Outline

• Differential Gene Expression.


• Test Statistics Null Distribution.

• Single-Step Procedures for Control of General Type I Error Rates, \( \theta(F_{V_n}) \).

• Step-Down Procedures for Control of the Family-Wise Error Rate.

• Augmentation Multiple Testing Procedures.

• Software: Bioconductor R Package \textit{multtest}.
References: Available on Course Website

   — *A summary or ”light” version of the four papers below, with details on software implementation.*

   — *A summary or ”light” version of the four papers below, with applications to: microarray experiments; ChIP-Chip experiments; Gene Ontology; genetic mapping with SNPs.*

References: Available on Course Website


Main methodological references, rather theoretical and detailed.
Differential Gene Expression

An important and common question in DNA microarray experiments is the identification of differentially expressed genes, i.e., genes whose expression measures are associated with possibly censored responses or covariates interest.

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.
**Differential Gene Expression**

**E.g.** Which genes have different mean expression levels in patients with ALL B-cell and ALL T-cell cancers?

**E.g.** Which genes have expression levels that are associated with patient survival?

**E.g.** Which genes are spatially differentially expressed in the mouse olfactory bulb?

**E.g.** Which genes are temporally differentially expressed in *S. Cerevisiae* during the cell-cycle?
Differential Gene Expression

The biological question of differential gene expression can be translated into a statistical multiple hypothesis testing problem, as follows.

- Define the data generating distribution and association parameters of interest.
  E.g. Means, differences in means, correlations, regression coefficients in linear models, Cox proportional hazards survival models, and time-series models.

- Translate the null hypothesis of no differential gene expression into a statement concerning the parameters of interest.
  E.g. Regression coefficients are equal to zero.

- Test for each gene, the null hypothesis of no association between the expression measures and the responses or covariates, using an appropriate test statistic. This involves estimating the parameters of interest and the variability of the corresponding estimators.
Differential Gene Expression

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.
Multiple Testing Procedure

A multiple testing procedure (MTP) involves the following.

- Selecting 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.
- Deriving rejection regions, i.e., cut-offs, for the test statistics under a joint null distribution that provides the desired control of the Type I error rate under the true, unknown data generating distribution.
- Deriving rejection regions that take into account the dependence structure among the test statistics.
- Using resampling methods to estimate the joint null distribution of the test statistics.
- Reporting adjusted $p$-values, which reflect the strength of the evidence against each null hypothesis in terms of the Type I error rate for the entire MTP.
Multiple Hypothesis Testing Problems in Genomics

- **Microarray experiments.** Identification of differentially expressed genes or genes with correlated expression profiles.

- **ChIP-Chip experiments.** Chromatin immunoprecipitation (ChIP) of DNA bound by transcription factor (TF) is followed by microarray hybridization (Chip) of IP enriched DNA. TF binding sites can be identified by testing whether each probe sequence is bound or not by the TF.

- **Gene Ontology (GO) annotation.** Test for associations between GO annotations and other properties of a genome (e.g., mean transcript levels in different cell populations).

- **Genetic mapping.** Test for associations between phenotypes and single nucleotide polymorphisms (SNP).
Multiple Hypothesis Testing Problems in Genomics

The above testing problems share the following general characteristics.

- Inference for high-dimensional, unknown multivariate distributions.

- Broad range of parameters of interest: e.g., regression coefficients in model relating patient survival to genome-wide transcript levels or DNA copy numbers, pairwise gene correlations between transcript levels, association measures between gene-property and gene-parameter profiles.

- Many null hypotheses: in the thousands.

- Complex dependence structure among test statistics: e.g., SNP Galois lattice, Gene Ontology directed acyclic graph (DAG).
Multiple Hypothesis Testing

Hypothesis testing is concerned with using observed data to test hypotheses, i.e., make decisions, regarding properties of the unknown data generating distribution.

- Data: \( X_1, \ldots, X_n \sim P \in \mathcal{M} \).
- Null hypotheses: \( H_0(m) = I(P \in \mathcal{M}(m)) \), \( \mathcal{M}(m) \subseteq \mathcal{M}, m = 1, \ldots, M \).
- Definition of Type I and II error rates.
- Test statistics: \( T_n = T(X_1, \ldots, X_n) \).
- Null distribution for the test statistics: \( Q_0 \) or estimate thereof, \( Q_{0n} \).
- Rejection regions: \( C_n = C(T_n, Q_{0n}, \alpha) \).
- Summaries of results: \( p \)-values, Type I and II error rates.
Data

Let $X_1, \ldots, X_n$ be a random sample of $n$ independent and identically distributed (i.i.d.) random variables, $X \sim P \in \mathcal{M}$, where the data generating distribution $P$ is known to be an element of a particular statistical model $\mathcal{M}$ (i.e., a set of possibly non-parametric distributions).
Null Hypotheses

General submodel null hypotheses. Define $M$ pairs of null and alternative hypotheses,

$$H_0(m) \equiv \{ P \in \mathcal{M}(m) \} \quad \text{vs.} \quad H_1(m) \equiv \{ P \notin \mathcal{M}(m) \},$$

in terms of a collection of $M$ submodels, $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \ldots, M$, for the data generating distribution $P$.

Let $\mathcal{H}_0 \equiv \{ m : H_0(m) = 1 \} = \{ m : P \in \mathcal{M}(m) \}$ be the set of $h_0 \equiv |\mathcal{H}_0|$ true null hypotheses and let $\mathcal{H}_1 \equiv \mathcal{H}_0^c$ be the set of $h_1 \equiv |\mathcal{H}_1| = M - h_0$ false null hypotheses (true positives).

This general representation covers tests of means, differences in means, correlations, and parameters in linear models, generalized linear models, survival models, time-series models, dose-response models, etc.
Null Hypotheses

Single-parameter null hypotheses. In many testing problems, the submodels concern parameters, i.e., functions of the data generating distribution $P$, $\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M)$. Two-sided tests

$$H_0(m) \equiv I(\psi(m) = \psi_0(m)), \quad m = 1, \ldots, M.$$  

One-sided tests

$$H_0(m) \equiv I(\psi(m) \leq \psi_0(m)), \quad m = 1, \ldots, M.$$  

The hypothesized null values, $\psi_0(m)$, are frequently zero.
Testing Procedure

A testing procedure is a data-driven rule for deciding whether or not to reject each of the $M$ null hypotheses $H_0(m)$, i.e., declare that $H_0(m)$ is false (zero) and hence $P \notin \mathcal{M}(m)$.

The decisions to reject or not the null hypotheses are based on test statistics, i.e., functions of the data, $T_n(m) = T(X_1, \ldots, X_n)(m)$. The testing procedure provides rejection regions, $\mathcal{C}_n(m)$, i.e., sets of values for the test statistics $T_n(m)$ that lead to the decisions to reject the null hypotheses $H_0(m)$:

"Reject $H_0(m)$ if $T_n(m) \in \mathcal{C}_n(m)$.”

In many testing problems, rejection regions are of the form $\mathcal{C}_n(m) = (c_n(m), \infty)$, where $c_n(m)$ are to-be-determined cut-offs, or critical values.
Test Statistics

For single-parameter null hypotheses, $H_0(m) = I(\psi(m) \leq \psi_0(m))$, one can use the following two types of test statistics.

**Difference statistics**

$$T_n(m) \equiv \text{Estimator} - \text{Null Value} = \sqrt{n}(\psi_n(m) - \psi_0(m)).$$

**$t$-statistics** (i.e., standardized difference statistics)

$$T_n(m) \equiv \frac{\text{Estimator} - \text{Null Value}}{\text{Standard Error}} = \frac{\sqrt{n}(\psi_n(m) - \psi_0(m))}{\sigma_n(m)}.$$

$\psi_n(m)$ denotes an asymptotically linear estimator of $\psi(m)$ and $\sigma_n(m)/\sqrt{n}$ a consistent estimator of the standard error of $\psi_n(m)$.

Test statistics for other types of null hypotheses include $F$-statistics, $\chi^2$-statistics, likelihood ratio statistics.

Let $Q_n(P)$ denote the (unknown) joint distribution of the $M$–vector of test statistics, $T_n = (T_n(m) : m = 1, \ldots, M)$. 
Test Statistics

One-sample $t$-statistics. Suppose we have $n$ i.i.d. random $M$-vectors, $X_1, \ldots, X_n \sim P$, from a population with mean parameter vector $\psi = E[X]$.

Null hypotheses. $H_0(m) = I(\psi(m) = \psi_0(m))$, $m = 1, \ldots, M$.

Test statistics.

$$T_n(m) = \sqrt{n} \frac{\bar{X}_n(m) - \psi_0(m)}{\sigma_n(m)}, \quad m = 1, \ldots, M,$$

where $\bar{X}_n(m) = \sum_i X_i(m)/n$ and $\sigma^2_n(m) = \sum_i (X_i(m) - \bar{X}_n(m))^2/n$ denote the sample means and variances, respectively.
Test Statistics

Two-sample $t$-statistics. Suppose we have $n_k$ i.i.d. random $M$–vectors, $X_{k,1}, \ldots, X_{k,n_k} \sim P_k$, from Population $k$, with mean parameter vector $\psi_k = E[X_k]$, $k = 1, 2$.

Null hypotheses. Equal population means

$$H_0(m) = I(\psi_2(m) = \psi_1(m)), \quad m = 1, \ldots, M.$$  

Test statistics. Two-sample Welch’s $t$-statistics

$$T_n(m) = \frac{\bar{X}_{2,n_2}(m) - \bar{X}_{1,n_1}(m)}{\sqrt{\sigma_{1,n_1}^2(m)/n_1 + \sigma_{2,n_2}^2(m)/n_2}}, \quad m = 1, \ldots, M,$$

where $\bar{X}_{k,n_k}(m)$ and $\sigma_{k,n_k}^2(m)$ denote sample means and variances for Population $k$, respectively. Can also use equal variance version.
Test Statistics

\((K\text{-sample}) F\text{-statistics}\). Suppose we have \(n_k\) i.i.d. random 
\(M\text{-vectors}, X_{k,1}, \ldots, X_{k,n_k} \sim P_k\), from Population \(k\), with mean parameter vector \(\psi_k = E[X_k], \ k = 1, \ldots, K\).

Null hypotheses. Equal population means

\[ H_0(m) = \mathbf{1}(\psi_1(m) = \psi_2(m) = \ldots = \psi_K(m)), \ m = 1, \ldots, M. \]

Test statistics.

\[ T_n(m) \equiv \frac{1/(K-1) \sum_{k=1}^K n_k \left( \bar{X}_{k,n_k}(m) - \bar{X}_n(m) \right)^2}{1/(n-K) \sum_{k=1}^K \sum_{i=1}^{n_k} (X_{k,i}(m) - \bar{X}_{k,n_k}(m))^2}, \ m = 1, \ldots, M, \]

where \(\bar{X}_{k,n_k}(m)\) denote the sample means for Population \(k\) and 
\(\bar{X}_n(m) \equiv \sum_k n_k \bar{X}_{k,n_k}(m) / \sum_k n_k\) denote the overall sample means.
Test Statistics

t-statistics for regression parameters in linear model. Suppose we have \( n \) i.i.d. observations \((W_1, X_1, Y_1), \ldots, (W_n, X_n, Y_n) \sim P\).

E.g. \( X \) is an \( M \)-vector of expression measures; \( Y \) is a quantitative clinical response; \( W \) is an \( L \)-vector of covariates.

Assume the following linear model

\[
E[Y|X(m), W] = \psi(m)_{1 \times 1}X(m)_{1 \times 1} + \beta(m)_{1 \times L}W_{L \times 1},
\]

\[
Var[Y|X(m), W] = \sigma^2(m).
\]
Test Statistics

Null hypotheses. No association between response $Y$ and variables $X(m)$, after adjusting for covariates $W$,

$$H_0(m) = I(\psi(m) = 0), \quad m = 1, \ldots, M.$$ 

Test statistics. $t$-statistics

$$T_n(m) = \sqrt{n} \frac{\psi_n(m) - 0}{\sigma_n(m)}, \quad m = 1, \ldots, M,$$

where $\psi_n(m)$ are the ordinary least squares (OLS) estimators of the regression coefficients $\psi(m)$ and $\sigma_n(m)/\sqrt{n}$ are the corresponding estimated standard errors.
Multiple Testing Procedure

A multiple testing procedure (MTP) provides rejection regions, $C_n(m)$, i.e., sets of values for the test statistics $T_n(m)$ that lead to the decision to reject the null hypotheses $H_0(m)$. In other words, a MTP produces a random subset $R_n$ of $R_n$ rejected hypotheses that estimates $\mathcal{H}_1$, the set of true positives,

$$R_n \equiv \{m : H_0(m) \text{ is rejected}\} = \{m : T_n(m) \in C_n(m)\}.$$

We may use the long notation $R_n = \mathcal{R}(T_n, Q_{0n}, \alpha)$ and $C_n(m) = \mathcal{C}(T_n, Q_{0n}, \alpha)(m)$ to emphasize that the MTP depends on:

- the data, $X_1, \ldots, X_n$, through the $M$–vector of test statistics $T_n$ (e.g., $t$-, $F$-, $\chi^2$-statistics);
- an $M$–variate null distribution, $Q_{0n}$, for the test statistics $T_n$;
- the nominal level, $\alpha$, of the MTP, i.e., the desired upper bound for a suitably defined Type I error rate $\theta$. 

\[\text{October 20, 2004}\]
Type I and Type II Errors

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

- A false positive, or Type I error, is committed by rejecting \((m \in \mathcal{R}_n)\) a true null hypothesis \((m \in \mathcal{H}_0)\).
  
  **E.g.** Report an association between phenotype and genotype when in truth there is no such association.

- A false negative, or Type II error, is committed when the testing procedure fails to reject \((m \notin \mathcal{R}_n)\) a false null hypothesis \((m \notin \mathcal{H}_0)\), cf. power.
  
  **E.g.** Fail to identify a true association between phenotype and genotype.
Type I and Type II Errors

Null hypotheses

<table>
<thead>
<tr>
<th>Not rejected, $\mathcal{R}_n^c$</th>
<th>Rejected, $\mathcal{R}_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>\mathcal{R}_n \cap \mathcal{H}_0</td>
</tr>
<tr>
<td>$U_n \equiv</td>
<td>\mathcal{R}_n^c \cap \mathcal{H}_1</td>
</tr>
</tbody>
</table>

(Type I errors)

$\mathcal{H}_0$ true,

Type II errors

$\mathcal{H}_1$ false,
Type I and Type II Errors

Ideally, one would like to simultaneously minimize both the chances of committing a Type I error and a Type II error. Unfortunately, this is not feasible and one seeks a trade-off between the two types of errors.

A standard approach is to specify an acceptable level $\alpha$ for the Type I error rate and derive testing procedures, i.e., rejection regions, that aim to minimize the Type II error rate, i.e., maximize power, within the class of tests with Type I error rate at most $\alpha$.

For single hypothesis testing, optimality results are available for particular types of data generating distributions, null hypotheses, and test statistics.
Type I Error Rates

We adopt a general definition of Type I error rates, as parameters, $\theta_n = \theta(F_{V_n,R_n})$, of the joint distribution $F_{V_n,R_n}$ of the numbers of Type I errors $V_n$ and rejected hypotheses $R_n$.

Such a general representation covers a broad class of Type I error rates, defined as tail probabilities,

$$\theta(F_{V_n,R_n}) = Pr(g(V_n, R_n) > q),$$

and expected values,

$$\theta(F_{V_n,R_n}) = E[g(V_n, R_n)],$$

for an arbitrary user-supplied function $g(V_n, R_n)$ and constant $q$.

As detailed below, special cases of interest include, $g(V_n, R_n) = V_n$ (gFWER, PCER) and $g(V_n, R_n) = V_n/R_n$ (TPPFP, FDR).
**Type I Error Rates:** \( g(V_n, R_n) = V_n \)

- **Generalized family-wise error rate** (gFWER), or probability of at least \((k + 1)\) Type I errors, \(k = 0, \ldots, (h_0 - 1)\),

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

When \(k = 0\), the gFWER is the usual **family-wise error rate**, FWER, controlled by the classical Bonferroni procedure.

- **Per-comparison error rate** (PCER), or expected proportion of Type I errors among the \(M\) tests,

\[
PCER \equiv \frac{1}{M} E[V_n] = \frac{1}{M} \int vdF_{V_n}(v).
\]
Type I Error Rates: \( g(V_n, R_n) = V_n/R_n \)

- Tail probabilities for the proportion of false positives (TPPFP) among the rejected hypotheses,
  \[ TPPFP(q) \equiv Pr(V_n/R_n > q), \quad q \in (0, 1). \]

- False discovery rate (FDR), or expected value of the proportion of false positives among the rejected hypotheses,
  \[ FDR \equiv E[V_n/R_n]. \]

Convention: \( V_n/R_n \equiv 0, \) if \( R_n = 0. \)

Error rates based on the proportion of false positives (vs. the number of false positives, as with gFWER) are especially appealing for the large-scale testing problems encountered in genomics, as they do not increase exponentially with the number of hypotheses.
**Adjusted \( p \)-Values**

The notion of \( p \)-value extends directly to multiple testing problems. Given a MTP, \( \mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) \), the adjusted \( p \)-value, \( \tilde{P}_{0n}(m) = \tilde{P}(T_n, Q_{0n})(m) \), for null hypothesis \( H_0(m) \), is defined as

\[
\tilde{P}_{0n}(m) \equiv \inf \{ \alpha \in [0, 1] : \text{Reject } H_0(m) \text{ at MTP level } \alpha, \text{ given } T_n \} = \inf \{ \alpha \in [0, 1] : m \in \mathcal{R}_n \} = \inf \{ \alpha \in [0, 1] : T_n(m) \in C_n(m) \}, \quad m = 1, \ldots, M.
\]

**E.g.** Bonferroni MTP for FWER control,

\[
\tilde{P}_{0n}(m) = \min(M P_{0n}(m), 1),
\]

where \( P_{0n}(m) = P(T_n(m), Q_{0n,m}) \) is the marginal, or unadjusted, \( p \)-value for \( H_0(m) \).
Adjusted $p$-Values

Advantages of reporting adjusted $p$-values.

- As in single hypothesis tests, the smaller the $p$-value, the stronger the evidence against the corresponding null hypothesis $\Rightarrow$ Reject $H_0(m)$ for small adjusted $p$-values $\tilde{P}_{0n}(m)$.

- Adjusted $p$-values reflect the strength of the evidence against each null hypothesis in terms of the Type I error rate for the entire MTP.

- Adjusted $p$-values can be defined for any Type I error rate, gFWER, FDR, TPPFP, etc.
Adjusted \( p \)-Values

- Adjusted \( p \)-values are flexible summaries of a MTP; the results of the MTP are provided for all nominal levels \( \alpha \) \( \implies \) do not need to choose the level ahead of time.

- We now have two equivalent representations of a MTP, in terms of rejections regions for the test statistics and in terms of adjusted \( p \)-values,

\[
\mathcal{R}_n = \{m : T_n(m) \in \mathcal{C}_n(m)\} = \{m : \tilde{P}_{0n}(m) \leq \alpha\}.
\]
Null Distribution

One of the main tasks in specifying a MTP is to derive rejection regions for the test statistics such that the Type I error rate is controlled at a desired level $\alpha$, i.e., such that

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

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

Note that the parameter $\theta(F_{V_n,R_n})$ 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 replaced by a null distribution $Q_0$ (or estimate thereof, $Q_{0n}$), in order to derive rejection regions, $C(T_n, Q_0, \alpha)(m)$, and resulting adjusted $p$-values, $\tilde{P}(T_n, Q_0)(m)$.

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 null distribution $Q_0$ does indeed provide the required control under the true distribution $Q_n(P)$. 
Null Distribution

For proper control, the null distribution $Q_0$ must be such that the Type I error rate under this assumed null distribution dominates the Type I error rate under the true distribution $Q_n(P)$. That is,

$$\theta(F_{V_n,R_n}) \leq \theta(F_{V_0,R_0}) \quad [\text{finite sample control}]$$

$$\limsup_{n \to \infty} \theta(F_{V_n,R_n}) \leq \theta(F_{V_0,R_0}) \quad [\text{asymptotic control}],$$

where $V_0$ and $R_0$ denote, respectively, the numbers of Type I errors and rejected hypotheses under the assumed null distribution $Q_0$. 
Null Distribution

For error rates $\theta(F_{V_n})$, defined as arbitrary parameters of the distribution of the number of Type I errors $V_n$, we propose as null distribution $Q_0$ the asymptotic distribution of the vector of null value shifted and scaled test statistics.

This null distribution does indeed provides the desired asymptotic control of the Type I error rate $\theta(F_{V_n})$, for general data generating distributions (with arbitrary dependence structures among variables), null hypotheses (defined in terms of submodels for the data generating distribution), and test statistics (e.g., $t$-statistics, $F$-statistics).

We propose resampling procedures (e.g., based on the non-parametric or model-based bootstrap) to conveniently obtain consistent estimators of the null distribution and the resulting test statistic cut-offs and adjusted $p$-values.
Null Distribution

t-statistics. For the test of single-parameter null hypotheses using t-statistics, the null values are $\lambda_0(m) = 0$ and $\tau_0(m) = 1$. The null distribution $Q_0$ is an $M$–variate Gaussian distribution

$$Q_0 = Q_0(P) \equiv N(0, \Sigma^*(P)).$$

For tests of means, where $\psi = \Psi(P) = E[X]$, $\Sigma^*(P)$ is simply the correlation matrix $Cor[X]$ for $X \sim P$. More generally, for an asymptotically linear estimator $\psi_n$, $\Sigma^*(P)$ is the correlation matrix of the vector influence curve (IC).

F-statistics. For testing the equality of $K$ population means using $F$-statistics, the null values are $\lambda_0(m) = 1$ and $\tau_0(m) = 2/(K - 1)$, under the assumption of equal variances in the different populations. The null distribution $Q_0$ is the joint distribution of a random $M$–vector of quadratic forms of Gaussian random variables.
Null Distribution

- We are only concerned with Type I error control under the true data generating distribution $P$. The notions of weak and strong control (and associated subset pivotality) 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 ($X \sim P_0 \in \bigcap_{m=1}^M M(m)$).

The latter practice does not necessarily provide proper control under the true distribution $P$, as the test statistics’ assumed null distribution $Q_n(P_0)$ and their true distribution $Q_n(P)$ may have different dependence structures (in the limit), for the true null hypotheses $\mathcal{H}_0$. 
Multiple Testing Procedures

Having selected a suitable test statistics null distribution $Q_0$ (or estimate thereof, $Q_{0n}$), there remains the main task of specifying rejection regions for each null hypothesis, i.e., cut-offs for each test statistic.

- **Common-cut-off procedures**: the same cut-off $c_0$ is used for each test statistic (cf. maxT).

  vs.

- **Common-quantile procedures**: the cut-offs are the $\delta_0$-quantiles of the marginal null distributions of the test statistics (cf. minP).

  Issue is balance: common-quantile, or $p$-value based, procedures put null hypotheses on an equal footing.

- **Single-step vs. stepwise procedures**.
Single-Step vs. Stepwise Procedures

- **Single-step procedures.** Each null hypothesis is evaluated using a rejection region that is independent of the results of the tests of other hypotheses – $C_n = C(Q_0, \alpha)$ (Bonferroni).

- **Stepwise procedures.**
  - The test procedure is applied to a sequence of successively smaller nested random (i.e., data-dependent) subsets of null hypotheses, defined by the ordering of the test statistics (common cut-offs) or unadjusted $p$-values (common-quantile cut-offs).
  - Rejection of a particular null hypothesis depends on the outcome of the tests of other hypotheses – $C_n = C(T_n, Q_0, \alpha)$. 
Step-Down vs. Step-Up Procedures

Among stepwise procedures, we distinguish between step-down and step-up procedures.

- **Step-down procedures.** Start with most significant hypothesis, as soon as one fails to reject a null hypothesis, no further hypotheses are rejected (Holm, 1979).

- **Step-up procedures.** Start with least significant hypothesis, as soon as one rejects a null hypothesis, reject all hypotheses that are more significant (Hochberg, 1986).
Multiple Testing Procedures

Next, we propose three main approaches for deriving rejection regions and corresponding adjusted $p$-values.

- **Single-step common-cut-off and common-quantile procedures** for control of general Type I error rates, $\theta(F_{V_n})$.

- **Step-down common-cut-off (maxT) and common-quantile (minP) procedures** for control of the FWER.

- **Augmentation procedures** for control of the gFWER and TPPFP, based on an initial FWER-controlling procedure, i.e., MTPs obtained by adding suitably chosen null hypotheses to the set of hypotheses already rejected by an initial MTP.
Single-Step Multiple Testing Procedures

Main idea. substitute control of the parameter $\theta(F_{V_n})$, for the unknown, true distribution $F_{V_n}$ of the number of Type I errors, by control of the corresponding parameter $\theta(F_{R_0})$, for the known, null distribution $F_{R_0}$ of the number of rejected hypotheses.

That is, consider single-step procedures of the form $R_n \equiv \{m : T_n(m) > c_n(m)\}$, where the cut-offs $c_n(m)$ are chosen so that

$$\theta(F_{R_0}) \leq \alpha,$$

for $R_0 \equiv \sum_{m=1}^{M} \mathbb{I}(Z(m) > c_n(m))$ and $Z \sim Q_0$.

Among the class of MTPs that satisfy $\theta(F_{R_0}) \leq \alpha$, consider two main approaches: procedures based on common cut-offs and procedures based on common-quantile cut-offs.
Single-Step Multiple Testing Procedures

**Single-step common-cut-off procedure.** The set of rejected hypotheses is of the form \( R_n \equiv \{ m : T_n(m) > c_0 \} \), where the common cut-off \( c_0 \) is the smallest (i.e., least conservative) value for which \( \theta(F_{R_0}) \leq \alpha \).

▷ **gFWER(k) control:** Procedure is based on the \((k + 1)\)st ordered test statistic. Adjusted \( p \)-values are given by

\[
\tilde{p}_{0n}(m) = Pr_{Q_0} (Z^\circ(k + 1) \geq t_n(m)), \quad m = 1, \ldots, M,
\]

where \( Z^\circ(m) \) denotes the \( m \)th ordered component of

\[
Z = (Z(m) : m = 1, \ldots, M) \sim Q_0,
\]
so that \( Z^\circ(1) \geq \ldots \geq Z^\circ(M) \).

▷ **FWER control \((k = 0)\):** Single-step maxT procedure, based on the maximum test statistic.
Single-Step Multiple Testing Procedures

Single-step common-quantile procedure.

The set of rejected hypotheses is of the form $\mathcal{R}_n \equiv \{m : T_n(m) > c_0(m)\}$, where $c_0(m) = Q_{0,m}^{-1}(\delta_0)$ is the $\delta_0$–quantile of the marginal null distribution $Q_{0,m}$ of the $m$th test statistic and $\delta_0$ is chosen as the smallest (i.e., least conservative) value for which $\theta(F_{R_0}) \leq \alpha$.

▷ $gFWER(k)$ control: Procedure based on the $(k + 1)$st ordered unadjusted $p$-value. Adjusted $p$-values are given by

$$\tilde{p}_{0n}(m) = Pr_{Q_0}(P_0^\circ(k + 1) \leq p_{0n}(m)), \quad m = 1, \ldots, M,$$

where $P_0(m) \equiv \bar{Q}_{0,m}(Z(m))$ and $P_{0n}(m) \equiv \bar{Q}_{0,m}(T_n(m))$ denote unadjusted $p$-values for $Z \sim Q_0$ and $T_n \sim Q_n$, respectively. $P_0^\circ(m)$ denote the $m$th ordered component of the $M$–vector of unadjusted $p$-values ($P_0(m) : m = 1, \ldots, M$), so that $P_0^\circ(1) \leq \ldots \leq P_0^\circ(M)$.

▷ FWER control ($k = 0$): Single-step minP procedure, based on the minimum unadjusted $p$-value.
Step-Down Multiple Testing Procedures

Main idea. Apply test procedure to a sequence of successively smaller nested random (i.e., data-dependent) subsets of null hypotheses, defined by the ordering of the test statistics (common cut-offs) or unadjusted $p$-values (common-quantile cut-offs).

Similar in spirit to their above single-step counterparts (special case $\theta(F_{Vn}) = 1 - F_{Vn}(0)$), with the important step-down distinction that hypotheses are considered successively, from most significant to least significant, with further tests depending on the outcome of earlier ones.
Step-Down Multiple Testing Procedures

**Step-down common-cut-off (maxT) procedure.** Let $O_n(m)$ denote the indices for the ordered test statistics $T_n(m)$, so that $T_n(O_n(1)) \geq \ldots \geq T_n(O_n(M))$. The adjusted $p$-values for the step-down maxT procedure are given by

\[
\tilde{p}_{0n}(o_n(m)) = \max_{h=1,\ldots,m} \left\{ \Pr_{Q_0} \left( \max_{l \in \overline{O}_n(h)} Z(l) \geq t_n(o_n(h)) \right) \right\},
\]

where $Z = (Z(m) : m = 1, \ldots, M) \sim Q_0$ and $\overline{O}_n(h) \equiv \{O_n(h), \ldots, O_n(M)\}$.

Thus, rather than being based solely on the distribution of the maximum test statistic over all $M$ hypotheses, the step-down common cut-offs and corresponding adjusted $p$-values are based on the distributions of maxima of test statistics over successively smaller nested random subsets of ordered null hypotheses, $\overline{O}_n(h)$.
Step-Down Multiple Testing Procedures

Step-down common-quantile (minP) procedure. Let $O_n(m)$ denote the indices for the ordered unadjusted $p$-values $P_{0n}(m)$, so that $P_{0n}(O_n(1)) \leq \ldots \leq P_{0n}(O_n(M))$. The adjusted $p$-values for the step-down minP procedure are given by

$$
\tilde{p}_{0n}(o_n(m)) = \max_{h=1,\ldots,m} \left\{ \Pr_{Q_0} \left( \min_{l \in \bar{O}_n(h)} P_0(l) \leq p_{0n}(o_n(h)) \right) \right\},
$$

where $P_0(m) = \bar{Q}_{0,m}(Z(m))$, $Z = (Z(m) : m = 1, \ldots, M) \sim Q_0$, and $\bar{O}_n(h) \equiv \{O_n(h), \ldots, O_n(M)\}$.

Thus, the step-down common-quantile cut-offs and corresponding adjusted $p$-values are based on the distributions of minima of unadjusted $p$-values over successively smaller nested random subsets of ordered null hypotheses, $\bar{O}_n(h)$. 
Augmentation Multiple Testing Procedures

Main idea. In order to control a new target Type I error rate, an augmentation multiple testing procedure (AMTP) is obtained by adding suitably chosen null hypotheses to the set of hypotheses already rejected by an initial MTP.

Given any initial procedure controlling the generalized family-wise error rate, $gFWER(k)$, we have derived AMTPs for controlling Type I error rates defined as tail probabilities and expected values for arbitrary functions $g(V_n, R_n)$ of the numbers of Type I errors $V_n$ and rejected hypotheses $R_n$.

$$Pr(V_n > k) \implies Pr(g(V_n, R_n) > q) \implies E[g(V_n, R_n)].$$

E.g. 1. $gFWER$: $g(V_n, R_n) = V_n$, number of Type I errors.
E.g. 2. TPPFP: $g(V_n, R_n) = V_n/R_n$, proportion of false positives among the rejected hypotheses.
Augmentation Multiple Testing Procedures

Figure 1: Augmentation MTP. gFWER control via FWER control.
Augmentation Multiple Testing Procedures

Advantages of AMTPs.

- Any gFWER-controlling MTP provides, without additional work, MTPs controlling a broad class of Type I error rates.
- Can build on the large pool of available FWER-controlling MTPs, such as the single-step and step-down maxT and minP procedures.
- The $gFWER(k)$ augmentation procedure, below, guarantees $k$ additional rejected hypotheses.
- Adjusted $p$-values are simply shifted versions of the adjusted $p$-values of the original MTP.
Augmentation Multiple Testing Procedures

• Denote the adjusted $p$-values for the initial MTP by $\tilde{P}_0n(m)$.

• Order the $M$ null hypotheses according to these $p$-values, from smallest to largest, that is, define indices $O_n(m)$, so that $\tilde{P}_0n(O_n(1)) \leq \ldots \leq \tilde{P}_0n(O_n(M))$.

• Then, for a nominal level $\alpha$ test, the initial MTP rejects the $R_n$ null hypotheses

$$R_n \equiv \{ m : \tilde{P}_0n(m) \leq \alpha \}.$$
Augmentation Multiple Testing Procedures

Augmentation procedure for controlling the gFWER:

**FWER \implies gFWER.** For control of \(gFWER(k)\) at level \(\alpha\), reject the \(R_n(\alpha)\) hypotheses specified by the initial FWER-controlling MTP, as well as the next \(A_n(\alpha)\) most significant null hypotheses,

\[
A_n(\alpha) = \min\{k, M - R_n(\alpha)\}.
\]

The adjusted \(p\)-values \(\tilde{P}_{0n}^+(O_n(m))\) for the new gFWER-controlling AMTP are simply \(k\)-shifted versions of the adjusted \(p\)-values of the initial FWER-controlling MTP, with the first \(k\) adjusted \(p\)-values set to zero. That is,

\[
\tilde{P}_{0n}^+(O_n(m)) = \begin{cases} 
0, & \text{if } m \leq k, \\
\tilde{P}_{0n}(O_n(m - k)), & \text{if } m > k.
\end{cases}
\]

The AMTP thus guarantees at least \(k\) rejected hypotheses.
Augmentation Multiple Testing Procedures

**Augmentation procedure for controlling the TPPFP:**

**FWER \implies TPPFP.** For control of $TPPF_P(q)$ at level $\alpha$, reject the $R_n(\alpha)$ hypotheses specified by the initial FWER-controlling MTP, as well as the next $A_n(\alpha)$ most significant null hypotheses,

$$A_n(\alpha) = \max \left\{ m \in \{0, \ldots, M - R_n(\alpha)\} : \frac{m}{m + R_n(\alpha)} \leq q \right\}$$

$$= \min \left\{ \left\lfloor \frac{q R_n(\alpha)}{1 - q} \right\rfloor, M - R_n(\alpha) \right\} ,$$

That is, keep rejecting null hypotheses until the ratio of additional rejections to the total number of rejections reaches the allowed proportion $q$ of false positives.
Augmentation Multiple Testing Procedures

The adjusted p-values $\tilde{P}_{0n}^+(O_n(m))$ for the new TPPFP-controlling AMTP are simply $mq$-shifted versions of the adjusted p-values of the initial FWER-controlling MTP, that is,

$$\tilde{P}_{0n}^+(O_n(m)) = \tilde{P}_{0n}(O_n([1-q)m]), \quad m = 1, \ldots, M.$$ 

Note that while the adjusted p-values for gFWER-controlling are shifted by a constant $k$, the shift $mq$ for the adjusted p-values of the TPPFP-controlling AMTP increases with $m$ as the hypotheses become less significant.

*Ceiling, $[x]$: least integer greater than or equal to $x$.*

*Floor, $\lfloor x \rfloor$: greatest integer less than or equal to $x$.***
Software: *multtest*

[www.bioconductor.org](http://www.bioconductor.org)

- **Multiple testing procedures.**
  - **FWER control:** Bonferroni, Holm (1979), Hochberg (1986), single-step and step-down maxT and minP.
  - **gFWER and TPPFP control:** augmentation of FWER-controlling procedures (van der Laan et al., 2004).

- **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 distribution.** Bootstrap and permutation.
Software: *multtest*

www.bioconductor.org

- **Output.** Parameter estimates, test statistics, unadjusted and adjusted \( p \)-values, cut-offs, confidence regions, estimated null distribution.

- **Graphical summaries.** Type I error rate vs. \# rejections, \# rejections vs. adjusted \( p \)-values, adjusted \( p \)-values vs. test statistics (“volcano” plots).

- **Software design.**
  - **Closures:** new tests can easily be added.
  - **Class/method** object-oriented programming.

Bioconductor R package: *multtest*.

Authors: Katherine S. Pollard, Yongchao Ge, Sandrine Dudoit.
Acknowledgments

Joint work with

- Mark J. van der Laan, Division of Biostatistics, UC Berkeley.
- Merrill D. Birkner, Division of Biostatistics, UC Berkeley.
- Sündüz Keleş, Departments of Biostatistics and Statistics, University of Wisconsin, Madison.
- Katherine S. Pollard, Center for Biomolecular Science & Engineering, UC Santa Cruz.
- Simon E. Cawley, Affymetrix.
- Mélanie Courtine, Laboratoire d’Informatique, Université Pierre et Marie Curie – Paris 6, France.