Inference Following Aggregate Level Hypothesis Testing

Ruth Heller

www.math.tau.ac.il/∼ruheller

Joint work with Nilanjan Chatterjee, Abba Krieger, and Jianxin Shi

A large scale genomic application

- Expression quantitative trait loci (eQTLs) studies aim to identify genetic variants associated with gene expression (eQTL SNPs).
- Within a single tissue may lack power to detect the association due to small sample size.
- The discovery power of eQTL SNPs predictive of gene expression across multiple tissues may be increased by aggregate testing across tissue types.
- For the n=17 tumor tissues in The Cancer Genome Atlas (TCGA) Project, we have $m = 7,732,750$ candidate cis-eQTL SNPs.

The cross-tissue eQTL dataset

The $m \times n = 7,732,750 \times 17$ matrix of p-values is our starting point:

	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LIHC	LUAD	LUSC	OV	PAAD	PRAD	SKCM	UCEC
rs10896016	0.013	0.733	0.266	0.361	0.922	0.007	0.996	0.023	0.140	0.016	0.000	0.129	0.067	0.257	0.141	0.016	0.592
rs1437891	0.455	0.000	0.002	0.902	0.547	0.000	0.520	0.778	0.000	0.344	0.001	0.303	0.163	0.642	0.005	0.415	0.429
rs13066873	0.002	0.000	0.001	0.007	0.544	0.014	0.008	0.003	0.001	0.010	0.000	0.041	0.010	0.043	0.064	0.000	0.002
rs2784574	0.022	0.621	0.874	0.058	0.305	0.507	0.285	0.654	0.693	0.080	0.074	0.086	0.696	0.462	0.922	0.983	0.707
rs11681508	0.109	0.161	0.106	0.928	0.684	0.499	0.739	0.449	0.137	0.601	0.862	0.608	0.844	0.583	0.750	0.528	0.000
rs224962	0.831	0.306	0.814	0.885	0.450	0.579	0.197	0.752	0.478	0.473	0.863	0.212	0.730	0.889	0.741	0.000	0.862
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Two goals for inference

1. To identify the SNPs that influence expression in at least one tissue.

	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LIHC	LUAD	LUSC	OV	PAAD	PRAD	SKCM	UCEC
rs10896016	0.013	0.733	0.266	0.361	0.922	0.007	0.996	0.023	0.140	0.016	0.000	0.129	0.067	0.257	0.141	0.016	0.592
rs1437891	0.455	0.000	0.002	0.902	0.547	0.000	0.520	0.778	0.000	0.344	0.001	0.303	0.163	0.642	0.005	0.415	0.429
rs13066873	0.002	0.000	0.001	0.007	0.544	0.014	0.008	0.003	0.001	0.010	0.000	0.041	0.010	0.043	0.064	0.000	0.002
rs2784574	0.022	0.621	0.874	0.058	0.305	0.507	0.285	0.654	0.693	0.080	0.074	0.086	0.696	0.462	0.922	0.983	0.707
rs11681508	0.109	0.161	0.106	0.928	0.684	0.499	0.739	0.449	0.137	0.601	0.862	0.608	0.844	0.583	0.750	0.528	0.000
rs224962	0.831	0.306	0.814	0.885	0.450	0.579	0.197	0.752	0.478	0.473	0.863	0.212	0.730	0.889	0.741	0.000	0.862
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Two goals for inference

1. To identify the SNPs that influence expression in at least one tissue.

	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LIHC	LUAD	LUSC	OV	PAAD	PRAD	SKCM	UCEC
rs10896016	0.013	0.733	0.266	0.361	0.922	0.007	0.996	0.023	0.140	0.016	0.000	0.129	0.067	0.257	0.141	0.016	0.592
rs1437891	0.455	0.000	0.002	0.902	0.547	0.000	0.520	0.778	0.000	0.344	0.001	0.303	0.163	0.642	0.005	0.415	0.429
rs13066873	0.002	0.000	0.001	0.007	0.544	0.014	0.008	0.003	0.001	0.010	0.000	0.041	0.010	0.043	0.064	0.000	0.002
rs2784574	0.022	0.621	0.874	0.058	0.305	0.507	0.285	0.654	0.693	0.080	0.074	0.086	0.696	0.462	0.922	0.983	0.707
rs11681508	0.109	0.161	0.106	0.928	0.684	0.499	0.739	0.449	0.137	0.601	0.862	0.608	0.844	0.583	0.750	0.528	0.000
rs224962	0.831	0.306	0.814	0.885	0.450	0.579	0.197	0.752	0.478	0.473	0.863	0.212	0.730	0.889	0.741	0.000	0.862
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2. For identified eQTL SNPs, identify the non-null tissues.

	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LIHC	LUAD	LUSC	OV	PAAD	PRAD	SKCM	UCEC
rs10896016	0.013	0.733	0.266	0.361	0.922	0.007	0.996	0.023	0.140	0.016	3.58e-5	0.129	0.067	0.257	0.141	0.016	0.592

	BLCA	**BRCA**	**COAD**	GBM	HNSC	**KIRC**	KIRP	LAML	**LGG**	LIHC	**LUAD**	LUSC	OV	PAAD	**PRAD**	SKCM	UCEC
rs1437891	0.455	2.98e-4	0.002	0.902	0.547	2.56e-7	0.520	0.778	4.54e-5	0.344	0.001	0.303	0.163	0.642	0.005	0.415	0.429

	BLCA	**BRCA**	**COAD**	**GBM**	HNSC	**KIRC**	**KIRP**	**LAML**	**LGG**	**LIHC**	**LUAD**	**LUSC**	**OV**	**PAAD**	PRAD	**SKCM**	**UCEC**
rs13066873	0.002	2.60e-4	0.001	0.007	0.544	0.014	0.008	0.003	0.001	0.010	1.24e-7	0.041	0.010	0.043	0.064	1.83e-4	0.002

Goal 1: meta-analysis

For feature (row) i:

- $H_{ij}, j = 1, \ldots, n$ are the n null hypotheses.
- $H_{iG} = \cap_{j=1}^{n} H_{ij}$ is the meta-analysis (global) null hypothesis.

Goal 1: meta-analysis

For feature (row) i:

- $H_{ij}, j = 1, \ldots, n$ are the n null hypotheses.
- $H_{iG} = \cap_{j=1}^{n} H_{ij}$ is the meta-analysis (global) null hypothesis.

The goal is to test $H_{1G}, \ldots, H_{mG}$, in order to identify the rows with signal in <u>at least one</u> column.

Goal 1: meta-analysis

For feature (row) i:

- $H_{ij}, j = 1, \ldots, n$ are the n null hypotheses.
- $H_{iG} = \cap_{j=1}^{n} H_{ij}$ is the meta-analysis (global) null hypothesis.

The goal is to test $H_{1G}, \ldots, H_{mG}$, in order to identify the rows with signal in <u>at least one</u> column.

A two-step process:

1. Pooling the evidence into an aggregate test (by row).

$$\begin{array}{ccc|c} p_{11} & \cdots & p_{1n} & \boxed{p_{1G}} \\ \vdots & \ddots & \vdots & \vdots \\ p_{m1} & \cdots & p_{mn} & \boxed{p_{mG}} \end{array}$$

2. Applying a multiple testing procedure on the aggregate test p-values

$$\boxed{p_{1G}, \ldots, p_{mG}}$$

Pooling strategies for the first step of the meta-analysis

For row i, P_{ij}, $j = 1, \ldots, n$ are the independent p-values, P_{iG} is the global null p-value.

- The Fisher and Pearson combining methods[1]:

$$p_{iG} = Pr(\chi^2_{2n} \geq -2 \sum_{j=1}^{n} \log p_{ij}).$$

$$p_{iG} = 2Pr\left[\chi^2_{2n} \geq \max\left\{-2\sum_{j=1}^{n} \log p^L_{ij}, -2\sum_{j=1}^{n} \log(1 - p^L_{ij})\right\}\right].$$

[1] Owen, 2009. Karl Pearson's meta-analysis revisited.

[2] Bhattacharjee et al., 2012. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits.

Pooling strategies for the first step of the meta-analysis

For row i, P_{ij}, $j = 1, \ldots, n$ are the independent p-values, P_{iG} is the global null p-value.

- The Fisher and Pearson combining methods[1]:

$$p_{iG} = Pr(\chi^2_{2n} \geq -2 \sum_{j=1}^{n} \log p_{ij}).$$

$$p_{iG} = 2Pr\left[\chi^2_{2n} \geq \max\left\{-2\sum_{j=1}^{n} \log p^L_{ij}, -2\sum_{j=1}^{n} \log(1 - p^L_{ij})\right\}\right].$$

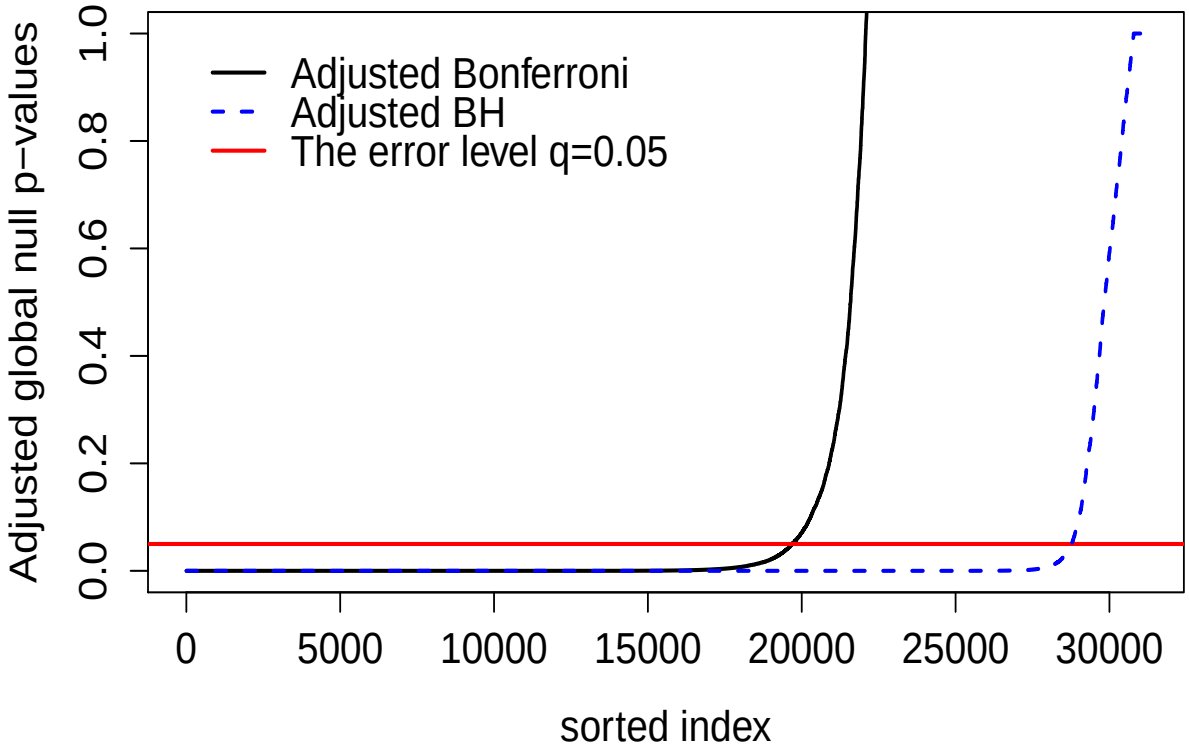
- The Stouffer combining method: $P_{iG} = 1 - \Phi(\sum_{j=1}^{n} z_{ij}/\sqrt{n})$, $z_{ij} = \Phi^{-1}(1 - p_{ij})$.
- Association analysis based on SubSETs (ASSET)[2]: significance based on $\sum_{j \in S_{\max}} w_j Z_j / \sqrt{|S_{\max}|}$, where $S_{\max}$ is the set with largest weighted Stouffer test statistic.
- ...

[1] Owen, 2009. Karl Pearson's meta-analysis revisited.

[2] Bhattacharjee et al., 2012. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits.

Results for the cross-tissue eQTL meta-analysis

using Pearson's p-values, adjusting for multiplicity by Bonferonni or BH[1].



[1] Benjamini and Hochberg, 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.

Goal 2: Inference following selection by aggregate level testing

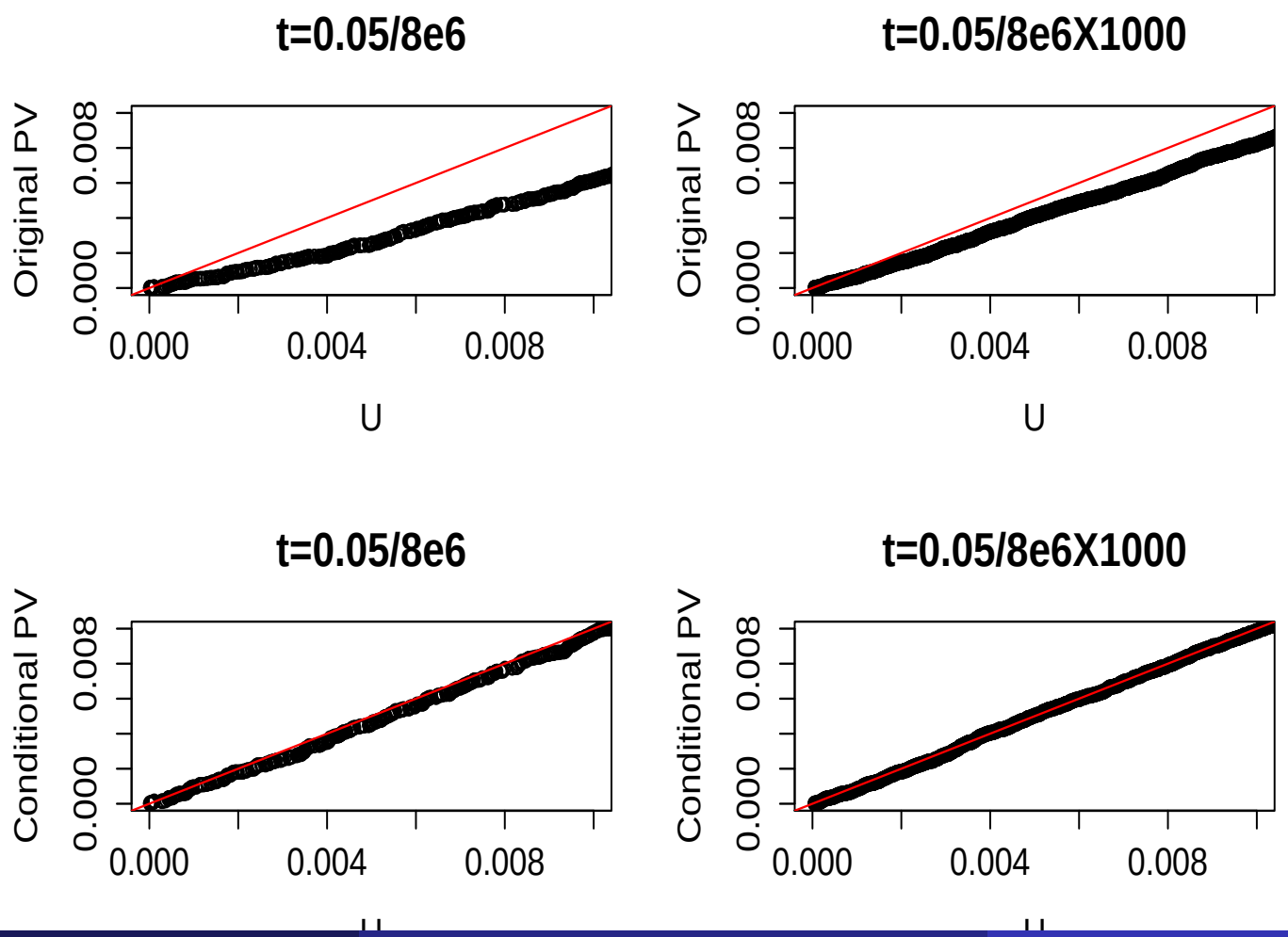
- In meta-analysis, aggregate level hypotheses testing is performed for powerful identification of rows with signal[1].
- A natural follow-up question is which studies contain signal within a discovered row.
- Testing $H_{i1}, \ldots, H_{in}$ following rejection of H_{iG} without accounting for the fact that H_{iG} was rejected using an aggregate-level test statistic, will produce biased inference [2] and hence an inflation of non-replicable results.

[1] Bhattacharjee et al., 2012. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits.

[2] Bogomolov and Benjamini, 2014. Selective inference on multiple families of hypotheses.

Distribution of a null p-value following selection

Figure: Given that the meta-analysis of $n = 20$ studies had Pearson's $P_G \leq t$, for a single null hypothesis the quantile plot of the conditional p-value (row 1) and naive p-value (row 2) versus the uniform.



Control over false positives

- When considering the family of all individual level hypotheses within all selected rows, regardless of row membership:
  - With overall FDR control, the false discovery proportions can be as high as one within a specific row.
  - With overall error control, the power may be low for large m.
- We suggest
  FWER/FDR control conditional on the row being selected. [1].
  - This type of false positive control is particularly important if a researcher conducts different follow-up studies for each selected row.
- A related goal: Controlling the average FWER/FDR over the selected rows[2].

[1] Heller, Chatterjee, Krieger, and Shi, 2016. Post-selection inference following aggregate level hypotheses testing in large scale genomic data.

[2] Benjamini and Bogomolov, 2014. Selective inference on multiple families of hypotheses.

The conditional error

- $\mathcal{S} \subseteq \{1, \ldots, m\}$ is the set of selected rows, e.g., all hypotheses rejected by Bonferroni/BH on the global null p-values.
- V_i = number of false discoveries for row i.
- R_i = number of discoveries for row i.
- The **conditional FWER** for row i is

$$E(I[V_i > 0] | i \in \mathcal{S}).$$

- The **conditional FDR** for row i is

$$E(V_i / \max\{R_i, 1\} | i \in \mathcal{S}).$$

Our approach for inference following row-selection

1. Compute the conditional p-values, conditional on being selected.
2. Apply a valid FWER/FDR controlling procedure on the conditional p-values.

Our approach for inference following row-selection

1. Compute the conditional p-values, conditional on being selected.
2. Apply a valid FWER/FDR controlling procedure on the conditional p-values.

Questions we address:

1. The row may contain both null and non-null p-values, so the probability of selection is not known even for the simplest rule $\{P_{iG} \leq t\}$. How can the conditional p-values be computed?
2. Even though the original p-values in a row are independent, the conditional p-values will be dependent.
   What is a valid FDR controlling procedure?

The conditional p-value computation for a selected row

Per column, we compute the p-value conditional on the event that the row was selected, holding all other p-values in the row fixed.

For example, for the first column:

$$p'_{i1} = p_{i1}/b_{i1}, \quad b_{i1} = \max\{p : p_{iG}(p, p_{i2}, \ldots, p_{in}) \leq t\}.$$

This is a valid p-value, since:

- P_{i1} is independent of $P_{i2}, \ldots, P_{in}$.
- if H_{i1} is null, then

$$P_{i1} \mid P_{iG} \leq t, P_{i2} = p_{i2}, \ldots, P_{in} = p_{in} \sim U(0, b_{i1}).$$

Properties of the conditional p-values

- If $P_{iG}(1, p_{i2}, \ldots, p_{in}) \leq t$, there is no correction: $p'_{i1} - p_{i1} = 0$.
- The ranking of the conditional p-values is the same as that of the original p-values, using Fisher's or Stouffer's combining method for aggregate testing.
- With Bonferroni-Holm/BH at level α on $p'_{i1}, \ldots, p'_{in}$, the conditional FWER/FDR is controlled.

Theoretical results for FDR control

Following selection of rows using a fixed cut-off

Theorem

If $p_{iG} \leq t_i$, then for the BH procedure at level α on $p'_{i1}, \ldots, p'_{in}$,

$$E(V_i / \max\{R_i, 1\} | i \in \mathcal{S}) \leq \frac{n_0(i)}{n} \alpha.$$

Equality follows if the global null p-value is Fisher's.

Following adaptive selection of rows, e.g., BH on $\{p_{iG}, i = 1, \ldots, m\}$

Theorem

Under row independence, if $p_{iG} \leq t(|\mathcal{S}|)$, then for the BH procedure at level α on $p'_{i1}, \ldots, p'_{in}$,

$$E(V_i / \max\{R_i, 1\} | i \in \mathcal{S}) \leq \frac{n_0(i)}{n} \alpha.$$

Proof when row selection is by a fixed cut-off

- Assume the first column is null.
- $I = 1$ if H_1 is rejected.
- $R =$ number of discoveries in the row.
- Using the representation of FDR from Benjamini and Yekutieli (2001)[3], the **conditional FDR** is

$$n_0 \sum_{k=1}^{n} \frac{1}{k} Pr\left(I = 1, R = k \mid p_G(P_1, P_2, \ldots, P_n) \leq t\right)$$

- We condition on $p_2, \ldots, p_n$ so that it is sufficient to show that

$$\sum_{k=1}^{n} \frac{1}{k} Pr\left(I = 1, R = k \mid p_G(P_1, p_2, \ldots, p_n) \leq t, P_2 = p_2, \ldots, P_n = p_n\right) \leq \frac{\alpha}{n}.$$

[3] Benjamini and Yekutieli, 2001. The control of the false discovery rate in multiple testing under dependency.

Proof when row selection is by a fixed cut-off

- $p_1' = p_1/b_1, \quad b_1 = \max\{p : p_G(p, p_2, \ldots, p_n) \leq t\}$.
- As p_1' increases $b_2, \ldots, b_n$ will be non-increasing.
- There must be $0 = a_0 < a_1 < \ldots < a_L = 1$ so that $R(p_1') = k_l$ for $a_{l-1} \leq p_1' \leq a_l, l = 1, \ldots, L$, where $k_1 > k_2 > \ldots > k_L$.
- Since we need $l = 1$, or $p_1' \leq R(p_1')\alpha/n$, there exists t such that

$$\sum_{k=1}^{n} \frac{1}{k} Pr\left(l = 1, R = k \mid p_G(P_1, p_2, \ldots, p_n) \leq t, P_2 = p_2, \ldots, P_n = p_n\right)$$

$$= \sum_{k=1}^{t-1} \frac{1}{k_l}(a_l - a_{l-1}) + \frac{1}{k_t}(k_t\alpha/n - a_{t-1}) \leq \frac{1}{k_t}\frac{k_t\alpha}{n} \leq \frac{\alpha}{n}.$$

Results for the cross-tissue eQTL analysis in TCGA

Table: The original two-sided p-values, conditional two-sided p-values, and BH-adjusted conditional two-sided p-values for each tissue, for three eQTL SNPs that differ in the number of post-selection discoveries.

	rs10896016-CTSW p-values			rs1437891-ASNSD1 p-values			rs13066873-LARS2 p-values		
	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$
BLCA	0.01259	0.29510	0.38590	0.45523	0.45523	0.64491	0.00199	0.00199	**0.00484**
BRCA	0.73273	0.73273	0.83043	0.00030	0.00804	**0.02278**	0.00026	0.00026	**0.00147**
COAD	0.26604	0.29510	0.38590	0.00231	0.00231	**0.02278**	0.00099	0.00099	**0.00362**
GBM	0.36091	0.29510	0.38590	0.90232	0.90232	0.90232	0.00716	0.00716	**0.01353**
HNSC	0.92247	0.92247	0.98012	0.54711	0.54711	0.66435	0.54393	0.54393	0.54393
KIRC	0.00743	0.29510	0.38590	2.56e-7	0.00804	**0.02278**	0.01362	0.01362	**0.01781**
KIRP	0.99577	0.99577	0.99577	0.51974	0.51974	0.66435	0.00834	0.00834	**0.01418**
LAML	0.02349	0.29510	0.38590	0.77827	0.77827	0.82691	0.00345	0.00345	**0.00733**
LGG	0.13963	0.29510	0.38590	0.00005	0.00804	**0.02278**	0.00107	0.00107	**0.00362**
LIHC	0.01575	0.29510	0.38590	0.34415	0.34415	0.64491	0.01007	0.01007	**0.01426**
LUAD	0.00004	0.29510	0.38590	0.00078	0.00804	**0.02278**	1.24e-7	1.24e-7	**2.11e-6**
LUSC	0.12911	0.29510	0.38590	0.30344	0.30344	0.64481	0.04074	0.04074	**0.04827**
OV	0.06658	0.29510	0.38590	0.16256	0.16256	0.39479	0.00961	0.00961	**0.01426**
PAAD	0.25674	0.25674	0.38590	0.64167	0.64167	0.72723	0.04259	0.04259	**0.04827**
PRAD	0.14091	0.29510	0.38590	0.00495	0.00804	**0.02278**	0.06407	0.06407	0.06807
SKCM	0.01577	0.29510	0.38590	0.41503	0.41503	0.64491	0.00018	0.00018	**0.00147**
UCEC	0.59226	0.59226	0.71917	0.42909	0.42909	0.64491	0.00167	0.00167	**0.00473**
p_{iG}	3×10^{-9}			2×10^{-10}			$< 10^{-20}$		

An existing alternative approach[1]

The BB selection adjusted procedure: apply an FWER/FDR controlling procedure within selected rows at level $\frac{|\mathcal{S}|}{m}\alpha$.

Theorem (Benjamini and Bogomolov, 2014)

If for each column, the set of p-values is PRDS on the subset of p-values corresponding to true null hypotheses, the selection is by fixed thresholding/BH on the global null p-values, and the procedure used for testing each selected row is level α (a) Bonferroni or (b) BH, then the select-adjusted procedure guarantees in case (a)

$$E\left(\frac{\sum_{i\in\mathcal{S}} I[V_i > 0]}{\max\{|\mathcal{S}|, 1\}}\right) \leq \alpha,$$

and in case (b)

$$E\left(\frac{\sum_{i\in\mathcal{S}} V_i / \max\{R_i, 1\}}{\max\{|\mathcal{S}|, 1\}}\right) \leq \alpha.$$

.

[1] Benjamini and Bogomolov, 2014. Selective inference on multiple families of hypotheses.

Results for the cross-tissue eQTL analysis in TCGA

The BB selection adjusted procedure applies the BH procedure on the original p-values at level $\frac{19{,}690}{7{,}732{,}750}0.05 = 0.00013$. With BB: no discoveries are made for the first eQTL SNP; a single discovery is made for the second and third eQTL SNP.

	rs10896016-CTSW p-values			rs1437891-ASNSD1 p-values			rs13066873-LARS2 p-values		
	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$
BLCA	0.01259	0.29510	0.38590	0.45523	0.45523	0.64491	0.00199	0.00199	**0.00484**
BRCA	0.73273	0.73273	0.83043	0.00030	0.00804	**0.02278**	0.00026	0.00026	**0.00147**
COAD	0.26604	0.29510	0.38590	0.00231	0.00231	**0.02278**	0.00099	0.00099	**0.00362**
GBM	0.36091	0.29510	0.38590	0.90232	0.90232	0.90232	0.00716	0.00716	**0.01353**
HNSC	0.92247	0.92247	0.98012	0.54711	0.54711	0.66435	0.54393	0.54393	0.54393
KIRC	0.00743	0.29510	0.38590	$2.56e-7$	0.00804	**0.02278**	0.01362	0.01362	**0.01781**
KIRP	0.99577	0.99577	0.99577	0.51974	0.51974	0.66435	0.00834	0.00834	**0.01418**
LAML	0.02349	0.29510	0.38590	0.77827	0.77827	0.82691	0.00345	0.00345	**0.00733**
LGG	0.13963	0.29510	0.38590	0.00005	0.00804	**0.02278**	0.00107	0.00107	**0.00362**
LIHC	0.01575	0.29510	0.38590	0.34415	0.34415	0.64491	0.01007	0.01007	**0.01426**
LUAD	0.00004	0.29510	0.38590	0.00078	0.00804	**0.02278**	$1.24e-7$	1.24e-7	$\mathbf{2.11e-6}$
LUSC	0.12911	0.29510	0.38590	0.30344	0.30344	0.64481	0.04074	0.04074	**0.04827**
OV	0.06658	0.29510	0.38590	0.16256	0.16256	0.39479	0.00961	0.00961	**0.01426**
PAAD	0.25674	0.25674	0.38590	0.64167	0.64167	0.72723	0.04259	0.04259	**0.04827**
PRAD	0.14091	0.29510	0.38590	0.00495	0.00804	**0.02278**	0.06407	0.06407	0.06807
SKCM	0.01577	0.29510	0.38590	0.41503	0.41503	0.64491	0.00018	0.00018	**0.00147**
UCEC	0.59226	0.59226	0.71917	0.42909	0.42909	0.64491	0.00167	0.00167	**0.00473**
p_{iG}	3×10^{-9}			2×10^{-10}			$< 10^{-20}$		

Comparison of approaches: error rate guarantees

Denote empirical control with ✓ and theoretical control with ✓.

With a fixed cut-off row-selection rule (e.g., Bonferroni)

	Approach	Average error	Conditional error for a nonnull row	Conditional error for a null row
Row independence	Naive	X	X	X
	BB	✓	✓	X
	conditional	✓	✓	✓
Row dependence	Naive	X	X	X
	BB	✓ PRDS	X	X
	conditional	✓ PRDS	✓	✓

Comparison of approaches: error rate guarantees

Denote empirical control with ✓ and theoretical control with ✓.

With a fixed cut-off row-selection rule (e.g., Bonferroni)

	Approach	Average error	Conditional error for a nonnull row	Conditional error for a null row
Row independence	Naive	X	X	X
	BB	✓	✓	X
	conditional	✓	✓	✓
Row dependence	Naive	X	X	X
	BB	✓ PRDS	X	X
	conditional	✓ PRDS	✓	✓

With a data-adaptive row-selection rule (e.g., BH)

	Approach	Average error	Conditional error for a nonnull row	Conditional error for a null row
Row independence	Naive	X	X	X
	BB	✓	✓	X
	conditional	✓	✓	✓
Row PRDS	Naive	X	X	X
	BB	✓	X	X
	conditional	✓	✓	X

Simulations with block dependence

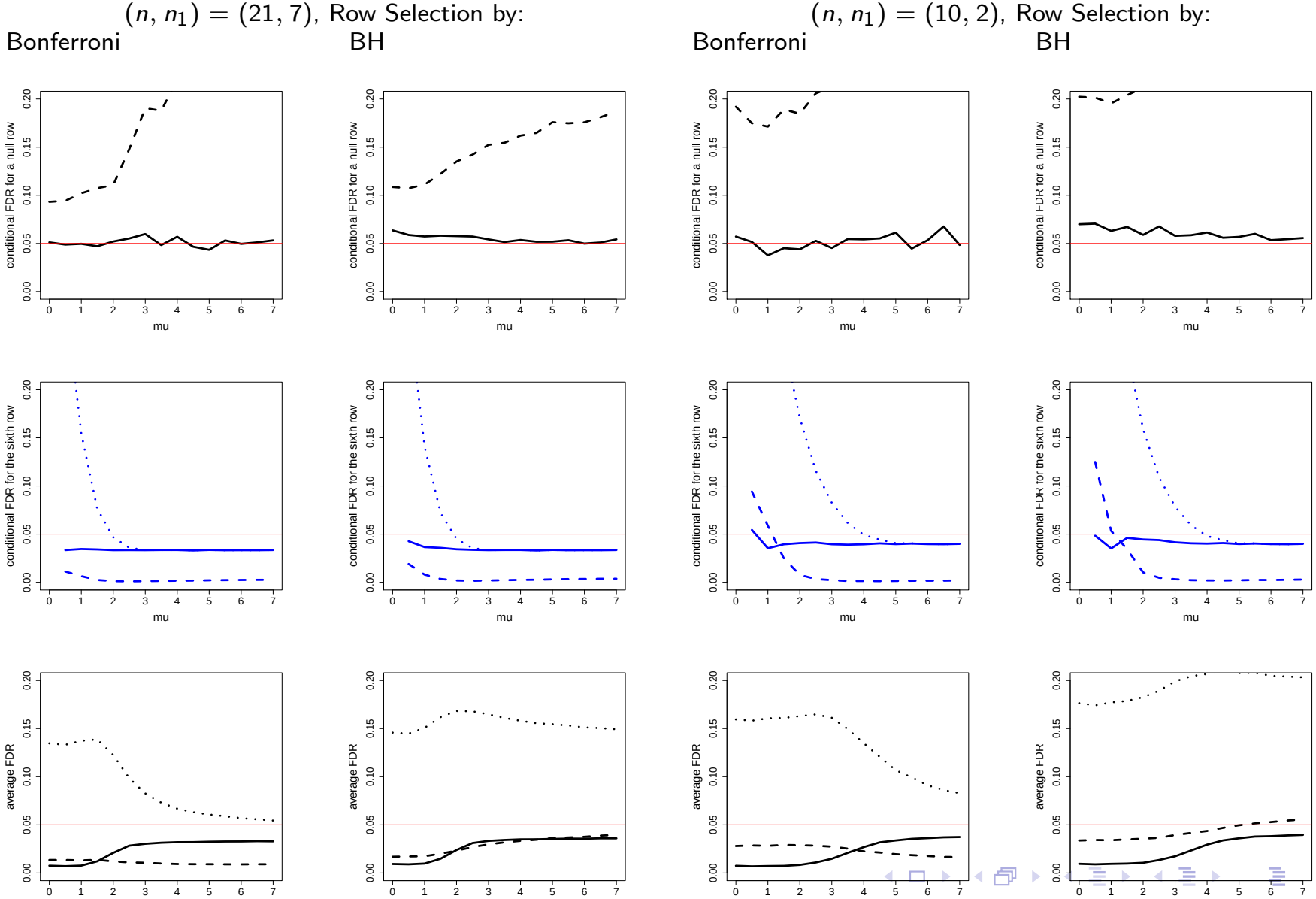
We consider 100 blocks of 11 rows, where the signal within non-null blocks is $N_{11}(\vec{\mu}, \Sigma)$ and within null blocks is $N_{11}(\vec{0}, \Sigma)$, where

$$\vec{\mu} = \begin{pmatrix} \rho^5\mu \\ \vdots \\ \rho\mu \\ \mu \\ \rho\mu \\ \vdots \\ \rho^5\mu \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{B-1} \\ \rho & 1 & \rho & \dots & \rho^{B-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{B-1} & \rho^{B-2} & \rho^{B-3} & \dots & 1 \end{pmatrix},$$

In n_1 studies there was one non-null block, and the remaining $n - n_1$ studies where all null:

$$\begin{pmatrix} \boxed{N_{11}(\vec{\mu}, \Sigma)} & \dots & \boxed{N_{11}(\vec{\mu}, \Sigma)} & N_{11}(\vec{0}, \Sigma) & \dots & N_{11}(\vec{0}, \Sigma) \\ N_{11}(\vec{0}, \Sigma) & \dots & N_{11}(\vec{0}, \Sigma) & N_{11}(\vec{0}, \Sigma) & \dots & N_{11}(\vec{0}, \Sigma) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$

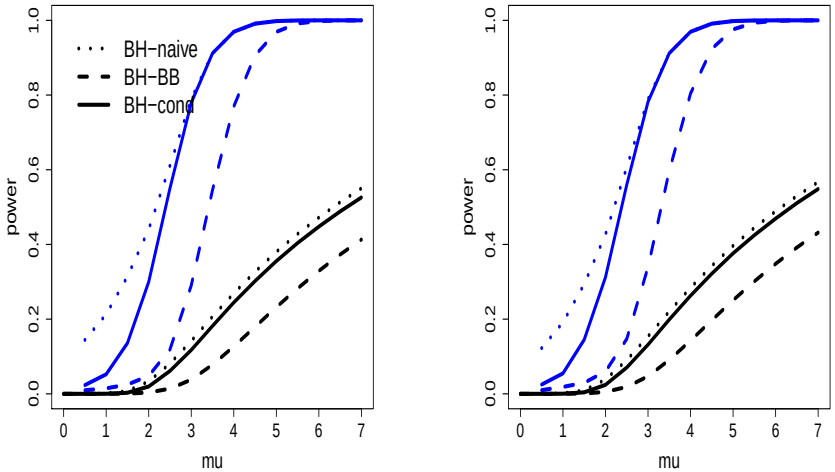
Results on error control: conditional approach (solid), BB (dashed), naive (dotted)



Results on power: conditional approach (solid), BB (dashed), naive (dotted)

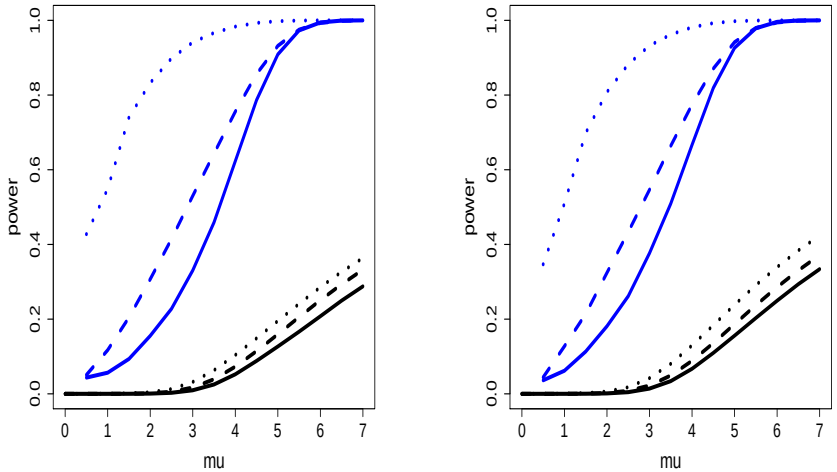
$(n, n_1) = (21, 7)$, Row Selection by:

Bonferroni BH



$(n, n_1) = (10, 2)$, Row Selection by:

Bonferroni BH



Comparison of valid approaches: power

- The Benjamini and Bogomolov (2014) approach has increased power as the fraction of selected rows, $|\mathcal{S}|/m$, increases.
- The conditional approach has increased power as the number of non-null columns in the row increases.
  - For moderate signal distributed sparsely within a row, and $|\mathcal{S}|/m$ not too small, the approach of Benjamini and Bogomolov (2014) may have better power.
  - For identification of eQTL SNPs in TCGA, since $|\mathcal{S}|/m$ is small and the signal is not very sparse across tissues, the conditional approach has greater power than the approach of Benjamini and Bogomolov (2014).

Summary

- In large scale analysis of genomic data, it is common to perform tests at an aggregate (row) level for powerful identification of the signal.
- Following row-selection, we presented a valid and powerful selection adjusted method for identification of columns/studies that drive the signal in the row[1].
- The choice of aggregate level test, and rule for row selection, affect the power of the meta-analysis as well as the post-selection inference.
  - For identification of eQTL SNPs, Bonferroni row-selection based on the Pearson global null p-values worked well.

[1] Heller, Chatterjee, Krieger, and Shi, 2016. Post-selection inference following aggregate level hypotheses testing in large scale genomic data.

An extension to dependent columns

- For dependent columns, with known dependence, we can compute valid p-values following row selection using the polyhedral lemma[2].
- An example application is GWAS, where aggregate tests are used for gene discovery and the dependence within the gene is known. An open question is inference at the variant level following gene-level association testing.
- We suggest valid conditional p-values for inference at the individual level, as well estimation of effect sizes, following selection by an aggregate test that takes the known dependence into account[3].

[2] Lee, Sun, Sun and Taylor, 2016. Exact post model selection inference, with application to the lasso.

[3] Heller, Meir, and Chatterjee, work in progress.

Some open questions

- Theoretical justification for average error control when the rows are dependent and the row selection is data-adaptive.
- Examination of the conditional error control when using a plug-in estimate of the fraction of nulls in a row with an FDR controlling procedure on the conditional p-values.
- Investigation of multi-layer strategies with the conditional approach (e.g., first identify sets of rows, then rows, then columns...)[4].

[4] Great progress has been recently made in controlling the FDR at multiple resolutions: Foygel Barber and Ramdas (2016). The p-filter: multi-layer FDR control for grouped hypotheses; Liu, Sarkar, Zhao (2016). A new approach to multiple testing of grouped hypotheses; Bogomolov, Peterson, Benjamini, Sabatti (2017). Testing hypotheses on a tree: new error rates and controlling strategies; Katsevich and Sabatti (2017). Multilayer Knockoff Filter: Controlled variable selection at multiple resolutions.