15

Multiple Testing Procedures: the multtest Package and Applications to Genomics

K. S. Pollard, S. Dudoit, and M. J. van der Laan

Abstract

The Bioconductor R package multtest implements widely applicable resampling-based single-step and stepwise multiple testing procedures (MTP) for controlling a broad class of Type I error rates. The current version of multtest provides MTPs for tests concerning means, differences in means, and regression parameters in linear and Cox proportional hazards models. Typical testing scenarios are illustrated by applying various MTPs implemented in multtest to the Acute Lymphoblastic Leukemia (ALL) data set of Chiaretti et al. (2004), with the aim of identifying genes whose expression measures are associated with (possibly censored) biological and clinical outcomes.

15.1 Introduction

Current statistical inference problems in biomedical and genomic data analysis routinely involve the simultaneous test of thousands, or even millions, of null hypotheses. Examples include:

- identification of differentially expressed genes in microarray experiments, i.e., genes whose expression measures are associated with possibly censored responses or covariates;
- tests of association between gene expression measures and Gene Ontology (GO) annotation;

- identification of transcription factor binding sites in ChIP-Chip experiments (Keleş et al., 2004);
- genetic mapping of complex traits using single nucleotide polymorphisms (SNP).

The above testing problems share the following general characteristics: inference for high-dimensional multivariate distributions, with complex and unknown dependence structures among variables; a broad range of parameters of interest, e.g. regression coefficients and correlations; many null hypotheses, in the thousands or even millions; complex dependence structures among test statistics.

Motivated by these applications, we have developed resampling-based single-step and stepwise multiple testing procedures (MTP) for controlling a broad class of Type I error rates. The main steps in applying a MTP are listed in the flowchart of Table 15.1. The different components of our multiple testing methodology are treated in detail in a collection of related articles (Dudoit et al., 2004a,b; Pollard and van der Laan, 2004; van der Laan et al., 2004a,b) and a book in preparation (Dudoit and van der Laan, 2004). In order to make this general methodology accessible, we have implemented several MTPs in the Bioconductor R package multtest, which is the subject of the current chapter. An expanded version of this chapter is available on-line as a technical report (Pollard et al., 2004).

15.2 Multiple hypothesis testing methodology

15.2.1 Multiple hypothesis testing framework

Hypothesis testing is concerned with using observed data to test hypotheses, i.e., make decisions, regarding properties of the unknown data generating distribution. For example, microarray experiments might be conducted on a sample of patients in order to identify genes whose expression levels are associated with survival. Below, we discuss in turn the main ingredients of a multiple testing problem.

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 an element of a particular *statistical model* $\mathcal{M}$ (i.e., a set of possibly non-parametric distributions). In a microarray experiment, for example, X is a vector of gene expression measurements, which we observe for each of n arrays.

Null and alternative hypotheses. Define M *null hypotheses* $H_0(m) \equiv \mathrm{I}[P \in \mathcal{M}(m)]$ in terms of a collection of *submodels*, $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \ldots, M$, for the data generating distribution P. The corresponding *alter-*

Table 15.1. *Multiple hypothesis testing flowchart.*

Provide data set

`MTP` arguments: `X`, `W`, `Y`, `Z`, `Z.incl`, and `Z.test`

⇓

Define parameters of interest, $\psi(m)$

⇓

Define null and alternative hypotheses, $H_0(m)$ **and** $H_1(m)$

⇓

Specify test statistics, $T_n(m)$

`MTP` arguments: `test`, `robust`, `standardize`, `alternative`, and `psi0`

⇓

Estimate test statistics null distribution, Q_{0n}

`MTP` arguments: `nulldist` and `B`

⇓

Select Type I error rate, $\theta(F_{V_n,R_n})$

`MTP` arguments: `typeone` and `alpha` (and also `k` and `q`)

⇓

Apply MTP

`MTP` argument: `method`

FWER	$Pr(V_n > 0)$	Single-step maxT procedure (sec. 15.2.3) Single-step minP procedure (sec. 15.2.3) Step-down maxT procedure (sec. 15.2.4) Step-down minP procedure (sec. 15.2.4)
gFWER	$Pr(V_n > k)$	Single-step $T(k+1)$ procedure (sec. 15.2.3) Single-step $P(k+1)$ procedure (sec. 15.2.3) Augmentation procedure (sec. 15.2.5)
TPPFP	$Pr(V_n/R_n > q)$	Augmentation procedure (sec. 15.2.5)
General	$\theta(F_{V_n})$	Single-step common cutoff procedure (sec. 15.2.3) Single-step common quantile procedure (sec. 15.2.3)

⇓

Summarize results

adjusted p-values, rejection regions, and confidence regions

`MTP` arguments: `get.adjp`, `get.cutoff`, and `get.cr`

native hypotheses are $H_1(m) \equiv \mathrm{I}[P \notin \mathcal{M}(m)]$. In many testing problems, the submodels concern *parameters*, i.e., functions of the data generating distribution P, $\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M)$, such as means, differences in means, correlation coefficients, and regression parameters.

Test statistics. A testing procedure is a *data-driven* rule for deciding whether or not to *reject* each of the M null hypotheses $H_0(m)$ based on an M-vector of *test statistics*, $T_n = (T_n(m) : m = 1, \ldots, M)$, that are functions of the observed data. Denote the typically unknown (finite sample) *joint distribution* of the test statistics T_n by $Q_n = Q_n(P)$.

Single-parameter null hypotheses are commonly tested using *t-statistics*, i.e., standardized differences,

$$T_n(m) \equiv \frac{\text{Estimator} - \text{Null value}}{\text{Standard error}} = \sqrt{n}\,\frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)}. \tag{15.1}$$

For tests of means, $T_n(m)$ is the usual one-sample or two-sample t-statistic, where $\psi_n(m)$ and $\sigma_n(m)$ are based on empirical means and variances, respectively. In some settings, it may be appropriate to use (unstandardized) *difference statistics*, $T_n(m) \equiv \sqrt{n}[\psi_n(m) - \psi_0(m)]$ (Pollard and van der Laan, 2004). Test statistics for other types of null hypotheses include F-statistics, χ^2-statistics, and likelihood ratio statistics.

Multiple testing procedure. A *multiple testing procedure* (MTP) provides *rejection regions*, $\mathcal{C}_n(m)$, i.e., sets of values for each test statistic $T_n(m)$ that lead to the decision to reject the null hypothesis $H_0(m)$. In other words, a MTP produces a random (i.e., data-dependent) subset $\mathcal{R}_n$ of rejected hypotheses that estimates the set of true positives,

$$\mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) \equiv \{m : H_0(m) \text{ is rejected}\} = \{m : T_n(m) \in \mathcal{C}_n(m)\}, \tag{15.2}$$

where the long notation $\mathcal{R}(T_n, Q_{0n}, \alpha)$ emphasizes that the MTP depends on: (i) the *data* through the *test statistics* T_n; (ii) a (estimated) test statistics *null distribution*, Q_{0n}, for deriving rejection regions; and (iii) the *nominal level* α, i.e., the desired upper bound for a suitably defined Type I error rate. Unless specified otherwise, it is assumed that large values of the test statistic, $T_n(m)$, provide evidence against the corresponding null hypothesis $H_0(m)$.

Example. Suppose that, as in the analysis of the ALL data set of Chiaretti et al. (2004) (Section 15.4), one is interested in identifying genes that are differentially expressed in two populations of ALL cancer patients, those with the B-cell subtype and those with the T-cell subtype. The data consist of random vectors X of microarray expression measures on M genes and an indicator Y for the ALL subtype (1 for B-cell, 0 for T-cell). Then, the parameter of interest is an M-vector of differences in mean expression mea-

Table 15.2. *Type I and Type II errors in multiple hypothesis testing.* $\mathcal{H}_0$ is the set of true null hypotheses, $\mathcal{H}_1$ is the set of false null hypotheses (i.e., true positives), and $\mathcal{R}_n$ is the set of rejected null hypotheses.

		Null hypotheses not rejected	Null hypotheses rejected	
Null hypotheses	true	$\lvert\mathcal{R}_n^c \cap \mathcal{H}_0\rvert$	$V_n = \lvert\mathcal{R}_n \cap \mathcal{H}_0\rvert$ (Type I errors)	$h_0 = \lvert\mathcal{H}_0\rvert$
	false	$U_n = \lvert\mathcal{R}_n^c \cap \mathcal{H}_1\rvert$ (Type II errors)	$\lvert\mathcal{R}_n \cap \mathcal{H}_1\rvert$	$h_1 = \lvert\mathcal{H}_1\rvert$
		$M - R_n$	$R_n = \lvert\mathcal{R}_n\rvert$	M

sures in the two populations, $\psi(m) = E[X(m)|Y = 1] - E[X(m)|Y = 0]$, $m = 1, \ldots, M$. To identify genes with higher mean expression measures in the B-cell compared to T-cell ALL subjects, one can test the one-sided null hypotheses $H_0(m) = \mathrm{I}[\psi(m) \leq 0]$ vs. the alternative hypotheses $H_1(m) = \mathrm{I}[\psi(m) > 0]$, using two-sample Welch t-statistics

$$T_n(m) \equiv \frac{\bar{X}_{1,n_1}(m) - \bar{X}_{0,n_0}(m)}{\sqrt{n_0^{-1}(m)\sigma^2_{0,n_0}(m) + n_1^{-1}(m)\sigma^2_{1,n_1}(m)}}, \tag{15.3}$$

where $n_k(m)$, $\bar{X}_{k,n_k}(m)$, and $\sigma^2_{k,n_k}(m)$ denote, respectively, the sample sizes, sample means, and sample variances, for patients with tumor subtype k, $k = 0, 1$.

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 a true null hypothesis, and a *false negative*, or *Type II error*, is committed when the test procedure fails to reject a false null hypothesis. The situation can be summarized by Table 15.2.

Type I error rates. When testing multiple hypotheses, there are many possible definitions for the Type I error rate and power of a test procedure. Accordingly, we define 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 (Dudoit et al., 2004b; Dudoit and van der Laan, 2004). Such a general representation covers the following commonly-used Type I error rates.

Generalized family-wise error rate (gFWER), or probability of at least $(k+1)$ Type I errors,

$$gFWER(k) \equiv Pr(V_n > k). \tag{15.4}$$

When $k = 0$, the gFWER is the usual *family-wise error rate* (FWER), or probability of at least one Type I error, $FWER \equiv Pr(V_n > 0)$.

Tail probabilities for the proportion of false positives (TPPFP) among the rejected hypotheses,

$$TPPFP(q) \equiv Pr(V_n/R_n > q), \qquad q \in (0,1). \tag{15.5}$$

False discovery rate (FDR), or expected value of the proportion of false positives among the rejected hypotheses (Benjamini and Hochberg, 1995),

$$FDR \equiv E[V_n/R_n]. \tag{15.6}$$

The convention that $V_n/R_n \equiv 0$ if $R_n = 0$ is used. Error rates based on the *proportion* of false positives (e.g., TPPFP and FDR) are especially appealing for large-scale testing problems such as those encountered in genomics, compared to error rates based on the *number* of false positives (e.g., gFWER), as they do not increase exponentially with the number of tested hypotheses.

Adjusted p-values. The notion of p-value extends directly to multiple testing problems, as follows. Given a MTP $\mathcal{R}_n(\alpha) = \mathcal{R}(T_n, Q_{0n}, \alpha)$, the *adjusted p-value* $\widetilde{P}_{0n}(m) = \widetilde{P}(T_n, Q_{0n})(m)$, for null hypothesis $H_0(m)$, is defined as the smallest Type I error level α at which one would reject $H_0(m)$, that is,

$$\begin{aligned} \widetilde{P}_{0n}(m) &\equiv \inf\{\alpha \in [0,1] : m \in \mathcal{R}_n(\alpha)\} \\ &= \inf\{\alpha \in [0,1] : T_n(m) \in \mathcal{C}_n(m)\}, \qquad m = 1, \ldots, M. \end{aligned} \tag{15.7}$$

As in single hypothesis tests, the smaller the adjusted p-value, the stronger the evidence against the corresponding null hypothesis. Reporting the results of a MTP in terms of adjusted p-values, as opposed to the binary decisions to reject or not the hypotheses, provides *flexible summaries* that can be used to compare different MTPs and do not require specifying the level α ahead of time.

Confidence regions. For the test of single-parameter null hypotheses and for any Type I error rate of the form $\theta(F_{V_n})$, Pollard and van der Laan (2004) and Dudoit and van der Laan (2004) provide results on the correspondence between single-step MTPs and θ-specific *confidence regions*.

15.2.2 Test statistics null distribution

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 $Q_n(P)$. For error rates $\theta(F_{V_n})$ (e.g., gFWER), defined as arbitrary parameters of the distribution of the number of Type I errors V_n, we propose as null distribution the asymptotic distribution $Q_0 = Q_0(P)$ of the M-vector Z_n of null value shifted and scaled test statistics (Dudoit and van der Laan, 2004; Dudoit et al., 2004b; Pollard and van der Laan, 2004; van der Laan et al., 2004b),

$$Z_n(m) \equiv \sqrt{\min\left\{1, \frac{\tau_0(m)}{Var[T_n(m)]}\right\}}\Big\{T_n(m) + \lambda_0(m) - E[T_n(m)]\Big\}. \quad (15.8)$$

For the test of single-parameter null hypotheses using t-statistics, the null values are $\lambda_0(m) = 0$ and $\tau_0(m) = 1$. 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. By shifting the test statistics $T_n(m)$ as in Equation (15.8), the number of Type I errors V_0 under the null distribution Q_0, is asymptotically stochastically greater than the number of Type I errors V_n under the true distribution $Q_n = Q_n(P)$.

Note that 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, Westfall and Young (Westfall and Young, 1993), p. 42-43) are therefore irrelevant to our approach. In addition, we propose a *null distribution for the test statistics*, $T_n \sim Q_0$, and not a data generating null distribution, $X \sim P_0 \in \cap_{m=1}^{M} \mathcal{M}(m)$. The latter practice does not necessarily provide proper Type I error control, 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.

Resampling procedures, such as the bootstrap procedure of section 15.2.2, may be used to conveniently obtain consistent estimators Q_{0n} of the null distribution Q_0 and of the corresponding test statistic cutoffs and adjusted p-values (Dudoit and van der Laan, 2004; Dudoit et al., 2004b; Pollard and van der Laan, 2004; van der Laan et al., 2004b). This bootstrap procedure is implemented in the internal function `boot.resample` and may be specified via the arguments `nulldist` and `B` of the main user-level function `MTP`.

Having selected a suitable test statistics null distribution, there remains the main task of specifying rejection regions for each null hypothesis, i.e., cutoffs for each test statistic, such that the Type I error rate is controlled at a desired level α. Next, we summarize the approaches to this task that

Bootstrap estimation of the null distribution Q_0

1. Let $P_n^\star$ denote an estimator of the data generating distribution P.
2. Generate B bootstrap samples, each consisting of n *i.i.d.* realizations of a random variable $X^\# \sim P_n^\star$. For the *non-parametric bootstrap*, samples of size n are drawn at random, with replacement from the observed data.
3. For the bth bootstrap sample, $b = 1, \ldots, B$, compute an M-vector of test statistics, and arrange these in an $M \times B$ matrix, $\mathbf{T}_n^\# = \left[T_n^\#(m, b)\right]$, with rows corresponding to the M null hypotheses and columns to the B bootstrap samples.
4. Compute row means, $E[{T_n}^\#(m, \cdot)]$, and row variances, $Var[{T_n}^\#(m, \cdot)]$, of the matrix $\mathbf{T}_n^\#$, to yield estimates of the true means $E[T_n(m)]$ and variances $Var[T_n(m)]$ of the test statistics, respectively.
5. Obtain an $M \times B$ matrix, $\mathbf{Z}_n^\# = \left[Z_n^\#(m, b)\right]$, of null value shifted and scaled bootstrap statistics $Z_n^\#(m, b)$, by row-shifting and scaling the matrix $\mathbf{T}_n^\#$ as in Equation (15.8) using the bootstrap estimates of $E[T_n(m)]$ and $Var[T_n(m)]$ and the user-supplied null values $\lambda_0(m)$ and $\tau_0(m)$.
6. The bootstrap estimate Q_{0n} of the null distribution Q_0 is the empirical distribution of the B columns $Z_n^\#(\cdot, b)$ of matrix $\mathbf{Z}_n^\#$.

have been implemented in the multtest package. The chosen procedure is specified using the `method` argument to the function `MTP`.

15.2.3 Single-step procedures for controlling general Type I error rates $\theta(F_{V_n})$

Control of a Type I error rate $\theta(F_{V_n})$ can be obtained by substituting the *known, null distribution* F_{R_0} of the number of rejected hypotheses for the *unknown, true distribution* F_{V_n} of the number of Type I errors. We propose the following single-step common cutoff and common quantile procedures (Dudoit et al., 2004b; Pollard and van der Laan, 2004).

General θ-controlling single-step common cutoff procedure

The set of rejected hypotheses is of the form $\mathcal{R}_n(\alpha) \equiv \{m : T_n(m) > c_0\}$, where the common cutoff c_0 is the *smallest* (i.e., least conservative) value for which $\theta(F_{R_0}) \leq \alpha$. For $gFWER(k)$ control, the procedure is based on the $(k+1)$*st ordered test statistic*. The adjusted p-values for the *single-step*

$T(k+1)$ *procedure* are given by

$$\widetilde{p}_{0n}(m) = Pr_{Q_0}\left(Z^\circ(k+1) \geq t_n(m)\right), \qquad m = 1, \ldots, M, \tag{15.9}$$

where $Z^\circ(m)$ denotes the mth ordered component of $Z = (Z(m) : m = 1, \ldots, M) \sim Q_0$, so that $Z^\circ(1) \geq \ldots \geq Z^\circ(M)$. For FWER control ($k = 0$), one recovers the *single-step maxT procedure*.

General θ-controlling single-step common quantile procedure

The set of rejected hypotheses is of the form $\mathcal{R}_n(\alpha) \equiv \{m : T_n(m) > c_0(m)\}$, where $c_0(m) = Q_{0,m}^{-1}(\delta_0)$ is the δ_0-quantile of the marginal null distribution $Q_{0,m}$ of the test statistic for the mth null hypothesis, i.e., the smallest value c such that $Q_{0,m}(c) = Pr_{Q_0}(Z(m) \leq c) \geq \delta_0$ for $Z \sim Q_0$. Here, δ_0 is chosen as the *smallest* (i.e., least conservative) value for which $\theta(F_{R_0}) \leq \alpha$.

For $gFWER(k)$ control, the procedure is based on the $(k+1)$*st ordered unadjusted p-value*. Specifically, let $\bar{Q}_{0,m} \equiv 1 - Q_{0,m}$ denote the survivor functions for the marginal null distributions $Q_{0,m}$ and define unadjusted p-values $P_0(m) \equiv \bar{Q}_{0,m}[Z(m)]$ and $P_{0n}(m) \equiv \bar{Q}_{0,m}[T_n(m)]$, for $Z \sim Q_0$ and $T_n \sim Q_n$, respectively. The adjusted p-values for the *single-step* $P(k+1)$ *procedure* are given by

$$\widetilde{p}_{0n}(m) = Pr_{Q_0}\left[P_0^\circ(k+1) \leq p_{0n}(m)\right], \qquad m = 1, \ldots, M, \tag{15.10}$$

where $P_0^\circ(m)$ denotes the mth ordered component of the M-vector of unadjusted p-values $P_0 = [P_0(m) : m = 1, \ldots, M]$, so that $P_0^\circ(1) \leq \ldots \leq P_0^\circ(M)$. For FWER control ($k = 0$), one recovers the *single-step minP procedure*.

15.2.4 Step-down procedures for controlling the family-wise error rate

Step-down MTPs consider hypotheses successively, from most significant to least significant, with further tests depending on the outcome of earlier ones. van der Laan et al. (2004b) propose step-down common cutoff (maxT) and common quantile (minP) procedures for controlling the family-wise error rate, FWER.

FWER-controlling step-down common cutoff (maxT) procedure

Let $O_n(m)$ denote the indices for the ordered test statistics $T_n(m)$, so that $T_n(O_n(1)) \geq \ldots \geq T_n(O_n(M))$. Consider the distributions of maxima of test statistics over the nested subsets of ordered null hypotheses $\overline{\mathcal{O}}_n(h) \equiv \{O_n(h), \ldots, O_n(M)\}$. The adjusted p-values are given by

$$\widetilde{p}_{0n}[o_n(m)] = \max_{h=1,\ldots,m} Pr_{Q_0}\left\{\max_{l \in \overline{\mathcal{O}}_n(h)} Z(l) \geq t_n[o_n(h)]\right\}, \tag{15.11}$$

where $Z = [Z(m) : m = 1, \ldots, M] \sim Q_0$.

FWER-controlling step-down common quantile (minP) procedure.
Let $O_n(m)$ denote the indices for the ordered unadjusted p-values $P_{0n}(m)$, so that $P_{0n}[O_n(1)] \leq \ldots \leq P_{0n}[O_n(M)]$. Consider the distributions of minima of unadjusted p-values over the nested subsets of ordered null hypotheses $\overline{\mathcal{O}}_n(h) \equiv \{O_n(h), \ldots, O_n(M)\}$. The adjusted p-values are given by

$$\widetilde{p}_{0n}(o_n(m)) = \max_{h=1,\ldots,m} Pr_{Q_0}\left\{\min_{l \in \overline{\iota}_n(h)} P_0(l) \leq p_{0n}[o_n(h)]\right\}, \qquad (15.12)$$

where $P_0(m) \equiv \bar{Q}_{0,m}[Z(m)]$ and $P_{0n}(m) \equiv \bar{Q}_{0,m}[T_n(m)]$, for $Z \sim Q_0$ and $T_n \sim Q_n$, respectively.

15.2.5 Augmentation multiple testing procedures for controlling tail probability error rates

van der Laan et al. (2004a), and subsequently Dudoit et al. (2004a) and Dudoit and van der Laan (2004), propose *augmentation multiple testing procedures* (AMTP), obtained by adding suitably chosen null hypotheses to the set of null hypotheses already rejected by an initial gFWER-controlling MTP. Adjusted p-values for the AMTP are shown to be simply shifted versions of the adjusted p-values of the original MTP. Denote the adjusted p-values for the initial FWER-controlling procedure $\mathcal{R}_n(\alpha)$ by $\widetilde{P}_{0n}(m)$. Order the M null hypotheses according to these p-values, from smallest to largest, that is, define indices $O_n(m)$, so that $\widetilde{P}_{0n}[O_n(1)] \leq \ldots \leq \widetilde{P}_{0n}[O_n(M)]$.

gFWER-controlling augmentation multiple testing procedure

For control of $gFWER(k)$ at level α, given an initial FWER-controlling procedure $\mathcal{R}_n(\alpha)$, reject the $R_n(\alpha) = |\mathcal{R}_n(\alpha)|$ null hypotheses specified by this MTP, as well as the next $A_n(\alpha)$ most significant hypotheses,

$$A_n(\alpha) = \min\{k, M - R_n(\alpha)\}. \qquad (15.13)$$

The adjusted p-values $\widetilde{P}^+_{0n}[O_n(m)]$ 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,

$$\widetilde{P}^+_{0n}[O_n(m)] = \begin{cases} 0, & \text{if } m \leq k \\ \widetilde{P}_{0n}[O_n(m-k)], & \text{if } m > k \end{cases}. \qquad (15.14)$$

The AMTP thus guarantees at least k rejected hypotheses.

TPPFP-controlling augmentation multiple testing procedure

For control of $TPPFP(q)$ at level α, given an initial FWER-controlling procedure $\mathcal{R}_n(\alpha)$, reject the $R_n(\alpha) = |\mathcal{R}_n(\alpha)|$ null hypotheses specified by this MTP, as well as the next $A_n(\alpha)$ most significant hypotheses,

$$\begin{aligned} A_n(\alpha) &= \max\left\{m \in \{0,\ldots,M - R_n(\alpha)\} : \frac{m}{m + R_n(\alpha)} \leq q\right\} \\ &= \min\left\{\left\lfloor \frac{qR_n(\alpha)}{1-q} \right\rfloor, M - R_n(\alpha)\right\}, \end{aligned} \tag{15.15}$$

where the *floor* $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, i.e., $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$. 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. The adjusted p-values $\widetilde{P}^+_{0n}[O_n(m)]$ for the new TPPFP-controlling AMTP are simply mq-shifted versions of the adjusted p-values of the initial FWER-controlling MTP. That is,

$$\widetilde{P}^+_{0n}(O_n(m)) = \widetilde{P}_{0n}(O_n(\lceil (1-q)m \rceil)), \qquad m = 1,\ldots,M, \tag{15.16}$$

where the *ceiling* $\lceil x \rceil$ denotes the least integer greater than or equal to x.

FDR-controlling procedures

Given any TPPFP-controlling procedure, van der Laan et al. (2004a) derive two simple (conservative) FDR-controlling procedures. The more general and conservative procedure controls the FDR at nominal level α, by controlling $TPPFP(\alpha/2)$ at level $\alpha/2$. The less conservative procedure controls the FDR at nominal level α, by controlling $TPPFP(1 - \sqrt{1-\alpha})$ at level $1 - \sqrt{1-\alpha}$. The reader is referred to the original article for details and proofs of FDR control (Section 2.4, Theorem 3). In what follows, we refer to these two MTPs as *conservative* and *restricted*, respectively.

15.3 Software implementation: R multtest package

The MTPs proposed in Sections 15.2.3 - 15.2.5 are implemented in the latest version of the Bioconductor R package multtest (Version 1.5.4). We stress that *all* the bootstrap-based MTPs implemented in multtest can be performed using the main user-level function `MTP`. Note that the multtest package also provides several simple, marginal FWER-controlling MTPs, available through the `mt.rawp2adjp` function, which takes a vector of unadjusted p-values as input and returns the corresponding adjusted p-values. For greater detail on multtest functions, the reader is referred to the pack-

age documentation, in the form of help files, e.g., `?MTP`, and vignettes, e.g., `openVignette("multtest")`.

15.3.1 Resampling-based multiple testing procedures: `MTP` function

The main user-level function for resampling-based multiple testing is `MTP`.

```
> args(MTP)

function (X, W = NULL, Y = NULL, Z = NULL, Z.incl = NULL,
  Z.test = NULL, na.rm = TRUE, test = "t.twosamp.unequalvar",
  robust = FALSE, standardize = TRUE, alternative = "two.sided",
  psi0 = 0, typeone = "fwer", k = 0, q = 0.1,
  fdr.method = "conservative", alpha = 0.05, nulldist = "boot",
  B = 1000, method = "ss.maxT", get.cr = FALSE, get.cutoff = FALSE,
  get.adjp = TRUE, keep.nulldist = FALSE, seed = NULL)
```

INPUT.

Data. The data, `X`, consist of a J-dimensional random vector, observed on each of n sampling units (patients, cell lines, mice, etc.). Other data components include weights `W`, a possibly censored continuous or polychotomous outcome `Y`, and additional covariates `Z`, whose use is specified with the arguments `Z.incl` and `Z.test`. The argument `na.rm` controls the treatment of missing values (`NA`). It is `TRUE` by default, so that an observation with a missing value in any of the data objects' jth component ($j = 1, \ldots, J$) is excluded from the computation of any test statistic based on this jth variable.

Test statistics. In the current implementation of multtest, the following test statistics are available through the argument `test`: one-sample t-statistics for tests of means; equal and unequal variance two-sample t-statistics for tests of differences in means; paired t-statistics; multi-sample F-statistics for tests of differences in means in one-way and two-way designs; t-statistics for tests of regression coefficients in linear models and Cox proportional hazards survival models. *Robust*, *rank-based* versions of the above test statistics can be specified by setting the argument `robust` to `TRUE` (the default value is `FALSE`).

Type I error rate. The `MTP` function controls by default the FWER (argument `typeone="fwer"`). Augmentation procedures (Section 15.2.5), controlling other Type I error rates such as the gFWER, TPPFP, and FDR, can be specified through the argument `typeone`. Details regarding the related arguments `k`, `q`, and `fdr.method` are available in the package documentation. The nominal level of the test is determined by the argument `alpha`, by default 0.05.

Test statistics null distribution. The test statistics null distribution is estimated by default using the non-parametric version of the bootstrap procedure of section 15.2.2 (argument `nulldist="boot"`). Permutation null distributions are also available via `nulldist="perm"`. The number of resampling steps is specified by the argument `B`, by default 1,000.

Multiple testing procedures. The `MTP` function implements the single-step and step-down (common cutoff) maxT and (common quantile) minP MTPs for FWER control, described in Sections 15.2.3 and 15.2.4, and specified through the argument `method`. In addition, augmentation procedures (AMTPs) are implemented in the functions `fwer2gfwer`, `fwer2tppfp`, and `fwer2fdr`, which take FWER adjusted p-values as input and return augmentation adjusted p-values for control of the gFWER, TPPFP, and FDR, respectively. These AMTPs can also be applied directly via the `typeone` argument of the main function `MTP`.

Output control. Additional arguments allow the user to specify which combination of MTP results should be returned.

OUTPUT.

The S4 class/method object-oriented programming approach was adopted to summarize the results of a MTP. The output of the `MTP` function is an instance of the *class MTP*, with the following *slots*,

```
> slotNames("MTP")

 [1] "statistic" "estimate"  "sampsize"  "rawp"
 [5] "adjp"      "conf.reg"  "cutoff"    "reject"
 [9] "nulldist"  "call"      "seed"
```

MTP results. An instance of the *MTP* class contains slots for the following MTP results: `statistic`, an M-vector of test statistics; `estimate`, an M-vector of estimated parameters; `rawp`, an M-vector of unadjusted p-values; `adjp`, an M-vector of adjusted p-values; `conf.reg`, lower and upper simultaneous confidence limits for the parameter vector; `cutoff`, cutoffs for the test statistics; `reject`, rejection indicators (`TRUE` for a rejected null hypothesis).

Null distribution. The `nulldist` slot contains the $M \times B$ *matrix* for the estimated test statistics null distribution.

Reproducibility. The slot `call` contains the call to the function `MTP`, and `seed` is an integer specifying the state of the random number generator used to create the resampled data sets.

15.3.2 Numerical and graphical summaries

The following *methods* were defined to operate on *MTP* instances and summarize the results of a MTP. The `print` method returns a description of an object of class *MTP*. The `summary` method returns a list with the the following components: `rejections`, number(s) of rejected hypotheses; `index`, indices for ordering the hypotheses according to significance; `summaries`, six number summaries of the distributions of the adjusted p-values, unadjusted p-values, test statistics, and parameter estimates. The `plot` method produces graphical summaries of the results of a MTP. The type of display may be specified via the `which` argument. Methods are also provided for subsetting (`[`) and conversion (`as.list`).

15.4 Applications: ALL microarray data set

15.4.1 ALL data package and initial gene filtering

We illustrate some of the functionality of the multtest package using the Acute Lymphoblastic Leukemia (ALL) microarray data set of Chiaretti et al. (2004), available in the data package ALL. The main object in this package is `ALL`, an instance of the class *exprSet*. The genes-by-subjects matrix of 12,625 Affymetrix *expression measures* (chip series HG-U95Av2) for each of 128 ALL patients is provided in the `exprs` slot of `ALL`. The `phenoData` slot contains 21 *phenotypes* (i.e., patient level responses and covariates) for each patient. Note that the expression measures have been obtained using the three-step robust multichip average (RMA) preprocessing method, implemented in the package affy. In particular, the expression measures have been subject to a base 2 logarithmic transformation. For greater detail, please consult the ALL package documentation and Appendix A.1.1.

```
> library("ALL")
> library("hgu95av2")
> data(ALL)
```

Our goal is to identify genes whose expression measures are associated with (possibly censored) biological and clinical outcomes such as: tumor cellular subtype (B-cell vs. T-cell), tumor molecular subtype (BCR/ABL, NEG, ALL1/AF4), and time to relapse. Alternative analyses of this data set are discussed in Chapters 10, 12, 16, 17, and 23. Before applying the MTPs, we perform initial gene filtering as in Chiaretti et al. (2004) and retain only those genes for which: (i) at least 20% of the subjects have a measured intensity of at least 100 and (ii) the coefficient of variation (i.e., the ratio of the standard deviation to the mean) of the intensities across samples is between 0.7 and 10. These two filtering criteria can be readily applied using functions from the genefilter package.

```
> ffun <- filterfun(pOverA(p = 0.2, A = 100), cv(a = 0.7,
+     b = 10))
> filt <- genefilter(2^exprs(ALL), ffun)
> filtALL <- ALL[filt, ]

> filtX <- exprs(filtALL)
> pheno <- pData(filtALL)
```

The new filtered data set, `filtALL`, contains expression measures on 431 genes, for 128 patients.

15.4.2 Association of expression measures and tumor cellular subtype: Two-sample t-statistics

In this example we examine use of FWER-controlling step-down minP MTP with two-sample Welch t-statistics and bootstrap null distribution.

Different tissues are involved in ALL tumors of the B-cell and T-cell subtypes. The phenotypic data include a variable, `BT`, which encodes the tissue type and stage of differentiation. In order to identify genes with higher mean expression measures in B-cell ALL patients compared to T-cell ALL patients, we create an indicator variable, `Bcell` (1 for B-cell, 0 for T-cell), and compute, for each gene, a two-sample Welch (unequal variance) t-statistic. We choose to control the FWER using the bootstrap-based step-down minP procedure with $B = 100$ bootstrap iterations, although more bootstrap iterations are recommended in practice.

```
> table(pData(ALL)$BT)

 B B1 B2 B3 B4  T T1 T2 T3 T4
 5 19 36 23 12  5  1 15 10  2

> Bcell <- rep(0, length(pData(ALL)$BT))
> Bcell[grep("B", as.character(pData(ALL)$BT))] <- 1

> seed <- 99
> BT.boot <- cache("BT.boot", MTP(X = filtX, Y = Bcell,
+     alternative = "greater", B = 100, method = "sd.minP",
+     seed = seed))

running bootstrap...
iteration = 100
```

Let us examine the results of the MTP stored in the object `BT.boot`.

```
> summary(BT.boot)

MTP:  sd.minP
Type I error rate:  fwer

  Level Rejections
1  0.05        273
```

```
            Min. 1st Qu. Median  Mean 3rd Qu.  Max.
adjp        0.00   0.000  0.000 0.364   1.000  1.00
rawp        0.00   0.000  0.000 0.354   1.000  1.00
statistic -34.40  -1.570  2.010 2.060   5.380 22.30
estimate   -4.66  -0.317  0.381 0.326   0.995  4.25
```

The `summary` method prints the name of the MTP (here, `sd.minP`, for step-down minP), the Type I error rate (here, `fwer`), the number of rejections at each Type I error rate level specified in `alpha` (here, 273 at level $\alpha = 0.05$), and six number summaries (mean and quantiles) of the adjusted p-values, unadjusted p-values, test statistics, and parameter estimates (here, difference in means).

The following commands may be used to obtain a list of genes that are differentially expressed in B-cell vs. T-cell ALL patients at nominal FWER level $\alpha = 0.05$, i.e., genes with adjusted p-values less than or equal to 0.05. Functions from the annotate and annaffy packages may then be used to obtain annotation information on these genes (e.g., gene names, PubMed abstracts, GO terms) and to generate HTML tables of the results (see Chapters 7 and 9). Here, we list the names of the first two genes only.

```
> BT.diff <- BT.boot@adjp <= 0.05
> BT.AffyID <- geneNames(filtALL)[BT.diff]
> mget(BT.AffyID[1:2], env = hgu95av2GENENAME)

$"1005_at"
[1] "dual specificity phosphatase 1"

$"1065_at"
[1] "fms-related tyrosine kinase 3"
```

Various graphical summaries of the results may be obtained using the `plot` method, by selecting appropriate values of the argument `which`. Figure 15.1 displays four such plots. We see (top left) that the number of rejections increases slightly when nominal FWER is greater than 0.6, and then increases quickly as FWER approaches 1. Similarly, the adjusted p-values for many genes are close to either 0 or 1 (top right) and the test statistics for genes with small p-values do not overlap with those for genes with p-values close to 1 (bottom left). Together these results indicate that there is a clear separation between the rejected and accepted hypotheses, i.e., between genes that are declared differentially expressed and those that are not.

```
> par(mfrow = c(2, 2))
> plot(BT.boot)
```

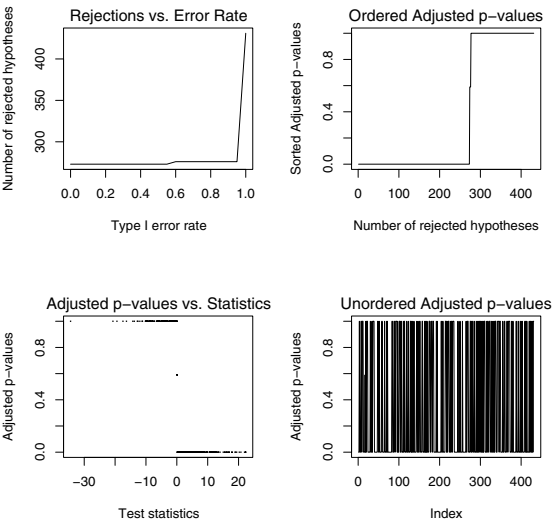

Figure 15.1. B-cell vs. T-cell ALL – FWER-controlling step-down minP MTP. By default, four graphical summaries are produced by the `plot` method for instances of the class *MTP*.

15.4.3 Augmentation procedures

In the context of microarray gene expression data analysis or other high-dimensional inference problems, one is often willing to tolerate some false positives, provided their number is small in comparison to the number of rejected hypotheses. In this case, the FWER is not a suitable choice of Type I error rate, and one should consider other rates that lead to larger sets of rejected hypotheses. The augmentation procedures of Section 15.2.5, implemented in the function `MTP`, allow one to reject additional hypotheses, while controlling an error rate such as the generalized family-wise error rate (gFWER), the tail probability for the proportion of false positives (TPPFP), or the false discovery rate (FDR). We illustrate the use of the `fwer2tppfp` and `fwer2fdr` functions, but note that the gFWER, TPPFP, and FDR can also be controlled directly using the main `MTP` function, with appropriate choices of arguments `typeone`, `k`, `q`, and `fdr.method`.

TPPFP control.

```
> q <- c(0.05, 0.1, 0.25)
> BT.tppfp <- fwer2tppfp(adjp = BT.boot@adjp, q = q)
> comp.tppfp <- cbind(BT.boot@adjp, BT.tppfp)
> mtps <- c("FWER", paste("TPPFP(", q, ")", sep = ""))
> mt.plot(adjp = comp.tppfp, teststat = BT.boot@statistic,
+     proc = mtps, leg = c(0.1, 430), col = 1:4,
```

```
+     lty = 1:4, lwd = 3)
> title("Comparison of TPPFP(q)-controlling AMTPs\n based on SD minP MTP")
```

Figure 15.2 (left) shows that, as expected, the number of rejections increases with the allowed proportion q of false positives when controlling $TPPFP(q)$ at a given level α.

FDR control. Given any TPPFP-controlling MTP, van der Laan et al. (2004a) derive two simple (conservative) FDR-controlling MTPs. Here, we compare these two FDR-controlling approaches, based on a TPPFP-controlling augmentation of the step-down minP procedure, to the marginal Benjamini and Hochberg (Benjamini and Hochberg, 1995) and Benjamini and Yekutieli (Benjamini and Yekutieli, 2001) procedures, implemented in the function `mt.rawp2adjp`. The following code chunk first computes adjusted p-values for the augmentation procedures, then for the marginal procedures, and finally makes a plot of the numbers of rejections vs. the nominal FDR for the four MTPs.

```
> BT.fdr <- fwer2fdr(adjp = BT.boot@adjp, method = "both")$adjp
> BT.marg.fdr <- mt.rawp2adjp(rawp = BT.boot@rawp,
+     proc = c("BY", "BH"))
> comp.fdr <- cbind(BT.fdr, BT.marg.fdr$adjp[
+     order(BT.marg.fdr$index), -1])
> mtps <- c("AMTP Cons", "AMTP Rest", "BY", "BH")
> mt.plot(adjp = comp.fdr, teststat = BT.boot@statistic,
+     proc = mtps, leg = c(0.1, 430), col = c(2,
+         2, 3, 3), lty = rep(1:2, 2), lwd = 3)
> title("Comparison of FDR-controlling MTPs")
```

Figure 15.2 (right) shows that the AMTPs based on conservative bounds for the FDR ("AMTP Cons" and "AMTP Rest") are more conservative than the Benjamini and Hochberg ("BH") MTP for nominal FDR less than 0.4, but less conservative than "BH" for larger FDR. The Benjamini and Yekutieli ("BY") MTP, a conservative version of the Benjamini and Hochberg MTP (with $\sim \log M$ penalty on the p-values), leads to the fewest rejections.

15.4.4 Association of expression measures and tumor molecular subtype: Multi-sample F-statistics

The phenotype data include a variable, `mol.bio`, which records chromosomal abnormalities, such as the BCR/ABL gene rearrangement; these abnormalities concern primarily patients with B-cell ALL and may be related to prognosis. To identify genes with differences in mean expression measures between different tumor molecular subtypes (BCR/ABL, NEG, ALL1/AF4, E2A/PBX1, p15/p16), within B-cell ALL subjects, one can perform a family of F-tests. Tumor subtypes with fewer than 10 subjects are removed from the analysis. Adjusted p-values and test statistic cutoffs

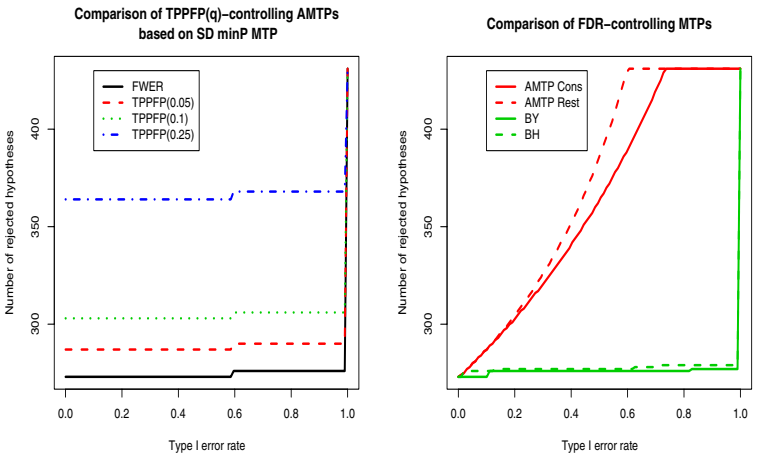

Figure 15.2. *B-cell vs. T-cell ALL – TPPFP and FDR-controlling AMTPs.* Plots of number of rejected hypotheses vs. nominal Type I error rate. *Left*: Comparison of TPPFP-controlling AMTPs, based on the FWER-controlling bootstrap-based step-down minP procedure, for different allowed proportions q of false positives. *Right*: Comparison of four FDR-controlling MTPs.

(for nominal levels α of 0.01 and 0.10) are computed as follows for the FWER-controlling bootstrap-based single-step maxT procedure.

```
> BX <- filtX[, Bcell == 1]
> Bpheno <- pheno[Bcell == 1, ]
> mb <- as.character(Bpheno$mol.biol)
> table(mb)

mb
ALL1/AF4  BCR/ABL E2A/PBX1       NEG  p15/p16
      10       37        5        42        1

> other <- c("E2A/PBX1", "p15/p16")
> mb.boot <- cache("mb.boot", MTP(X = BX[, !(mb %in%
+     other)], Y = mb[!(mb %in% other)], test = "f",
+     alpha = c(0.01, 0.1), B = 100, get.cutoff = TRUE,
+     seed = seed))

running bootstrap...
iteration = 100

> mb.rej <- summary(mb.boot)$rejections

> mb.rej

  Level Rejections
1  0.01        416
2  0.10        418
```

For control of the FWER at nominal level $\alpha = 0.01$, the bootstrap-based single-step maxT procedure with F-statistics identifies 416 genes as having significant differences in mean expression measures between tumor molecular subtypes.

15.4.5 Association of expression measures and time to relapse: Cox t-statistics

The bootstrap-based MTPs implemented in the main `MTP` function (`nulldist="boot"`) allow the test of hypotheses concerning regression parameters in models for which the subset pivotality condition may not hold (e.g., logistic and Cox proportional hazards models). The phenotype information in the ALL package includes the original remission status of the ALL patients (`remission` variable in the *data.frame* `pData(ALL)`). There are 66 B-cell ALL subjects who experienced original complete remission (`remission="CR"`) and who were followed up for remission status at a later date. We apply the single-step maxT procedure to test for a significant association between expression measures and time to relapse amongst these 66 subjects, adjusting for sex. Note that most of the code below is concerned with extracting the (censored) time to relapse outcome and covariates from slots of the *exprSet* instance `ALL`.

```
> cr.ind <- (Bpheno$remission == "CR")
> cr.pheno <- Bpheno[cr.ind, ]
> times <- strptime(cr.pheno$"date last seen", "%m/%d/%Y") -
+     strptime(cr.pheno$date.cr, "%m/%d/%Y")
> time.ind <- !is.na(times)
> times <- times[time.ind]
> cens <- ((1:length(times)) %in% grep("CR", cr.pheno[time.ind,
+     "f.u"]))
> rel.times <- Surv(times, !cens)
> patients <- (1:ncol(BX))[cr.ind][time.ind]
> relX <- BX[, patients]
> relZ <- Bpheno[patients, ]

> cox.boot <- cache("cox.boot", MTP(X = relX, Y = rel.times,
+     Z = relZ, Z.incl = "sex", Z.test = NULL, test = "coxph.YvsXZ",
+     B = 100, get.cr = TRUE, seed = seed))
```

For control of the FWER at nominal level $\alpha = 0.05$, the bootstrap-based single-step maxT procedure identifies 22 genes whose expression measures are significantly associated with time to relapse. Using the function `mget`, we examine the names of these genes.

```
> cox.diff <- cox.boot@adjp <= 0.05
> sum(cox.diff)

[1] 22
```

```
> cox.AffyID <- geneNames(filtALL)[cox.diff]
> mget(cox.AffyID, env = hgu95av2GENENAME)

$"106_at"
[1] "runt-related transcription factor 3"

$"1403_s_at"
[1] "chemokine (C-C motif) ligand 5"

$"182_at"
[1] "inositol 1,4,5-triphosphate receptor, type 3"

$"286_at"
[1] "histone 2, H2aa"

$"296_at"
[1] "tubulin, beta 2"

$"33232_at"
[1] "cysteine-rich protein 1 (intestinal)"

$"34308_at"
[1] "histone 1, H2ac"

$"35127_at"
[1] "histone 1, H2ae"

$"36638_at"
[1] "connective tissue growth factor"

$"37027_at"
[1] "AHNAK nucleoprotein (desmoyokin)"

$"37218_at"
[1] "BTG family, member 3"

$"37343_at"
[1] "inositol 1,4,5-triphosphate receptor, type 3"

$"38124_at"
[1] "midkine (neurite growth-promoting factor 2)"

$"39182_at"
[1] "epithelial membrane protein 3"

$"39317_at"
[1] "cytidine monophosphate-N-acetylneuraminic acid
   hydroxylase (CMP-N-acetylneuraminate monooxygenase)"
```

```
$"39331_at"
[1] "tubulin, beta 2"

$"39338_at"
[1] "S100 calcium binding protein A10 (annexin II ligand,
   calpactin I, light polypeptide (p11))"

$"40147_at"
[1] "vesicle amine transport protein 1 homolog (T californica)"

$"40567_at"
[1] "tubulin, alpha 3"

$"40729_s_at"
[1] "allograft inflammatory factor 1"

$"41071_at"
[1] "serine protease inhibitor, Kazal type 2 (acrosin-
   trypsin inhibitor)"

$"41164_at"
[1] "immunoglobulin heavy constant mu"
```

Figure 15.3 is a plot of the Cox regression coefficient estimates (circles) and corresponding confidence regions (text indicating the level) for the five genes with the smallest adjusted p-values. The plot illustrates that the level $\alpha = 0.05$ confidence regions corresponding to the significant gene does not include the null value $\psi_0 = 0$ for the Cox regression parameters (red line). The confidence regions for the next four genes, do include 0.

```
> plot(cox.boot, which = 5, top = 5, sub.caption = NULL)
> abline(h = 0, col = "red")
```

15.5 Discussion

The multtest package implements resampling-based multiple testing procedures that can be applied to a broad range of testing problems in biomedical and genomic data analysis. Ongoing efforts involve expanding the class of MTPs implemented in multtest, enhancing software design and the user interface, and increasing computational efficiency. Specifically, regarding the offering of MTPs, we envisage the following new developments.

- Expanding the class of available tests, by adding test statistic closures for tests of correlations, quantiles, and parameters in generalized linear models (e.g., logistic regression).

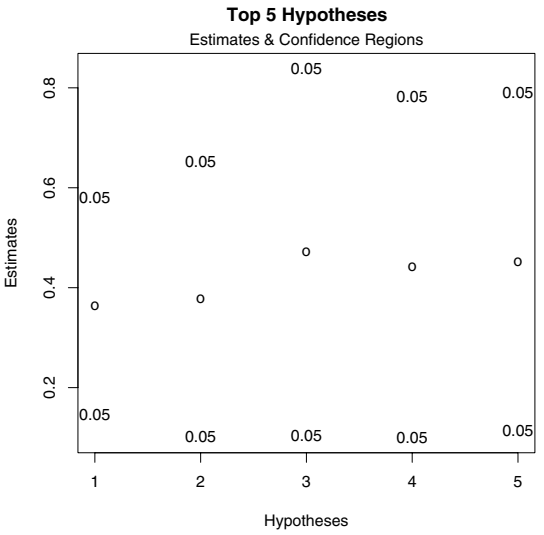

Figure 15.3. Time to relapse – FWER-controlling single-step maxT MTP. Plot of Cox regression coefficient estimates and corresponding confidence intervals for the fifteen genes with the smallest adjusted p-values, based on the FWER-controlling bootstrap-based single-step maxT procedure (`plot` method, `which=5`).

- Expanding the class of resampling-based estimators for the test statistics null distribution (e.g., parametric bootstrap, Bayesian bootstrap), possibly using a function closure approach.
- Providing parameter confidence regions and test statistic cutoffs for other Type I error rates than the FWER.
- Implementing the new augmentation multiple testing procedures proposed in Dudoit et al. (2004a) and Dudoit and van der Laan (2004), for controlling tail probabilities $Pr(g(V_n, R_n) > q)$ for an arbitrary function $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n.

Efforts regarding software design and the user interface include the following.

- Providing a formula interface for a symbolic description of the tests to be performed (cf. model specification in `lm`).
- Providing an `update` method for objects of class *MTP*, to facilitate the reuse of available estimates of the null distribution when implementing new MTPs.
- Extending the *MTP* class to keep track of results for several MTPs.