Multiple Testing Procedures with Applications to Genomics

Sandrine Dudoit

Division of Biostatistics and Department of Statistics

University of California, Berkeley

www.stat.berkeley.edu/~sandrine

MCP 2007: The 5th International Conference on Multiple Comparison Procedures

Vienna, Austria

July 09–11, 2007

Acknowledgements

- Mark J. van der Laan, Division of Biostatistics and Department of Statistics, UC Berkeley.
- Houston N. Gilbert, Graduate Group in Biostatistics, UC Berkeley.
- Alan E. Hubbard, Division of Biostatistics, UC Berkeley.
- Sündüz Keleş, Departments of Biostatistics and Statistics, University of Wisconsin, Madison.
- Katherine S. Pollard, Genome Center and Department of Statistics, UC Davis.

References

This presentation is based on the forthcoming book

S. Dudoit and M. J. van der Laan (2007). *Multiple Testing Procedures with Applications to Genomics*, Springer Series in Statistics.

Related articles, lecture notes, and software may be downloaded from Sandrine Dudoit's website

www.stat.berkeley.edu/~sandrine

and Mark van der Laan's website

www.stat.berkeley.edu/~laan.

Outline

Sandrine Dudoit.

- Motivation: Multiple Hypothesis Testing Problems in Genomics.
- Overview of Main Contributions: Multiple Hypothesis Testing Framework and Procedures.
- Application: Multiple Tests of Association with Biological Annotation Metadata.
- Software Implementation: Bioconductor R Package multtest.

Mark van der Laan.

- Test Statistics Null Distribution.
- Resampling-Based Empirical Bayes Multiple Testing Procedures.

Houston Gilbert. FDR-Controlling Resampling-Based Empirical Bayes Multiple Testing Procedures: Simulation Study and Application to Microarray-Based Genetic Mapping of Gene Expression in *S. cerevisiae*.

Multiple Hypothesis Testing Problems in Genomics

- High-throughput microarray gene expression analysis.
  - Identification of differentially expressed (DE) genes. Testing for associations between gene expression measures and possibly censored biological and clinical covariates and outcomes.
  - Identification of co-expressed (CE) genes. Testing for associations in the expression measures of sets of genes across biological conditions.
- Biological annotation metadata analysis. Testing for associations between gene expression measures and biological annotation metadata.
  E.g. Gene Ontology (GO) terms; PubMed abstracts.

Multiple Hypothesis Testing Problems in Genomics

- Protein sequence analysis. Testing for associations between phenotypes and codon/amino acid mutations.

  E.g. Association between viral replication capacity and HIV-1 sequence variation.

- Genetic mapping of complex traits. Testing for associations between (sets of) phenotypes and genotypes.

  E.g. Phenotypes: Affectedness status; transcript (i.e., mRNA) levels.
  Genotypes: Single nucleotide polymorphisms (SNP); SNP haplotypes; microsatellite marker genotypes.

- Mass-spectroscopy. Testing for associations between phenotypes and protein mass-spectroscopy measures.

  E.g. Association between leukemia class (ALL vs. AML) and mass-to-charge ratios.

Multiple Hypothesis Testing Problems in Genomics

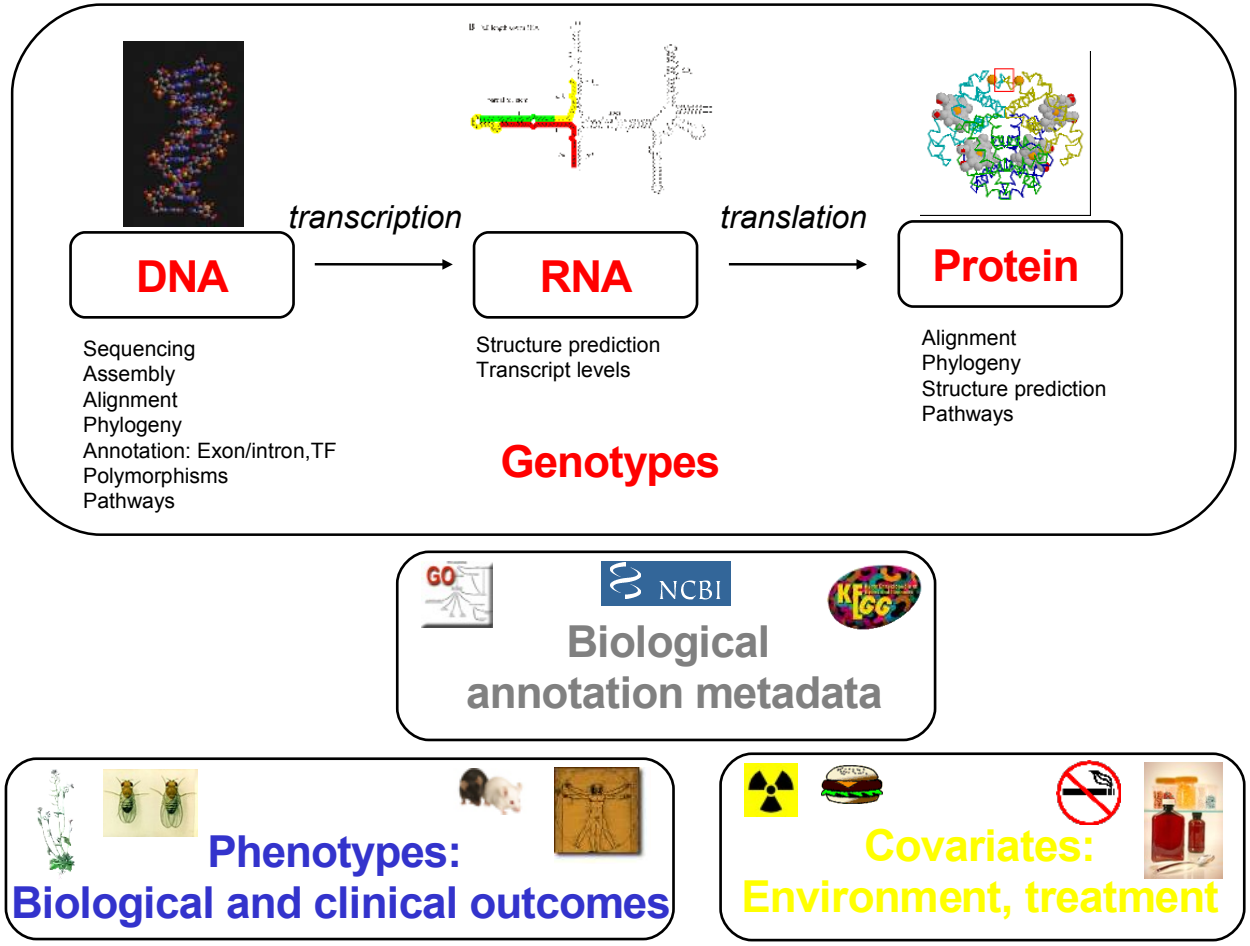

Figure 1: *Biomedical and genomic data.*

Multiple Hypothesis Testing Problems in Genomics

- Inference for high-dimensional multivariate distributions, with complex and unknown dependence structures among variables.
- Broad range of parameters of interest.

  E.g. Regression coefficients in non-linear models relating patient survival data to genome-wide transcript levels, DNA copy numbers, or SNP genotypes;

  measures of association between GO annotation and parameters of the distribution of microarray expression measures;

  pairwise correlation coefficients between transcript levels.
- Many null hypotheses, in the thousands or even millions.
- Complex and unknown dependence structures among test statistics.

  E.g. Directed acyclic graph (DAG) structure of GO terms;

  Galois lattice for multilocus composite SNP genotypes.

Main Contributions: General and Unified Framework

Motivation.

- Large-scale multiple testing problems, e.g., genomics.
- Limitations of existing multiple testing methods, in terms of scope, Type I error rates, marginal nature, distributional assumptions, etc.

Main contributions.

- Foundations of a general and unified methodology for multiple hypothesis testing.
- Resampling-based joint multiple testing procedures (MTP) for controlling a broad class of Type I error rates, such as generalized tail probabilities and generalized expected values for arbitrary functions of the numbers of Type I errors and rejected hypotheses.
- Software implementation: Bioconductor R package multtest.

Main Contributions: General and Unified Framework

General and unified framework for multiple hypothesis testing.

- General definition of null and alternative hypotheses in terms of submodels for the data generating distribution,

$$H_0(m) \equiv \mathrm{I}\left(P \in \mathcal{M}(m)\right) \quad \text{vs.} \quad H_1(m) \equiv \mathrm{I}\left(P \notin \mathcal{M}(m)\right). \qquad (1)$$

- General definition of test statistics and rejection regions.
- General definition of Type I error rates (and power) as arbitrary parameters $\Theta(F_{V_n,R_n})$ of the joint distribution of the numbers of Type I errors V_n and rejected hypotheses R_n.
- General definition of adjusted p-values and parameter confidence regions for arbitrary Type I error rates.

Dudoit and van der Laan (2007, Chapter 1), Dudoit et al. (2004b), Pollard and van der Laan (2004).

Multiple Hypothesis Testing Framework: Error Rates

Table 1: *Type I and Type II errors in multiple hypothesis testing.*

Null hypotheses

	Non-rejected, $\mathcal{R}_n^c$	**Rejected,** $\mathcal{R}_n$	
True, $\mathcal{H}_0$	$W_n = \lvert\mathcal{R}_n^c \cap \mathcal{H}_0\rvert$	$V_n = \lvert\mathcal{R}_n \cap \mathcal{H}_0\rvert$	h_0
False, $\mathcal{H}_1$	$U_n = \lvert\mathcal{R}_n^c \cap \mathcal{H}_1\rvert$	$S_n = \lvert\mathcal{R}_n \cap \mathcal{H}_1\rvert$	h_1
	$M - R_n$	R_n	M

Type I errors: $\mathcal{R}_n \cap \mathcal{H}_0$

Type II errors: $\mathcal{R}_n^c \cap \mathcal{H}_1$

Multiple Hypothesis Testing Framework: Error Rates

Type I error rates. Define a Type I error rate as a parameter $\theta_n = \Theta(F_{V_n,R_n})$ of the joint distribution F_{V_n,R_n} of the numbers of Type I errors $V_n = |\mathcal{R}_n \cap \mathcal{H}_0|$ and rejected hypotheses $R_n = |\mathcal{R}_n|$.

Specifically, consider generalized tail probability (gTP) error rates,

$$gTP(q,g) \equiv \Pr(g(V_n, R_n) > q), \tag{2}$$

and generalized expected value (gEV) error rates,

$$gEV(g) \equiv \mathrm{E}[g(V_n, R_n)], \tag{3}$$

for arbitrary functions $g(V_n, R_n)$ of the numbers of Type I errors V_n and rejected hypotheses R_n.

Multiple Hypothesis Testing Framework: Error Rates

Number of false positives, $g(v, r) = v$.

Generalized family-wise error rate (gFWER):

$$gFWER(k) = \Pr(V_n > k).$$

Per-family error rate (PFER):

$$PFER = \mathrm{E}[V_n].$$

Proportion of false positives among the rejected hypotheses, $g(v, r) = v/r$.

Tail probability for the proportion of false positives (TPPFP):

$$TPPFP(q) = \Pr(V_n/R_n > q).$$

False discovery rate (FDR):

$$FDR = \mathrm{E}[V_n/R_n].$$

Type I error rates based on the proportion of false positives among the rejected hypotheses are particularly appealing for large-scale testing problems, as they do not increase exponentially with the number of tested hypotheses.

Multiple Hypothesis Testing Framework: Adjusted p-Values

As in the case of single hypothesis testing, the results of a multiple testing procedure $\mathcal{R}_n(\alpha)$ are reported in terms of the following quantities.

- Rejection regions for the test statistics.
- Confidence regions for the parameters of interest.
- Adjusted p-values. The adjusted p-value $\widetilde{P}_{0n}(m)$, for null hypothesis $H_0(m)$, is the smallest nominal Type I error level α of the multiple hypothesis testing procedure $\mathcal{R}_n(\alpha)$ at which one would reject $H_0(m)$, given the data. That is,

$$\begin{aligned} \widetilde{P}_{0n}(m) &\equiv \inf\{\alpha \in [0,1] : \text{Reject } H_0(m) \text{ at nominal MTP level } \alpha\} \\ &= \inf\{\alpha \in [0,1] : m \in \mathcal{R}_n(\alpha)\}, \qquad m = 1, \ldots, M. \end{aligned} \tag{4}$$

Multiple Hypothesis Testing Framework: Adjusted p-Values

Reporting the results of a MTP in terms of adjusted p-values, as opposed to only rejection or not of the null hypotheses, offers several advantages.

- Adjusted p-values can be defined for any Type I error rate (e.g., gFWER, TPPFP, or FDR).

  E.g. FWER-controlling single-step Bonferroni (1936) MTP:

$$\widetilde{P}_{0n}(m) = \min\left\{M\, P_{0n}(m), 1\right\}.$$

  FDR-controlling step-up Benjamini and Hochberg (1995) MTP:

$$\widetilde{P}_{0n}(O_n(m)) = \min_{h=m,\ldots,M}\left\{\min\left\{\frac{M}{h}\, P_{0n}(O_n(h)), 1\right\}\right\}.$$

- The smaller the adjusted p-value $\widetilde{P}_{0n}(m)$, the stronger the evidence against the corresponding null hypothesis $H_0(m)$. Thus, one rejects $H_0(m)$ for small adjusted p-values $\widetilde{P}_{0n}(m)$.

Multiple Hypothesis Testing Framework: Adjusted p-Values

- They reflect the strength of the evidence against each null hypothesis in terms of the Type I error rate for the entire MTP.
- They are flexible summaries of a MTP, in the sense that results are supplied for all Type I error levels α, i.e., the level α need not be chosen ahead of time.
- They provide convenient benchmarks to compare different MTPs, whereby smaller adjusted p-values indicate a less conservative procedure.
- Plots of sorted adjusted p-values allow investigators to examine sets of rejected hypotheses associated with various Type I error rates (e.g., gFWER, TPPFP, or FDR) and nominal levels α. Such plots provide tools to decide on an appropriate combination of the number of rejected hypotheses and tolerable false positive rate for a particular experiment and available resources.

Main Contributions: Test Statistics Null Distribution

Test statistics null distribution.

- General characterization of a proper null distribution in terms of null domination conditions for the test statistics for the true null hypotheses.
- Explicit construction of two main types of test statistics null distributions.
  - Null shift and scale-transformed test statistics null distribution, based on user-supplied upper bounds for the means and variances of the test statistics for the true null hypotheses.
  - Null quantile-transformed test statistics null distribution, based on user-supplied marginal test statistics null distributions.
- Resampling procedures (e.g., non-parametric and model-based bootstrap) for consistent estimation of the null distribution and of the corresponding test statistic cut-offs, parameter confidence regions, and adjusted p-values.

Main Contributions: Test Statistics Null Distribution

- Only concerned with controlling the Type I error rate under the true data generating distribution.

  The concepts of weak and strong control of a Type I error rate are therefore irrelevant.

- Directly consider a null distribution for the test statistics rather than a data generating null distribution.

  The latter approach does not necessarily provide proper Type I error control under the true distribution.

✏ Dudoit and van der Laan (2007, Chapter 2), Dudoit et al. (2004b), van der Laan and Hubbard (2006), Pollard and van der Laan (2004).

☞ *... more in Mark van der Laan's presentation.*

Main Contributions: Multiple Testing Procedures

Joint multiple testing procedures.

- Joint single-step common-cut-off and common-quantile procedures for controlling general Type I error rates $\Theta(F_{V_n})$, defined as arbitrary parameters of the distribution of the number of Type I errors V_n.
  E.g. Generalized family-wise error rate (gFWER),
  $gFWER(k) = 1 - F_{V_n}(k) = \Pr(V_n > k)$.
  ✏ Dudoit and van der Laan (2007, Chapter 4), Dudoit et al. (2004b), Pollard and van der Laan (2004).

- Joint step-down common-cut-off (maxT) and common-quantile (minP) procedures for controlling the family-wise error rate (FWER),
  $FWER = gFWER(0) = 1 - F_{V_n}(0) = \Pr(V_n > 0)$.
  ✏ Dudoit and van der Laan (2007, Chapter 5), van der Laan et al. (2004a).

Main Contributions: Multiple Testing Procedures

- (Marginal/joint single-step/stepwise common-cut-off/common-quantile) augmentation multiple testing procedures (AMTP) for controlling generalized tail probability (gTP) error rates, $gTP(q, g) = \Pr(g(V_n, R_n) > q)$, based on an initial gFWER-controlling procedure.
  E.g. gFWER: $g(v, r) = v$; TPPFP: $g(v, r) = v/r$.
  ✏ Dudoit and van der Laan (2007, Chapter 6), Dudoit et al. (2004a), van der Laan et al. (2004b).

- Joint resampling-based empirical Bayes procedures for controlling generalized tail probability and generalized expected value error rates.
  ✏ Dudoit and van der Laan (2007, Chapter 7), van der Laan et al. (2005).

☞ *... more in Mark van der Laan's and Houston Gilbert's presentations.*

$\Theta(F_{V_n})$-Controlling Single-Step Procedures

Three-step road map. We have proposed a road map that leads to

- the general characterization and explicit construction of a proper test statistics null distribution Q_0;
- joint single-step procedures for controlling Type I error rates defined as arbitrary parameters $\Theta(F_{V_n})$ of the distribution of the number of Type I errors V_n (e.g., gFWER).

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

✏ Dudoit and van der Laan (2007, Chapter 4), Dudoit et al. (2004b), Pollard and van der Laan (2004).

$\Theta(F_{V_n})$-Controlling Single-Step Procedures

Procedure 1 [Three-step road map for controlling Type I error rates $\Theta(F_{V_n})$]

1. *Null domination conditions for the Type I error rates $\Theta(F_{V_n})$ and $\Theta(F_{V_0})$.*
   Select a test statistics null distribution Q_0 such that

$$(\limsup_{n\to\infty}) \qquad \Theta(F_{V_n}) \leq \Theta(F_{V_0}). \tag{NDΘ}$$

2. *Monotonicity of the Type I error rate mapping Θ.*

$$\Theta(F_{V_0}) \leq \Theta(F_{R_0}). \tag{5}$$

3. *Control of $\Theta(F_{R_0})$. Select rejection regions so that*

$$\Theta(F_{R_0}) \leq \alpha. \tag{6}$$

$\Theta(F_{V_n})$-Controlling Single-Step Procedures

Specifically, consider single-step procedures with one-sided rejection regions, so that

$$\mathcal{R}_n = \{m : T_n(m) > c(m)\}.$$

Among the family of MTPs that satisfy the Type I error constraint

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

for $R_0 = R(c|Q_0) = \sum_{m=1}^{M} \mathrm{I}(Z(m) > c(m))$ and $Z \sim Q_0$, we have explicitly derived two types of procedures:

- Procedure 2, based on a common cut-off for all test statistics;
- Procedure 3, with common-quantile cut-offs for the test statistics.

$\Theta(F_{V_n})$-Controlling Single-Step Procedures

Procedure 2 [$\Theta(F_{V_n})$-controlling single-step common-cut-off procedure]

For controlling the Type I error rate $\Theta(F_{V_n})$ at level α, the set of rejected null hypotheses is of the form $\mathcal{R}_n = \{m : T_n(m) > c_0\}$, where the common cut-off c_0 is the smallest (i.e., least conservative) value for which $\Theta(F_{R_0}) \leq \alpha$. Adjusted p-values are given by

$$\widetilde{P}_{0n}(m) = \Theta(F_{R(T_n(m)^{(M)}|Q_0)}), \qquad m = 1, \ldots, M, \tag{7}$$

where $T_n(m)^{(M)}$ denotes an M-vector of common cut-offs equal to $T_n(m)$.

$\Theta(F_{V_n})$-Controlling Single-Step Procedures

- gFWER control, i.e., $\Theta(F_{V_n}) = 1 - F_{V_n}(k)$: single-step $T(k+1)$ procedure, based on the $(k+1)$st largest test statistic,

$$\widetilde{p}_{0n}(m) = \Pr_{Q_0}\left(Z^{\circ}(k+1) \geq t_n(m)\right), \qquad m = 1, \ldots, M. \tag{8}$$

- FWER control, i.e., $\Theta(F_{V_n}) = 1 - F_{V_n}(0)$: single-step maxT procedure, based on the maximum test statistic,

$$\widetilde{p}_{0n}(m) = \Pr_{Q_0}\left(\max_{m=1,\ldots,M} Z(m) \geq t_n(m)\right), \qquad m = 1, \ldots, M. \tag{9}$$

$\Theta(F_{V_n})$-Controlling Single-Step Procedures

Procedure 3 [$\Theta(F_{V_n})$-controlling single-step common-quantileprocedure]

For controlling the Type I error rate $\Theta(F_{V_n})$ at level α, the set of rejected null hypotheses is of the form $\mathcal{R}_n = \{m : T_n(m) > c_0(m)\}$, where $c_0(m) = Q_{0,m}^{-1}(\delta_0) = \inf\{z \in \mathbb{R} : Q_{0,m}(z) \geq \delta_0\}$ is the δ_0-quantile of the marginal null distribution $Q_{0,m}$. The common quantile probability δ_0 is chosen as the smallest (i.e., least conservative) value for which $\Theta(F_{R_0}) \leq \alpha$. Adjusted p-values are given by

$$\widetilde{P}_{0n}(m) = \Theta(F_{R(q_0^{-1}(1-P_{0n}(m)))|Q_0)}), \qquad m = 1, \ldots, M, \tag{10}$$

where $P_{0n}(m)$ is the unadjusted p-value for null hypothesis $H_0(m)$,

$$P_{0n}(m) = \bar{Q}_{0,m}(T_n(m)) = 1 - Q_{0,m}(T_n(m)), \tag{11}$$

and $q_0^{-1}(\delta) = (Q_{0,m}^{-1}(\delta) : m = 1, \ldots, M)$ denotes an M-vector of δ-quantiles for the marginal null distributions $Q_{0,m}$.

$\Theta(F_{V_n})$-Controlling Single-Step Procedures

- gFWER control, i.e., $\Theta(F_{V_n}) = 1 - F_{V_n}(k)$: single-step $P(k+1)$ procedure, based on the $(k+1)$st smallest unadjusted p-value,

$$\widetilde{p}_{0n}(m) = \Pr_{Q_0}\left(P_0^\circ(k+1) \leq p_{0n}(m)\right), \qquad m = 1, \ldots, M. \tag{12}$$

- FWER control, i.e., $\Theta(F_{V_n}) = 1 - F_{V_n}(0)$: single-step minP procedure, based on the minimum unadjusted p-value,

$$\widetilde{p}_{0n}(m) = \Pr_{Q_0}\left(\min_{m=1,\ldots,M} P_0(m) \leq p_{0n}(m)\right), \qquad m = 1, \ldots, M. \tag{13}$$

gTP-Controlling Augmentation Procedures

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

Given any initial gFWER-controlling MTP, we have derived AMTPs for controlling generalized tail probability (gTP) error rates,
$gTP(q, g) = \Pr(g(V_n, R_n) > q)$, for arbitrary functions $g(V_n, R_n)$ of the numbers of Type I errors V_n and rejected hypotheses R_n.

gFWER	AMTP	gTP
$\mathcal{R}_n(\alpha)$	$\Longrightarrow$	$\mathcal{R}_n^+(\alpha) = \mathcal{R}_n(\alpha) \cup \mathcal{A}_n(\alpha)$
$\Pr(V_n > k_0) \leq \alpha$		$\Pr(g(V_n^+, R_n^+) > q) \leq \alpha$

✏ Dudoit and van der Laan (2007, Chapter 6), Dudoit et al. (2004a), van der Laan et al. (2004b).

gTP-Controlling Augmentation Procedures

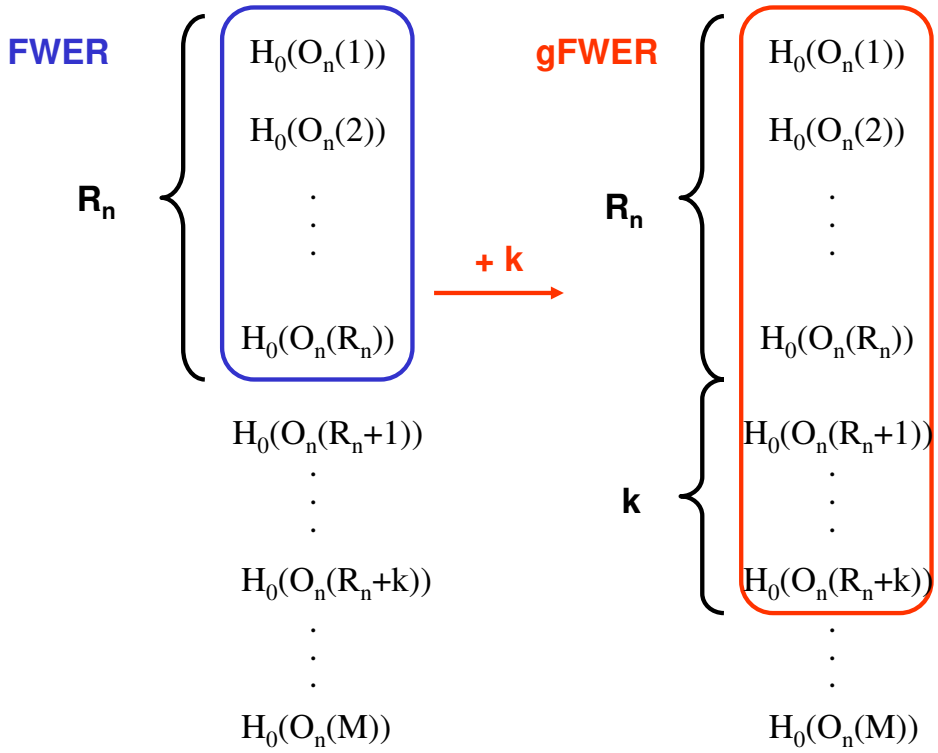

Figure 2: *Augmentation multiple testing procedures.* $gFWER(k)$-control via FWER control.

gTP-Controlling Augmentation Procedures

Procedure 4 [gTP-controlling augmentation multiple testing procedure]

Consider any $gFWER(k_0)$*-controlling procedure* $\mathcal{R}_n(\alpha)$*, with adjusted* p*-values* $\widetilde{P}_{0n}(m)$ *and indices* $O_n(m)$ *so that* $\widetilde{P}_{0n}(O_n(1)) \leq \cdots \leq \widetilde{P}_{0n}(O_n(M))$*. This initial gFWER-controlling procedure rejects the following* $R_n(\alpha) = |\mathcal{R}_n(\alpha)|$ *null hypotheses,*

$$\mathcal{R}_n(\alpha) = \left\{m : \widetilde{P}_{0n}(m) \leq \alpha\right\} = \{O_n(m) : m = 1, \ldots, R_n(\alpha)\}\,.$$

For controlling $gTP(q, g)$ *at level* α*, the augmentation multiple testing procedure rejects the* $R_n(\alpha)$ *null hypotheses specified by the initial gFWER-controlling MTP, as well as the next* $A_n(\alpha)$ *most significant null hypotheses,*

where

$$A_n(\alpha) \equiv \max\{m \in \{0, \ldots, M - R_n(\alpha)\} : g(k_0 + m, R_n(\alpha) + m) \leq q\}. \quad (14)$$

The set of rejected null hypotheses for the gTP-controlling AMTP is

$$\mathcal{R}_n^+(\alpha) \equiv \{O_n(m) : m = 1, \ldots, R_n(\alpha) + A_n(\alpha)\} \quad (15)$$

and the adjusted p-values satisfy

$$\widetilde{P}_{0n}(O_n(m)) = \widetilde{P}_{0n}^+(O_n(S_n(m))), \quad (16)$$

where $S_n : \{1, \ldots, M\} \to \{1, \ldots, M\}$ is an integer shift function defined by

$$S_n(m) \equiv R_n^+(\widetilde{P}_{0n}(O_n(m))) = m + A_n(\widetilde{P}_{0n}(O_n(m))). \quad (17)$$

gTP-Controlling Augmentation Procedures

Intuitively, a gTP-controlling AMTP keeps rejecting null hypotheses until $g(k_0 + m, R_n + m)$ reaches the bound q for false positives.

The adjusted p-values for the AMTP are shifted versions of the adjusted p-values of the initial gFWER-controlling MTP.

gFWER-controlling AMTP. $(g(v, r) = v, k_0 = 0)$

$$\widetilde{P}^{+}_{0n}(O_n(m)) = \begin{cases} 0, & \text{if } m \leq k \\ \widetilde{P}_{0n}(O_n(m-k)), & \text{if } m > k \end{cases}. \tag{18}$$

TPPFP-controlling AMTP. $(g(v, r) = v/r, k_0 = 0)$

$$\widetilde{P}^{+}_{0n}(O_n(m)) = \widetilde{P}_{0n}(O_n(\lceil (1-q)m \rceil)), \qquad m = 1, \ldots, M. \tag{19}$$

gTP-Controlling Augmentation Procedures

- Any gFWER-controlling procedure (marginal/joint single-step/stepwise common-cut-off/common-quantile) provides immediately and trivially MTPs controlling a broad class of Type I error rates, e.g., gFWER and TPPFP.
- One can build on the large pool of available FWER-controlling MTPs, such as the single-step and step-down maxT and minP procedures.
- Adjusted p-values for an AMTP are simply shifted versions of the ordered adjusted p-values for the initial MTP.
- $gFWER(k)$-controlling AMTP guarantees at least k rejected null hypotheses.
- AMTPs augment the set of null hypotheses rejected by an initial MTP conservatively, in the sense that every additional rejected hypothesis is counted as a false positive.
- Unlike many procedures controlling the proportion of false positives, which assume either independence or specific dependence structures for the joint distribution of the test statistics, AMTPs provide gTP control for general data generating distributions, i.e., arbitrary joint distributions for the test statistics.

Resampling-Based Empirical Bayes Procedures

Many commonly-used MTPs share the following two conservative features.

- Test statistics joint distribution. The unknown test statistics joint distribution is replaced by a null distribution that satisfies null domination conditions.
  E.g. $\Theta(F_{V_n}) \leq \Theta(F_{V_0})$ as in Step 1 of the road map of Procedure 1 and $\Theta(F_{V_n})$-controlling single-step Procedures 2 and 3.

- Set of true null hypotheses. The unknown set of true null hypotheses $\mathcal{H}_0$ is replaced by the complete set of null hypotheses $\{1, \ldots, M\}$.
  E.g. $\Theta(F_{V_0}) \leq \Theta(F_{R_0})$ as in Step 3 of the road map of Procedure 1 and $\Theta(F_{V_n})$-controlling single-step Procedures 2 and 3;
  counting every additional rejected hypothesis as a Type I error in Equation (14) of gTP-controlling augmentation Procedure 4;
  controlling FDR at level $(h_0/M)\alpha \leq \alpha$ as in step-up Benjamini and Hochberg (1995) procedure.

Resampling-Based Empirical Bayes Procedures

In order to achieve more power regarding the second point, one can adopt an empirical Bayes approach and generate random guessed sets of true null hypotheses $\mathcal{H}_{0n}$ under a suitable distribution $Q_{0n}^{\mathcal{H}}$.

We have provided a general characterization and explicit constructions for resampling-based empirical Bayes procedures that control generalized tail probability and generalized expected value error rates, e.g., FDR.

✏ Dudoit and van der Laan (2007, Chapter 7), van der Laan et al. (2005).

☞ *... more in Mark van der Laan's and Houston Gilbert's presentations.*

Main Contributions: General and Unified Framework

N. B. Compared to previously-proposed approaches, our multiple testing procedures based on a null-transformed test statistics null distribution offer the following advantages.

- General and unified framework for multiple hypothesis testing.
- Proper Type I error control for general
  - data generating distributions, with arbitrary dependence structures among variables;
  - null hypotheses, defined in terms of submodels for the data generating distribution;
  - test statistics, e.g., t-statistics, χ^2-statistics, F-statistics;
  - Type I error rates, such as, generalized tail probabilities and generalized expected values for arbitrary functions of the numbers of Type I errors and rejected hypotheses.

Main Contributions: General and Unified Framework

- Do not rely on restrictive and questionable assumptions on the joint distribution of the test statistics, such as, independence, Simes' Inequality, subset pivotality.
- Account for the joint distribution of the test statistics $\Longrightarrow$ more power than procedures based solely on marginal distributions, i.e., unadjusted p-values.
- Report results using adjusted p-values.

Main Contributions: General and Unified Framework

Table 2: *Multiple hypothesis testing flowchart.*

Specify data generating distribution and parameters of interest

$$P, \psi = (\psi(j) : j = 1, \ldots, J)$$

$\Downarrow$

Define null and alternative hypotheses

$$H_0(m) = \mathrm{I}(P \in \mathcal{M}(m)) \text{ and } H_1(m) = \mathrm{I}(P \notin \mathcal{M}(m))$$

$\Downarrow$

Specify test statistics

$$T_n = (T_n(m) : m = 1, \ldots, M)$$

$\Downarrow$

Estimate test statistics null distribution

$$Q_{0n}$$

$\Downarrow$

Select Type I error rate

$$\Theta(F_{V_n, R_n})$$

$\Downarrow$

Apply MTP

$\Downarrow$

Summarize results

Adjusted p-values, rejection regions, and confidence regions

Main Contributions: General and Unified Framework

Apply MTP

FWER	$\Pr(V_n > 0)$	Single-step common-cut-off maxT
		Single-step common-quantile minP
		Step-down common-cut-off maxT
		Step-down common-quantile minP
		Resampling-based empirical Bayes
gFWER	$\Pr(V_n > k)$	Single-step common-cut-off $T(k+1)$
		Single-step common-quantile $P(k+1)$
		Augmentation
		Resampling-based empirical Bayes
General	$\Theta(F_{V_n})$	Single-step common-cut-off
		Single-step common-quantile
		Resampling-based empirical Bayes
TPPFP	$\Pr(V_n / R_n > q)$	Augmentation
		Resampling-based empirical Bayes
gTP	$\Pr(g(V_n, R_n) > q)$	Augmentation
		Resampling-based empirical Bayes
FDR	$\mathrm{E}[V_n / R_n]$	TPPFP-based
		Resampling-based empirical Bayes
gEV	$\mathrm{E}[g(V_n, R_n)]$	gTP-based
		Resampling-based empirical Bayes
General	$\Theta(F_{g(V_n, R_n)})$	Resampling-based empirical Bayes

Main Contributions: General and Unified Framework

- Data generating distribution: $P \in \mathcal{M}$.
- Parameters: $\psi = (\psi(j) : j = 1, \ldots, J)$, where $\psi(j) = \Psi(P)(j)$.
- Null and alternative hypotheses: $H_0(m) = \mathrm{I}\left(P \in \mathcal{M}(m)\right)$ and $H_1(m) = \mathrm{I}\left(P \notin \mathcal{M}(m)\right)$, where $\mathcal{M}(m) \subseteq \mathcal{M}, m = 1, \ldots, M$.
- Data and empirical distribution: $\mathcal{X}_n = \{X_i : i = 1, \ldots, n\} \overset{IID}{\sim} P, P_n$.
- Test statistics: $T_n = (T_n(m) : m = 1, \ldots, M)$, where $T_n(m) = T(m; \mathcal{X}_n) = T(m; P_n)$.
- Test statistics null distribution: Q_0 (or estimator thereof, Q_{0n}).
- Multiple testing procedure and rejection regions: $\mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) = \{m : T_n(m) \in \mathcal{C}_n(m)\} = \{m : H_0(m)$ is rejected$\}$.
- Type I error rate: $\theta_n = \Theta(F_{V_n, R_n})$, where $V_n = |\mathcal{R}_n \cap \mathcal{H}_0|$ = # Type I errors and $R_n = |\mathcal{R}_n|$ = # rejected hypotheses.
- Type II error rate/power: $\vartheta_n = \Theta(F_{U_n, R_n})$, where $U_n = |\mathcal{R}_n^c \cap \mathcal{H}_1|$ = # Type II errors.
- Summaries of results: Adjusted p-values, test statistic rejection regions, parameter confidence regions.

Tests of Association with Biological Annotation Metadata

Experimental data, such as microarray gene expression measures, gain much in relevance from their association with biological annotation metadata, i.e., data on data.

E.g. GenBank sequences, GO terms, KEGG pathways, PubMed abstracts.

A challenging and fascinating area of research for statisticians concerns the development of methods for relating experimental data to the wealth of metadata available publicly on the WWW.

Tasks include accessing and pre-processing the data, making inference from these data, and summarizing and interpreting the results.

Dudoit and van der Laan (2007, Chapter 10), Dudoit et al. (2007).

Tests of Association with Biological Annotation Metadata

In this context, an important class of statistical problems involves testing for associations between

- gene-annotation profiles, i.e., known fixed features of a genome,
- gene-parameter profiles, i.e., unknown parameters of the distribution of variable features of this genome in a population of interest.

Here, features of a genome are said to be fixed, if they remain constant among population units. In contrast, variable features are allowed to differ among population units.

The parameter of interest then corresponds to measures of association between known gene-annotation profiles and unknown gene-parameter profiles.

Tests of Association with Biological Annotation Metadata

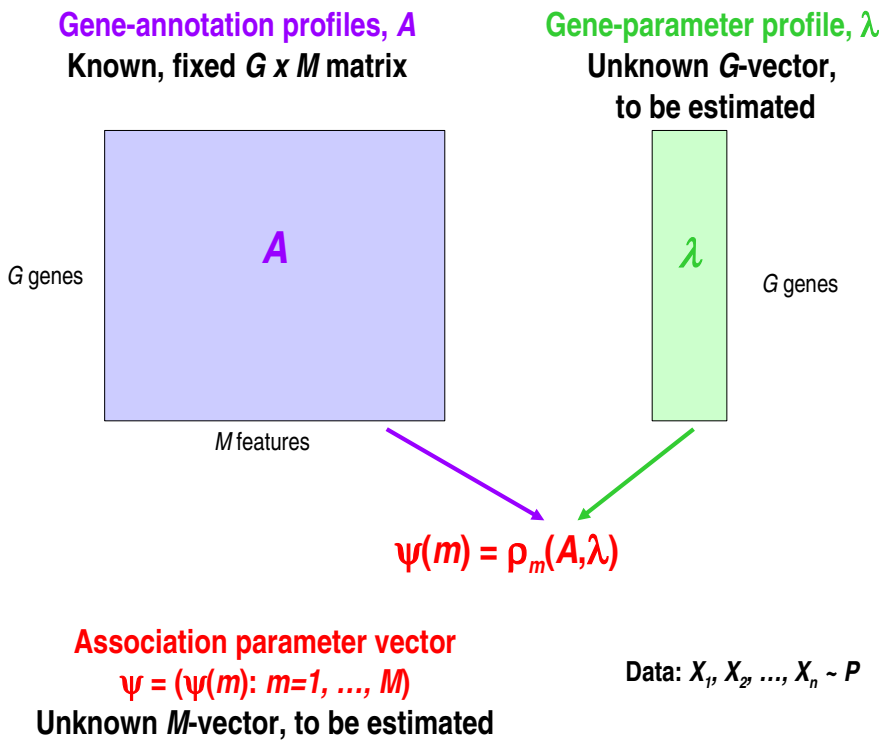

Figure 3: *Parameters for tests of association with biological annotation metadata.*

Tests of Association with Biological Annotation Metadata

Gene-annotation profiles. Gene-annotation profiles refer to features of a genome that are assumed to be known and constant among population units.

Fixed features of interest typically consist of gene annotation metadata, that reflect current knowledge on gene properties, such as, nucleotide and protein sequences, regulation, and function.

E.g. Gene Ontology (GO, `www.geneontology.org`) annotation.
Gene pathway membership (e.g., Kyoto Encyclopedia of Genes and Genomes, KEGG, `www.genome.ad.jp/kegg`).
Gene regulation by particular transcription factors, presence or absence of certain motifs in gene control region (e.g., Transcription Factor DataBase, TRANSFAC, `www.gene-regulation.com`).
Exon/intron counts/lengths/nucleotide distributions.

N. B. Features are fixed in time only for a given version/release of the corresponding database(s).

Tests of Association with Biological Annotation Metadata

Let $A = (A(g,m) : g = 1, \ldots, G;\ m = 1, \ldots, M)$ denote a $G \times M$ gene-annotation matrix, providing data on M features for G genes in an organism of interest.

Row $A(g, \cdot) = (A(g,m) : m = 1, \ldots, M)$ is an M-dimensional gene-specific feature vector for the gth gene.

Column $A(\cdot, m) = (A(g,m) : g = 1, \ldots, G)$ is a G-dimensional gene-annotation profile for the mth feature.

Tests of Association with Biological Annotation Metadata

Gene-parameter profiles. Gene-parameter profiles concern the distribution of variable features of a genome in a well-defined population.

Gene-specific variables of interest reflect cellular type/state/activity under particular conditions and include microarray measures of transcript levels and comparative genomic hybridization (CGH) measures of DNA copy numbers.

E.g. Vector of genome-wide mean transcript levels in a population of heat-shocked yeast cells.
Vector of regression coefficients relating survival to genome-wide transcript levels or DNA copy numbers in a population of cancer patients.

Tests of Association with Biological Annotation Metadata

Let $X = (X(j) : j = 1, \ldots, J) \sim P \in \mathcal{M}$ denote a J-dimensional random vector, with data generating distribution P belonging to a (possibly non-parametric) model $\mathcal{M}$.

Let the parameter mapping $\Lambda : \mathcal{M} \to \mathbb{R}^G$ define a G-dimensional gene-parameter profile, $\Lambda(P) = \lambda = (\lambda(g) : g = 1, \ldots, G) \in \mathbb{R}^G$.

While gene-annotation profiles are known and fixed, gene-parameter profiles are typically unknown and need to be estimated, e.g., from a microarray experiment involving a sample of population units.

Tests of Association with Biological Annotation Metadata

The association parameter of interest is an M-vector

$$\psi = (\psi(m) : m = 1, \ldots, M) = \rho(A, \lambda), \tag{20}$$

of association measures between the gene-annotation profiles A and a gene-parameter profile λ.

In the simplest case, one could define the M association parameters univariately, i.e., let

$$\psi(m) = \rho_m(A(\cdot, m), \lambda), \tag{21}$$

where $\rho_m(\cdot, \cdot)$ provides a measure of association between two G-vectors (e.g., t-statistic, χ^2-statistic, Pearson correlation coefficient).

Tests of Association with Biological Annotation Metadata

Our approach to multiple tests of association with biological annotation metadata differs in a number of important ways from current approaches, such as those developed for inference with GO.

General gene-annotation profiles.

Existing approaches typically consider binary gene-annotation profiles, e.g., vectors of indicators of GO term annotation.

Our general definition of gene-annotation profiles allows consideration of arbitrary qualitative and quantitative fixed features of a genome, e.g., membership of genes to any number of pathways or clusters, exon/intron counts/lengths/nucleotide distributions, mean transcript levels.

Tests of Association with Biological Annotation Metadata

General gene-parameter profiles.

Existing approaches typically consider binary gene-parameter profiles, e.g., vectors of indicators of differential expression.

Our general definition of gene-parameter profiles allows consideration of a much broader class of testing problems, concerning arbitrary qualitative and quantitative parameters, such as, differences in mean expression levels or regression coefficients relating expression levels to clinical outcomes.

Estimated gene-parameter profiles.

Existing approaches typically assume known gene-parameter profiles. For example, the list of DE genes from a microarray experiment is usually treated as known and fixed in subsequent analyses with GO, while in fact it corresponds to an unknown and estimated parameter.

Distinguishing between the definition of a parameter and inference concerning this parameter provides a more rigorous and general formulation of the statistical question.

Tests of Association with Biological Annotation Metadata

General tests of association.

Common approaches to tests of association with GO annotation are typically limited to tests of independence in 2×2 contingency tables (e.g., based on the hypergeometric distribution, Fisher's exact test). Rows correspond to gene annotation with a given GO term (fixed binary gene-annotation profile) and columns to a gene property of interest, such as differential expression (treated as a fixed binary gene-parameter profile).

Our approach allows consideration of a broader class of biological testing problems, while properly accounting for the fact that gene-parameter profiles are usually unknown and replaced by a random (i.e., data-driven) estimator.

Tests of Association between GO Annotation and DE in ALL

Our proposed approach to tests of association with biological annotation metadata is illustrated using the acute lymphoblastic leukemia (ALL) microarray dataset of Chiaretti et al. (2004), with the aim of relating Gene Ontology (GO) annotation to differential expression (DE) among ALL samples.

The BCR/ABL fusion is the molecular analogue of the Philadelphia chromosome, one of the most frequent cytogenetic abnormalities in human leukemias. A number of recent articles have investigated the prognostic relevance of the BCR/ABL fusion in adult ALL of the B-cell lineage.

Consider the following two related questions.

- Identifying differentially expressed genes between B-cell ALL with the BCR/ABL fusion and cytogenetically normal NEG B-cell ALL.
- Identifying GO terms associated with differential expression.

Tests of Association between GO Annotation and DE in ALL

Acute lymphoblastic leukemia dataset. The Chiaretti et al. (2004) ALL dataset comprises, for each of 128 ALL cell samples,

- 12,625 microarray expression measures (Affymetrix chip series HG-U95Av2);
- 21 phenotypes (i.e., covariates and outcomes).

Focus on

- $n = 79$ B-cell ALL cell samples of the BCR/ABL and NEG molecular types;
- $G = 2,071$ filtered genes with unique Entrez Gene IDs.

Tests of Association between GO Annotation and DE in ALL

The Gene Ontology. The Gene Ontology (GO) Consortium (www.geneontology.org) provides ontologies, i.e., structured and controlled vocabularies, to describe gene products in terms of their associated biological processes (BP), cellular components (CC), and molecular functions (MF).

For each of the three ontologies, GO terms are organized in a directed acyclic graph (DAG), i.e., a directed graph (one-way edges) containing no cycles (no path starts and ends at the same vertex).

The GO Consortium and other organizations provide mappings between GO terms and genes in various organisms.

Tests of Association between GO Annotation and DE in ALL

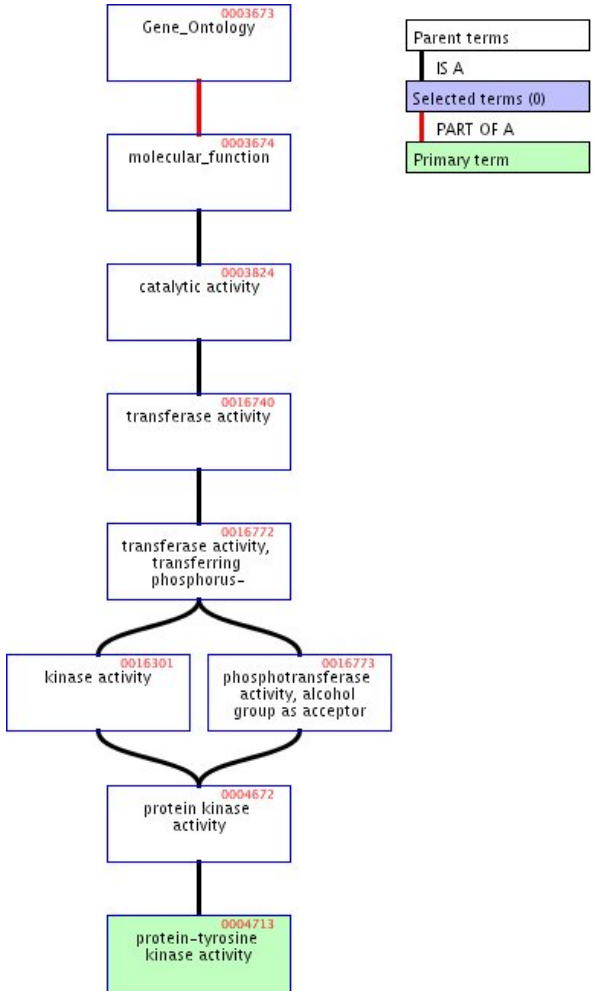

Figure 4: *Gene Ontology.* Portion of the MF DAG for the GO term *protein-tyrosine kinase activity* (GO:0004713). EBI QuickGO browser (www.ebi.ac.uk/ego).

Tests of Association between GO Annotation and DE in ALL

GO gene-annotation profiles. For each of the three gene ontologies, assemble a $G \times M$ binary gene-annotation matrix A, indicating for each gene g whether it is annotated with each GO term m,

$$A(g,m) = \begin{cases} 1, & \text{if gene } g \text{ is annotated with GO term } m \\ 0, & \text{otherwise} \end{cases}.$$

Consider only the M GO terms annotating at least 10 of the $G = 2,071$ filtered genes.

Ontology	**Number of terms,** M
Biological Process	367
Cellular Component	81
Molecular Function	185

Tests of Association between GO Annotation and DE in ALL

Null hypotheses. Test the M null hypotheses of no association between GO gene-annotation profiles $A(\cdot, m)$ and a DE gene-parameter profile λ.

	Testing scenario		
	Continuous		**Binary**
	MT[t, t]	**MT[d, t]**	**MT[$\neq, \chi$]**
DE gene-parameter, λ	Standardized	Unstandardized	Indicator
	difference of means		
GO gene-annotation, A	Indicator		Indicator
DE and GO association, ρ	t-statistic		χ^2-statistic
Null hypotheses, $H_0(m)$	$\rho(A(\cdot, m), \lambda) = 0$		$\rho(A(\cdot, m), \lambda) \leq 1$

Tests of Association between GO Annotation and DE in ALL

Testing scenarios MT[t, t] and MT[d, t].

- DE gene-parameter profile, λ: Continuous
  Standardized/Unstandardized differences of means in BCR/ABL vs. NEG B-cell ALL
  Estimator: Two-sample Welch t-statistics/differences of empirical means
- GO gene-annotation profiles, A: Binary
- DE and GO association measure, ρ: Two-sample Welch t-statistic

Testing scenario MT[$\neq, \chi$]. (The usual approach.)

- DE gene-parameter profile, λ: Binary
  Indicators of DE between BCR/ABL and NEG B-cell ALL
  Estimator: Based on adjusted p-values for FWER-controlling permutation-based single-step maxT procedure, with a pre-specified proportion of DE genes (MT[$\neq, \chi : \gamma G$]) or significance level (MT[$\neq, \chi : \alpha$])
- GO gene-annotation profiles, A: Binary
- DE and GO association measure, ρ: χ^2-statistic

Tests of Association between GO Annotation and DE in ALL

Test statistics. Unstandardized difference statistics:

$$T_n(m) = \sqrt{n}(\psi_n(m) - \psi_0(m)).$$

Test statistics null distribution. Non-parametric bootstrap estimator of the null shift-transformed test statistics null distribution, $B = 5,000$.

Multiple testing procedure. FWER-controlling single-step maxT procedure.

Tests of Association between GO Annotation and DE in ALL

Results: DE between BCR/ABL and NEG B-cell ALL. Two-sided tests using two-sample Welch t-statistics and FWER-controlling bootstrap-based single-step maxT MTP.

- 16 DE genes at nominal FWER level $\alpha = 0.05$.
- DE genes tend to be over-expressed in cell samples with the BCR/ABL fusion (14/16 positive t-statistics).
- The ABL1 gene shows the most over-expression in BCR/ABL cell samples.
- DE genes appear to be related to apoptosis or oncogenesis.

Tests of Association between GO Annotation and DE in ALL

Results: Association between GO annotation and DE in ALL.

- Adjusted p-values tend to be quite large, with only a handful of GO terms identified as being significantly associated with BCR/ABL vs. NEG DE.
- Little overlap between the binary and continuous testing scenarios.
- Testing scenarios based on binary DE gene-parameter profiles tend to be more conservative than scenarios based on continuous profiles and lack robustness with respect to the somewhat arbitrary DE/non-DE gene dichotomization (i.e., the number of DE genes).
- Testing scenarios based on standardized and unstandardized continuous DE gene-parameter profiles lead to very similar results.
- GO terms associated with BCR/ABL vs. NEG DE tend to concentrate in certain branches of the DAGs.
- Some of the genes annotated with the identified GO terms have been linked to the BCR/ABL proto-oncogene and have been suggested as potential targets for molecular therapies of leukemia.

Tests of Association between GO Annotation and DE in ALL

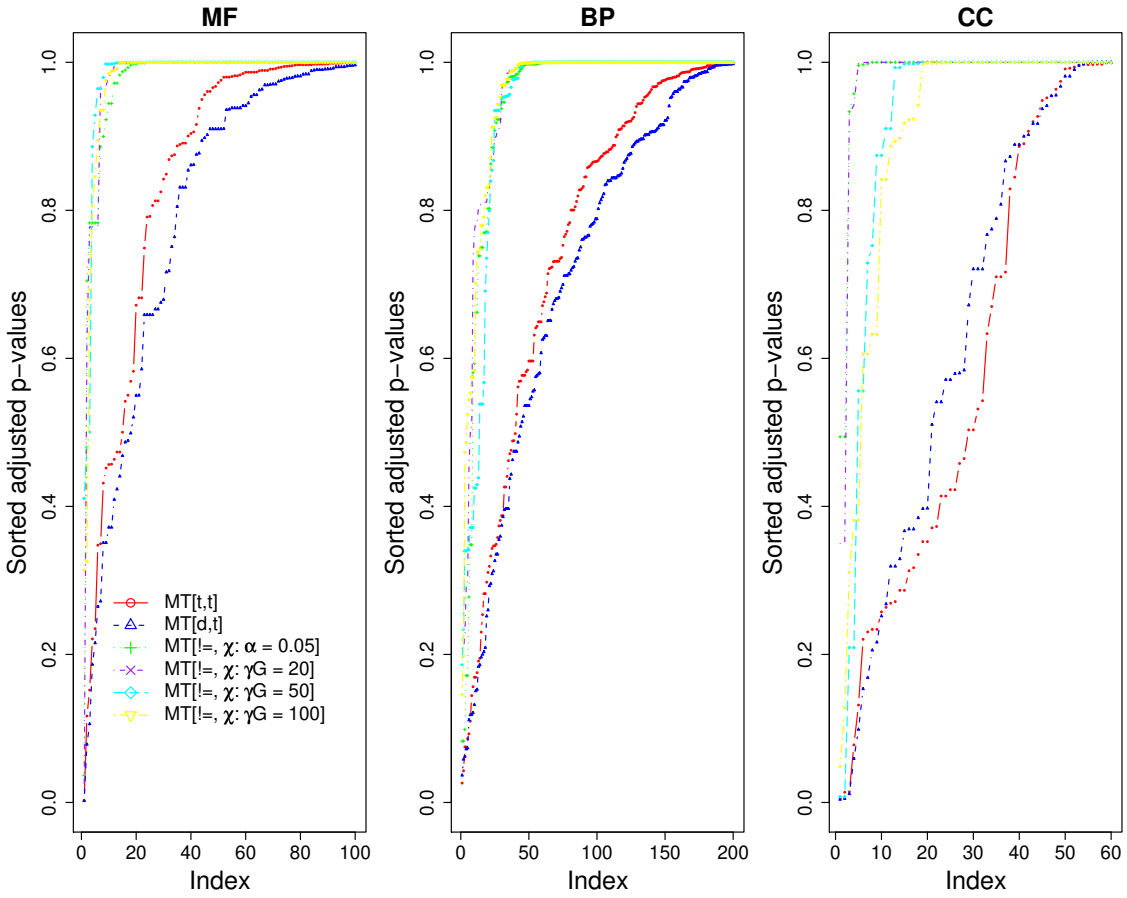

Figure 5: *GO terms associated with BCR/ABL vs. NEG DE, adjusted p-values.*

Tests of Association between GO Annotation and DE in ALL

Table 3: *GO terms associated with BCR/ABL vs. NEG DE.*

	Nominal FWER level, α								
	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
MT[t, t**]**	2	6	14	3	4	5	1	1	3
MT[d, t**]**	1	5	16	3	5	7	1	2	4
MT[$\neq, \chi : \alpha = 0.05$**]**	0	3	5	0	0	0	1	1	1
MT[$\neq, \chi : \gamma G = 20$**]**	0	0	0	0	0	0	1	1	1
MT[$\neq, \chi : \gamma G = 50$**]**	0	0	1	2	2	2	0	0	0
MT[$\neq, \chi : \gamma G = 100$**]**	0	0	2	1	1	2	0	0	0
	BP			**CC**			**MF**		

Tests of Association between GO Annotation and DE in ALL

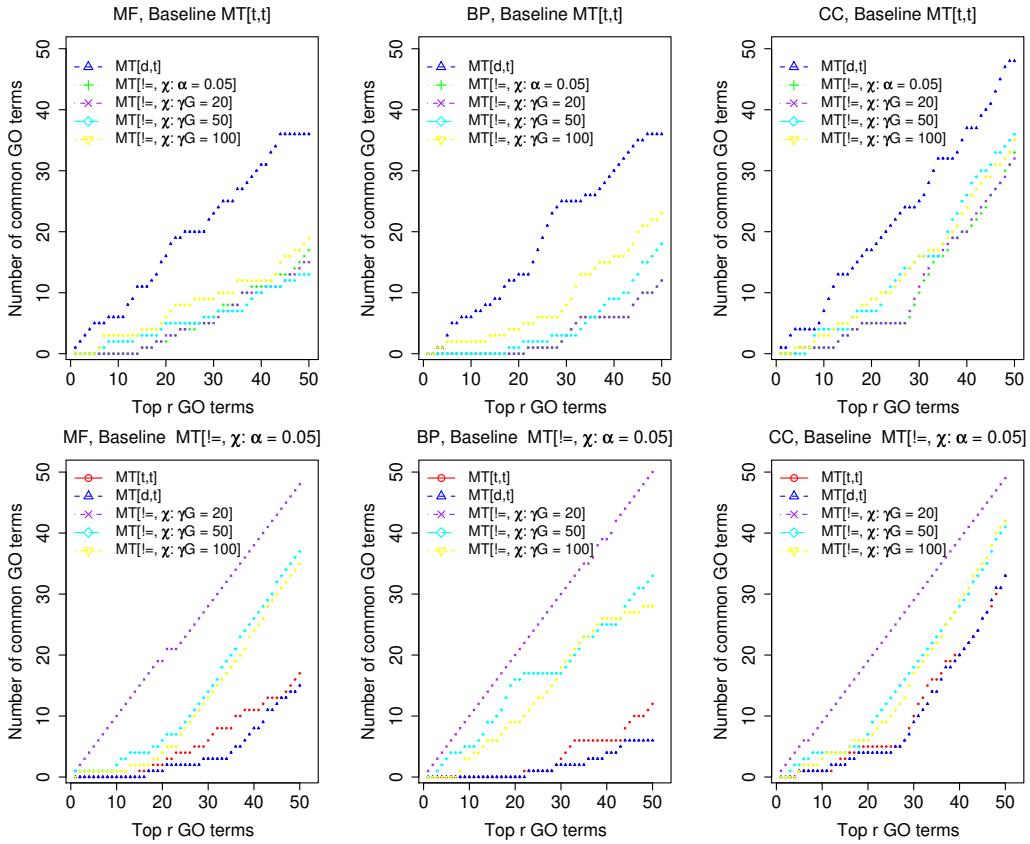

Figure 6: *GO terms associated with BCR/ABL vs. NEG DE, common terms between testing scenarios.*

Tests of Association between GO Annotation and DE in ALL

Table 4: *GO terms associated with BCR/ABL vs. NEG DE, top 20 BP GO terms.*

BP, Scenario MT[t, t]

GO term ID	GO term	$A_1(m)$	$\widetilde{P}_{0n}(m)$
GO:008152	*metabolism*	1076	2.6e-02
GO:044237	*cellular metabolism*	1045	4.3e-02
GO:009058	*biosynthesis*	187	7.5e-02
GO:044238	*primary metabolism*	1002	7.5e-02
GO:044249	*cellular biosynthesis*	169	8.6e-02
GO:006091	*generation of precursor metabolites and energy*	98	9.3e-02
GO:019882	*antigen presentation*	15	1.1e-01
GO:030333	*antigen processing*	14	1.4e-01
GO:006916	*anti-apoptosis*	21	1.6e-01
GO:043066	*negative regulation of apoptosis*	26	1.7e-01
GO:043069	*negative regulation of programmed cell death*	26	1.7e-01
GO:007154	*cell communication*	390	1.8e-01
GO:006457	*protein folding*	52	1.9e-01
GO:007165	*signal transduction*	351	1.9e-01
GO:000226	*microtubule cytoskeleton organization and biogenesis*	14	2.3e-01
GO:006082	*organic acid metabolism*	65	2.5e-01
GO:006163	*purine nucleotide metabolism*	29	2.8e-01
GO:007155	*cell adhesion*	59	2.8e-01
GO:007028	*cytoplasm organization and biogenesis*	10	3.0e-01
GO:019752	*carboxylic acid metabolism*	63	3.1e-01

Tests of Association between GO Annotation and DE in ALL

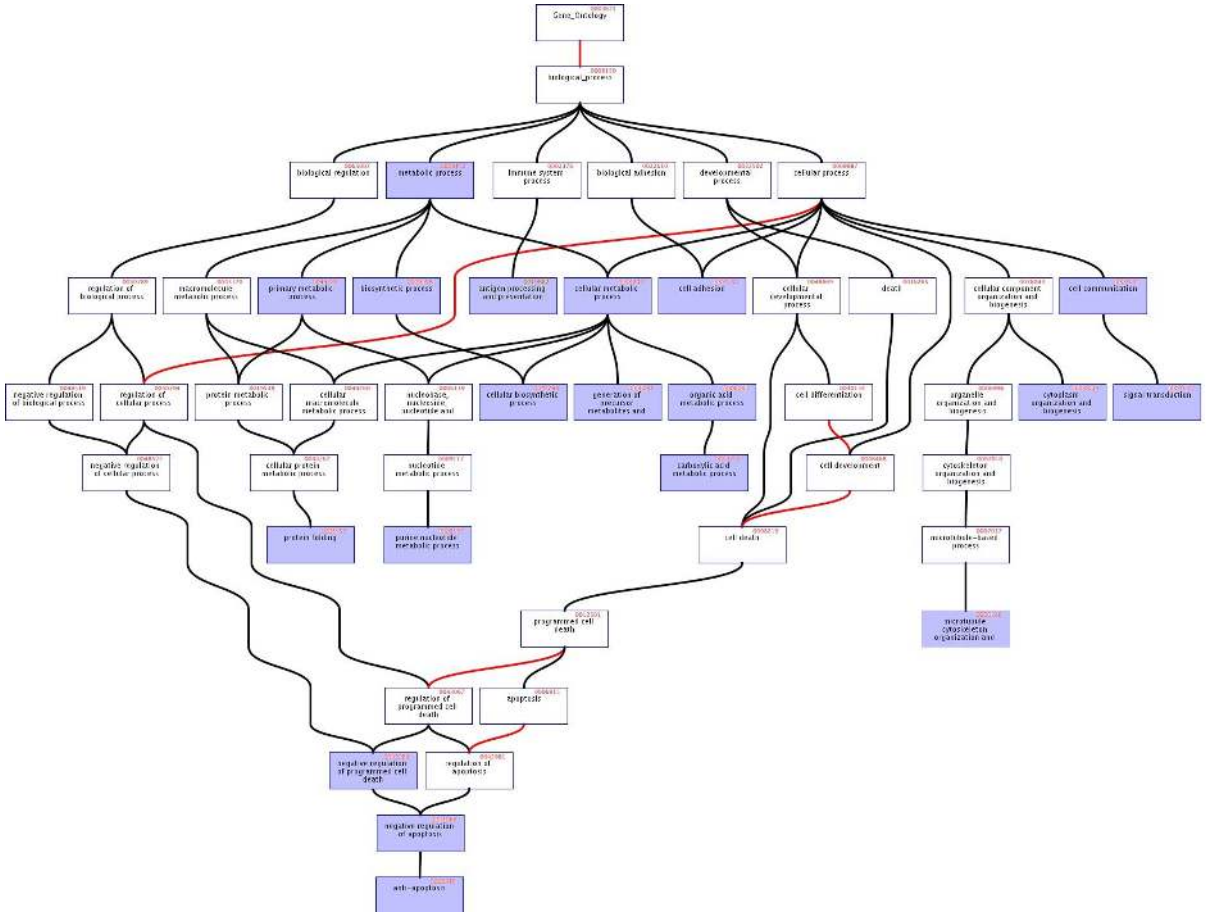

Figure 7: *GO terms associated with BCR/ABL vs. NEG DE, DAG for top 20 BP GO terms.*

Tests of Association between GO Annotation and DE in ALL

Table 5: *GO terms associated with BCR/ABL vs. NEG DE, top 20 CC GO terms.*

CC, Scenario MT[t, t]

GO term ID	GO term	$A_1(m)$	$\widetilde{P}_{0n}(m)$
GO:0005840	*ribosome*	25	5.6e-03
GO:0030529	*ribonucleoprotein complex*	77	1.4e-02
GO:0005830	*cytosolic ribosome (sensu Eukaryota)*	11	1.4e-02
GO:0043234	*protein complex*	334	7.8e-02
GO:0005886	*plasma membrane*	200	1.3e-01
GO:0005829	*cytosol*	78	2.2e-01
GO:0005737	*cytoplasm*	578	2.3e-01
GO:0005887	*integral to plasma membrane*	125	2.3e-01
GO:0031226	*intrinsic to plasma membrane*	125	2.3e-01
GO:0019866	*inner membrane*	37	2.6e-01
GO:0005743	*mitochondrial inner membrane*	28	2.6e-01
GO:0005746	*mitochondrial electron transport chain*	11	2.7e-01
GO:0000502	*proteasome complex (sensu Eukaryota)*	26	2.7e-01
GO:0000323	*lytic vacuole*	28	2.9e-01
GO:0005764	*lysosome*	28	2.9e-01
GO:0005576	*extracellular region*	54	3.1e-01
GO:0005773	*vacuole*	29	3.2e-01
GO:0005622	*intracellular*	1152	3.4e-01
GO:0043228	*non-membrane-bound organelle*	218	3.5e-01
GO:0043232	*intracellular non-membrane-bound organelle*	218	3.5e-01

Tests of Association between GO Annotation and DE in ALL

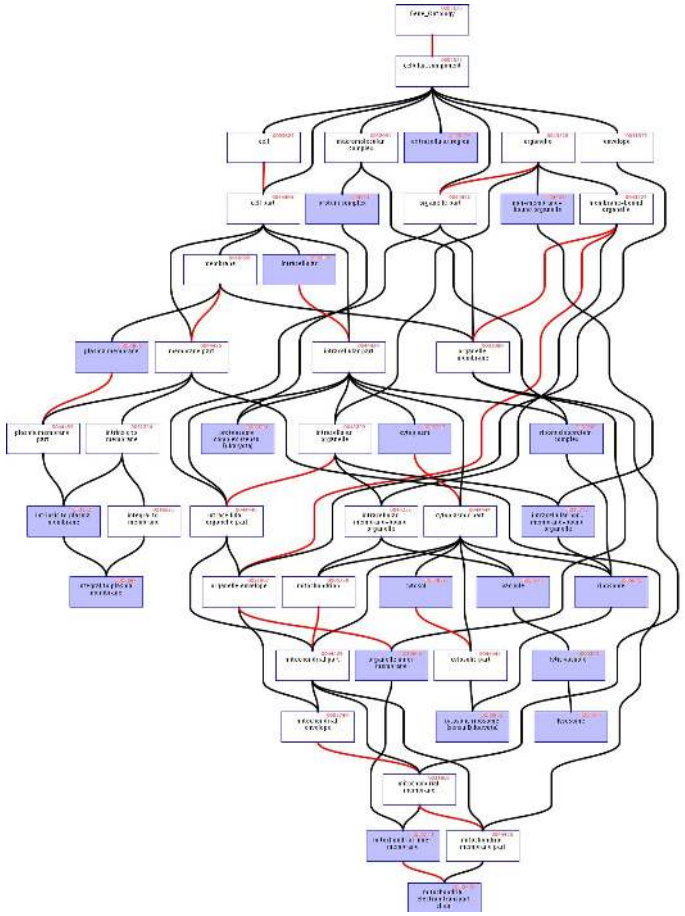

Figure 8: *GO terms associated with BCR/ABL vs. NEG DE, DAG for top 20 CC GO terms.*

Tests of Association between GO Annotation and DE in ALL

Table 6: *GO terms associated with BCR/ABL vs. NEG DE, top 20 MF GO terms.*

MF, Scenario MT[t, t]

GO term ID	GO term	$A_1(m)$	$\widetilde{P}_{0n}(m)$
GO:0003735	*structural constituent of ribosome*	24	2.4e-03
GO:0003723	*RNA binding*	143	1.2e-01
GO:0048037	*cofactor binding*	11	1.5e-01
GO:0051082	*unfolded protein binding*	47	2.2e-01
GO:0016853	*isomerase activity*	28	2.3e-01
GO:0016491	*oxidoreductase activity*	89	3.5e-01
GO:0005509	*calcium ion binding*	69	3.5e-01
GO:0015399	*primary active transporter activity*	57	4.3e-01
GO:0004872	*receptor activity*	101	4.5e-01
GO:0004871	*signal transducer activity*	242	4.6e-01
GO:0016765	*transferase activity, transferring alkyl or aryl (other than methyl) groups*	10	4.6e-01
GO:0016860	*intramolecular oxidoreductase activity*	13	4.6e-01
GO:0016614	*oxidoreductase activity, acting on CH-OH group of donors*	18	4.7e-01
GO:0016616	*oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor*	18	4.7e-01
GO:0043169	*cation binding*	230	5.0e-01
GO:0005489	*electron transporter activity*	47	5.4e-01
GO:0005386	*carrier activity*	73	5.5e-01
GO:0004888	*transmembrane receptor activity*	59	5.7e-01
GO:0003824	*catalytic activity*	635	5.8e-01
GO:0003676	*nucleic acid binding*	449	6.7e-01

Tests of Association between GO Annotation and DE in ALL

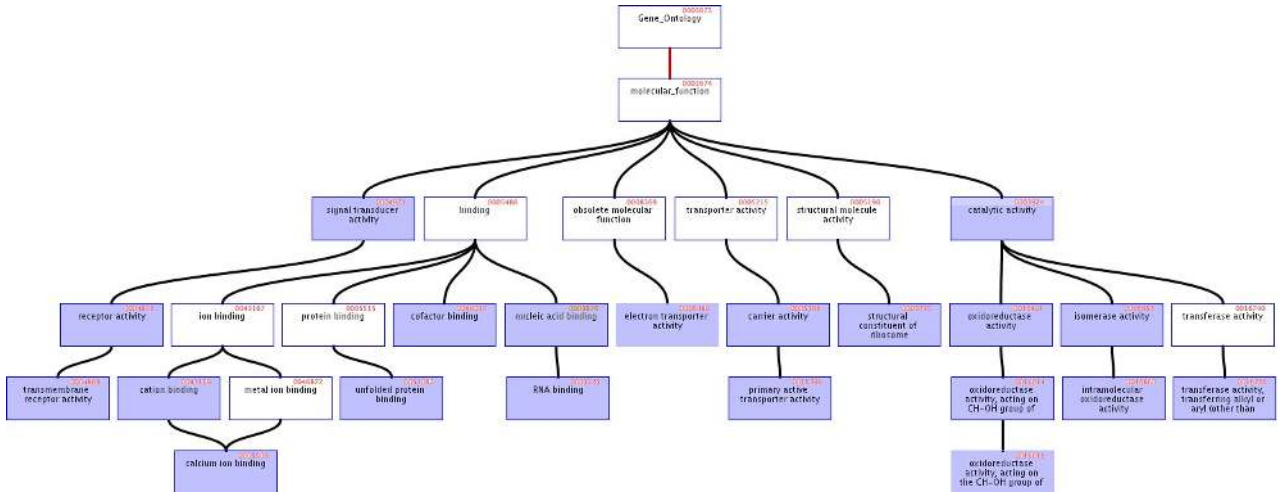

Figure 9: *GO terms associated with BCR/ABL vs. NEG DE, DAG for top 20 MF GO terms.*

Tests of Association between GO Annotation and DE in ALL

Conclusions.

- Choice of gene-parameter profile λ for measuring differential expression between BCR/ABL and NEG B-cell ALL has a large impact on the list of identified GO terms.
- Testing scenarios based on binary DE gene-parameter profiles – still the norm for combined GO annotation and microarray data analyses – have clear limitations.
- Importance of defining proper parameters, i.e., GO gene-annotation profiles A, DE gene-parameter profile λ, and association measure ρ for GO annotation and DE.

Tests of Association between GO Annotation and DE in ALL

Ongoing efforts.

- Other Type I error rates and multiple testing procedures, e.g., gTP- and gEV-controlling resampling-based empirical Bayes procedures.
- More general and biologically pertinent multivariate association measures ρ, that take into account the DAG structure of GO terms by considering the gene-annotation profiles of offspring or ancestor terms.
- Better numerical and graphical approaches for representing and interpreting the multiple testing results, e.g., the lists of GO terms and associated adjusted p-values.
- Software implementation in Bioconductor R package.

Tests of Association between GO Annotation and DE in ALL

Pre-processing and filtering.

- Three-step robust multichip average (RMA) pre-processing for all 128 ALL samples (Bolstad et al., 2005).
- Base 2 logarithmic transformation.
- Intensity-based filtering (von Heydebreck et al., 2004). Retain only probes with: (i) fluorescence intensities greater than 100 (absolute scale) for at least 25% of the 79 cell samples and (ii) interquartile range (IQR) of the fluorescence intensities for the 79 cell samples greater than 0.5 (log base 2 scale).
- Average the expression measures of multiple probes mapping to the same gene (i.e., same Entrez Gene ID).

Tests of Association between GO Annotation and DE in ALL

R and Bioconductor packages. (R Release 2.2.1; Bioconductor Release 1.7)

- multtest (Version 1.8.0): Resampling-based multiple testing procedures.
- ALL (Version 1.0.2): Microarray expression measures and phenotypes for Chiaretti et al. (2004) ALL study.
- annotate (Version 1.8.0): General annotation software package.
- annaffy (Version 1.2.0): Annotating and generating HTML reports for Affymetrix chip data.
- hgu95av2 (Version 1.10.0): Affymetrix chip-specific metadata package.
- GO (Version 1.10.0): GO-specific metadata package.

Software Implementation: Bioconductor R Package multtest

The multiple testing procedures developed in Dudoit and van der Laan (2007) and related articles are implemented in the R package multtest, released as part of the Bioconductor Project, an open-source software project for the analysis of biomedical and genomic data.

Please consult the package documentation (e.g., helpfiles, manuals) and book chapters for details.

Bioconductor R package: multtest.
Authors: Katherine S. Pollard, Yongchao Ge, and Sandrine Dudoit.
URL: www.bioconductor.org.


Dudoit and van der Laan (2007, Section 13.1), Pollard et al. (2005).

Software Implementation: Bioconductor R Package multtest

Test statistics. t-statistics for tests of regression coefficients in linear models and Cox proportional hazards survival models;
F-statistics for tests of equality of means in one-way and two-way designs.

Weighted and robust rank-based versions of the above test statistics are implemented.

Test statistics null distribution. Bootstrap null shift and scale-transformed; permutation.

Software Implementation: Bioconductor R Package multtest

Multiple testing procedures.

- FWER control.
  - Marginal single-step Bonferroni (1936), step-down Holm (1979), and step-up Hochberg (1988).
  - Joint single-step maxT and minP (Ch. 4 in Dudoit and van der Laan, 2007; Dudoit et al., 2004b; Pollard and van der Laan, 2004).
  - Joint step-down maxT and minP (Ch. 5 in Dudoit and van der Laan, 2007; van der Laan et al., 2004a).
- gFWER and TPPFP control. Augmentation multiple testing procedures (Ch. 6 in Dudoit and van der Laan, 2007; van der Laan et al., 2004b).
- FDR control.
  - Marginal step-up Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001).
  - TPPFP-based (Ch. 6 in Dudoit and van der Laan, 2007; van der Laan et al., 2004b).

Software Implementation: Bioconductor R Package multtest

- Numerical summaries. Parameter estimates; test statistics; unadjusted and adjusted p-values; test statistic cut-offs; parameter confidence regions; estimated test statistics null distribution.
- Graphical summaries. Type I error rates vs. # rejections; # rejections vs. adjusted p-values; adjusted p-values vs. test statistics ("volcano" plots).
- Software design.
  - Function closure. Allow uniform data input for all MTPs; facilitate the extension of the package's functionality, by implementing, for example, new types of test statistics.
  - Class/method object-oriented programming. Represent and operate on the results of multiple testing procedures.

Software Implementation: SAS Macros

SAS macros are available to compute the following components of a MTP:

- t-statistics;
- non-parametric bootstrap estimates of the null shift and scale-transformed test statistics null distribution;
- adjusted p-values for the FWER-controlling single-step maxT procedure;
- adjusted p-values for the gFWER- and TPPFP-controlling augmentation procedures.

Author: M. D. Birkner.

URL: `www.stat.berkeley.edu/~sandrine/MTBook`.

Dudoit and van der Laan (2007, Section 13.2), Birkner et al. (2005).

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

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

M. D. Birkner, K. S. Pollard, M. J. van der Laan, and S. Dudoit. Multiple testing procedures and applications to genomics. Technical Report 168, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720-7360, 2005. URL `www.bepress.com/ucbbiostat/paper168`.

B. M. Bolstad, R. A. Irizarry, L. Gautier, and Z. Wu. Preprocessing high-density oligonucleotide arrays. In R. C. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 2, pages 13–32. Springer, New York, 2005. URL `www.bioconductor.org/pubs/docs/mogr`.

C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, pages 3–62, 1936.

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 and M. J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer, New York, 2007.

S. Dudoit, M. J. van der Laan, and M. D. Birkner. Multiple testing procedures for controlling tail probability error rates. Technical Report 166, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720-7360, 2004a. URL www.bepress.com/ucbbiostat/paper166.

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, 2004b. URL www.bepress.com/sagmb/vol3/iss1/art13.

S. Dudoit, S. Keleş, and M. J. van der Laan. Multiple tests of association with biological annotation metadata. In D. Nolan and T. P. Speed, editors, *A Festschrift for David Freedman*, Institute of Mathematical Statistics, Lecture Notes – Monograph Series. 2007. (To appear).

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. *Scandinavian Journal of Statistics*, 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.

K. S. Pollard, S. Dudoit, and M. J. van der Laan. Multiple testing procedures: The multtest package and applications to genomics. In R. C. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 15, pages 249–271. Springer, New York, 2005. URL `www.bioconductor.org/pubs/docs/mogr`, `www.bepress.com/ucbbiostat/paper164`.

M. J. van der Laan and A. E. Hubbard. Quantile-function based null distribution in resampling based multiple testing. *Statistical Applications in Genetics and Molecular Biology*, 5(1):Article 14, 2006. URL `www.bepress.com/sagmb/vol5/iss1/art14`.

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, 2004a. URL `www.bepress.com/sagmb/vol3/iss1/art14`.

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, 2004b. URL www.bepress.com/sagmb/vol3/iss1/art15.

M. J. van der Laan, M. D. Birkner, and A. E. Hubbard. Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 29, 2005. URL www.bepress.com/sagmb/vol4/iss1/art29.

A. von Heydebreck, W. Huber, and R. Gentleman. Differential expression with the Bioconductor Project. Technical Report 7, Bioconductor Project Working Papers, 2004. URL www.bepress.com/bioconductor/paper7.

