Applications of Multiple Testing Procedures

Katherine S. Pollard[1], Sandrine Dudoit[2], Mark J. van der Laan[3]

October 6, 2004

1. Center for Biomolecular Science and Engineering, University of California, Santa Cruz, `http://lowelab.ucsc.edu/katie/`
2. Division of Biostatistics, University of California, Berkeley, `http://www.stat.berkeley.edu/~sandrine/`
3. Department of Statistics and Division of Biostatistics, University of California, Berkeley, `http://www.stat.berkeley.edu/~laan/`

Contents

1 Overview

The Bioconductor R package *multtest* implements widely applicable resampling-based single-step and stepwise multiple testing procedures (MTP) for con-

trolling a broad class of Type I error rates, in testing problems involving general data generating distributions (with arbitrary dependence structures among variables), null hypotheses, and test statistics Dudoit and van der Laan (2004); Dudoit et al. (2004); van der Laan et al. (2004b,a); Pollard and van der Laan (2004). A key feature of these MTPs is the test statistics null distribution (rather than data generating null distribution) used to derive rejection regions (i.e., cut-offs) for the test statistics and the resulting adjusted p-values. For general null hypotheses, defined in terms of submodels for the data generating distribution, this null distribution is the asymptotic distribution of the vector of null value shifted and scaled test statistics. The current version of *multtest* provides MTPs for null hypotheses concerning means, differences in means, and regression parameters in linear,and Cox proportional hazards models. Both non-parametric bootstrap and permutation estimators of the test statistics (t- or F-statistics) null distribution are available. Procedures are provided to control Type I error rates defined as tail probabilities and expected values of arbitrary functions of the numbers of Type I errors, V_n, and rejected hypotheses, R_n. These error rates include: the generalized family-wise error rate, $gFWER(k) = Pr(V_n > k)$, or chance of at least $(k + 1)$ false positives (the special case $k = 0$ corresponds to the usual family-wise error rate, FWER); tail probabilities $TPPFP(q) = Pr(V_n/R_n > q)$ for the proportion of false positives among the rejected hypotheses; the false discovery rate, $FDR = E[V_n/R_n]$. Single-step and step-down common-cut-off (maxT) and common-quantile (minP) procedures, that take into account the joint distribution of the test statistics, are implemented to control the FWER. In addition, augmentation procedures are provided to control the gFWER and TPPFP, based on *any* initial FWER-controlling procedure. The results of a multiple testing procedure are summarized using rejection regions for the test statistics, confidence regions for the parameters of interest, and adjusted p-values.

The modular design of the *multtest* package allows interested users to readily extend the package functionality by inserting additional functions for test statistics and testing procedures. A class/method object-oriented programming approach was adopted to summarize the results of a MTP.

The multiple testing procedures are applied to the Acute Lymphoblastic Leukemia (ALL) dataset of Chiaretti et al. Chiaretti et al. (2004), available in the R package *ALL*, to identify genes whose expression measures are associated with (possibly censored) biological and clinical outcomes such as: cytogenetic test status (normal vs. abnormal), tumor molecular subtype (BCR/ABL, NEG, ALL1/AF4, E2A/PBX1, p15/p16, NUP-98), and patient survival.

2 Getting started

Installing the package. To install the *multtest* package, first download the appropriate file for your platform from the Bioconductor website `http://www.bioconductor.org/`. For Windows, start R and select the `Packages` menu, then `Install package from local zip file...`. Find and highlight the location of the zip file and click on `open`. For Linux/Unix, use the usual command `R CMD INSTALL` or set the option `CRAN` to your nearest mirror site and use the command `install.packages` from within an R session.

Loading the package. To load the *multtest* package in your R session, type `library(multtest)`.

Help files. Detailed information on *multtest* package functions can be obtained in the help files. For example, to view the help file for the function `MTP` in a browser, use `help.start` followed by `? MTP`.

Case study. We illustrate some of the functionality of the *multtest* package using the Acute Lymphoblastic Leukemia (ALL) microarray dataset of Chiaretti et al. Chiaretti et al. (2004). Available in the data package *ALL*, this dataset includes 21 phenotypes and 12,625 Affymetrix gene expression measures (chip series hgu95av2), for each of 128 ALL patients. The expression measures have been jointly normalized using RMA. To view a description of the experiments and data, type `? ALL`.

Sweave. This document was generated using the `Sweave` function from the R *tools* package. The source (.Rnw) file is in the `/inst/doc` directory of the *multtest* package.

3 Software Application: ALL microarray dataset

3.1

The main user-level function for resampling-based multiple testing is `MTP`. Its input/output and usage are described in the accompanying vignette (MTP). Here, we illustrate some of the functionality of the *multtest* package using the Acute Lymphoblastic Leukemia (ALL) microarray dataset of Chiaretti et al. Chiaretti et al. (2004), available in the data package *ALL*. We begin by loading the necessary packages.

```
> library(Biobase)
> library(multtest)
> library(genefilter)
> library(reposTools)
```

We use the *reposTools* package to get the necessary data pacakges.

```
> z <- try(getReposEntry("http://www.bioconductor.org/data/experimental/repos"))
> try(install.packages2("ALL", repEntry = z))

> library(ALL)
> try(install.packages2("hgu95av2"))

> library(hgu95av2)
```

3.2 *ALL* data package and initial gene filtering

The Acute Lymphoblastic Leukemia (ALL) microarray dataset of Chiaretti et al. Chiaretti et al. (2004) consists of 21 *phenotypes* (i.e., patient level responses and covariates) and 12,625 Affymetrix *gene expression measures* (chip series HGU95Av2), for each of 128 ALL patients. For greater detail, please consult the *ALL* package documentation. The main object in this package is `ALL`, an instance of the class *exprSet*, which contains the expression measures, phenotypes, and gene annotation information. The genes-by-subjects matrix of expression measures is provided in the `exprs` slot of `ALL` and the phenotype data are stored in the `phenoData` slot. Note that the expression measures have been obtained using the three-step robust multichip average (RMA) pre-processing method, implemented in the package *affy*. In particular, the expression measures have been subject to a base 2 logarithmic transformation.

```
> data(ALL)
> class(ALL)

[1] "exprSet"
attr(,"package")
[1] "Biobase"

> slotNames(ALL)

[1] "exprs"       "se.exprs"    "phenoData"   "description" "annotation"
[6] "notes"
```

```
> show(ALL)

Expression Set (exprSet) with
        12625 genes
        128 samples
                 phenoData object with 21 variables and 128 cases
         varLabels
                cod:  Patient ID
                diagnosis:  Date of diagnosis
                sex:  Gender of the patient
                age:  Age of the patient at entry
                BT:  does the patient have B-cell or T-cell ALL
                remission:  Complete remission(CR), refractory(REF) or NA. Derived fro
                CR:  Original remisson data
                date.cr:  Date complete remission if achieved
                t(4;11):  did the patient have t(4;11) translocation. Derived from cit
                t(9;22):  did the patient have t(9;22) translocation. Derived from cit
                cyto.normal:  Was cytogenetic test normal? Derived from citog
                citog:  original citogenetics data, deletions or t(4;11), t(9;22) statu
                mol.biol:  molecular biology
                fusion protein:  which of p190, p210 or p190/210 for bcr/able
                mdr:  multi-drug resistant
                kinet:  ploidy: either diploid or hyperd.
                ccr:  Continuous complete remission? Derived from f.u
                relapse:  Relapse? Derived from f.u
                transplant:  did the patient receive a bone marrow transplant? Derived
                f.u:  follow up data available
                date last seen:  date patient was last seen

> names(varLabels(ALL))

 [1] "cod"            "diagnosis"      "sex"            "age"
 [5] "BT"             "remission"      "CR"             "date.cr"
 [9] "t(4;11)"        "t(9;22)"        "cyto.normal"    "citog"
[13] "mol.biol"       "fusion protein" "mdr"            "kinet"
[17] "ccr"            "relapse"        "transplant"     "f.u"
[21] "date last seen"

> X <- exprs(ALL)
> pheno <- pData(ALL)
```

Our goal is to identify genes whose expression measures are associated with (possibly censored) biological and clinical outcomes such as: cytogenetic test status (normal vs. abnormal), tumor molecular subtype (BCR/ABL, NEG, ALL1/AF4, E2A/PBX1, p15/p16, NUP-98), and time to relapse. However, before applying the multiple testing procedures discussed in Section **??**, above, we perform initial gene filtering as in Chiaretti et al. 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^X, ffun)
> filtX <- X[filt, ]
> dim(filtX)
```

```
[1] 431 128
```

```
> filtALL <- ALL[filt, ]
```

3.3 Association of expression measures and cytogenetic test status: two-sample t-statistics

Step-down minP FWER-controlling MTP with two-sample Welch t-statistics and bootstrap null distribution The phenotype data include an indicator variable, `cyto.normal`, for cytogenetic test status (1 for normal vs. 0 for abnormal). To identify genes with higher mean expression measures in the abnormal compared to the normal cytogenetics subjects, one-sided two-sample t-tests can be performed. We choose to use the Welch t-statistic and to control the FWER using the bootstrap-based step-down minP procedure with $B = 10000$ bootstrap iterations.

```
> seed <- 99
> cyto.boot <- MTP(X = filtALL, Y = "cyto.normal", alternative = "less",
+     B = 10000, method = "sd.minP", seed = seed)
```

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

```
> class(cyto.boot)
```

```
[1] "MTP"
attr(,"package")
[1] "multtest"

> slotNames(cyto.boot)

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

> print(cyto.boot)

        Multiple Testing Procedure

Object of class:  MTP
sample size = 128
number of hypotheses = 431

test statistics = t.twosamp.unequalvar
type I error rate = fwer
nominal level alpha = 0.05
multiple testing procedure = sd.minP

Call: MTP(X = filtALL, Y = "cyto.normal", alternative = "less", B = 10000,
    method = "sd.minP", seed = seed)

Slots:
            Class    Mode Length Dimension
statistic numeric numeric    431
estimate  numeric numeric    431
sampsize  numeric numeric      1
rawp      numeric numeric    431
adjp      numeric numeric    431
conf.reg    array logical      0     0,0,0
cutoff     matrix logical      0       0,0
reject     matrix logical    431     431,1
nulldist   matrix logical      0       0,0
call         call    call      7
seed      integer numeric      1

> summary(cyto.boot)
```

```
MTP:  sd.minP
Type I error rate:  fwer

  Level Rejections
1  0.05         12

            Min. 1st Qu.    Median     Mean 3rd Qu.   Max.
adjp       0.000  0.9990  1.000000  0.90540  1.0000 1.0000
rawp       0.000  0.0791  0.460800  0.48180  0.8741 1.0000
statistic -2.922 -0.7589 -0.038150 -0.11920  0.5946 2.5620
estimate  -1.083 -0.2015 -0.008719 -0.03249  0.1467 0.6709
```

The following commands may be used to obtain a list of genes that are differentially expressed in normal vs. abnormal cytogenetics 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.

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

[1] 12

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

$"1433_g_at"
[1] " MAD, mothers against decapentaplegic homolog 3 (Drosophila)"

$"32542_at"
[1] " four and a half LIM domains 1"

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

$"33284_at"
[1] " myeloperoxidase"

$"33891_at"
```

```
[1] " chloride intracellular channel 4"

$"37539_at"
[1] " RalGDS-like gene"

$"38487_at"
[1] " stabilin 1"

$"38944_at"
[1] " MAD, mothers against decapentaplegic homolog 3 (Drosophila)"

$"39712_at"
[1] " S100 calcium binding protein A13"

$"40607_at"
[1] " dihydropyrimidinase-like 2"

$"40888_f_at"
[1] " eukaryotic translation elongation factor 1 alpha 1"

$"41470_at"
[1] " prominin 1"
```

Various graphical summaries of the results may be obtained using the `plot` method, by selecting appropriate values of the argument `which` (Figure 1).

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

Marginal FWER-controlling MTPs with two-sample Welch t-statistics and bootstrap null distribution Given a vector of unadjusted p-values, the `mt.rawp2adjp` function computes adjusted p-values for the marginal FWER-controlling MTPs of Bonferroni, Holm Holm (1979), Hochberg Hochberg (1988), and Šidák Šidák (1967), discussed in detail in Dudoit et al. Dudoit et al. (2003). The `mt.plot` function may then be used to compare the different procedures in terms of their adjusted p-values.

```
> marg <- c("Bonferroni", "Holm", "Hochberg", "SidakSS", "SidakSD")
> cyto.marg <- mt.rawp2adjp(rawp = cyto.boot@rawp, proc = marg)
```

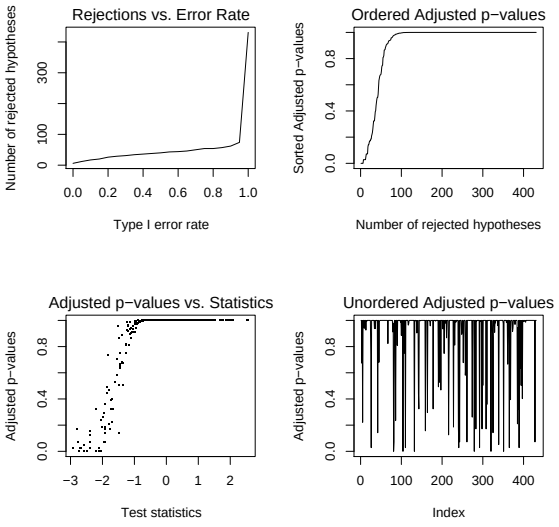

Figure 1: *Cytogenetic test status — Step-down minP FWER-controlling MTP.* By default, four graphical summaries are produced by the `plot` method for instances of the class *MTP*.

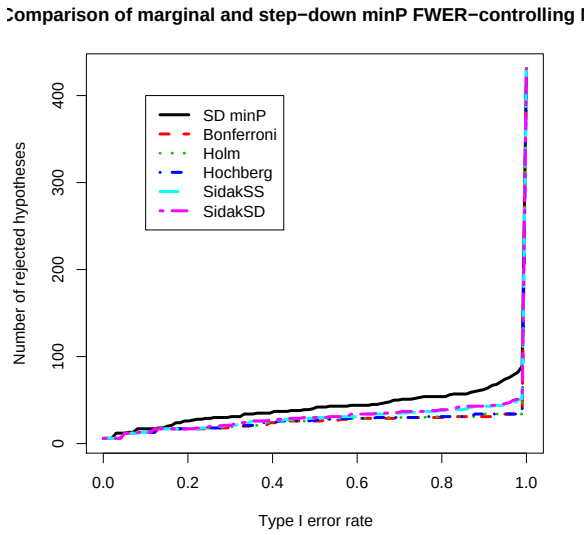

Figure 2: *Cytogenetic test status — Marginal vs. joint FWER-controlling MTPs.* Plot of number of rejected hypotheses vs. nominal Type I error rate for comparing bootstrap-based marginal and step-down minP FWER-controlling MTPs.

```
> comp.marg <- cbind(cyto.boot@adjp, cyto.marg$adjp[order(cyto.marg$index),
+     -1])

> par(mfrow = c(1, 1))
> mt.plot(adjp = comp.marg, teststat = cyto.boot@statistic, proc = c("SD minP",
+     marg), leg = c(0.1, 400), col = 1:6, lty = 1:6, lwd = 3)
> title("Comparison of marginal and step-down minP FWER-controlling MTPs")
```

In this dataset, most of the FWER-controlling MTPs perform similarly, making very few rejections at nominal Type I error rates near zero. As expected, the bootstrap-based step-down minP procedure, which takes into account the joint distribution of the test statistics, leads to slightly more rejections than the marginal methods (Figure 2). The results also illustrate that stepwise MTPs are less conservative than their single-step analogues (e.g., Holm and Hochberg vs. Bonferroni; step-down Šidák vs. single-step Šidák).

Step-down minP FWER-controlling MTP with two-sample Welch t-statistics and permutation null distribution Because the sample sizes are not equal for the two cytogenetic groups and the expression measures may have different covariance structures in the two populations, we expect the bootstrap and permutation null distributions to yield different sets of rejected hypotheses (Pollard & van der Laan Pollard and van der Laan (2004)). To compare the two approaches, we apply the permutation-based step-down minP procedure, first using the old `mt.minP` function and then using the new `MTP` function (which calls `mt.minP`). Please note that while the `MTP` and `mt.minP` functions produce the same results, these are presented in a different manner. In particular, for the new function `MTP`, the results (e.g., test statistics, parameter estimates, unadjusted p-values, adjusted p-values, cut-offs) are given in the original order of the null hypotheses, while in the `mt.minP` function, the hypotheses are sorted first according to their adjusted p-values, next their unadjusted p-values, and finally their test statistics. In addition, the new function `MTP` implements a broader range of MTPs and has adopted the S4 class/method design for representing and summarizing the results of a MTP.

```
> set.seed(99)
> NAs <- is.na(pheno$cyto.normal)
> cyto.perm.old <- mt.minP(X = filtX[, !NAs], classlabel = pheno$cyto.normal[!NAs],
+     side = "lower", B = 10000)

> names(cyto.perm.old)

[1] "index"    "teststat" "rawp"     "adjp"     "plower"

> sum(cyto.perm.old$adjp <= 0.05)

[1] 431

> set.seed(99)
> cyto.perm.new <- MTP(X = filtX, Y = pheno$cyto.normal, alternative = "less",
+     nulldist = "perm", B = 10000, method = "sd.minP")

> summary(cyto.perm.new)

MTP:  sd.minP
Type I error rate:  fwer
```

```
  Level Rejections
1  0.05        431

           Min.    1st Qu.     Median       Mean    3rd Qu.     Max.
adjp        0.0  0.000e+00 5.264e-315  0.0037750  7.813e-03 0.007813
rawp        0.0  0.000e+00 5.263e-315  0.0002957  2.473e-05 0.007031
statistic -15.6 -5.976e-15  0.000e+00 -0.1615000 2.658e-315 8.712000
estimate     NA         NA         NA        NaN         NA       NA
```

```
> sum(cyto.perm.new@adjp <= 0.05)
```

```
[1] 431
```

```
> sum(cyto.perm.new@adjp <= 0.05 & cyto.boot@adjp <= 0.05)
```

```
[1] 12
```

At nominal FWER level $\alpha = 0.05$, the permutation step-down minP procedure identifies 431 genes as differentially expressed between patients with normal and abnormal cytogenetic test status. In contrast, the bootstrap version of the step-down minP procedure identifies 12 differentially expressed genes.

Step-down minP FWER-controlling MTP with robust two-sample t-statistics and bootstrap null distribution The Wilcoxon rank sum statistic (also known as the Mann-Whitney statistic) is a robust alternative to the usual two-sample t-statistic. Note that only $B = 1000$ bootstrap iterations are used in this example, although more are recommended in practice.

```
> cyto.wilcox <- MTP(X = filtALL, Y = "cyto.normal", robust = TRUE,
+     alternative = "less", B = 1000, method = "sd.minP", seed = seed)
```

```
running bootstrap...
iteration = 100 200 300 400 500 600 700 800 900 1000
```

```
> sum(cyto.wilcox@adjp <= 0.05)
```

```
[1] 16
```

```
> sum(cyto.wilcox@adjp <= 0.05 & cyto.boot@adjp <= 0.05)
```

```
[1] 7
```

At nominal FWER level $\alpha = 0.05$, the bootstrap step-down minP MTP based on the robust Wilcoxon test statistic identifies 16 genes as differentially expressed, compared to 12 genes for the same MTP based on the Welch t-statistic. 7 genes are identified by both procedures.

3.4 Augmentation procedures for gFWER, TPPFP, and FDR control

In the context of microarray gene expression data analysis or other high-dimensional inference problems, one is often willing to accept 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 **??**, 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 of the proportion of false positives (TPPFP), or the false discovery rate (FDR). We illustrate the use of the `fwer2gfwer`, `fwer2tppfp`, and `fwer2fdr` functions, but note that the gFWER, TPPFP, and FDR can also be controlled directly using the `MTP` function with appropriate choices of arguments `typeone`, `k`, `q`, and `fdr.method`.

gFWER control

```
> k <- c(5, 10, 50, 100)
> cyto.gfwer <- fwer2gfwer(adjp = cyto.boot@adjp, k = k)
> comp.gfwer <- cbind(cyto.boot@adjp, cyto.gfwer)
> mtps <- paste("gFWER(", c(0, k), ")", sep = "")
> mt.plot(adjp = comp.gfwer, teststat = cyto.boot@statistic, proc = mtps,
+     leg = c(0.1, 400), col = 1:5, lty = 1:5, lwd = 3)
> title("Comparison of gFWER(k)-controlling AMTPs based on SD minP MTP")
```

For gFWER-controlling AMTPs, Figure 3 illustrates that the number of rejected hypotheses increases linearly with the number k of allowed false positives, for nominal levels α such that the initial FWER-controlling MTP does not reject more than $M - k$ hypotheses. That is, the curve for the $gFWER(k)$–controlling AMTP is obtained from that of the initial FWER-controlling procedure by a simple vertical shift of k.

TPPFP control

```
> q <- c(0.05, 0.1, 0.5)
> cyto.tppfp <- fwer2tppfp(adjp = cyto.boot@adjp, q = q)
> comp.tppfp <- cbind(cyto.boot@adjp, cyto.tppfp)
> mtps <- c("FWER", paste("TPPFP(", q, ")", sep = ""))
> mt.plot(adjp = comp.tppfp, teststat = cyto.boot@statistic, proc = mtps,
+     leg = c(0.1, 400), col = 1:4, lty = 1:4, lwd = 3)
> title("Comparison of TPPFP(q)-controlling AMTPs based on SD minP MTP")
```

For TPPFP control, Figure 4 shows that, as expected, the number of rejections, while controlling $TPPFP(q)$ at a given level α, increases with the allowed proportion q of false positives, though not linearly. Furthermore, for the ALL dataset, the increases in the number of rejections are not very large.

FDR control Given any TPPFP-controlling MTP, van der Laan et al. 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 & Hochberg Benjamini and Hochberg (1995) and Benjamini & Yekutieli Benjamini and Yekutieli (2001) procedures, implemented in the function `mt.rawp2adjp`.

```
> cyto.fdr <- fwer2fdr(adjp = cyto.boot@adjp, method = "both")$adjp
> cyto.marg.fdr <- mt.rawp2adjp(rawp = cyto.boot@rawp, proc = c("BY",
+     "BH"))
> comp.fdr <- cbind(cyto.fdr, cyto.marg.fdr$adjp[order(cyto.marg.fdr$index),
+     -1])
> mtps <- c("AMTP Cons", "AMTP Rest", "BY", "BH")
> mt.plot(adjp = comp.fdr, teststat = cyto.boot@statistic, proc = mtps,
+     leg = c(0.1, 400), col = c(2, 2, 3, 3), lty = rep(1:2, 2),
+     lwd = 3)
> title("Comparison of FDR-controlling MTPs")
```

Figure 5 shows that for most values of the nominal FDR level α, the usual Benjamini & Hochberg ("BH") MTP leads by far to the largest number of rejected hypotheses. The Benjamini & Yekutieli ("BY") MTP, a conservative version of the Benjamini & Hochberg MTP (with $\sim \log M$ penalty on the p-values), leads to much fewer rejections. The AMTPs based on conservative bounds for the FDR ("AMTP Cons" and "AMTP Rest") are much more

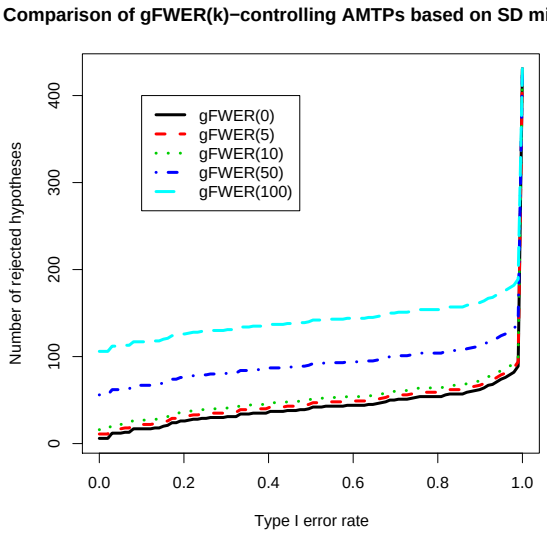

Figure 3: *Cytogenetic test status — gFWER-controlling AMTPs.* Plot of number of rejected hypotheses vs. nominal Type I error rate for comparing gFWER-controlling AMTPs, based on the bootstrap step-down minP FWER-controlling procedure, with different allowed numbers k of false positives.

conservative than the Benjamini & Hochberg MTP and only lead to an increased number of rejections for very high nominal FDR levels.

3.5 Association of expression measures and tumor molecular subtype: multi-sample F-statistics

To identify genes with differences in mean expression measures between different tumor molecular subtypes (BCR/ABL, NEG, ALL1/AF4, E2A/PBX1, p15/p16, NUP-98), one can perform a family of F-tests. Tumor subtypes with fewer than 10 subjects are merged into one group. Adjusted p-values and test statistic cut-offs (for nominal levels α of 0.01 and 0.1) are computed as follows for the bootstrap-based single-step maxT FWER-controlling procedure.

```
> mb <- as.character(pheno$mol.biol)
> table(mb)
```

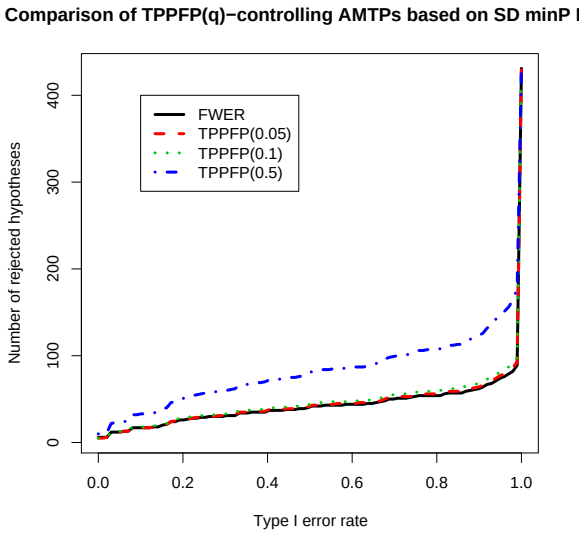

Figure 4: *Cytogenetic test status — TPPFP-controlling AMTPs.* Plot of number of rejected hypotheses vs. nominal Type I error rate for comparing TPPFP-controlling AMTPs, based on the bootstrap step-down minP FWER-controlling procedure, with different allowed proportions q of false positives.

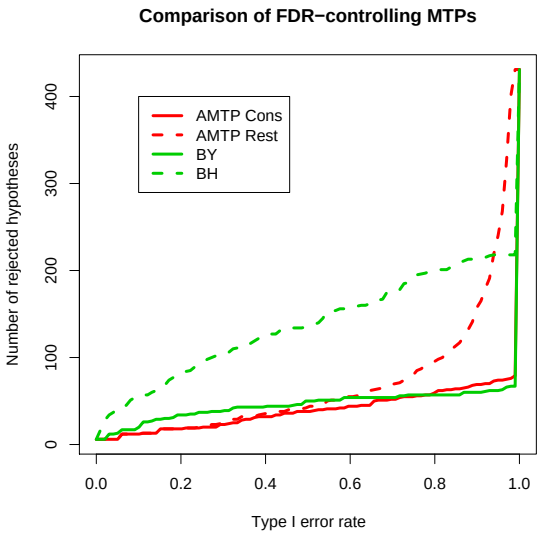

Figure 5: *Cytogenetic test status — FDR-controlling MTPs.* Plot of number of rejected hypotheses vs. nominal Type I error rate for comparing four FDR-controlling MTPs.

```
mb
ALL1/AF4  BCR/ABL E2A/PBX1      NEG   NUP-98  p15/p16
      10       37        5       74        1        1

> other <- c("E2A/PBX1", "NUP-98", "p15/p16")
> mb[mb %in% other] <- "other"
> table(mb)

mb
ALL1/AF4  BCR/ABL      NEG    other
      10       37       74        7

> mb.boot <- MTP(X = filtX, Y = mb, test = "f", alpha = c(0.01,
+     0.1), B = 10000, get.cutoff = TRUE, seed = seed)
```

Let us examine the results of the MTP.

```
> summary(mb.boot)

MTP:  ss.maxT
Type I error rate:  fwer

  Level Rejections
1  0.01        427
2  0.10        429

             Min. 1st Qu. Median     Mean 3rd Qu.    Max.
adjp      0.00000   0.000  0.000 0.001214   0.000  0.3702
rawp      0.00000   0.000  0.000 0.000000   0.000  0.0000
statistic 0.06911   1.816  4.375 5.922000   8.283 36.1300
estimate       NA      NA     NA      NaN      NA      NA

> mb.diff <- mb.boot@adjp <= 0.01
> sum(mb.diff)

[1] 427

> sum(mb.boot@statistic >= mb.boot@cutoff[, "alpha=0.01"] & mb.diff)

[1] 427
```

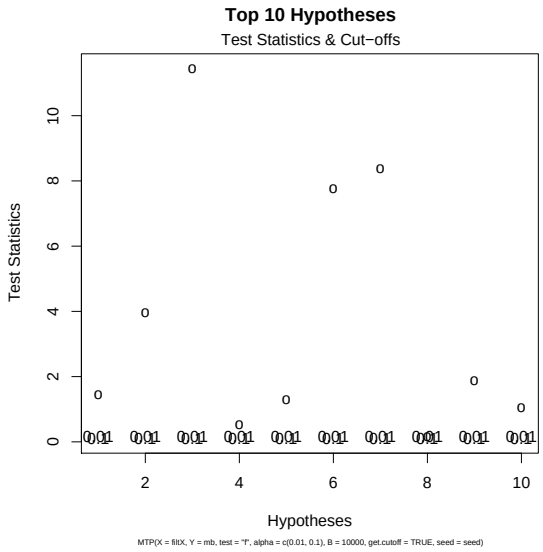

Figure 6: *Tumor molecular subtype — Single-step maxT FWER-controlling MTP.* Plot of F-statistics and corresponding cut-offs for the 10 genes with the smallest adjusted p-values, based on the bootstrap single-step maxT FWER-controlling procedure (`plot` method, `which=6`).

For control of the FWER at nominal level $\alpha = 0.01$, the bootstrap-based single-step maxT procedure with F-statistics identifies 427 genes (i.e., almost all the 431 filtered genes) as having significant differences in mean expression measures between tumor molecular subtypes. This set can be identified through either adjusted p-values or cut-offs for the test statistics. The plot of test statistics and corresponding cut-offs in Figure 6 illustrates that the F-statistics for the 10 genes with the smallest adjusted p-values are much larger than expected by chance under the null distribution.

```
> plot(mb.boot, which = 6)
```

3.6 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 88 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 88 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`. Only $B = 5000$ bootstrap iterations are used in this example.

```
> cr.ind <- pheno$remission == "CR"
> cr.pheno <- pheno[cr.ind, ]
> times <- strptime(cr.pheno$"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(filtX))[cr.ind][time.ind]
> relX <- filtX[, patients]
> relZ <- pheno[patients, ]

> cox.boot <- MTP(X = relX, Y = rel.times, Z = relZ, Z.incl = "sex",
+     Z.test = NULL, test = "coxph.YvsXZ", B = 5000, get.cr = TRUE,
+     seed = seed)

> summary(cox.boot)

MTP:  ss.maxT
Type I error rate:  fwer

  Level Rejections
1  0.05          7

             Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
adjp       0.0002  0.9758 1.00000 0.90890  1.0000 1.0000
rawp       0.0000  0.0414 0.11200 0.14370  0.2253 0.4312
```

```
statistic -2.2780 -0.2521 0.37040 0.41410  1.1010 4.0490
estimate  -0.5374 -0.0279 0.04581 0.04583  0.1231 0.3972
```

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

```
[1] 7
```

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

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

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

$"33412_at"
[1] " lectin, galactoside-binding, soluble, 1 (galectin 1)"

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

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

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

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

```
> plot(cox.boot, which = 5)
> abline(h = 0, col = 2, lwd = 2)
```

For control of the FWER at nominal level $\alpha = 0.05$, the bootstrap-based single-step maxT procedure identifies 7 genes whose expression measures are significantly associated with time to relapse. Equivalently, Figure 7 illustrates that the level $\alpha = 0.05$ confidence regions corresponding to these 7 genes do not include the null value $\psi_0 = 0$ for the Cox regression parameters (indicated by red horizontal line). The confidence intervals for the next four genes barely cover $\psi_0 = 0$.

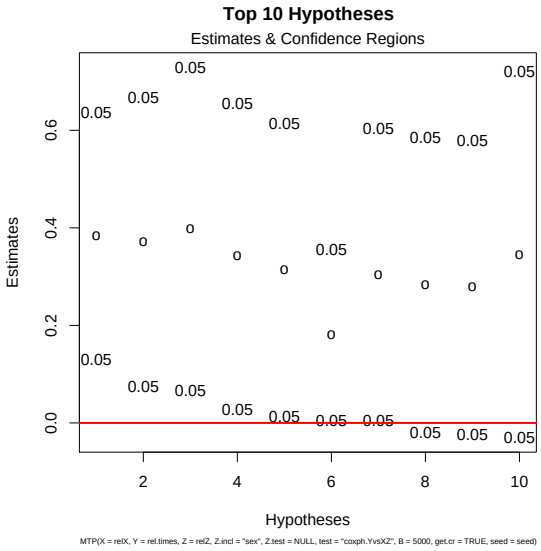

Figure 7: *Time to relapse — Single-step maxT FWER-controlling MTP.* Plot of Cox regression coefficient estimates and corresponding confidence intervals for the 10 genes with the smallest adjusted p-values, based on the bootstrap single-step maxT FWER-controlling procedure (`plot` method, `which=5`).

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSSB*, 57:289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple hypothesis testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz, and R. Foa. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–2778, 2004.

S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures and Applications to Genomics*. Springer, 2004. (In preparation).

S. Dudoit, M. J. van der Laan, and K. S. Pollard. Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13, 2004. URL `www.bepress.com/sagmb/vol3/iss1/art13`.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.

S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6:65–70, 1979.

K. S. Pollard and M. J. van der Laan. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, 125(1–2):85–100, 2004.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15, 2004a. URL `www.bepress.com/sagmb/vol3/iss1/art15`.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical*

Applications in Genetics and Molecular Biology, 3(1):Article 14, 2004b. URL `www.bepress.com/sagmb/vol3/iss1/art14`.

Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62: 626–633, 1967.