Multiple Hypothesis Testing in Microarray Data Analysis

Sandrine Dudoit

jointly with

Mark van der Laan and Katie Pollard

Division of Biostatistics, UC Berkeley

www.stat.berkeley.edu/~sandrine

CBMB and QB3 Short Course: Analysis of Gene Expression Microarray Data
Genentech Hall Auditorium, Mission Bay, UCSF
November 15, 2003

©Copyright 2003, all rights reserved
Outline

• Motivation: Microarray data analysis.
• Multiple hypothesis testing framework.
• Test statistics null distribution.
• Single-step procedures.
• Control of gFWER via control of FWER.
• Step-down procedures.
• Application: Alizadeh et al. (2000) DLBCL microarray data; Golub et al. (1999) ALL/AML microarray data.
• Software: Bioconductor R `multtest` package.
• Ongoing work.
Motivation: Microarray data analysis

General question: Relate genome-wide microarray measures to biological and clinical covariates and outcomes. That is, make inference about the joint distribution of these variables.

— High-dimensional, unknown multivariate distributions.
— Large and complex parameter spaces.

Covariates: treatment, dose, time, demographic variables, occurrence of DNA sequence motifs, SNP genotypes, etc.

Outcomes: affectedness/unaﬀectedness, quantitative trait, tumor class, metastasis indicator, response to treatment, time to recurrence, patient survival, etc.

— Polychotomous or continuous.
— Censored or uncensored.
**Motivation: Microarray data analysis**

**Differential gene expression**: Identify genes whose expression levels are associated with a response or covariate of interest.

**Estimation**: Estimate parameter of interest for the joint distribution of the microarray measures and responses/covariates. Estimate variability of the estimators.

**Testing**: Test for each gene a null hypothesis concerning a gene-specific parameter of interest.

**E.g.** Parameters of interest include means, differences in means, correlations, interactions, regression parameters, survival probabilities.
Multiple hypothesis testing

• Large multiplicity problem: thousands of hypotheses are tested simultaneously!
  – Increased chance of false positives (i.e., Type I errors).
  – E.g., chance of at least one $p$–value $< \alpha$ for $m$ independent tests is $1 - (1 - \alpha)^m$ and converges to one as $m$ increases. For $m = 1,000$ and $\alpha = 0.01$, this chance is 0.9999568!
  – Individual $p$–values of 0.01 no longer correspond to significant findings.

• Use a multiple testing procedure (MTP) that accounts for the simultaneous test of multiple hypotheses by considering the joint distribution of the test statistics for each hypothesis.
Multiple hypothesis testing

- Define a suitable set of null hypotheses: Which subset of genes to consider? What is the parameter of interest?
- Compute an appropriate test statistic for each null hypothesis.
- Select a Type I error rate that corresponds to a suitable form of control of false positives for the particular application: e.g., FWER, PCER, FDR.
- Use an MTP and associated null joint distribution for the test statistics that provide the desired control of the Type I error rate under the true unknown data generating distribution.
- Use resampling methods to estimate the unknown (null) joint distribution of the test statistics.
- Report adjusted p–values for each hypothesis, which reflect the overall Type I error rate of the MTP.
Model

Data. Let $X_1, \ldots, X_n$ be $n$ independent and identically distributed (i.i.d.) random $d$-vectors, $X \sim P \in \mathcal{M}$, where $\mathcal{M}$ is a model (possibly non-parametric) for the data generating distribution $P$.

E.g. $X_i = (W_i, Z_i)$: gene expression measures $W_i$ and biological and clinical outcomes $Z_i$ for patient $i$, $i = 1, \ldots, n$.

The dimension of the data vector $X$ ($d$) is usually much larger than the sample size ($n$):

$$n \ll d.$$
Parameters of interest

Parameters. Functions of the unknown data generating distribution $P$: $\mu = (\mu(j) : j = 1, \ldots, m)$, where $\mu(j) = \mu_j(P) \in \mathbb{R}$ and typically $m \geq d$.

Our goal is to use the data, $X_1, \ldots, X_n$, to make inference, i.e., estimate and/or test hypotheses, concerning these unknown parameters.

Parameters can be means, differences in means, correlations, and refer to linear models, generalized linear models, survival models (e.g., Cox proportional hazards model), time-series models, dose-response models, etc.
Null hypotheses

General submodel null hypotheses. Defined in terms of a collection of \( m \) submodels, \( \mathcal{M}_j \subseteq \mathcal{M} \), for the data generating distribution \( P \).

\[
H_{0j} \equiv I(P \in \mathcal{M}_j)
\]

vs.

\[
H_{1j} \equiv I(P \notin \mathcal{M}_j), \quad j = 1, \ldots, m.
\]

Thus, \( H_{0j} \) is true, i.e., \( H_{0j} = 1 \), if \( P \in \mathcal{M}_j \) and false otherwise.

**E.g. 1.** \( H_{0j} \): gene \( j \) has equal mean expression levels in \( K \) different types of tumors.

**E.g. 2.** \( H_{0j} \): gene \( j \) is not associated with survival for a particular type of cancer.

**E.g. 3.** \( H_{0,(g_1,g_2)} \): the expression levels of gene pair \((g_1, g_2)\) are not correlated in a particular timecourse experiment.
Null hypotheses

True null hypotheses

\[ S_0 = S_0(P) \equiv \{ j : H_{0j} \text{ is true} \} = \{ j : P \in \mathcal{M}_j \}, \]
\[ m_0 \equiv |S_0|. \]

False null hypotheses

\[ S_0^c = S_0^c(P) \equiv \{ j : H_{0j} \text{ is false} \} = \{ j : P \notin \mathcal{M}_j \}, \]
\[ m_1 \equiv |S_0^c| = m - m_1. \]

E.g. \( S_0 \) is the set of non-differentially expressed genes and \( S_0^c \) is the set of differentially expressed genes.

Our goal is to estimate \( S_0 \) and \( S_0^c \), i.e., determines which genes belong to each of these sets.
Test statistics

Rejection decisions are based on an \( m \)-vector of test statistics, 
\[ T_n = (T_n(j) : j = 1, \ldots, m) \sim Q_n(P) = Q_n, \]
that are functions of the data, \( X_1, \ldots, X_n \).

We assume that large values of \( T_n(j) \) provide evidence against null hypothesis \( H_{0j} \), \( j = 1, \ldots, m \). For two-sided tests, one can take absolute values of the test statistics.

E.g. Reject \( H_{0j} \), i.e., declare gene \( j \) differentially expressed, if \( T_n(j) > c_j \), where \( c = (c_j : j = 1, \ldots, m) \) denotes an \( m \)-vector of possibly random cut-offs.

Notation. Use lower case letters (e.g., \( x, t_n, \tilde{p}_n(j) \)) for realizations of random variables (e.g., \( X, T_n, \tilde{P}_n(j) \)).
Test statistics

Single parameter null hypotheses.

\[ H_{0j} = I(\mu(j) \leq \mu_0(j)) \quad \text{vs.} \quad H_{1j} = I(\mu(j) > \mu_0(j)), \]
\[ j = 1, \ldots, m. \]

Difference statistics

\[ D_n(j) \equiv (\text{Estimator} - \text{Null Value}) = \sqrt{n}(\mu_n(j) - \mu_0(j)). \]

\textit{t}-statistics (i.e., standardized difference)

\[ T_n(j) \equiv \frac{\text{Estimator} - \text{Null Value}}{\text{Standard Error}} = \sqrt{n} \frac{\mu_n(j) - \mu_0(j)}{\sigma_n(j)}. \]
Multiple testing procedures

A multiple testing procedure (MTP) produces a set $S_n$ of rejected hypotheses that estimates $S^c_0$, the set of false null hypotheses

$$S_n = S(T_n, Q_0, \alpha) \equiv \{ j : H_{0j} \text{ is rejected} \} \subseteq \{1, \ldots, m\},$$

where the set $S_n$ (or $\hat{S}^c_0$) depends on

- the data, $X_1, \ldots, X_n$, through the test statistics $T_n$;
- a null distribution, $Q_0$, for the test statistics $T_n$ — used to compute cut-offs for these $T_n$ (or corresponding $p$-values);
- the nominal level $\alpha$ of the MTP, i.e., the desired upper bound for a suitably defined Type I error rate.

**E.g.** Set of genes identified as differentially expressed:

$$S_n \equiv \{ j : T_n(j) > c_j \},$$

where $c = (c_j : j = 1, \ldots, m)$ denotes an $m$-vector of possibly random cut-offs, computed under $Q_0$. 
Type I and Type II errors

In any testing problem, two types of errors can be committed.

- **Type I error** or **false positive**: \( S_n \cap S_0 \),
i.e., reject \((S_n)\) a true null hypothesis \((S_0)\).
  
  **E.g.** *Say gene is differentially expressed when in truth it isn’t.*

- **Type II error** or **false negative**: \( S_n^c \cap S_0^c \),
i.e., fail to reject \((S_n^c)\) a false null hypothesis \((S_0^c)\).
  
  **Cf.** *power: \( S_n \cap S_0^c \).*
  
  **E.g.** *Fail to identify a truly differentially expressed gene.*

Denote the number of Type I errors by \( V_n \equiv |S_n \cap S_0| \) and the corresponding cumulative distribution function (c.d.f.) by \( F_{V_n} \).
Type I and Type II errors

Null hypotheses

<table>
<thead>
<tr>
<th>Not rejected</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td>S_n^c \cap S_0</td>
</tr>
</tbody>
</table>

Null hypotheses

| False | \(|S_n^c \cap S_0^c|\) | \(|S_n \cap S_0^c|\) | \(U_n = |S_n^c \cap S_0^c|\) | \(m_1 = |S_0^c|\) |

| \(|S_n^c|\) | \(R_n = |S_n|\) | \(m\) |

Adapted from Benjamini & Hochberg (1995).
Type I error rates

Consider Type I error rates, $\theta(F_{V_n})$, that are functions, i.e., parameters, of the distribution $F_{V_n}$ of the number of false positives.

1. **Per-comparison error rate (PCER)**, or expected proportion of Type I errors among the $m$ tests,

   \[ PCER \equiv E[V_n]/m = \int vdF_{V_n}(v)/m. \]

2. **Per-family error rate (PFER)**, or expected number of Type I errors,

   \[ PFER \equiv E[V_n] = \int vdF_{V_n}(v). \]

3. **Median-based per-family error rate (mPFER)**, or median number of Type I errors,

   \[ mPFER \equiv \text{Median}(V_n) = F_{V_n}^{-1}(1/2). \]
Type I error rates

4. Family-wise error rate (FWER), or probability of at least one Type I error,

\[ \text{FWER} \equiv \Pr(V_n > 0) = 1 - F_{V_n}(0). \]

5. Generalized family-wise error rate (gFWER), or probability of at least \( k + 1 \) Type I errors, \( k = 0, \ldots, m_0 - 1 \),

\[ \text{gFWER} \equiv \Pr(V_n > k) = 1 - F_{V_n}(k). \]

When \( k = 0 \), the gFWER is the usual family-wise error rate, FWER.
Type I error rates

Make the following two assumptions for the mapping $\theta : F \rightarrow \theta(F)$, defining a Type I error rate as a parameter corresponding to a c.d.f $F$ on $\{0, \ldots, m\}$.

Monotonicity.

If $F_1 \geq F_2$, then $\theta(F_1) \leq \theta(F_2)$.

Uniform continuity. For two sequences of c.d.f’s, $\{F_n\}$ and $\{G_n\}$,

if $d(F_n, G_n) \rightarrow 0$, then $\theta(F_n) - \theta(G_n) \rightarrow 0$, as $n \rightarrow \infty$.

Here, given two c.d.f’s, $F_1$ and $F_2$ on $\{0, \ldots, m\}$, the distance measure $d$ is defined by $d(F_1, F_2) \equiv \max_{x \in \{0, \ldots, m\}} | F_1(x) - F_2(x) |$. 
Type I error rates

N.B. The false discovery rate (FDR) of Benjamini & Hochberg (1995) cannot be represented as a parameter $\theta(F_{V_n})$, because it also involves the distribution of the number of rejections, $R_n$.

The FDR is the expected proportion of Type I errors among the rejected hypotheses, i.e.,

$$FDR \equiv E[V_n/R_n],$$

with the convention that $V_n/R_n = 0$, if $R_n = 0$. 
Adjusted $p$-values

In an MTP, the adjusted $p$-value, $\tilde{P}_n(j) = \tilde{P}(j, T_n, Q_0)$, for null hypothesis $H_{0j}$, is defined as

$$\tilde{P}_n(j) \equiv \inf \{ \alpha \in [0, 1] : \text{Reject } H_{0j} \text{ at MTP level } \alpha \}.$$ 

E.g. Bonferroni for FWER control: $\tilde{P}_n(j) = \min(mP_n(j), 1)$, where $P_n(j)$ is the unadjusted $p$-value for $H_{0j}$.

- Can be defined for any Type I error rate: FWER, FDR, etc.
- Reflect the strength of the evidence against each null hypothesis in terms of the Type I error rate for the entire MTP.
- **Flexible**: The results of an MTP are provided for all levels $\alpha \Rightarrow$ do not need to choose the level ahead of time.
Adjusted \( p \)-values

We now have two representations for an MTP, in terms of

- **cut-offs** for the test statistics \( T_n(j) \), \( c_j = c_j(T_n, Q_0, \alpha) \):

\[
S(T_n, Q_0, \alpha) = \{ j : T_n(j) > c_j \};
\]

- **adjusted \( p \)-values**, \( \tilde{P}_n(j) = \tilde{P}(j, T_n, Q_0) \):

\[
S(T_n, Q_0, \alpha) = \{ j : \tilde{P}_n(j) \leq \alpha \}.\]
Null distribution

We are concerned with MTPs that provide (finite sample or asymptotic) control of a Type I error rate, \( \theta(F_{V_n}) \), at level \( \alpha \). That is,

\[
\theta(F_{V_n}) \leq \alpha \quad \text{[finite sample control]}
\]

\[
\limsup_{n \to \infty} \theta(F_{V_n}) \leq \alpha \quad \text{[asymptotic control]},
\]

where the parameter \( \theta(F_{V_n}) \) is defined under the true distribution \( Q_n(P) \) for the test statistics \( T_n \), i.e., the distribution for \( T_n \) that corresponds to the true underlying data generating distribution \( P \).
Null distribution

In practice, however, the true distribution $Q_n = Q_n(P)$, for the test statistics $T_n$ is unknown and estimated by a null distribution $Q_0$, in order to derive cut-offs for these test statistics $T_n$ (or corresponding adjusted $p$-values).

N.B. The choice of null distribution $Q_0$ is crucial, in order to ensure that (finite sample or asymptotic) control of the Type I error rate under the assumed $Q_0$ does indeed provide the required control under the true distribution $Q_n(P)$. 
Null distribution

For proper control of the Type I error rate, the assumed null distribution $Q_0$ must be such that

$$\theta(F_{V_n}) \leq \theta(F_{V_0}) \quad \text{[finite sample control]}$$

$$\lim \sup_{n \to \infty} \theta(F_{V_n}) \leq \theta(F_{V_0}) \quad \text{[asymptotic control]},$$

where

$$V_n \equiv \text{Number of Type I errors under } T_n \sim Q_n(P) \text{ — truth},$$

$$V_0 \equiv \text{Number of Type I errors under } T_n \sim Q_0 \text{ — assumed.}$$
Null distribution

The required asymptotic control of the Type I error rate follows from the following general asymptotic null domination condition on the distribution for the number of Type I errors $F_{V_n}$

$$\lim_{n \to \infty} \inf F_{V_n}(x) \geq F_{V_0}(x), \quad \forall x \in \{0, \ldots, m\}. \quad (1)$$

In particular, (1) holds if the $S_0$-specific distribution of $Q_0$ equals or dominates the $S_0$-specific distribution of $Q_n = Q_n(P)$, i.e.,

$$\lim_{n \to \infty} Q_{n,S_0} \geq Q_{0,S_0}.$$

More specific (i.e., less stringent) forms of null domination can be derived for given definitions of the Type I error rate.

Finite sample analogues of null domination can be formulated.
Null distribution

We propose a general construction for a null distribution $Q_0$ for the test statistics $T_n$. This null distribution

- satisfies the asymptotic null domination condition

$$\lim_{n \to \infty} \inf F_{V_n}(x) \geq F_{V_0}(x), \quad \forall x \in \{0, \ldots, m\},$$

- and provides asymptotic control of general Type I error rates $\theta(F_{V_n})$,

$$\lim_{n \to \infty} \sup \theta(F_{V_n}) \leq \alpha,$$

- in single-step and step-down procedures,

- for testing general null hypotheses $H_{0j} = I(P \in \mathcal{M}_j)$, corresponding to submodels $\mathcal{M}_j \subseteq \mathcal{M}$ for the data generating distribution $P$. 
Null distribution

**Theorem 0. General construction for test statistics null distribution \( Q_0 \).**

Suppose there exists \( m \)-vectors of null values \( \theta_0 \in IR^m \) and \( \tau_0 \in IR^{+m} \), such that, for \( j \in S_0 \),

\[
\limsup_{n \to \infty} ET_n(j) \leq \theta_0(j) \quad \text{and} \quad \limsup_{n \to \infty} Var[T_n(j)] \leq \tau_0(j).
\]

Then, let \( \nu_0n(j) \equiv \sqrt{\min\left(1, \frac{\tau_0(j)}{Var[T_n(j)]}\right)} \) and define null-value shifted and scaled statistics \( Z_n(j) \)

\[
Z_n(j) \equiv \nu_0n(j)(T_n(j) + \theta_0(j) - ET_n(j)).
\]

Suppose that \( Z_n \overset{D}{\Rightarrow} Z \sim Q_0(P) \). Then, \( Q_0 = Q_0(P) \) satisfies the asymptotic null domination condition in (1).
Null distribution

t-statistics. For tests of single parameter null hypotheses, using t-statistics for asymptotically linear estimators, the null values are \( \theta_0(j) = 0 \) and \( \tau_0(j) = 1 \), and the null distribution \( Q_0 = Q_0(P) \) for \( T_n \) corresponds to the asymptotic distribution of the standardized estimators \( \mu_n \)

\[
Q_0(P) \equiv N(0, \rho(P)),
\]

where \( \rho(P) \) denotes the correlation matrix of the vector influence curve.

That is, a correct null distribution is the projection of the limit of \( Q_n(P) \) onto the space of mean zero distributions.

E.g. For tests of means, where \( \mu = E[X] \), \( \rho(P) \) is simply the correlation matrix for \( X \sim P \).
Null distribution

N.B. The following important points distinguish our approach from existing approaches to Type I error rate control in MTP.

- We are only concerned with control under the true data generating distribution $P$. The notions of weak and strong control are therefore irrelevant to our approach.

- We propose a null distribution for the test statistics $(T_n \sim Q_0)$, and not a data generating null distribution $(X \sim P_0)$. The latter practice does not necessarily provide proper control under the true distribution $P$, as the assumed null distribution $Q_n(P_0)$ and the true distribution $Q_n(P)$ for the test statistics $T_n$ may have different limits: $\lim_n Q_n(P_0) \neq \lim_n Q_n(P)$ (or $\rho_{S_0}(P_0) \neq \rho_{S_0}(P)$).
Estimation of null distribution

In practice, since $P$ is unknown, then so is the null distribution $Q_0 = Q_0(P)$. One can estimate $Q_0$ by $Q_{0n}$ as follows.

- **Bootstrap null distribution**: $Z_n^# \sim Q_{0n}$, based on $X_1^#, \ldots, X_n^# \sim P_n$, very general.

- **Test statistics specific null distribution**
  - Single parameter null hypotheses: $Q_{0n} = N(0, \rho_n)$, where $\rho_n$ is an estimator of the correlation matrix $\rho(P)$ of the vector IC. E.g., for means, $\rho(P)$ can be estimated directly from the data, in other cases, $\rho(P)$ can be estimated with the bootstrap.
  - Multiple parameter null hypotheses: $F$-specific distribution (quadratic forms of Gaussian random variables).

- **Data generating null distribution**: $Q_{0n} = Q_n(P_{0n})$, where $P_{0n}$ is an estimator of $P_0 \in \cap_{j=0}^m \mathcal{M}_j$, e.g., using permutation.

**Problem.** May have $\lim_n Q_{n}(P_{0n}) \neq \lim_n Q_n(P)$. 
Bootstrap estimation of null distribution

**Procedure 0. Bootstrap estimation of null distribution $Q_0$.**

1. Generate $B$ (non-parametric or parametric) bootstrap samples.

2. For each bootstrap sample, compute an $m$-vector of test statistics, $T_n^b = (T_n^b(j) : j = 1, \ldots, m)$, which can be arranged in an $m \times B$ matrix, $\mathbf{T} = (T_n^b(j))$, with rows corresponding to the $m$ hypotheses and columns to the $B$ bootstrap samples.

3. Compute row means and variances of the matrix $\mathbf{T}$, to yield estimates of $ET_n(j)$ and $Var[T_n(j)]$, $j = 1, \ldots, m$.

4. Obtain an $m \times B$ matrix, $\mathbf{Z} = (Z_n^b(j))$, of null-value shifted and scaled statistics, $Z_n^b(j)$, by row shifting and scaling the matrix $\mathbf{T}$ using the bootstrap estimates of $ET_n(j)$ and $Var[T_n(j)]$ and the user-supplied null-values $\theta_0(j)$ and $\tau_0(j)$.

5. The bootstrap estimate $Q_{0n}$ of the null distribution $Q_0$ is the empirical distribution of the columns $Z_n^b$ of matrix $\mathbf{Z}$. 
Theorem 1. Asymptotic control for procedures based on consistent estimator of null distribution $Q_0$.

Let $Q_{0n}$ be a consistent estimator of the null distribution $Q_0$, i.e., $Q_{0n}$ converges weakly to $Q_0$ (e.g., from bootstrap).

Consider a single-step or step-down MTP, $S(T_n, Q_0, \alpha)$, based on cut-offs $c(T_n, Q_0, \alpha)$, that provides asymptotic control of the Type I error rate $\theta(F_{V_n})$ at level $\alpha$. Let $c(T_n, Q_{0n}, \alpha)$ and $S(T_n, Q_{0n}, \alpha)$ denote the corresponding cut-offs and MTP based on the estimator $Q_{0n}$. Then, under regularity conditions for the null distribution $Q_0$,

$$c_j(T_n, Q_{0n}, \alpha) - c_j(T_n, Q_0, \alpha) \xrightarrow{P} 0 \quad \text{as } n \to \infty, \quad \forall j = 1, \ldots, m.$$ 

Consequently, the MTP $S(T_n, Q_{0n}, \alpha)$ also provides asymptotic control of the Type I error rate $\theta(F_{V_n})$ at level $\alpha$. 
Single-step vs. stepwise procedures

One distinguishes between two main approaches for choosing a cut-off vector $c = (c_j : j = 1, \ldots, m)$.

- **Single-step procedures**: equivalent adjustments for all $H_{0j}$, i.e., cut-offs only depend on null distribution $Q_0$ and level $\alpha$: $c = c(Q_0, \alpha)$. (Bonferroni)

- **Stepwise procedures**: adjustments depend on the observed data, i.e., cut-offs also depend on test statistics $T_n$: $c = c(T_n, Q_0, \alpha)$.
  - **Step-down**: start with most significant hypothesis, as soon as one fails to reject a null hypothesis, no further hypotheses are rejected. (Holm, 1979).
  - **Step-up**: start with least significant hypothesis, as soon as one rejects a null hypothesis, reject all hypotheses that are more significant. (Hochberg, 1986).
Single-step multiple testing procedures

Given a null distribution $Q_0$, a mapping $\theta : F \rightarrow \theta(F)$ defining the Type I error rate, and a target or nominal level $\alpha$, consider single-step procedures of the form

$$S(T_n, Q_0, \alpha) = \{j : T_n(j) > c_j(Q_0, \alpha)\},$$

based on two definitions for the $m$-vector of cut-offs $c(Q_0, \alpha) = (c_j(Q_0, \alpha) : j = 1, \ldots, m)$.

Common cut-off. The same cut-off $c_j(Q_0, \alpha) \equiv c_0$ is used for all $m$ test statistics.

Common quantile. The cut-offs for the test statistics are chosen as the $(1 - \delta_0)$-quantiles, i.e., $Pr_{Q_0}(T_n(j) \leq c_j(Q_0, \alpha)) = 1 - \delta_0$. 

Sandrine Dudoit  CBMB/QB3, November 15, 2003  Page 34
Single-step multiple testing procedures

Figure 1: Common cut-off vs. common quantile procedures.
Single-step multiple testing procedures

Let
\[ R_0 = \sum_{j=1}^{m} I(T_n(j) > c_j) = \text{Total } \# \text{ rejections for } T_n \sim Q_0. \]

**Single-step common cut-off procedure.**
Choose common cut-off \( c_j(Q_0, \alpha) \equiv c_0 \) as the smallest cut-off such that \( \theta(F_{R_0}) \leq \alpha. \)

**Single-step common quantile procedure.** Choose common quantiles corresponding to the largest \( \delta_0 \) such that \( \theta(F_{R_0}) \leq \alpha. \)

**E.g.** For FWER, \( Pr(R_0 > 0) \leq \alpha, \) for PCER, \( E[R_0]/m \leq \alpha \)
Single-step multiple testing procedures

**Theorem 2.** Asymptotic control of Type I error $\theta(F_{V_n})$ for single-step procedures.

Consider a null distribution $Q_0$ as in Theorem 0 and a Type I error rate $\theta(F_{V_n})$ defined in terms of a monotone and uniformly continuous mapping $\theta : F \rightarrow \theta(F)$. Define a single-step procedure

$$S_n = S(T_n, Q_0, \alpha) = \{ j : T_n(j) > c_j(Q_0, \alpha) \}$$

based on common cut-offs or common quantile cut-offs $c_j(Q_0, \alpha)$. Then, $S_n$ provides asymptotic control of the Type I error rate $\theta(F_{V_n})$ at level $\alpha$

$$\lim_{n \to \infty} \sup \theta(F_{V_n}) \leq \alpha.$$
Single-step multiple testing procedures: gFWER

Single-step common cut-off procedure for gFWER.

\[ \tilde{p}_n(j) = Pr_{Q_0}(T_n^\circ(k + 1) \geq t_n(j)), \]

where \( T_n^\circ(k + 1) \) denotes the \( k \)th most significant test statistic, that is, \( T_n^\circ(1) \geq T_n^\circ(2) \geq \ldots \geq T_n^\circ(m) \).

Single-step common quantile procedure for gFWER.

\[ \tilde{p}_n(j) = Pr_{Q_0}(P_n^\circ(k + 1) \leq p_n(j)), \]

where \( P_n^\circ(k + 1) \) denotes the \( k \)th most significant unadjusted \( p \)-value, that is, \( P_n^\circ(1) \leq P_n^\circ(2) \leq \ldots \leq P_n^\circ(m) \).

For FWER control \( (k = 0) \), our general procedure reduces to the single-step maxT and single-step minP procedures, based on the maximum test statistic \( T_n^\circ(1) \) and the minimum unadjusted \( p \)-value \( P_n^\circ(1) \), respectively.
Control of gFWER via control of FWER

One can derive a gFWER-controlling procedure using a simple modification of an FWER-controlling procedure.

The main idea is to follow the FWER-controlling procedure exactly until the first non-rejection and then reject this hypothesis and also the next \( k - 1 \) most significant hypotheses.

**Advantages.** (i) Only requires working with the FWER. (ii) Guarantees at least \( k \) rejected hypotheses.
Figure 2: *Control of gFWER via control of FWER.*
Control of gFWER via control of FWER

**Proof.** Let $V_n^0$ and $V_n^k$ denote the number of Type I errors for the FWER- and gFWER-controlling procedures, respectively. Then,

$$V_n^k \leq V_n^0 + k,$$

so that

$$Pr(V_n^k > k) \leq Pr(V_n^0 + k > k) = Pr(V_n^0 > 0).$$
Control of gFWER via control of FWER

Let $O_n(j)$ denote the indices for the ordered test statistics (or unadjusted $p$-values), so that $O_n(1)$ corresponds to the most significant hypothesis, and $O_n(m)$ to the least significant hypothesis. That is, $T_n(O_n(1)) \geq \ldots \geq T_n(O_n(m))$.

The adjusted $p$-values, $\tilde{P}^k_n$, for the gFWER procedure are obtained by simply shifting by $k$ the adjusted $p$-values, $\tilde{P}^0_n$, for the FWER procedure

$$
\tilde{P}^k_n(O_n(j)) = \begin{cases} 
0, & j = 1, \ldots, k, \\
\tilde{P}^0_n(O_n(j - k)), & j = k + 1, \ldots, m. 
\end{cases}
$$
Step-down multiple testing procedures

Step-down maxT procedure for FWER control.
Let $T_n^o(j)$ be the ordered test statistics, $T_n^o(1) \geq \ldots \geq T_n^o(m)$, and $O_n(j)$ the indices for these order statistics, so that $T_n^o(j) = T_n(O_n(j))$. For $Z \sim Q_0$ and subsets $\overline{O}_n(j) \equiv \{O_n(j), \ldots, O_n(m)\}$, define quantiles

$$C_n(j) \equiv (1 - \alpha)-\text{quantile of max}_{k \in \overline{O}_n(j)} Z(k)$$

and step-down cut-offs

$$C_n^o(1) \equiv C_n(1), \ C_n^o(j) \equiv \begin{cases} C_n(j), & \text{if } T_n^o(j - 1) > C_n^o(j - 1) \\ \infty, & \text{otherwise} \end{cases}, \ j = 2, \ldots, m.$$ 

Reject $H_{0,O_n(j)}$ if $T_n^o(j) > C_n^o(j), \ j = 1, \ldots, m.$
Step-down multiple testing procedures

The adjusted $p$-value for hypothesis $H_{0,O_n(j)}$ is given by

$$\tilde{P}_n(O_n(j)) = \max_{k=1,\ldots,j} \left\{ Pr_{Q_0} \left( \max_{l \in \{O_n(k),\ldots,O_n(m)\}} Z(l) \geq T_n(O_n(k)) \right) \right\}. $$

That is, the MTP is based on the distribution of successive maxima of the test statistics $T_n(j)$, over subsets determined by their observed ordering.
Step-down multiple testing procedures

An analogue of the step-down maxT procedure can be defined based on the distribution of successive minima of the unadjusted \( p \)-values \( P_n(j) \).

**Step-down minP procedure for FWER control.**
The adjusted \( p \)-value for hypothesis \( H_{0,0_n(j)} \) is given by

\[
\tilde{P}_n(O_n(j)) = \max_{k=1,\ldots,j} \left\{ Pr_{Q_0} \left( \min_{l \in \{0_n(k),\ldots,0_n(m)\}} \bar{Q}_0 l(Z(l)) \leq P_n(O_n(k)) \right) \right\},
\]

where \( O_n(j) \) denote the indices of the ordered unadjusted \( p \)-values, i.e., \( P_n(O_n(1)) \leq \ldots \leq P_n(O_n(m)) \).
Step-down multiple testing procedures

**Theorem 3.** Asymptotic control of FWER (and gFWER) for step-down procedures.
Consider a null distribution $Q_0$ as in Theorem 0. Assume that, with probability one in the limit, the most significant test statistics correspond to the false null hypotheses, i.e.,

$$\lim_{n} Pr(\{O_n(1), \ldots, O_n(m_1)\} = S_0^c) = 1.$$  

Then, the step-down maxT and step-down minP procedures provide asymptotic control of the FWER at level $\alpha$

$$\lim_{n \to \infty} \sup Pr(V_n > 0) \leq \alpha.$$
Alizadeh et al. (2000) DLBCL microarray dataset

- Target mRNA samples from $n = 40$ Diffuse Large B-Cell Lymphoma (DLBCL) patients.
- Expression levels of $m = 13,412$ clones measured with cDNA microarrays (relative to a pooled control).
- Patients belong to two molecularly distinct disease groups:
  - $n_1 = 21$ Activated B-like (Population 1);
  - $n_2 = 19$ Germinal center (GC) B-like (Population 2).
- Survival time $T$ measured on each patient.
- Pre-processing:
  - $\log_2(R/G)$;
  - replace missing values with average $\log_2(R/G)$ for that gene;
  - truncate ratios exceeding 20-fold to $\pm \log_2(20)$. 
Alizadeh et al. (2000) DLBCL microarray dataset

- **Model:** $n_k$ i.i.d. random $m$-vectors, $X_{k,1}, \ldots, X_{k,n_k} \sim P_k$, of expression measures from patient Population $k$, with mean parameter vector $\mu_k = E[X_k], \ k = 1, 2$.

- **Null hypotheses:** Equal mean expression levels in GC B-like (Popn. 2) and activated B-like (Popn. 1) patients.

  $$H_{0j} = I(\mu(j) \equiv \mu_2(j) - \mu_1(j) = 0), \ j = 1, \ldots, m.$$  

- **Test statistics:** Welch’s two-sample $t$-statistics (unequal variances).

- **Type I error rate:** $FWER = Pr(V_n > 0), \ \alpha = 0.05$.

- **Multiple testing procedures:**
  1. Single-step common quantile (minP), with non-parametric bootstrap null distribution, $Q_{0n}$.
  2. Single-step common quantile (minP), with permutation null distribution, $Q_n(P_{0n})$.
  3. Bonferroni, with nominal $t$-distribution.
Alizadeh et al. (2000) DLBCL microarray dataset

Table 1: *Test difference in mean expression levels for GC B-like and activated B-like patients.* Number of rejected null hypotheses (out of $m = 13,412$) for three MTPs for FWER control.

<table>
<thead>
<tr>
<th>MTP</th>
<th>Cut-offs</th>
<th>Null distribution</th>
<th># rejections ($R_n$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Single-step minP</td>
<td>Bootstrap</td>
<td>186</td>
</tr>
<tr>
<td>2</td>
<td>Single-step minP</td>
<td>Permutation</td>
<td>287</td>
</tr>
<tr>
<td>3</td>
<td>Bonferroni</td>
<td>Nominal t</td>
<td>32</td>
</tr>
</tbody>
</table>

- All 32 genes identified as differentially expressed by MTP3 were also identified by MTP 1 and MTP 2.
- MTP1 and MTP2 identified 156 genes in common.
Alizadeh et al. (2000) DLBCL microarray dataset

- **Model**: Logistic regression model for each gene

  \[ Pr(\text{GC B-like} \mid X(j)) = \frac{e^{\beta_0(j) + \beta_1(j) \times X(j)}}{1 + e^{\beta_0(j) + \beta_1(j) \times X(j)}}, \quad j = 1, \ldots, m. \]

- **Null hypotheses**: No association between expression measure \( X(j) \) of gene \( j \) and patient group: \( H_{0j} = I(\beta_1(j) = 0), \quad j = 1, \ldots, m. \)

- **Test statistics**: \( T_n(j) = \sqrt{n} \beta_{1n}(j) \).

- **Type I error rate**: \( gFWER = Pr(V_n > k), \quad k = 1, \ldots, 200, \alpha = 0.05. \)

- **Multiple testing procedures**:
  1. **Direct gFWER single-step common quantile procedure**, i.e., single-step based on \( P_n^\circ(k + 1) \), with non-parametric bootstrap null distribution, \( Q_{0n} \).
  2. **Via FWER single-step common quantile procedure**, i.e., single-step minP based on \( P_n^\circ(1) \), with non-parametric bootstrap null distribution, \( Q_{0n} \).
Alizadeh et al. (2000) DLBCL microarray dataset

Table 2: Test for slope in logistic regression model for DLBCL class and expression measures. Number of rejected null hypotheses (out of \( m = 13,412 \)) for two MTPs for gFWER control.

<table>
<thead>
<tr>
<th></th>
<th>( k )</th>
<th>0</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct gFWER</td>
<td>( R_n^k )</td>
<td>303</td>
<td>303</td>
<td>303</td>
<td>471</td>
<td>553</td>
</tr>
<tr>
<td>Via FWER</td>
<td>( R_n^0 + k )</td>
<td>303</td>
<td>313</td>
<td>353</td>
<td>403</td>
<td>503</td>
</tr>
</tbody>
</table>

Here, \( R_n^k \) denotes the number of rejected hypotheses for a procedure directly controlling the gFWER. For FWER, \( k = 0 \).
Golub et al. (1999) ALL/AML microarray dataset

- Target mRNA samples from $n = 38$ leukemia patients (learning set).

- Expression levels of $d = 6,817$ human genes measured with Affymetrix hu6800 chip.

- Patients belong to two molecularly distinct disease groups:
  - $n_1 = 27$ acute lymphoblastic leukemia (ALL) (Population 1);
  - $n_2 = 11$ acute myeloid leukemia (AML) (Population 2).

- Pre-processing:
  - thresholding: floor of 100 and ceiling of 16,000;
  - filtering: exclusion of genes with $\frac{\text{max}}{\text{min}} \leq 5$ or $(\text{max} - \text{min}) \leq 500$;
  - base 10 logarithmic transformation.
Golub et al. (1999) ALL/AML microarray dataset

- **Model:** $n_k$ i.i.d. random $m$-vectors, $X_{k,1}, \ldots, X_{k,n_k} \sim P_k$, of expression measures from patient Population $k$, with mean parameter vector $\mu_k = E[X_k]$, $k = 1, 2$.

- **Null hypotheses:** Equal mean expression levels in AML (Popn. 2) and ALL (Popn. 1) patients.

  \[ H_{0j} = I(\mu(j) \equiv \mu_2(j) - \mu_1(j) = 0), \quad j = 1, \ldots, m. \]

- **Test statistics:** Welch’s two-sample $t$-statistics.

- **Multiple testing procedures:** Permutation null distribution.
  - FWER: Bonferroni, Holm (1979), Hochberg (1986), step-down maxT.
  - PCER: SAM, unadjusted $p$-values.
Golub et al. (1999) ALL/AML microarray dataset
Golub et al. (1999) ALL/AML microarray dataset
Software: multtest
www.bioconductor.org

Bioconductor R package: multtest.
Authors: Yongchao Ge, Katie Pollard, Sandrine Dudoit.

- **Test statistics.** $t$-statistics and $F$-statistics (one- and two-factor) for linear models and Cox PH model. Includes non-parametric and weighted statistics.

- **Null distributions.** Permutation and bootstrap.

- **Multiple testing procedures.**
  - **FWER control:** Bonferroni, Holm (1979), Hochberg (1986), single-step and step-down maxT and minP.
Software: multtest
www.bioconductor.org

- **Output.** Parameter estimates, test statistics, unadjusted and adjusted \( p \)-values, cut-off vector, confidence intervals, null distribution.

- **Plots.** Type I error rate vs. \# rejections, \# rejections vs. adjusted \( p \)-values, adjusted \( p \)-values vs. test statistics ("volcano").

- **Software design.**
  - Closures: new tests can easily be added.
  - Class/method OOP.
Ongoing work

- Applications to more complex testing problems: e.g., parameters for censored survival time models, contrasts in designed microarray experiments.
- Comparison of different null distributions: general proposal from Theorem 0 vs. test statistic specific.
- Comparison of single-step vs. step-down procedures.
- Test optimality in terms of power: Optimize power among class of MTPs; Extensions of univariate concepts?
References


