Family-wise Error Rate Control in QTL Mapping and Gene Ontology Graphs

with Remarks on Family Selection

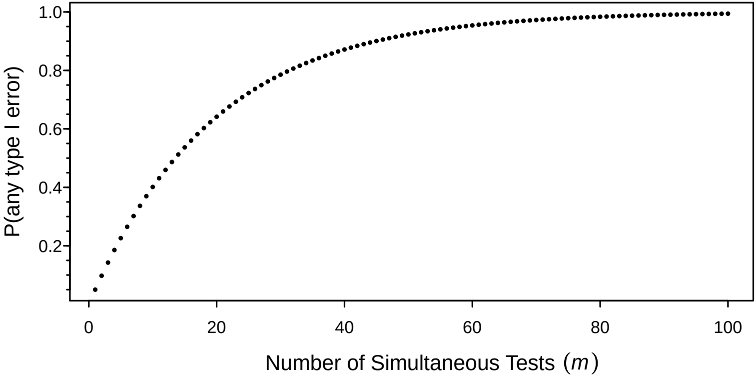
Garrett Saunders – Dissertation Defense

April 15, 2014

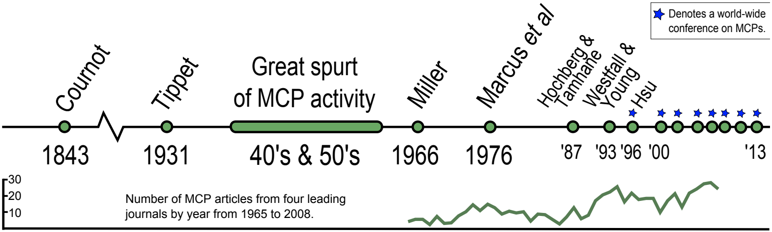
Contents

1. Introduction

"If enough statistics are computed, some of them will be sure to show structure." (Diaconis 1985)



1. Introduction



1. Introduction

"The Vitality of [the] field in the future – as a research area – depends upon [the researcher's] ability to continue and address the real needs of statistical analysis in current problems" (Benjamini, 2010).

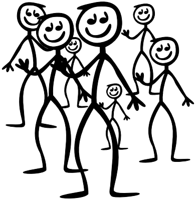
1. Introduction

	Declared non-significant	*Declared significant*	*Total*
True null hypothesis	$\mathbf{U}$	$\mathbf{V}$	m_0
False null hypothesis	$\mathbf{T}$	$\mathbf{S}$	$m - m_0$
	$m - \mathbf{R}$	$\mathbf{R}$	m

- The Per Comparison Error Rate (PCER): $E(V/m)$
- The Familywise Error Rate (FWER): $P(V \geq 1)$
- The False Discovery Rate (FDR): $E(V/R)$

1. Introduction

Selecting a Family of Hypotheses



- A subjective, but important decision.
- “Any collection of inferences for which it is meaningful to take into account some combined measure of errors.” (Hochberg & Tamhane 1987)
- Gatekeeping (Bretz *et al.* 2009)

1. Introduction

The Bonferroni Adjustment: *test each* H_i *at level* α/m

- Boole's Inequality:

$$P(A \cup B) \leq P(A) + P(B) \quad \ldots \text{or generally} \quad P\left(\bigcup A_i\right) \leq \sum P(A_i)$$

- Let R_i denote the event that hypothesis H_i is rejected.
- Then, if $P_{H_i}(R_i) = \alpha_i$,

$$\text{FWER} = P_{H_i}\left(\bigcup R_i\right) \leq \sum P_{H_i}(R_i) = \sum \alpha_i$$

- If $\alpha_i = \alpha/m$ for all i, then FWER $\leq \alpha$.

1. Introduction

Weighted Bonferroni Adjustment: *test H_i at level α_i, s.t. $\sum \alpha_i \leq \alpha$*

- Since

$$\text{FWER} = P_{H_i}\left(\bigcup R_i\right) \leq \sum P_{H_i}(R_i) = \sum \alpha_i$$

- So long as

$$\sum \alpha_i \leq \alpha,$$

- then FWER $\leq \alpha$.

1. Introduction

Holm's Sequential Bonferroni: *test ordered* $H_{(j)}$ *at level* $\alpha/(m-j+1)$

- Let P_i denoted the p-value for H_i.
- Let $I \subseteq \{1,\ldots,m\}$ index the true H_i, $|I| = k \leq m$.
- Then,

$$
\begin{aligned}
P\left(P_i > \frac{\alpha}{k} \text{ for all } i \in I\right) &= 1 - P\left(P_i \leq \frac{\alpha}{k} \text{ for some } i \in I\right) \\
&\geq 1 - \sum_{i \in I} P\left(P_i \leq \frac{\alpha}{k}\right) \\
&\geq 1 - k\frac{\alpha}{k} = 1 - \alpha.
\end{aligned}
$$

- Since $m - j + 1 \geq k$, FWER $\leq \alpha$.

1. Introduction

Closed Testing: *reject* w_i *iff all* $w_j \subset w_i$ *are rejected at level* α

- Let W be a set of hypotheses.
- W is *closed under intersection* if:
    - for any two hypotheses $H_i, H_j \in W$,
    - $w = H_i \cap H_j$ is also in W.

An example

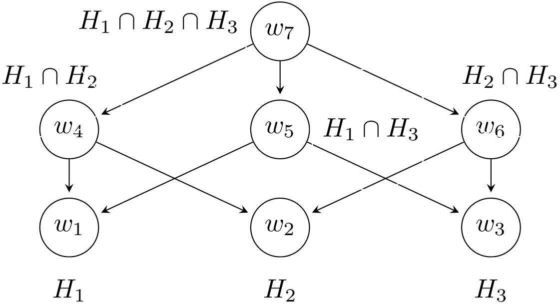
- Consider the elementary hypotheses H_1, H_2, and H_3.
- Let $w_1 = H_1$, $w_2 = H_2$, $w_3 = H_3$, and

  $w_4 = H_1 \cap H_2$, $w_5 = H_1 \cap H_3$, $w_6 = H_2 \cap H_3$, and $w_7 = H_1 \cap H_2 \cap H_3$
- $W = \{w_1, \ldots, w_7\}$ is a set of hypotheses closed under intersections.

1. Introduction

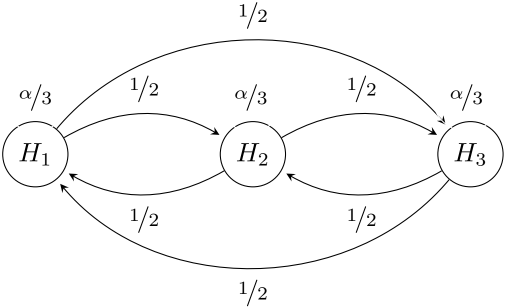
Closed Testing: *reject* w_i *iff all* $w_j \subset w_i$ *are rejected at level* α

- $W = \{w_1, \ldots, w_7\}$ is a set of hypotheses closed under intersections.



1. Introduction

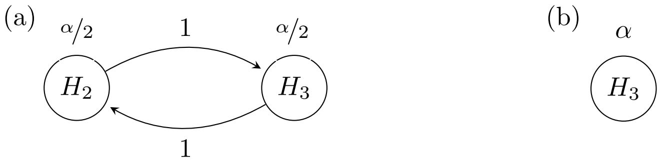
Generalized Weighted Bonferroni Testing



- $\sum_{i=1}^{m} a_i \leq a$
- $0 \leq g_{ij} \leq 1$, $g_{ii} = 0$, and $\sum_{k=1}^{m} g_{ik} \leq 1$ for all $i, j = 1, \ldots, m$.

1. Introduction

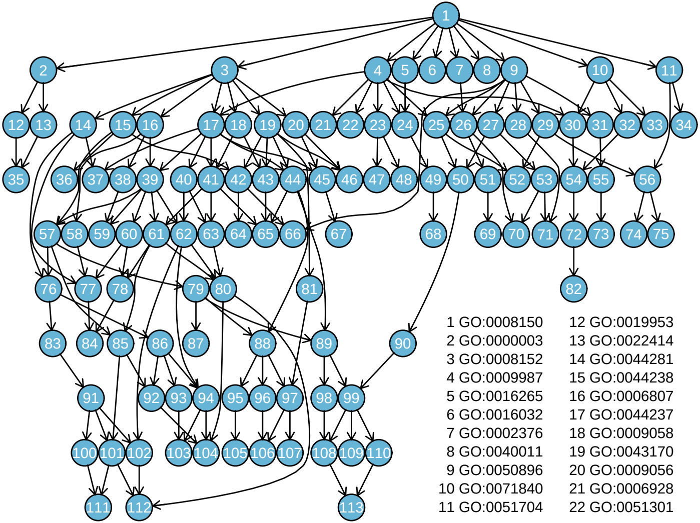
Generalized Weighted Bonferroni Testing



- $\sum_{i=1}^{m} a_i \leq a$
- $0 \leq g_{ij} \leq 1$, $g_{ii} = 0$, and $\sum_{k=1}^{m} g_{ik} \leq 1$ for all $i, j = 1, \ldots, m$.

Contents

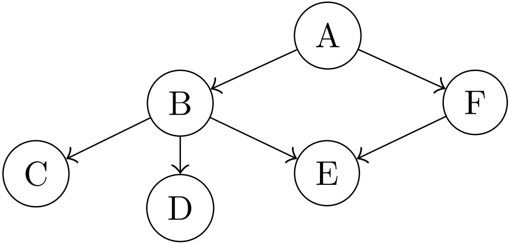
2. FWER Control within GO Graphs



1	GO:0008150	12	GO:0019953
2	GO:0000003	13	GO:0022414
3	GO:0008152	14	GO:0044281
4	GO:0009987	15	GO:0044238
5	GO:0016265	16	GO:0006807
6	GO:0016032	17	GO:0044237
7	GO:0002376	18	GO:0009058
8	GO:0040011	19	GO:0043170
9	GO:0050896	20	GO:0009056
10	GO:0071840	21	GO:0006928
11	GO:0051704	22	GO:0051301

2. FWER Control within GO Graphs

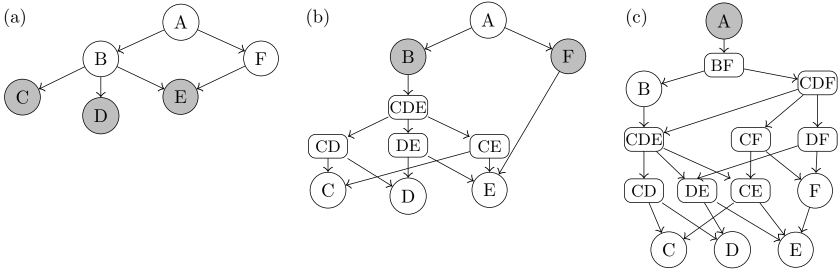
(Toy) Example GO Graph



2. FWER Control within GO Graphs

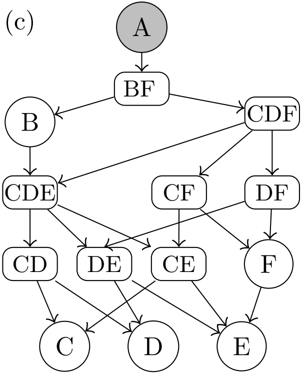
Focus Level Method (Goeman and Mansmann 2008)

- Applies a top-down and a bottom-up approach.

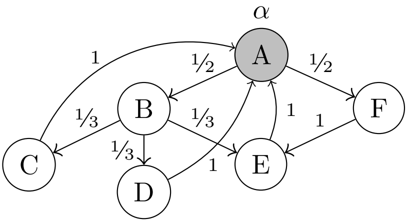


2. FWER Control within GO Graphs

Focus Level Method



Short Focus Level



2. FWER Control within GO Graphs

Table: Summary of power calculations for Simulation 1.

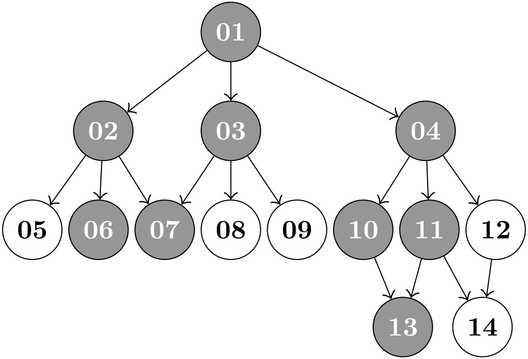
		Node						Mean Computation Time (sec)
n	Method	A	B	F	C	D	E	
5	FL	0.447	0.428	0.132	0.142	0.135	0.130	0.426134
	SFL	0.447	0.366	0.120	0.092	0.083	0.122	0.001778
20	FL	0.574	0.567	0.180	0.186	0.192	0.179	0.102097
	SFL	0.574	0.552	0.178	0.184	0.188	0.179	0.001789
100	FL	0.642	0.635	0.202	0.220	0.207	0.201	0.355848
	SFL	0.642	0.623	0.201	0.217	0.204	0.201	0.001793

FL: Focus Level

SFL: Short Focus Level

2. FWER Control within GO Graphs

Simulation 2



(The closure of this graph contains 574 nodes.)

2. FWER Control within GO Graphs

Table: Results of the power analysis under Simulation 2.

	GO:01	GO:02	GO:03	GO:04	GO:06	GO:07	GO:10	GO:11	GO:13
FL	0.995	0.968	0.890	0.462	0.512	0.872	0.380	0.399	0.344
SFL	0.995	0.988	0.952	0.543	0.837	0.949	0.489	0.476	0.445

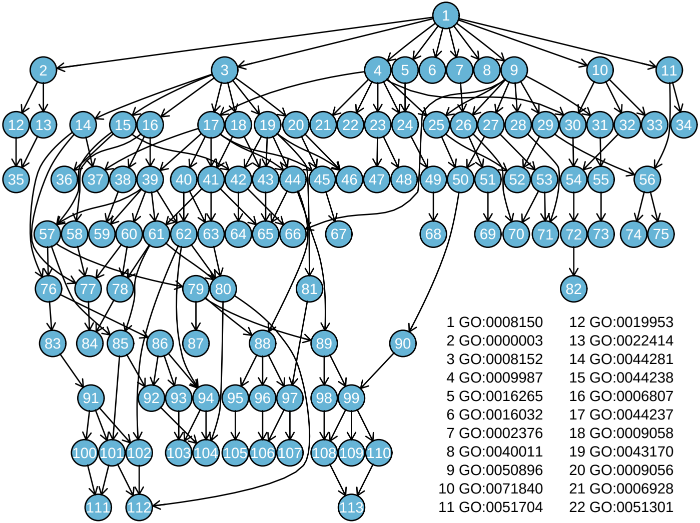
FL: Focus Level

SFL: Short Focus Level

	Computation Time
FL	3:42:938
SFL	0:00:015

2. FWER Control within GO Graphs

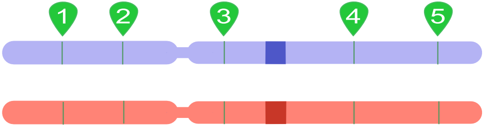
(Computation took 3 minutes and 23 seconds. Original graph contained 5,687 nodes.)



1	GO:0008150	12	GO:0019953
2	GO:0000003	13	GO:0022414
3	GO:0008152	14	GO:0044281
4	GO:0009987	15	GO:0044238
5	GO:0016265	16	GO:0006807
6	GO:0016032	17	GO:0044237
7	GO:0002376	18	GO:0009058
8	GO:0040011	19	GO:0043170
9	GO:0050896	20	GO:0009056
10	GO:0071840	21	GO:0006928
11	GO:0051704	22	GO:0051301

Contents

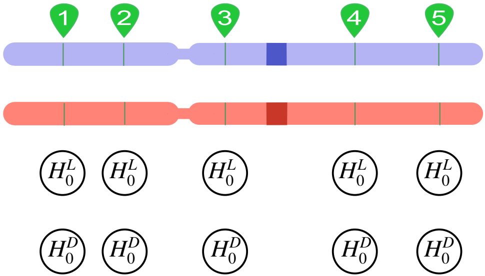
3. FWER Control in LD QTL Mapping



Want to know if:

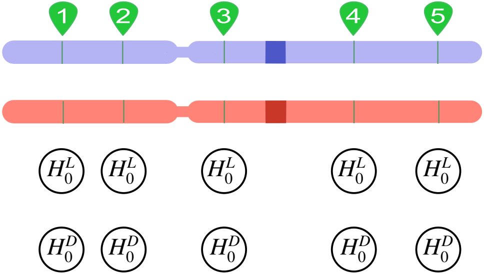
- a QTL exists,
- the QTL is linked to any markers.

3. FWER Control in LD QTL Mapping



- H_0^L: No QTL exists.
- H_0^D: QTL is unlinked with marker.

3. FWER Control in LD QTL Mapping



$$L(p, q, D, \mu_1, \ldots, \mu_G, \sigma | Y, M) = \prod_{i=1}^{n} \sum_{g=1}^{G} \omega_{g|M_i}(p, q, D) f(Y_i | \mu_g, \sigma)$$

(D is not identifiable under H_0^L)

3. FWER Control in LD QTL Mapping

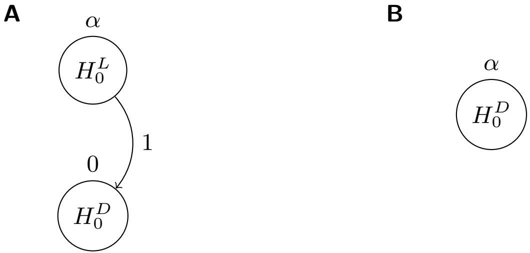


Figure: **A**) Demonstration of the GBA testing scheme for a single marker. **B**) The updated graph after finding H_0^L significant.

3. FWER Control in LD QTL Mapping

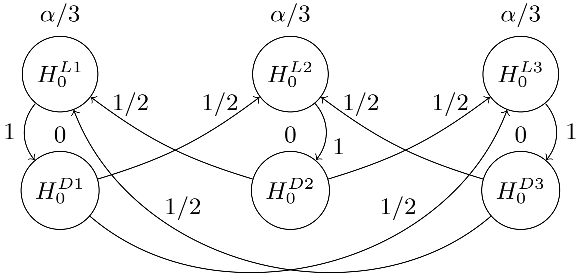


Figure: Demonstration of the hierarchy of the GBA testing scheme for three markers.

3. FWER Control in LD QTL Mapping

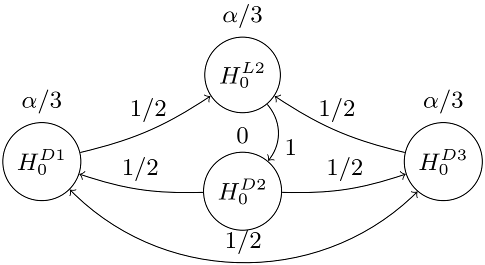


Figure: Demonstration of the GBA testing scheme for three markers assuming that hypotheses H_0^{L1} and H_0^{L3} from the initial graph in Figure 2 are rejected.

3. FWER Control in LD QTL Mapping

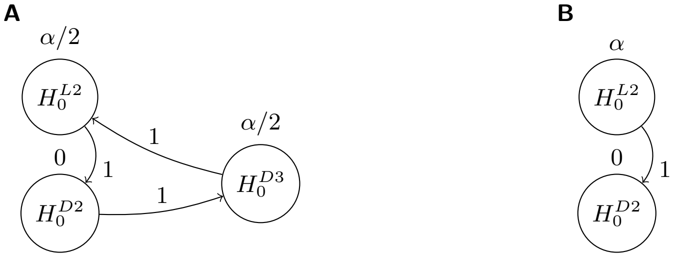


Figure: **A**) The updated graph from Figure 3 assuming the hypothesis H_0^{D1} of Figure 3 is rejected at the $\alpha/3$-level. **B**) Graph resulting from the rejection of the hypothesis H_0^{D3} at the $\alpha/2$-level.

3. FWER Control in LD QTL Mapping

Conditions under which the GBA simplifies to an IUT (+Holm).

- Let H_0^U denote the union hypothesis $H_0^L \cup H_0^D$.
- Let P^U denote the p-value for the IUT of H_0^U.
- Let k denote the marker with $\arg\min P_i^U < \alpha/m$.
- Then,

$$mp_k^L = \tilde{P}_k^L \leq \tilde{P}_k^D = m\max\{p_k^L, p_k^D\} = mP_k^M = \tilde{P}_k^M \leq \alpha$$

- where m is number of markers, p_k^L and p_k^D are raw p-values for marker k, $\tilde{P}_k$ are GBA adjusted p-values for marker k, and P_k^M denotes the raw IUT p-value for marker k.

3. FWER Control in LD QTL Mapping

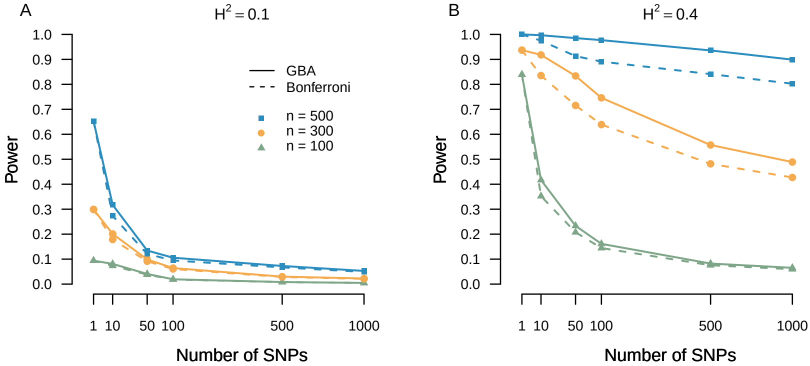


Figure: Power comparison between the graphical Bonferroni adjustment (GBA) and standard Bonferroni adjustment under different sample size, number of SNPs, and heritability (**A**: $H^2 = 0.1$, **B**: $H^2 = 0.4$).

3. FWER Control in LD QTL Mapping

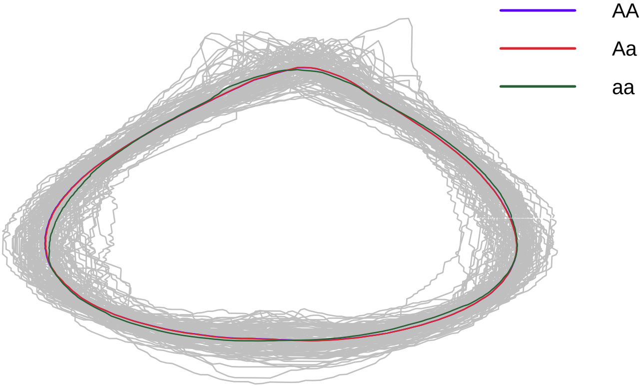


Figure: The control of leaf shape for different genotypes (AA, Aa, aa) of the QTL identified by marker 1 on PC 4.

3. FWER Control in LD QTL Mapping

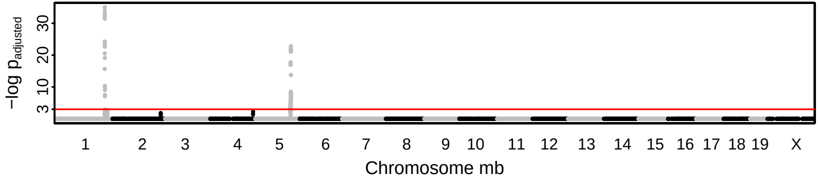


Figure: The negative log of the GBA-adjusted p-values for H_0^D for each SNP in the mouse HDL cholesterol QTL mapping project.

▸ Link

Contents

4. QTL Mapping: Hypotheses and Approaches

- H_A: a linked QTL.

$$L(p, q, D, \mu_1, \ldots, \mu_G, \sigma | Y, M) = \prod_{i=1}^{n} \sum_{g=1}^{G} \omega_{g|M_i}(p, q, D) f(Y_i | \mu_g, \sigma). \tag{1}$$

- H_0^2: an unlinked QTL.

$$L(q, \mu_1, \ldots, \mu_G, \sigma | Y) = \prod_{i=1}^{n} \sum_{g=1}^{G} \omega_g(q) f(Y_i | \mu_g, \sigma) \tag{2}$$

- H_0^1 : no QTL.

$$L(\mu, \sigma | Y) = \prod_{i=1}^{n} f(Y_i | \mu, \sigma). \tag{3}$$

4. QTL Mapping: Hypotheses and Approaches

Test for association between QTL and phenotype Y.

H_0^L: $\mu_1 = \mu_2 = \mu_3 \equiv \mu$ vs

H_1^L: one of the equalities above does not hold.

4. QTL Mapping: Hypotheses and Approaches

Test for linkage between SNP and QTL.

$$H_0^D : D = 0 \quad \text{vs} \quad H_1^D : D \neq 0. \tag{4}$$

$$\chi_D^2 = \frac{n\hat{D}^2}{\hat{p}(1-\hat{p})\hat{q}(1-\hat{q})} \sim \chi_1^2 \tag{5}$$

4. QTL Mapping: Hypotheses and Approaches

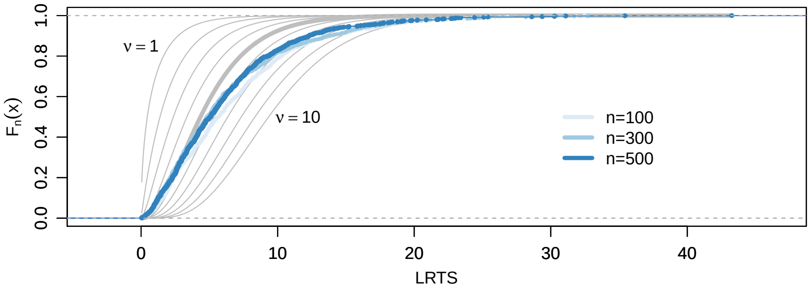


Figure: Likelihood Ratio Test of H_0^1 against H_A for synthetic data simulated under the null hypothesis of no QTL, H_0^1.

4. QTL Mapping: Hypotheses and Approaches

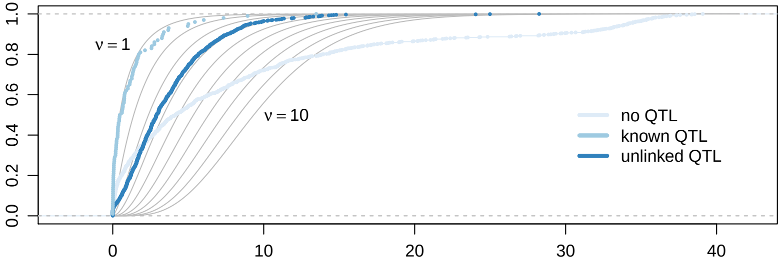


Figure: The empirical cumulative density functions corresponding to the test of $D = 0$ for three scenarios.

4. QTL Mapping: Hypotheses and Approaches

The (Bivariate) Null Kernel Method

1. Simulate s data sets, each of size n, based on the model assumptions of the joint hypothesis H_0.
2. Calculate T_i and U_i for $i = 1, \ldots, s$.
3. Estimate the joint density $\hat{f}$ of T and U using a kernel density estimation technique on the T_i and U_i.
4. Compute the cdf $\hat{F}$ of $\hat{f}$ by $\hat{F}(c) = \int_{A(c)} \hat{f}$, where $A(c) = \{(t,u) | \hat{f}(t,u) \geq c\}$.
5. The joint p-value for the calculated statistics $\hat{t}$ and $\hat{u}$ can then be obtained by the formula $p = 1 - \hat{F}(\hat{f}(\hat{t}, \hat{u}))$.

4. QTL Mapping: Hypotheses and Approaches

The Null Kernel Method vs. Hotelling's T^2 (bivariate test of location)

$$T^2 = n(\bar{\mathbf{X}} - \mu_0)' S^{-1} (\bar{\mathbf{X}} - \mu_0) \sim 2(n-1)/(n-2) F_{2,n-2} \qquad (6)$$

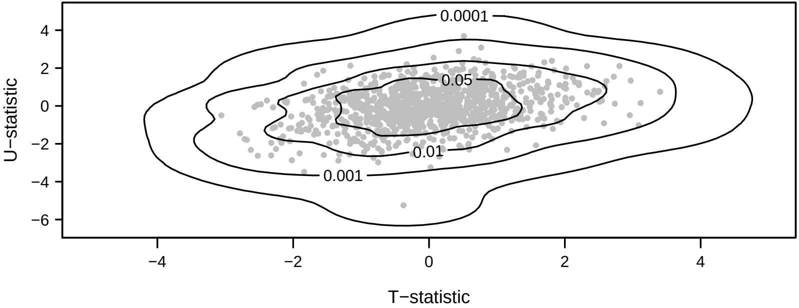


Figure: Visualization of the Null Kernel method as applied to a sample of 1,000 T and U statistics simulated under the bivariate normal null distribution with zero mean, unit variances, and covariances of 0.3.

4. QTL Mapping: Hypotheses and Approaches

The Null Kernel Method vs. Hotelling's T^2 (bivariate test of location)

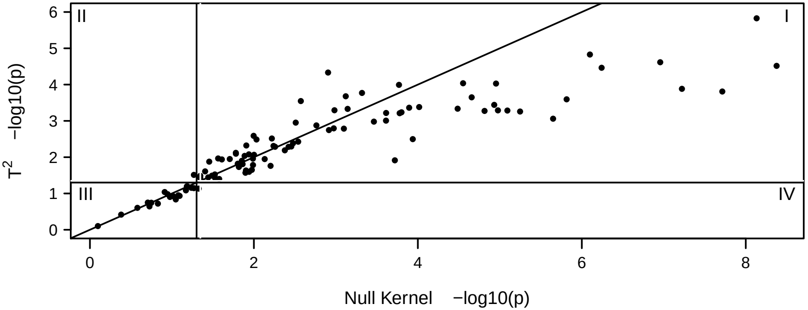


Figure: Comparison of P-values ($-\log_{10}(p)$) obtained from either the Null Kernel method or Hotelling's T^2 test.

4. QTL Mapping: Hypotheses and Approaches

The Null Kernel Method vs. Hotelling's T^2 (bivariate test of location)

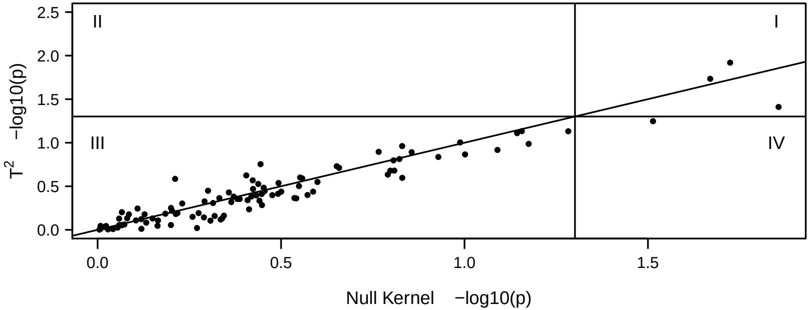


Figure: The $-\log_{10}$ of the P-values from the Null Kernel and Hotelling's T^2 methods for data simulated consistent with the null hypothesis.

4. QTL Mapping: Hypotheses and Approaches

QTL mapping Simulation Study

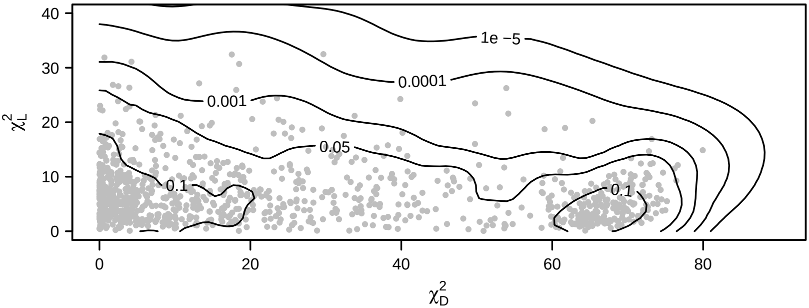


Figure: Visualization of the Null Kernel estimated (null) density for the bivariate data corresponding to the test of H_0^D, χ^2_D, and H_0^L, χ^2_L.

4. QTL Mapping: Hypotheses and Approaches

QTL mapping Simulation Study

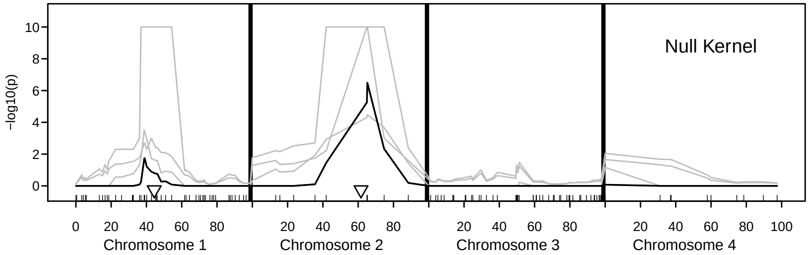


Figure: The resulting adjusted P-values from each of the permutation, simulation, and theoretical approaches against the results of the Null Kernel method.

4. QTL Mapping: Hypotheses and Approaches

Mice HDL QTL mapping study

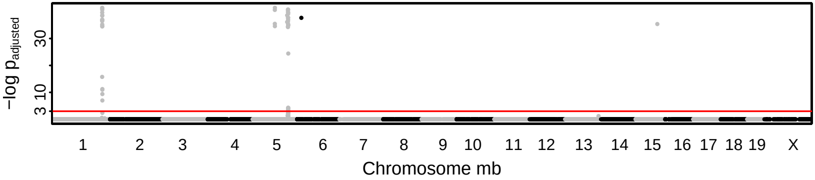


Figure: The negative log of the Holm adjusted P-values for the Null Kernel approach.

4. QTL Mapping: Hypotheses and Approaches

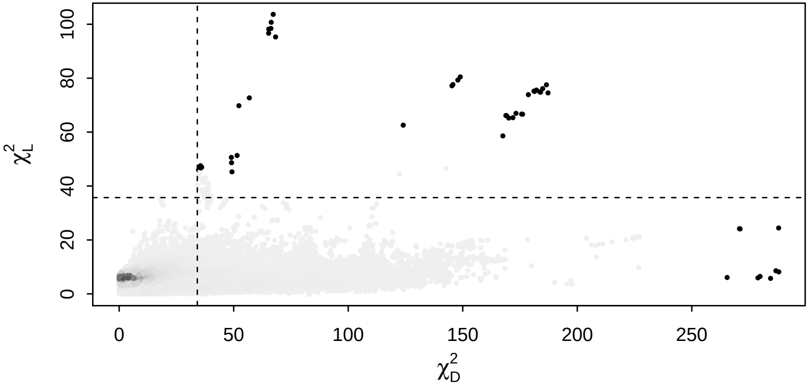


Figure: The joint plot of the observed test statistics for the mouse HDL QTL mapping data.

Contents

5. Discussion

“Professional statisticians... bear an obligation to offer alternatives (or entirely new approaches) that meet real needs and are practical as well.” (J. W. Tukey, as quoted in Benjamini and Braun 2001)

5. Discussion

1. Introduction
2. FWER Control within Gene Ontology Graphs
    - Extended GBA methods to Restricted Hypotheses (Theorem 1).
    - Introduced the Short Focus Level method (code in: `mvGST`).
    - Quantified the computational advantage.
3. A Power Improving Multiplicity Correction for Large-Scale SNP Selection in LD Based QTL Mapping
    - Introduced a GBA approach for LD based QTL mapping.
    - Protects model identifiability and strong FWER control.
    - Quantified the power increase numerically and practically.
4. QTL Mapping: Hypotheses and Approaches
    - Detailed problems of χ^2 assumptions in LD based QTL mapping.
    - Introduced the Null Kernel method.
    - Showed power and computational advantages of the NK method.
5. Discussion

Acknowledgements

This work was supported by

- Utah Agricultural Experiment Station (UAES) project number UTA01062, associated with the W2112 multi-state project “Reproductive Performance in Domestic Ruminants”
- Utah State University VPR Research Catalyst Grant.