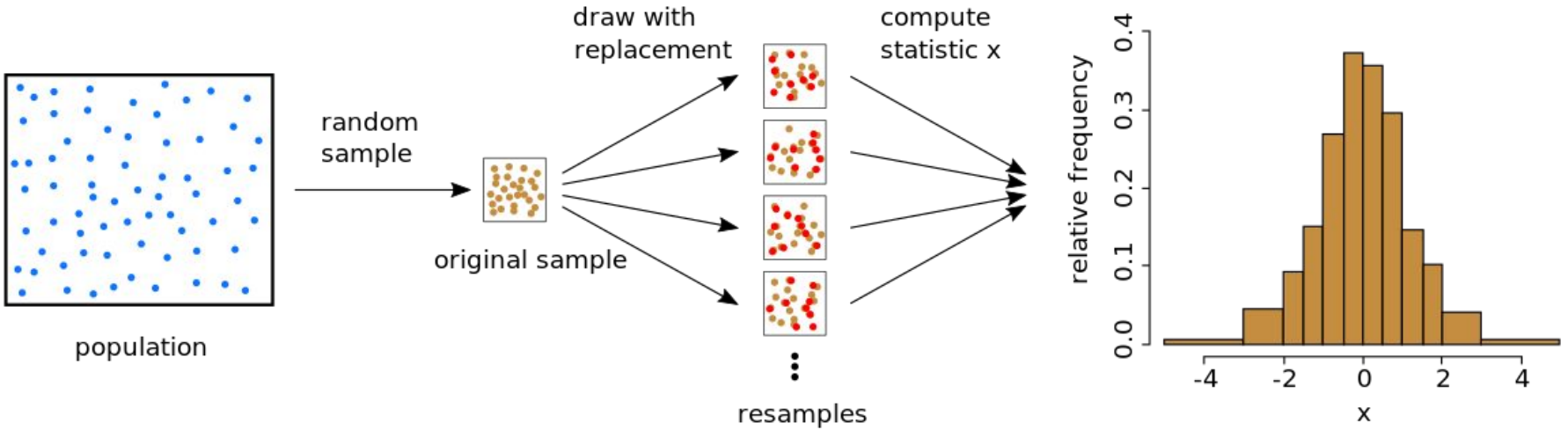
$$\bar{x}_{(i)} = \frac{1}{n-1} \sum_{j \in [n], j \neq i} x_j, \qquad i = 1, \ldots, n.$$

$$\bar{x}_{\text{jack}} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_{(i)}.$$

Given a sample of size n, a jackknife estimator can be built by aggregating the parameter estimates from each subsample of size (n − 1) obtained by omitting one observation.

Efron (1982), Wikipedia

Bootstrap: random sampling with replacement



The basic idea of bootstrapping is that inference about a population from sample data (sample → population) can be modeled by *resampling* the sample data and performing inference about a sample from resampled data (resampled → sample).

As the population is unknown, the true error in a sample statistic against its population value is unknown. In bootstrap-resamples, the 'population' is in fact the sample, and this is known; hence the quality of inference of the 'true' sample from resampled data (resampled → sample) is measurable.

Efron (1979), Wikipedia

Bootstrap, further variations

Parametric bootstrap:

Fit a parametric model and sampling from the fitted model

Smooth bootstrap:

Add random noise to each resampled observations. i.e., sampling from a kernel density estimate of the data

Resampling residual in regression

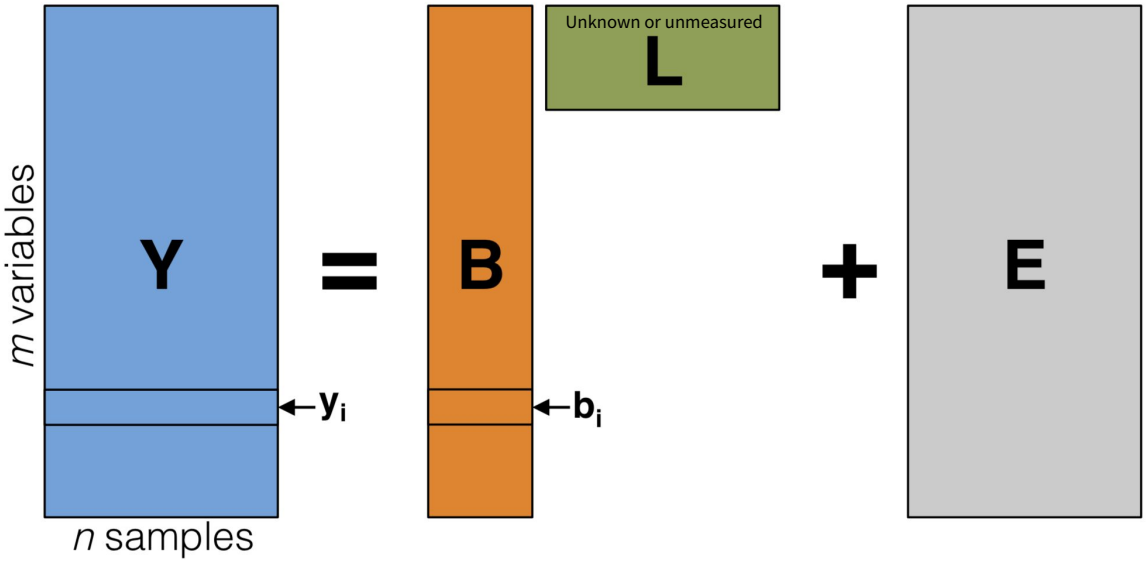
1. Fit the model and retain the fitted values $\widehat{y}_i$ and the residuals $\widehat{\varepsilon}_i = y_i - \widehat{y}_i, (i = 1, \ldots, n)$.
2. For each pair, (x_i, y_i), in which x_i is the (possibly multivariate) explanatory variable, add a randomly resampled residual, $\widehat{\varepsilon}_j$, to the fitted value $\widehat{y}_i$. In other words, create synthetic response variables $y_i^* = \widehat{y}_i + \widehat{\varepsilon}_j$ where j is selected randomly from the list (1, ..., n) for every i.
3. Refit the model using the fictitious response variables y_i^*, and retain the quantities of interest (often the parameters, $\widehat{\mu}_i^*$, estimated from the synthetic y_i^*).
4. Repeat steps 2 and 3 a large number of times.

Efron (1979), Wikipedia

Latent Variable Model

Statistical & principled model for unsupervised learning

$$Y = B\,L + E$$



Latent variable as a conditional expectation

Expected influence of **z** on **Y** by E[**Y**|**z**],

$$\mathbf{Y} = \mathrm{E}[\mathbf{Y}|\mathbf{z}] + \mathbf{E}$$

Estimate **L**(**z**), that is a row basis for E[**Y**|**z**]

This low dimensional **L**(**z**) is **manifestation of the latent variables in the data**.

When clear in context, we simply write **L**. Most of time, **z** is not knowable.

General latent variable models

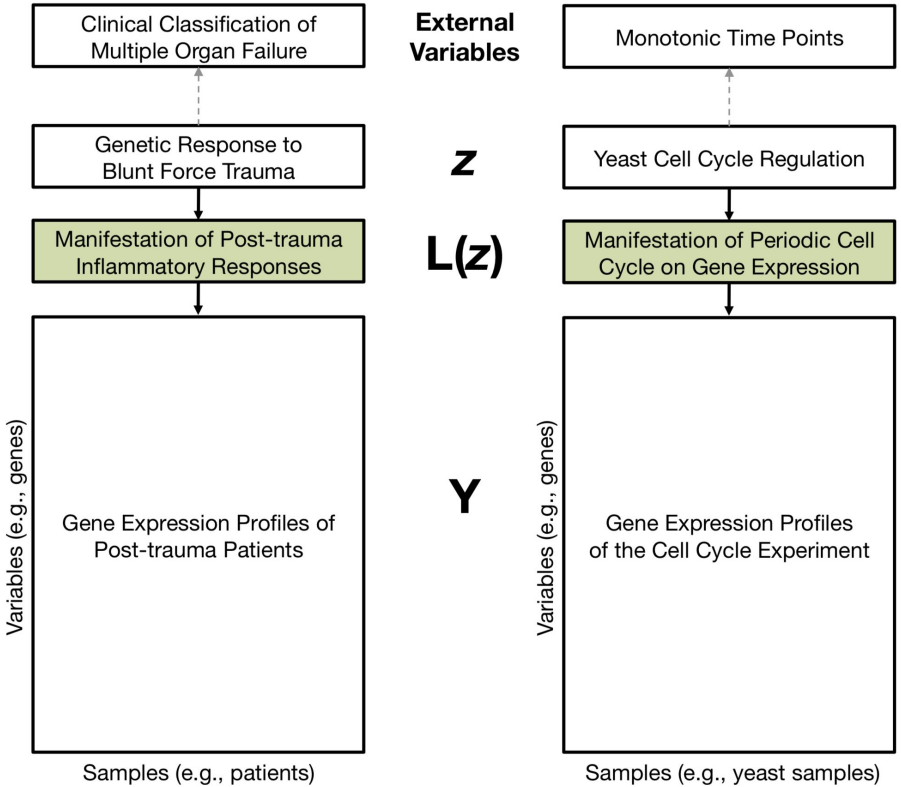
$$\mathbf{Y} = \mathbf{BL} + \mathbf{E}$$

L is estimate by the top r PCs (or related quantities), which we call V^T_r resulting in

$$\mathbf{Y} = \mathbf{\Gamma V}^T_r + \mathbf{E'}$$

E' denotes that it's an error term but distinct from the original **E**

Manifestation of latent variables



Molecular signatures ~ latent variables

In a right context, latent variables may contains molecular information about the experiments, environments, or diseases.

Even when some information is captured by labels (e.g., potential candidates for independent variables), they may not be accurate or precise.

In absence of independent variables, we would like to conduct significance testing with respect to latent variables (analogous to one in a linear model)

General latent variable models

$$\mathbf{Y} = \mathbf{BL} + \mathbf{E}$$

$$= \mathbf{\Gamma V}^{T}_{r} + \mathbf{E'}$$

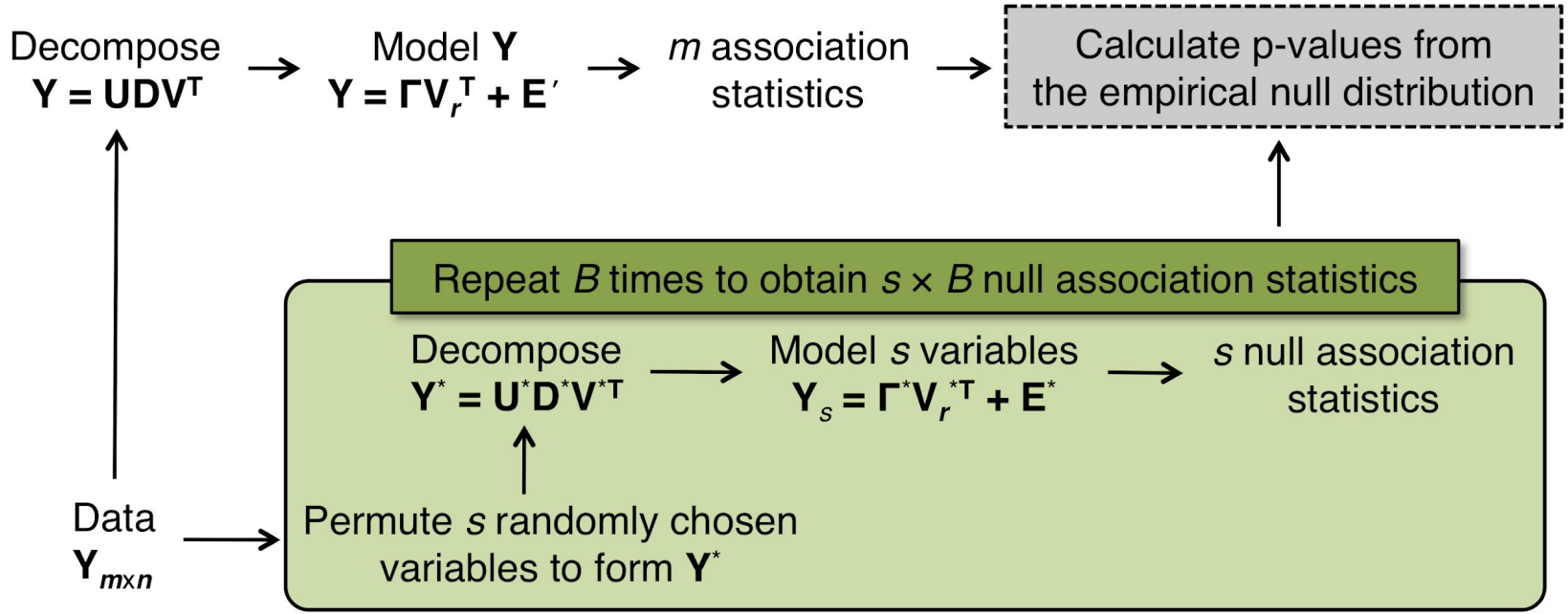
- Directly using $\mathbf{V}^{T}_{r}$ isn't testing the association between **Y** and **L**.
- $\mathbf{V}^{T}_{r}$ depends on **Y**.
- The association between **Y** and $\mathbf{V}^{T}_{r}$ are likely to be significant

Association tests between Y and L

Goal: which observed variables (e.g., genes) are statistically significantly related to the latent variables?

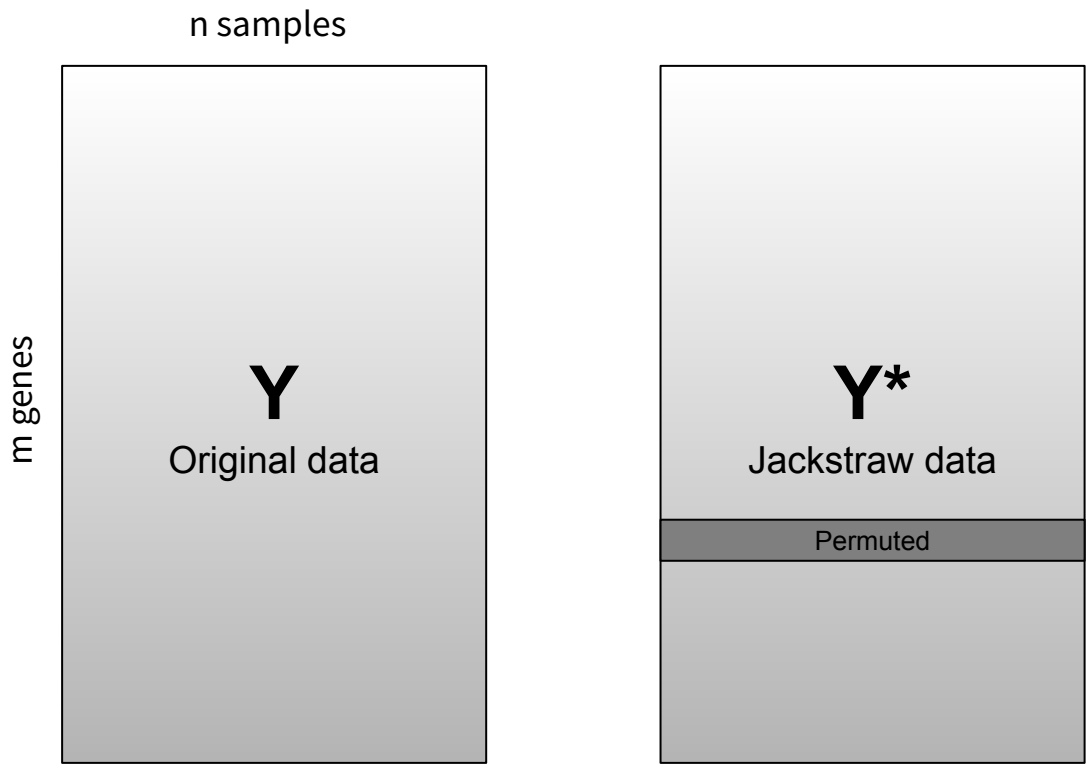
Solution: use the latent variable estimates (e.g., PCs) in association tests, while accounting for the fact that the latent variable estimates depend on the data

Jackstraw procedures



Chung and Storey (2015)

Jackstraw s=1



Simulation study

$$Y = BL + E$$

Simulating a case vs. control gene expression study, where 5% of genes are associated with L

Number of variables m = 1000

Number of observation n = 20

Proportion of null variables = .95

Latent variable function form L ~ dichotomous shift 

Non-null coefficients $\sim_{i.i.d}$ Uniform(0,1)

Illustrated example

$$\mathbf{Y} = \mathbf{BL} + \mathbf{E}$$

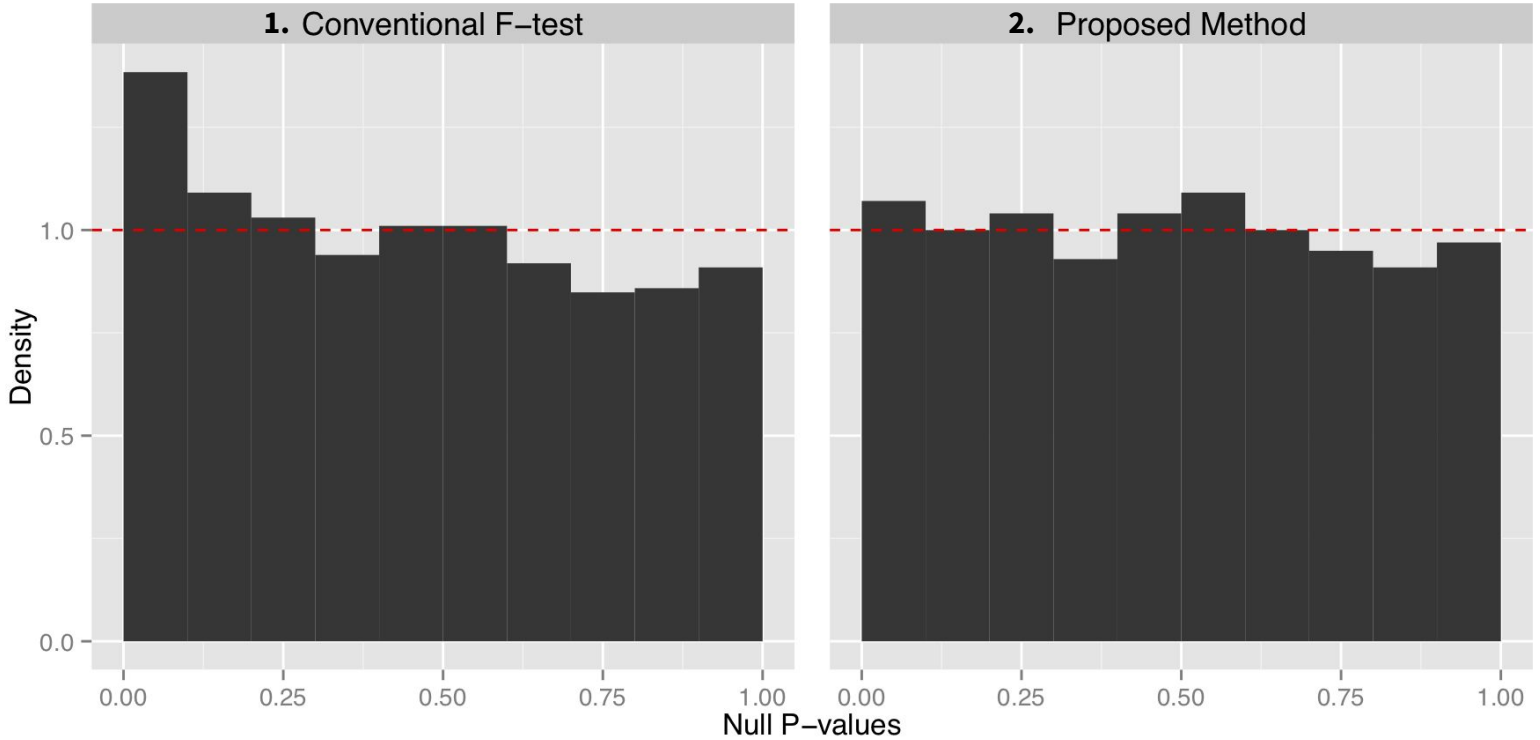
$$= \mathbf{\Gamma V}^{\mathsf{T}}_{r} + \mathbf{E'}$$

Simulate a data Y from where B, L, E that are generated following certain distributions
Compare two approaches, by looking at p-values associated with null variables (null p-values)

1. Naive association testing (e.g., conventional linear model) between **Y** and $\mathbf{V}^{\mathsf{T}}_{r}$
2. Correct association testing (e.g., the jackstraw) between **Y** and **L**

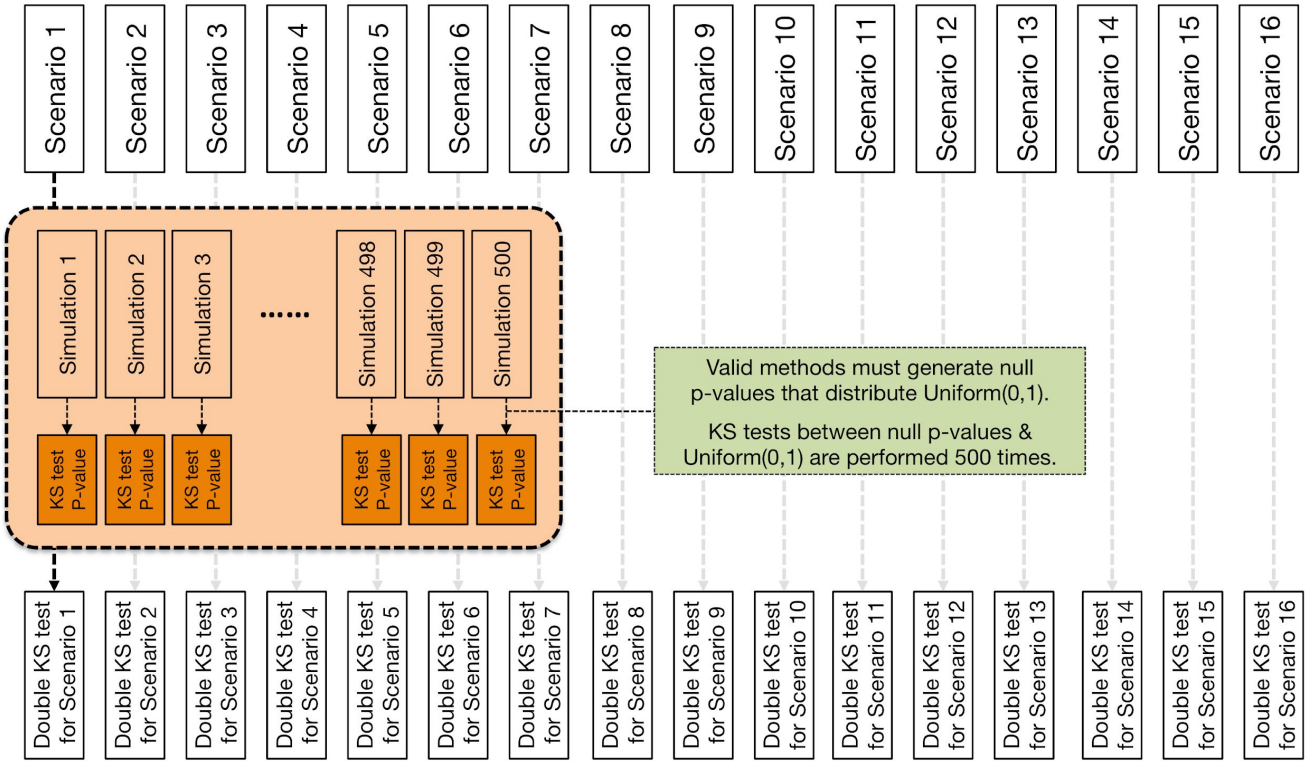
Null p-values evaluation

Evaluate statistical tests by looking at p-values associated with null variables (null p-values)



Red dashed lines indicates a theoretically correct distribution

To be sure, try diverse scenarios and use a Kolmogorov–Smirnov (KS) test for uniformity



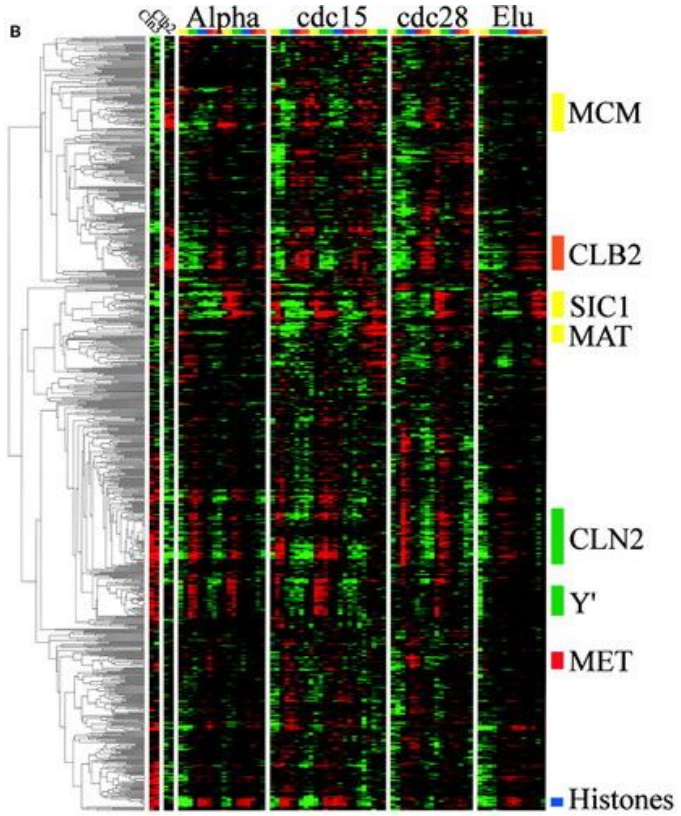
Application in gene expression study

The cell cycle patterns are needed to identify genes under regulation, but existing conventions are ad-hoc and arbitrary.

Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. P. T. Spellman et al. (1998)

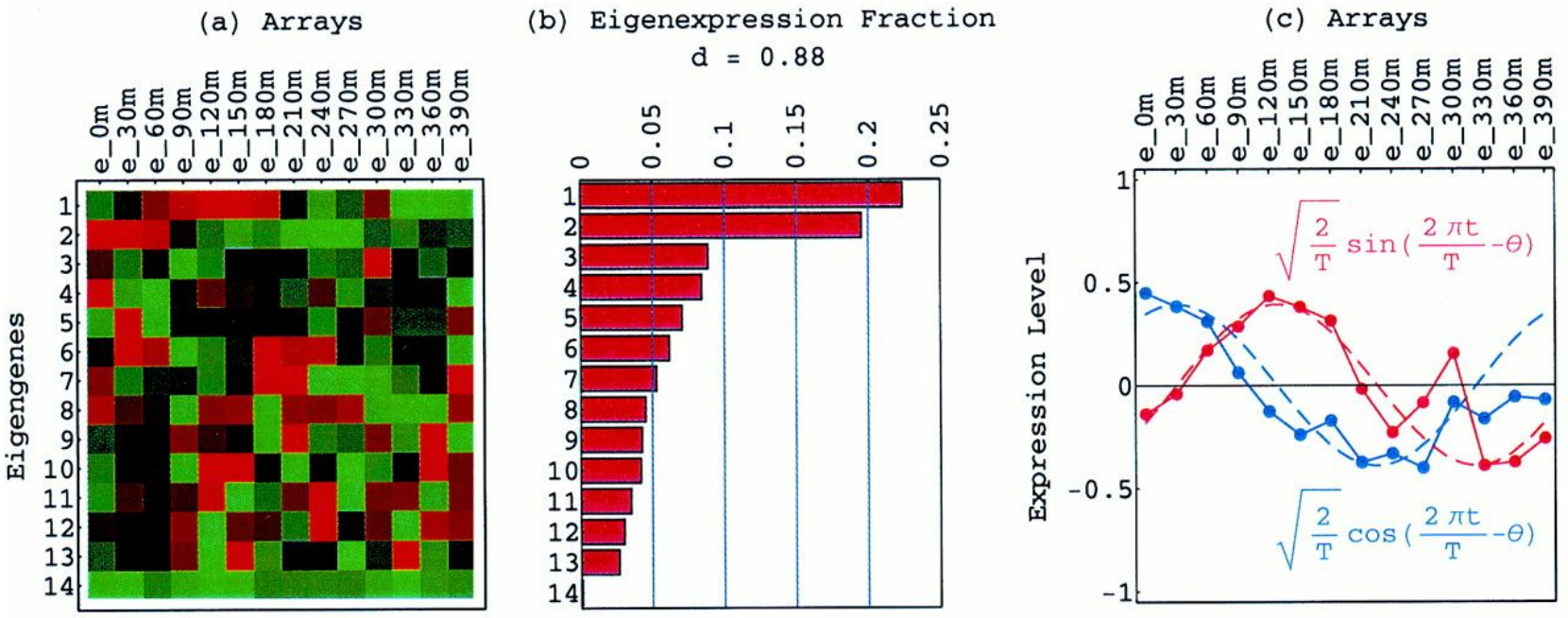
The cell cycle patterns can be estimated from gene expression data.

Singular value decomposition for genome-wide expression data processing and modeling by O. Alter et al. (2000)



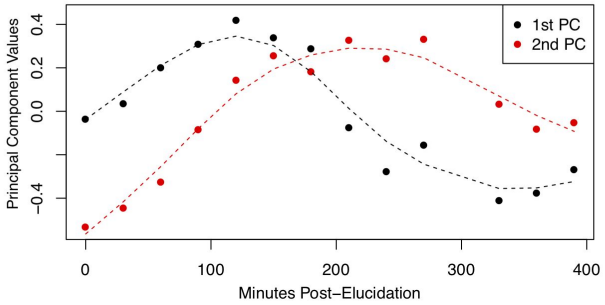
(A) Gene expression patterns for cell cycle–regulated genes. The 800 genes are ordered by the times at which they reach peak expression.

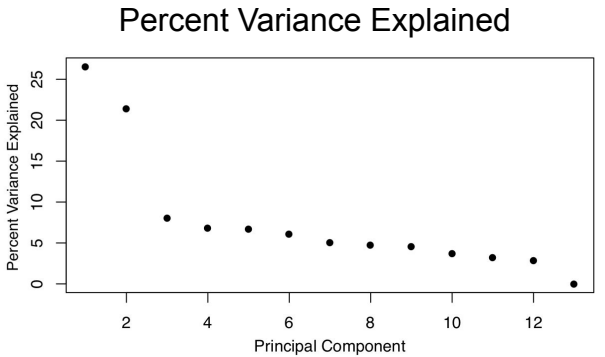
SVD/PCA of the yeast cell cycle data

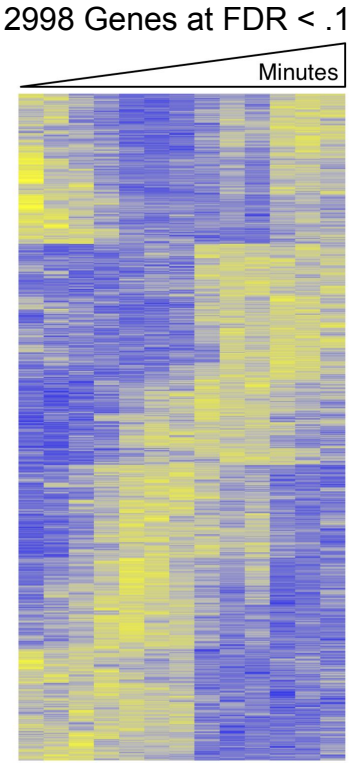


Alter et al 2000

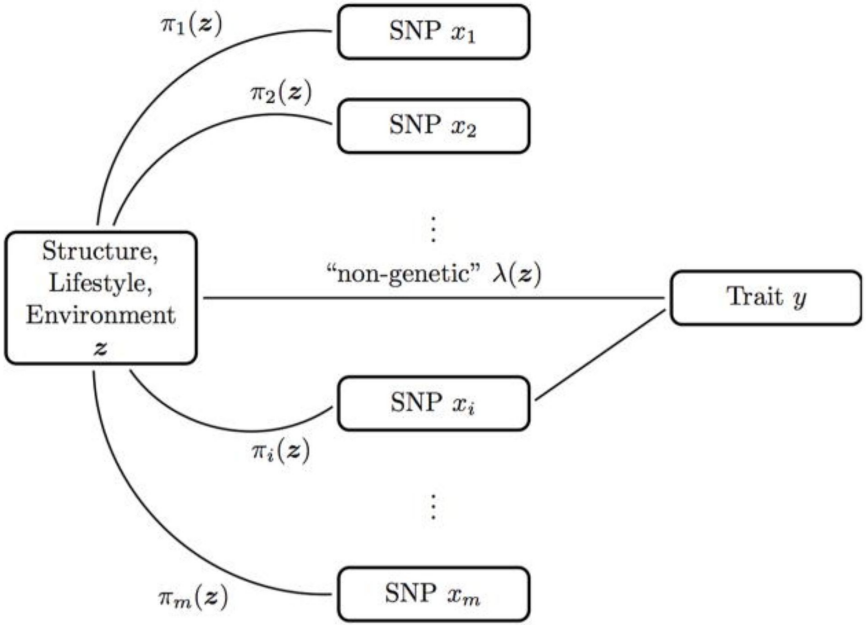
Apply the jackstraw test on the top 2 PCs







Application in population genomics



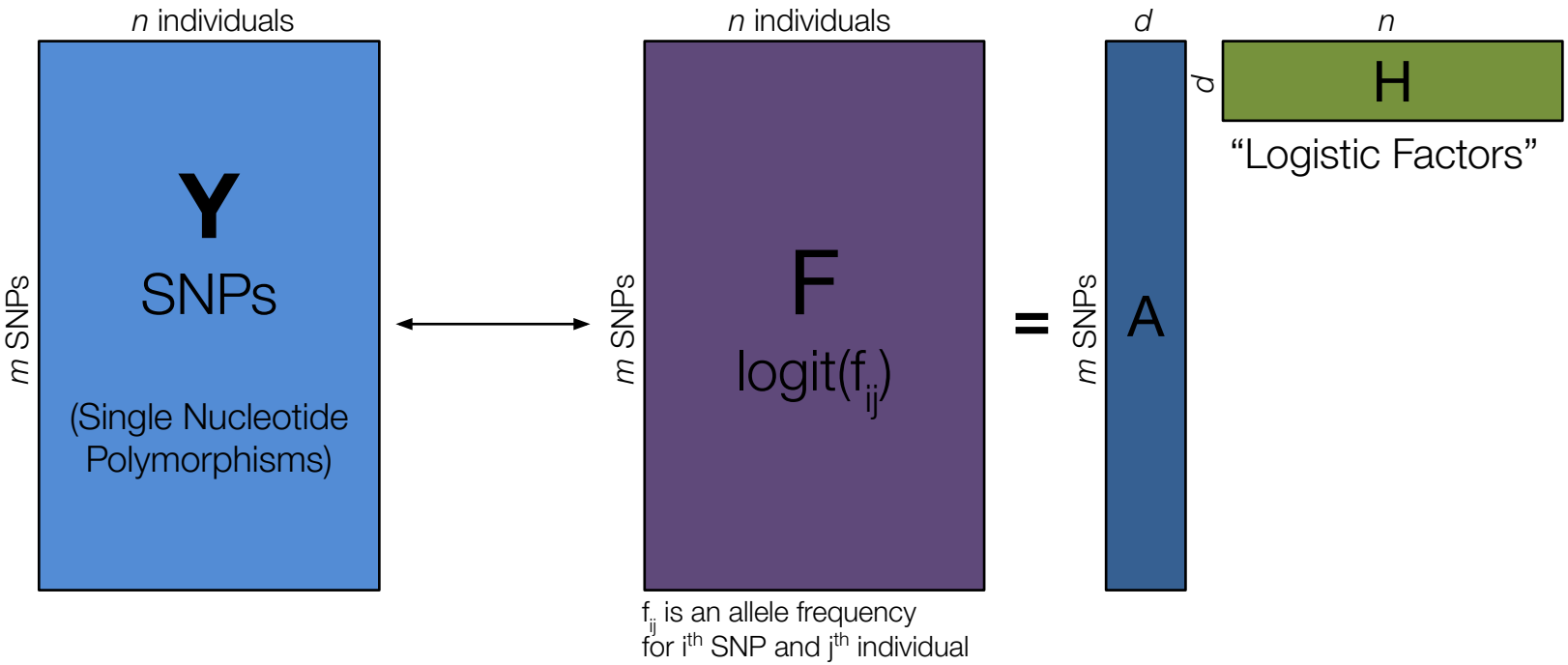
Geographical or self-report ancestry labels are ad-hoc, arbitrary, and sometimes misleading. Lewontin (1972)

Population structures are estimated from SNPs. Patterson et al. (2006), Novembre & Stephens (2008), Hao et al. (2013)

Use the jackstraw to identify SNPs (or genetic elements) that are associated with that estimated population structure.

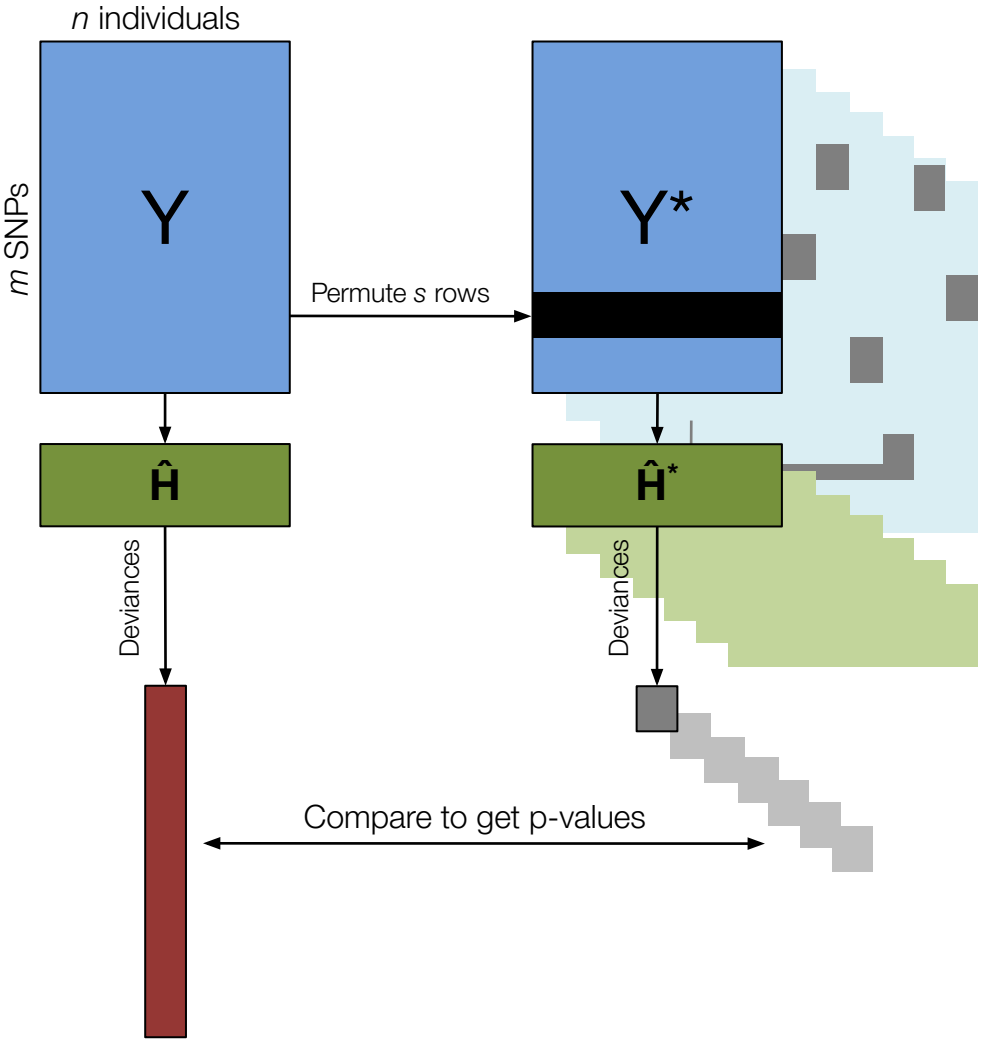
Population Structure in SNPs

Logistic Factor Analysis (Hao *et al.* 2013)

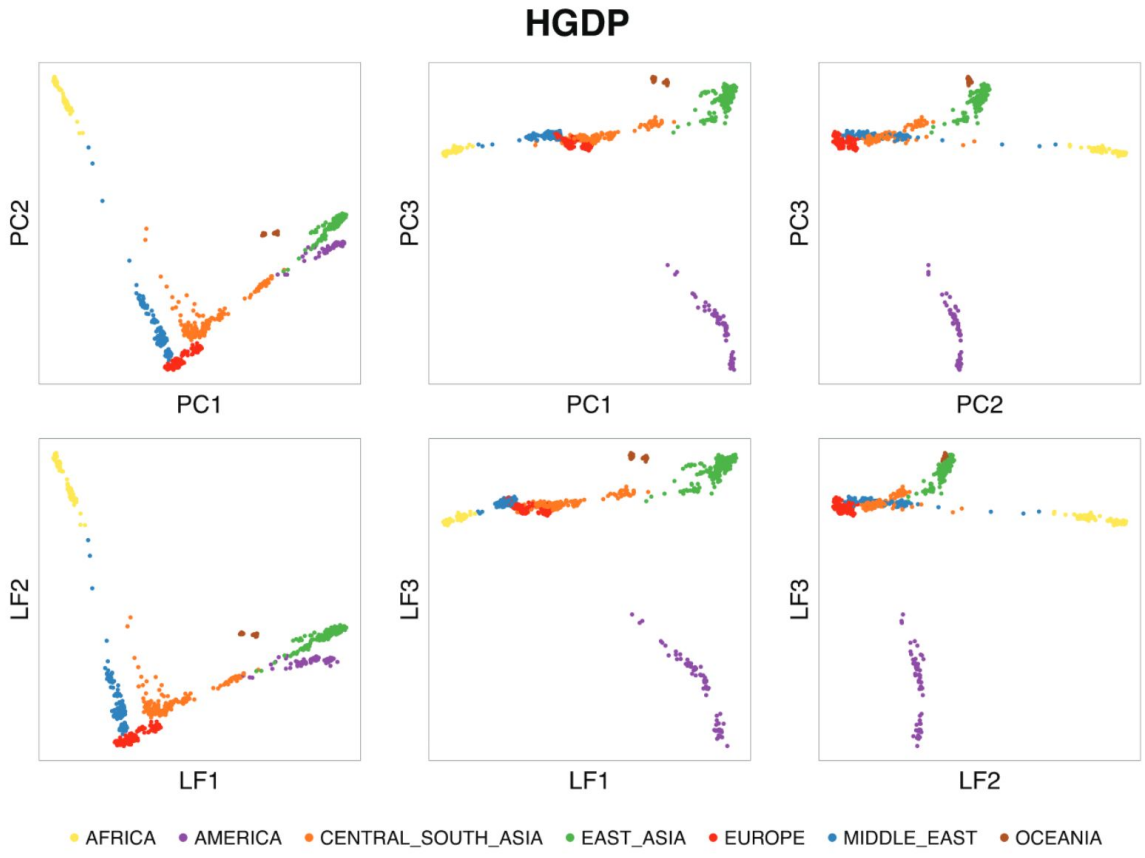


Jackstraw for LFA

Enables statistical testing of association between genetical population structure and SNPs



Human Genome Diversity Project

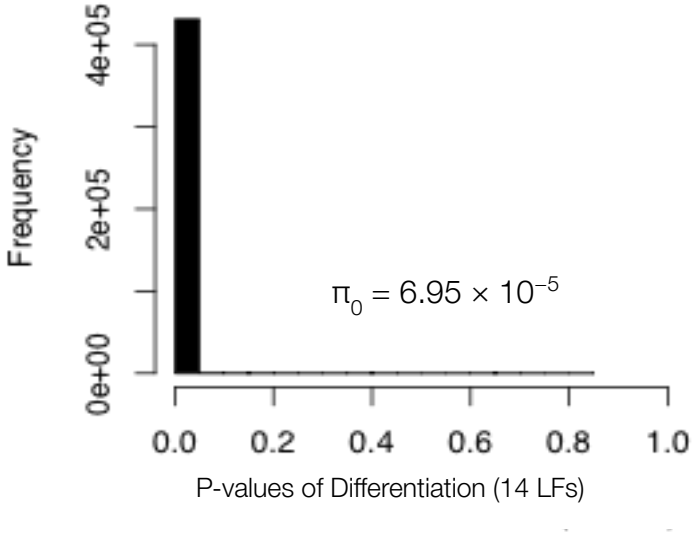


← Population structures are estimated with logistic factors.

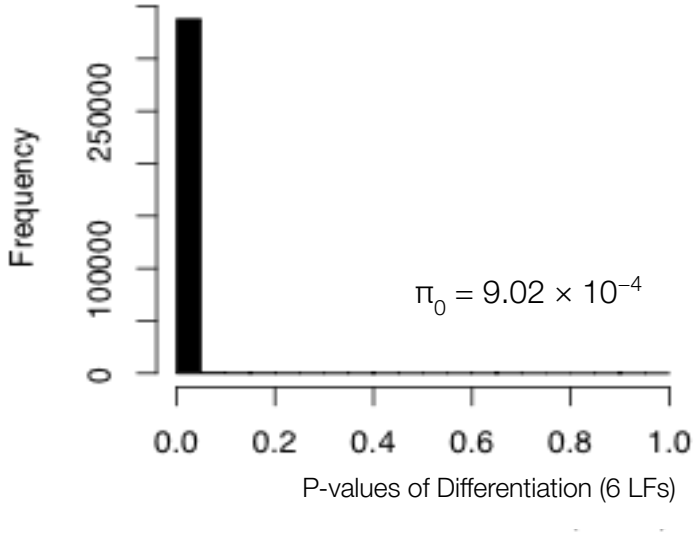
Do association tests (e.g., differentiation) using the jackstraw

Pervasive and weak genetic differentiation

Human Genome Diversity Project (HGDP)
m = 431345 SNPs for n = 940 individuals



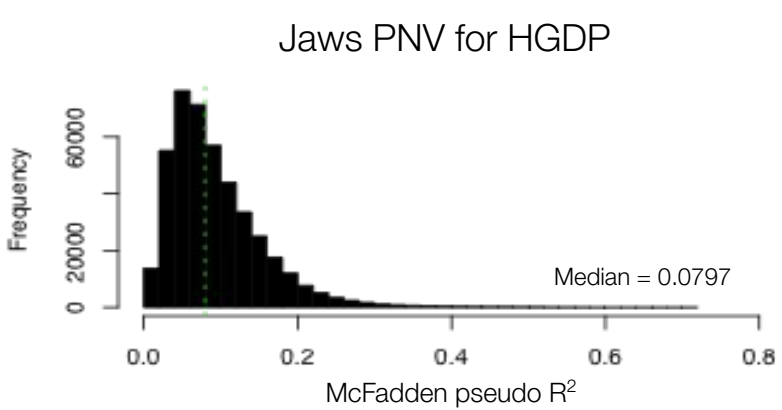
Thousand Genome Project (TGP)
m = 339100 SNPs for n = 1500 individuals

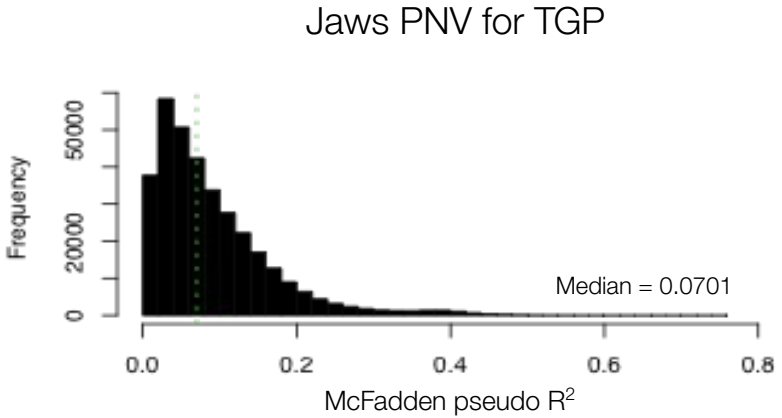


HGDP Data from Cann *et al.* (2002) and Rosenberg *et al.* (2002)

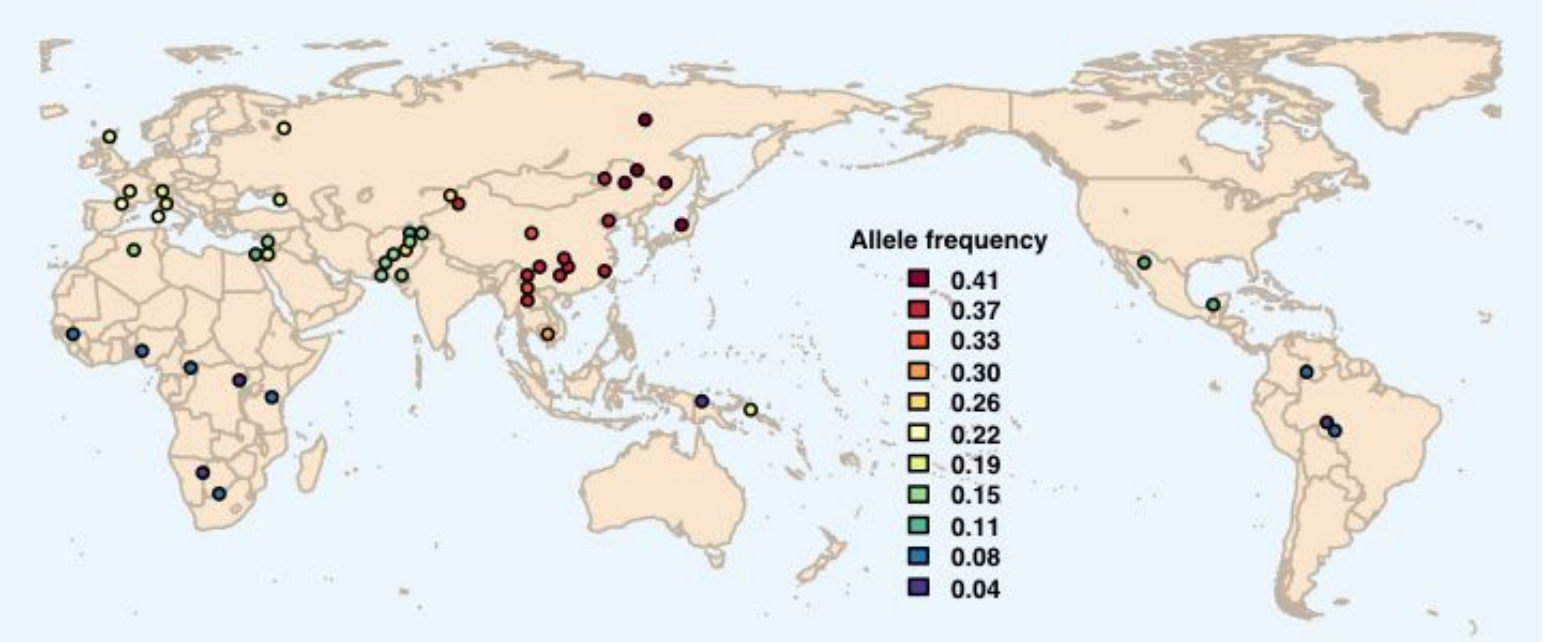
TGP Data from The 1000 Genomes Project Consortium (2012)

Pseudo R-squared measures for Population Structure



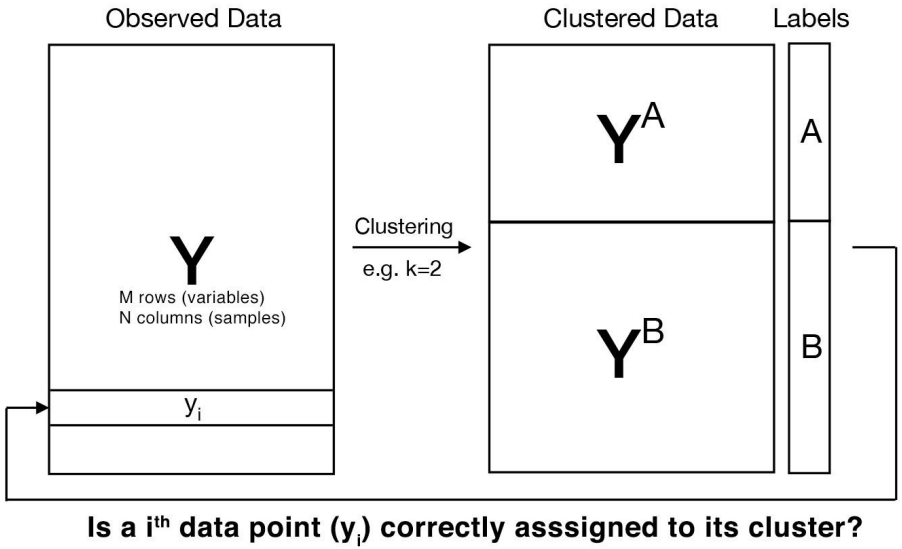


SNP with Median Differentiation



SNP ID rs2836463 in HGDP

Jackstraw for clustering



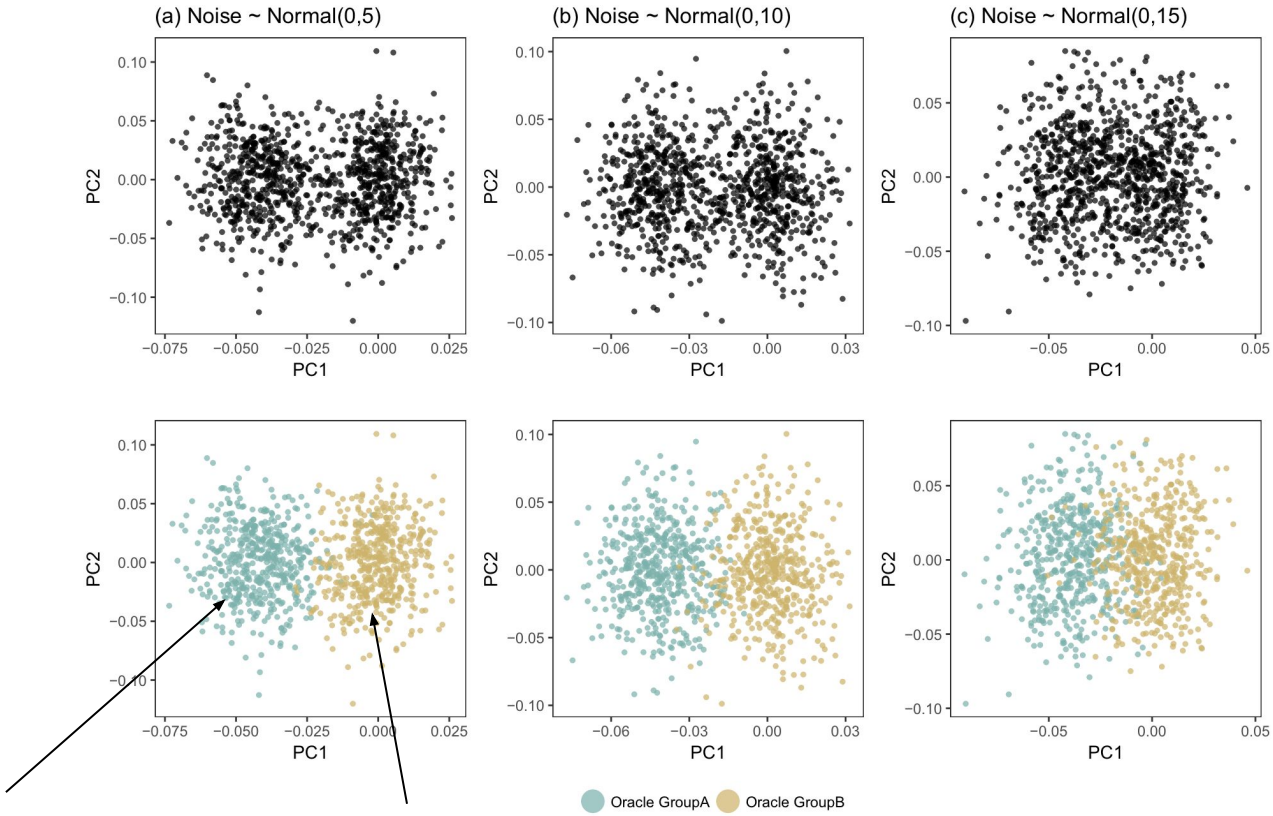
Unsupervised clustering provides “clusters” of data that are substantially distinct.

It can be seen in the context of latent variable models, with categorical latent variables

How can we evaluate whether an individual member is correctly assigned to its cluster

* in practice, already use silhouette analysis, gap statistics, or other methods to find optimal clustering

Chung (2020)



Two distinct latent groups. But as a greater level of noise is introduced, they do not get separated easily

Consider that m variables form K subpopulations.
For k=1,...,K, a mutually exclusive subset of cells (m_k out of m) are assigned to kth cluster.
Samples within the kth cluster summarized by their center (or other representative) $\mathbf{c}_k(\mathbf{Y})$ for k=1,...K.
Consider there exist unobserved centers $\mathbf{l}_k$ and coefficients $\mathbf{b}_k$ for k=1,...,K. Then, the data are modeled as:

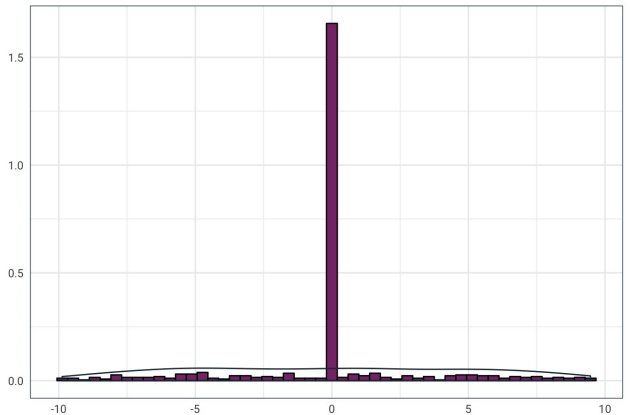
$$\mathbf{Y} = \mathbf{BL} + \mathbf{E}$$

The spike-and-slab model for introduces zero-one latent variable γ_k with initial inclusion probabilities.

$$\mathbf{b}_k = \gamma_k \boldsymbol{\beta}_k$$

$\gamma_{i,k}$ is 1 if an ith sample is associated with $\mathbf{l}_k$. Otherwise, 0.
$\boldsymbol{\beta}_k$ may take on a continuous distribution, quantifying the relationship between $\mathbf{L}$ and $\mathbf{Y}$.

e.g.

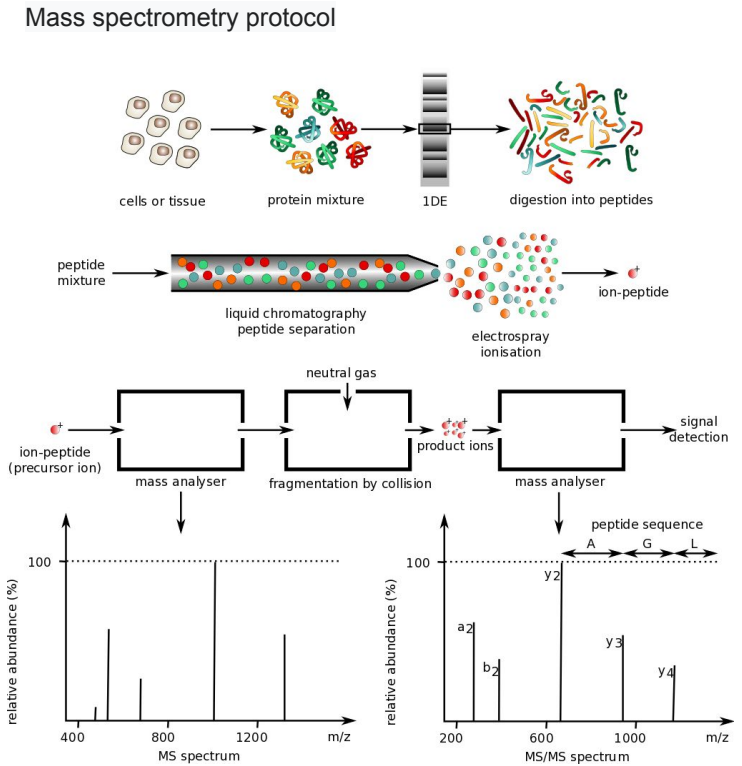


Application in proteomics

Molecular signatures in hypertrophy is hypothesized to exist. Related to clinical significance for heart remodeling and cardiovascular diseases

We are studying its manifestation on cysteine oxidative post-translational modifications (O-PTM) of proteomes

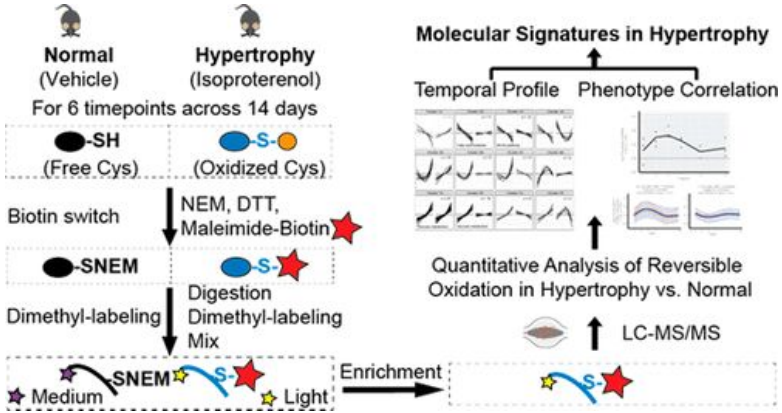
After clustering O-PTM signatures, how can we evaluate their memberships?

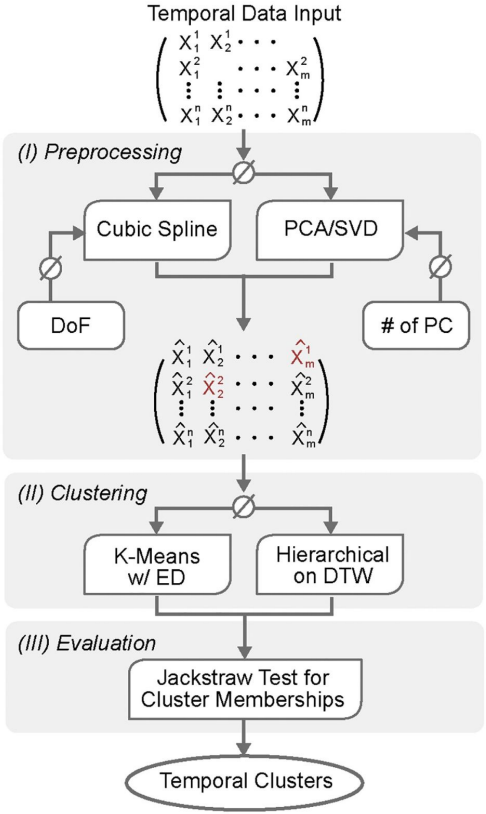


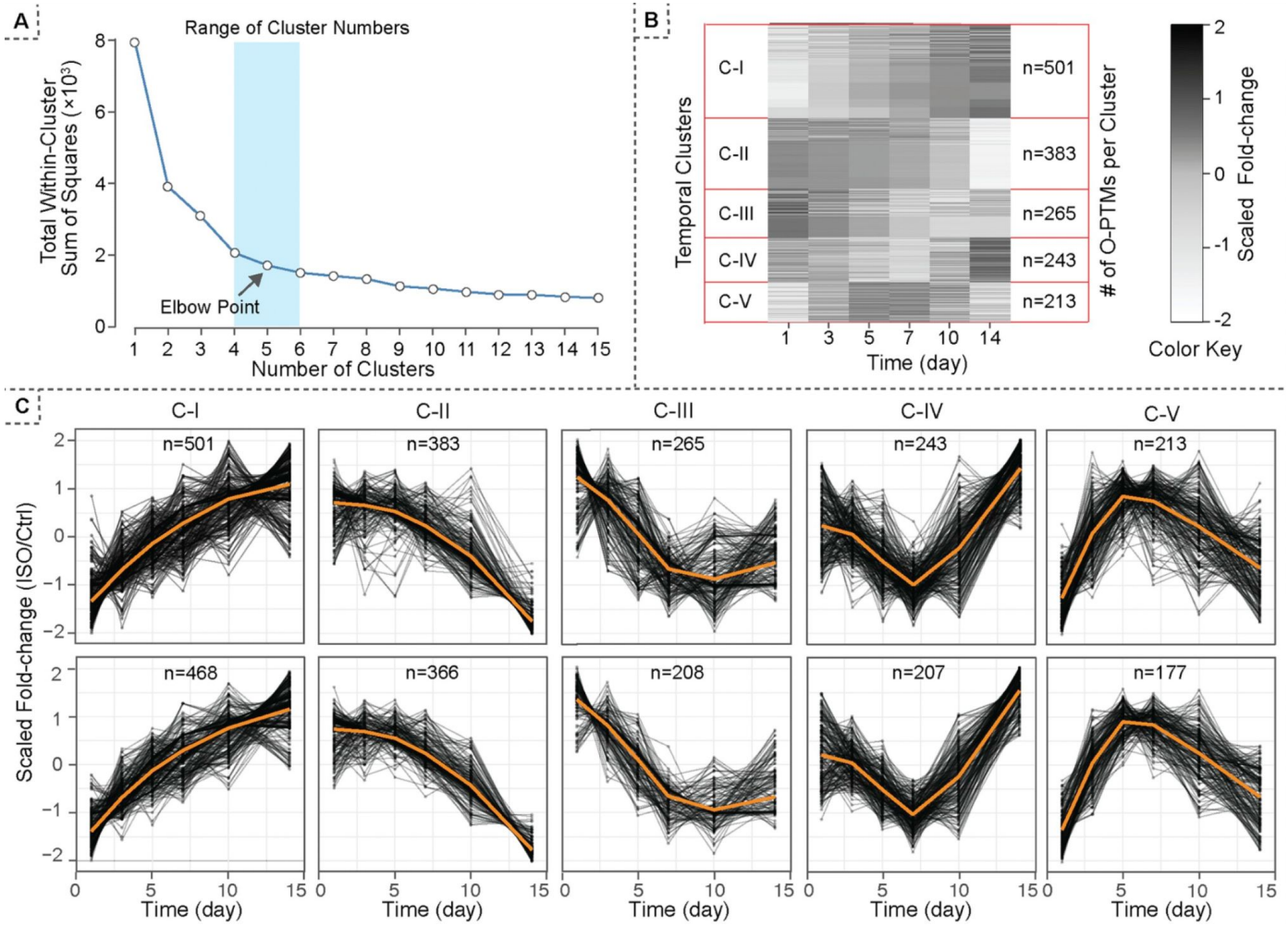
Wikipedia

Data analysis pipeline

The temporal changes of cysteine O-PTMs across the myocardial proteome were captured over time using a mouse model of cardiac hypertrophy



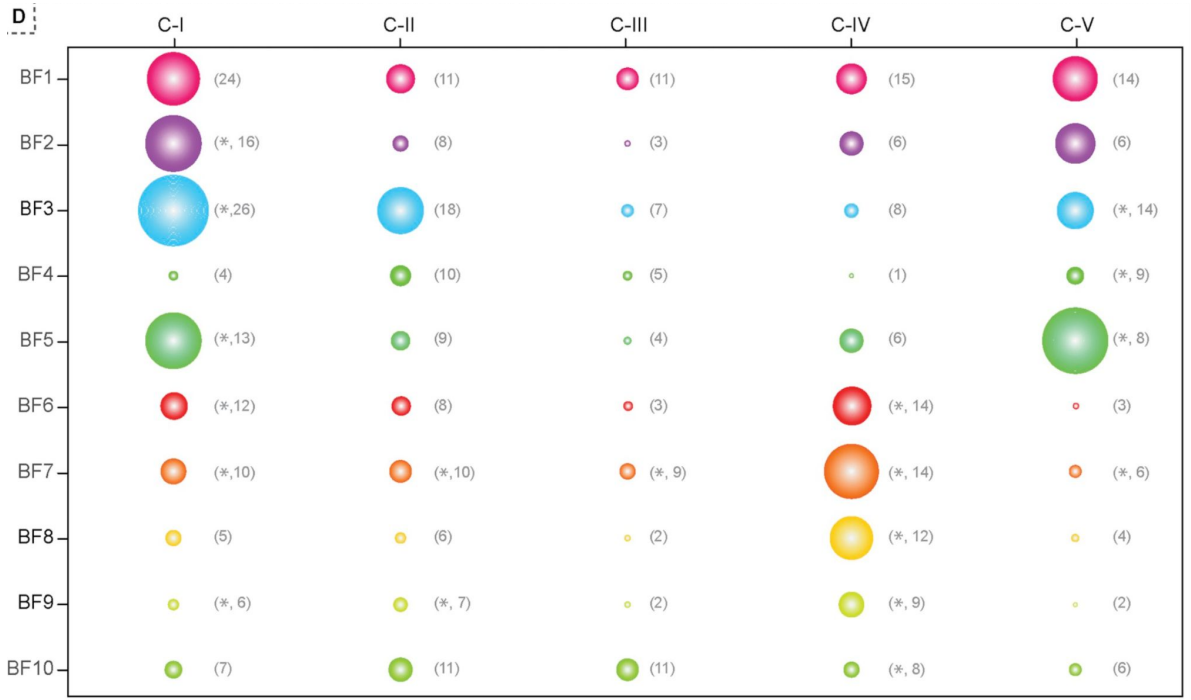




1605 O-PTMs

1426 O-PTMs with PIP > 80%

BF1, neutrophil degranulation
BF2, response to elevated platelet cytosolic Ca2+
BF3, extracellular matrix organization;
BF4, protein translation
BF5, post-translational protein phosphorylation;
BF6, glucose metabolism;
BF7, pyruvate metabolism and citric acid (TCA) cycle
BF 8, respiratory electron transport;
BF9, branched-chain amino acid (BCAA) catabolism
BF10, fatty acid metabolism.



Radius of circle: occurance of O-PTMs; (*, n): significant FDR<0.05 and number of proteins.

Protein O-PTMs of temporal significance were further annotated by their temporal patterns (as shown in five clusters) and their biological functions (BFs as shown in 10 essential pathways). Each circle represents a cluster of O-PTMs sharing both the temporal pattern and BF attribute. The occurrences of O-PTMs (a radius of a circle), the false discovery rate (*, FDR < 0.05), and the number of proteins (n) are labelled for each O-PTM cluster.