Multiple hypothesis Testing with applications to genomics

B. Tallur

Séminaire Symbiose

26 septembre 2008

Plan de l'exposé

- Test statistique simple
- Hypothèses nulle et alternative, région critique, erreurs de type1 et type2, p-value, puissance de test
- Qu'est-ce que les tests multiples?
- Problèmes posés en génomique en termes de tests multiples

Un test staistique classique

- Hypothèse nulle: hypothèse qui sera rejetée ou non en appliquant une règle de décision
- Hypothèse alternative: c'est elle qu'on accepte en cas de rejet de l'hypothèse nulle
- Région critique: définit la règle de décision en fonction de la valeur observée d'une statistique
- Risque d'erreur de type1 α: probabilité de rejeter une hypothèse nulle qui, en réalité, est vraie
- Risque d'erreur de type2 β: probabilité de ne pas rejeter une hypothèse nulle qui, en réalité, est fausse
- Puissance du test (1- β): probabilité de rejeter une hypothèse nulle qui, en réalité, est fausse

Un Exemple: Lancer de pièce

On veut tester si une pièce de monnaie est « équilibrée »

Hypothèse nulle à tester: la probabilité d'obtenir « pile » =0,5

Méthode classique consiste à fixer un seuil pour α et définir la région critique en fonction de ce risque

Si on fixe α=0,05 on rejette l'hypothèse injustement avec une probabilité de 5%

En répétant le même test 100 fois, on sait que théoriquement on va conclure 5 fois que la pièce n'est pas équilibrée sachant qu'en vérité elle l'est

Lancer de pièce (suite)

- On lance la pièce 10 fois et on décide de conclure que la pièce est truquée si on obtient 9 fois ou plus de pile.
- on a Prob[9 pile ouplus]=0,0107 (c'est la p-value associée à ce test)
- Avec un seuil de 0,05 on rejette notre hypothèse nulle

On lance 100 pièces

- Pour une pièce donnée, le risque d'erreur de type 1 est bien 0,0107
- Mais la proba que une des 100 pièces (peu importe laquelle) soit rejetée sera plus élevée
- Supposant toutes les pièces équilibrées, Pr[toutes les pièces soient acceptées] $=(1-0,0107)^{100} = 0,34$
- Dans ce cas, la proba de conclure injustement qu'une pièce est truquée est donc plus élevée que dans le cas d'un test classique
- Prob[rejeter au moins une pièce injustement] = 1-0,34 = 0,66

Genomic Context

- Microarray experiments
- Expression levels of thousands of genes are measured simultaneously
- An important question: identification of differentially expressed genes, i.e., genes whose expression levels are associated with a response or a covariate of interest

Covariates and responses

Covariates may be:

- polytomous (nominal) e.g., cell type, treatment/control status, drug type
- Continuous e.g., dose of a drug, time

Responses could be, for example, censored survival times or other clinical outcomes

Differential expression and hypothesis testing

- The biological question of differential expression can be restated statistcally as a problem of *multiple hypothesis testing*: the simultaneous test for each gene, of the null *hypothesis of no association* between the expression levels and the responses or covariates

Two types of errors

- *False positive*, or Type I error: declaring that a gene is differentially expressed when it is not
- *False negative*, or Type II error: test fails to identify a truly differentially expressed gene

Problems

- When many hypotheses are tested and each test has a specified Type I error probability, the chances of committing some Type I errors increases with the number of hypotheses

Requirements

- Need to define an appropriate Type I error rate and devising powerful multiple testing procedures that control this error rate and account for the joint distribution of the test statistics

Microarray Experiments

- *m* genes (variables or features)
- *n* samples (observations)
- For each sample a response or covariate is recorded
- The data for sample i consist of a response y_i and a gene expression profile

$X_i = (x_{1i}, \ldots, x_{mi})$

The data

- The expression data are stored in a mxn matrix $\mathbf{X}=(x_{ji})$ with rows corresponding to genes and columns to samples

The pairs (X_i, y_i) are considered as a random sample from a population of interest

- The goal is to infer about the population, specifically, to test hypotheses concerning the joint distribution of $\mathbf{X}=(x_1, \ldots., x_m)$ and Y

Differential expression

- *Simultaneous* test for each gene j
- *Null hypothesis*: H_j no association between X_j and Y (in some cases, more specific null hypothesis may be of equal mean expression levels in two populations of cells as opposed to identical distributions)

A standard approach

Consists of two steps

- Computing a test statistic T_j for each gene j
- Applying a *multiple testing procedure* to determine which hypotheses to reject while *controlling a suitably defined Type I error*
- H_j is rejected based on astatistic T_j

Set -up

- *m* null hypotheses Hj, j=1, ...,m
- Let R be the number of rejected hypotheses

	Not rejected	Rejected	
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	m_1
Total	m-R	R	m

Set-ut (2)

- R is an observable random variable
- S, T, U and V are unobservable r.v.'s
- Ideally, one would like to minimize V (the number of false positives or type I errors) and T (the number of false negatives or type II errors)

A univariate case

- While testing a single hypothesis Hj, the procedure is: Specify level α for the Type I error rate and seek tests which minimize Type II error rate i.e. maximize power, within the class of tests with Type I error rate $\leq \alpha$.

$$\Pr\left(\left|T_j\right| \geq c_\alpha / H_j\right) \leq \alpha$$

Where c_α is the critical value

Generalization of Type I error rate to multiple testing situation

- PCER (Per-Comparison Error Rate):

$PCER = E(V)/m$

- PFER (Per-Family Error Rate):

$PFER = E(V)$

- FWER (Family-wise Error Rate):

$FWER = Pr(V \geq 1)$

- FDR (False Discovery Rate):

$FDR = E(Q)$ with $Q = V/R$ if $R>0$ and $Q=0$ if $R=0$.

i.e. FDR is the expected proportion of type I errors among rejected hypotheses

Type I Error rates (2)

- These error rates are defined under the true and unknown data generating distribution of $\mathbf{X}=(x_1, \ldots., x_m)$ and Y
- They depend upon which specific subset

$$\Lambda_0 \subseteq \{1,2,...,m\}$$

Of null hypotheses is true for this distribution

Strong and weak control of Type I error rates

- *Strong control* refers to the control of Type I error rate for *any subset* $\Lambda_0 \subseteq \{1,2,...,m\}$

of true null hypotheses

- *Weak control* refers to the control of Type I error rate only *when all the null hupotheses are true* i.e. under the complete null hypothesis

$$H_0^C = \bigcap_{j=1}^{m} H_j$$

- In general the weak control is unsatisfactory because the complete null hypothesis is not realistic.

Coparison of Type I error rates

- In general, for a given multiple testing procedure,

$PCER \leq FWER \leq PFER$

Power of the test

- Within the class of multiple testing procedures that control a given Type I error rate at a level α, one seeks procedures that maximize power i.e. minimize Type II error rate

Generalization of power to multiple testing

The three common definitions are:

1. $Pr(S \geq 1) = Pr(T \leq m_1 - 1)$ i.e. probability of rejecting at least one false hypothesis;
2. $E(S)/m_1$ = average power or average probability of rejecting false null hypotheses;
3. Probability of rejecting all false null hypotheses, $Pr(S = m_1) = Pr(T = 0)$

When the family of tests consists of pairwise mean comparisons, thes quantities are called *any-pair* power, *per-pair* power and *all-pairs* power respectively

p-values

- consider a single hypothesis, say H_1, and a test statistic T_1. The Critical region S_α of size α is such that
- $Pr(T_1 \in S_\alpha \mid H_1) = \alpha$
- the p-value p_1 associated with the test of H_1 is the level at which the hypothesis would be just rejected, given t_1
- $p_1 = \inf\{\alpha: t_1 \in S_\alpha\} = Pr\{|T_1| \geq |t_1| \mid H_1\}$

Adjusted p-values (multiple testing situation)

- Let t_j and $p_j = Pr\{|T_j| \geq |t_j| \mid H_j\}$ denote respectively, the tset statistic and the unadjusted p-value for hypothesis H_j (gene j), j=1,2,...,m.
- Given any test procedure, the adjusted p-value corresponding to the single hupothesis H_j can be defined as the nominal level of the entire test procedure at which H_j would just be rejected, given the values of all test statistics

Types of multiple testing procedures

- Single-step procedures: each hypothesis is evaluated using a critical value that is independent of the results of the tests of other hypotheses

Bonferroni procedure rejects any hypothesis with unadjusted p-value $\leq \alpha/m$

The corresponding adjusted p-values are given by

$$\tilde{p}_j = \min(mp_j, 1)$$

(strong control of FWER)

Single-step procedures (2)

- Šidák procedure: it is exact for protecting the FWER under complete null hypothesis, when the unadjusted p-values are independently distributed as U[0,1]
- Adjusted p-values are given by

$$\tilde{p}_j = 1 - (1 - p_j)^m$$

- Single-step procedures are simple to implement but they tend to be conservative

Step-down procedures

- Let $p_{r_1} \le p_{r2} \le \ldots \le p_{rm}$ denote the observed ordered unadjusted p-values and let $H_{r_1} \le H_{r2} \le \ldots \le H_{rm}$ denote the corresponding null hypotheses.
- *The Holm's procedure* (strong control of FWER at level α):
- Define $j^* = \min\left\{j : p_{r_j} > \alpha/(m-j+1)\right\}$
- And reject hypotheses H_{r_j} ,for j=1,2,..j^*-1
- If no such j^* exists, reject all hypotheses

Step-down procedure

- Adjusted p-value for Holm's procedure is

$$\tilde{p}_{r_j} = \max_{k=1,2,\ldots j} \left\{ \min\left((m-k+1) p_{r_k}, 1 \right) \right\}$$

- Holm's procedure is less conservative than the standard Bonferroni procedure
- Other step-down procedures
  - Step-down Šidák , minP, maxT

Step-up procedures

- Hypotheses that correspond to the least significant test statistics are considered successively, with further tests dependent on the outcomes of the earlier ones. As soon as one hypothesis is rejected, all remaining hypotheses are rejected

Step-up procedures (2)

- Hochberg adjusted p-values are:

$$\tilde{p}_{r_j} = \min_{k=j,\ldots m} \left\{ \min\left((m-K+1)p_{r_k}, 1 \right) \right\}$$

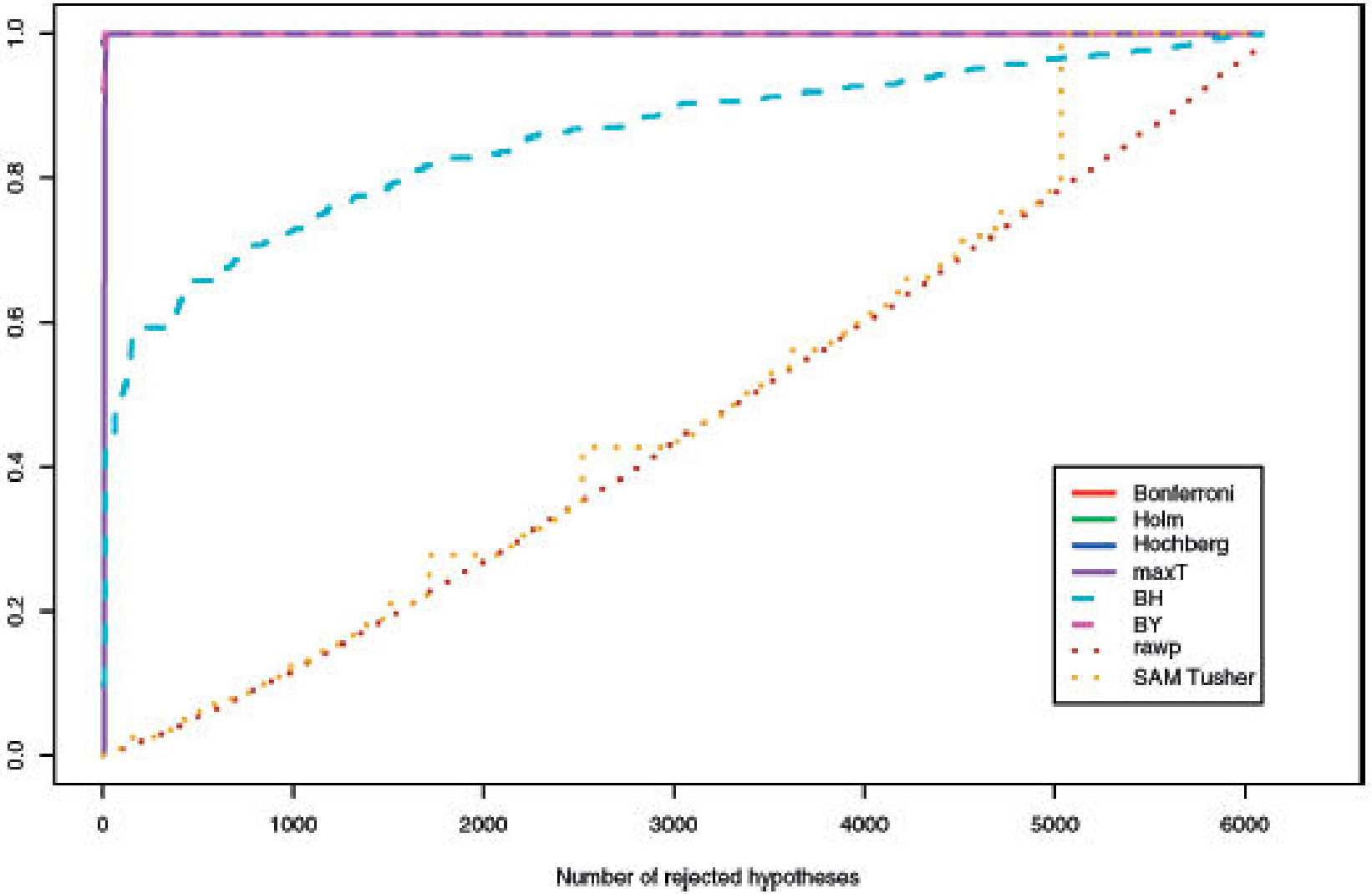
- There are other step-up prcedures such as
- Benjamini and Hochberg (1995)
- Benjamini and Yekutieli (2001)

Estimation of adjusted p-values

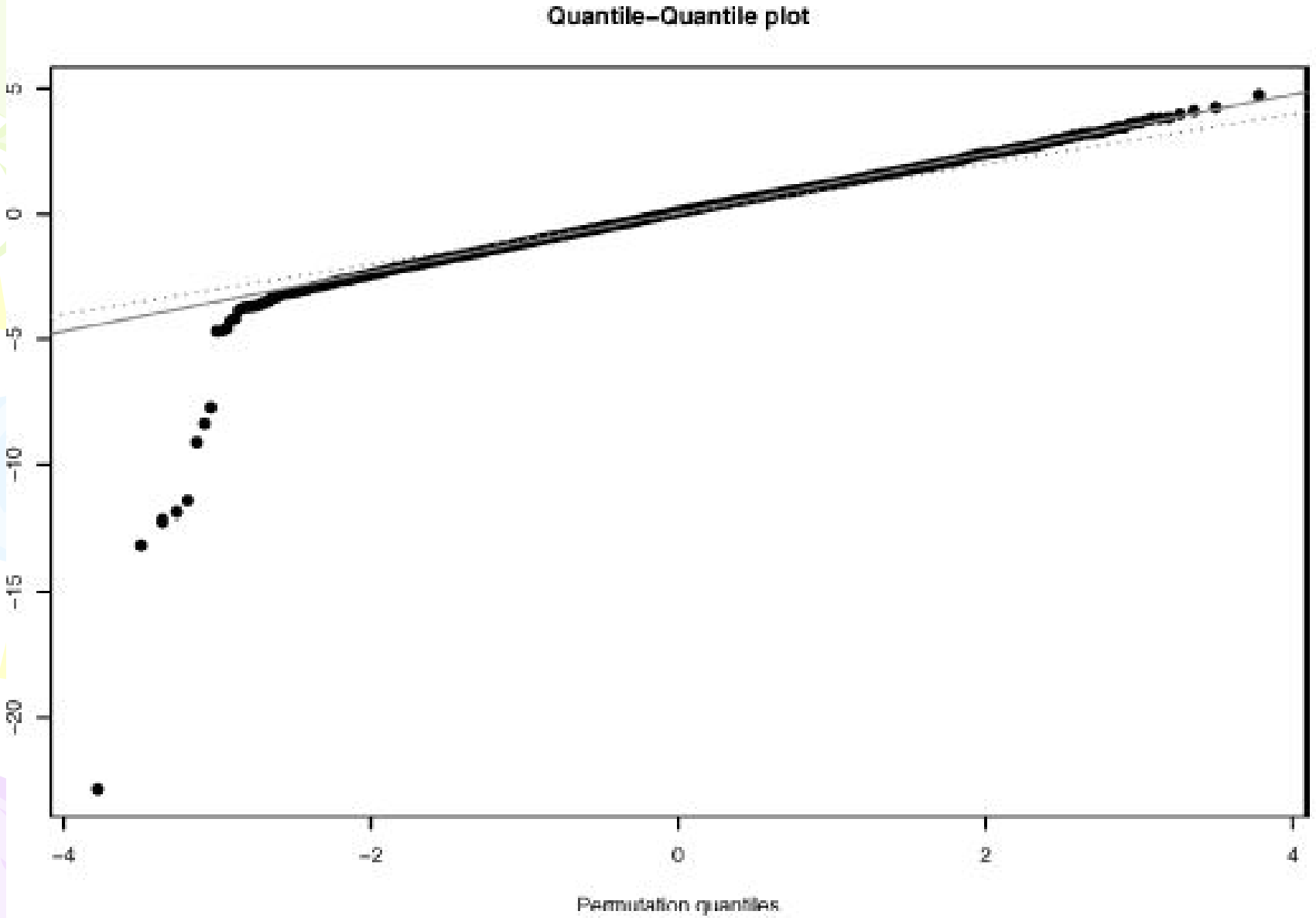
- Permutation methods are proposed to estimate the marginal or joint distribution of the test statistics.
- (I wil not talk about these algorithms today, may be next time…)

Comparison on real and simulated data

- Dudoit et al. have compared different multiple testing procedures on several real and simulated data
- Apolipoprotein AI experiment of Callow et al. (2000) as part of the study of lipid metabolism in mice

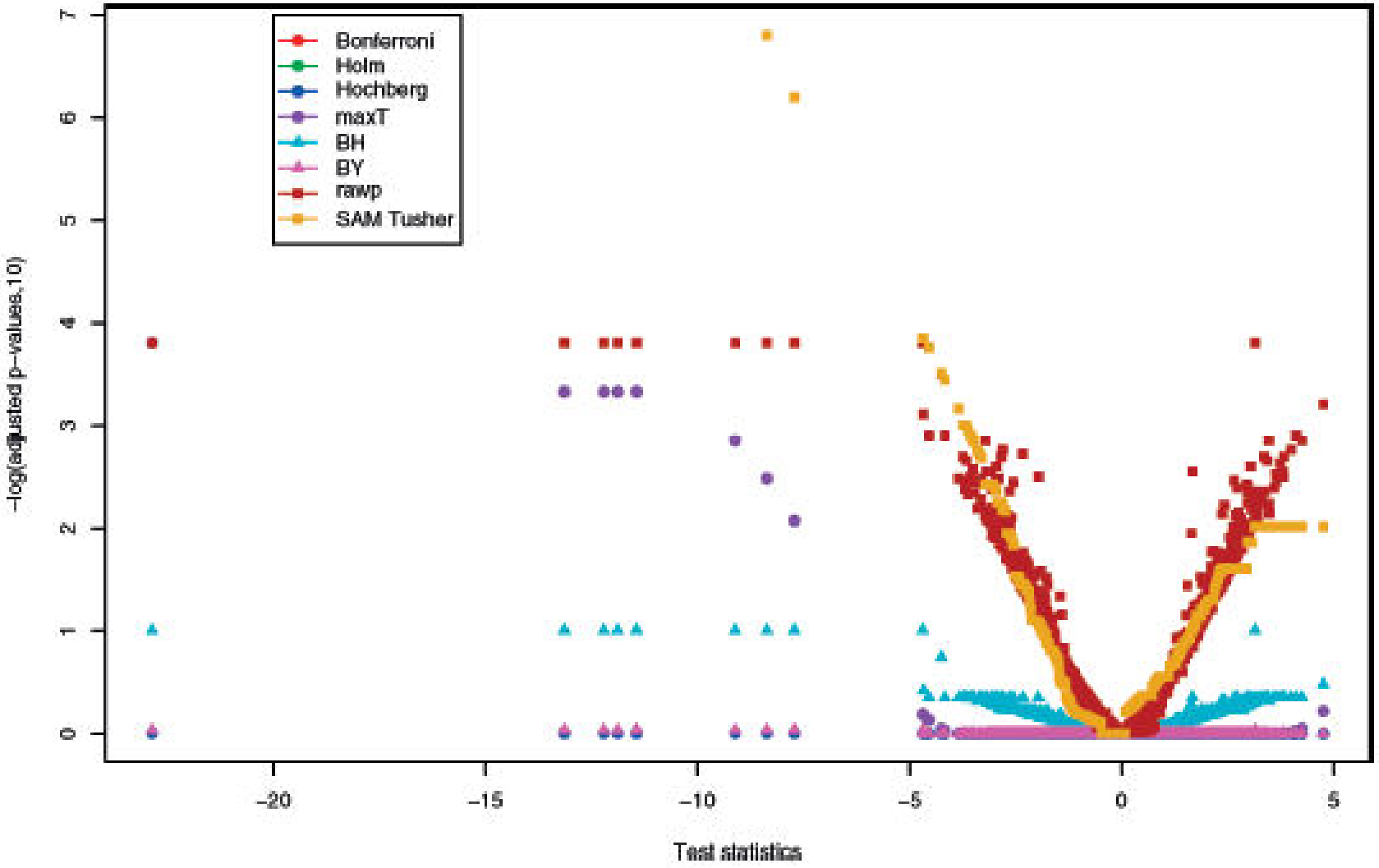

Bonferroni
Holm
Hochberg
maxT
BH
BY
rawp
SAM Tusher
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
6000
Number of rejected hypotheses

(a)

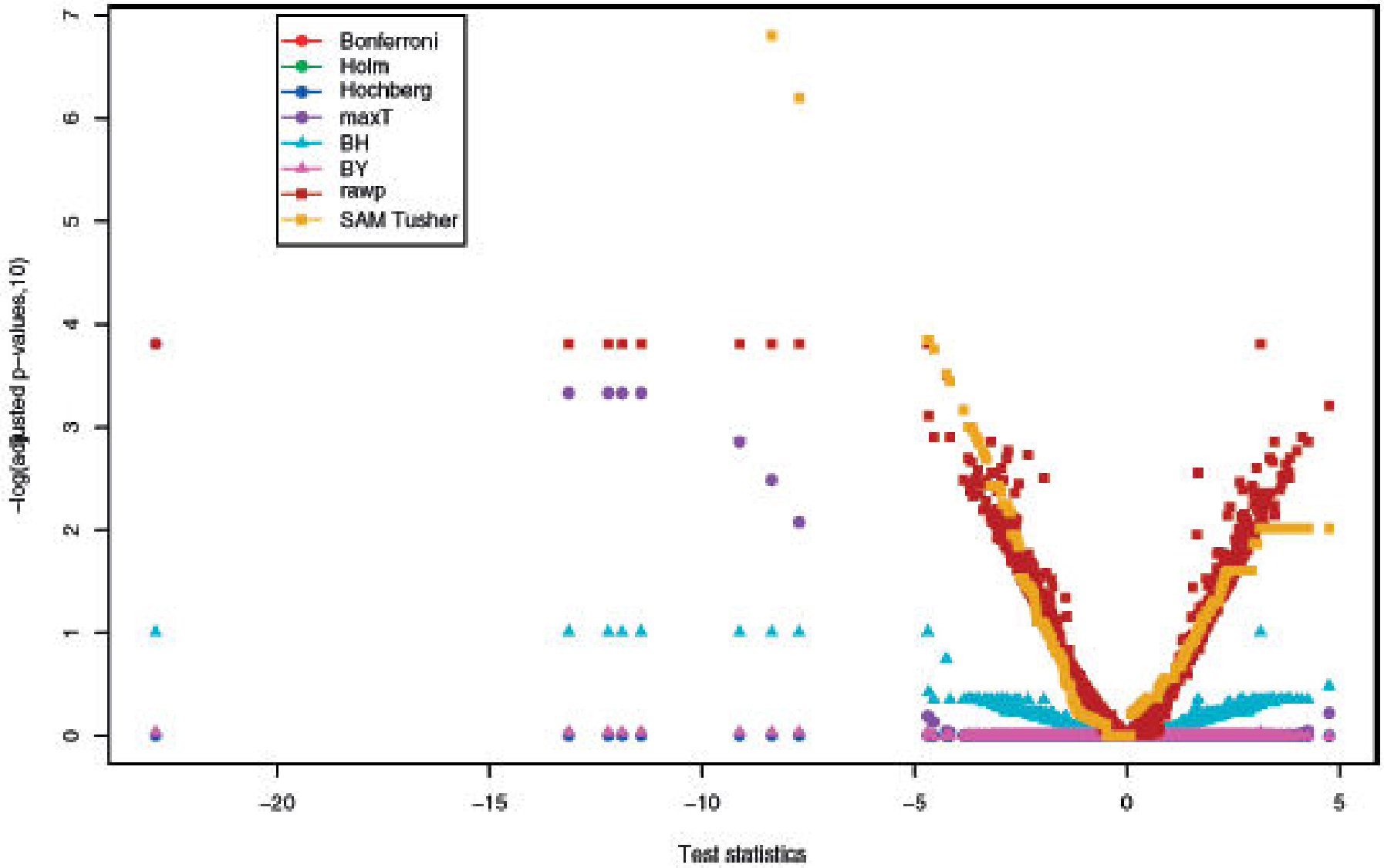

Quantile–Quantile plot
5
0
−5
−10
−15
−20
−4
−2
0
2
4
Permutation quantiles

(d)

(a)


Bonferroni
Holm
Hochberg
maxT
BH
BY
rawp
SAM Tusher
−log(adjusted p−values,10)
0
1
2
3
4
5
6
7
−20
−15
−10
−5
0
5
Test statistics

(c)


Bonferroni
Holm
Hochberg
maxT
BH
BY
rawp
SAM Tusher
−log(adjusted p−values,10)
0
1
2
3
4
5
6
7
−20
−15
−10
−5
0
5
Test statistics

(c)

References

- Sandrine Dudoit, Juliet Popper Shaffer and Jennifer C Boldrik; « Multiple hypothesis testing in microarray experiments »; Statistical Science, 2003, vol. 18, No 1, 71 – 103
- Merrill D. Birkner, Katherine S. Pollard, Mark J. van der Laan, Sandrine Dudoit; « Multiple testing procedures andapplications to genomics »; 2005, paper 168, university of California, Berkley Division of Biostatistics working papers series
- Shaffer, J.P.; « Multiple hypothesis testing: A review »; 1986, Annual review of psychology, 46, 561 – 584.

References

- Yoav. Benjamini, Yosef Hochberg
- Controlling false discovery rate: A practical and powerful approach to multiple testign, JRSSS (1995)
- J.D Storey, R. Tibshirani

Statistical significance for genomewide studies, PNAS, Aug 5, 2003