Testing Sets of Genomic Features aka Pathways

Mitra Ebrahimpoor

Joint work with Jelle Goeman

Data integration workshop, Oct-2019

Medical Statistics,
Leiden University Medical Center

Leiden University Medical Center

Background

Where we start...

The typical result of an experiment:

A list of deferentially expressed features.

The typical result of an experiment:

A list of deferentially expressed features.

Accession Number	Description	Fold change AB/DC	p-value
IPI00791534	Solute carrier family 4, anion exchanger member 1	49.91	0.033
IPI00022418	Isoform 1 of fibronectin	33.71	0.006
IPI00020101	Histone H2B type 1-C/E/F/G/I	5.47	0.013
IPI00026314	Isoform 1 of gelsolin	5.33	0.048
IPI00418431	ASPN protein	5.21	0.009
IPI00021405	Isoform A of prelamin-A/C	5.19	0.009
IPI00009802	Isoform V0 of versican core protein	5.07	0.003
IPI00418169	Isoform 2 of annexin A2	4.98	0.0003
IPI00453473	Histone H4	4.46	0.024
IPI00022200	Isoform 1 of collagen alpha-3(VI) chain	1.25	0.036
IPI00410241	Periostin isoform thy6	- 1.67	0.047
IPI00218585	Isoform 2 of Periostin	- 1.71	0.034
IPI00006114	Pigment epithelium-derived factor	- 1.71	0.044
IPI00007960	Isoform 1 of periostin	- 1.83	0.012

(Salmon et al., Journal of Proteomics:2013)

Where we start...

The typical result of an experiment:

A list of deferentially expressed features.

Biological theory is about groups of features that are ...

–Involved in a biological process

–Involved in a series of chemical reaction

$\rightarrow$ Biological theory is about sets of features.

Advantages

- Easier to interpret in terms of biology
- Has more statistical power

  Ref: [Subramanian et al., 2005] & [Khatri et al., 2012]

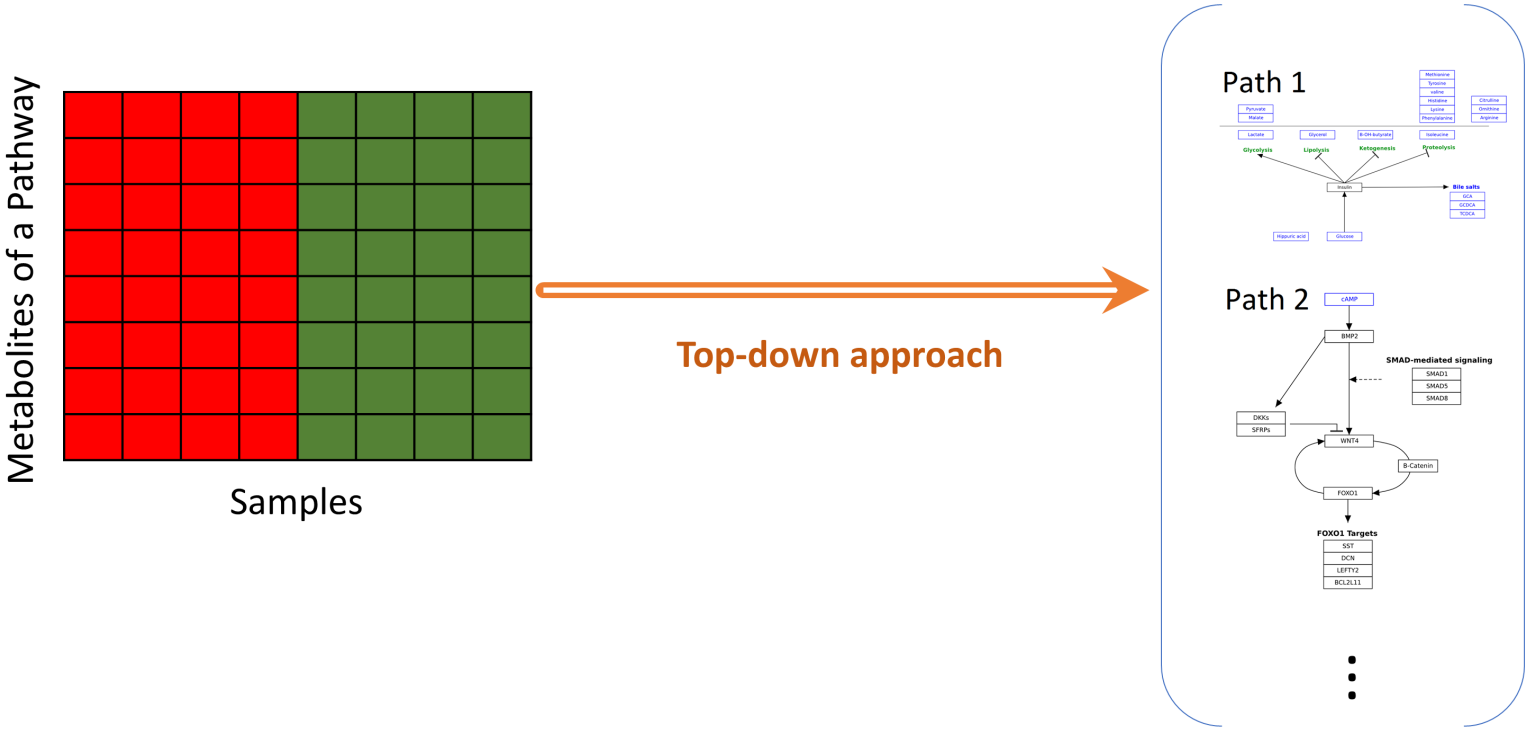
Existing Methods

Testing Pathways : Approach

Two approaches

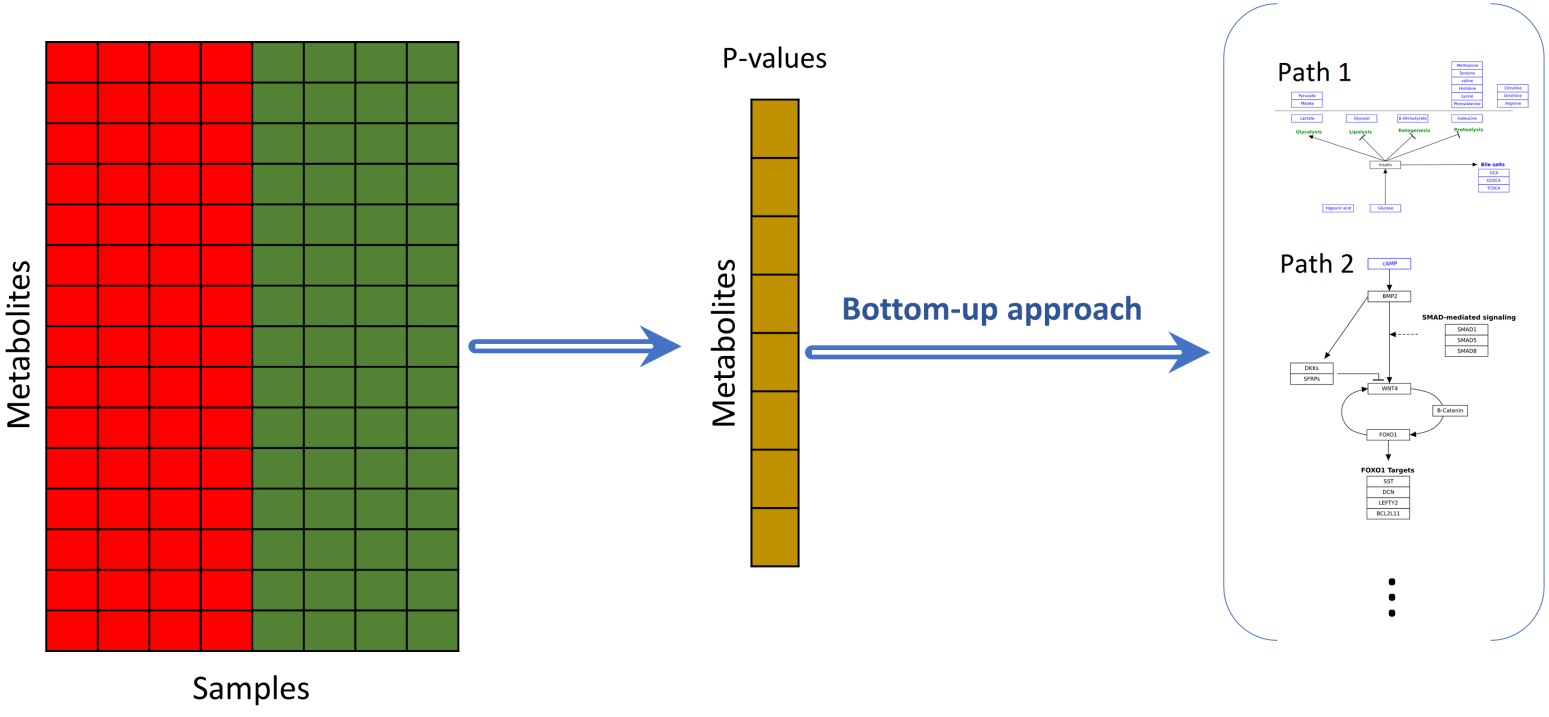
Two approaches

- Top-down: Model the metabolites jointly at the pathway level e.g. Globaltest



Two approaches

- Bottom-up: Test the metabolites separately and combine the p-values e.g. Fisher's exact test



SEA follows a Bottomn-up approach.

Notation

- m features
- 2^m possible feature-sets identified by S

There are many sources to define S.

Notation

- m features
- 2^m possible feature-sets identified by S

There are many sources to define S.

Notation

- m features
- 2^m possible feature-sets identified by S

There are many sources to define S.

Notation

- Subset $\mathbb{T}$ are the truly active features (A)
- $A(S) = |S \cap \mathbb{T}|$
- $\pi(S) = A(S)/|S|$

Notation

- Subset $\mathbb{T}$ are the truly active features (A)
- $A(S) = |S \cap \mathbb{T}|$
- $\pi(S) = A(S)/|S|$

Null Hypothesis of Interest for testing feature-set S

– self-contained null hypothesis : Are there *any* active features in the set?

$$H_0^{SC}(S) : \pi(S) = 0$$

Notation

- Subset $\mathbb{T}$ are the truly active features (A)
- $A(S) = |S \cap \mathbb{T}|$
- $\pi(S) = A(S)/|S|$

Null Hypothesis of Interest for testing feature-set S

– self-contained null hypothesis : Are there *any* active features in the set?

$$H_0^{SC}(S) : \pi(S) = 0$$

– competitive null hypothesis: Are the features in the selected set at most as active as the *features in the background*?

$$H_0^{Comp}(S) : \pi(S) \leq \pi(\mathbb{S})$$

Notation

- Subset $\mathbb{T}$ are the truly active features (A)
- $A(S) = |S \cap \mathbb{T}|$
- $\pi(S) = A(S)/|S|$

Null Hypothesis of Interest for testing feature-set S

– self-contained null hypothesis : Are there *any* active features in the set?

$$H_0^{SC}(S) : \pi(S) = 0$$

– competitive null hypothesis: Are the features in the selected set at most as active as the *features in the background*?

$$H_0^{Comp}(S) : \pi(S) \leq \pi(\mathbb{S})$$

"Same results if there are few active features."

Self-contained Method

Are there any active features in the selected set?

Self-contained Method

Are there any active features in the selected set?

- Example: Global test, PLAGE, FORGE

Are there any active features in the selected set?

- Example: Global test, PLAGE, FORGE

Global test : example

- Association with response y
- Measured features $=$ a set of covariates $(x_1, \ldots, x_p)$
- Clinical covariates: $(z_1, \ldots, z_m)$
- Linear predictor: $\alpha + \sum_{j=1}^{m} z_i \gamma_j + \sum_{i=1}^{p} x_i \beta_i$
- Null hypothesis:

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$

Competitive Method

Are the features in the selected set at most as active as the background?

Competitive Method

Are the features in the selected set at most as active as the background?

- Example: Fisher's exact test, GSA, SAFE, GSEA

Are the features in the selected set at most as active as the background?

- Example: Fisher's exact test, GSA, SAFE, GSEA

Fisher's exact test: example
20000 features; 100 feature-sets; 200 Sig. features

Could arise due to chance: $p = 0.26$

	Sig. features	*non-Sig. features*	*total*
in feature-set	2	98	100
not in feature-set	198	19702	19900
total	200	19800	20000

Could not arise due to chance: $p = 0.0005$

	Sig. features	*non-Sig. features*	*total*
in feature-set	6	94	100
not in feature-set	194	21706	19900
total	200	19800	20000

Self-contained or Competitive?

Self-contained

- lower specificity

Self-contained or Competitive?

Self-contained

- lower specificity

Competitive

- Seems biologically more relevant
- Not over-powered
- Permutes the features

Where we end up

Current methods

Current methods

- No method tests the competitive null in fact

Current methods

- No method tests the competitive null in fact

  Ref: [Debrabant, B. 2017] & [Maciejewski, H. 2014]

- Competitive methods depend on the independence among features

Current methods

- No method tests the competitive null in fact

  Ref: [Debrabant, B. 2017] & [Maciejewski, H. 2014]

- Competitive methods depend on the independence among features

  Ref: [Goeman, J. 2007]

Current methods

- No method tests the competitive null in fact

  Ref: [Debrabant, B. 2017] & [Maciejewski, H. 2014]

- Competitive methods depend on the independence among features

  Ref: [Goeman, J. 2007]

- All methods depend on the choice of database and feature-set

Current methods

- No method tests the competitive null in fact

  Ref: [Debrabant, B. 2017] & [Maciejewski, H. 2014]

- Competitive methods depend on the independence among features

  Ref: [Goeman, J. 2007]

- All methods depend on the choice of database and feature-set

- All methods need multiple comparisons corrections with multiple feature-sets

Simultaneous Enrichment Analysis (SEA)

Unified null hypothesis

$$H_0^U(S, c) : \pi(S) \leq c$$

Unified null hypothesis

$$H_0^U(S, c) : \pi(S) \leq c$$

Self-contained

$$H_0^{self}(S) : \pi(S) = 0$$

The Unified null hypothesis

Unified null hypothesis

$$H_0^U(S, c) : \pi(S) \leq c$$

Self-contained

$$H_0^{self}(S) : \pi(S) = 0$$

Competitive

$$H_0^{Comp}(S) : \pi(S) \leq \pi(\mathbb{S})$$

Hommel (1988)

The set S of hypotheses is rejected by Simes if and only if:

$$\min_i \frac{|S|}{i} p_{(i:S)} < \alpha$$

Assumption:

Certain form of dependence among features

Same assumption required for using BH

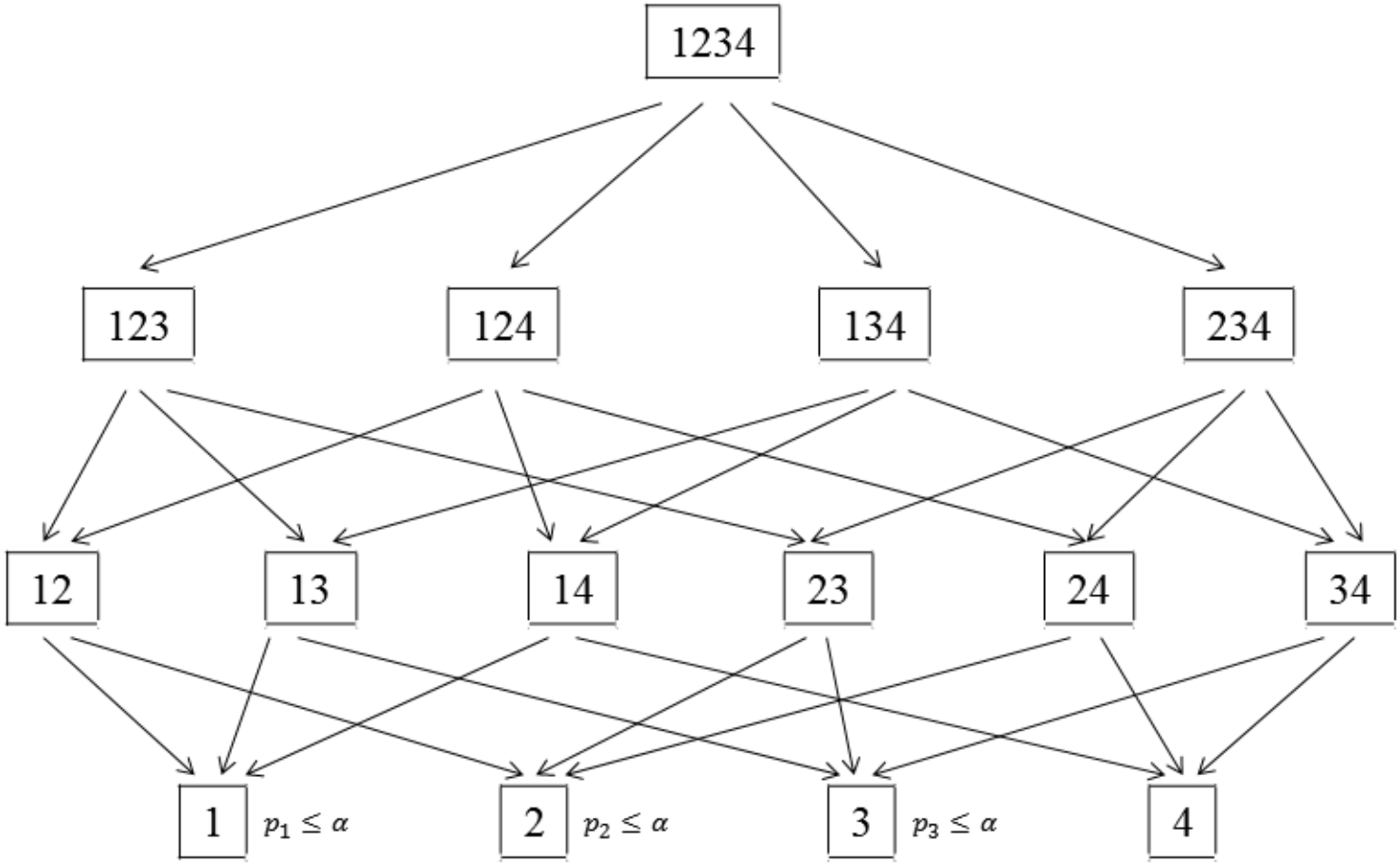


Figure 1: Closed Testig Example: $\alpha/2 < p1 \leq p2 \leq p3 \leq 2\alpha/3$, and $p4 > \alpha$

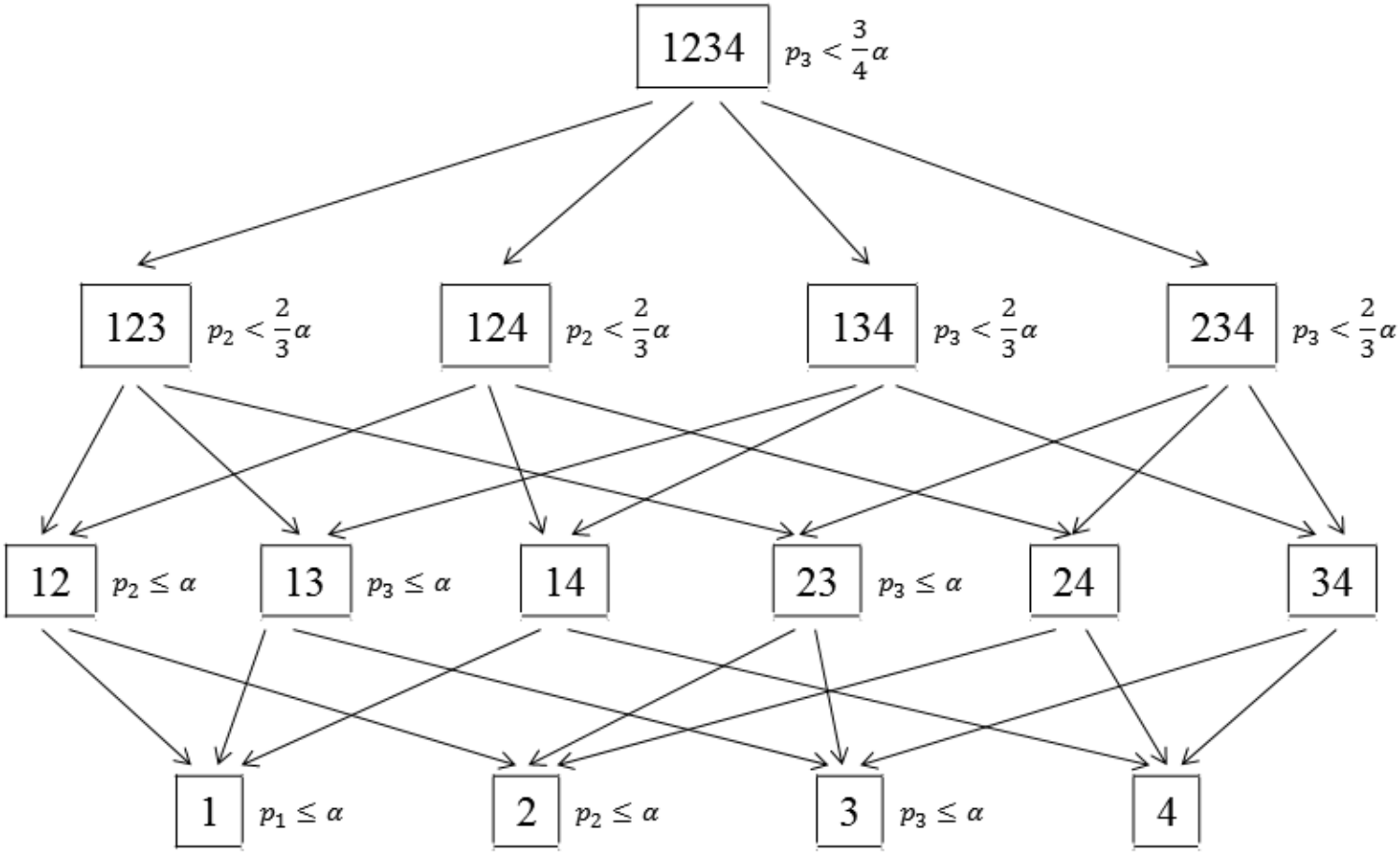


Figure 2: Closed Testig Example: $\alpha/2 < p1 \leq p2 \leq p3 \leq 2\alpha/3$, and $p4 > \alpha$

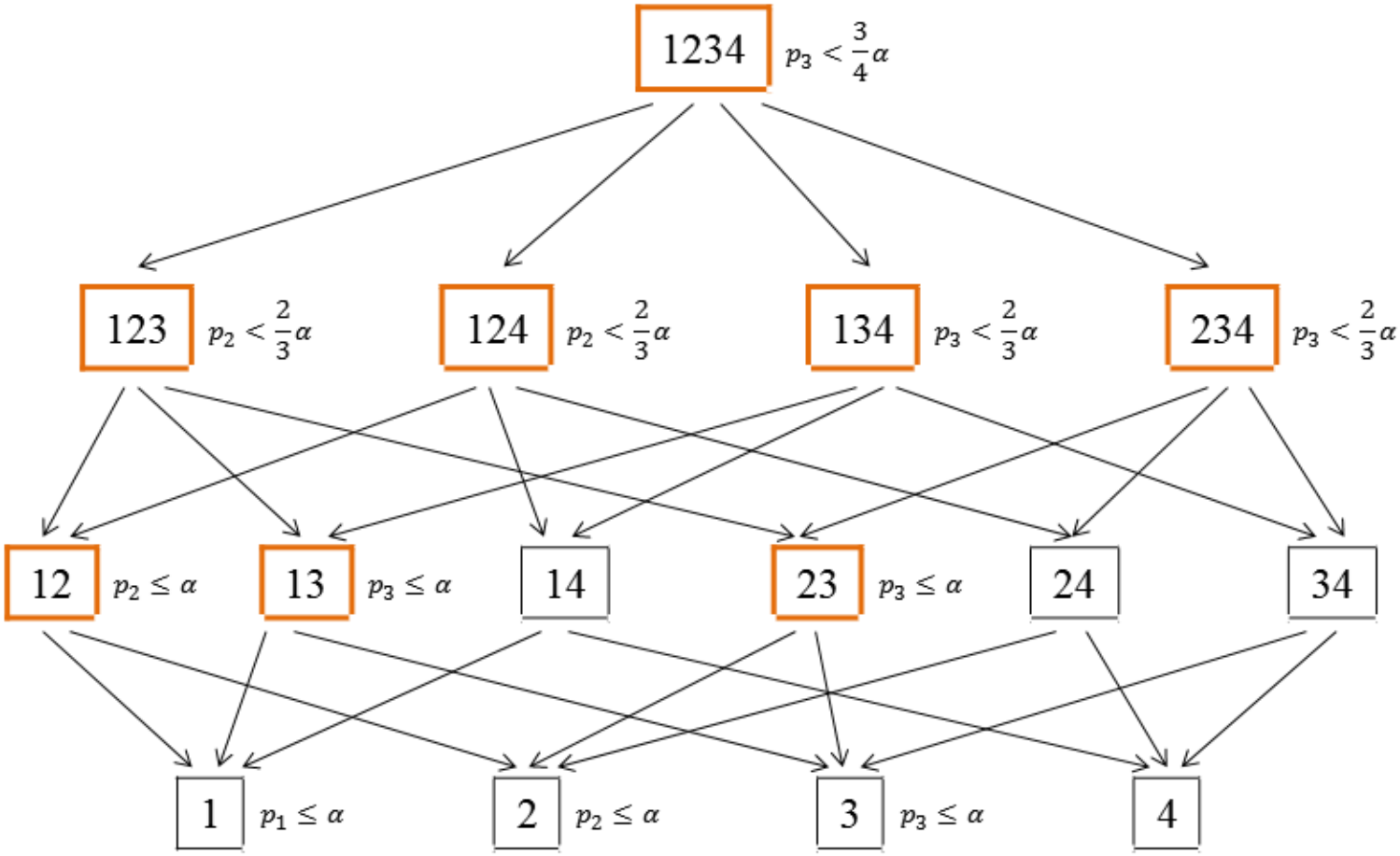


Figure 3: Closed Testig Example: $\alpha/2 < p1 \leq p2 \leq p3 \leq 2\alpha/3$, and $p4 > \alpha$

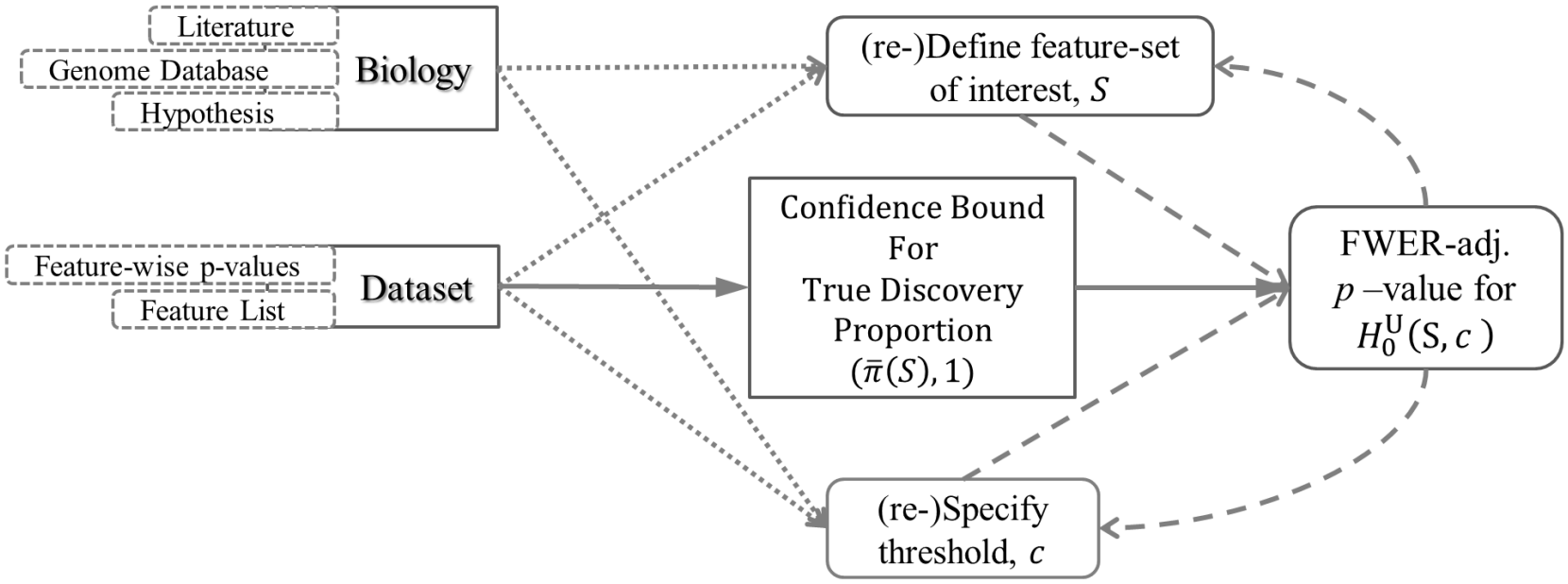
Goeman et al. (2016)

$$P\left(\begin{array}{c}\hat{\pi}(S_1) \leq \pi(S_1)\\ \hat{\pi}(S_2) \leq \pi(S_2)\\ \cdot\\ \cdot\\ \cdot\\ \hat{\pi}(S_{2^m}) \leq \pi(S_{2^m})\end{array}\right) \geq 1-\alpha$$

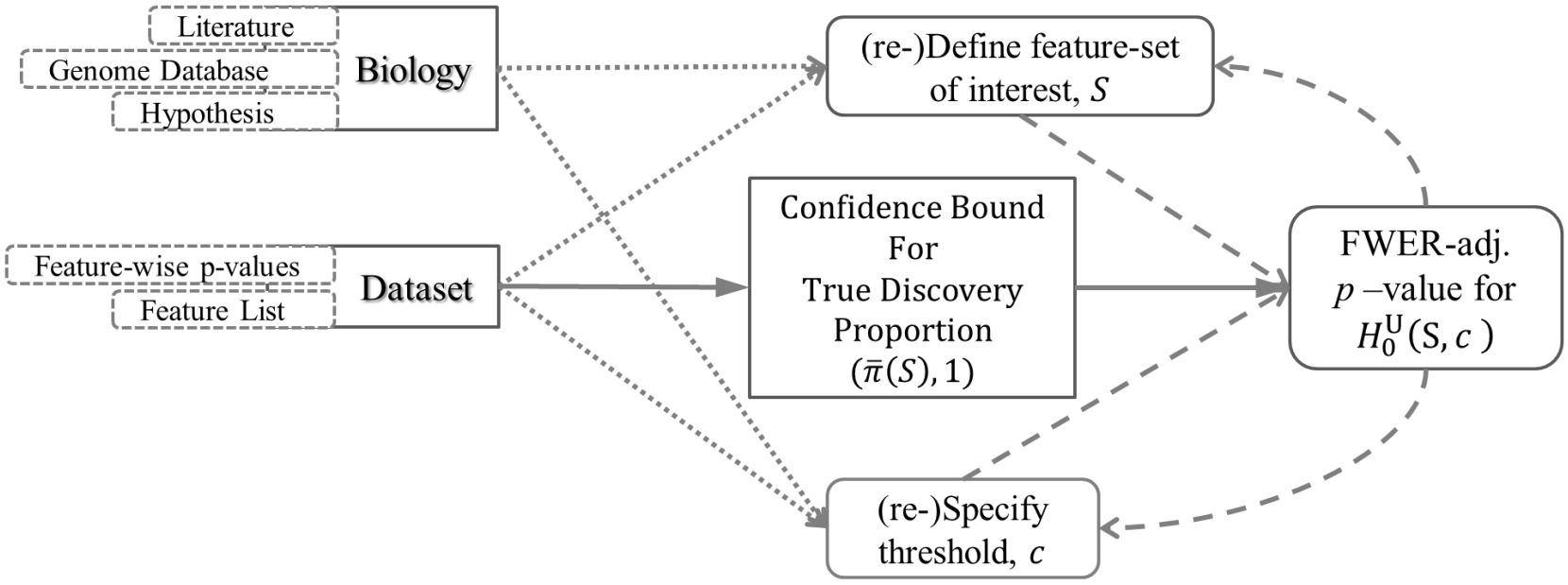
- Overall FWER control
- Simultaneity over all S:

  Choice of S after analysis

  S can be revised as many times as desired


Literature
Genome Database
Hypothesis
Biology
Feature-wise p-values
Feature List
Dataset
(re-)Define feature-set of interest, S
Confidence Bound For True Discovery Proportion $(\bar{\pi}(S), 1)$
(re-)Specify threshold, c
FWER-adj. p –value for $H_0^U(S, c)$

SEA pipeline



- Get raw p-values per feature
- Choose S and c
- Build a confidence interval for $\bar{\pi}(S)$
- Check if the CI includes c
- Calculate the adjusted p-value
- Modify S and/or c as needed

Discussion: Properties of SEA

- Valid competitive testing
- Combined feature-wise and feature-set testing
- Flexibility $\times$ 3
- Short computation time
- Power comparable to classical feature-set testing

Discussion: Properties of SEA

- Valid competitive testing
- Combined feature-wise and feature-set testing
- Flexibility $\times$ 3
- Short computation time
- Power comparable to classical feature-set testing

- An exploratory tool

That is all!

Any questions?