Selective Inference in Genetics

Chiara Sabatti

Biomedical Data Science and Statistics

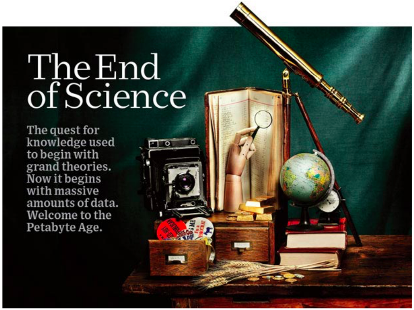
Luminy, July 2017

Acknowledgements

- Yoav Benjamini
- **Malgorzata Bogdan**
- Marina Bogomolov
- Damian Brzyski
- Emmanuel Candes
- Eugene Katsevich
- Christine Peterson
- Snigdha Panigrahi
- David Siegmund
- Laurel Stell
- Piotr Sobczyk
- Asaf Weinstein
- Junjie Zhu

- GTEx consortium
- FinMedSeq consortium
- Funding from NIH

Statistics in the era of big data


The End of Science
The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. Welcome to the Petabyte Age.

Wired Magazine, issue 16.07

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

This is the religion of big data. As a believer, you see ethics and laws in a different light than the non-believers. You also believe that you are part of a new scientific movement which does away with annoying things such as making hypotheses and the assumptions behind traditional statistical techniques. ***No need to ask questions, just collect lots of data and let it speak*****.**

Gil Press,
Forbes, September 2014

A specific view-point: gene mapping

Which genetic variants influence phenotypes of interest?

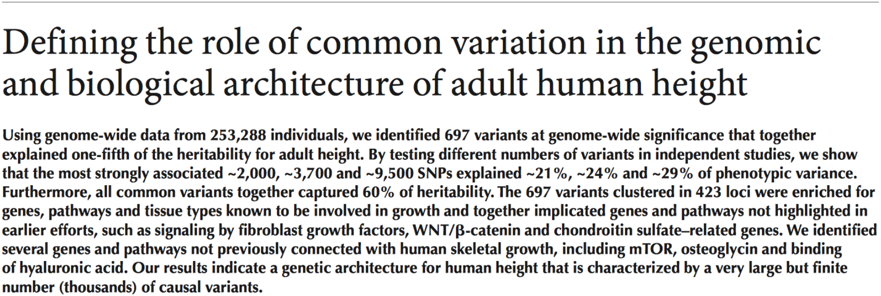
- **Genes** identify a protein: they tell us what is the **biological pathway** involved in disease and proteins can be useful **drug targets**
- Knowing which variants are important makes it possible to provide **genetic counseling**
- Understanding which **groups of phenotypes** are influenced by the same variants help us to refine our diagnostic tools and gives us a handle on mechanism.

Looking across the entire genome

This was never hypothesis driven research. The space of hypotheses tested has been defined by our ability to probe genetic variation.

- Even when the locations in the genome one was able to probe where only a handful (different protein types, as blood groups), geneticists decided that an association had to have a p-value $< 10^{-4}$ to be significant
- When genome-scans became a possibility, we looked at monogenic disease, and aimed to control FWER, the probability of making at least one false discovery.
- The landscape has changed as we start looking at a large number of phenotypes, many of which might be complex.

An example of polygenic trait: height



Defining the role of common variation in the genomic and biological architecture of adult human height

Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/β-catenin and chondroitin sulfate–related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

Nature Genetics, **46**: 1173–1186, 2014

A study example: Genotype-Tissue Expression (GTEx)

TRAITS:

- Expression is measured for 20,000-30,000 genes
- In 44 tissues

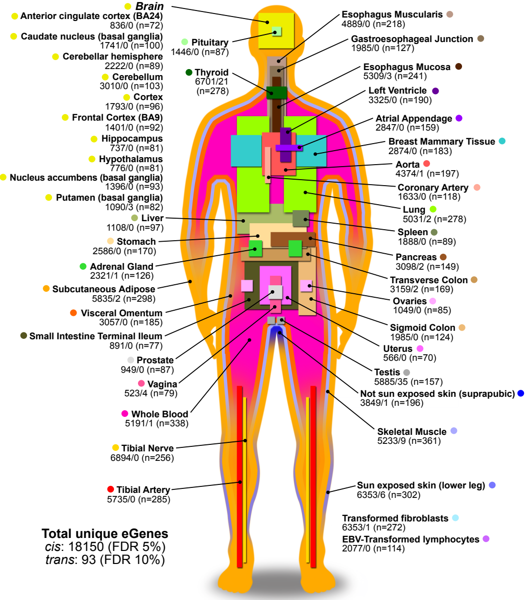
GENOTYPES:

- ≈ 11 million Single Nucleotide Polymorphisms (SNPs)

GOALS:

- **Identify the loci where genetic variation influences the expression levels of genes (eQTL) in one tissue**
- eGene: a gene whose expression appears to be genetically regulated
- Understand when this regulation is **conserved across tissues**
- Understand which **specific variants** underly this regulation (eSNPs)

A study example: Genotype-Tissue Expression (GTEx)



● ***Brain***

● **Anterior cingulate cortex (BA24)** 836/0 (n=72)

● **Caudate nucleus (basal ganglia)** 1741/0 (n=100)

● **Cerebellar hemisphere** 2222/0 (n=89)

● **Cerebellum** 3010/0 (n=103)

● **Cortex** 1793/0 (n=96)

● **Frontal Cortex (BA9)** 1401/0 (n=92)

● **Hippocampus** 737/0 (n=81)

● **Hypothalamus** 776/0 (n=81)

● **Nucleus accumbens (basal ganglia)** 1396/0 (n=93)

● **Putamen (basal ganglia)** 1090/3 (n=82)

● **Liver** 1108/0 (n=97)

● **Stomach** 2586/0 (n=170)

● **Adrenal Gland** 2321/1 (n=126)

● **Subcutaneous Adipose** 5835/2 (n=298)

● **Visceral Omentum** 3057/0 (n=185)

● **Small Intestine Terminal Ileum** 891/0 (n=77)

● **Prostate** 949/0 (n=87)

● **Vagina** 523/4 (n=79)

● **Whole Blood** 5191/1 (n=338)

● **Tibial Nerve** 6894/0 (n=256)

● **Tibial Artery** 5735/0 (n=285)

● **Pituitary** 1446/0 (n=87)

● **Thyroid** 6701/21 (n=278)

Esophagus Muscularis ● 4889/0 (n=218)

Gastroesophageal Junction ● 1985/0 (n=127)

Esophagus Mucosa ● 5309/3 (n=241)

Left Ventricle ● 3325/0 (n=190)

Atrial Appendage ● 2847/0 (n=159)

Breast Mammary Tissue ● 2874/0 (n=183)

Aorta ● 4374/1 (n=197)

Coronary Artery ● 1633/0 (n=118)

Lung ● 5031/2 (n=278)

Spleen ● 1888/0 (n=89)

Pancreas ● 3098/2 (n=149)

Transverse Colon ● 3159/2 (n=169)

Ovaries ● 1049/0 (n=85)

Sigmoid Colon ● 1985/0 (n=124)

Uterus ● 566/0 (n=70)

Testis ● 5885/35 (n=157)

Not sun exposed skin (suprapubic) ● 3849/1 (n=196)

Skeletal Muscle ● 5233/9 (n=361)

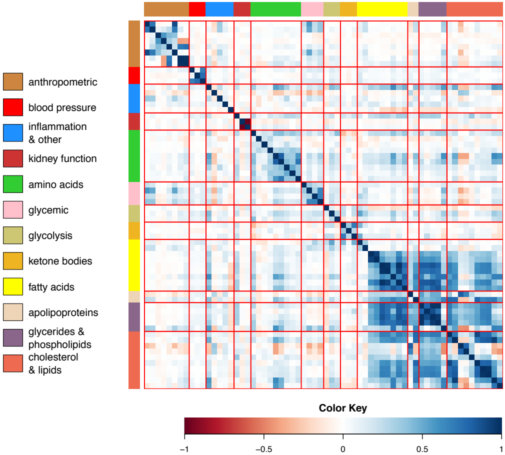
Sun exposed skin (lower leg) ● 6353/6 (n=302)

Transformed fibroblasts ● 6353/1 (n=272)

EBV-Transformed lymphocytes ● 2077/0 (n=114)

Total unique eGenes
cis: 18150 (FDR 5%)
trans: 93 (FDR 10%)

Another study example: the genetics of metabolites



- Exome resequencing
- ≈ 20,000 subjects of Finnish descent
- Detail phenotypes defined with Nuclear Magnetic Resonance (NMR)
- Which genes are important for metabolic syndrome?
- Which variants in these genes?
- Which groups of phenotypes are influenced by the same variants?

Discoveries and FDR

In studies with many expected discoveries, it is natural to want to control FDR

Discoveries and FDR

In studies with many expected discoveries, it is natural to want to control FDR

$\Longrightarrow$ There are many possible discoveries

- **Genetic loci** that are relevant for one phenotype
- **Genes** that influence one phenotype
- **Genetic variants** that influence one phenotype
- Any of the above resolutions, for **groups of phenotypes**

Discoveries and FDR

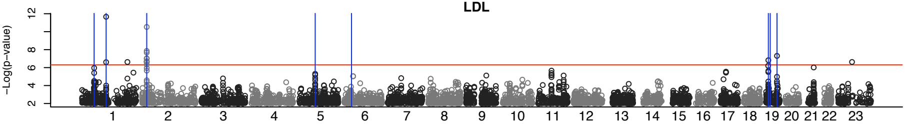
In studies with many expected discoveries, it is natural to want to control FDR

$\Longrightarrow$ There are many possible discoveries

- **Genetic loci** that are relevant for one phenotype
- **Genes** that influence one phenotype
- **Genetic variants** that influence one phenotype
- Any of the above resolutions, for **groups of phenotypes**

$\Longrightarrow$ As controlling FDR is controlling a property on average over the selected, we need to make sure that we pay attention to what the **selections** are.

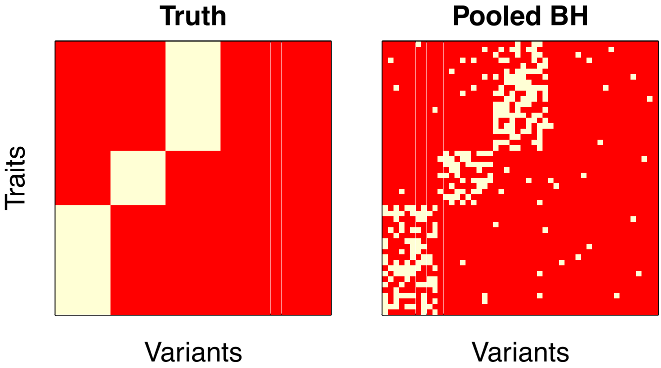
Ex 1. Discovering loci is not the same as discovering variants



A possibly large number of correct rejections at some location can inflate the denominator in the definition of false discovery rate, hence artificially creating a small false discovery rate, and lowering the barrier to possible false detections at distant locations
Siegmund, Yakir and Zhang (2011)

See also Perone et al. (2004) and Benjamini and Heller (2007).

Ex 2. variants with effects on some traits



Controlling FDR of trait× SNPs discoveries, does not result in controlling the FDR of SNP discoveries

See also Foygel Barber & Ramdas (2015)

Practice in eQTL (*cis*): selection

- The analysis is based on tests for association between the expression of each gene (in each tissue) and each of the genotyped variants

- Reporting the results of such tests is too lengthy, so focus is on aggregate discoveries: eGenes, eSNPs

- FDR is the criteria of choice, but it has been noted that applying FDR controlling procedures to each of the gene×SNP hypotheses leads to an inflation of scientifically interesting discoveries.

- Analysis is carried out in multiple steps
  1. First **eGenes** are identified aggregating signals from all SNP, and controlling FDR of eGenes.
  2. One attempts to identify exactly which variants matter in the neighborhood of the gene (fine mapping), using model selection techniques coupled with some stringent criteria to guarantee consistency of results.
     Ex. The least significant variant added to the model for **one gene** has to have a p-value smaller than that of the least significant eGene **across the entire genome**.

Genetics as a playground for selective inference

- See yesterday survey by Benjamini
- Multiple talks here and work by others

Two stories about FDR control in the presence of structure

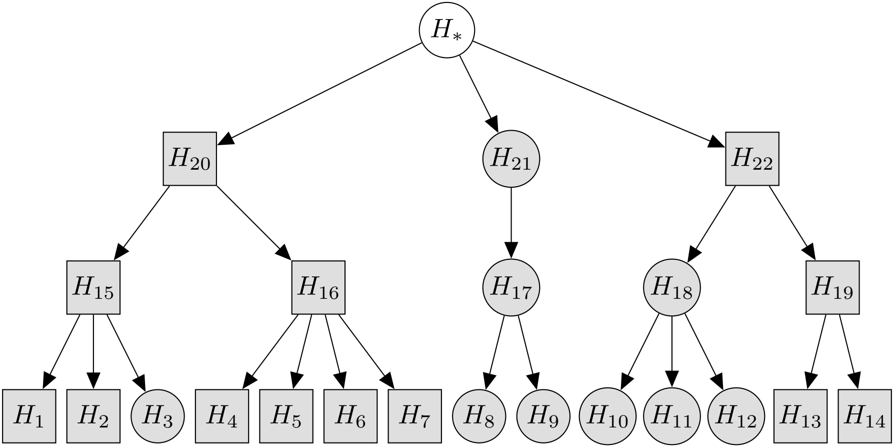
Testing hypotheses on a tree

Bogomolov, Peterson, Benjamini and S. (2017) arXiv:1705.07529

Controlled variable selection at multiple resolutions

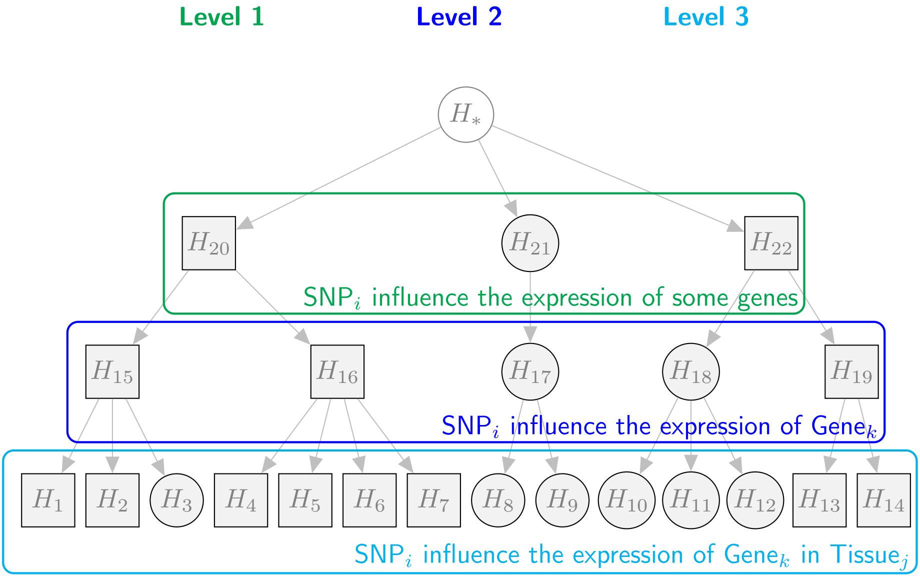
Katsevich and S. (2017) arXiv:1706.09375

A tree of families of hypotheses



- Each hypothesis i at level ℓ is parent to a family of hypotheses $\mathcal{F}_i^{\ell+1}$ at level $\ell + 1$
- ○ = true null hypotheses □ = false null
- Parent hypotheses are true if the intersection of descendant hypotheses is true

A tree of hypotheses — GTEx example



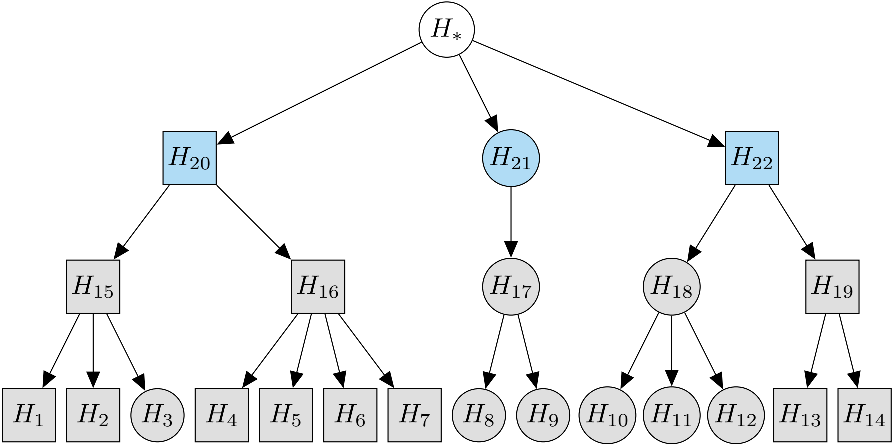
P-values for hypotheses on the tree

- Assume p-values p_i for H_i in highest level L (finest scale).
- P-values for hypotheses in lower levels $\ell < L$ can be obtained with any valid rule. In particular, they can be obtained combining the p-values of children.
- Ex. Simes' rule.
  Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(|\mathcal{F}_i^{\ell+1}|)}$ be the ordered p-values for the hypotheses in the family $\mathcal{F}_i^{\ell+1}$ indexed by H_i

$$p_i = \min_t p_{(t)} \times \frac{|\mathcal{F}_i^{\ell+1}|}{t}$$

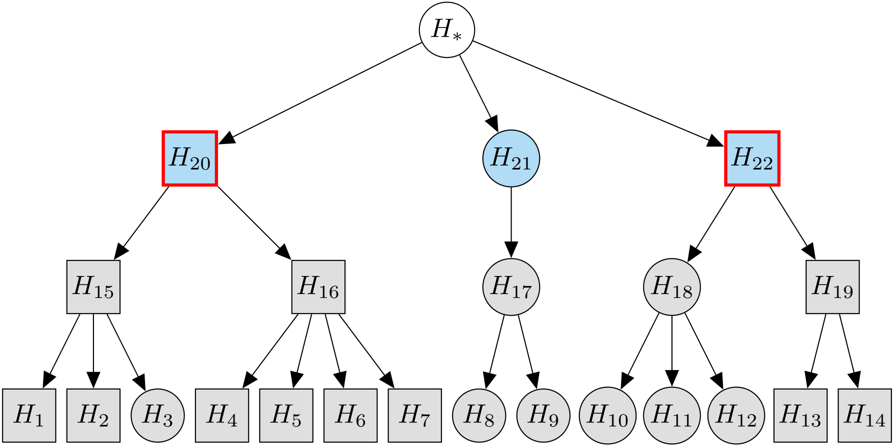
Hierarchical testing – Level 1

(see Yekutieli, 2008)



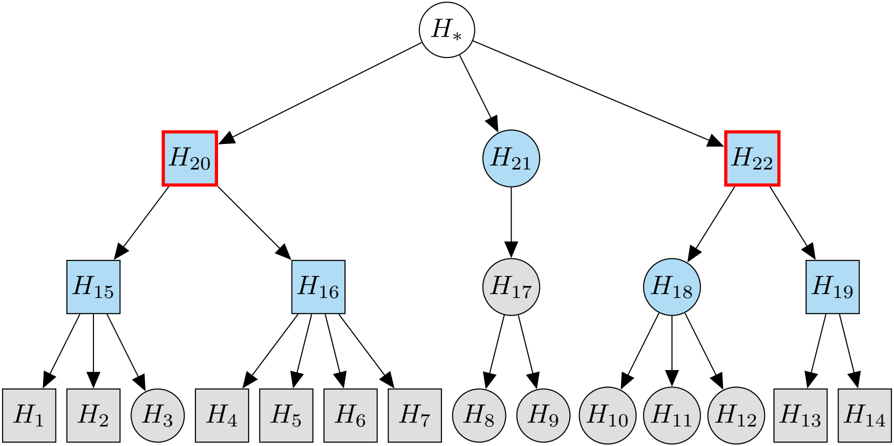
Hierarchical testing – Level 1, rejections

(see Yekutieli, 2008)



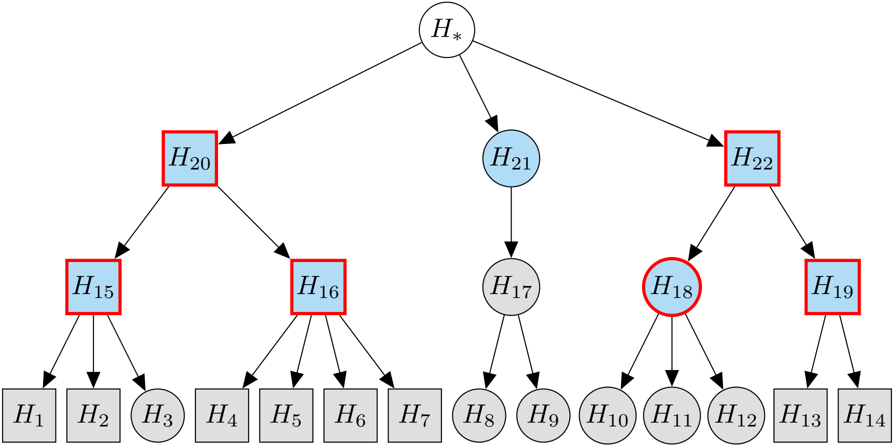
Hierarchical testing – Level 2

(see Yekutieli, 2008)



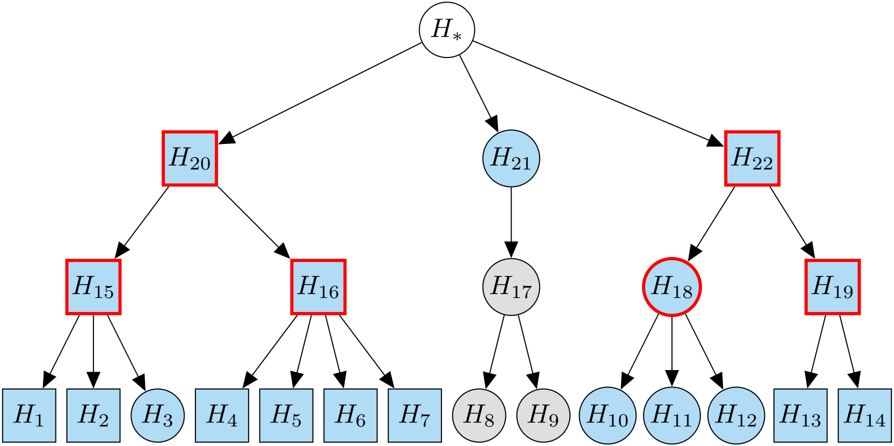
Hierarchical testing – Level 2, rejections

(see Yekutieli, 2008)



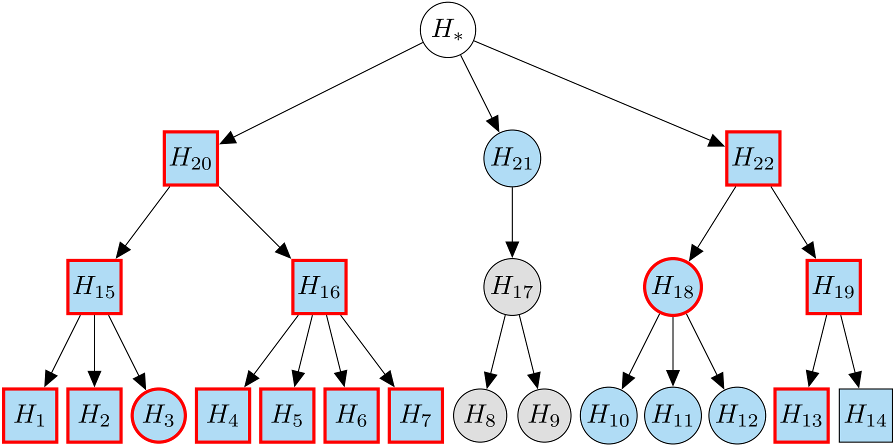
Hierarchical testing – Level 3

(see Yekutieli, 2008)

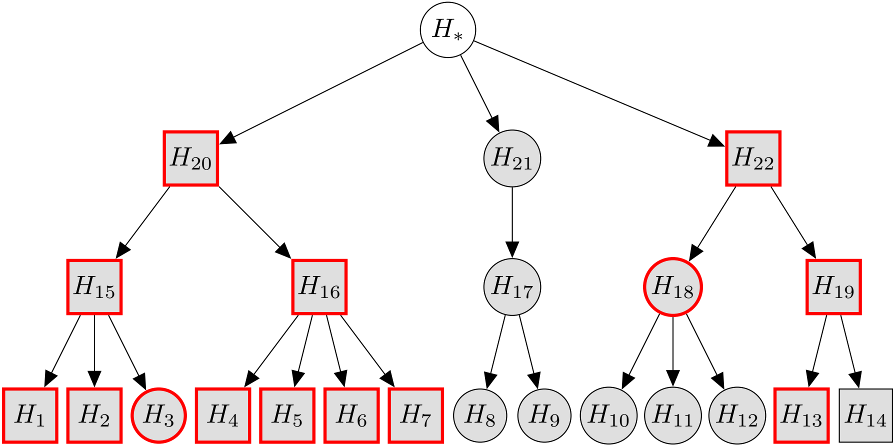


Hierarchical testing – Level 3, rejections

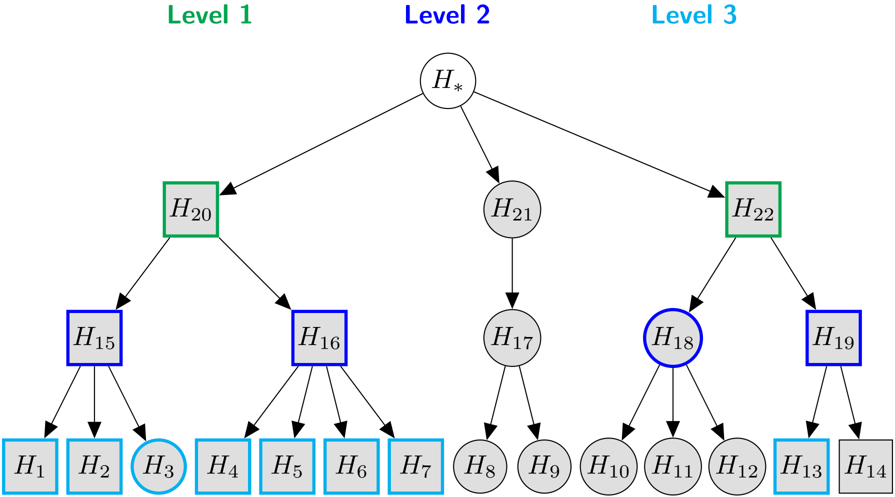
(see Yekutieli, 2008)



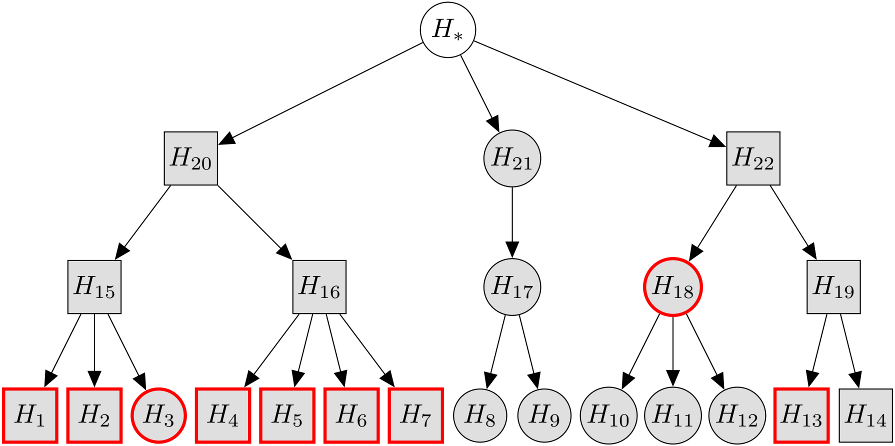
All discoveries



Level specific discoveries



Outer node discoveries



Notions of global error

Yekutieli (2008)

- Tree FDR
- Level specific FDR
- Outer node FDR

Can be controlled with BH when tests are independent across levels

Benjamini & Bogomolov (2014)

- Define a new notion of average error rate over the selected families
- Consider only 2 levels trees
- Strategy that offer control for any type of dependence
- **Our work, extension to general trees**

Heller, Chatterjee, Krieger, and Shi (2016)

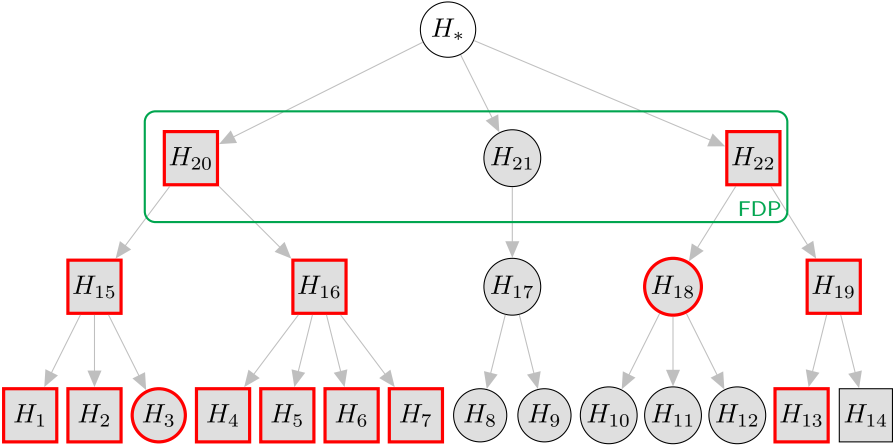
- Talk today...

Barber & Ramdas (2016)

- Non hierarchical testing
- Level specific FDR control
- Restricted to Simes' combination rule

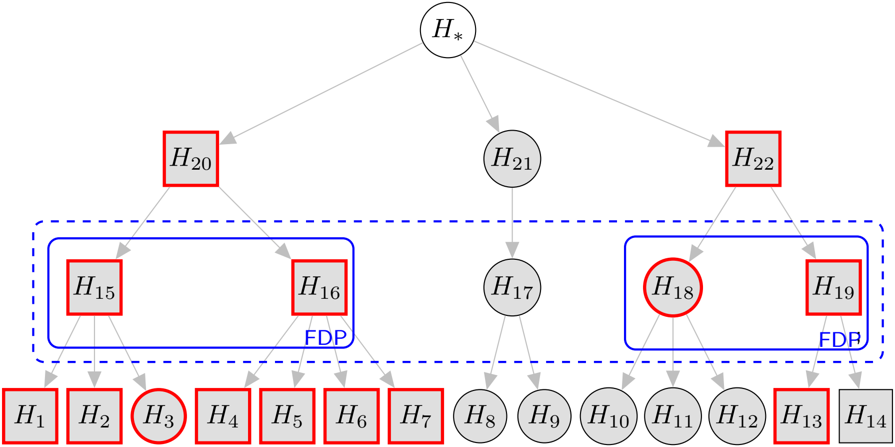
Level 1 selective FDR

Expected value of the FDP at level 1



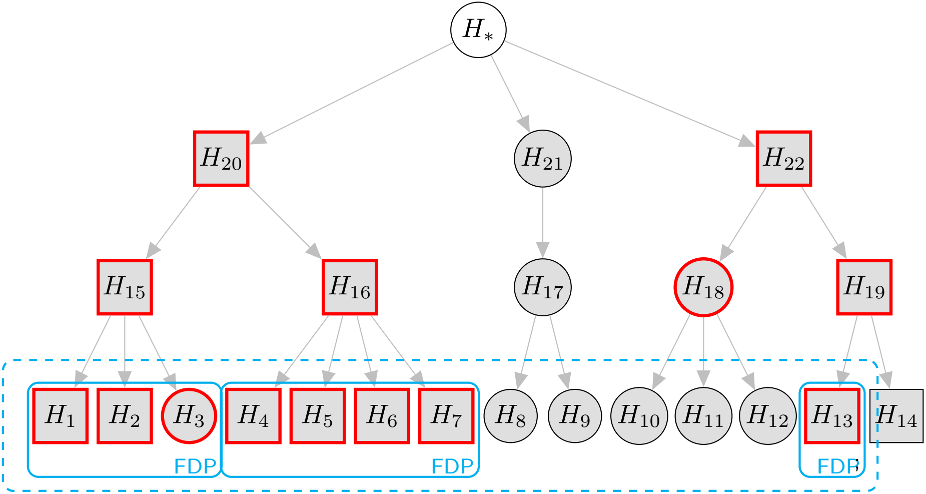
Level 2 selective FDR

Expected value of the weighted average FDP across all families tested at level 2



Level 3 selective FDR

Expected value of the weighted average FDP across all families tested at level 3



Target global error: selective FDR

$$\begin{aligned} \mathrm{sFDR}^{\ell} &= \mathbb{E}(\overline{\mathrm{sFDP}}^{\ell}) \\ \overline{\mathrm{sFDP}}^{\ell} &= \sum_{\mathcal{F}_i^{\ell} \text{ is tested}} w_i \mathrm{FDP}_{\mathcal{F}_i^{\ell}} \end{aligned}$$

- sFDR^1 is equal to FDR^1
- $\mathrm{sFDR}^2 = \mathbb{E}(\sum_{i \in \mathcal{S}^1} \mathrm{FDP}_i / |\mathcal{S}^1|)$ (Benjamini & Bogomolov 2014)
- generically,

$$\overline{\mathrm{sFDP}}^{\ell} = \sum_{i_1 \in \mathcal{S}^1} \frac{1}{|\mathcal{S}^1|} \sum_{i_2 \in \mathcal{S}^2_{i_1}} \frac{1}{|\mathcal{S}^2_{i_1}|} \cdots \sum_{i_{\ell-1} \in \mathcal{S}^{\ell-1}_{i_{\ell-2}}} \frac{1}{|\mathcal{S}^{\ell-1}_{i_{\ell-2}}|} \mathrm{FDP}_{\mathcal{F}^{\ell}_{i_{\ell-1}}}$$

with $|\mathcal{S}_i^{\ell}|$ the max between 1 and the number of rejected hypotheses in family $\mathcal{F}_i^{\ell}$.

Selective FDR

- Statements can be made about the discoveries at each level
- It incorporates the order of testing
- The error rate definition guarantees a coherence among discoveries
  - There is no multiplicity across levels that is unaccounted for
  - Fine scale discoveries can only follow coarser scale ones
  - Discoveries at one level might not be followed up by discoveries at the next level (different from p-filter)
- "Consistency" across levels: control of sFDR$^{\ell}$ guarantees control of sFDR$^{\ell-1}$ **if** whenever a parent hypothesis is rejected at least one of the hypotheses in the family it indexes is rejected.

Testing procedure: TreeBH

```
Input  : The target levels for error rates for sFDR$^\ell$, $\ell = 1, \dots, L$: $q_0^{(1)}, \dots, q_0^{(L)}$
The $p$-values for all the hypotheses in the tree
begin
    $S^0 \longleftarrow i_0$
    $q_{i_0} = q_0^{(1)}$
    for $\ell = 1, \dots, L$ do
        $\mathcal{S}^\ell \longleftarrow \emptyset$
    end
end
for $\ell = 1, \dots, L$ do
    while $\mathcal{S}^{\ell-1} \neq \emptyset$ do
        for $i \in \mathcal{S}^{\ell-1}$ do
            Apply the BH procedure at level $q_i$ on the $p$-values of family $\mathcal{F}_i^\ell$
            $\mathcal{S}^\ell \longleftarrow \mathcal{S}^\ell \bigcup \mathcal{S}_i^\ell$
            for $j \in \mathcal{S}_i^\ell$ do
                $q_j \longleftarrow q_0^{(\ell+1)} \times q_i / q_0^{(\ell)} \times |\mathcal{S}_i^\ell| / |\mathcal{F}_i^\ell|$
            end
        end
    end
end
Output: The set of all the rejected hypotheses, $\bigcup_{\ell=1}^{L} \mathcal{S}^\ell$.
```

Selective FDR control

A1 The BH procedure is valid for the dependence among the p-values within each family, e.g. the p-values are independent or satisfy the positive regression dependence on the subset of true null hypotheses property (PRDS).

A2 For each level $\ell \in \{2, \ldots, L\}$ and each family $\mathcal{F}_i^\ell$, the p-values of the hypotheses in $\mathcal{F}_i^\ell$ are independent of the p-values of the hypotheses at levels $1, \ldots, \ell - 1$ which are not ancestors of the family $\mathcal{F}_i^\ell$.

Theorem

If assumptions A1 and A2 hold, the TreeBH procedure with input parameters $(q^{(1)}, \ldots, q^{(L)})$, guarantees for each $\ell \in \{1, \ldots, L\}$ that sFDR$^\ell \leq q^{(\ell)}$.

Selective FDR control

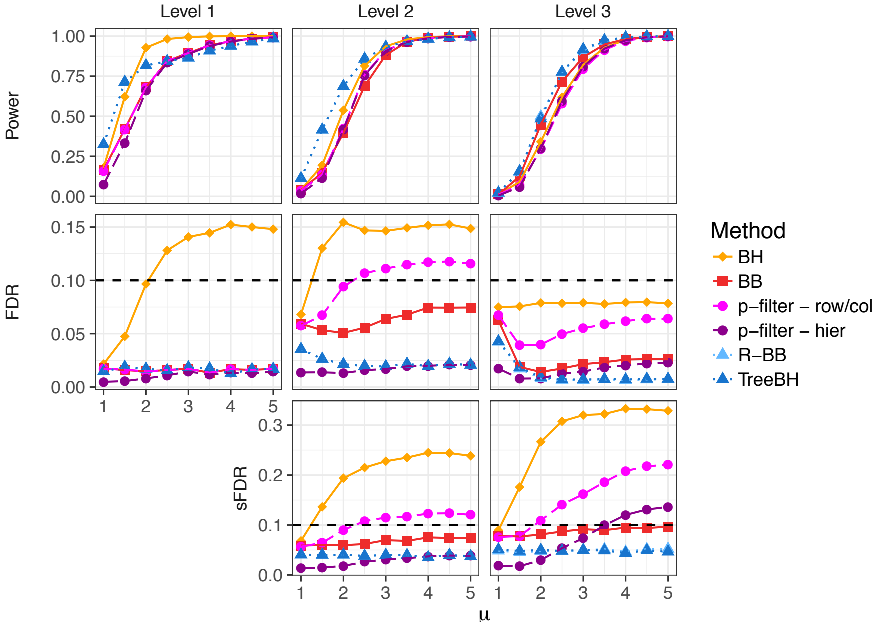
- Measures of error other than FDP can be used for each family
- Combining error controlling procedures with the appropriate level adjustment leads to control
- In each level, the target q is adjusted by S/M, where M represents the total number of hypotheses and S the number of selected ones.
- Driving force is Marina Bogomolov

A didactic example

- Level 1 discoveries: rows
- Level 2 discoveries: groups of columns within rows
- Level 3 discoveries: individual hypotheses
- red: non null hypotheses

$H_{1,1,1}$	$H_{1,1,2}$	$H_{1,2,1}$	$H_{1,2,2}$	$H_{1,3,1}$	$H_{1,3,2}$	$H_{1,4,1}$	$H_{1,4,2}$	$H_{1,5,1}$	$H_{1,5,2}$	$H_{1,6,1}$	$H_{1,6,2}$	$\dots$	$H_{1,6,90}$
$H_{2,1,1}$	$H_{2,1,2}$	$H_{2,2,1}$	$H_{2,2,2}$	$H_{2,3,1}$	$H_{2,3,2}$	$H_{2,4,1}$	$H_{2,4,2}$	$H_{2,5,1}$	$H_{2,5,2}$	$H_{2,6,1}$	$H_{2,6,2}$	$\dots$	$H_{2,6,90}$
$H_{3,1,1}$	$H_{3,1,2}$	$H_{3,2,1}$	$H_{3,2,2}$	$H_{3,3,1}$	$H_{3,3,2}$	$H_{3,4,1}$	$H_{3,4,2}$	$H_{3,5,1}$	$H_{3,5,2}$	$H_{3,6,1}$	$H_{3,6,2}$	$\dots$	$H_{3,6,90}$
$H_{4,1,1}$	$H_{4,1,2}$	$H_{4,2,1}$	$H_{4,2,2}$	$H_{4,3,1}$	$H_{4,3,2}$	$H_{4,4,1}$	$H_{4,4,2}$	$H_{4,5,1}$	$H_{4,5,2}$	$H_{4,6,1}$	$H_{4,6,2}$	$\dots$	$H_{4,6,90}$
$H_{5,1,1}$	$H_{5,1,2}$	$H_{5,2,1}$	$H_{5,2,2}$	$H_{5,3,1}$	$H_{5,3,2}$	$H_{5,4,1}$	$H_{5,4,2}$	$H_{5,5,1}$	$H_{5,5,2}$	$H_{5,6,1}$	$H_{5,6,2}$	$\dots$	$H_{5,6,90}$
$H_{6,1,1}$	$H_{6,1,2}$	$H_{6,2,1}$	$H_{6,2,2}$	$H_{6,3,1}$	$H_{6,3,2}$	$H_{6,4,1}$	$H_{6,4,2}$	$H_{6,5,1}$	$H_{6,5,2}$	$H_{6,6,1}$	$H_{6,6,2}$	$\dots$	$H_{6,6,90}$

- The families $\{H_{i,6,j}, j = 1, \dots 90\}$ contain a lot more hypotheses.
- Family $\{H_{1,6,j}, j = 1, \dots 90\}$ contains many non nulls $\Longrightarrow$ one false discovery at level 3 can generate a much higher family FDP then overall FDP
- The families are very homogeneous $\Longrightarrow$ testing within families should be more powerful.


Level 1
Level 2
Level 3
Power
1.00
0.75
0.50
0.25
0.00
FDR
0.15
0.10
0.05
0.00
sFDR
0.3
0.2
0.1
0.0
1
2
3
4
5
μ
Method
BH
BB
p−filter − row/col
p−filter − hier
R−BB
TreeBH

GTEx data

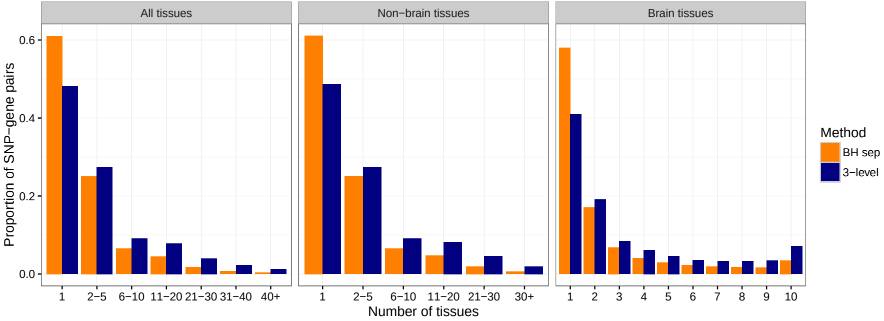
- The Phase 1 data release includes 450 subjects and 44 tissues with at least 60 sample
- Gene expression was measured for around 21,000–34,000 genes per tissue
- genotypes were estimated for around 11 million SNPs. We focus on the set of reasonably independent SNPs filtered to have local $R^2 < 0.5$. After tissue-specific QC, this set includes between 250,000 and 300,000 SNPs per tissue for each of the 44 tissues, with a total of 305,820 SNPs passing QC in at least one tissue.
- Consider *cis* analysis, and the following hierarchy

Level 1	SNP
Level 2	SNP× Gene
Level 3	SNP×Gene×Tissue.

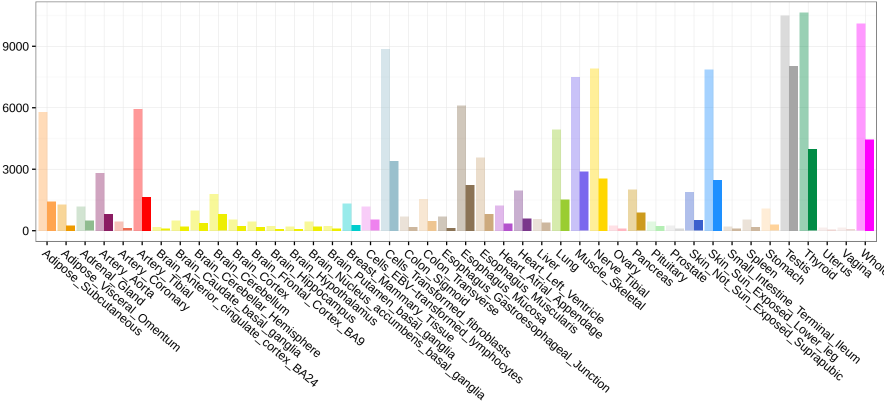
Comparing the results of different procedures

	BH sep	BH pooled	3-level
# eSNPs	9.1×10^4	8.6×10^4	4.5×10^4
% eSNPs	30%	28%	15%
# SNP-gene pairs	1.9×10^5	1.8×10^5	9.3×10^4
# genes per eSNP	2.1	2.0	2.1
# SNP-gene-tissue triplets	6.4×10^5	6.2×10^5	5.1×10^6
# tissues per SNP-gene pair	3.3	3.5	5.4
% SNP-gene pairs 1 tissue only	61%	61%	48%

Number if tissues in which one eSNP is active



Number of tissue-specific eSNPs



Transparent BH sep

Solid hierarchical testing

Model selection with multi-layer FDR control

Set-up (convenience)

$\boldsymbol{y} \in \mathbb{R}^{n\times 1}$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$
- $\boldsymbol{\beta}$ is sparse
- $n \geq p$

GOAL: select $\mathcal{S} \subset \{1, \ldots, p\}$ to approximate the underlying set $\{j : \beta_j \neq 0\}$ of "important" variables.

- $\mathcal{S}$ is interpreted at L different "resolutions" or **layers** of inference.
- Each layer ℓ partitions the hypotheses in disjoint groups $\mathcal{A}_g^\ell$:

$$\bigcup_g^{G_\ell} \mathcal{A}_g^\ell = \{1, \ldots, p\}$$

- Selection set $\mathcal{S}$ induces a selection set at each layer

$$\mathcal{S}_\ell = \{g = 1, \ldots, G_\ell : \mathcal{S} \cap \mathcal{A}_g^\ell \neq \varnothing\}$$

Ex. Service et al. (2014)

A resequencing study with the goal of identifying which variants, in which genes, in which loci are important for some phenotypes

Multiple regression is important to select specific variants.

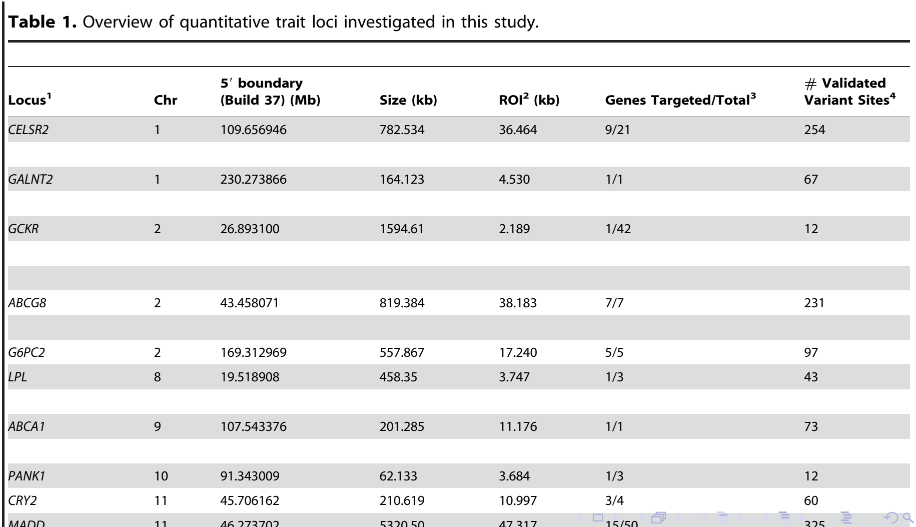
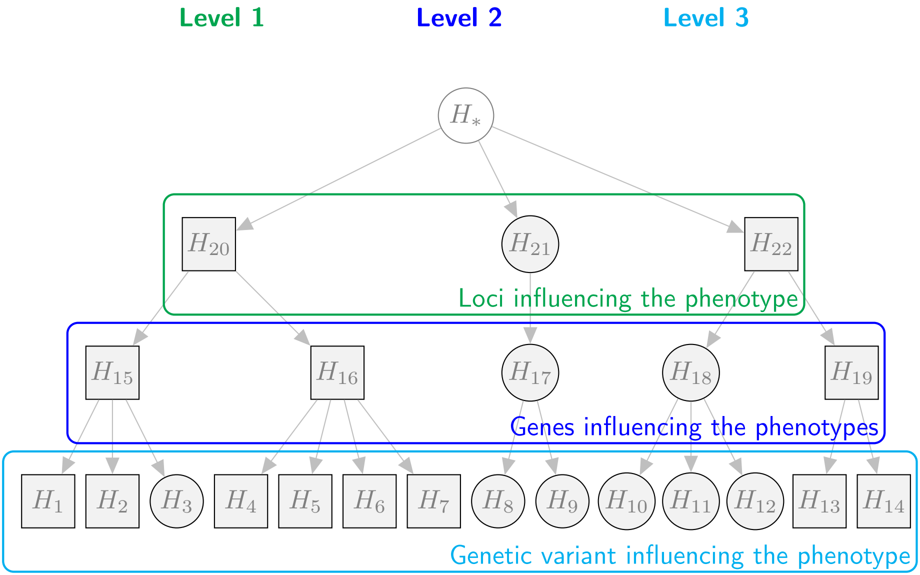


Table 1. Overview of quantitative trait loci investigated in this study.

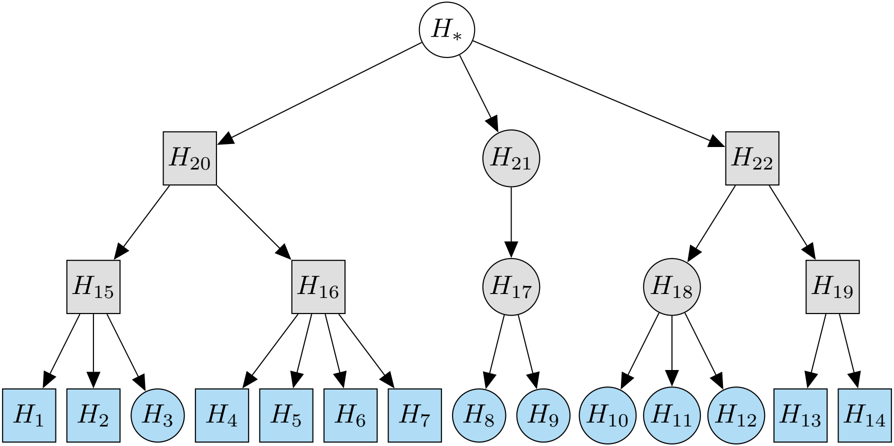
Locus[1]	**Chr**	**5′ boundary (Build 37) (Mb)**	**Size (kb)**	**ROI**[2] **(kb)**	**Genes Targeted/Total**[3]	**# Validated Variant Sites**[4]
CELSR2	1	109.656946	782.534	36.464	9/21	254
GALNT2	1	230.273866	164.123	4.530	1/1	67
GCKR	2	26.893100	1594.61	2.189	1/42	12
ABCG8	2	43.458071	819.384	38.183	7/7	231
G6PC2	2	169.312969	557.867	17.240	5/5	97
LPL	8	19.518908	458.35	3.747	1/3	43
ABCA1	9	107.543376	201.285	11.176	1/1	73
PANK1	10	91.343009	62.133	3.684	1/3	12
CRY2	11	45.706162	210.619	10.997	3/4	60
MADD	11	46.273702	5320.50	47.317	15/50	325

Service et al. example



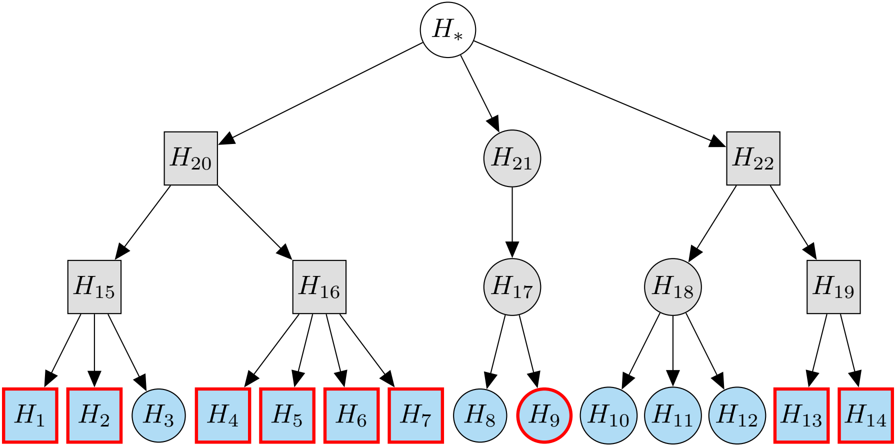
Selections from the finest resolution – no need for hierarchy

(see Barber & Ramdas, 2016)



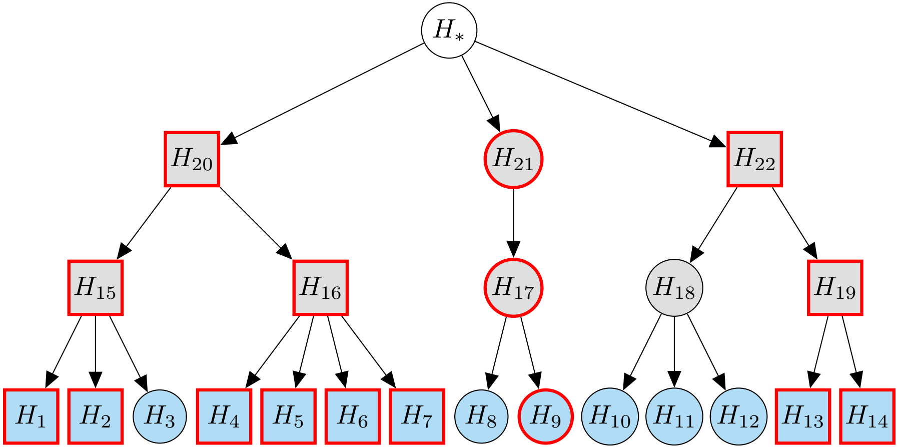
Selections at finest scale

(see Barber & Ramdas, 2016)



Implied rejections at coarser scales

(see Barber & Ramdas, 2016)



Multilayer FDR control

- We focus on level specific FDR
- The rejections at each layer are consistent by definition

Definition

A selection procedure obeys *multilayer FDR control* at levels $q_1, \ldots, q_L$ for each of the layers if

$$\mathrm{FDR}_\ell = \mathbb{E}\left[\frac{|\mathcal{S}_\ell \cap \mathcal{H}_0^\ell|}{|\mathcal{S}_\ell|}\right] \leq q_\ell \text{ for all } \ell.$$

How to achieve this?

Capitalize on a series of strategies recently proposed

- p-filter by Barber & Ramdas (2016)
- Knockoffs: Barber & Candes (2015), Candes et al. (2016)
- group knockoffs: Barber & Dai (2016)

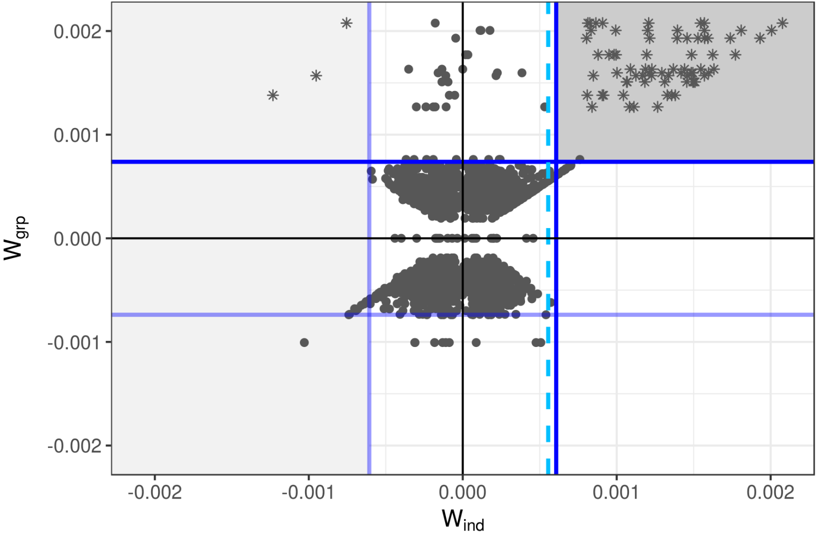
Proving that the strategy actually works required some careful work by Eugene Katsevich

An idea of the algorithm

Framework 1: Multilayer Knockoff Filter

```
Data: $\boldsymbol{X}$, $\boldsymbol{y}$, partitions $\{\mathcal{A}_g^\ell\}_{g,\ell}$ with $g = 1, \ldots, G_\ell$ and $\ell = 1, \ldots, L$, FDR target
      levels $q_1, \ldots, q_L$
for $\ell = 1$ to $L$ do
    Construct group knockoff features $\widetilde{\boldsymbol{X}}^\ell$;
    Construct group knockoff statistics $\boldsymbol{W}^\ell = (W_1^\ell, \ldots, W_{G_\ell}^\ell) = w^\ell([\boldsymbol{X}\ \widetilde{\boldsymbol{X}}^\ell], \boldsymbol{y})$
     satisfying the sign-flip property;
end
For $\boldsymbol{t} = (t_1, \ldots, t_L)$, define $\mathcal{S}(\boldsymbol{t}) = \{j : W_{g(j,\ell)}^\ell \geq t_\ell\ \forall \ell\}$;
For each $\ell$, let $\widehat{V}_\ell(t_\ell) = 1 + |\{g : W_g^\ell \leq -t_\ell\}|$ and define $\widehat{\mathsf{FDP}}_\ell(\boldsymbol{t}) = \frac{\widehat{V}_\ell(t_\ell)}{|\mathcal{S}_\ell(\boldsymbol{t})|}$;
Find $\boldsymbol{t}^* = \min\{\boldsymbol{t} : \widehat{\mathsf{FDP}}_\ell(t) \leq q_\ell\ \forall \ell\}$;
Result: Selection set $\mathcal{S} = \mathcal{S}(\boldsymbol{t}^*)$.
```

Illustration



Properties

Theorem

For any valid construction of group knockoff statistics, the MKF method satisfies

$$\text{FDR}_\ell \le c \cdot q_\ell \qquad \textit{for all } \ell,$$

where $c = 1.93$.

- The knockoff statistics can have arbitrary dependence across layers.
- The constant factor c appears not relevant in practice.
- A similar result holds for a generalization of p-filter.

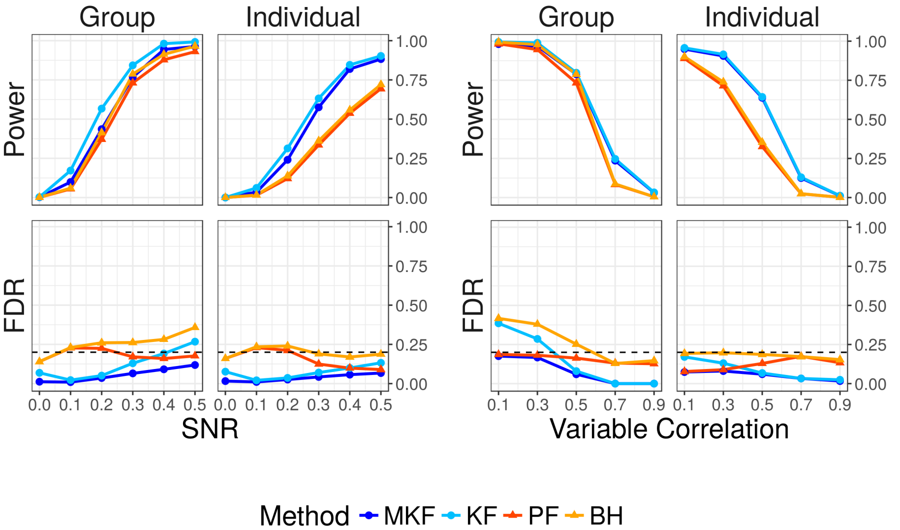
Simulation – set up

- $n = 4500$, $p = 2000$.
- $\boldsymbol{X}$ generated row-wise from AR(1) process with correlation ρ
- $\boldsymbol{y}$ generated from low-dimensional linear model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{\beta}$ has 75 non-null elements
- $L = 2$ with an individual layer and a group layer
- 200 groups of size 10 each
- the non null β_is are in 20 groups

Simulation – results



Analysis of the Service et al. (2014) dataset

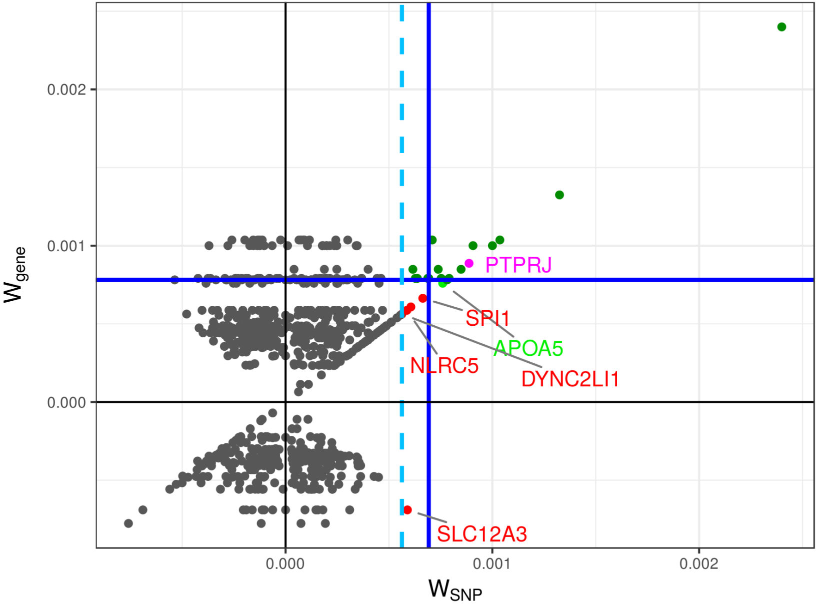
Data

- n=5335 individuals
- p= 768 genetic variants
- G= 85 genes

Methods compared

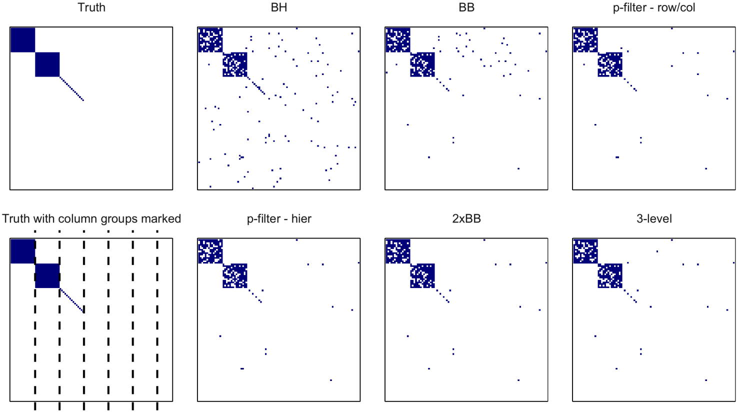
- MKF with $q_{SNP} = q_{gene} = 0.1$
- KF with $q_{SNP} = 0.1$

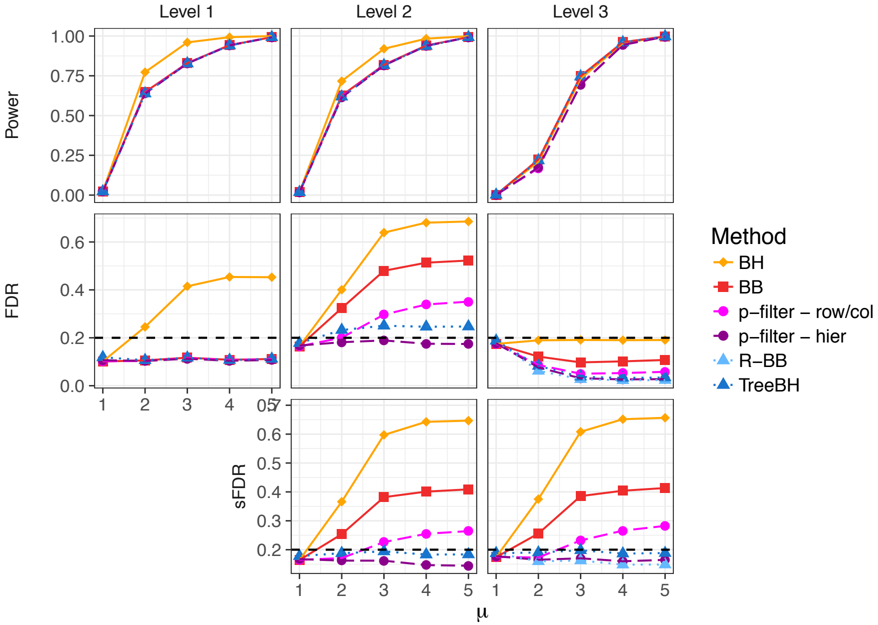
Results


0.002
0.001
0.000
W_{gene}
PTPRJ
SPI1
APOA5
NLRC5
DYNC2LI1
SLC12A3
0.000
0.001
0.002
W_{SNP}

Thank you!

An example from the p-filter paper





Level 1
Level 2
Level 3
Power
1.00
0.75
0.50
0.25
0.00
FDR
0.6
0.4
0.2
0.0
1
2
3
4
5
sFDR
0.7
0.6
0.5
0.4
0.3
0.2
μ
Method
BH
BB
p−filter – row/col
p−filter – hier
R−BB
TreeBH